第 2 回長期大型研究助成

奈良県立医科大学

「生体 5 次元インフォマティクス国際統合ハブ」

～「生体5次元情報」を解読する医工計測技術を創出する「知・もの・人」づくり計画
Virtual Human InformatiX (V-iCliniX) Project～

最終成果報告書

公益財団法人
中谷医工計測技術振興財団

中谷医工計測技術振興財団

第2回長期大型研究助成　奈良県立医科大学
最終成果報告書

目　　次

1．代表ご挨拶 …… 1

2．参画機関 …… 5

3．総括 …… 9

3－Ⅰ．V-iCliniX Concept「知・もの・人」づくり …… 11

3－Ⅱ．V-iCliniX Organization …… 12

3－Ⅲ．Timeline（プロジェクト時系列）…… 12

3－Ⅳ．「知」の成果…… 13

3－Ⅴ．「もの」の成果…… 16

3－Ⅵ．「人」の成果…… 16

3－Ⅶ．Beyond V-iCliniX …… 23

4．謝辞 …… 25

5．Appendix …… 29

5－1　発表論文

5－1－1　佐藤教授グループ …… 31

5－1－2　森准教授グループ …… 141

5－2　シンポジウム資料

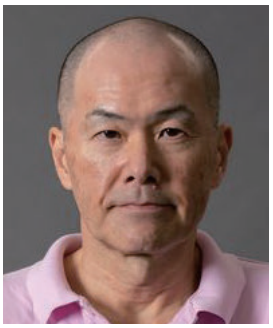
5－2－1　キックオフシンポジウム …… 255

5－2－2　中間成果発表会 …… 259

5－2－3　最終成果報告会 …… 281

1．代 表 ご 挨 拶

ご挨拶



21世紀初頭にヒト全ゲノムがシーケンスされ、ヒトのブループリントが明かになりました。そして、現在、このブループリントに刻まれている生体内情報を解読し、そしてその情報を制御し、書き換えることで、これまでのヒトの歴史における時間スケールをはるかに超える健康長寿の実現を目指して様々な研究がグローバルに展開されています。

私は、これを実現するためには、数学や数理情報学（生体情報の解読)、エンジニアリング（生体情報を制御し書き換える）の概念、方法論、そしてツールを生命科学・医学へ導入することが必須であり、これからは当たり前であると信じていています。

さらに、生体内情報をこれまでの４次元（時空間）スケールをさらに拡張した多次元での場における生体内情報と捉えることが重要と考えています。そこで、本研究計画では、生体内の時空間へ時空間同士を繋ぐもう一つの軸を加えた生体５次元情報を、医工計測技術を用いて解読し制御することを目指した研究計画を提案しました。

そして、その提案が中谷医工計測技術振興財団に評価され、本研究計画の実施が実現しました。研究においては、ただ研究するだけでは、その限られた期間のみの研究に終わってしまいます。研究とは、長期間継続されることで、その真の価値が見えてきます。また、その価値が社会的なインパクトというフットプリントをつけるまではさらに時間がかかります。したがって、本研究計画では、助成期間の５年間という時間を超えて、本研究計画が継続され発展できるよう、若手の研究者そして研究者の卵、さらに学生の育成も主たる柱のひとつとして展開してきました。

この期間中、特に最初のプロジェクト立ち上げには数多くの試行錯誤がありましたが、中谷医工計測技術振興財団の関係者各位、審査委員会の先生方、奈良県立医科大学の同僚そしてバックヤードの方々、共同研究機関の研究者そしてそれぞれの機関のバックヤードの方々、数多くの方々に支えられ協力して頂くことで、無事任期を終えることができました。

私は、この５年間は、あくまで新たな取り組みのキックオフ期間だったと考えております。この５年間の失敗も含めて得られたことを基盤にこれからはしっかりした土台を作り、その上へ着実に積み木を積み上げるといったバトンがBeyond（次の時期・取り組み）へ渡せたと信じています。

本報告書では、これまで５年間、本研究計画が何にどのように取り組んだかが幅広い方々へお伝えできればと存じます。今後、５年、10年、20年、その先も、このプロジェクトの科学的、技術的、人的な子孫、そしてそれらが社会で生み出す価値観（価値そのものよりは価値観が重要と信じています）をこれからも見守って頂けたらと存じます。

以上をもって挨拶とさせて頂きます。

佐藤匠徳（ナルトク）（Thomas N. Sato）

２．参　画　機　関

研 究 拠 点：

奈良県立医科大学Ｖ-iCliniX講座

研究代表者：

佐藤匠徳（Ｔｈｏｍａｓ Ｎ. Ｓａｔｏ）

共同研究機関：

国際電気通信基礎技術研究所（ＡＴＲ）

東京工業大学

国立研究開発法人物質・材料研究機構（ＮＩＭＳ）

東京都立大学

３．総　　括

３－Ⅰ．V-iCliniX Concept「知・もの・人」づくり

V-iCliniX 計画では、従来の生体計測による 4 次元（時空間）情報に、人体の多階層（分子－細胞－器官－多器官網－個体）ネットワークが織りなす「第５の軸」を加えた「生体５次元情報」を解読できる技術を、異分野研究者集団がつくるバーチャル研究チームで創出しました。

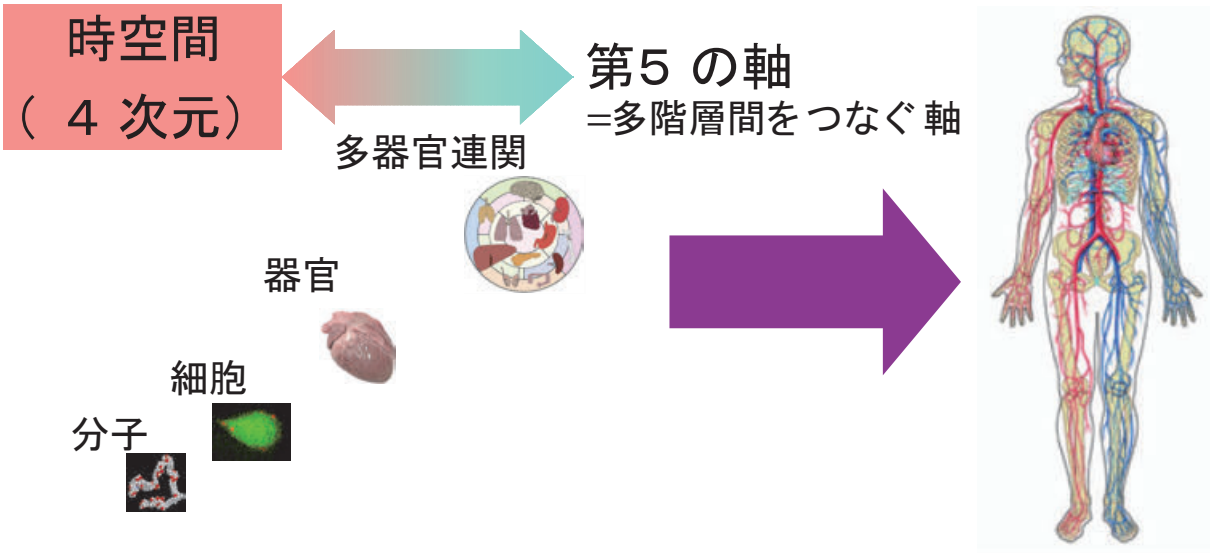


図１．生体５次元情報

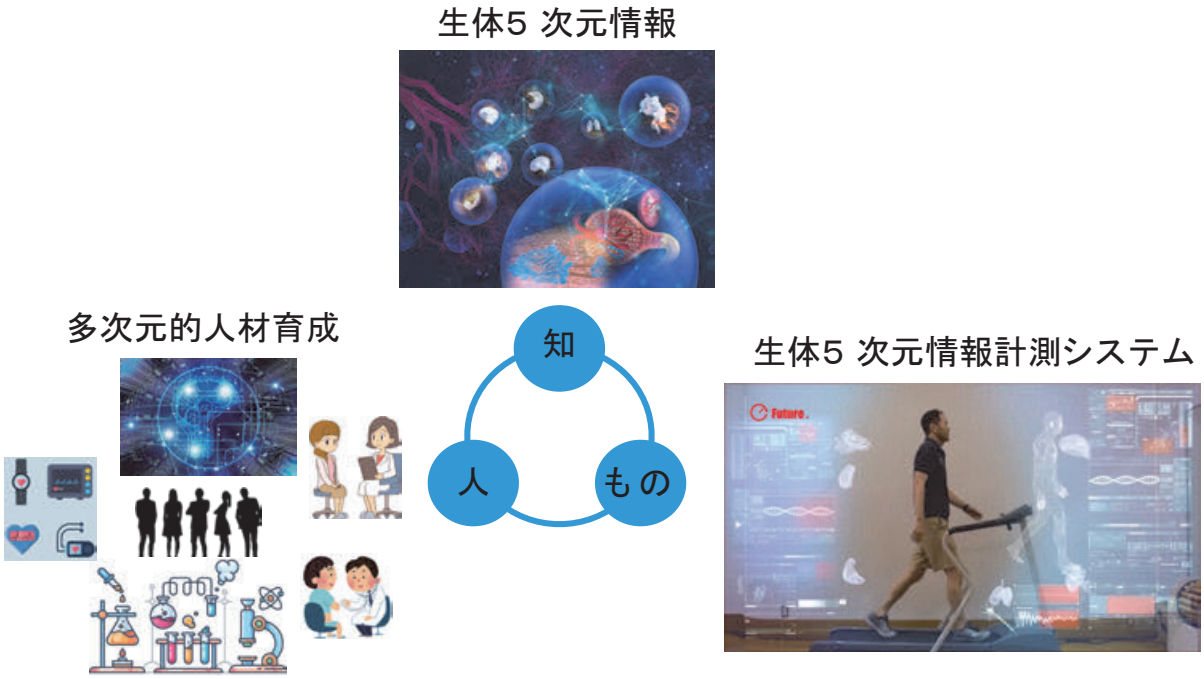


図２．「知・もの・人」づくり

３－Ⅱ．V-iCliniX Organization

V-iCliniX 計画では、奈良県立医科大学 V-iCliniX 講座佐藤グループ（臨床研究、プロジェクトマネージメント）、ATR 佐藤匠徳特別研究所佐藤グループ（動物モデル、数理情報）、東京工業大学徳田グループ（体内埋め込み型デバイス）、NIMS 土井グループ（体内センサ）、東京都立大学成川グループ（蛍光プロテインエンジニアリング）といった異分野研究集団が有機的に交わるバーチャル研究チームを形成しプロジェクトを実施しました。

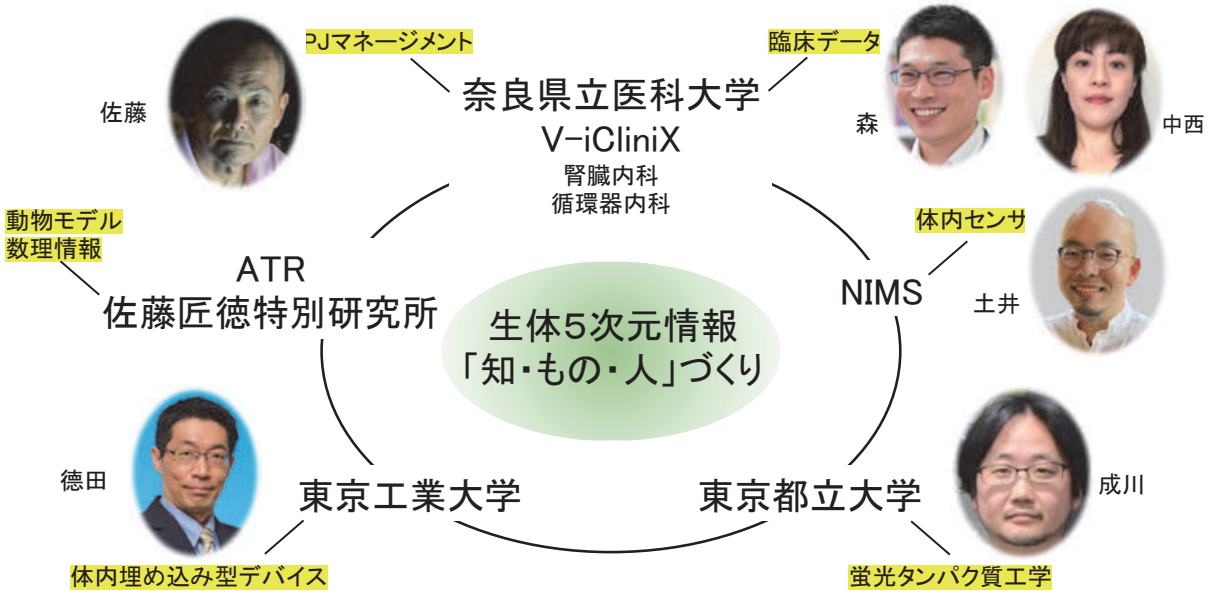


図３．V-iCliniX 計画体制

３－Ⅲ．Timeline（プロジェクト時系列）

数学、物理学、化学という分野は、いくつかの基本原理があり、その原理に基づいて体系的に観察された（る）自然現象をそれぞれの分野における現象の見方（Lens：レンズ）で説明する学問です。一方、生命科学や医学においては、現象の観察の説明に基づいた学問で、数学、物理学、化学における基本原理に相当する体系的な原理は「進化論」「セントラルドグマ」とまだ数少ないというのが現状です。しかし、21 世紀に入り、ヒトゲノムの解読、その後、体内全細胞カタログ、という生物学における百科事典作りが進んでいます（図 4）。これらの生命科学における網百科事典を数理情報学の概念及び手法をフルに活用し解析することで、生体における多次元情報（ここでは 5 次元情報と定義しました）を紐解くことで、数学、物理学、化学における基本原理に相当する体系的な原理を生命科学において見出すことを目的と、V-iCliniX 計画を考案しました。V-iCliniX 計画のアイデアは、もともと 2013 年に佐藤が代表で採択された「ERATO 佐藤ライブ予測制御プロジェクト」で具現化され、その後、2019 年度からは本 V-iCliniX 計画に引き継がれました（図 4）。さらに、これらの公的研究費の成果を事業化するために、佐藤が Founder で Karydo TherapeutiX 株式会社を 20215 年 10 月に起業しました。また、「ERATO 佐藤ライブ予測制御プロジェクト」及び「V-iCliniX 計画」を非連続的に次レベルへ発展させるため 2021 年 12 月からは防衛装備庁 ATLA の「体内精密情報デジタルツインシステム」プロジェクトが発足しています（図 4）。

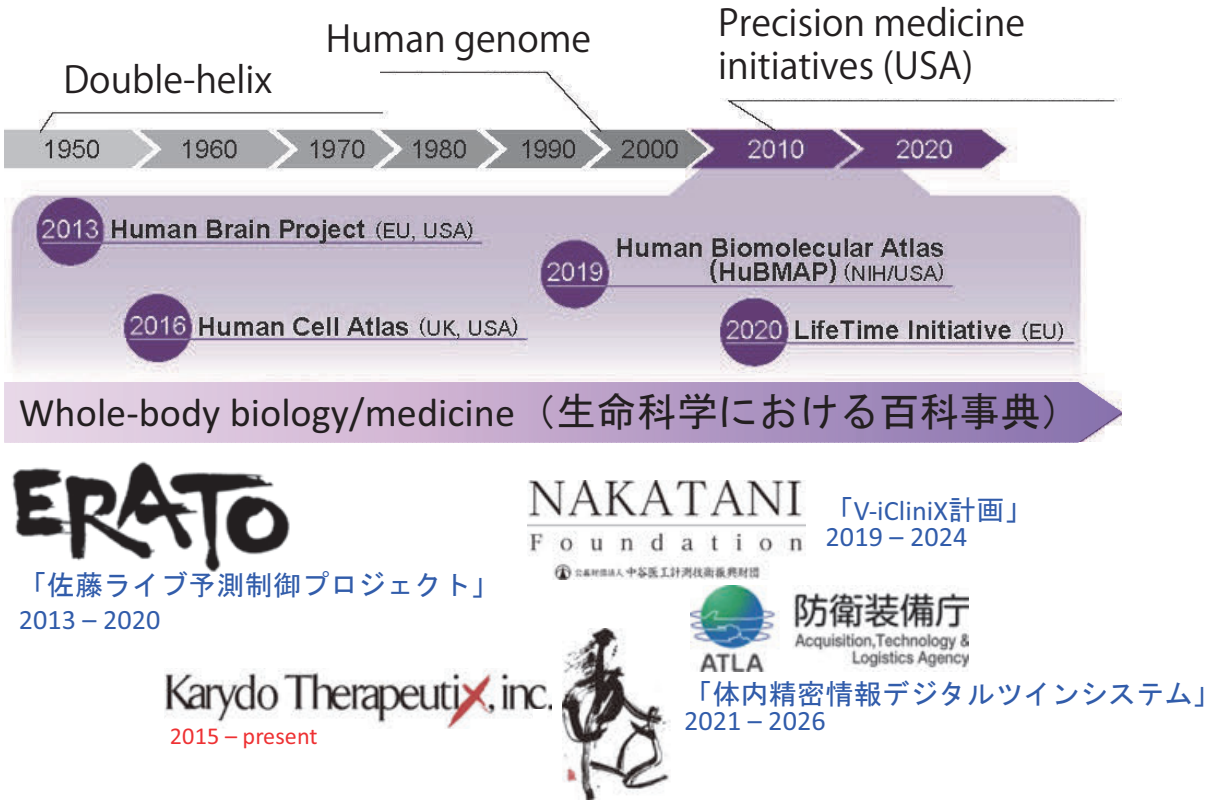


図４．Timeline

３－Ⅳ.「知」の成果

医薬品の効能及び副作用予測 AI：Humanized mouse database

医薬品や健康素材の効果及び副作用を全身網羅的に予測できるインテリジェントシステムを研究開発しました。

（論文）

1）Kozawa, S., Sagawa, F., Endo, S., De Almeida, G.M., Mitsuishi, Y., and Sato, T.N.（2020）. Predicting Human Clinical Outcomes Using Mouse Multi-Organ Transcriptome. iScience *23*, 100791. 10.1016/j.isci.2019.100791.

※ 5-1-1-1）P33 参照

疾患分類 AI

各疾患間の multimodal な情報（ゲノム、腸内細菌叢、各臓器における遺伝子発現変化、血中バイオマーカー、など）の類似性をもとに疾患を部類する AI を研究開発しました。

（論文）

2）Kozawa, S., Yokoyama, H., Urayama, K., Tejima, K., Doi, H., Takagi, S., and Sato, T.N.（2023）. Latent disease similarities and therapeutic repurposing possibilities uncovered by multi-modal generative topic modeling of human diseases. Bioinformatics Advances. 10.1093/bioadv/vbad047.

※ 5-1-1-2）P64 参照

疾患の多臓器連関メカニズム生成 AI

各疾患の multimodal な情報（ゲノム、腸内細菌叢、各臓器における遺伝子発現変化、血中バイオマーカー、など）から、その疾患のこれまで埋もれていた多臓器連関による発症・重症化メカニズムを見出し生成する AI システムを研究開発しました。

（論文）

3）Kozawa, S., Tejima, K., Takagi, S., Kuroda, M., Nogami-Itoh, M., Kitamura, H., Niwa, T., Ogura, T., Natsume-Kitatani, Y., and Sato, T.N.（2023）. Latent inter-organ mechanism of idiopathic pulmonary fibrosis unveiled by a generative computational approach. Scientific reports *13*, 21981. 10.1038/s41598-023-49281-0.

※ 5-1-1-3）P82 参照

臓器 RNA-seq データの Deconvolution

ヘテロな複数の細胞種から構成されている臓器の RNA-seq データをシングルセルへ分解において各細胞種の細胞内 RNA 量の違いを考慮できる方法を研究開発した。

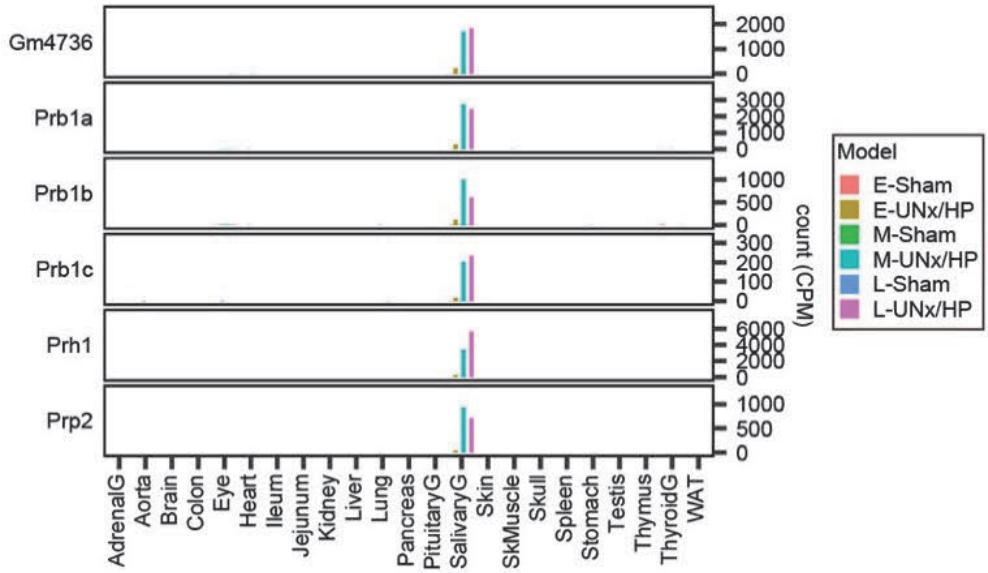
（論文）

4）Tejima, K., Kozawa, S., and Sato, T.N.（2024）Cell type-specific weighting-factors to solve solid organs-specific limitations of single cell RNA-sequencing. bioRxiv .
doi: https://doi.org/10.1101/2020.07.05.188276

※ 5-1-1-4）P98 参照

唾液腺からの液性因子群によるリン恒常性制御

唾液腺から分泌され、体内のリン恒常性の制御に関わっているプロリンリッチなタンパク質ファミリーを発見し、その機能メカニズム及びヒトにおける慢性腎疾患（CKD）の発症・重症化への関与を見出した。



（論文）

投稿中

その他

上記以外に、新規な生体機能維持機構などの発見を達成し、この成果は現在論文投稿準備中です。また、V-iCliniX 講座の准教授の森グループは以下の成果を論文として発表しました。

（森グループ）

1）Matsubayashi M, Sakaguchi YM, Sahara Y, Nanaura H, Kikuchi S, Asghari A, Bui L, Kobashigawa S, Nakanishi M, Nagata R, Matsui TK, Kashino G, Hasegawa M, Takasawa S, Eriguchi M, Tsuruya K, Nagamori S, Sugie K, Nakagawa T, Takasato M, Umetani M, Mori E. 27-Hydroxycholesterol regulates human SLC22A12 gene expression through estrogen receptor action. ***FASEB J*** 2021; 35（1）: e21262. doi: 10.1096/fj.202002077R.

※ 5-1-2-1）P143 参照

2）Iwasa N, Matsui TK, Iguchi N, Kinugawa K, Morikawa N, Sakaguchi YM, Shiota T, Kobashigawa S, Nakanishi M, Matsubayashi M, Nagata R, Kikuchi S, Tanaka T, Eura N, Kiriyama T, Izumi T, Saito K, Kataoka H, Saito Y, Kimura W, Wanaka A, Nishimura Y, Mori E, Sugie K. Gene Expression Profiles of Human Cerebral Organoids Identify PPAR Pathway and PKM2 as Key Markers for Oxygen-Glucose Deprivation and Reoxygenation. ***Front Cell Neurosci*** 2021; 15: 605030. doi: 10.3389/fncel.2021.605030.

※ 5-1-2-2）P158 参照

3）Nanaura H, Kawamukai H, Fujiwara A, Uehara T, Aiba Y, Nakanishi M, Shiota T, Hibino M, Wiriyasermkul P, Kikuchi S, Nagata R, Matsubayashi M, Shinkai Y, Niwa T, Mannen T, Morikawa N, Iguchi N, Kiriyama T, Morishima K, Inoue R, Sugiyama M, Oda T, Kodera N, Toma-Fukai S, Sato M, Taguchi H, Nagamori S, Shoji O, Ishimori K, Matsumura H, Sugie K, Saio T, Yoshizawa T, Mori E. C9orf72-derived arginine-rich poly-dipeptides impede phase modifiers. ***Nat Commun*** 2021; 12（1）: 5301. doi: 10.1038/s41467-021-25560-0.

※ 5-1-2-3）P174 参照

4）Tomo Shiota, Riko Nagata, Sotaro Kikuchi, Hitoki Nanaura, Masaya Matsubayashi, Mari Nakanishi, Shinko Kobashigawa, Noriyoshi Isozyme, Takao Kiriyama, Kazuaki Nagayama, Kazuma Sugie, Yoshito Yamashiro and Eiichiro Mori（2022）C9orf72-Derived Proline:Arginine Poly-Dipeptides Modulate Cytoskeleton and Mechanical Stress Response. *Frontiers in Cell and Developmental Biology, doi.org/10.3389/fcell.2022.750829*

※ 5-1-2-4）P186 参照

5）Iguchi N, Isozumi N, Hattori Y, Imamura T, So M, Nanaura H, Kiriyama T, Eura N, Yamaoka M, Nakanishi M, Mori M, Ohki S, Kumeta H, Koga H, Watabe M, Mabuchi T, Kanemura S, Okumura M, Yoshizawa T, Ota I, Suzuki N, Aoki M, Yamashiro Y, Saio T, Sugie K, Mori E. Zinc finger domains bind low-complexity domain polymers. ***bioRxiv*** 2023.10.29.564599.

※ 5-1-2-5）P198 参照

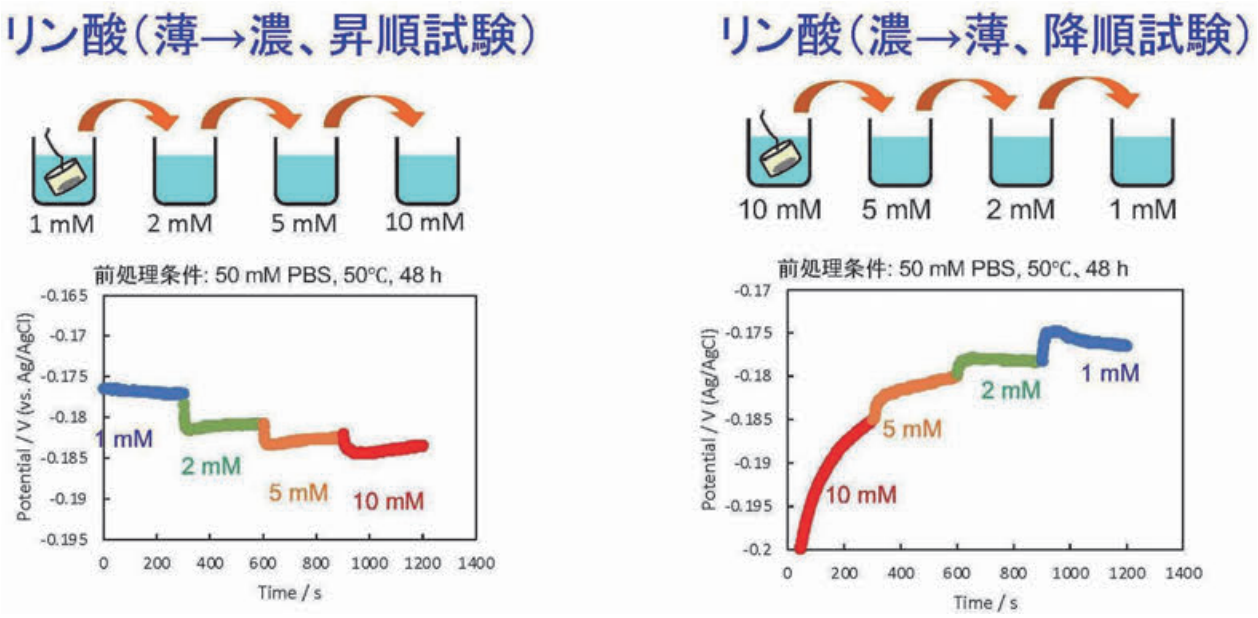
6）Sakaguchi YM, Wiriyasermkul P, Matsubayashi M, Miyasaka M, Sakaguchi N, Sahara Y, Takasato M, Kinugawa K, Sugie K, Eriguchi M, Tsuruya K, Kuniyasu H, Nagamori S, Mori E. Identification of three distinct cell populations for urate excretion in human kidney. ***bioRxiv*** 2023.06.29.545826.

※ 5-1-2-6）P221 参照

３－Ⅴ.「もの」の成果

体内リン酸センサ開発

体内のリン恒常性は、慢性腎疾患（CKD）、骨疾患、循環器系疾患、老化などへ深く関わっている。したがって、体内のリン酸濃度の変化をリアルタイムで計測できる技術は、これらの疾患や老化予防へ重要となる。そこで、リン酸濃度の変化をリアルタイムで電気化学的に計測できる分子センサを開発した。



３－Ⅵ.「人」の成果

V-iCliniX セミナーシリーズ

第 1 回
2019 年 10 月 8 日
講師：美島 健二 先生 昭和大学歯学部　口腔病態診断科学講座　口腔病理学部門　教授
セミナータイトル：唾液腺の発生研究と再生医療への応用

第 2 回
2020 年 5 月 26 日
講師：飯田 薫子 先生 お茶の水女子大学　基幹研究院　自然科学系　教授
セミナータイトル：機能性食品因子による生体制御と代謝疾患への応用

第 3 回
2020 年 6 月 16 日
講師：辻 真博 先生（国研）科学技術振興機構（JST）　研究開発戦略　フェロー
セミナータイトル：健康・医療研究開発動向の現状と展望　“①世界のライフサイエンス・臨床医学研究の潮流、②再生医療からデザイナー細胞へ、③デジタル医療”

第 4 回
2020 年 7 月 22 日
講師：矢作 直也 先生 筑波大学　医療医学系　ニュートリゲノミクスリサーチグループ / 内分泌代謝・糖尿病内科　グループ代表 / 准教授
セミナータイトル：in vivo イメージングから見えてきたニュートリゲノミクスの世界

第 5 回
2020 年 9 月 23 日
講師：成川 礼 先生 静岡大学　学術院　理学専攻　光生物学研究室　講師
セミナータイトル：UV から遠赤色光まで網羅する光受容体・シアノバクテリオクロムの色調節機構解明とその応用利用

第 6 回
2021 年 6 月 22 日
講師：冨田 峻介 先生 産業技術総合研究所　健康医工学研究部門　ナノバイオデバイス研究グループ　上級主任研究員
セミナータイトル：Chemical tongue：ヒトの味覚を模倣したバイオ試料認識技術

第 7 回
2021 年 9 月 28 日
講師：田仲 真紀子 先生 電気通信大学大学院　情報理工学研究科　基盤理工学専攻　准教授
セミナータイトル：光を用いた DNA 集合状態の機能探索　～化学の視点からの DNA 研究～

第 8 回
2021 年 11 月 15 日
講師：福山 真央 先生 東北大学　多元物質科学研究所　ナノ・マイクロ計測化学研究分野　助教
セミナータイトル：ナノ・マイクロサイズの分析化学　～微小現象の理解に向けて～

第 9 回
2022 年 5 月 17 日
講師：長崎 幸夫 先生 筑波大学　数理物質系　物質工学域　教授
セミナータイトル：自己組織化を基盤とする新しい創薬

第 10 回
2022 年 9 月 20 日
講師：佐藤 守俊 先生　東京大学大学院　総合文化研究科　教授
セミナータイトル：生命現象の光操作技術の創出

第 11 回
2022 年 11 月 29 日
講師：笹川 清隆 先生　奈良先端科学技術大学院大学　先端科学技術研究　准教授
セミナータイトル：脳機能計測に向けた整体埋植イメージンセンサの開発

シンポジウム

キックオフシンポジウム

5-2-1 P255 参照

2019 年 5 月 16 日@奈良県立医科大学

The 1st International Symposium on Human InformatiX

2020 年 2 月 27・28 日 @ATR（ハイブリッド）

Special Lecture:

Shinichi Nishikawa, M.D., Ph.D.



Keynote Speakers:

Michael Stumpf, Ph.D.
（U. Melburne, Australia）



Roeland Merks, Ph.D.
（Leiden U., Netherlands）



Johan Björkegren, M.D., Ph.D.
（Mt. Sinai Med. Sch., USA
/Karolinska Inst., Sweden）



Xia Yang, Ph.D.
（UCLA, USA）



Invited Speakers:

Anton Krantz, Ph.D.
(UCSD, USA)



Mariko Okada, Ph.D.
(Osaka U., Japan)



Yayoi Natsume, Ph.D.
(National Inst. Biomed. Innovation, Health and Nutrition, Japan)



Momoko Hayamizu, M.D., Ph.D.
(The Institute of Statistical Mathematics, Japan)



Momoko Horikoshi, M.D., Ph.D.
(Riken., Japan)



Minako Imamura, Ph.D.
(Ryuku U., Japan)



Ichigaku Takigawa, Ph.D.
(Riken, Japan)

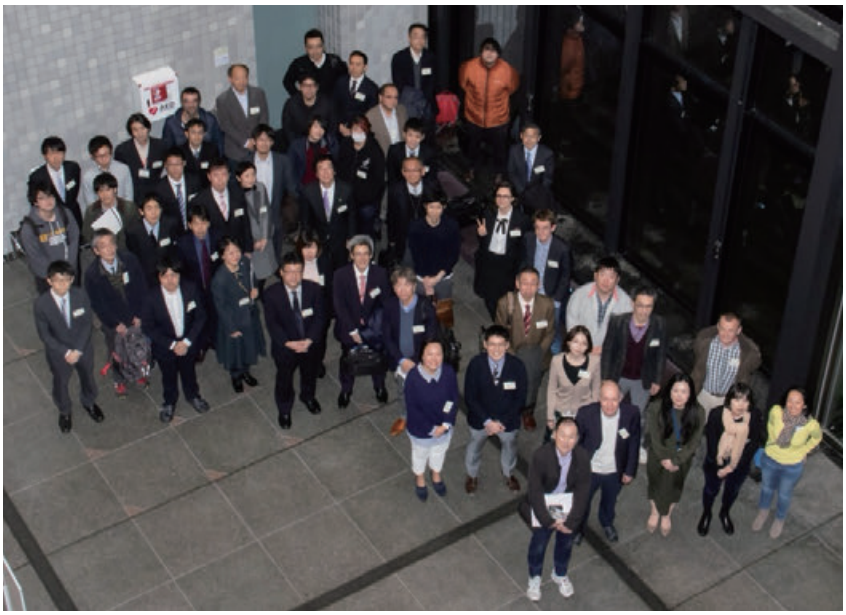


Siew-Kee (Amanda) Low, Ph.D.
(Japanese Foundation for Cancer Res., Japan)



Aki Minoda, Ph.D.
(Riken, Japan)





中間成果報告会　　5-2-2 P259 参照

The 2nd International Symposium on Human InformatiX: *Sensing, Modeling and Controlling Omics in Biomedicine*
2022 年 10 月 24 日@大阪府立国際会議場（ハイブリッド）

Keynote Speaker

Dr. Cindy Reinhart-King, Professor, Biomedical Engineering, Vanderbilt University, USA; President, The Biomedical Engineering Society, USA



V-iCliniX 最終成果報告会　　5-2-3 P281 参照

2024 年 2 月 16 日@奈良県立医科大学
V-iCliniX & Beyond

学生　研究インターン



土井穂隆
（奈良県立医科大学 2 年次～ 6 年次）
2019/09/01 – 2023/03/31



上野雄平
（奈良県立医科大学 4 年次～ 6 年次）
2020/04/01 – 2023/03/31



横山弘奈
（奈良県立医科大学 1 年次～ 6 年次）
2020/02/01 – 現在



寺島美憂
（岡山大学医学部 1 年次～ 5 年次）
（中谷財団　国際交流プログラム OG）
2021/06/01 – 現在

張夏帆
（中谷財団　国際交流プログラム OG）
大阪大学理学部生物を 2021 年 3 月卒業後、2021 年 4 月～ 2021 年 7 月までインターン@ ATR。2021 年 8 月より米国ジョンズホプキンス大学医工学修士学科へ留学

中谷医工計測技術振興財団　国際学生交流プログラム（RIES）Lectures

Thomas N. Sato. Lecture to RIES Summer Students（Japan and US）Online、1 月 4 日 , 2021.

Thomas N. Sato. Lecture to RIES Summer Students（Japan and US）TKP ガーデンシティー大阪 OAP タワー、6 月 11 日 , 2022.

Thomas N. Sato. Lecture to RIES Summer Students（Japan and US）KKR ホテル大阪、6 月 17 日 , 2023.

招待講演

佐藤匠徳

データサイエンスと未来～ AI が拓く科学技術と産学連携～（京都大学　稲盛ホール）「Biomedical Data Science による研究開発：ERATO プロジェクトから事業化まで」（佐藤匠徳）2019 年 3 月 12 日

第 46 回日本毒性学会学術年会（アステイ徳島）「Humanized Mouse-Database（hMDB）: Intelligent drug development platform integrating in vivo assay system and AI-based predictions of AE, PK, IND and DR」（佐藤匠徳）2019 年 6 月 27 日

第 20 回日本毒性学会生涯教育講習会（あわぎんホール　徳島県）「Virtual Human InformatiX Project：ヒト生体 5 次元情報ネットワークの計測・解読から開ける世界」（佐藤匠徳）2019 年 6 月 29 日

テクノロジーインパクト 2030 第 4 期「医療・健康・バイオ× AI」が切り開く未来」
（東京）「医療・健康・バイオ× AI が切り開く未来」（佐藤匠徳）2019 年 11 月 11 日

第 10 回疾患バイオリソースセンターセミナー
（東京医科歯科大学）「Virtual Human InformatiX 計画 :
Deconvolution, Reconstitution, Prediction of Human Biology and the Applications to Real-World Medicine」
（佐藤匠徳）2019 年 12 月 3 日

ヘルスケアデバイス実装研究会第 11 回公開研究会　高齢化社会に向けてエレクトロニクス、人に役立つエレクトロニクス、ヘルスケアデバイス（京都工芸繊維大学）「Virtual Human InformatiX 計画 :
Deconvolution, Reconstitution, Prediction of Human Biology and the Applications to Real-World Medicine」
（佐藤匠徳）2019 年 12 月 3 日

第 85 回日本循環器学会学術集会
（The 85th Annual Scientific Meeting of the Japanese Circulation Society）
会長特別企画における Keynote 講演（パシフィコ横浜・オンラインのハイブリット開催）「Digital Twins in Disease Science and Patient Journey」（佐藤匠徳）2021 年 3 月 26 日

幹細胞を用いた化学物質リスク情報共有化コンソーシアム 2021 年度年会
特別講演（オンライン開催）「多階層疾患デジタルツインによるバーチャル医薬品開発：探索研究から臨床試験シミュレーションまで」（佐藤匠徳）2021 年 4 月 16 日

第 17 回 CEM フォーラム（オンライン）「生体精密情報デジタルツインシステム」（佐藤匠徳）2021 年 10 月 9 日

第 118 回　ファインケミカルズ研究会例会（KKR ホテル 大阪大会議室）「体内精密情報デジタルツイン」（佐藤匠徳）2022 年 9 月 7 日

Cardiovascular and Metabolic Week 2022（CVMW2022）（第 39 回国際心臓研究学会日本部会、第 6 回日本循環器学会基礎研究フォーラム、第 30 回日本血管生物学会学術集会　合同）（ステーションコンファレンス東京：オンライン講演）
基調講演「Precision Human Digital-Twins System」（佐藤匠徳）2022 年 12 月 16 日

香川大学講義「医学と研究」（香川大学医学研究科：オンライン講演）「Precision Human Digital-Twins System」（佐藤匠徳）2024 年 1 月 11 日

森英一朗

森 英一朗 . 神経難病の病態解明と創薬への取り組み . 東京医科歯科大学 第 15 回 CBIR 若手シンポ・第 76 回大学院セミナー , 東京 , 2 月 17 日 , 2023.

森 英一朗 . *C9orf72* 変異に関連した相分離異常と疾患発症 . 第 21 回神経・筋の免疫疾患を考える会 , 大阪 , 2 月 17 日 , 2023.

森 英一朗 . 相分離の駆動・制御・破綻と神経難病 . 第 4 回生体内超分子システム研究会 , 川崎 , 2 月 27 日 , 2023.

森 英一朗 . 神経変性疾患における相分離破綻モデル . 日本小脳学会 第 13 回学術集会･総会 , 小平 , 3 月 25 日 , 2023.

森 英一朗 . 分子の動きを捉えて薬を作る？（D2-10）. コロナで変わった・進んだバイオと化学～生体分子の集合体と相分離が拓く新しい化学～ , 第 13 回 CSJ 化学フェスタ 2023, 東京 , 2023 年 10 月 18 日 .

Mori E. Recognizing cross- β polymers. 1S06m: Molecular mechanism of motor neuron disease unravelled: achievements from the interdisciplinary collaboration, The 46th Annual Meeting of the Japan Neuroscience Society, Sendai, August 1st, 2023.

森 英一朗 . 相分離生物学の基礎と神経変性疾患への応用可能性（S4-1）. シンポジウム 4「神経変性疾患の相分離」, 第 17 回パーキンソン病・運動障害疾患コングレス , 大阪 , 2023 年 7 月 21 日 .

森 英一朗 . 相分離メガネがあぶりだした生命科学の課題（WS3-01）. WS3「遅延制御による相分離現象の理解」（共催：学術変革領域研究 B「遅延制御超分子化学」), 第 23 回蛋白質科学会 , 名古屋 , 2023 年 7 月 5 日 .

一般向けセミナー

ATR オープンハウス　テーマ講演「Virtual Human InformatiX Project: ヒト生体 5 次元情報ネットワークの計測・解読から拓ける世界」（佐藤匠徳）2019 年 11 月 1 日

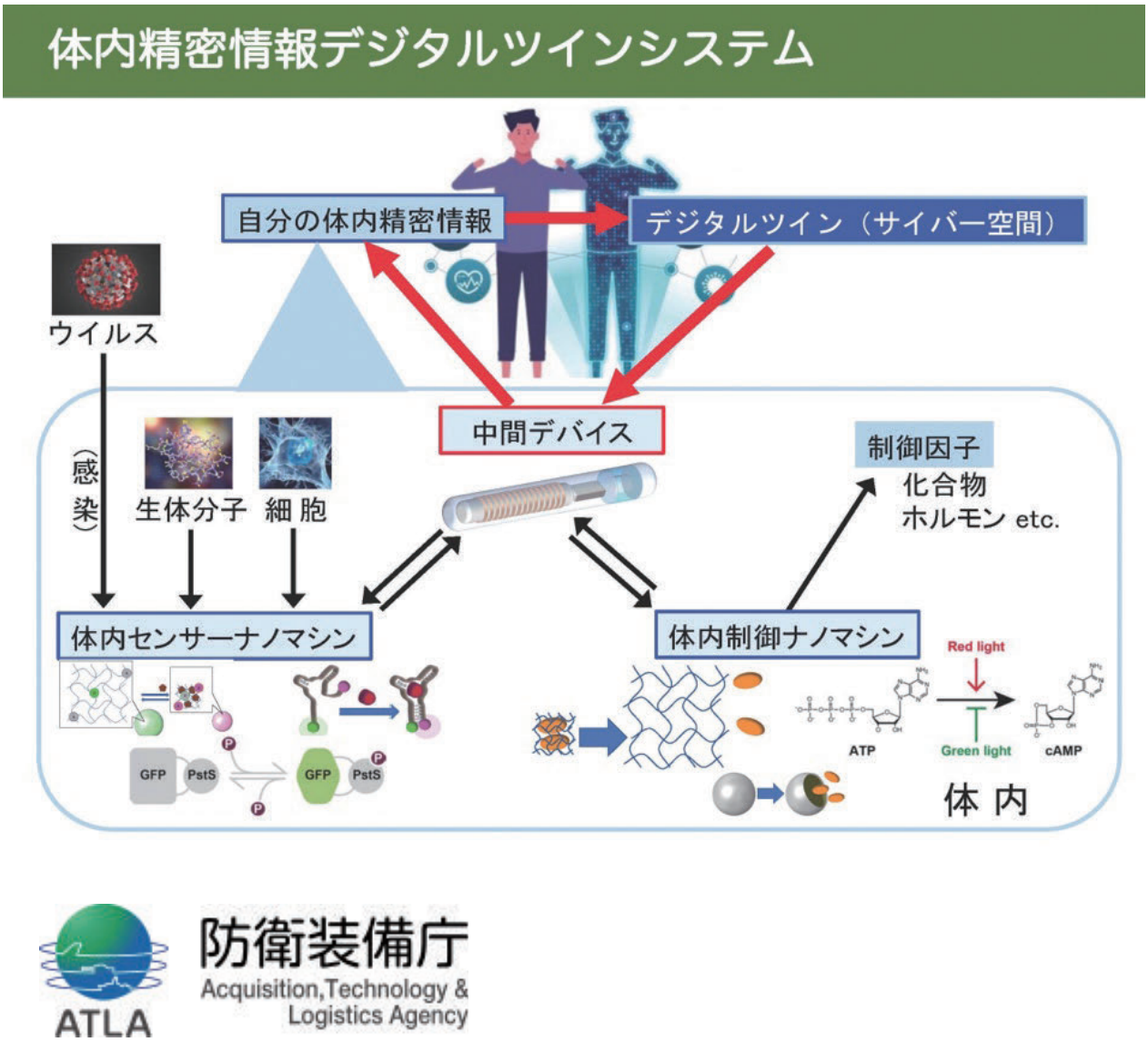
ATR オープンハウス　テーマ講演「体内精密情報デジタルツイン」（佐藤匠徳）2021 年 11 月 11 日

ATR オープンハウス　テーマ講演「体内精密情報デジタルツインシステム」（佐藤匠徳）2022 年 10 月 7 日

3－Ⅶ. Beyond V-iCliniX

V-iCliniX 終了後は、以下 3 つの取り組みにより、V-iCliniX の成果を発展させさ社会へ還元しています。

1.「体内精密情報デジタルツインシステム」防衛装備庁 ATLA



2. Karydo TherapeutiX 株式会社で事業化



3. 森グループの成果は、奈良県立医科大学　未来基礎医学講座及び森が起業したモルミル株式会社（molmir）で、「分子を見る」をキーワードに基礎研究から事業化を推進させています





４．謝　　辞

謝　辞

これまでの５年間、中谷財団、奈良県立医科大学、そして、財団関係者各位、技術評価委員の先生方、V-iCliniX及び共同研究のメンバー、国内外のアカデミア・企業の数多くの方々に大変お世話になりました。深く御礼申し上げます。みなさまのお力添え無くしては、本プロジェクトを立ち上げ、完走させることはできませんでした。ありがとうございます。これからも、前述のBeyond V-iCliniXの取り組みにより、V-iCliniXの成果を発展させ、社会へ還元していく所存でございます。引き続きよろしくお願い致します。

佐藤匠徳（ナルトク）（Thomas N. Sato）

5. Appendix

5-1 発 表 論 文

5-1-1　佐藤教授グループ

1）…… 33
2）…… 64
3）…… 82
4）…… 98

1）Kozawa, S., Sagawa, F., Endo, S., De Almeida, G.M., Mitsuishi, Y., and Sato, T.N. (2020). Predicting Human Clinical Outcomes Using Mouse Multi-Organ Transcriptome. iScience 23, 100791. 10.1016/j.isci.2019.100791.

2）Kozawa, S., Yokoyama, H., Urayama, K., Tejima, K., Doi, H., Takagi, S., and Sato, T.N. (2023). Latent disease similarities and therapeutic repurposing possibilities uncovered by multi-modal generative topic modeling of human diseases. Bioinformatics Advances. 10.1093/bioadv/vbad047.

3）Kozawa, S., Tejima, K., Takagi, S., Kuroda, M., Nogami-Itoh, M., Kitamura, H., Niwa, T., Ogura, T., Natsume-Kitatani, Y., and Sato, T.N. (2023). Latent inter-organ mechanism of idiopathic pulmonary fibrosis unveiled by a generative computational approach. Scientific reports *13*, 21981. 10.1038/s41598-023-49281-0.

4）Tejima, K., Kozawa, S., and Sato, T.N. (2024) Cell type-specific weighting-factors to solve solid organs-specific limitations of single cell RNA-sequencing. bioRxiv . doi: https://doi.org/10.1101/2020.07.05.188276

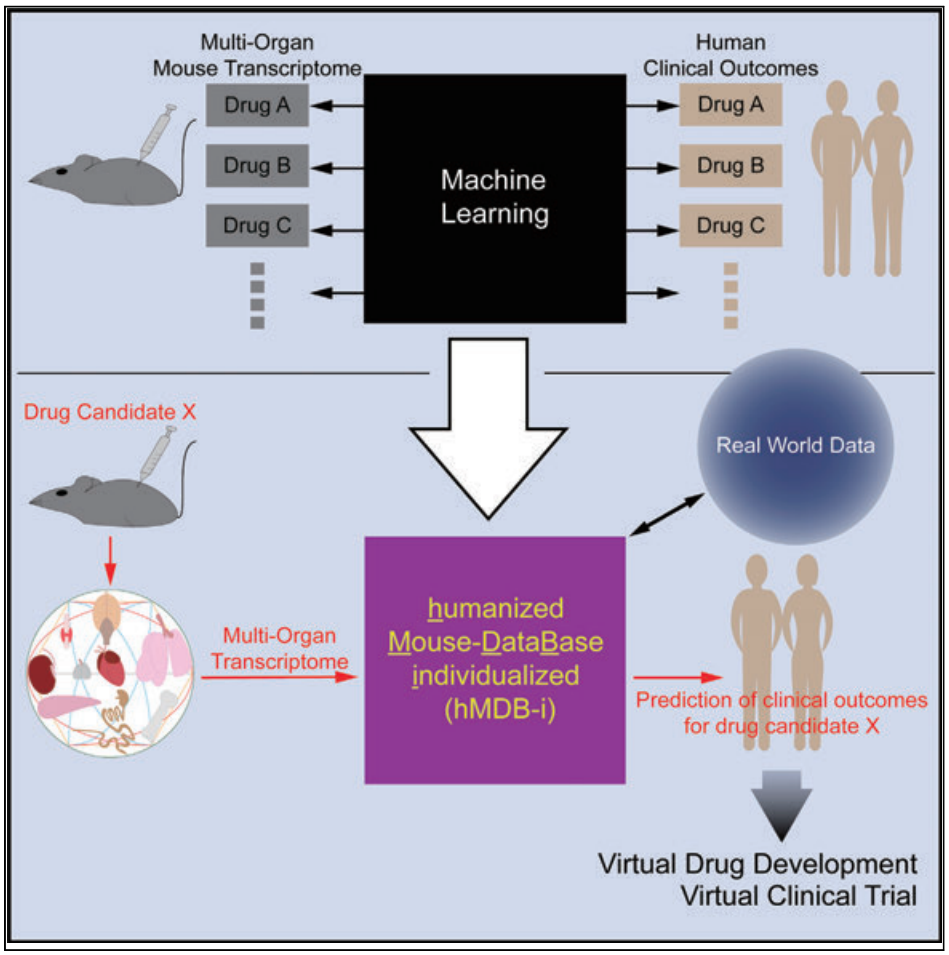
1)

iScience

CellPress

Article

Predicting Human Clinical Outcomes Using Mouse Multi-Organ Transcriptome



Satoshi Kozawa, Fumihiko Sagawa, Satsuki Endo, Glicia Maria De Almeida, Yuto Mitsuishi, Thomas N. Sato

island1005@gmail.com

HIGHLIGHTS

Mouse-transcriptome humanized by machine learning predicts human clinical outcomes

This system is referred to as humanized Mouse DataBase individualized, hMDB-i

hMDB-i predicts drug adverse events and therapeutic indications

hMDB-i is a bias-free and modality-independent system for virtual drug development

DATA AND CODE AVAILABILITY

GSE142068

Kozawa et al., iScience 23, 100791
February 21, 2020 © 2019 The Author(s).
https://doi.org/10.1016/j.isci.2019.100791

iScience

CellPress

Article

Predicting Human Clinical Outcomes Using Mouse Multi-Organ Transcriptome

Satoshi Kozawa,[1,2,3] Fumihiko Sagawa,[1,2,3] Satsuki Endo,[1,2,3] Glicia Maria De Almeida,[1] Yuto Mitsuishi,[1] and Thomas N. Sato[1,2,3,4,5,6,7,*]

SUMMARY

Approximately 90% of pre-clinically validated drugs fail in clinical trials owing to unanticipated clinical outcomes, costing over several hundred million US dollars per drug. Despite such critical importance, translating pre-clinical data to clinical outcomes remain a major challenge. Herein, we designed a modality-independent and unbiased approach to predict clinical outcomes of drugs. The approach exploits their multi-organ transcriptome patterns induced in mice and a unique mouse-transcriptome database "humanized" by machine learning algorithms and human clinical outcome datasets. The cross-validation with small-molecule, antibody, and peptide drugs shows effective and efficient identification of the previously known outcomes of 5,519 adverse events and 11,312 therapeutic indications. In addition, the approach is adaptable to deducing potential molecular mechanisms underlying these outcomes. Furthermore, the approach identifies previously unsuspected repositioning targets. These results, together with the fact that it requires no prior structural or mechanistic information of drugs, illustrate its versatile applications to drug development process.

INTRODUCTION

Unexpected adverse events (AEs) and/or the lack of the expected efficacy in human subjects fails drug development. The average success rate of drug candidates through clinical trials is 13.8% (Wong et al., 2019). This low success rate costs US$161M–1.8B per drug candidate on drug developers (e.g., pharmaceutical and biotech companies) (Morgan et al., 2011), leading to drug price hike and rising medical cost. Hence, effective prediction of clinical outcomes from the pre-clinical studies improves the success rate of drug development and reduces the drug price and medical cost.

A major impediment in drug development is the unpredictability of human outcomes on the basis of the pre-clinical results using cells, organs, and/or animal models. To overcome this problem, numerous technological solutions, both experimental and computational approaches, are currently undertaken. The uses of human cells and organs *in vitro* and *in vivo* models are among such approaches. Human cells such as induced human pluripotent cells (human iPSCs) (Elitt et al., 2018; Ko and Gelb, 2014; Meseguer-Ripolles et al., 2018) and an organ(s)-on-a-chip consisting of human cells (Oleaga et al., 2016; Rezaei Kolahchi et al., 2016) are currently utilized as drug screening and *in vitro* toxicity assays. As *in vivo* models, partially humanized mouse models, such as those where the liver is nearly 100% composed of human cells, is exploited (Tateno et al., 2004). In addition to such experimental approaches, computational tools are also invented and used. In particular, applications of machine learning algorithms for predicting clinical outcomes are in fashion (Shah et al., 2019; Vamathevan et al., 2019). They exploit the big-data sets representing structural and functional features of drugs and their target information.

Although both of such experimental and computational approaches have shown some success and promise, there are certain limitations with these existing approaches. The existing machine learning approaches require prior knowledge about the characteristics (such as structure) of drugs and their mechanisms of actions (such as target molecules). In addition, many of the computational approaches are specialized for the drugs of specific modality such as small-molecule compounds (i.e., mixed structures and mechanisms). Hence, they have difficulty dealing with the mixture of compounds. The existing experimental approaches are often "biased", as they can design assay system according to what the testers want to examine. For example, the testers need to know which organ(s) or phenotype(s) to examine for the drug effect(s) prior to designing the experiment(s). Furthermore, these presently available experimental and computational approaches fail to recapitulate the organismal level body-wide drug effects in human. Hence, an approach that recapitulates organismal biology but does not require any prior knowledge about the drug structure or

[1]Karydo TherapeutiX, Inc., Kyoto, Japan

[2]ERATO Sato Live Bio-Forecasting Project, Kyoto, Japan

[3]The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

[4]Biomedical Engineering, Cornell University, Ithaca, NY, USA

[5]Centenary Institute, Sydney, Australia

[6]V-iClinix Laboratory, Nara Medical University, Nara, Japan

[7]Lead Contact

*Correspondence: island1005@gmail.com

https://doi.org/10.1016/j.isci.2019.100791

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

mechanisms (i.e., modality-independent and unbiased) is expected to advance and facilitate the drug development process.

At the whole-body level, drugs could act on not only one or a few specific organ(s), but also multiple organs showing complex effects (Berger and Iyengar, 2011). Drugs could also act on one organ that influences the functions of other organs via inter-organ cross talks (Droujinine and Perrimon, 2013, 2016). Hence, measuring the drug effects on large number of organs may provide useful features that could be exploited in predicting the drug effects at the whole-body level. However, this assay is only possible with animal models but not with human subjects.

We previously conducted multi-organ multi-model transcriptome studies and identified model-specific and organ-specific transcriptome patterns in mice (Kozawa et al., 2018). Furthermore, weighted correlation network analysis (WGCNA) of the gene expression across multiple organs identified a number of putative organ-to-organ cross talks, some of which we genetically validated (Kozawa et al., 2018). On the basis of this previous study, we considered that such multi-organ transcriptome patterns could be exploited as the unique features of drugs at the whole-body level to predict its clinical outcomes. Furthermore, if successful, the approach provides a modality-independent and unbiased system requiring no prior knowledge of the structures/mechanisms of drugs, as the only requirement is that the drugs can be administered to mouse.

In the current study, we develop and show the performance of such system. We design a hybrid approach of experimental and machine-learning methods to utilize the multi-organ transcriptome patterns induced by drugs in mice to predict their clinical outcomes in human. We evaluate its prediction performance for both AEs and therapeutic indications by cross-validation. In addition, we also test whether the approach could be adapted to deducing molecular and cellular mechanisms underlying such predictions. Furthermore, we illustrate a possibility of applying this approach to identifying drug repositioning targets.

RESULTS

"Humanizing" Mouse Datasets

The concept of our approach is described in Figure 1. First, we generate transcriptome data across 24 organs from the mice to which we administer each drug with known clinical outcomes. Next, for each human clinical criterion (e.g., AE, therapeutic indication), we train a machine learning model with the 24-organ transcriptome data induced by the drug in mice and its known outcomes in human. The outcome data are classified according to the demographic profiles (e.g., sex, age) of the individuals. Consequently, the drug-induced 24-organ mouse transcriptome patterns are associated with the individualized human clinical outcomes in the machine learning model/database, hence referred to as "humanized Mouse-DataBase, individualized (hMDB-i)." To predict clinical outcomes of a new drug candidate X, we generate the 24-organ transcriptome data from the mice to which X is administered. Such data are used as the input data into hMDB-i to predict X's putative individualized clinical outcomes (i.e., the output).

AEs Prediction by Support Vector Machine Algorithm

We generated the 24-organ transcriptome datasets with 15 drugs of various modalities (i.e., small molecules, antibody, and peptide) with diverse known therapeutic indications (see Transparent Methods, Figure 1 and Table S1). The incidence of AEs reported for each drug in each sex and each age group were compiled from US Food and Drug (FDA) Adverse Event Reporting System (FAERS). For each AE (a total of 5,519 AEs is reported for one or more of the 15 drugs), we train a support vector machine (SVM) algorithm with the drug-induced 24-organ transcriptome patterns to predict the individualized outcomes.

We performed cross-validation to evaluate the effectiveness of this approach to predict the outcomes of the AEs (see Transparent Methods, Figure 2A). We first train an SVM algorithm with the 14 drug-induced 24-organ mouse transcriptome data. We then input the 24-organ mouse transcriptome data of the omitted drug as "an unseen drug data" (see Transparent Methods, Figure 2A). We performed such evaluations for all 15 drugs by omitting one drug at a time from the training data. We used an SVM classifier to predict whether each AE would be observed by each drug for each sex and age-group (see Transparent Methods). The scores of accuracy, precision, recall, and F-measure of the prediction result of the AEs reported for each sex and age group for each drug are summarized in Figure 2B and Table S2.

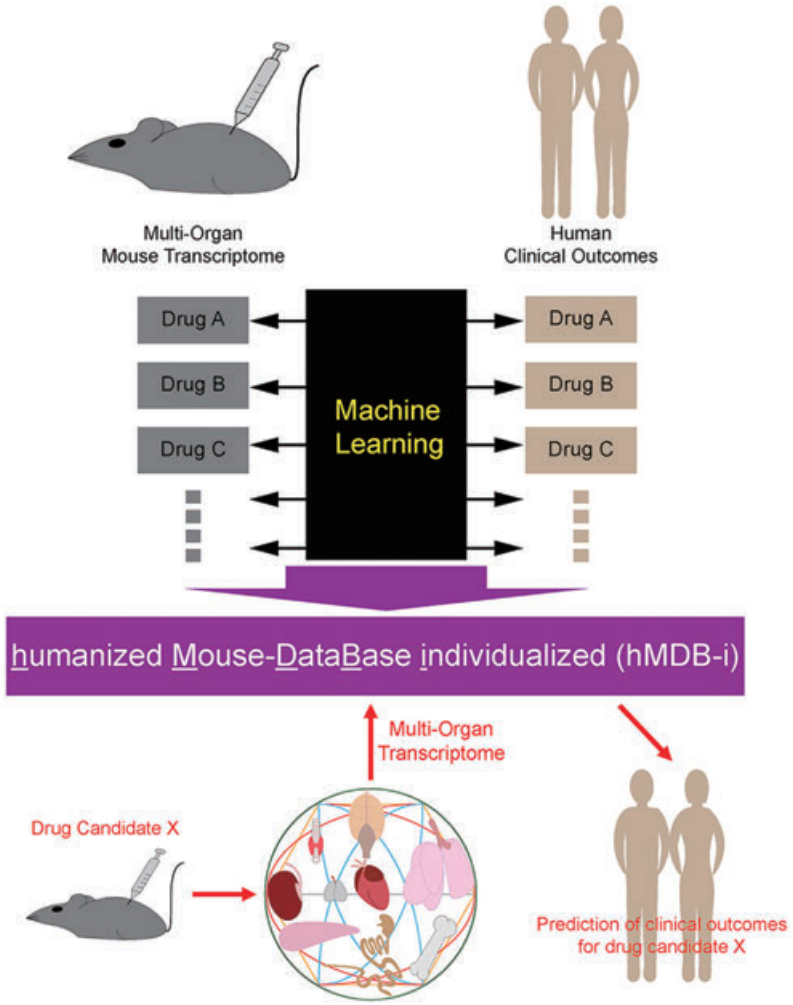


Figure 1. Conceptual Framework of hMDB-i
See text for the description.

The result shows that the accuracy scores are over 0.7 for the 200 of 264 sex/age-groups across the 15 drugs and over 0.9 for the 52 sex/age-groups. This indicates that more than 70% of the AEs predicted to either appear or not appear for the drug are indeed reported or not reported, respectively, in these sex/age-groups. The precision scores are over 0.5 for the 156 of the 264 sex/age-groups, indicating that more than 50% of the AEs predicted to appear for the drug are indeed reported in these sex/age-groups. If the question is whether an AE predicted to appear for a drug is indeed reported in at least one or more of the sex/age group, 13 drugs (alendronate, acetaminophen, aripiprazole, cisplatin, clozapine, doxycycline, empagliflozin, lenalidomide, olanzapine, evolocumab [Repatha], risedronate, sofosbuvir, teriparatide) show the precision scores of over 0.5 (see "All" row in each drug in Figure 2B), indicating that more than 50% of the AEs predicted to appear for these drugs are indeed reported at some sex/age groups. Furthermore, for nine drugs (alendronate, acetaminophen, aripiprazole, cisplatin, clozapine, doxycycline, lenalidomide, olanzapine, teriparatide), the precision scores of 0.9–1.0 are found at least in one or more the sex/age groups. The recall score indicates how many of the reported AEs are indeed predicted for the drug in each sex/age group. The recall scores are relatively lower across all drugs and sex/age groups, indicating that the approach misses many of the reported AEs. However, for such drugs (asenapine and lurasidone), the recall scores of 0.7–1.0 are found for certain sex/age groups. Sample size differences among different

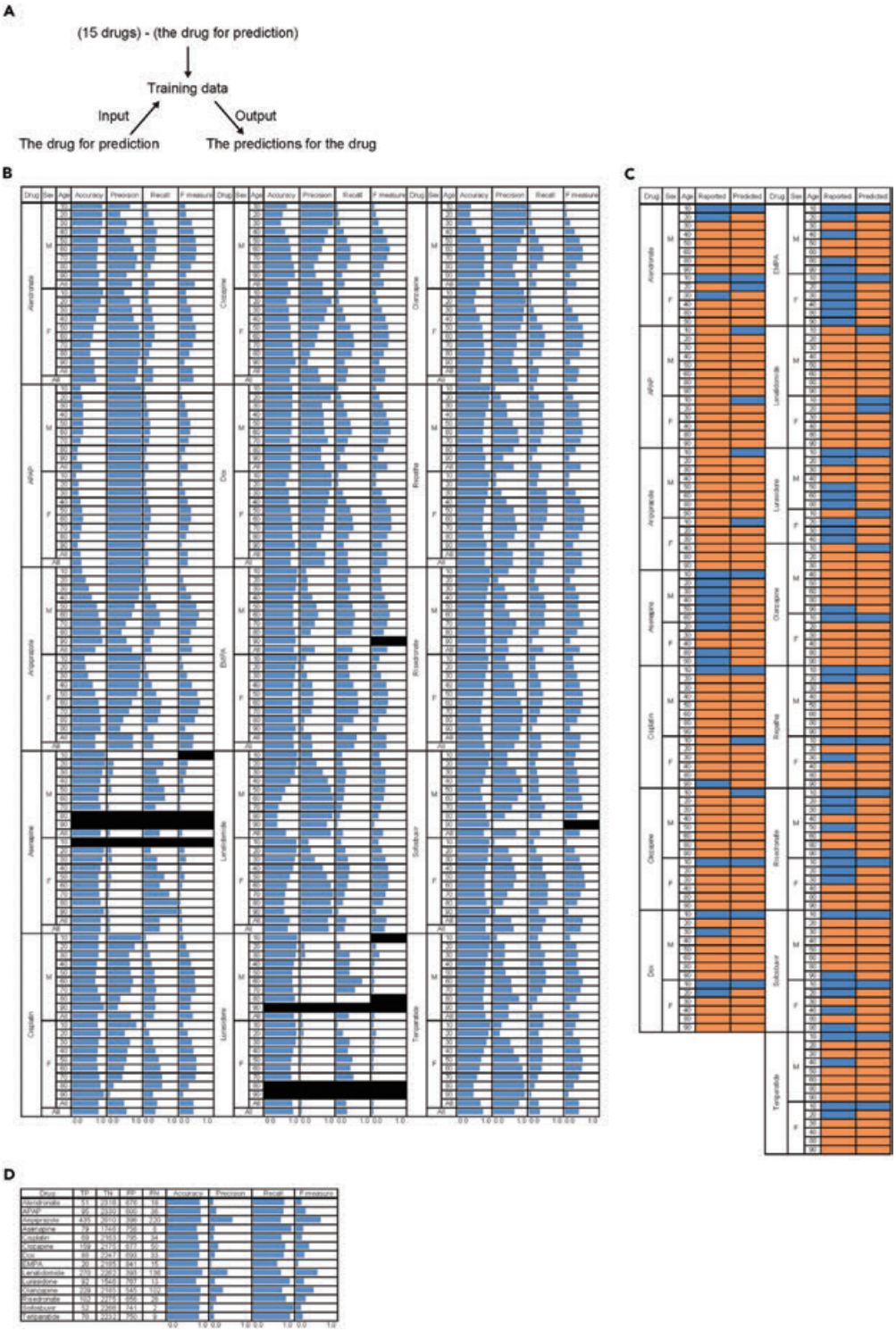

A
(15 drugs) - (the drug for prediction)
Training data
Input
Output
The drug for prediction
The predictions for the drug
B
C
D

Figure 2. Summary of the AE Prediction Results by SVM/hMDB-i

(A) Hold-out scheme.

(B) Confusion matrix summarizing the AE prediction results for each drug. Scores of accuracy (Accuracy), precision (Precision), recall (Recall), and F-measure (F measure) for each drug (Drug) are indicated by bars (light blue) for each age group (Age) and sex (Sex). The scale of the scores (0.0–1.0) are indicated at the bottom. The black filledboxes across all scores indicate that no AEs are reported for this drug in this sex/age group. When the denominator is 0 in the F-measure calculation, the score is incalculable, hence shown as black filled box.

(C) The prediction results of death event for each drug (Drug) are summarized for each age group (Age) and sex (Sex) using FAERS. The reported outcomes (Reported) are shown as orange and light-blue filled boxes indicating death event is reported and not-reported, respectively. The prediction results (Predicted) are shown as orange and light-blue filled boxes indicating death event is predicted to occur and not-to-occur, respectively. The age groups are 10: 10–19 years old, 20: 20–29 years old, 30: 30–39 years old, 40: 40–49 years old, 50: 50–59 years old, 60: 60–69 years old, 70: 70–79 years old, 80: 80–89 years old, 90: 90–99 years old, All: 10–99 years old. F, females; M, males; APAP, Acetaminophen; Dox, doxycycline; EMPA, empagliflozin; Repatha, evolocumab. The raw data are available as Table S2.

(D) The confusion matrix showing the prediction result using SIDER4.1 reference database. The raw data are available as Table S3.

sex/age groups and the number of the reports for each AE may significantly impact on the accuracy, precision, and recall scores. In fact, the number of the reported AEs for asenapine and lurasidone is significantly smaller than those of the other drugs and their precision is lower.

In FAERS, many of the AEs (e.g., diarrhea, abdominal pain) are common among many drugs. The outcomes of rare but serious AEs (SAEs) are often more important to know in advance to clinical trials. Hence, we assessed the prediction performance on SAEs. The top of such SAE hierarchy is "death event" (Sonawane et al., 2018). The result shows that the correct outcomes of death event are predicted for 148 of 204 sex/age groups across the 15 drugs (Figure 2C). The death events are reported in a total of 141 sex/age groups, and our prediction approach missed its incidence only in 12 groups (Figure 2C).

To examine potential influences of reference-specific errors, biases, and/or cofounders that are often found in any real-world data such as FAERS (Harpaz et al., 2013; Tatonetti et al., 2012), we validated our prediction with another reference, Side Effect Resource (SIDER) (http://sideeffects.embl.de). Several differences between FAERS and SIDER must be noted. SIDER4.1 version does not include evolocumab (Repatha). The outcomes reported in SIDER4.1 are not classified according to sex/age groups. The SEs and AEs are not necessarily represented by the same terms. Considering these differences, we examined the outcome predictions for 14 drugs (i.e., 15 drugs minus evolocumab) without the sex/age group classifications. In addition, the SEs/AEs common to both reference databases were analyzed. The result, 0.677–0.827 accuracy, 0.023–0.523 precision, and 0.571–0.963 recall, is comparable with that with FAERS (Figure 2D, the full and raw data for the confusion matrix is available as Table S3), further supporting the effectiveness of our method.

Next, we evaluated the necessity of multiple organ transcriptome. Figure 3 shows the SAEs that the 24-organ transcriptome correctly predicts the outcome for each drug (Figure 3 and Table S4). The predictions of these SAEs by individual-organ transcriptome data in the cross-validation scheme show variable results among different organs (Figure 3). Although this result indicates that some of the individual-organ datasets are sufficient to predict the correct outcomes, which individual-organ dataset(s) is(are) necessary varies among which SAEs to be predicted (Figure 3). Hence, it appears beneficial to collect all 24-organ transcriptome data.

AE Prediction by Random Forest Algorithm

We tested another machine learning algorithm, random forest (RF), to predict the AE outcomes. We compared the prediction results for death event outcomes (Figure 4 and Table S5). The result shows that both SVM and RF provides the same predictions for all sex/age groups (Figure 4A). The calculations of accuracy, precision, recall, and F-measure scores by both algorithms show the similar results for all 15 drugs (Figure 4B). For all drugs except asenapine, empagliflozin, and lurasidone, all scores of accuracy/precision/recall are 0.5–1.0. Although the accuracy and the precision scores for asenapine, empagliflozin, and lurasidone are 0.2–0.4, the recall scores for these three drugs by both SVM and RF are 1.0 (asenapine, empagliflozin) and 0.667 (lurasidone), indicating that the occurrence of death event by these three drugs are efficiently predicted across all sex/age groups (i.e., not missing the possible occurrence of death event). The results indicate that both algorithms are equivalent in predicting AE outcomes.

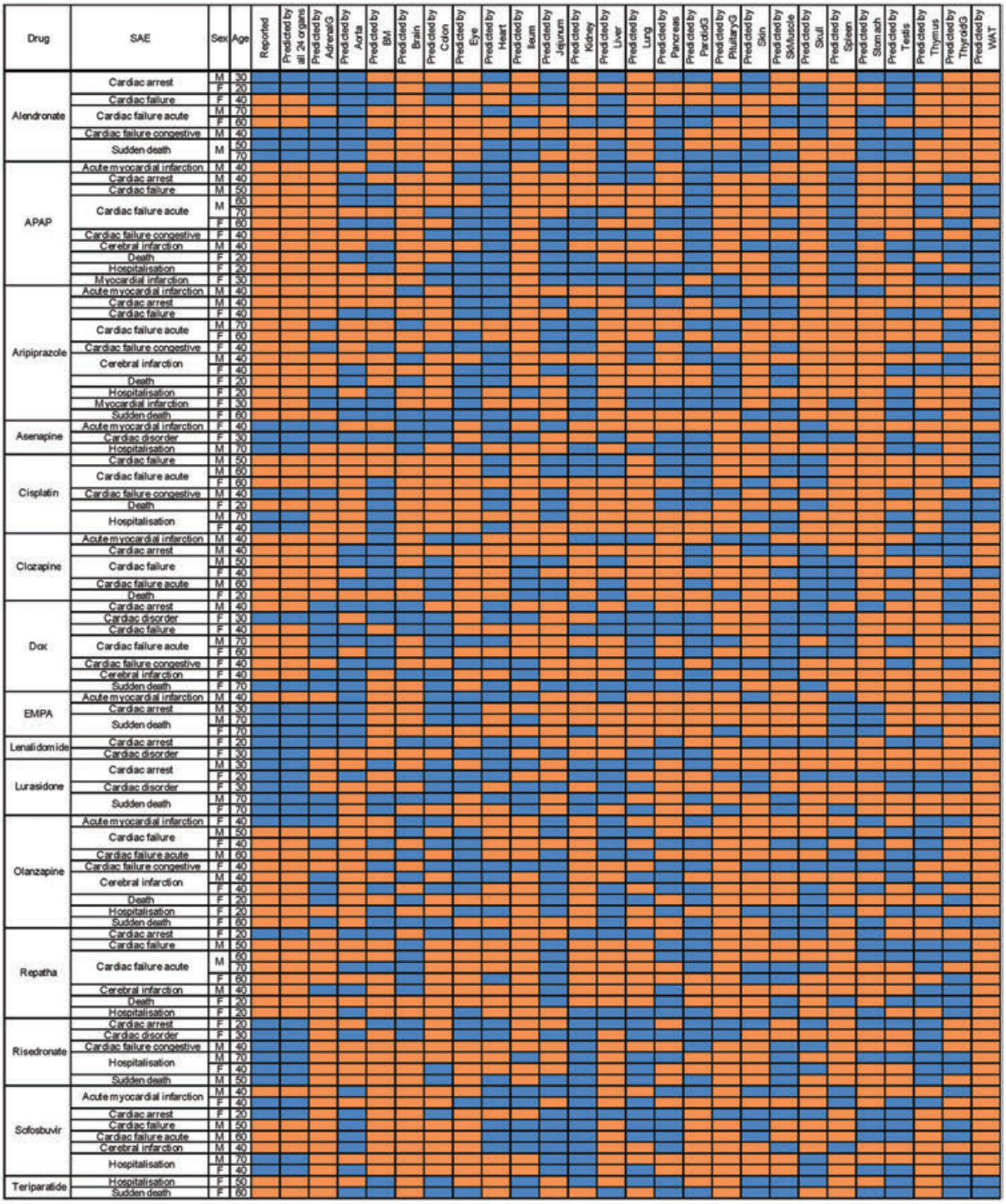


Figure 3. The Comparison of the Prediction Results by the 24-Organ and Individual-Organ Transcriptome Data

Each serious adverse event is described (SAE). The reported outcomes (Reported) are shown as orange and light-blue filled boxes indicating the SAE is reported and not-reported, respectively. The prediction results (Predicted) are shown as orange and light-blue filled boxes indicating the SAE is predicted to occur and not-to-occur, respectively. The prediction results of with the 24-organ (Predicted by all 24 organs) and individual-organ (Predicted by *organ name*)

Figure 3. *Continued*

data are shown. AdrenalG, adrenal gland; BM, bone marrow; WAT, gonadal white adipose tissue; ParotidG, parotid gland; PituitaryG, pituitary gland; SkMuscle, skeletal muscle; ThyroidG, thyroid gland. The age groups are 10: 10–19 years old, 20: 20–29 years old, 30: 30–39 years old, 40: 40–49 years old, 50: 50–59 years old, 60: 60–69 years old, 70: 70–79 years old, 80: 80–89 years old, 90: 90–99 years old. F, females; M, males; APAP, Acetaminophen; Dox, doxycycline; EMPA, empagliflozin; Repatha, evolocumab. The result is also included in Table S4.

AE Prediction by Majority Decision Framework

Although the performance with two different algorithms (SVM and RF) were equivalent (Figure 4), we obtained differential results with the 24-organ and the individual-organ approaches (Figures 2 and 3). In the case of the individual-organ approach, which individual-organ dataset(s) to be used was critical for predicting the correct outcomes (Figure 3). Hence, it may be beneficial to conduct predictions with both the 24-organ and all individual-organs to maximize the effectiveness of the hMDB-i approach. For this purpose, we designed a majority decision framework and evaluated its effectiveness for predicting death event outcomes with alendronate and lenalidomide in female/20s group where the SVM approach with the 24-organ transcriptome dataset alone failed to predict its occurrence (see Methods for the detailed description, Figure 4C). The result shows that this majority decision framework effectively predicts the correct outcomes (Figure 4C).

AE Prediction by Link-Prediction Framework

Next, we tested the power of hMDB-i for the prediction of AE outcomes in a different framework. The relationship between the adverse effects and therapeutic indications has previously been exploited for clinical outcome predictions (Zhang et al., 2013). Hence, we examined whether such data could be utilized to enhance the prediction capability of hMDB-i. For this purpose, we devised a link-prediction (LP) framework as shown in Figure 5A (see Transparent Methods for the detailed description). Three training datasets are used: (1) the 24-organ transcriptome data of each of the 15 drugs, (2) AEs reported for each of all drugs in FAERS, and (3) all AEs reported in FAERS for each of all indications in FAERS (Figure 5A). In the LP framework, one-class SVM algorithm is used to train the models and the presence/absence of a link of an untrained drug (Drug Candidate X in Figure 5A), based on its 24-organ transcriptome pattern, to each AE is determined (see Transparent Methods for the detailed description).

The LP prediction of all AEs for three drugs, alendronate, clozapine, and evolocumab (Repatha), was conducted, and the result is summarized as a confusion matrix (Figure 5B, Table S6). Although the accuracy and precision scores are slightly better with hMDB-i alone (SVM), the LP framework (LP) improves hMDB-i in the recall scores (Figure 5B). This result indicates two properties of these two approaches: (1) The AE outcomes predicted by hMDB-i alone is more likely to be observed than those with the LP framework; (2) The LP framework is superior to hMDB-i alone in not missing the occurrence of potential AEs. The examples of such superior property of the LP framework are shown in Figure 5C listing some of the reported SAEs that are missed by the hMDB-i alone but captured by the LP framework. The full list of such can be found in Table S6.

Biological Insights into the AE Mechanisms

Insights into the biological mechanisms underlying AEs provide opportunities for designing the strategies to reduce the incidence of AEs during the drug development processes. For this purpose, RF algorithm is useful as it calculates feature importance of organ-gene datasets for the correct outcome predictions. Tables 1 and 2 show organs and cellular/molecular pathways that were found important for predicting the correct outcomes of the death event for males at 50s for aripiprazole. Top 5 and Top 8 organ-pathway combinations identified by REACTOME (Table 1) and KEGG (Table 2), respectively, are shown. The results for this and other drugs and age/sex groups are summarized in Tables S7 and S8. The result indicates a possibility that the hMDB-i is also useful for deducing mechanisms underlying the predicted AE outcomes.

Predicting the Outcome Dynamics over Ages

Information regarding the quantitative differences of the AE occurrence among the sex/age group(s) would be beneficial for designing efficient clinical trials. It allows for selecting target subjects with less chance of observing an AE in the trial. Hence, we next evaluated the performance for predicting quantitative changes for the outcome incidences of AEs (see Transparent Methods for the detailed description, Figure 6, Table S9). For this purpose, we applied support vector regressor (SVR) and random forest regressor (RFR) algorithms to the hMDB-i framework. The number of death event reports changes over age to

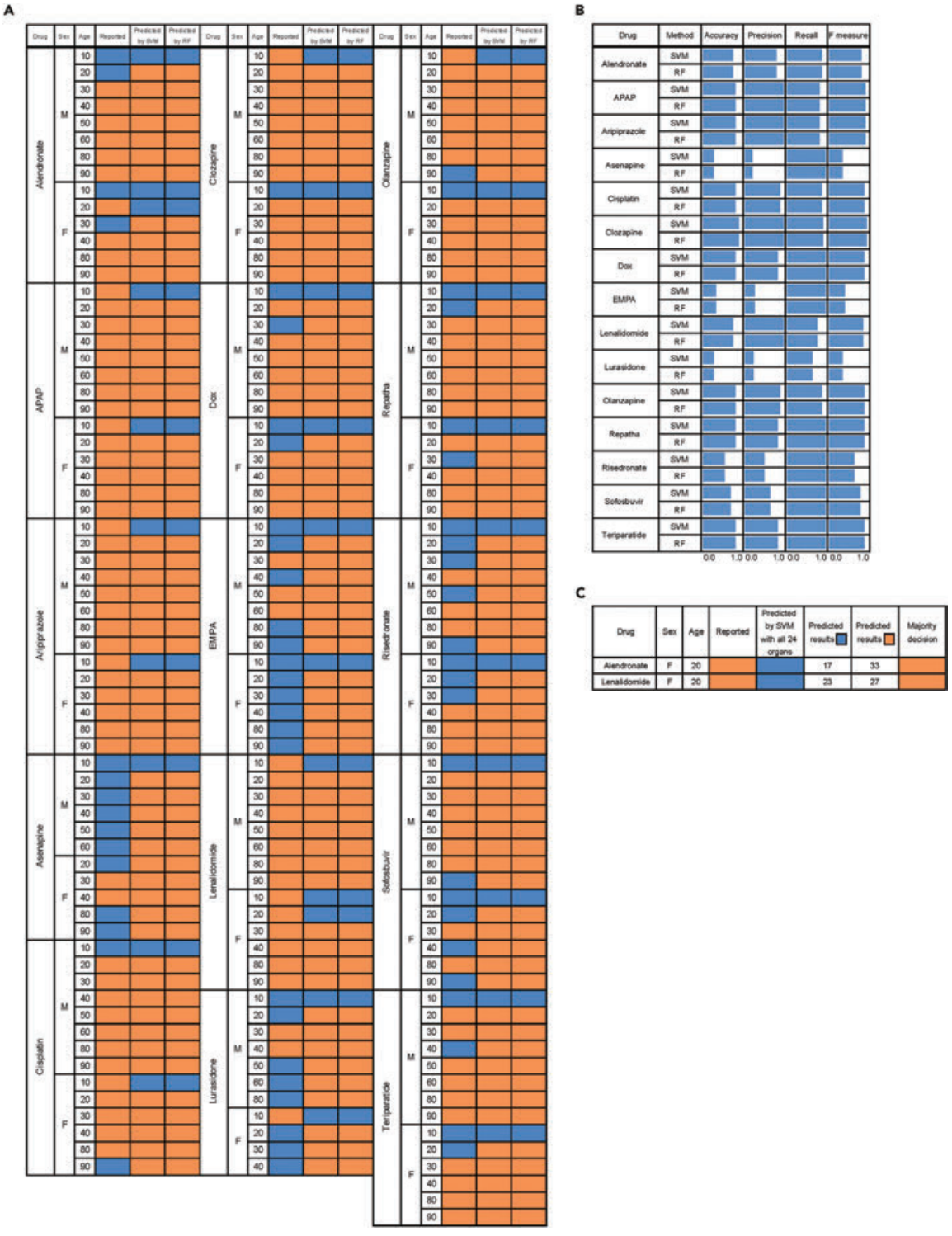

A
B
C
Drug
Method
Accuracy
Precision
Recall
F measure
Alendronate
APAP
Aripiprazole
Asenapine
Cisplatin
Clozapine
Dox
EMPA
Lenalidomide
Lurasidone
Olanzapine
Repatha
Risedronate
Sofosbuvir
Teriparatide
SVM
RF
Sex
Age
Reported
Predicted by SVM
Predicted by RF
Predicted by SVM with all 24 organs
Predicted results
Majority decision
17
33
23
27

Figure 4. The Comparison of Death Event Prediction Results by SVM and RF Algorithms

(A) The prediction results of death event for each drug by SVM and RF algorithms are summarized for each age group (Age) and sex (Sex). The reported outcomes (Reported) are shown as orange and light-blue filled boxes indicating death event is reported and not-reported, respectively. The prediction results (Predicted by SVM/RF) are shown as orange and light-blue filled boxes indicating death event is predicted to occur and not-to-occur, respectively. The age groups are 10: 10–19 years old, 20: 20–29 years old, 30: 30–39 years old, 40: 40–49 years old, 50: 50–59 years old, 60: 60–69 years old, 70: 70–79 years old, 80: 80–89 years old, 90: 90–99 years old. F, females; M, males.

(B) The prediction results for each drug by SVM and RF are summarized as a confusion matrix. Scores of accuracy (Accuracy), precision (Precision), recall (Recall), and F-measure (F measure) for each drug (Drug) are indicated by bars (light blue) for each age group (Age) and sex (Sex). The scale of the scores (0.0–1.0) are indicated at the bottom. APAP, Acetaminophen; Dox, doxycycline; EMPA, empagliflozin; Repatha, evolocumab. The results of the RF prediction are also available as Table S5.

(C) Majority decision approach for death event prediction. "Voting" (i.e., majority decision) by both SVM and RF methods using the 24-organ and all individual-organ transcriptome data are conducted. The reported outcomes (Reported) are shown as orange filled box indicating death event is reported. The prediction results with the 24-organ data (Predicted by SVM with all 24 organs) are shown as light-blue filled boxes indicating "death event" is predicted to not-to-occur. The majority decisions (Majority decision) by counting all votes are shown as orange filled box voting for death event to occur. The number of "votes" for "to-occur" (Predicted results) and "not-to-occur" (Predicted results) is indicated by the numbers.

variable degrees for each drug ("Reported" in Figure 6). There also appears that the sex difference exists (F versus M in Figure 6). The comparison of the predictions by both algorithms (SVR and RFR) to the reported (Reported) incidence (death/all reported AEs) shows that the predictions by both algorithms efficiently capture the general trends of the quantitative changes of death event incidence over age for drugs/sex such as alendronate/female, doxycycline/male, aripiprazole/female (Figure 6, Table S9). On the other hand, the predictions for other drugs (e.g., clozapine: both sexes, empagliflozin:female, teriparatide:female) appear less efficient (Figure 6, Table S9). Although further improvements are necessary, the result suggests that the approach could be exploited to select specific sex/age groups as target subjects in clinical trials of at least some drugs.

Predicting Therapeutic Indications

Identifying an appropriate therapeutic indication(s) (TIs) is also critical in drug development. Hence, we evaluated the utility of the multi-organ transcriptome datasets for predicting TIs. For this purpose, we adapted the LP framework (Figure 7A). Three training datasets are used: (1) the 24-organ transcriptome data of each drug of the 15 drugs, (2) all indications reported for each of the 15 drugs in FAERS, and (3) incidence of all AEs reported in FAERS for each of indications in FAERS (Figure 7A). In the LP framework, one-class SVM algorithm is used to train the models and the presence/absence of a link of an untrained drug (Drug Candidate X in Figure 7A), based on its 24-organ transcriptome data, to each TI is determined (see Transparent Methods for the detailed description, Figure 7A).

We evaluated the performance of this framework by omitting the data of one drug from all training datasets at a time and repeating this for all 15 drugs to evaluate the performance of predicting potential TIs of each omitted drug (Figure 7B). The result is shown as a confusion matrix (Figure 7B), and the full list is available as Table S10. The accuracy scores are high (>0.78) for all 15 drugs (Figure 7B), indicating that more than 78% of the indications or non-indications predicted are indeed reported or not to be reported as the indications for each drug, respectively. The recall scores are also high (>0.8) for alendronate, aripiprazole, asenapine, clozapine, empagliflozin, lurasidone, olanzapine, evolocumab (Repatha), risedronate, sofosbuvir, and teriparatide (Figure 7B), indicating that over 80% of the reported indications are predicted for these 11 drugs by the method. The recall score of doxycycline is 0.527 (Figure 7B), indicating approximately 50% of the reported indications are predicted for this drug by the method. The recall scores are low for acetaminophen (0.141), cisplatin (0), and lenalidomide (0) (Figure 7B), indicating that this method fails to capture many reported indications for these drugs, as expressed by relatively smaller number of false-negatives (FNs) as compared with that of true-positives (TPs) for these drugs (Figure 7B). Only acetaminophen (APAP) shows high precision score (1.000), and all the others show low precision scores (<0.35). Both cisplatin and lenalidomide show 0 TP and 0 FN, thus the precision and the F-measure scores were unable to be calculated (Figure 7B). Such low precision scores for many drugs are mainly due to a large number of false-positives (FPs) as compared with that of TPs (Figure 7B). The result illustrates a possibly useful application of this LP framework to the prediction of drug TIs (see further in Discussion).

Application to Drug Repositioning

We also evaluated the utility of the multi-organ transcriptome datasets for drug repositioning. In the scheme described in Figure 7B, the FP TIs (yellow highlight in Figure 7B) (the full list as Table S10)

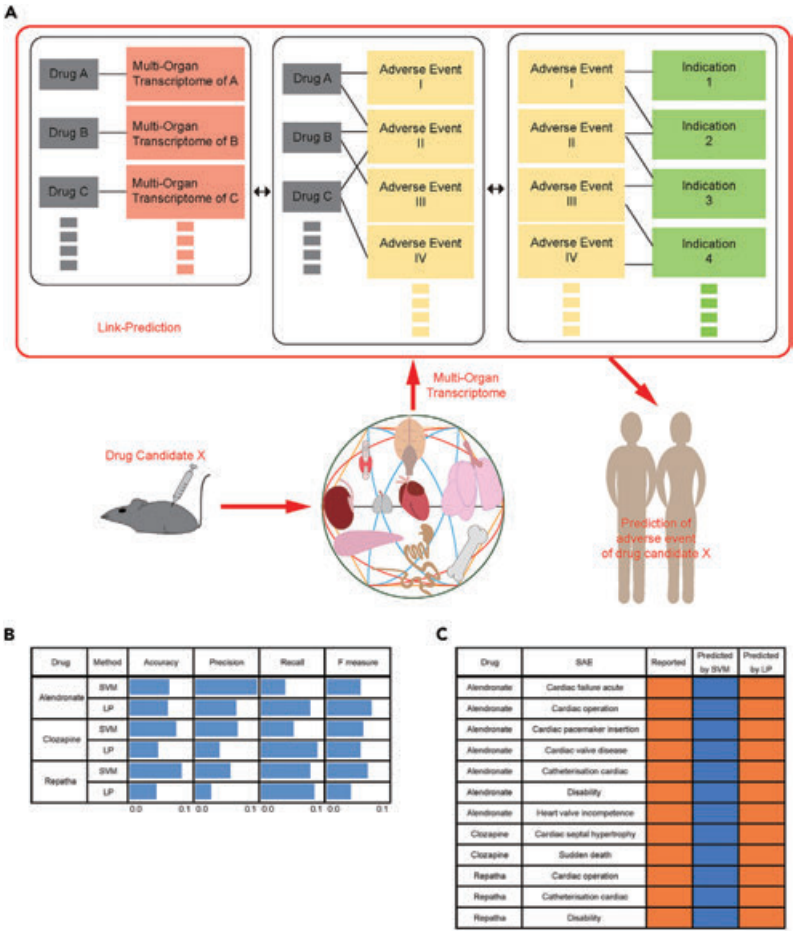


Figure 5. AE Prediction by LP Framework

(A) LP framework to predict AEs. See the text for the description.

(B) The comparison of all AEs prediction by SVM/hMDB-i and the LP approaches. The results are summarized as a confusion matrix. Scores of accuracy (Accuracy), precision (Precision), recall (Recall), and F-measure (F measure) for each drug (Drug) are indicated by bars (light blue) for each method (Methods). The scale of the scores (0.0–1.0) are indicated at the bottom. SVM, SVM/hMDB-i framework; LP, LP framework. The raw data are also available as Table S6.

(C) The list of SAEs that are correctly predicted for alendronate, clozapine, and evolocumab (Repatha) by the LP but not by SVM/hMDB-i framework. The reported outcomes (Reported) are shown as orange and light-blue filled boxes indicating the SAE is reported and not-reported, respectively. The prediction results (Predicted by SVM/LP) are shown as orange and light-blue filled boxes indicating the SAE is predicted to occur and not-to-occur, respectively. The results are also available as Table S6.

could include repositioning targets for the drugs (see further in Discussion). However, in drug repositioning, drug X does not exist. Hence, we used the same LP framework but did not omit the data of any drugs, i.e., the datasets of the drug of the prediction target is also included in the training datasets (Figure 7C).

The result is shown as a confusion matrix (Figure 7C) and the full list is available as Table S11. The result shows the increased number of TPs and decreased number of FNs for all drugs, resulting in the improved recall scores (Figure 7C). In this scheme, both accuracy and recall scores for all drugs are 0.770–1.000 (Figure 7C), indicating the approach can capture over 77% of both reported and non-reported indications. The precision scores for all drugs remain relatively low due to the large number of FPs as compared with that of

Organ	Pathway	Entities FDR
Stomach	Detoxification of reactive oxygen species	0.00038
AdrenalG	PPARA activates gene expression	0.00062
AdrenalG	Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	0.00062
AdrenalG	Lipophagy	0.00282
WAT	Chaperone mediated autophagy	0.00282

Table 1. Enriched Pathways in REACTOME
The top 5-organ pathways for aripiprazole/male/50s are shown.

TPs (Figure 7C and Table S11). The indications found in FP (yellow highlight in Figure 7C) (the full list as Table S11) could include repositioning targets (see further in Discussion).

As the number of FPs are relatively large, calculating decision function values of each indication found in FP in both schemes could facilitate the ranking of each indication according to their likeliness as repositioning targets. Hence, the decision function values of all indications in both Figures 7B and 7C are shown in Tables S10 and S11. These tables could serve as a useful reference for selecting potential repositioning targets for further investigational evaluations (see further in Discussion).

We also validated our indication prediction with SIDER4.1 version. The result, 0.402–0.998 accuracy and 0.135–1 recall (omitting cisplatin and lenalidomide for the reason described above for the FAERS result in the leave-one-out cross-validation), is comparable with that with FAERS (Figures 7B and 7C, the full and raw data for the confusion matrix are available as Tables S12 and S13). The precision is 0.003–0.069, appears slightly lower than those with FAERS (Figures 7B and 7C), owing to the larger number of FPs. This may be caused by the larger total number of indications reported in FAERS (11,312 indications) as compared with that in SIDER (2,088 indications), contributing to the reduced precision with SIDER reference.

Side-by-Side Comparisons of Our Methods to Other Multi-Features-Based Prediction Methods

Some other *in silico* approaches predicting SEs/AEs are previously reported (Pauwels et al., 2011; Wang et al., 2016). In particular, Wang et al. reported an SEs/AEs prediction method by combining multi-features (cell morphological features, MACCS chemical fragments, L1000 landmark genes) with adverse drug reactions labels (http://maayanlab.net/SEP-L1000/index.html#) (Wang et al., 2016). We compared the SEs/AEs prediction performance of our method with theirs side by side (Figure 8A, Table S14). Of 15 drugs that we analyzed, 9 drugs (alendronate, aripiprazole, acetaminophen, asenapine, cisplatin, clozapine, doxycycline, lenalidomide, olanzapine) are covered by their method. Their method is applicable to only small-molecule drugs, hence SEs/AEs of biologics/antibody drugs (evolocumab/Repatha in our study) or peptide drugs (teriparatide in our study) can be predicted only by our method, illustrating the advantage of our drug modality-independent approach.

We performed side-by-side comparisons of our method with theirs with these nine drugs represented by both methods (Figure 8A, Table S14). With seven drugs (alendronate, acetaminophen, asenapine, cisplatin, clozapine, doxycycline, olanzapine), our method detects far more SEs/AEs labeled in SIDER or the processed FAERS (Figure 8A, Table S14). Even with the other two drugs (aripiprazole, lenalidomide), our method identifies a large number of SEs/AEs (aripiprazole: 93 in SIDER and 37 in the processed FAERS, lenalidomide: 71 in SIDER and 50 in processed FAERS) that are missed by their method (Figure 8A, Table S14).

Prediction of TIs of small-molecule drugs on the basis of their structural information, targets, and interactions is previously reported (Li and Lu, 2012). We performed side-by-side comparisons of our method to theirs in predicting TIs (Figure 8B, Table S15). Of 15 drugs that we studied, 6 drugs (alendronate,

Organ	Pathway	P, Adjusted
Spleen	Glycosphingolipid biosynthesis—ganglio series	0.00174
Eye	Salivary secretion	0.00904
AdrenalG	PPAR signaling pathway	0.01008
Stomach	Glutathione metabolism	0.01032
Stomach	Thyroid hormone synthesis	0.01032
Stomach	Arachidonic acid metabolism	0.01032
ThyroidG	Proteasome	0.01066
ThyroidG	Epstein-Barr virus infection	0.02666

Table 2. Enriched Pathways in KEGG
The top 8-organ pathways for aripiprazole/male/50s are shown.

acetaminophen, cisplatin, clozapine, doxycycline, risedronate) are covered by their method. Like the SEs/AEs prediction, our method is modality free; hence, we can predict TIs of biologics/antibody drugs (evolocumab/Repatha) and peptide drugs (teriparatide), which their small-molecule-based prediction method fails. This again illustrates the advantage of our drug modality-independent approach. We performed side-by-side comparisons of our method to theirs with these six drugs represented by both methods (Figure 8B, Table S15) and show that our method detects far more TIs labeled in SIDER or FAERS with all six drugs (Figure 8B, Table S15).

DISCUSSION

The difficulty of predicting human clinical outcomes during pre-clinical studies is the major hindrance in drug development. This is mainly due to the animal models versus human difference in the outcomes. To overcome this problem, numerous experimental and computational approaches are currently used (Elitt et al., 2018; Ko and Gelb, 2014; Meseguer-Ripolles et al., 2018; Oleaga et al., 2016; Rezaei Kolahchi et al., 2016; Shah et al., 2019; Tateno et al., 2004; Vamathevan et al., 2019). Some approaches use transcriptome and/or other omics data to predict clinical outcomes (Ho et al., 2016). However, they use only cells and limited types of organs/tissues and none exploit the whole-body level biological features such as the body-wide multi-organ transcriptome data for clinical outcome prediction.

In the current study, we exploit multi-organ transcriptome datasets in mice to predict human clinical outcomes (Figure 1). We "humanize" the mouse data by training machine learning models with multi-organ transcriptome data derived from the mouse and human clinical datasets (Figure 1). Our evaluation illustrates advantages and effectiveness of this approach in several applications to the outcome prediction of both AEs and TIs (Figures 2, 3, 4, 5, 6, and 7).

Several other *in silico* approaches predicting SEs/AEs and TIs have been previously reported (Li and Lu, 2012; Pauwels et al., 2011; Wang et al., 2016). The side-by-side comparisons of our methods with theirs shows that our method detects far more SEs/AEs and TIs than theirs with most of the 15 drugs evaluated here (Figure 8). Even with the drugs that both methods are comparable (aripiprazole and lenalidomide in the SE/AE prediction), our method detected the outcomes that are missed by theirs (Figure 8A). These results illustrate a usefulness of exploiting multi-organ transcriptome datasets in predicting drug SEs/AEs and TIs.

Potential associations of SEs with diseases are computed on the basis of drug fingerprints/targets and their known SEs/potential disease indications and reported as an integrated table (https://astro.temple.edu/~tua87106/druganalysis.html) (Zhang et al., 2013). Although, by using this table, they suggest that drug-repositioning candidates could be identified, they fail to explicitly tabulate specific drug-repositioning targets or validate them. In this paper, we exploit potential SE/AE-TI association and combined it with our body-wide multi-organ transcriptome approach in the LP framework to compute drug TIs and explicitly tabulate them (Figure 7). We, furthermore, evaluate the effectiveness and the efficiency of our method by cross-validation (Figure 7B).

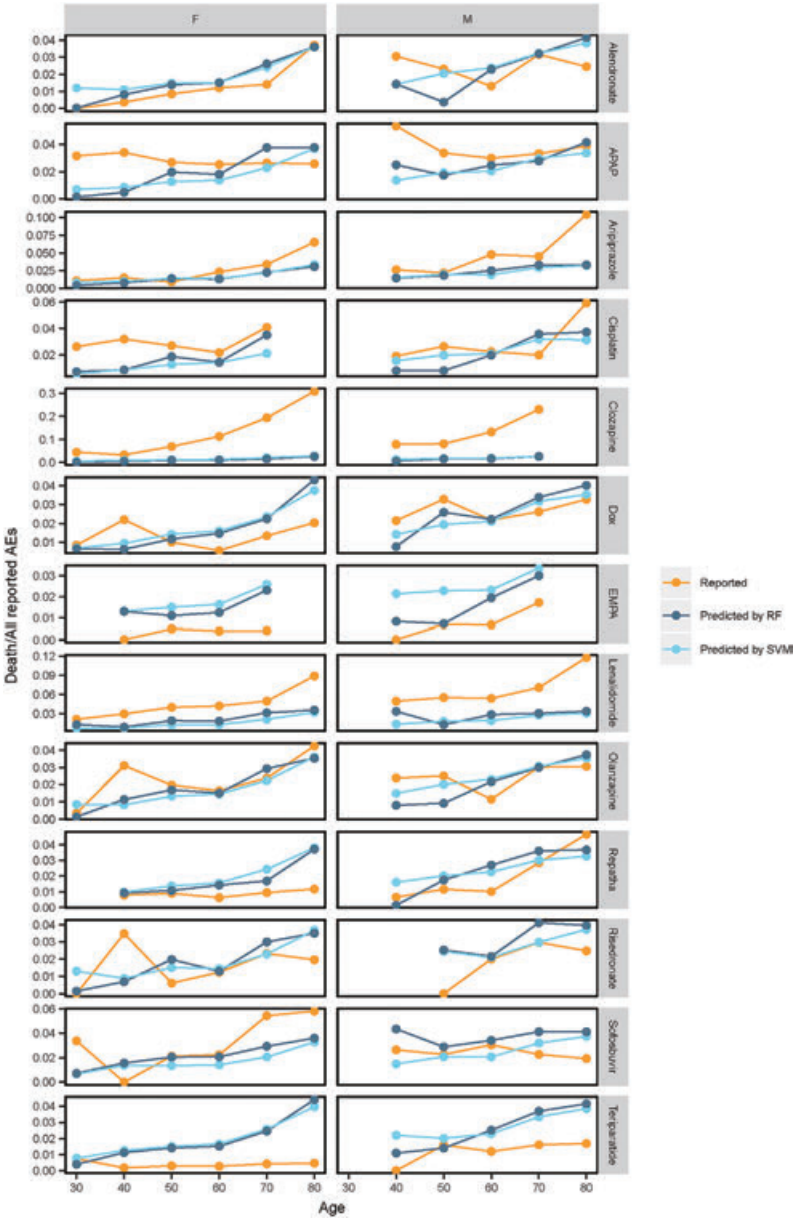
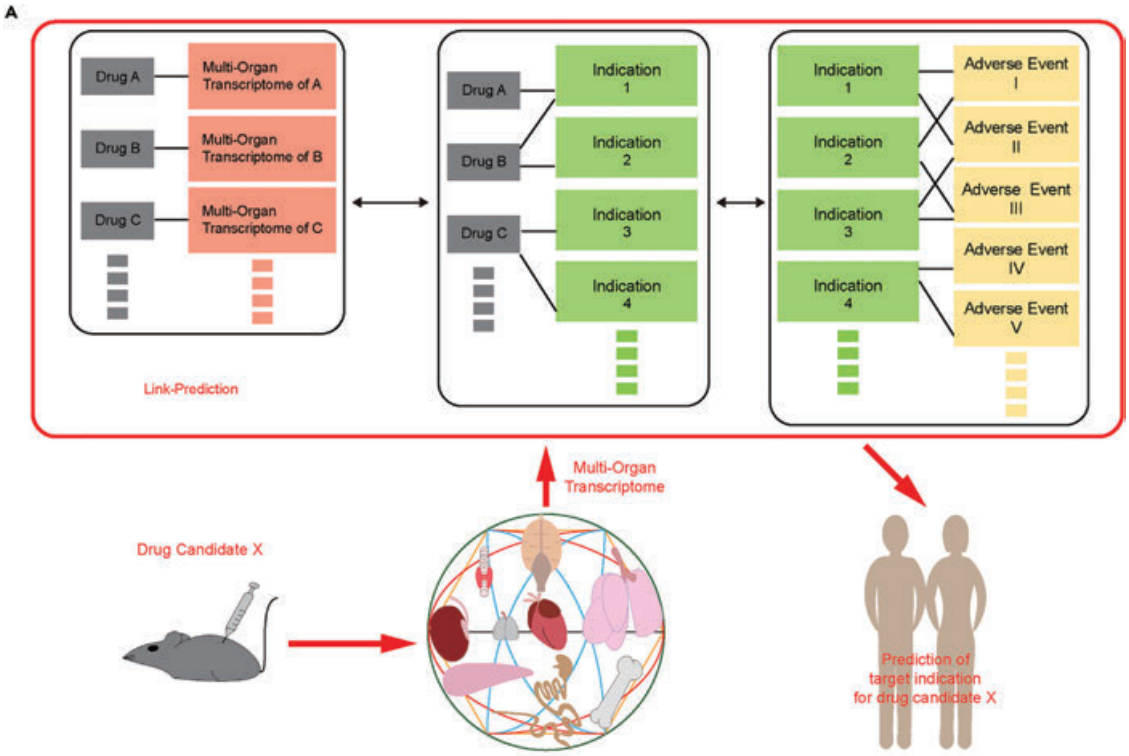
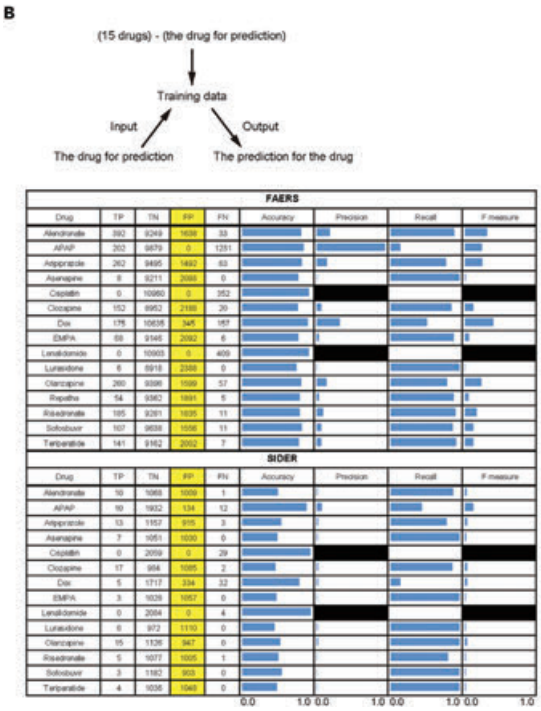


Figure 6. Prediction of Quantitative Dynamics of Death Events Over Age

The results of predicting quantitative dynamics of death events at each age group for each drug are summarized. The results by SVM regressor (SVM) and RF regressor (RF) are shown. The age groups are 30: 30–39 years old, 40: 40–49 years old, 50: 50–59 years old, 60: 60–69 years old, 70: 70–79 years old, 80: 80–89 years old. The results are also available as Table S9.

There are two major advantages of our approach over the other *in silico* prediction methods. The first is its drug-modality independence. This is illustrated by the effective predictions for small-molecule (alendronate, acetaminophen, aripiprazole, asenapine, cisplatin, clozapine, doxycycline, empagliflozin, lenalidomide, lurasidone, olanzapine, risedronate, sofosbuvir), antibody (evolocumab/Repatha), and peptide (teriparatide) drugs. In principle, our approach is expected to be as effective for other modalities such as nucleic acid, gene, and cell therapeutics and also for their mixtures, as the approach does not use


A
Drug A
Drug B
Drug C
Multi-Organ Transcriptome of A
Multi-Organ Transcriptome of B
Multi-Organ Transcriptome of C
Link-Prediction
Indication 1
Indication 2
Indication 3
Indication 4
Adverse Event I
Adverse Event II
Adverse Event III
Adverse Event IV
Adverse Event V
Multi-Organ Transcriptome
Drug Candidate X
Prediction of target indication for drug candidate X


B
(15 drugs) - (the drug for prediction)
Training data
Input
Output
The drug for prediction
The prediction for the drug
FAERS
SIDER

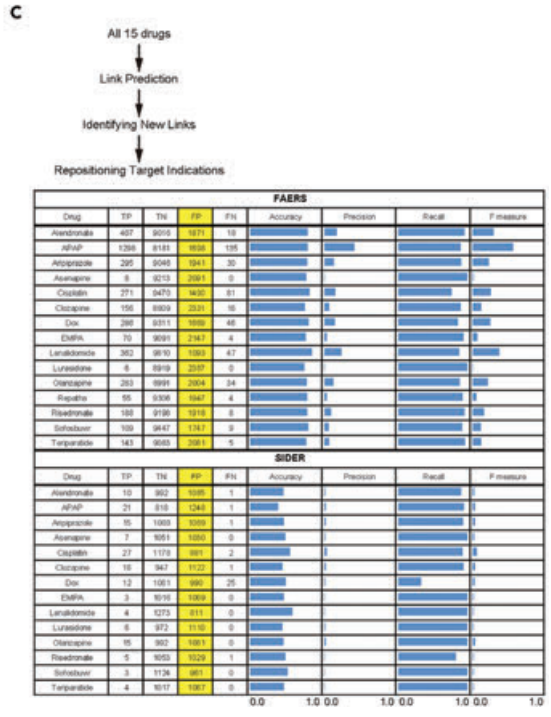

C
All 15 drugs
Link Prediction
Identifying New Links
Repositioning Target Indications
FAERS
SIDER

Figure 7. Prediction of Therapeutic Indications

(A) LP framework to predict therapeutic indications. See text for the detailed description.

(B) Cross-validation with FAERS and SIDER references. The hold-out scheme is shown at the top. The results are shown as a confusion matrix. The scores are indicated as light blue bars. Scores of precision and F-measure for cisplatin and lenalidomide are incalculable (black filled cells) as both TP and FP scores are 0. The results are also available as Tables S10 (FAERS) and S12 (SIDER).

(C) Repositioning prediction. The scheme is shown at the top. The results with FAERS and SIDER references are shown as a confusion matrix. The scores are indicated as light blue bars. TP, true positive; TN, true negative; FP, false positive; FN, false negative. The scale of the scores (0.0–1.0) are indicated at the bottom. FP as potential repositioning targets are yellow highlighted. The results are also available as Tables S11 (FAERS) and S13 (SIDER).

any structural features of the drugs for the predictions. The other advantage is the fact that the approach requires no prior knowledge about the molecular or cellular targets of the drugs or the mechanisms of their actions. Hence, the only requirement of our approach is that the test drug can be administered to the mouse.

In our approach, the effective prediction is critically dependent on identifying common transcriptome features between the drugs in the models and the target/test drug (i.e., drug X). The 15-drug group in the current study consists of those with diverse AE and indication patterns. Despite such complex clinical outcome patterns, the datasets from such a small number of drugs are surprisingly sufficient to effectively predict the outcomes of many AEs and TIs for most of the drugs in the cross-validation studies (Figures 2, 3, 4, 5, 6, 7, and 8). This may be due to the presence of many possibly useful features in the multi-organ transcriptome data for the predictions that could not be identified in the transcriptome or other omics data of the cells and the limited types of organs/tissues.

Although the use of the 24-organ transcriptome data (i.e., hMDB-i approach) is generally effective in predicting AEs, we find the differential performance by using all 24-organ data together or using them as individual-organ datasets (Figure 3). We also observed differential performance depending on which of the 24 individual-organ datasets is used (Figure 3). To address these issues, we suggest that majority decision approach could be beneficial and illustrate such examples (Figure 4C). Furthermore, the use of hMDB-i approach without and within the LP framework appears to influence the prediction consequence (Figures 5B and 5C). The accuracy and precision scores are better without the LP framework, but the recall scores are better with the LP framework. Hence, it may be beneficial to select the framework (i.e., with/without the LP framework) according to the objective of the prediction (e.g., to be more confident about the predicted AEs versus to capture all possible AEs).

Drug repositioning is a strategy to develop the existing drug for an additional therapeutic indication(s) (Pushpakom et al., 2019). The advantage is that the properties (such as pharmacokinetics) of such drugs are already determined. Furthermore, their safety in human is secured; hence, phase I trials could often be bypassed (Pushpakom et al., 2019). Such advantage of drug repositioning results in the overall cost- and time-saving in the development, as compared with new drug development (Pushpakom et al., 2019). We examined the potential utility of our approach to predict TIs for drug repositioning (Figure 7). The potential therapeutic targets for the repositioning of each drug are those that are labeled as "FP" in Figures 7B and 7C and Tables S10 and S11. The examination of the list identified many common FP indications shared among the drugs of the same or related mechanisms of actions (Tables S10 and S11). For alendronate and risedronate, both bisphosphonate medications for treating osteoporosis, Addison's disease and amyloidosis are among those predicated as FP targets with high decision function values. For clozapine and aripiprazole, both atypical anti-psychotic medications antagonizing related neurotransmitter receptors, coagulopathy and ulcer are examples of such FP targets. Although their investigational validations must await the future clinical trials, such targets shared among the related medications could be among the leading candidates for the repositioning.

In the study of predicting TIs, we show that both cisplatin and lenalidomide failed to identify true positive targets (Figure 7B). On the other hand, including these two drugs in the training data for the same prediction framework resulted in the improved identification of the true-positive targets (Figure 7C). These results may suggest that the gene expression patterns of these two drugs include distinguishing features that are unique to each of these two drugs; hence, such features could not be exploited by the scheme shown in Figure 7B (i.e., the data of these drugs are omitted from the training data for their predictions). This point is supported by the fact that the inclusion of the transcriptome data of these two drugs in the training datasets lead to the effective outcome predictions for the other drugs (Figure 7B). Furthermore, such unique

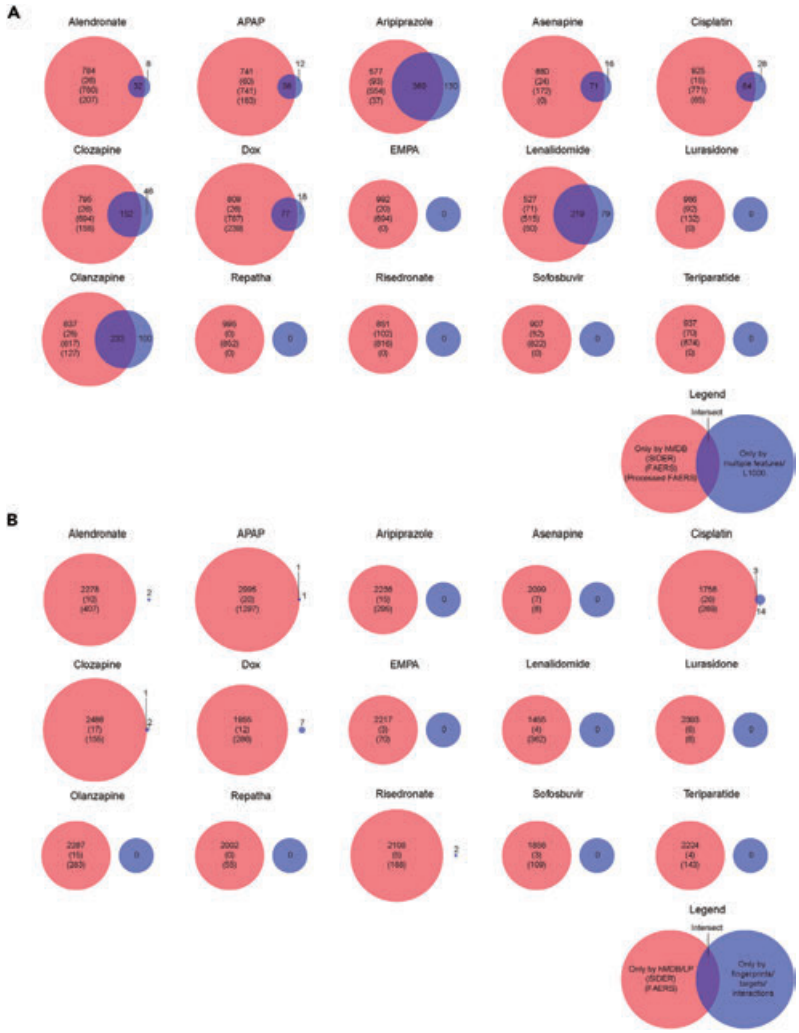


Figure 8. Side-by-Side Comparisons of Our Methods with Other Multi-Features-Based Prediction Methods

The comparisons are shown as Venn diagrams for SE/AE (A) and TI (B) predictions. The description of the diagrams is shown as "Legend" diagram located at the right bottom in each panel. The number of the SEs/AEs and TIs predicted only by our method (A: hMDB, B: hMDB/LP), only by the existing methods (A: multiple-features/L1000, B: fingerprints/targets/Interactions), and by both methods are indicated accordingly in the diagrams. Our methods (A: hMDB, B: hMDB/LP) are compared with multiple-features/L1000 method (A) and fingerprints/targets/Interactions method (B), respectively. The predictions for the drugs that are not represented or cannot be predicted by the other methods are indicated as 0. Of SEs/AEs and TIs predicted only by our methods, those labeled in SIDER, FAERS, and the processed FAERS databases are also indicated. SEs for evolocumab/Repatha do not exist in SIDER; hence, it is indicated as 0. The table forms of the Venn diagrams and their full searchable raw data are available as Tables S14 (A) and S15 (B).

features can be effectively exploited when included in the training data, resulting in the successful identification of the true-positives (Figure 7C).

In drug development and clinical trials, it would be important and beneficial to understand the mechanisms of the outcomes. Such mechanistic understanding could serve as a basis for further improvement on the drug designs and also designing the pre-clinical and clinical studies. Furthermore, availability of the mechanisms facilitates the regulatory approval processes and also improves ethical responsibilities providing an explanation to the study subjects and patients. Using RF algorithm allows us to identify

specific molecular/cellular pathways that contributed to the AE predictions of the drug, and the examples are shown in Tables 1 and 2 and the full lists are available in Tables S7 and S8. Although such pathways are those modulated by the drugs in mice and may not necessarily be the same in human, many are conserved across animal species. Hence, they could serve as at least, if not at all, as clues for deducing possible mechanisms underlying the AEs and/or generating working hypothesis for further experimental and clinical studies for the validation.

The current study illustrates beneficial utilities of applying body-wide multi-organ transcriptome data obtained from mice to predict human clinical outcomes. The further inclusion of other drugs' multi-organ transcriptome and their clinical outcome data in hMDB-i and other relevant data, such as individuals' genotypes and lifestyle information, as training datasets may improve the current performance where some limitations are observed with some drugs and/or some outcomes. With such refinements in the training datasets, further improvements in the individualized predictions is expected in the future, contributing to the realization of the ultimate virtual precision medicine.

Limitations of the Study

In the prediction study of therapeutic indications, we used all therapeutic indications associated with each drug in FAERS as training data. Thus, no distinction can be made between the effective and ineffective treatments. Furthermore, all indications for which each drug was used are included as the training data for that particular drug, i.e., the indications for which other drug(s) was/were used for the purpose of treatments and drug X as simply an accompanying drug were included as "reported" therapeutic indications for drug X in the training data. For example, acetaminophen is used to attenuate pain associated with osteoporosis; hence, both pain and osteoporosis were included as the indications for this drug. Furthermore, we also used all AE labels associated with each drug and/or indication. Hence, no distinctions were made between those of causative relationships and of simple association. Currently, using FAERS, it is difficult to make such distinctions for all drugs in a systematic manner. Some distinction may be possible by using other public database such as clinicaltrials.gov, but the number of reports is limited and not sufficient for our machine-learning approach. Despite such compromise, the fact that our approach predicts many human clinical outcomes according to the quality of the input training data indicates that, with further refinement in the input training datasets, it is reasonable to expect that our method generates more sophisticated predictions. In fact, we validated our predictions with a drug-label based reference, SIDER. The results show that both SE/AE and TI predictions with SIDER are comparable with those with FAERS (Figures 2D, 7B, and 7C, Tables S3, S12, and S13). These results further support the effectiveness and usefulness of our approach.

Owing to the experimental cost necessary to perform the 24-organ transcriptome analyses, the current study is limited to the evaluation of 15 drugs. There are 3,732 approved drugs (https://www.drugbank.ca/stats), and it would be extremely expensive to perform the 24-organ transcriptome analyses with most, if not all, of these drugs. With the current cost for the transcriptome analyses, we estimate such cost would be US$20–30 million. This is expensive for one laboratory scale but possibly realistic as a consortium effort. Considering the expensiveness of drug development, this might be a worthy investment. In fact, we may not need all drug data, as we show even the 15-drug datasets were sufficient for many of the outcome predictions, despite their diverse modalities and TIs (Figures 2, 3, 4, 5, 6, 7, and 8).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

The accession number for the RNA-seq data reported in this paper is GEO: GSE142068. Hyperparameters are available at https://www.hmdb.karydo-tx.com/. The code for our algorithm is available for non-profit use with Material Transfer Agreement.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.100791.

ACKNOWLEDGMENTS

We thank S. Keiko, R. Takahashi, and E. Kojima for administrative assistance. We are also grateful to the members of Karydo TherapeutiX, Inc. and Sato lab at ATR for advice and discussion throughout the course of this work. This work was supported by funds from Public Interest Incorporated Foundation KYOTO Industrial Support Organization 21 (T.N.S.) and JST ERATO Grant Number JPMJER1303 (T.N.S.).

AUTHOR CONTRIBUTIONS

T.N.S. conceived the idea of the project, designed the overall experiments and *in silico* analyses, and supervised the overall research project. S.K. designed and performed the *in silico* analyses. F.S., S.E., G.M.D.A., and Y.M. contributed to designing the experiments and performed the experiments. T.N.S., S.K., F.S., and S.E. wrote the manuscript.

DECLARATION OF INTERESTS

T.N.S. is the inventor of the patents filed on hMDB-i and is the scientific founder of and holds shares of Karydo TherapeutiX, Inc., which owns the right to this patent. S.K., S.F., S.E., G.M.D.A., and Y.M. are employees of Karydo TherapeutiX, Inc. Correspondence and requests for materials should be addressed to T.N.S (island1005@gmail.com).

Received: August 31, 2019
Revised: October 17, 2019
Accepted: December 16, 2019
Published: January 9, 2020

REFERENCES

Berger, S.I., and Iyengar, R. (2011). Role of systems pharmacology in understanding drug adverse events. Wiley Interdiscip. Rev. Syst. Biol. Med. *3*, 129–135.

Droujinine, I.A., and Perrimon, N. (2013). Defining the interorgan communication network: systemic coordination of organismal cellular processes under homeostasis and localized stress. Front. Cell. Infect. Microbiol. *3*, 82.

Droujinine, I.A., and Perrimon, N. (2016). Interorgan communication pathways in physiology: focus on Drosophila. Annu. Rev. Genet. *50*, 539–570.

Elitt, M.S., Barbar, L., and Tesar, P.J. (2018). Drug screening for human genetic diseases using iPSC models. Hum. Mol. Genet. *27*, R89–R98.

Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P., and Shah, N.H. (2013). Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin. Pharmacol. Ther. *93*, 539–546.

Ho, T.B., Le, L., Thai, D.T., and Taewijit, S. (2016). Data-driven approach to detect and predict adverse drug reactions. Curr. Pharm. Des. *22*, 3498–3526.

Ko, H.C., and Gelb, B.D. (2014). Concise review: drug discovery in the age of the induced pluripotent stem cell. Stem Cells Transl. Med. *3*, 500–509.

Kozawa, S., Ueda, R., Urayama, K., Sagawa, F., Endo, S., Shiizaki, K., Kurosu, H., Maria de Almeida, G., Hasan, S.M., Nakazato, K., et al. (2018). The body-wide transcriptome landscape of disease models. iScience *2*, 238–268.

Li, J., and Lu, Z. (2012). A new method for computational drug repositioning using drug pairwise similarity. Proceedings (IEEE Int. Conf. Bioinformatics Biomed.) *2012*, 1–4.

Meseguer-Ripolles, J., Khetani, S.R., Blanco, J.G., Iredale, M., and Hay, D.C. (2018). Correction to: pluripotent stem cell-derived human tissue: platforms to evaluate drug metabolism and safety. AAPS J. *20*, 30.

Morgan, S., Grootendorst, P., Lexchin, J., Cunningham, C., and Greyson, D. (2011). The cost of drug development: a systematic review. Health Policy *100*, 4–17.

Oleaga, C., Bernabini, C., Smith, A.S., Srinivasan, B., Jackson, M., McLamb, W., Platt, V., Bridges, R., Cai, Y., Santhanam, N., et al. (2016). Multi-Organ toxicity demonstration in a functional human in vitro system composed of four organs. Sci. Rep. *6*, 20030.

Pauwels, E., Stoven, V., and Yamanishi, Y. (2011). Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics *12*, 169.

Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., et al. (2019). Drug repurposing: progress, challenges and recommendations. Nat. Rev. Drug Discov. *18*, 41–58.

Rezaei Kolahchi, A., Khadem Mohtaram, N., Pezeshgi Modarres, H., Mohammadi, M.H., Geraili, A., Jafari, P., Akbari, M., and Sanati-Nezhad, A. (2016). Microfluidic-based multi-organ platforms for drug discovery. Micromachines (Basel) *7*, 162.

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., Ringel, M., and Schork, N. (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Digit. Med. *2*, 69.

Sonawane, K.B., Cheng, N., and Hansen, R.A. (2018). Serious adverse drug events reported to the FDA: analysis of the FDA adverse event reporting system 2006-2014 database. J. Manag. Care Spec. Pharm. *24*, 682–690.

Tateno, C., Yoshizane, Y., Saito, N., Kataoka, M., Utoh, R., Yamasaki, C., Tachibana, A., Soeno, Y., Asahina, K., Hino, H., et al. (2004). Near completely humanized liver in mice shows human-type metabolic responses to drugs. Am. J. Pathol. *165*, 901–912.

Tatonetti, N.P., Ye, P.P., Daneshjou, R., and Altman, R.B. (2012). Data-driven prediction of drug effects and interactions. Sci. Transl. Med. *4*, 125ra131.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.

Wang, Z., Clark, N.R., and Ma'ayan, A. (2016). Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics *32*, 2338–2345.

Wong, C.H., Siah, K.W., and Lo, A.W. (2019). Estimation of clinical trial success rates and related parameters. Biostatistics *20*, 273–286.

Zhang, P., Wang, F., Hu, J., and Sorrentino, R. (2013). Exploring the relationship between drug side-effects and therapeutic indications. AMIA Annu. Symp. Proc. *2013*, 1568–1577.

iScience, Volume *23*

Supplemental Information

Predicting Human Clinical Outcomes

Using Mouse Multi-Organ Transcriptome

Satoshi Kozawa, Fumihiko Sagawa, Satsuki Endo, Glicia Maria De Almeida, Yuto Mitsuishi, and Thomas N. Sato

Supplementary Figure

Figure S1

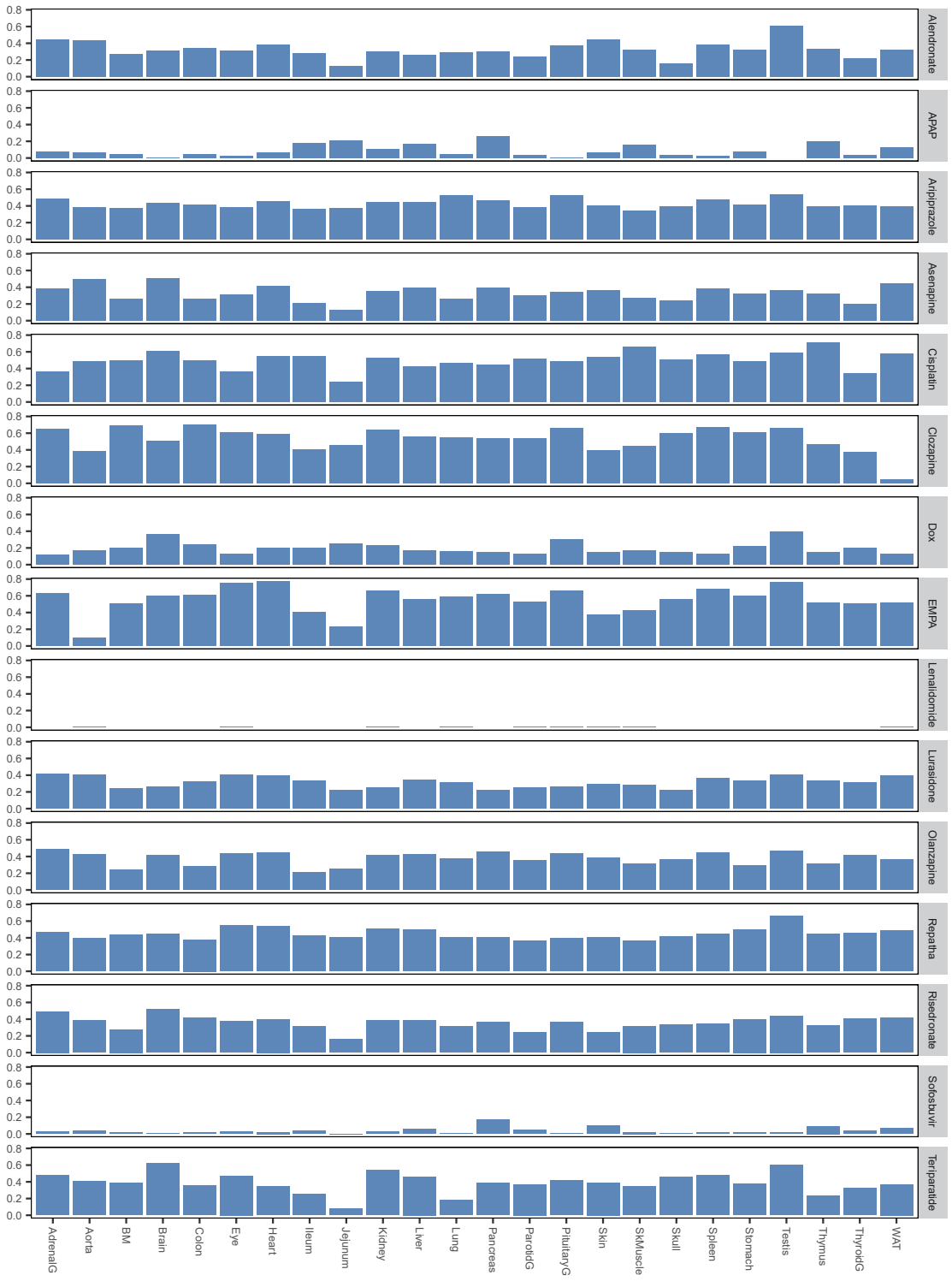


Figure S1. The patterns of differentially expressed genes in each organ of the mice treated with each drug, Related to Figure 1. Y-axis: (the number of differentially expressed genes for the drug) / (the number of differentially expressed genes for all drugs) for each organ. The differentially expressed genes are those with adjusted $p<0.0001$ by DESeq2 analysis. The DESeq2 data are available as Supplementary Table 1. The raw RNAseq sequence data are available at https://www.hmdb.karydo-tx.com/hMDB_freeDL/.

Supplementary Table Legends

Table S1. The DESeq2 data for Figure S1, Related to Figure 1.

DE gene: the number of differentially expressed genes.

Table S2. The raw data of Figure 2B

The binary labelling system is applied to each AE: 1 when the AE is reported/predicted to occur and 0 when the AE is not-reported/predicted not-to-occur. The empty cells indicate that none of the AEs are reported for this drug in this sex/age-group.

Table S3. The raw data of Figure 2D

The binary labelling system is applied to each side effect (SE): 1 when the SE is reported/predicted to occur and 0 when the SE is not-reported/predicted not-to-occur.

Table S4. The raw data of Figure 3

The binary labelling system is applied to each AE: 1 when the AE is reported/predicted to occur and 0 when the AE is not-reported/predicted not-to-occur.

Table S5. The raw data of Figures 4A&B.

The binary labelling system is applied to each AE: 1 when the AE is reported/predicted to occur and 0 when the AE is not-reported/predicted not-to-occur.

Table S6. The raw data of Figures 5B&C

The binary labelling system is applied to each AE: 1 when the AE is reported/predicted to occur and 0 when the AE is not-reported/predicted not-to-occur.

Table S7. The raw data of Table 1

All pathways of FDR<0.05 are included.

Table S8. The raw data of Table 2

All pathways of adjusted $p<0.05$ are included.

Table S9. The raw data of Figure 6

The actual number of reports in FAERS (Reported #) is shown for each drug, AE, sex, age-group in the table.

Table S10. The raw data of Figure 7B (vs. FAERS)

Decision function value for each indication is indicated.

Table S11. The raw data of Figure 7C (vs. FAERS)

Decision function value for each indication is indicated.

Table S12. The raw data of Figure 7B (vs. SIDER)

1 when the TI is reported/predicted to occur and 0 when the TI is not-reported/predicted not-to-occur.

Table S13. The raw data of Figure 7C (vs. SIDER)

1 when the TI is reported/predicted to occur and 0 when the TI is not-reported/predicted not-to-occur.

Table S14. The raw data of Figure 8A

SEs predicted only by our method (highlighted by orange), only by the multiple features/L1000 method (highlighted by sky blue) and by both methods (highlighted by light purple) are indicated.

Table S15. The raw data of Figure 8B

TIs predicted only by our method (highlighted by orange), only by the fingerprints/targets/interactions method (highlighted by sky blue) and by both methods (highlighted by light purple) are indicated.

Transparent Methods

Mice and 24-organ transcriptome analyses

The all mouse studies and the transcriptome analyses are performed as previously reported (Kozawa et al., 2018). The 24 organs are adrenal glands, aorta, bone marrow cells (BM), brain, colon, eyes, heart, ileum, jejunum, left kidney, liver, lung, pancreas, parotid glands, pituitary glands, skeletal muscle, skin, skull, spleen, stomach, left testis, thymus, thyroid glands, and gonadal white adipose tissue (WAT). All animal protocols were approved by the Animal Care and Use Committee of Karydo TherapeutiX, Inc. (Approved Number: AN20170001KTX, AN20170010KTX, AN20170014KTX, AN20180012KTX and AN20190006KTX). All animals were housed in a temperature-controlled room at around 25 °C with 12 h light/dark cycle and allowed free access to water and normal foods (CE-2, CLEA Japan, Inc., Tokyo, Japan). For the transcriptome analyses, we used QuantSeq 3'mRNA-Seq Library Prep Kit for Illumina (FWD) (cat#015.384, LEXOGEN) and Illumina NextSeq 500 (75bp single-read, ca. 400 million reads/run, NextSeq 500/550 High Output Kit v2.5, cat#20024906).

Drug administration and harvesting organs

Alendronate: Male C57BL/6NCrSlc mice at 11 weeks old received subcutaneous injection of alendronate sodium salt trihydrate (Wako) in PBS (Nacalai Tesque) at 1.0 mg/kg dosage every 3 to 4 days for 8 days. The drug was prepared fresh everytime. Organs were harvested in the afternoon on the 8th day after the drug administration.

Acetaminophen: Male C57BL/6NCrSlc mice at 10 weeks old were fasted for 12 hours with free access to water during the whole time. Immediately after the fasting period, mice received single intraperitoneal injection of acetaminophen (Wako) in saline (Otsuka Pharmaceutical) at 300 mg/kg dosage. After the injection, mice were allowed the free access to the normal foods. Organs were harvested 2 hours after the drug administration before noon.

Aripiprazole: Male C57BL/6NJcl mice at 11 weeks old received single intraperitoneal injection of aripiprazole (Sigma-Aldrich) in 0.5%(w/v) carboxymethylcellulose 400 solution (Wako) at 0.3 mg/kg dosage. Organs were harvested 2 hours after the drug administration in the afternoon.

Asenapine: Male C57BL/6NCrSlc mice at 11 weeks old received single subcutaneous injection of asenapine maleate (Chemscene) in saline at 0.3 mg/kg dosage. Organs were harvested 2 hours after the drug administration in the afternoon.

Cisplatin: Male C57BL/6NJcl mice at 11 weeks old received single intraperitoneal injection of cisplatin (Bristol-Myers Squibb) at 20 mg/kg dosage. Organs were harvested in the afternoon on the 3rd day after the drug administration.

Clozapine: Male C57BL/6NJcl mice at 11 weeks old received single subcutaneous injection of clozapine (Sigma-Aldrich) at 0.3 mg/kg dosage. Clozapine was first dissolved in acetic acid before diluted in saline and the pH was adjusted to 6 with 1M NaOH. Organs were harvested 2 hours after the drug administration in the afternoon.

Doxycycline: Male C57BL/6NCrSlc mice at 9 weeks old were allowed to a free access to the doxycycline solution containing 2 mg/mL doxycycline hydrochloride n-hydrate (Wako) and 5% sucrose (Nacalai Tesque) in RO water or a vehicle solution containing 5% sucrose in RO water for 2 weeks. The new bottle was prepared weekly for 2 weeks. Organs were harvested in the afternoon on the 13th day after the drug administration.

Empagliflozin: Male C57BL/6NCrSlc mice at 10 weeks old received oral (gavage) administration of empagliflozin (Toronto research chemicals) in 0.5% carboxymethylcellulose at 10 mg/kg everyday for 2 weeks. The drug was prepared fresh at every administration. Organs were harvested 2 weeks after the drug administration in the afternoon.

Lenalidomide: Male C57BL/6NCrSlc mice at 8 weeks old received oral (gavage) administration of lenalidomide (Wako) in 0.5% carboxymethylcellulose and 0.25% Tween-80 (Nacalai Tesque) at 50 mg/kg or a vehicle solution (0.5% carboxymethylcellulose and 0.25% Tween-80) everyday for 69 days. The drug was prepared fresh at every administration. Organs were harvested 69 days after the drug administration in the afternoon.

Lurasidone: Male C57BL/6NCrSlc mice at 11 weeks old received single oral (gavage) administration of lurasidone hydrochloride (Medchemexpress) in 0.5% carboxymethylcellulose at 0.3 mg/kg. Organs were harvested 2 hours after the drug administration in the afternoon.

Olanzapine: Male C57BL/6NCrSlc mice at 11 weeks old received single oral (gavage) administration of olanzapine (Tokyo Chemical Industry) in 0.5% carboxymethylcellulose at 0.3 mg/kg. Organs were harvested 2 hours after the drug administration in the afternoon.

Evolocumab (Repatha®): Male C57BL/6NJcl mice at 11 weeks old received subcutaneous injection of Repatha® (Astellas) in saline at 10 mg/kg every 10 days for 4 weeks. Organs were harvested 4 weeks after the drug administration in the afternoon.

Risedronate: Male C57BL/6NCrSlc mice at 11 weeks old received oral (gavage) administration of risedronate sodium salt (Cayman Chemical Company) in PBS at 10 mg/kg dosage every other day for 8 days. The drug was prepared fresh at every administration. Organs were harvested 8 days after the drug administration in the afternoon.

Sofosbuvir: Male C57BL/6NCrSlc mice at 7 weeks old received intraperitoneal injection of sofosbuvir (LKT) at 20 mg/kg dosage everyday for 10 days. Sofosbuvir was first diluted in DMSO (Nacalai Tesque) and then further diluted x100 in PBS prior to the administration (the final concentration is 1.0% DMSO/PBS). Organs were harvested 10 days after the drug administration in the afternoon.

Teriparatide: Male C57BL/6NCrSlc mice at 10 weeks old received subcutaneous injection of parathyroid hormone fragment 1-34 human (teriparatide) (Sigma-Aldrich) in saline at 40 μg/kg dosage or a vehicle solution (saline) everyday for 4 weeks. Organs were harvested 4 weeks after the drug administration in the afternoon.

Wild type (WT): Organs of male C57BL/6NJcl mice at 11 weeks old were harvested in the afternoon.

Differential gene expression data for the prediction

Differential gene expression data in each organ obtained from the mice to which each drug was administered were used as the features of each drug in the machine learning frameworks. The RNA-seq data processing (mapping and counting transcripts) were conducted as previously described(Kozawa et al., 2018). The mapping on mouse genome was conducted on mm10 with TopHat2. The differentially expressed genes in each organ of the drug-administered vs. vehicle-control (doxycycline and lenalidomide) or WT (all the other drugs) mice were identified by DESeq2 (1.22.1). Biological

duplicates (i.e., two mice for drug-administered and for controls) were analyzed for all organs and the drug-administered models. We selected the differentially expressed genes of which adjusted p-values < 1e-4 for each organ as the features of each drug. The combination of the organ and the $\log_2$fold values of all the selected genes combined from all organs (the 24-organ framework) or individual-organs (the individual-organ framework) are used as the input data for each drug. For SVM algorithm, we applied principle component analysis (PCA) to the organ-transcriptome data to reduce data dimensionality.

Prediction of AEs by SVM and RF

AEs were predicted by two machine learning algorithms, support vector machine (SVM) and random forest (RF). To predict an AE for a drug, we trained machine learning models by the transcriptome data for 14 drugs (i.e., the 15 drugs - the prediction-target drug) and the AE data reported for the 14 drugs in FAERS (Figure 2A). To obtain an AE prediction for the target drug, we input the organ-transcriptome data of the target drug into the trained machine learning models (Figure 2A). This training/prediction procedure was repeated for all drugs and all AEs. The transcriptome data are differential gene expression data described in "Differential gene expression data for the prediction" section. To obtain AE data, we downloaded raw AE reports from FAERS (https://www.fda.gov/drugs/surveillance/fda-adverse-event-reporting-system-faers) (2014 3rd quarter to 2017 4th quarter) and extracted the reports associated with the 15 drugs for each sex and age-groups. A total of 5,519 AEs is reported for one or more of the 15 drugs (Table S2). For each sex and age-group, the AEs reported in one or more the 15 drugs are used as all AEs for that particular sex and age-group (Table S2). The number of all AEs for each group are: 727 (M:10s), 1,147 (M:20s), 1,339 (M:30s), 1,747 (M:40s), 2,441 (M:50s), 2,618 (M:60s), 2,169 (M:70s), 1,189 (M:80s), 263 (M:90s), 784 (F:10s), 1,289 (F:20s), 1,687 (F:30s), 2,149 (F:40s), 2,828 (F:50s), 3,188 (F:60s), 2,807 (F:70s), 1,863 (F:80s), 573 (F:90s) (Supplementary Table 2). The binary labelling system was applied to each AE: 1 when the AE is reported/predicted to occur and 0 when the AE is not-reported/predicted not-to-occur. AEs that are 1 or 0 for all drugs in a sex/age-group were omitted from the training or prediction as their outcomes are naturally predictable. In addition, AEs that are 1 or 0 for all drugs except for the prediction-target drug in a sex/age-group were also omitted as their outcomes are, in principle, unpredictable. SVM was performed on R library 'e1071'

with default parameters. RF was performed on Python package scikitlearn. RF parameter 'n estimator' was set to 10, and the other parameters ('max_feature', 'max_depth' and 'min_samples_leaf') were decided by leave-one-out cross-validation for each AE, sex- and sex-group.

Majority decision approach for AE prediction

This approach is applied to predicting the outcomes of death event for alendronate/female/20s and lenalidomide/female/20s where the prediction results by SVM with all 24-organs failed. For the SVM method, there are 25 "votes" – SVM with the 24-organs (1 vote) and SVM with each of the 24 individual-organs (24 votes). For the RF method, there are also 25 votes – RF with the 24-organs (1 vote) and RF with each of the 24 individual-organs (24 votes). The number of the prediction votes for "to-occur" and "not-to-occur" is counted and summed. The prediction that won the majority votes become the prediction.

Prediction of AEs by LP

The prediction of AEs was also performed by solving a LP problem using one-class SVM. Let $\boldsymbol{g}_A$, $\boldsymbol{g}_B$ be the differential gene expression data described in "Differential gene expression data for the prediction" section without PCA of drug A and drug B respectively, and let $\boldsymbol{a}_i = (a_{\mathrm{i}1}, a_{\mathrm{i}2}, \ldots, a_{iM})$, $\boldsymbol{a}_{ii}$ be vectors in which each element represents the number of reports for indication 1, indication 2, …, indication M that are co-reported with AE i and AE ii, respectively.

In the training phase of one-class SVM to predict AEs for drug X, we made combined vector for the drug A – AE i combination $\boldsymbol{g}_A\boldsymbol{a}_i$ as a positive-link sample when more than 10 records exist for drug A inducing AE i in FAERS, and input all positive sample vectors without those of drug X to a one-class SVM. The kernel function of one-class SVM was designed as below,

$$k(\boldsymbol{g}_A\boldsymbol{a}_i, \boldsymbol{g}_B\boldsymbol{a}_{ii}) = \langle \boldsymbol{g}_A, \boldsymbol{g}_B \rangle \langle \boldsymbol{a}_i, \boldsymbol{a}_{ii} \rangle$$

where $\langle \cdot, \cdot \rangle$ represents operator that scales each vector to be that l2-norm equals 1 and take inner product between both scaled vectors. In the test phase of one-class SVM to predict the AEs for drug X, we input a combined vector for the target drug-AE combination $\boldsymbol{g}\boldsymbol{a}$ and ask whether the AE occurs with the drug

(i.e., the presence or the absence of the drug-AE link in the LP problem), and SVM returns 1 (occur) or -1 (not occur). We repeat this prediction all AEs reported at FAERS. One-class SVM was performed on Python package 'scikit-learn' with parameter nu=0.1.

REACTOME and KEGG analyses

REACTOME and KEGG analyses were performed on the gene sets of each organ with the feature importance greater than 0 in the RF prediction of the death event. REACTOME pathway analyses was performed REACTOME web API (https://reactome.org/dev/analysis) and KEGG pathway analyses was performed on R library 'clusterR'.

Prediction of AEs by SVR and RFR

For the quantitative prediction (Figure 7), the incidence of an AE for a drug and each sex/age-group was calculated as (the number of the reports of the AE for a sex/age-group)/(the number of the reports of all AEs for all sex/age-groups) for each drug. These values were used as target variables with two machine learning algorithms, support vector regressor (SVR) and random forest regressor (RFR) to train the models. The features of the two algorithms are differential gene expression data described in "Differential gene expression data for the prediction" section. While the actual incidence ranges from 0.0 to 1.0, the output values from the trained machine learning models range from $-\infty$ to ∞. To address this issue, we used the output of logit function (inverse function of sigmoid function) when inputting the incidence values. SVM was performed on R library 'e1071' with default parameters. RF was performed on Python package scikit-learn. RF parameter 'n_esitmator' was set to 10, and other parameters ('max_feature', 'max_depth' and 'min_samples_leaf') were decided by leave-one-out cross-validation for each AE, sex and age-group.

Prediction of therapeutic indications

The LP of therapeutic indications was performed by one-class SVM as follows: Let $\boldsymbol{g}_A, \boldsymbol{g}_B$ be the differential gene expression data described in "Differential gene expression data for the prediction". Section without PCA for drug A and drug B, respectively, and let $\boldsymbol{d}_1 = (d_{1\mathrm{i}}, d_{1\mathrm{ii}}, \ldots, d_{1N})$, $\boldsymbol{d}_2$ be

vectors that element represents the number of reports AE i, AE ii, …, AE N that be co-reported indication 1 and indication 2, respectively. In the training phase of one-class SVM to predict the indications for drug X, we made combined vector for the drug A - indication 1 combination $\boldsymbol{g}_A\boldsymbol{d}_1$ as a positive sample when more than 10 records exist for drug A taken by patients with the indication 1 in FAERS, and input all positive sample vectors without those for drug X to a one-class SVM. The kernel function of the one-class SVM was designed as below,

$$k(\boldsymbol{g}_A\boldsymbol{d}_1, \boldsymbol{g}_B\boldsymbol{d}_2) = \langle \boldsymbol{g}_A, \boldsymbol{g}_B \rangle \langle \boldsymbol{d}_1, \boldsymbol{d}_2 \rangle$$

where $\langle \cdot, \cdot \rangle$ represents operator that scales each vector to be that l2-norm equals 1 and takes inner product between both scaled vectors. In the test phase of one-class SVM to predict the indications for drug X, we input a combined vector for target drug-indication combination $\boldsymbol{gd}$ and ask whether the drug can treat the indication (i.e., the presence or the absence of the drug-indication link in the LP problem), and SVM returns 1 (can treat) or -1 (cannot treat). We repeat prediction for all indications. One-class SVM was performed on Python package 'scikit-learn' with parameter nu=0.1.

Validation with SIDER

The Side effects (SEs) and indication datasets reported in SIDER4.1 (http://sideeffects.embl.de) were compiled for each drug. The total number of SEs and indications reported (labeled) for the 14 drugs in this study (the 15 drugs minus evolocumab) are 3,061 and 2,088, respectively. SIDER does not classify the incidence of each SE according to sex/age-groups. Hence, all sex/age-group division-based predictions with FAERS are combined for each drug. When the outcomes of an AE in any one or more of the sex/age-groups show positive for a drug in the FAERS-prediction, we scored it as positive for that AE and for this drug, and scored the accuracy, precision and recall using the reported/labeled outcomes in SIDER4.1.

Supplementary References

Kozawa, S., Ueda, R., Urayama, K., Sagawa, F., Endo, S., Shiizaki, K., Kurosu, H., Maria de Almeida, G., Hasan, S.M., Nakazato, K*., et al.* (2018). The Body-wide Transcriptome Landscape of Disease Models. iScience *2*, 238-268.

2)

Bioinformatics Advances, 2023, vbad047
https://doi.org/10.1093/bioadv/vbad047
Advance Access Publication Date: 12 April 2023
Original Paper

OXFORD

Data and text mining

Latent disease similarities and therapeutic repurposing possibilities uncovered by multi-modal generative topic modeling of human diseases

Satoshi Kozawa[1,2,3], Hirona Yokoyama[1,2,4], Kyoji Urayama[1,2,3], Kengo Tejima[1,2,3], Hotaka Doi[1,2,4], Shunki Takagi[1,2] and Thomas N. Sato [1,2,3,4,*]

[1]Karydo TherapeutiX, Inc., Kyoto 619-0288, Japan, [2]The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International (ATR), Kyoto 619-0288, Japan, [3]ERATO Sato-Live Bio-Forecasting Project, Japan Science and Technology Agency (JST), Kyoto 619-0288, Japan and [4]V-iCliniX Laboratory, Nara Medical University, Nara 634-8521, Japan

*To whom correspondence should be addressed.
Associate Editor: Cecilia Arighi

Received on November 11, 2022; revised on March 14, 2023; editorial decision on March 23, 2023; accepted on March 31, 2023

Abstract

Motivation: Human diseases are characterized by multiple features such as their pathophysiological, molecular and genetic changes. The rapid expansion of such multi-modal disease-omics space provides an opportunity to re-classify diverse human diseases and to uncover their latent molecular similarities, which could be exploited to repurpose a therapeutic-target for one disease to another.

Results: Herein, we probe this underexplored space by soft-clustering 6955 human diseases by multi-modal generative topic modeling. Focusing on chronic kidney disease and myocardial infarction, two most life-threatening diseases, unveiled are their previously underrecognized molecular similarities to neoplasia and mental/neurological-disorders, and 69 repurposable therapeutic-targets for these diseases. Using an edit-distance-based pathway-classifier, we also find molecular pathways by which these targets could elicit their clinical effects. Importantly, for the 17 targets, the evidence for their therapeutic usefulness is retrospectively found in the pre-clinical and clinical space, illustrating the effectiveness of the method, and suggesting its broader applications across diverse human diseases.

Availability and implementation: The code reported in this article is available at: https://github.com/skozawa170301ktx/MultiModalDiseaseModeling

Contact: island1005@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Human diseases are characterized by alterations in a multitude of features: genetics, molecular, cellular, inter-organ pathways, histopathology, physiology, microbiota, etc. Recently, this disease-omics data space is rapidly expanding and becoming readily available, enabling the comprehensive characterizations of diverse human diseases (Hasin *et al.*, 2017; Perakakis *et al.*, 2018; Reel *et al.*, 2021).

For example, GWAS catalog (https://www.ebi.ac.uk/gwas/) and other similar databases provide a comprehensive list of genetic factors associated with thousands of human diseases and traits. KEGG (https://www.genome.jp/kegg/), Reactome (https://reactome.org) and other similar databases describe comprehensive molecular pathways. DisGeNET (https://www.disgenet.org) compiles altered gene expression, biomarkers, post-translational modifications, genetic factors, drug-targets, etc. and their association with human diseases. Disbiome (https://disbiome.ugent.be/home) tabulates human disease-associated microbiota. There are also large numbers of drug-related databases (SIDER: http://sideeffects.embl.de, DrugBank: https://go.drugbank.com, FAERS: https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard, etc.) that comprehensively list therapeutic-indications, side-effects/adverse-events, targets, etc. of the drugs. Moreover, we can virtually identify cells and organs that express genes/proteins of interest at The Human Protein Atlas (https://www.proteinatlas.org), Human Cell Atlas (https://www.humancellatlas.org) and other similar open-resources.

Hence, this rapidly expanding multi-modal disease-omics data space provides an opportunity to re-classify diverse human diseases according to their multi-modal similarity metrics. Furthermore, this approach could find previously underrecognized disease–disease similarities.

The pioneering study built on the graph theory provided an overview of disease–disease similarities according to their single modality

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

features, genetic variabilities (Goh *et al.*, 2007). Since then, more sophisticated network-based and other approaches have evolved to characterize multi-modal nature of human diseases (Barabási *et al.*, 2011; García Del Valle *et al.*, 2021; Li *et al.*, 2021; Menche *et al.*, 2015; Perakakis *et al.*, 2018; Reel *et al.*, 2021). Despite the development of such tools and methods, the ever-expanding multi-modal disease-omics space remains under-explored. Hence, further in-depth probing of this data space is expected to uncover latent molecular mechanisms underlying non-classical under-recognized disease–disease similarities.

As an approach that could integrate multiple types of features to classify human diseases and measure their similarity metrics, topic modeling was brought to our attention. This algorithm has been applied to categorize social media information (Zheng *et al.*, 2014), and also to image annotation and classification and computer vision (Roller and im Walde, 2013). This approach has also recently been applied to the classification of clinical notes (Wen *et al.*, 2021) and RNA dual-omics (RNA, microRNA) data (Valle *et al.*, 2022).

Based on these previous reports, we considered a use of the multi-modal topic modeling to re-classify diverse types of human diseases according to their multi-modal disease-omics features. Using this multi-modal omics-based soft-clustering of human diseases, we aim to identify molecular targets that could be repurposed from one disease to another for its treatment and/or detection (i.e. latent disease-omics features/therapeutic-targets). Furthermore, the identification of unexpected repurposable therapeutic targets may unveil previously underrecognized disease–disease similarities (i.e. latent disease similarities). This approach differs from that of 'drug-repurposing' where a drug, not a disease omics-feature, for a disease is repurposed to treat another disease according to the similarities of drug characteristics (e.g. the similarities of drug structures, drug-targets, etc.) (Al-Saleem *et al.*, 2021; Bisgin *et al.*, 2012; Hooshmand *et al.*, 2021; Malas *et al.*, 2019; Panchapakesan and Pollock, 2018; Park, 2019; Pushpakom *et al.*, 2019; Schuler *et al.*, 2022).

Hence, in this article, we report a multi-modal generative topic modeling-based method that is suitable for re-classifying human diseases to identify repurposable molecular therapeutic-targets and to unveil latent disease–disease similarities. We also illustrate its applications to two globally most life-threatening human diseases, chronic kidney diseases (CKD) (Chen *et al.*, 2019) and myocardial infarction (MI) (Anderson and Morrow, 2017).

2 Methods

2.1 General overview of the multi-modal generative topic modeling approach in this study

The general overview of the approach in this study is shown in Figure 1A. The details are described in the following sections (Sections 2.2 and 2.3). Let us explain the overall approach by taking an example of predicting latent omics features and disease-similarities of CKD. The training dataset consists of 6955 human diseases, each of which is described by three disease omics modalities, AlteredExpression (Ae), Biomarker (Bm) and GeneticVariation (Gv) (see Section 2.2 for the details). Each modality contains distinct types of omics features that characterize the corresponding disease. First, we remove Ae modality for CKD, leaving the CKD training dataset with only two modalities, Bm and Gv. All the other diseases (6954 diseases) remain labeled with three modalities. By the multi-modal generative topic modeling (see Section 2.3 for the details), we predict the omics features of the missing Ae modality for CKD. We repeat this step for all three modalities (Ae, Bm, Gv) for CKD. Next, from the predicted omics features for all three modalities for CKD, we remove those included in the modalities of the diseases that are apparently related to CKD (e.g. renal diseases, coronary diseases, diabetes, etc.) (see Section 2.3.5 for the specific list of the diseases). Consequently, the remaining features represent the 'latent omics features' of CKD. We then identify the disease-labels of these latent omics features in the training dataset, and they constitute the list of the diseases of which similarity to CKD is 'latent' (i.e. underrecognized). Hence, this relationship is referred to as 'latent disease similarity' for CKD.

2.2 Multi-modal disease-omics datasets

The multi-modal disease-omics features for human diseases used in this study are AlteredExpression (Ae), Biomarker (Bm), GeneticVariation (Gv) data from DisGeNET v7.0. (https://www.disgenet.org/downloads) (Piñero *et al.*, 2017) and Microbiota (Mb) from Disbiome (version on 11th of November 2020, https://disbiome.ugent.be/home) (Janssens *et al.*, 2018). Ae is the list of genes and proteins of which changes in expressions are attributed to a disease(s). Bm is the list of biomarkers which are attributed to a disease(s). Gv is the list of genes of which mutations are reported for a disease(s). Mb is the list of microbial organisms of which changes in abundance are reported for a disease(s). We chose these molecular omics features to characterize and model human diseases as they are amenable to therapeutic-targeting and/or disease-detection. A total of 6955 and 158 human diseases are found labeled by all the combinations of the Ae/Bm/Gv and the Ae/Bm/Gv/Mb modalities, respectively.

To match disease names acquired from multiple data sources, we added UMLS IDs to the disease names. The UMLS ID annotation was performed by 'UMLS_AUI.extract_terminology ("ICD10")' function from Python library 'PyMedTermino (version 0.3.3)' (Lamy *et al.*, 2015). Prior to performing this function, ''s' was replaced by a blank space in disease names. Following the assignment of the UMLS ID annotations, the UMLS IDs were combined by string '|' if these UMLS IDs have the same disease names. For the disease names where this UMLS ID annotation method failed, the actual disease names in the datasets were used and only those with the exact matching names were combined. For the Ae/Bm/Gv and the Ae/Bm/Gv/Mb combinations, a total of 1809 and 116 diseases are convertible to the UMLS IDs, respectively.

2.3 Multi-modal generative topic modeling of human diseases

2.3.1 The multi-modal generative topic modeling and prediction of latent omics features

The multi-modal generative topic modeling that we employed is based on Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003). The side-by-side comparison between the classical LDA and our multi-modal generative topic modeling is shown in Figure 1B.

The classical LDA is commonly used to soft-clustering documents by using words in the documents (referred to as 'bag-of-words'). Our multi-modal generative topic modeling is developed to soft-cluster human diseases based on their multi-modal disease-omics features (indicated as 'components' in the algorithm diagram) (e.g. Ae, Bm, Gv, Mb, etc.). In our multi-modal generative topic modeling, the human diseases and omics-features correspond to 'documents' and 'words' in the classical LDA model, respectively. Therefore, the human diseases are soft-clustered according to the co-occurrence of omics-features between the diseases. Furthermore, the model is generated by 'cross-referencing' the probabilistic distributions of the features of each other's modality datasets, not by a simple concatenation of the multi-modal datasets. This is necessary as the omics features across the different modalities represent distinct types of the data—for example, Ae and Mb consist of the transcripts (i.e. mRNA) and microbial organisms, respectively. Therefore, the concatenation of such multi-modal data leads to inaccurate representation of human diseases.

In our multi-modal generative topic modeling, all modalities (i.e. Ae, Bm, Gv, Mb in this study) for a given disease are designed to exhibit the same probabilistic distribution pattern across the topics. Thus, the human diseases are soft-clustered according to the overall probabilistic distributions of the multiple modalities.

The probabilistic generative topic model using multi-modal datasets is built as follows (see also the right panel in Figure 1B for the graphical description): Let $w_{di}^{(\cdot)}$ be the i-th disease-omics component [we use 'component(s)' in the algorithm/program, but it is the same as the disease-omics feature(s)] of disease d acquired from the

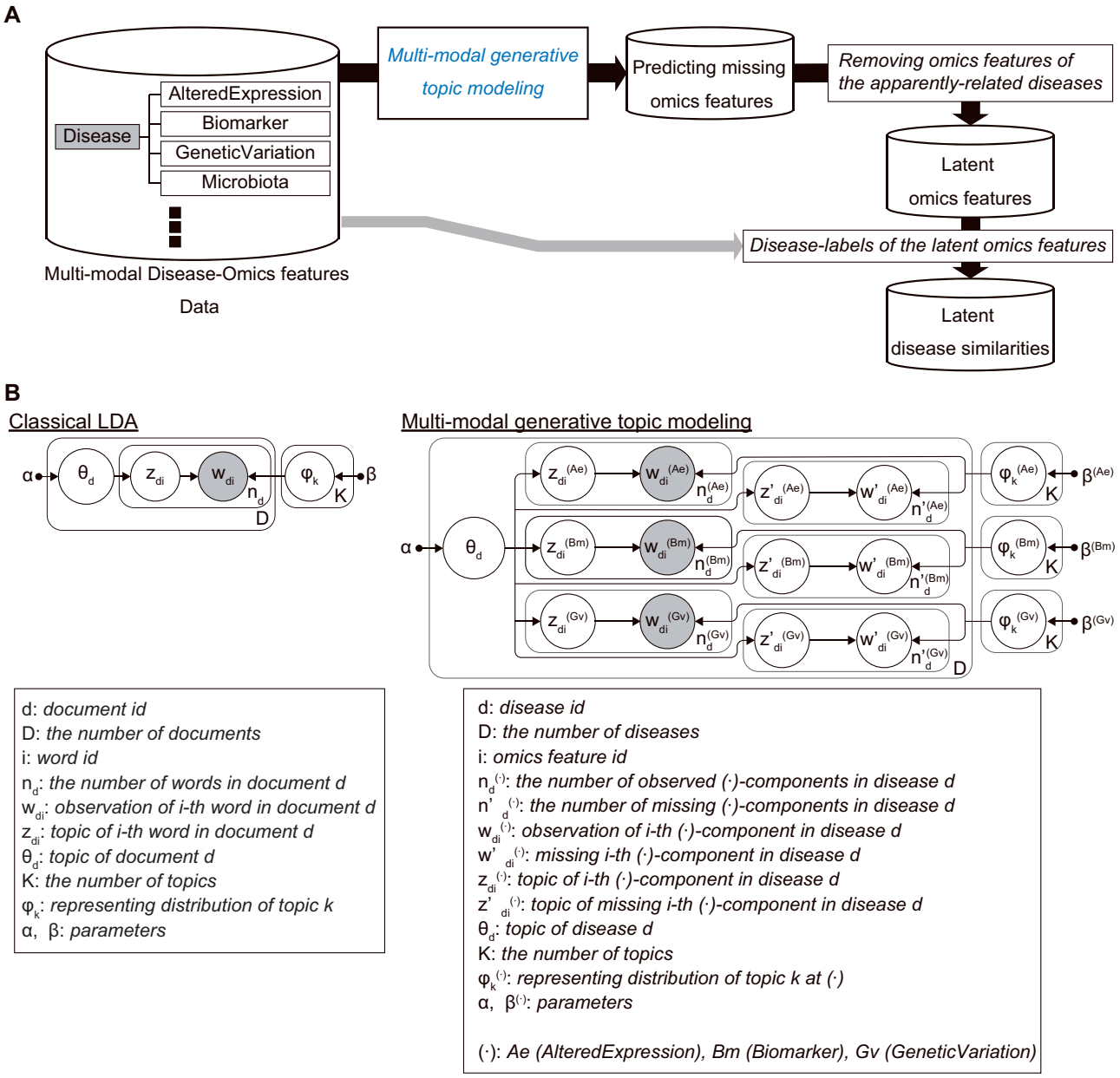


Fig. 1. Schematic diagram of the multi-modal generative topic-modeling. (**A**) The general overview of the approach. Using the comprehensive multi-modal disease-omics datasets, the human diseases are soft-clustered by the multi-modal generative topic modeling according to their multi-modal similarity metrics. Next, the features of a modality of a disease-of-interest are intentionally removed (indicated as 'missing omics features' in the figure). These missing features are then predicted by the multi-modal generative topic model. From these predicted features, those of the apparently related diseases (e.g. myocardial infarction versus heart failure, chronic kidney disease versus renal failure, etc.) in the original datasets are then removed, leaving the unexpected features ('latent omics features') of the target diseases. As the result, the origin of the disease(s) of the latent omics features in the databases represents underrecognized latent disease–disease similarity. The datasets and their uses/analyses are illustrated as 'cylinders' and 'boxes', respectively. (**B**) The side-by-side comparison between the classical LDA algorithm and the multi-modal generative topic modeling introduced in this article. See also Sections 2.1, 2.2 and 2.3 for the details

modality (i.e. Ae, Bm, Gv or Mb) dataset $(\cdot)$. Let $z_{di}^{(\cdot)}$ be the topic number of $w_{di}^{(\cdot)}$, $\boldsymbol{\theta}_d$ be the topic probability of disease d, and $\boldsymbol{\phi}_k^{(\cdot)}$ be the occurrence probability of disease-omics component v of topic k of $(\cdot)$ dataset. The topics number K is determined as described in Section 2.3.3. The joint distribution of these variables is defined as follows

$$\begin{aligned} &p(\boldsymbol{w}, z, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \left[\prod_{(\cdot)}\prod_d\prod_i p\left(w_{di}^{(\cdot)}|z_{di}^{(\cdot)}, \boldsymbol{\phi}^{(\cdot)}\right)\right]\left[\prod_{(\cdot)}\prod_d\prod_i p\left(z_{di}^{(\cdot)}|\boldsymbol{\theta}_d\right)\right] \\ &\times \left[\prod_d p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})\right]\left[\prod_{(\cdot)}\prod_k p\left(\boldsymbol{\phi}_k^{(\cdot)}|\boldsymbol{\beta}^{(\cdot)}\right)\right]. \end{aligned} \quad (1)$$

The exact formula of the elements of the joint distribution is described as:

$$\begin{aligned} &p\left(w_{di}^{(\cdot)}|z_{di}^{(\cdot)}, \boldsymbol{\phi}^{(\cdot)}\right) \\ &= Multinomial\left(w_{di}^{(\cdot)}|\boldsymbol{\phi}^{(\cdot)}\right) = \prod_{k=1}^{K}\prod_{v=1}^{V^{(\cdot)}}\phi_{kv}^{(\cdot)}\delta\left(z_{di}^{(\cdot)}=k\right)\delta\left(w_{di}^{(\cdot)}=v\right), \end{aligned} \quad (2)$$

$$p\left(z_{di}^{(\cdot)}|\boldsymbol{\theta}_d\right) = Multinomial\left(z_{di}^{(\cdot)}|\boldsymbol{\theta}_d\right) = \prod_{k=1}^{K}\theta_{dk}^{\delta\left(z_{di}^{(\cdot)}=k\right)}, \quad (3)$$

$$p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) = Dirichlet(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1}, \quad (4)$$

$$p\left(\boldsymbol{\phi}_k^{(\cdot)}|\boldsymbol{\beta}^{(\cdot)}\right) = Dirichlet\left(\boldsymbol{\phi}_k^{(\cdot)}|\boldsymbol{\beta}^{(\cdot)}\right) = \frac{\Gamma\left(\sum_v \beta_v^{(\cdot)}\right)}{\prod_v^{V^{(\cdot)}} \Gamma\left(\beta_v^{(\cdot)}\right)} \prod_{v=1}^{V^{(\cdot)}} {\phi^{(\cdot)}}_{kv}^{\beta_v^{(\cdot)} - 1}, \quad (5)$$

where $\boldsymbol{\alpha} \in R^{+K}$ and $\boldsymbol{\beta}^{(\cdot)} \in R^{+V^{(\cdot)}}$ are hyper parameters which are set to vectors having the elements of 0.1, $V^{(\cdot)}$ is the total number of unique components across all diseases for the corresponding $(\cdot)$ dataset, $\Gamma(\cdot)$ is the Gamma function and $\delta(\cdot)$ is the Kronecker delta function.

2.3.2 Estimate of the topic distribution and disease-omics components

Based on the model, we estimate the posterior distributions of the variables $z_{di}^{(\cdot)}$, $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k^{(\cdot)}$, and also estimate a part of $w_{di}^{(\cdot)}$ (it will be referred to as $w_{di}^{\prime(\cdot)}$ that is missing in a $(\cdot)$ dataset) by using Gibbs sampling method. Based on the generative model built as described in 2.3.1, the conditional distributions of the variable are calculated as follows:

$$p\left(z_{di}^{(\cdot)} = k | w_{di}^{(\cdot)} = v, \boldsymbol{\theta}_d, \boldsymbol{\phi}^{(\cdot)}\right) = \frac{\theta_{dk} \phi_{kv}^{(\cdot)}}{\sum_{k'} \theta_{dk'} \phi_{k'v}^{(\cdot)}}, \quad (6)$$

$$p\left(\boldsymbol{\theta}_d|\boldsymbol{z}^{(\cdot)}, \boldsymbol{\alpha}\right) = Dirichlet\left(\boldsymbol{\theta}_d|\boldsymbol{\alpha}'\right) \propto \prod_{k=1}^{K} \theta_{dk}^{\alpha_k + \sum_i \delta\left(z_{di}^{(\cdot)} = k\right) - 1}, \quad (7)$$

$$\begin{aligned} &p\left(\boldsymbol{\phi}_k^{(\cdot)}|\boldsymbol{z}^{(\cdot)},\ \boldsymbol{w}^{(\cdot)},\ \boldsymbol{\beta}^{(\cdot)}\right) \\ &= Dirichlet\left(\boldsymbol{\phi}_k^{(\cdot)}|\boldsymbol{\beta}^{(\cdot)'}\right) \propto \prod_v {\phi^{(\cdot)}}_{kv}^{\beta_k^{(\cdot)} + \sum_d \sum_i \delta\left(z_{di}^{(\cdot)} = k\right)\ \delta\left(w_{di}^{(\cdot)} = v\right) - 1}, \end{aligned} \quad (8)$$

$$p\left(w_{di}^{(\cdot)} = v | z_{di}^{(\cdot)} = k,\ \phi^{(\cdot)}\right) = \phi_{kv}^{(\cdot)}. \quad (9)$$

From these conditional distributions, we sample $z_{di}^{(\cdot)}$,$\boldsymbol{\theta}_d$, $\boldsymbol{\phi}_k^{(\cdot)}$ and $w_{di}^{\prime(\cdot)}$, until their values converge. We found the topic distribution converges after 1000 to 2000 iterations with any combinations of the diseases and modalities, indicating that at least 2000 iterations are required. Therefore, to be sufficient enough, the number of iterations is set to 5000 in this article. After the completion of sampling, we estimate the value of each variable by averaging the sampled values from the conditional distributions. The initial values of the variables are set as follows: The initial values of θ_d and $\boldsymbol{\phi}_k^{(\cdot)}$ were set to the uniform distribution. The initial value of $w_{di}^{\prime(\cdot)}$ is set to the disease-omics component that is sampled from $\boldsymbol{\phi}_k^{(\cdot)}$ at random. The total number of $w_{di}^{\prime(\cdot)}$ of disease d is decided by sampling from the binominal distribution. The parameters for the binominal distribution are estimated by the maximum likelihood estimation method using the observed data. The estimated value of $\boldsymbol{\theta}_d$ represents the probability of topics at disease d. The likeliness of the missing values of disease d in $(\cdot)$ dataset can be inferred by sampling frequency of $w_{di}^{\prime(\cdot)}$.

2.3.3 Computation of the topics number K

The topics number K is selected by the Louvain method (Blondel *et al.*, 2008). To apply this method, an undirected graph per each of the Ae, Bm, Gv, Mb datasets is constructed by using 'NetworkX' (https://networkx.org/). In each graph, if two diseases (i.e. nodes) share disease-omics components, we allow an edge between them. The edges are weighted according to the number of shared disease-omics components. The Louvain method is then applied 20 times to each graph (i.e. Ae, Bm, Gv, Mb modality), and the most frequently obtained number of communities was selected for the community number for each modality. The maximum community number for the combination of the modalities (Ae/Bm/Gv or Ae/Bm/Gv/Mb) is then selected as the K for the corresponding modality combination. The computed topics numbers are 6 and 3 for the Ae/Bm/Gv and Ae/Bm/Gv/Mb combinations, respectively (Supplementary Fig. S1). The Louvain method is performed by Python package 'python-louvain' (https://python-louvain.readthedocs.io/en/latest/).

2.3.4 Performance evaluation of the multi-modal generative topic modeling method

The performance of the multi-modal generative topic modeling is evaluated by 'leave-one-modality-out' per each disease method and then by calculating their AUC scores. The input is the Ae/Bm/Gv or Ae/Bm/Gv/Mb dataset where single modality components (e.g. Ae, Bm, Gv, Mb) are purposely left-out for each disease. We then performed the multi-modal generative topic modeling as described in the previous sections on each of these input datasets. The likeliness of the missing disease-omics features is determined by the sampling frequency of $w_{di}^{\prime(\cdot)}$, where $(\cdot)$ is either Ae/Bm/Gv or Ae/Bm/Gv/Mb combination. Hence, the likeliness is the prediction probability of the missing disease-omics component $w_{di}^{\prime(\cdot)}$. The label is defined whether each of the disease-omics components is present or not (i.e. binary labeling) in the original dataset. The AUC scores are calculated from these prediction probability values and the labels for the disease in each of the $(\cdot)$ datasets. The calculation of the AUC scores was performed by function 'roc_auc_score()' in package 'scikit-learn' (https://scikit-learn.org/stable/) (Pedregosa *et al.*, 2011). The Youden's index (Youden, 1950) was used as the cut-off threshold for the sampling frequency of $w_{di}^{\prime(\cdot)}$. Youden's index is a ROC curve-based thresholding method. The thresholds for each modality for each disease were computed as follows: (i) true positive rate (TPR) and false positive rate (FPR) were computed using the function 'roc_curve()' in the package 'scikit-learn' (https://scikit-learn.org/stable/), (ii) the Youden's index was then calculated by the formula, TPR-FPR for each threshold and (iii) for each modality of each disease, the threshold which corresponds to the maximum Youden's index was selected as the cut-off threshold for the corresponding modality for each disease.

2.3.5 Identification of latent disease-omics features and disease-similarity

The latent disease-omics features are identified by removing the features derived from the diseases (in the training dataset) that are apparently related to the target disease(s). For CKD and/or MI as the target diseases, the removed are the features derived from the diseases of which names contain the following terms: 'heart', '**cardi**'(e.g. myocardial), 'athero**' (e.g. atherosclerosis), 'arterio**' (e.g. arteriosclerosis), 'coronary', 'kidney', 'renal', 'nephro**' (e.g. nephropathy), 'glomer**' (e.g. glomerular), 'diabe**' (e.g. diabetes), 'vascul**' (e.g. vascular), 'capil**' (e.g. capillary), 'hypertens**' (e.g. hypertension) (** could be any characters).

Next, the latent disease-similarity is determined as follows: The disease-labels of these latent disease omics features in the training dataset are identified. Hence, they represent the diseases of which similarity to the target diseases (i.e. CKD and/or MI in this study) is latent. Therefore, this disease–disease relationship is referred to as 'latent disease-similarity'.

2.4 Characterization of the latent disease-omics features

2.4.1 Organ/cell expression enrichment analysis

To find the specific organ/cell-expression patterns for the identified latent disease-omics features, we performed human organ/cell enrichment analyses using THE HUMAN PROTEIN ATLAS v 21.1. (https://www.proteinatlas.org) (Thul *et al.*, 2017; Uhlén *et al.*, 2015). The table was downloaded from '25. Data from the Human Protein Atlas in tab-separated format' in the 'DOWNLOADABLE DATA' page (https://www.proteinatlas.org/about/download). The enrichment analysis was performed using chi-square test of independence to evaluate the statistical significance of the enriched expression in the specific organ(s)/cell(s) detected for the genes of interest. We performed the test by making the 2 × 2 contingency table consisting of the appearance frequency of the genes of interest and that of the genes of interest in each organ/cell. This table was used as the input to perform the chi-square test of independence

using the Python function 'scipy.stats.chi2_contingency()' (https://docs.scipy.org/doc/scipy/index.html).

2.4.2 KEGG enrichment analysis

To find the specific KEGG pathways for the predicted latent disease-omics features, we performed KEGG enrichment analysis using KEGG database (https://www.genome.jp/kegg/) (Kanehisa and Goto, 2000). KEGG enrichment analysis was performed by R function 'enrichKEGG()' in the package 'clusterProfiler' (Yu *et al.*, 2012). For inputting enrichKEGG(), the symbol names of the genes were converted to Entrez IDs using R function 'bitr()' in the package 'clusterProfiler' (Yu *et al.*, 2012).

2.4.3 GO enrichment analysis

To find the specific gene ontology terms for the predicted latent disease-omics features, we performed GO enrichment analysis using GO database (http://geneontology.org/) (Ashburner *et al.*, 2000; Carbon *et al.*, 2021). GO enrichment analysis was performed by R function 'enrichGO()' in package 'clusterProfiler' (Yu *et al.*, 2012). For inputting enrichGO(), the symbol names of the genes were converted to Entrez IDs using R function 'bitr()' in the package 'clusterProfiler' (Yu *et al.*, 2012).

2.5 Edit-distance-based classifier

Edit-distance-based method was designed for two purposes:

1. To measure the relatedness of the sequential orders of the molecular components in the molecular pathways triggered by the predicted latent omics-features/components to the known molecular pathways for their target disease (CKD or MI in this article).
2. To infer putative side-effects that could result from targeting the predicted latent omics-features/components.

The overall design of the method, consisting of multiple modules, is schematically shown in Figure 7. We employed two methods, Method A and Method B. Each module is as described below:

2.5.1 Path extraction

Each path is a sequence of KEGG components extracted from the KEGG pathway. In the KEGG database, all components are systematically labeled as identifiers (e.g. hsa:3065). Therefore, we use the identifiers, allowing the reliable computation of the distances. Each path starts at the node of a KEGG component described as a drug target at the DrugBank (https://go.drugbank.com/) and ends at the node of a KEGG component with an outdegree of 0. The therapeutic indications (TIs) and side-effects (SEs) for each drug for the drug target(s) are from SIDER 4.1 (http://sideeffects.embl.de/). This design results in the paths where the starting nodes (i.e. molecular targets of the drugs) are labeled with TIs and SEs. The holdout validation was performed with the training versus test data as 9:1.

2.5.2 Similarity computation by edit-distance

The similarity between two paths N and M is calculated using the edit-distance (Levenshtein distance) method as follows:

$$Similarity(N,M) = 1 - \frac{EditDistance(N,M)}{max\big(length(N),\ length(M)\big)}, \tag{10}$$

where $length(X)$ is the number of components of path X. The edit-distance between paths N and M is calculated by considering a path as a word and a component as a character, using the dynamic programming algorithm (Navarro, 2001) as follows:

$$\begin{aligned} C_{i,0} &= i \quad 0 \le i \le length(N), \\ C_{0,j} &= j0 \le j \le length(M), \end{aligned} \tag{11}$$

$$x = \begin{cases} 0 & if\ N_i = M_j \\ 1 & if\ N_i \neq M_j \end{cases}, \tag{12}$$

$$C_{i,j} = min\big(C_{i-1,j} + 1,\ \ C_{i,j-1} + 1,\ \ C_{i-1,j-1} + x\big), \tag{13}$$

$$EditDistance(N,M) = C_{length(N),length(M)}, \tag{14}$$

where N_i is the ith component of path N and M_j is the jth component of path M.

2.5.3 Feature selection of the paths for TI/SE by PCA (Method A)

For the target TI or SE, we calculated the similarity matrix A between the paths with a starting node for the selected TI or SE and the paths for the training or prediction (Fig. 7A). Each component of matrix A is the pairwise similarity value calculated by the edit-distance as described in Section 2.5.2. The dimension of matrix A is reduced by the principal component analysis (PCA) (Wold *et al.*, 1987). The principal components are selected until the cumulative contribution reaches 99%. The PCA result was used as the features of the paths for the training or prediction for the target TIs or SEs.

2.5.4 Feature selection of the paths for TI/SE by clustering (Method B)

For each drug target x, we calculated the similarity matrix B between the paths with x as a starting node and all paths (including the paths with x themselves) (Fig. 7B). Each component of matrix B is the pairwise similarity value calculated by the edit-distance as described in Section 2.5.2. The correlation matrix C of the paths with a starting node x was calculated by considering the row of matrix B as the vector of the path. A clustering for these paths was performed by the Python function 'scipy.cluster.hierarchy.average()' (https://scipy.org/) using matrix C. The clusters are composed of paths whose distance from each other is less than or equal to the threshold of 0.2. Each of the clusters (e.g. Cluster 0, Cluster 1, etc.) is then linked to each path. Next, for each path with a starting node corresponding to the selected target TI or SE, its starting node is paired with its cluster number (Fig. 7C). The sampling is performed as follows. If the multiplicity of the pair(s) is greater than ceil(500/n), where n is the number of the pairs excluding duplicates, the corresponding paths were sampled to ceil(500/n). If the multiplicity of the pair(s) is less than or equal to ceil(500/n), the corresponding paths were used as they are. The ceil(500/n) was selected as the maximum value within our hardware capacity. The similarity matrix between the paths obtained by this method and paths for the training or prediction was calculated and used as the feature of the selected TI or SE. Each component of the similarity matrix is the pairwise similarity value calculated by the edit-distance as described in Section 2.5.2.

2.5.5 Training of binary classifier

For the target TI or SE, the paths for the training were downsampled to the ratio of 1(positive):1(negative) (the paths for the corresponding TI or SE versus the paths that do not correspond to the target TI or SE). Random under-sampling was performed on Python package 'imblearn' (https://imbalanced-learn.org/stable/index.html) (Lemaitre *et al.*, 2017). These downsampled paths were converted to features by using 'Method A' or 'Method B' as described in 2.5.3 or 2.5.4, respectively. Ten models determining whether or not given paths are related to each of the TIs or SEs were developed for bagging using the downsampled path-TI (or SE) features and trained by LightGBM (Ke *et al.*, 2017). This training process was conducted for all TIs and SEs. The hyperparameters of LightGBM were tuned by Python function 'optuna.integration.lightgbm()' (https://optuna.org/) using 20% of the downsampled data.

2.5.6 Prediction of TIs and SEs

The paths with a drug-target component as the starting node were extracted from the KEGG pathways as described in the Section 2.5.1. The feature of paths was calculated by Methods A or B as described in Sections 2.5.3 or 2.5.4, respectively. These features were used as input data (Fig. 7D). For each TI or SE, if the predicted values for a given path show ≥0.5 with 6 or more of the 10 models,

the path was determined as 'true (1)', and if the predicted values show otherwise, it was determined as 'false (0)'.

2.6 The retrospective validation in the pre-clinical and clinical data space

To further evaluate the likeliness of the repurposability of the predicted target molecules for CKD and/or MI biomarkers and/or therapeutic-targets, we conducted comprehensive pre-clinical and clinical data-mining to retrospectively find any such implications in Google, PubMed, Cinicaltrials.gov (https://www.clinicaltrials.gov) data space.

3 Results

3.1 Multi-modal generative topic modeling of human diseases and its prediction performance

In this study, we developed a multi-modal generative topic modeling method that is applicable to multi-modal disease-omics data of human diseases (Fig. 1 and Sections 2.3.1, 2.3.2 and 2.3.3). We evaluated its prediction performance using Ae, Bm, Gv for 6955 human diseases derived from DisGeNET (https://www.disgenet.org). The prediction performance-validation is conducted by randomly dividing them into the 6665 diseases as training data and the remaining 290 diseases as test data encompassing all 6955 diseases. For each set of the test diseases, we removed their single modality features as prediction targets. We tested one modality at a time for each set of the test diseases. The prediction performance for the removed (i.e. missing in the test data) modality features in the test data is measured by Area Under receiver operating characteristic Curve (AUC) (Fig. 2A, Supplementary Table S1, see also Section 2.3.5). The results show that AUCs for approximately 88% (6125/6955), 92% (6381/6955), 84% (5812/6955) of the diseases are >0.8 for Ae, Bm, Gv, respectively, supporting the effectiveness of the method.

To test the modality-scalability of the method, we added another modality dataset, disease-microbiota (Mb) dataset from Disbiome (https://disbiome.ugent.be/home) to the above Ae/Bm/Gv combination. This expanded combination of datasets enabled us to soft-

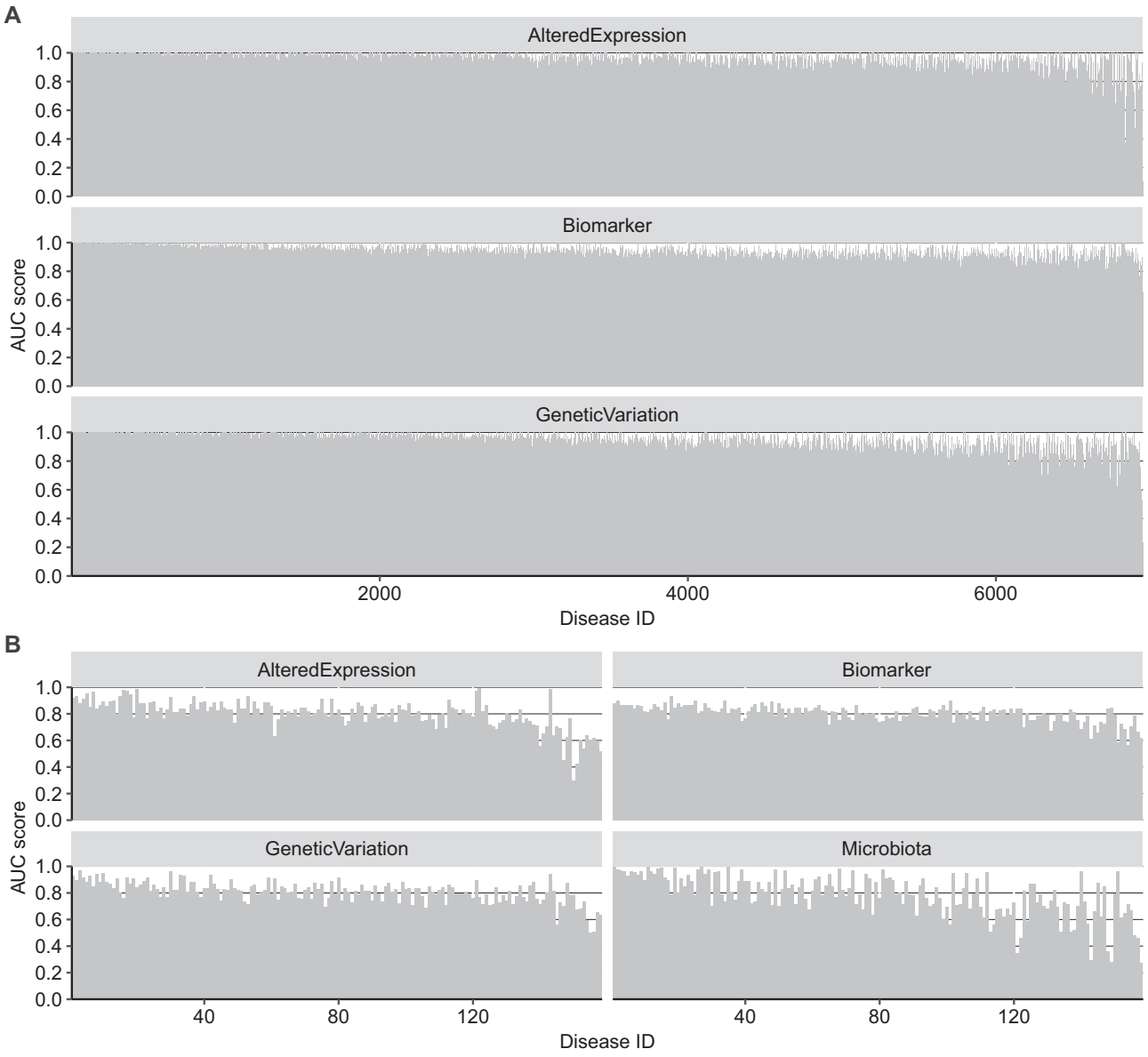


Fig. 2. The AUC scores of the cross validations. (**A**) The AUC scores for the AlteredExpression, Biomarker, GeneticVariation modalities across 6955 human diseases. The raw data are available as Supplementary Table S1. (**B**) The AUC scores for the AlteredExpression, Biomarker, GeneticVariation, Microbiota modalities across 158 human diseases. The raw data are available as Supplementary Table S2

cluster 158 diseases all of which are annotated with the Ae, Bm, Gv and Mb features. The performance is evaluated by leaving-one-modality-of-a-disease-out validation and by computing the AUC for each modality (i.e. Ae, Bm, Gv, Mb) of each disease (Fig. 2B, Supplementary Table S2). The result shows that AUCs for approximately 59% (93/158), 58% (91/158), 61% (96/158), 50% (79/158) of the diseases are >0.8 for Ae, Bm, Gv, Mb, respectively. Despite the significant reduction of the training data (6955 diseases for the three-way modalities versus 158 diseases for the four-way modalities), more than 50% of the diseases exhibit the AUCs of > 0.8, supporting the reasonable modality-scalability of the method.

3.2 Inference of repurposable molecular targets for CKD and MI and their similarities to other diseases

Using this performance-validated method, we aimed to identify repurposable molecular targets for treating and/or detecting two globally most life-threatening diseases, CKD and MI. For this purpose, we further tested the performance of the method specifically for CKD and MI using the above-described Ae/Bm/Gv datasets encompassing 6955 diseases including CKD and MI. To infer repurposable Ae, Bm, Gv features from the other diseases as therapeutic-targets for CKD and/or MI, we purposely removed Ae, Bm or Gv features from CKD or MI and predicted the missing features. The result shows AUCs >0.8 for all Ae, Bm and Gv features of both diseases (Fig. 3, Supplementary Tables S3–S9).

From this list, the features that are present in the training datasets (i.e. those correctly predicted) were first removed, leaving those that are absent in the CKD or MI data. Next, we used Youden's index to select those that are considered as 'statistically positive' by this criterion (Supplementary Table S10, see also Section 2.3.4). Through these selections, left are features that are absent in the training datasets for the corresponding diseases and regarded as statistically significant (Supplementary Tables S11–S16). Further selection was conducted by removing those that are labeled with other kidney/renal and cardiac/heart/cardiovascular related diseases (e.g. renal failure, heart failure, etc.), as we could easily postulate, without any computational-methods, their repurposability to CKD and/or MI therapeutics. Through this additional selection step, we obtained a list of 30 and 57 molecular therapeutic candidates for CKD and MI, respectively, out of which 18 are shared by the two (Supplementary Tables S17 and S18). These candidates are particularly enriched in Ae, Bm and/or Gv of neoplasia (e.g. neoplasms, malignant neoplasms, neoplasm metastasis, malignant neoplasm of breast, primary malignant neoplasm, liver carcinoma, etc.) and mental/neurological disorders (e.g. schizophrenia, seizures, epilepsy, intellectual disability, etc.) (Fig. 4, Supplementary Table S19), unveiling their molecular similarities to the renal and/or cardiovascular diseases such as CKD and MI. Considering the relatively high AUCs of this inference method (Fig. 3), this possibility is further supported.

3.3 Therapeutic mechanisms of the inferred targets

To gain mechanistic insights into the putative therapeutic actions of the predicted target molecules, we identified their expression patterns in the human body using a comprehensive human protein/gene expression database (Fig. 5, see also Section 2.4.1). While the expression of each target is found across multiple organs and cell-types (Fig. 5A and B, Supplementary Tables S20 and S21), the CKD and MI targets are enriched in the liver and the brain, respectively (Fig. 5C, Supplementary Table S22). At the single cell-level, we observed some enrichment in the hepatocytes and bipolar cells for CKD and MI, respectively (Fig. 5D, Supplementary Table S23). The result suggests that these organs/cells may serve as therapeutic targets for the respective disease.

Further mechanistic insights were gained by the enrichment analyses of biological pathways and functions using KEGG and GO databases (Fig. 6, Supplementary Tables S24–S28, see also Sections 2.4.2 and 2.4.3). The analyses found the enrichment of the MI targets in neural-pathways and -functions. These analyses, together with the expression pattern results, suggest that the nervous system functions and pathways are potential therapeutic targets for MI. In contrast, no enrichments are found for the CKD targets, instead they are sparsely encompassed across multiple biological pathways and functions (Supplementary Tables S25–S28).

Therapeutic-targeting elicits both favorable and unfavorable effects. The former is the therapeutic indications (TIs) and the latter is side-effects (SEs). Hence, we designed an analysis to infer these effects triggered by targeting the identified latent disease-omics features. For this purpose, we employed an edit-distance-based machine-learning classifier method (Fig. 7, see also Section 2.5). This classifier uses the edit-distance, specifically Levenshtein distance, to measure the similarity metrics between the pathways. This method measures the similarities of all possible pathways of the target candidates to each pathway downstream of the clinically approved drugs for each disease are computed. This is repeated for all pair-wise combinations for each disease and the computed edit-distances are used as input data for the corresponding disease-pathways classifier.

We, first, applied this method to the CKD and MI classifiers to determine the extent of the similarities between the pathways downstream of the identified latent disease-omics feature molecules and those of clinically approved drugs for CKD and/or MI. The hold-out validation shows that this method is highly reliable as indicated by the high prediction performance measures (i.e. accuracy scores > 0.94, precision scores > 0.71, recall scores > 0.85, F1 scores > 0.81) for both CKD and MI (Supplementary Tables S29 and S30).

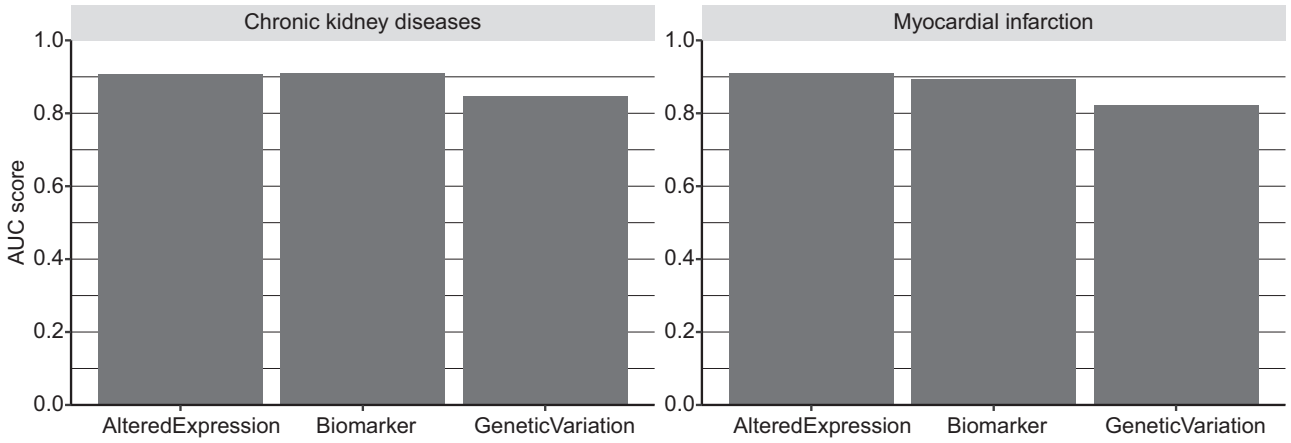


Fig. 3. The AUC scores of each modality for the chronic kidney disease and the myocardial infarction. The AUC scores are calculated for AlteredExpression, Biomarker, GeneticVariation modalities for each disease (chronic kidney disease, myocardial infarction) and shown as bar graphs. The bars are shown as the AUC scores of 10× leave-one-modality/disease-out cross-validations. The raw data are available as Supplementary Tables S3–S18

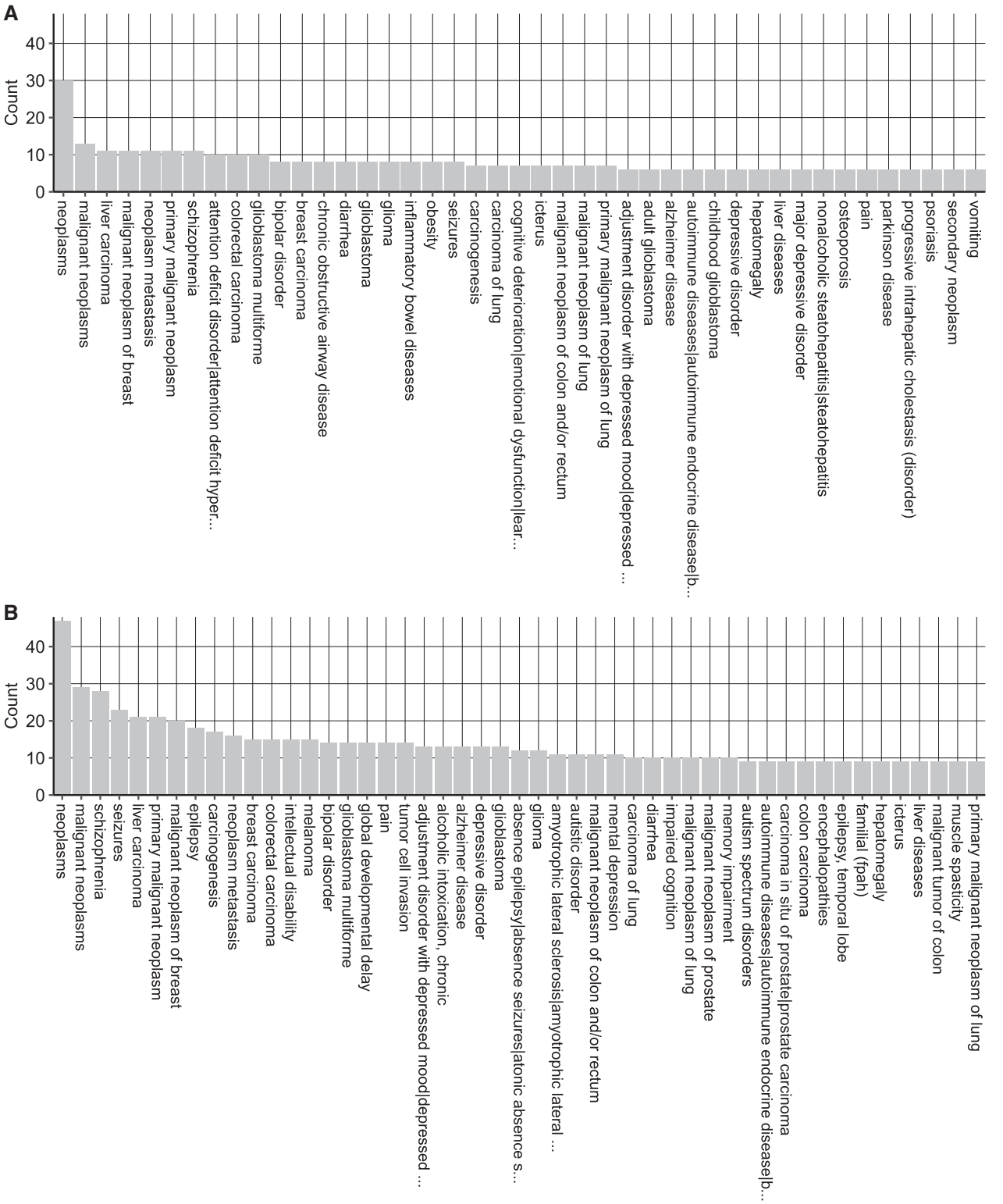


Fig. 4. Enrichment analysis of the original human diseases in the DisGeNET database from which the latent disease-omics features for CKD and MI are identified. (**A**) CKD, (**B**) MI. The numbers of the predicted latent disease-omics features that appear in each disease are indicated as 'count'. The long disease names are cut short and indicated as '.' at their ends. The raw data are available as Supplementary Table S19

Based on this highly reliable performance of the method, we applied it to infer molecular pathways impacted by the identified latent therapeutic targets (i.e. latent disease-omics features) (Fig. 8, Supplementary Tables S31–S34). This analysis found none of the CKD candidates share their pathways with the known CKD targets. Moreover, none belong to the same KEGG pathways. For MI, two candidates (ASL, LAMTOR1) are predicted as outside the known MI pathways, nor do they belong to the same KEGG pathways.



Fig. 5. The organ/cell expression patterns of the predicted disease-omics features for CKD and MI. (**A**) The heatmap indicating the expression levels of each predicted target (bottom) in each organ (left) for each disease (top). The raw data are available as Supplementary Table S20. (**B**) The heatmap indicating the expression levels for each predicted target (bottom) in each cell-type (left) for each disease (top). The raw data are available as Supplementary Table S21. (**C**) The organ enrichment analysis result for the predicted targets for each disease (top). The enrichment levels are indicated as –log10(q-values). The q-value ($q$value)=0.05 is indicated as a red line in each graph. The raw data are available as Supplementary Table S22. (**D**) The cell-type enrichment analysis result for the predicted targets for each disease (top). The enrichment levels are indicated as –log10(q-values). The q-value ($q$value) = 0.05 is indicated as a red line in each graph. The raw data are available as Supplementary Table S23

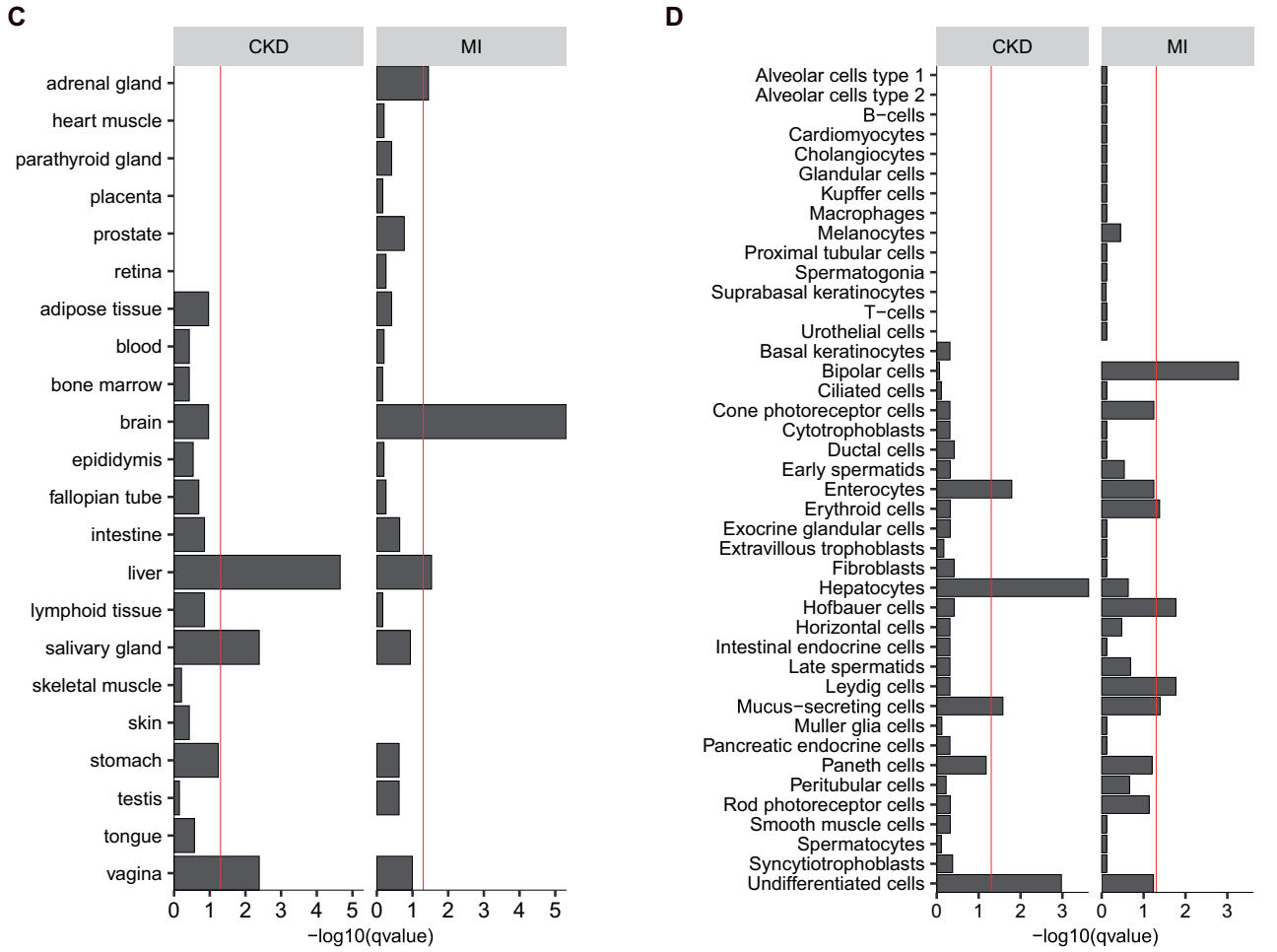


Fig. 5. (Continued)

Additionally, nine other candidates (AK2, GABRA5, GABRB3, GALT, GRM7, PILRA, PRKG2, PYGL, GNPAT) are also predicted as outside the known MI pathways, although they belong to the same KEGG pathways as the MI pathways. In contrast, two MI candidates (CHRM1, GRM3) and the known MI targets share the parts of their pathways.

Next, we examined potential SEs resulting from targeting the identified latent disease-omics feature molecules (Fig. 9, Supplementary Table S35). In this study, we focused on the 176 serious adverse outcomes. The hold-out validation shows F1 scores > 0.5 for 124 out of the 176 SEs, suggesting that this prediction method is relatively useful. This prediction found four candidates (AK2, ASL, PILRA, PYGL) that are free of the selected 176 serious adverse outcomes, suggesting that they are less harmful therapeutic-targets.

3.4 Retrospective pre-clinical and clinical validation of the 69 targets

Next, we searched the publications and Clinicaltrials.gov (https://www.clinicaltrials.gov) for the predicted therapeutic utility of the 69 targets (Table 1, Supplementary Table S36, see also Section 2.6). This examination found pre-clinical and/or clinical evidence supporting the therapeutic usefulness of the 17 out of 69 targets for renal and/or cardiovascular diseases, providing the independent and retrospective validation of their predicted therapeutic repurposing possibilities.

These 17 targets are ATP8B1 (1/69), ASL (3/69), GUCY2C (4/69), GALT (8/69), BLOC1S2 (9/69), AEBP1 (10/69), CD79B (24/69), EDS8 (26/69), PYGL (29/69), IL36RN (30/69), LMF1 (35/69), THRSP (50/69), SPATC1L (53/69), H3-3B (54/69), DRD5 (56/69), UBE2D1 (57/69), AK2 (62/69), where the numbers in () are the ranking of each target among the 69 predicted targets (e.g. 1/69 indicating the 1st among the 69, 3/69 indicating the 3rd among the 69, etc.) according to their predicted probabilities (Supplementary Tables S4–S9).

The pre-clinical and clinical evidence include experimental results with knockout mice of the target gene(s) and the outcomes obtained from clinical-trials, respectively. For example, knockout mice for AK2, an MI candidate, are reported to show cardiac dysfunctions (Zhang *et al.*, 2021). This mouse study suggests a role of AK2 in cardiac development and/or function. Hence, AK2 could serve as a therapeutic target for cardiovascular diseases such as MI.

Deficiency of ASL, a candidate for both CKD and MI, is a rare genetic disorder resulting in argininosuccinic aciduria, a defective urea cycle condition leading to the insufficient breakdown/removal of nitrogen from the body, and consequently the patients develop hypertension (Kho *et al.*, 2018). Hypertension is a known risk factor for both renal and cardiovascular diseases including CKD and MI (Clemmer *et al.*, 2022; Ku *et al.*, 2019). A clinical trial (NCT02252770) was conducted to evaluate the benefit of a nitric oxide dietary supplementation to argininosuccinic aciduria patients, but no outcomes are yet reported.

A cross-transplantation study using the kidneys from the knockout mice of DRD5, a CKD candidate, shows hypertension and cardiac dysfunctions in this mouse model (Asico *et al.*, 2011). Furthermore, both hypertension and cardiac dysfunctions are known risk factors for renal diseases such as CKD (Clemmer *et al.*, 2022; Ku *et al.*, 2019), hence, supporting the prediction of DRD5 as a therapeutic candidate for CKD.

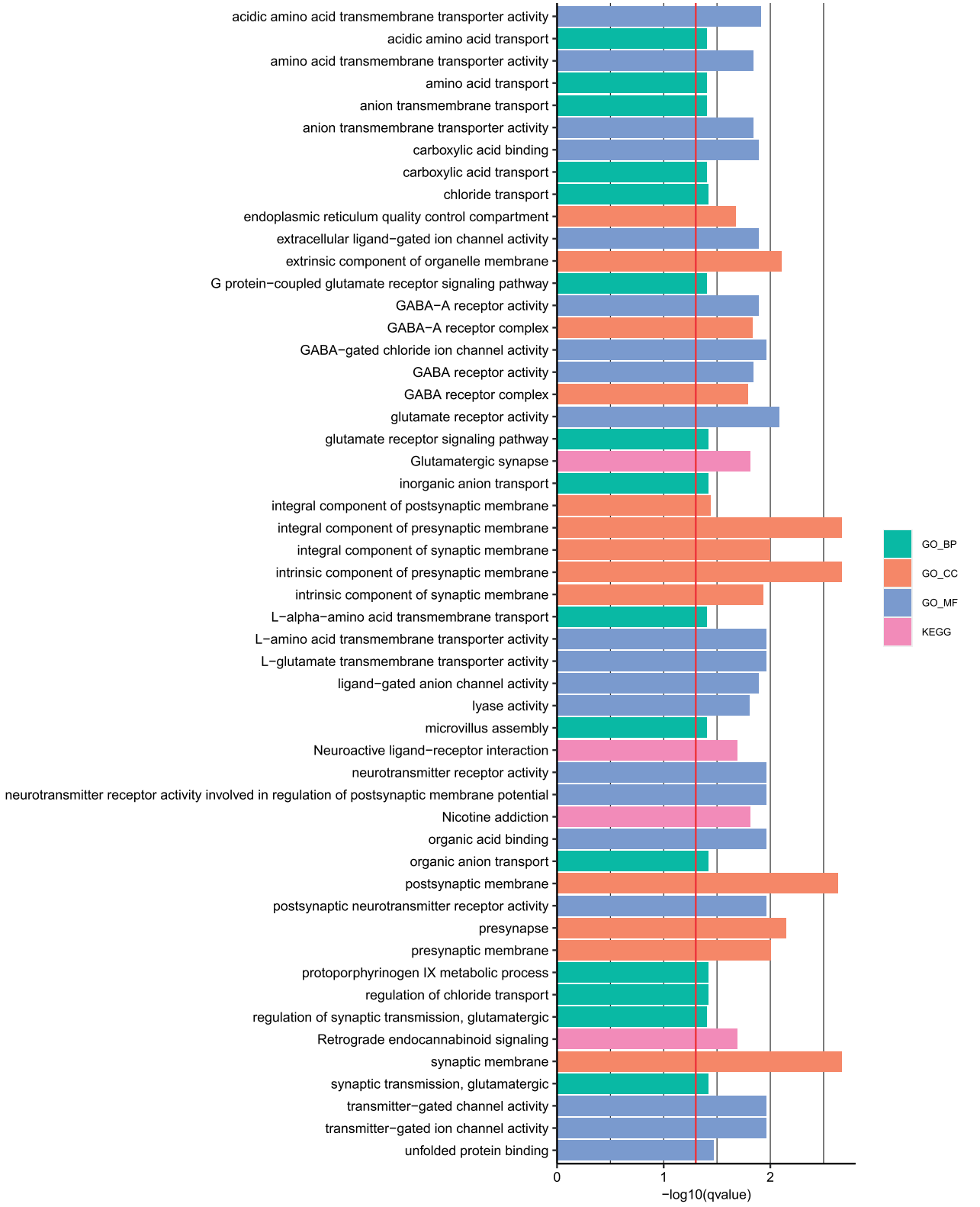


Fig. 6. KEGG and GO enrichment analyses of the predicted disease-omics features for MI. The enrichment values for KEGG pathways and GO terms are shown as bars. The enrichment value is indicated as –log10(q-values). Those with q-values < 0.05 are shown. The q-value (qvalue)=0.05 is indicated as a red line in each graph. The raw data are available as Supplementary Tables S24–S28

GUCY2C, guanylate cyclase 2C, is predicted as a therapeutic target for both CKD and MI (Table 1). In the clinical trial (NCT03217591), therapeutic effects of a soluble guanylate cyclase stimulator, IW-1973 (a.k.a., Praliciguat) for diabetic nephropathy/diabetic kidney diseases were evaluated. The outcomes show several metrics supporting further investigation of Praliciguat for diabetic kidney diseases (Hanrahan *et al.*, 2020).

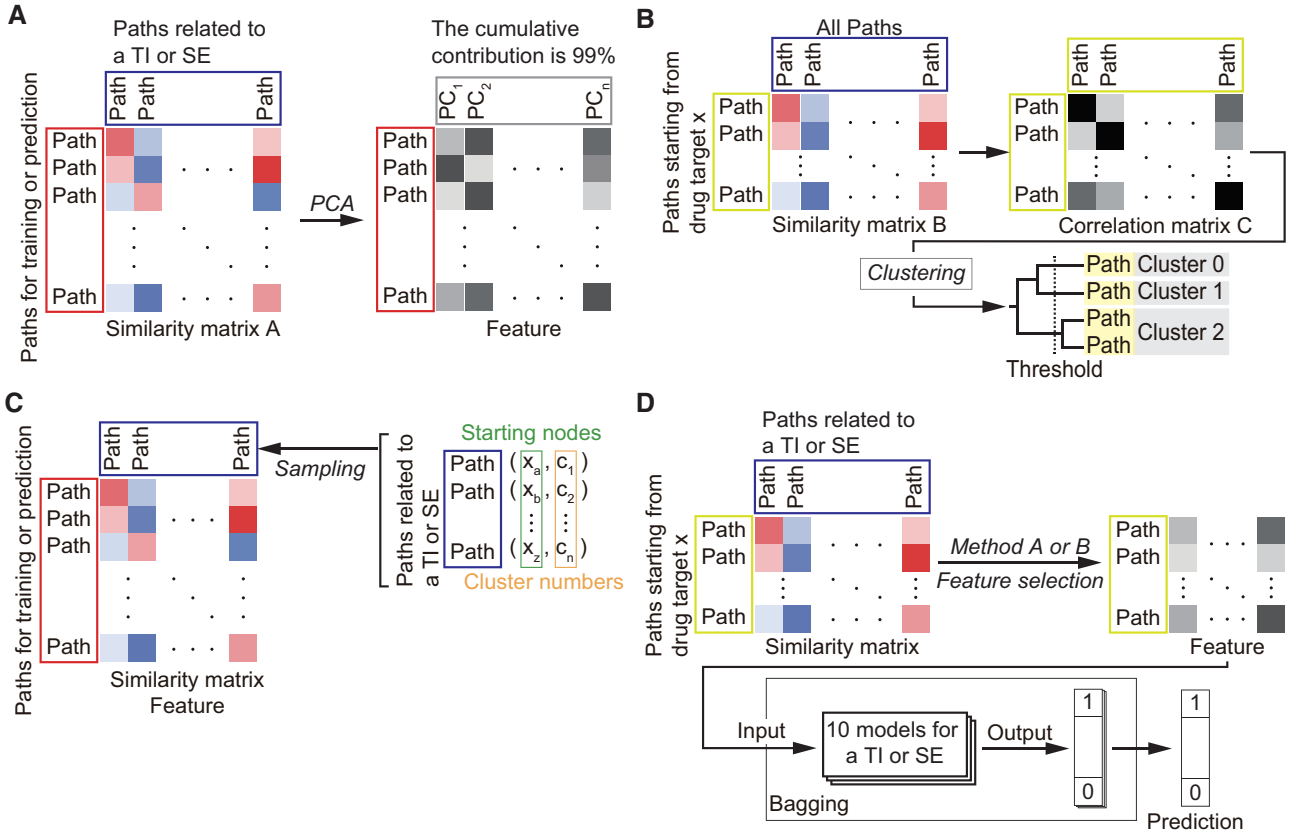


Fig. 7. Schematic diagram of the edit-distance-based classifier. (**A**) The feature selection module for Method A. (**B**) The path—cluster linking module for Method B. (**C**) The feature selection module for Method B. (**D**) The prediction module for Methods A and B. See 'Edit-distance-based classifier' in the Section 2.5 for the detailed step-by-step description. The raw data are available as Supplementary Tables S29 and S30

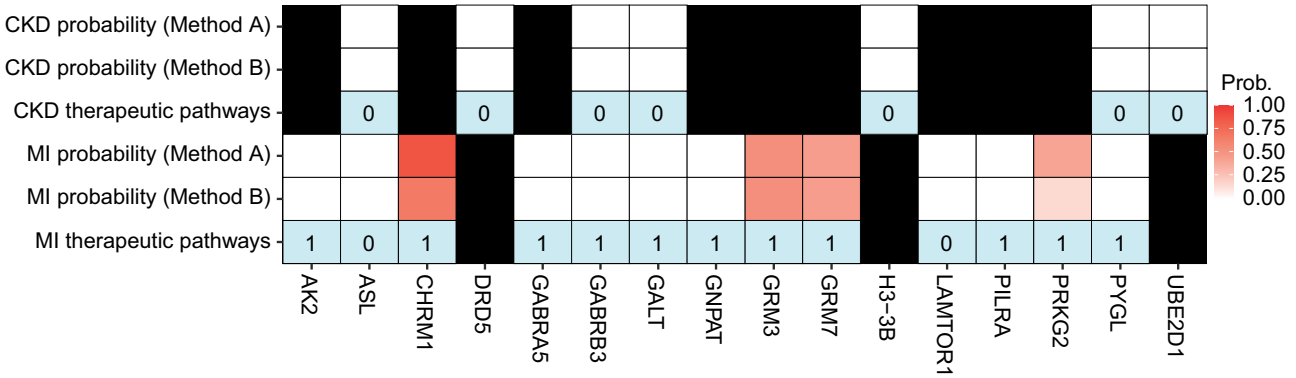


Fig. 8. The deduced therapeutic pathways of the predicted targets. The probabilities that the identified latent omics-features elicit their predicted therapeutic effect(s) via the clinically approved therapeutic pathways for each target disease (CKD: chronic kidney disease, MI: myocardial infarction) are shown as heatmap (CKD/MI probability on the left). For which of the diseases (CKD versus MI) each target is predicted are indicated by open (predicted for) or filled (not-predicted for) cells. Whether the targets are within or outside the same KEGG pathway(s) of the known targets of the clinically approved drugs for the corresponding target disease are indicated as '1' and '0', respectively, in the corresponding cell (CKD/MI therapeutic pathways on the left). The raw data are available as Supplementary Tables S31–S34

Unilateral ureteral-obstruction in mice results in the upregulation of H3-3B, a CKD candidate, in the kidneys (Shindo *et al.*, 2018). In addition, a knockdown experiment of histone cell cycle regulation defective homolog A (HIRA) in a normal rat kidney cell (NRK-52) causes the decreased H3-B3 expression and increased fibrogenesis (Shindo *et al.*, 2018). Furthermore, in patients with IgA nephropathy, H3-3B immune-stains positively correlate with kidney fibrosis (Shindo *et al.*, 2018). These results support a therapeutic candidacy of H3-3B for renal diseases including CKD.

Genetic mutation of PYGL, a candidate for both CKD and MI, in human prevents effective glycogen breakdown in the liver leading to glycogen storage diseases (Zhan *et al.*, 2021). While no clinical implications for CKD or MI or other renal/cardiovascular diseases in the patients of these conditions are recorded (NCT02385162), our expression analysis (Fig. 5) suggests that the liver is a potential therapeutic target for both CKD and MI. Hence, PYGL could serve as a therapeutic target for CKD, MI and/or other renal/cardiovascular diseases.

UBE2D1, a CKD candidate, is inferred as a potential biomarker for diabetes-related sepsis by a machine-learning pipeline using public databases (Wang *et al.*, 2022). Diabetes is a known risk factor for CKD (Shahbazian and Rezaii, 2013), hence, further supporting the candidacy of UBE2D1 for a CKD target as deduced herein.

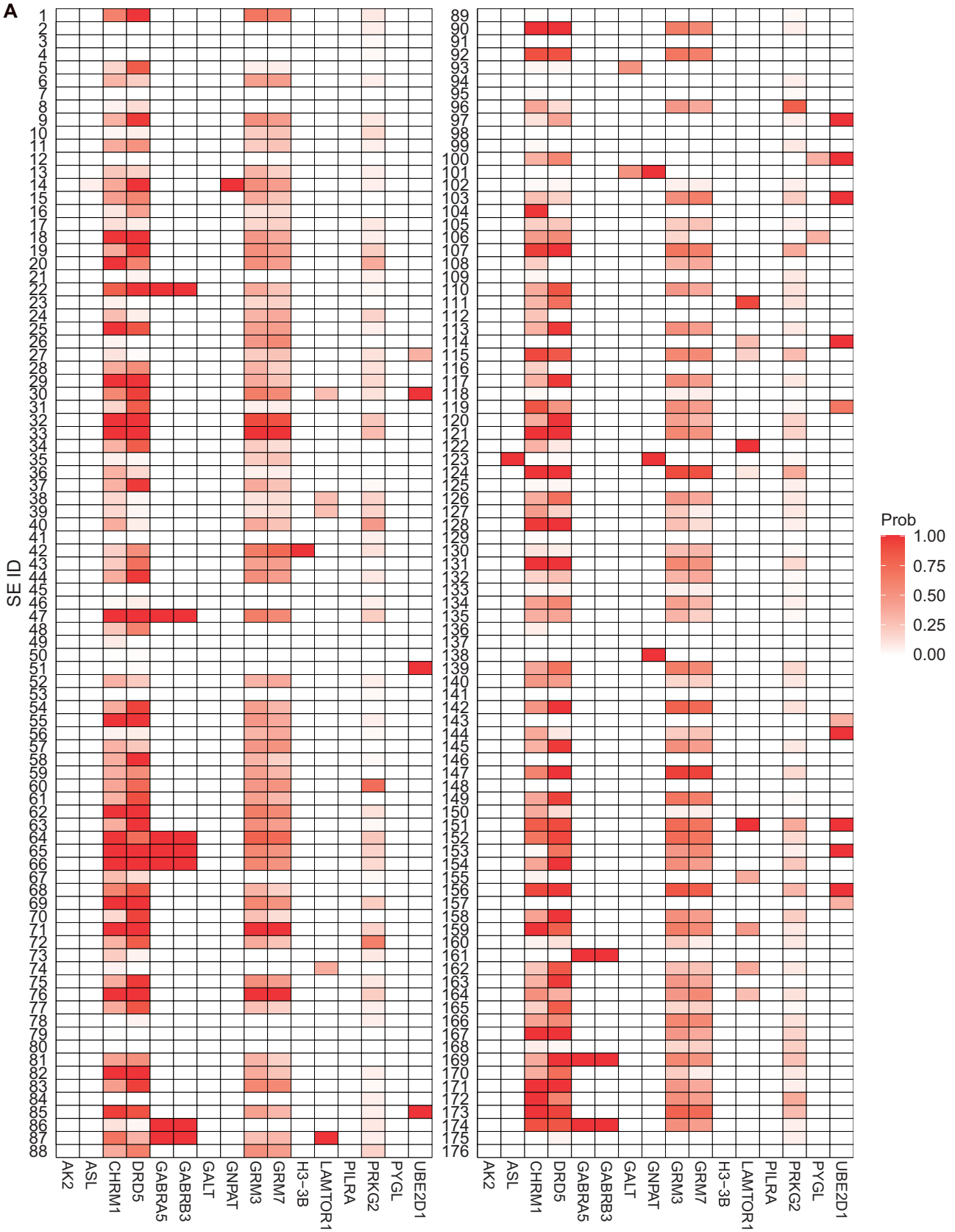


Fig. 9. The side-effects inferred by the edit-distance-based classifier for the predicted disease omics-features upon their therapeutic-targeting. (**A**) The probabilities for the 176 serious side-effects (left) are indicated as the heatmap for the CKD targets (bottom). The raw data are available as Supplementary Table S35. (**B**) The probabilities for the 176 serious side-effects (left) are indicated as the heatmap for the MI targets (bottom). The raw data are available as Supplementary Table S35. Each side-effect is indicated as ID number (SE_ID) on the left of each panel and they are fully described in Supplementary Table S35

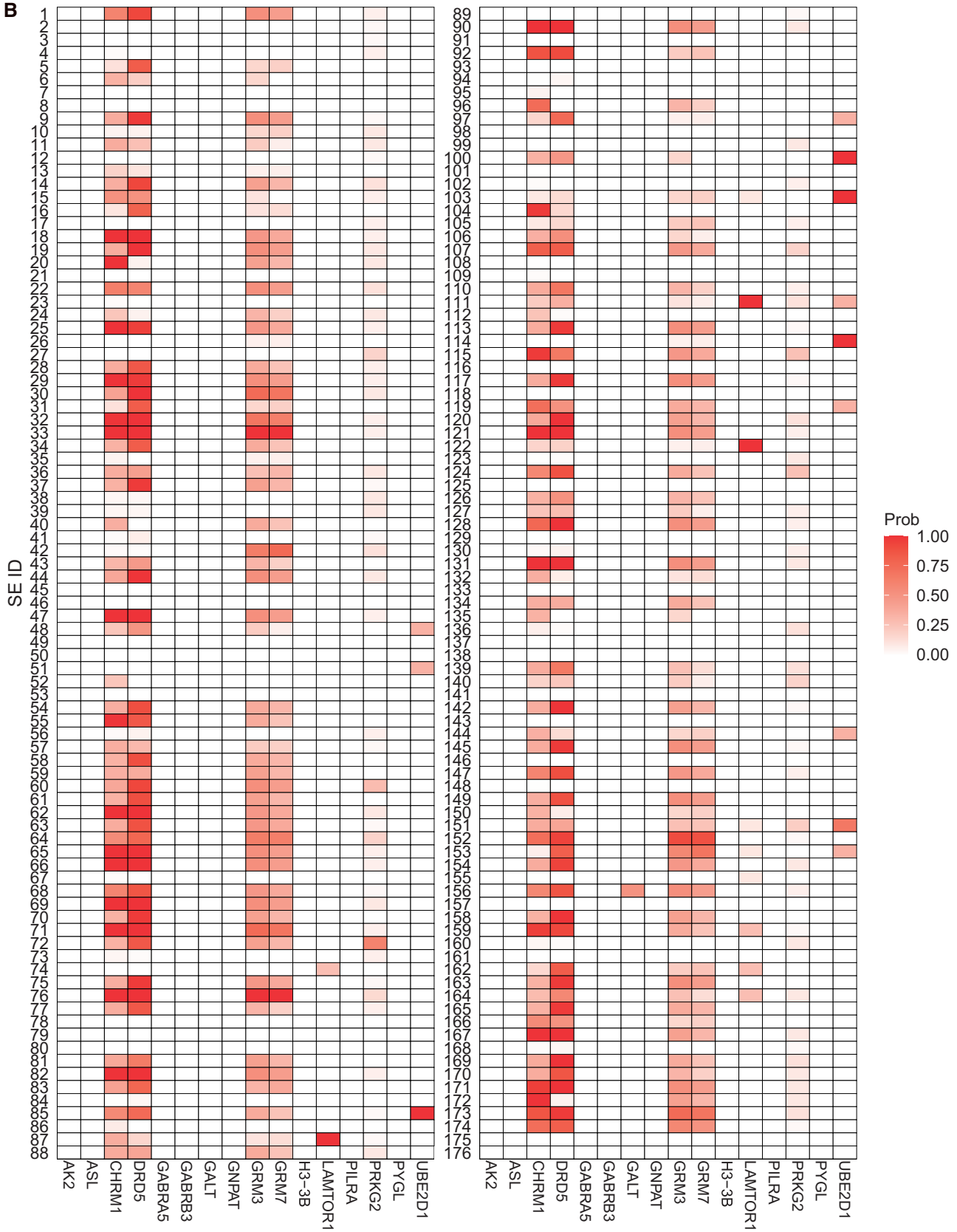


Fig. 9. (Continued)

These retrospective and independent validation results in Table 1 further support the therapeutic possibilities of the candidates reported in this study for CKD and/or MI and/or other renal and/or cardiovascular diseases. At the same time, they also strengthen the possibility that the predicated-targets without the retrospective evidence (i.e. those with blank in both the Publication and ClinicalTrials.gov cells) would represent new disease-omics features for treating and/or detecting CKD and/or MI.

4 Discussion

In this article, we applied a multi-modal soft-clustering method to the multiple disease-omics datasets and uncovered latent molecular similarities across 6955 human diseases (Figs 1–4). By exploiting these molecular similarities, we identified 69 targets that could be therapeutically repurposed for CKD and/or MI treatments (Table 1). The comprehensive omics analyses, in combination with an edit-distance-based classifier, found their underlying therapeutic mechanisms (Figs 5, 6, 8 and 9).

Importantly, we found the evidence retrospectively supporting the predicted therapeutic utility of the 17 targets in the pre-clinical and clinical data space (Table 1).

Recently, other soft-clustering methods for multi-modal data are reported (Yan *et al.*, 2021; Zhang *et al.*, 2022). While their utility or effectiveness with multi-modal biomedical data remains unknown, they may provide an additional framework to the analysis of the multi-modal disease-omics data studied in this article.

Our primary motivation of this study is to identify repurposable disease-omics molecular features that could be repurposed from a disease to another disease to its treatment and/or detection on the basis of multi-modal disease–disease similarities that are underrecognized in the conventional disease classification. Toward this goal, we used the soft-clustering of human diseases by multi-modal generative topic modeling to detect subtle differences in the multi-modal features of the diseases (Fig. 1). As a result, we were able to develop an algorithm that exhibits mostly AUC > 0.8 for predicting the missing modality features of 6955 human diseases (Fig. 2A). In this study, we tested this method with three modalities, Ae, Bm, Gv, for 6955 diseases (Fig. 2A), and four modalities, Ae, Bm, Gv, Mb, for 158 diseases (Fig. 2B). The result shows both sets result in virtually equivalent performance, suggesting the scalability of the method with additional modalities.

This method identified molecular features shared by CKD/MI and non-renal/non-cardiovascular diseases such as neoplasia and mental/neurological disorders (Fig. 4), indicating a latent underlying mechanism shared among these diseases. The neoplasia can be regarded as a partial cellular reprogramming, as it is accompanied by the aberrant activations of large number of genes (Buganim *et al.*, 2012; Suvà *et al.*, 2013; Ward and Thompson, 2012; Xing *et al.*, 2020). This phenomenon could be reflected in the molecular similarities between the renal/cardiovascular diseases (e.g. CKD and MI) and the neoplasia. It is also recently reported that MI accelerates breast cancer via innate immune reprogramming (Koelwyn *et al.*, 2020). This clinical observation might be a consequence of their molecular and mechanistic similarities as predicted in this study. Moreover, various clinical observations also suggest that CKD and cancer are mutual risk-factors, but without any clear molecular mechanisms (Wong *et al.*, 2016). Hence, it is possible that the predicted molecular mechanisms/pathways shared by CKD and neoplasia reported in this article may be an underlying molecular mechanism of these clinical observations.

Virtually all peripheral organs such as the liver, the kidney, the heart, etc. are under the control of neural inputs and these organs feedback their physiological information to the neural organ such as the brain (Imai *et al.*, 2008; Underwood and Altounian, 2021). Hence, such inter-organ neural feed-forward and feed-back loops could be reflected in the similar molecular features and underlying mechanisms of the renal/cardiovascular diseases (e.g. CKD and MI) and the mental/neurological disorders as predicted by the method reported herein. In support of this possibility, many mental disorders are prevalent in CKD patients (Simoes *et al.*, 2019). Furthermore, myocardial infarction is often followed by deteriorated mental health conditions (De Hert *et al.*, 2018; Lloyd, 1987). Despite such clinical evidence, no concrete molecular mechanisms explaining these clinical observations remain unknown. Thus, the common molecular mechanisms and/or pathways described in this study could be the ones.

This study identified 69 molecules that could be targeted for the treatment and/or detection of CKD and/or MI treatments on the basis of the similarities of CKD and/or MI to neoplasia and/or mental/neurological disorders (Table 1). Their expression patterns and KEGG/GO analyses indicate they are enriched in the brain and the metabolic organs such as the liver and their physiological functions (Figs 5 and 6). These results are coherent with the molecular similarities between the CKD/MI and mental/neurological disorders described in this article. They are also consistent with the fact that many of the renal/cardiovascular diseases including CKD and MI are broadly regarded as metabolic and life-style diseases (Sharifi-Rad *et al.*, 2020; Thomas *et al.*, 2011).

The edit-distance-based classifier shows two types of therapeutic mechanisms by which these 69 candidates could elicit their effects in the treatments of CKD, MI and/or other renal-/cardiovascular-diseases. Those that function via the pathways that are also targeted by the drugs approved for the corresponding diseases (i.e. CKD, MI), and the others that function independently of them (Fig. 8). The independent pathways may be a part of the previously unknown molecular mechanisms underlying the corresponding disease(s). In this case, their therapeutic-targeting could lead to the development of 'first-in-class' drugs for the corresponding diseases. In contrast, those within the already-targeted pathways are activated or inhibited by the existing drugs. Hence, they could be further developed by adding new indications for the diseases that are described as molecularly similar in this article.

The edit-distance-based classifier is also applied to evaluate putative SEs that could accompany the therapeutic targeting of these candidate molecules (Fig. 9). The result shows the four (AK2, ASL, PILRA, PYGL) are less harmful targets. This analysis provides useful information for selecting out those that are likely less toxic, prior to spending labor-, time- and cost-intensive pre-clinical and clinical studies during the therapeutic development.

The likeliness of the repurposability of the predicted CKD and MI targets is further strengthened by the retrospective finding of the therapeutic implications of the 17 targets in the pre-clinical and clinical-trials data space, despite their absence in the training datasets (Table 1). In addition to these molecular targets with the retrospective evidence, we also found those without any retrospective evidence and they may represent new molecular features that could therapeutically be developed to treat and/or detect CKD and/or MI.

In this study, we introduced a multi-modal generative topic modeling approach to find repurposable molecular targets and their use to unveil latent disease–disease similarities, and the characterization of their suitability for therapeutic development by the edit-distance-based method. While the results show many promises, there are some limitations that must be noted:

1. The multi-modal generative topic modeling computes the disease similarities according to the known multi-modal features of one or more of the human diseases. Hence, the features unlinked to any of the diseases are excluded from the outputs.
2. A caution must be paid to the interpretation of the cell-type expression (Fig. 5B) and the enrichment results (Fig. 5D). This is due to the apparent biases in the cell-type representations in the currently available human single-cell transcriptome databases,
3. All paths in the edit-distance classifier are generated with a target as the starting node. Hence, if the target is the most downstream component, it cannot be analyzed by this approach.
4. The edit-distance classifier reported herein is a mechanism-based predictor. Hence, any 'off-target' effects are out of scope.

Table 1. The retrospective validation of the predicted therapeutic targets in the pre-clinical and clinical data space

CKD	MI	Targets	Publication	ClinicalTrials.gov
	MI	ADRM1		
CKD	MI	AEBP1	DOI:10.1186/s12967-021-03000-3	
	MI	AK2	DOI:10.1016/j.bbrc.2021.01.097	
CKD	MI	ASL		NCT02252770
CKD	MI	ATP8B1		NCT02094222
	MI	BARX1		
CKD	MI	BLOC1S2	DOI:10.1038/cdd.2015.128	
	MI	C15orf41		
	MI	C19orf12		
	MI	C2-AS1		
	MI	CCDC115		
	MI	CD79B		NCT04668365
	MI	CHRM1		
CKD		CRNKL1		
CKD		DRD5	DOI:10.1681/ASN.2010050533	
	MI	ECT		
	MI	EDS8		NCT02165085
CKD	MI	EPB41		
	MI	ERLEC1		
CKD		ERVK-19		
CKD		FAM13A		
	MI	GABRA5		
CKD	MI	GABRB3		
CKD	MI	GALT		NCT02519504
	MI	GNPAT		
	MI	GRIK1		
	MI	GRM3		
	MI	GRM7		
CKD	MI	GUCY2C		NCT03217591
CKD		H3-3B	DOI:10.1038/s41598-018-32518-8	
	MI	HBB-LCR		
	MI	HBE1		
	MI	HM13		
CKD	MI	HSD3B7		
CKD		HTR3A		
	MI	IKZF2		
CKD		IL36RN	DOI:10.1016/j.kint.2017.09.017	
CKD	MI	IREB2		
	MI	LAMTOR1		
CKD		LINC01185		
	MI	LMF1		NCT03912181
CKD	MI	NARS1		
	MI	NDUFB2		
CKD		NME8		
	MI	ODAM		
	MI	P3H3		
	MI	PCDH19		
	MI	PILRA		
	MI	PRKG2		
CKD	MI	PYGL		NCT02385162
	MI	RANBP2		
	MI	RHAG		
	MI	SARM1		
	MI	SLC17A7		
CKD	MI	SLC25A13		
	MI	SLC6A5		
CKD	MI	SPAG8		
	MI	SPAST		
CKD	MI	SPATC1L	DOI:10.3892/mmr.2017.7334	
	MI	SPNS1		
	MI	STK26		
CKD	MI	SYN3		
CKD		THRSP	DOI:10.1155/2014/520281	

(continued)

Table 1. (continued)

CKD	MI	Targets	Publication	ClinicalTrials.gov
CKD	MI	TRP-AGG2-6		
CKD		UBE2D1	DOI:10.1155/2022/9463717	
CKD	MI	UGGT1		
	MI	UROS		
	MI	UST		
CKD		ZDHHC13		

Note: The publications and clinical-trials are indicated as DOI and clinicaltrial.gov NCT numbers, respectively. The blank cells are those without publications or clinicaltrials.gov evidence. The raw data are available as Supplementary Table S36.

5. The approach introduced here does not consider 'druggability' of the identified molecules, which are also important factors to determine whether the targets are repurposable (Owens, 2007).

Despite such limitations, the retrospective validation (Table 1), together with the high AUC scores obtained by the cross-validation (Figs 2 and 3) and the modality scalability (Fig. 2), demonstrates the effectiveness of the method in uncovering latent disease–disease similarities and therapeutic repurposing possibilities across diverse diseases and modalities. Hence, the method is expected to be effective, not only for CKD or MI, but also for other types of diseases and with different and/or additional combinations of disease feature modalities.

Acknowledgements

The authors thank K. Sugisaka, R. Takahashi, R. Kitaura and R. Ishikawa for their administrative assistance. They are also grateful to the members of Karydo TherapeutiX, Inc. and Sato lab at ATR for advice and discussion throughout the course of this work.

Author contributions

Satoshi Kozawa (Data curation [Equal], Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Writing—original draft [Supporting], Writing—review & editing [Supporting]), Hirona Yokoyama (Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Writing—original draft [Supporting], Writing—review & editing [Supporting]), Kyoji Urayama (Data curation [Supporting], Investigation [Supporting], Writing—review & editing [Supporting]), Kengo Tejima (Data curation [Supporting], Investigation [Supporting], Writing—review & editing [Supporting]), Hotaka Doi (Data curation [Supporting], Writing—review & editing [Supporting]), Shunki Takagi (Investigation [Supporting], Writing—review & editing [Supporting]), and Thomas N. Sato (Conceptualization [Lead], Data curation [Equal], Formal analysis [Equal], Funding acquisition [Lead], Investigation [Equal], Methodology [Equal], Project administration [Lead], Supervision [Lead], Writing—original draft [Lead], Writing—review & editing [Lead])

Software and data availability

The code and curated data reported in this article are available at: https://github.com/skozawa170301ktx/MultiModalDiseaseModeling.

Funding

This work was supported in part by Innovative Science and Technology Initiative for Security Grant Number JPJ004596 ATLA Japan (T.N.S.), JST ERATO Grant Number JPMJER1303 Japan (T.N.S.), Nakatani Foundation (T.N.S.) and AMED under Grant Number JP21he2102002 (T.N.S.).

Conflict of interests: All authors are employees of Karydo TherapeutiX, Inc.

References

Al-Saleem,J. *et al.* (2021) Knowledge graph-based approaches to drug repurposing for COVID-19. *J. Chem. Inf. Model.*, **61**, 4058–4067.
Anderson,J.L. and Morrow,D.A. (2017) Acute myocardial infarction. *N. Engl. J. Med.*, **376**, 2053–2064.
Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
Asico,L. *et al.* (2011) Lack of renal dopamine D5 receptors promotes hypertension. *J. Am. Soc. Nephrol.*, **22**, 82–89.
Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.
Bisgin,H. *et al.* (2012) Investigating drug repositioning opportunities in FDA drug labels through topic modeling. *BMC Bioinformatics*, **13**, S6.
Blei,D.M. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.
Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.*, **2008**, P10008.
Buganim,Y. *et al.* (2012) Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, **150**, 1209–1222.
Carbon,S. *et al.* (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
Chen,T.K. *et al.* (2019) Chronic kidney disease diagnosis and management: a review. *JAMA*, **322**, 1294–1304.
Clemmer,J.S. *et al.* (2022) Physiological mechanisms of hypertension and cardiovascular disease in end-stage kidney disease. *Curr. Hypertens. Rep.*, **24**, 413–424.
De Hert,M. *et al.* (2018) The intriguing relationship between coronary heart disease and mental disorders. *Dialogues Clin. Neurosci.*, **20**, 31–40.
García del Valle,E.P. *et al.* (2021) DisMaNET: a network-based tool to cross map disease vocabularies. *Comput. Methods Programs Biomed.*, **207**, 106233.
Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. USA*, **104**, 8685–8690.
Hanrahan,J.P. *et al.* (2020) Effects of the soluble guanylate cyclase stimulator praliciguat in diabetic kidney disease a randomized placebo-controlled clinical trial. *Clin. J. Am. Soc. Nephrol.*, **16**, 59–69.
Hasin,Y. *et al.* (2017) Multi-omics approaches to disease. *Genome Biol.*, **18**, 1–15.
Hooshmand,S.A. *et al.* (2021) A multimodal deep learning-based drug repurposing approach for treatment of COVID-19. *Mol. Divers.*, **25**, 1717–1730.
Imai,J. *et al.* (2008) Regulation of pancreatic beta cell mass by neuronal signals from the liver. *Science*, **322**, 1250–1254.
Janssens,Y. *et al.* (2018) Disbiome database: linking the microbiome to disease. *BMC Microbiol.*, **18**, 1–6.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Ke,G. *et al.* (2017) LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, **30**.
Kho,J. *et al.* (2018) Argininosuccinate lyase deficiency causes an endothelial-dependent form of hypertension. *Am. J. Hum. Genet.*, **103**, 276–287.
Koelwyn,G.J. *et al.* (2020) Myocardial infarction accelerates breast cancer via innate immune reprogramming. *Nat. Med.*, **26**, 1452–1458.
Ku,E. *et al.* (2019) Hypertension in CKD: core curriculum 2019. *Am. J. Kidney Dis.*, **74**, 120–131.
Lamy,J.B. *et al.* (2015) PyMedTermino: an open-source generic API for advanced terminology services. *Stud. Health Technol. Inform.*, **210**, 924–928.

Lemaitre,G. *et al.* (2017) Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.*, **18**, 1–5.
Li,Y. *et al.* (2021) Evaluating disease similarity based on gene network reconstruction and representation. *Bioinformatics*, **37**, 3579–3587.
Lloyd,G.G. (1987) Myocardial infarction and mental illness: a review. *J. R. Soc. Med.*, **80**, 101–104.
Malas,T.B. *et al.* (2019) Drug prioritization using the semantic properties of a knowledge graph. *Sci. Rep.*, **9**, 9, 1–10.
Menche,J. *et al.* (2015) Uncovering disease–disease relationships through the incomplete interactome. *Science*, **347**, 1257601.
Navarro,G. (2001) A guided tour to approximate string matching. *ACM Comput. Surv.*, **33**, 31–88.
Owens,J. (2007) Determining druggability. *Nat. Rev. Drug Discov.*, **6**, 187–187.
Panchapakesan,U. and Pollock,C. (2018) Drug repurposing in kidney disease. *Kidney Int.*, **94**, 40–48.
Park,K. (2019) A review of computational drug repurposing. *Transl. Clin. Pharmacol.*, **27**, 59–63.
Pedregosa,F. F. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
Perakakis,N. *et al.* (2018) Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism*, **87**, A1–A9.
Piñero,J. *et al.* (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
Pushpakom,S. *et al.* (2019) Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.*, **18**, 41–58.
Reel,P.S. *et al.* (2021) Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.*, **49**, 107739.
Roller,S. and im Walde,S.S. (2013) A multimodal LDA model integrating textual, cognitive and visual modalities. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, pp. 1146–1157. Association for Computational Linguistics.
Schuler,J. *et al.* (2022) Evaluating the performance of drug-repurposing technologies. *Drug Discov. Today*, **27**, 49–64.
Shahbazian,H. and Rezaii,I. (2013) Diabetic kidney disease; review of the current knowledge. *J. Renal Inj. Prev.*, **2**, 73–80.
Sharifi-Rad,J. *et al.* (2020) Diet, lifestyle and cardiovascular diseases: linking pathophysiology to cardioprotective effects of natural bioactive compounds. *Int. J. Environ. Res. Public Health*, **17**, 2326.
Shindo,T. *et al.* (2018) TGF-β1 promotes expression of fibrosis-related genes through the induction of histone variant H3.3 and histone chaperone HIRA. *Sci. Rep.*, **8**, 1–13.
Silva,A.C.S. *et al.* (2019) Neuropsychiatric disorders in chronic kidney disease. *Front Pharmacol*, **10**, 932.
Suvà,M.L. *et al.* (2013) Epigenetic reprogramming in cancer. *Science (1979)*, **340**, 1567–1570.
Thomas,G. *et al.* (2011) Metabolic syndrome and kidney disease: a systematic review and meta-analysis. *Clin. J. Am. Soc. Nephrol.*, **6**, 2364–2373.
Thul,P.J. *et al.* (2017) A subcellular map of the human proteome. *Science (1979)*, **356**, eaal3321.
Uhlén,M. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419.
Underwood,E. and Altounian,V. (2021) A sense of self. *Science*, **372**, 1142–1145.
Valle,F. *et al.* (2022) Multiomics topic modeling for breast cancer classification. *Cancers (Basel)*, **14**, 1150.
Wang,X. *et al.* (2022) UBE2D1 and COX7C as potential biomarkers of diabetes-related sepsis. *Biomed. Res. Int.*, **2022**, 1.
Ward,P.S. and Thompson,C.B. (2012) Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell.*, **21**, 297–308.
Wen,Z. *et al.* (2021) Mining heterogeneous clinical notes by multi-modal latent topic model. *PLoS ONE*, **16**, e0249622.
Wold,S. *et al.* (1987) Principal component analysis. *Chemometr. Intell. Lab. Syst.*, **2**, 37–52.
Wong,G. *et al.* (2016) Chronic kidney disease and the risk of cancer: an individual patient data meta-analysis of 32,057 participants from six prospective studies. *BMC Cancer*, **16**, 488.
Xing,Q.R. *et al.* (2020) Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. *Sci. Adv.*, **6**, eaba1190.
Yan,A. *et al.* (2021) A clustering algorithm for multi-modal heterogeneous big data with abnormal data. *Front. Neurorobot.*, **15**, 64.
Youden,W.J. (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–35.
Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
Zhan,Q. *et al.* (2021) Glycogen storage disease type VI with a novel PYGL mutation: two case reports and literature review. *Medicine (Baltimore)*, **100**, e25520.
Zhang,S. *et al.* (2021) Adenylate kinase AK2 isoform integral in embryo and adult heart homeostasis. *Biochem. Biophys. Res. Commun.*, **546**, 59–64.
Zhang,T. *et al.* (2022) Weighted Bayesian Gaussian mixture model for roadside LiDAR object detection. *arXiv*, arXiv:2204.09804.
Zheng,Y. *et al.* (2014) Topic modeling of multimodal data: an autoregressive approach. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1370–1377.

3)

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Latent inter-organ mechanism of idiopathic pulmonary fibrosis unveiled by a generative computational approach

Satoshi Kozawa[1,2,8], Kengo Tejima[1,2,8], Shunki Takagi[1,2,8], Masataka Kuroda[3,4], Mari Nogami-Itoh[3], Hideya Kitamura[5], Takashi Niwa[5], Takashi Ogura[5], Yayoi Natsume-Kitatani[3,6] & Thomas N. Sato[1,2,7]✉

Idiopathic pulmonary fibrosis (IPF) is a chronic and progressive disease characterized by complex lung pathogenesis affecting approximately three million people worldwide. While the molecular and cellular details of the IPF mechanism is emerging, our current understanding is centered around the lung itself. On the other hand, many human diseases are the products of complex multi-organ interactions. Hence, we postulate that a dysfunctional crosstalk of the lung with other organs plays a causative role in the onset, progression and/or complications of IPF. In this study, we employed a generative computational approach to identify such inter-organ mechanism of IPF. This approach found unexpected molecular relatedness of IPF to neoplasm, diabetes, Alzheimer's disease, obesity, atherosclerosis, and arteriosclerosis. Furthermore, as a potential mechanism underlying this relatedness, we uncovered a putative molecular crosstalk system across the lung and the liver. In this inter-organ system, a secreted protein, kininogen 1, from hepatocytes in the liver interacts with its receptor, bradykinin receptor B1 in the lung. This ligand–receptor interaction across the liver and the lung leads to the activation of calmodulin pathways in the lung, leading to the activation of interleukin 6 and phosphoenolpyruvate carboxykinase 1 pathway across these organs. Importantly, we retrospectively identified several pre-clinical and clinical evidence supporting this inter-organ mechanism of IPF. In conclusion, such feedforward and feedback loop system across the lung and the liver provides a unique opportunity for the development of the treatment and/or diagnosis of IPF. Furthermore, the result illustrates a generative computational framework for machine-mediated synthesis of mechanisms that facilitates and complements the traditional experimental approaches in biomedical sciences.

Idiopathic pulmonary fibrosis (IPF) is a chronic disease characterized by scarring in the interstitium of the lung, affecting 3–9 and 4 or less per 100,000 person-years in North America/Europe and South America/East-Asia, respectively[1,2]. Both the incidence and poor prognosis of IPF increase with age[3,4]. Specifically, the median age of the newly diagnosed is 62 years-old and their prognosis is poor—3–5 years of survival rate.

There are two Food and Drug Administration (FDA)-approved drugs for the treatment of IPF: nintedanib and pirfenidone[1,2]. Nintedanib is a tyrosine kinase inhibitor. Pirfenidone is an inhibitor of transforming growth factor (TGF)-beta production and downstream signaling, collagen synthesis and fibroblast proliferation. Hence, these drugs are regarded as pleiotropic anti-fibrosis drugs. Currently there are no IPF-specific therapeutics. Furthermore, the precise IPF diagnosis requires complex and multiple-types of tests as its overlapping pathologies with other interstitial lung fibrosis diseases[1,2]. These are in part due to the complexity of the IPF pathogenesis and to its ill-defined cellular and molecular mechanisms.

[1]Karydo TherapeutiX, Inc., 2-2-2 Hikaridai, Seika-Cho, Soraku-Gun, Kyoto 619-0288, Japan. [2]The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan. [3]National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan. [4]Mitsubishi Tanabe Pharma Corporation, Kanagawa, Japan. [5]Kanagawa Cardiovascular and Respiratory Center, Kanagawa, Japan. [6]Institute of Advanced Medical Sciences, Tokushima University, Tokushima, Japan. [7]V-iCliniX Laboratory, Nara Medical University, Nara, Japan. [8]These authors contributed equally: Satoshi Kozawa, Kengo Tejima and Shunki Takagi. ✉email: island1005@gmail.com

While IPF was classically considered an inflammatory disease, a new picture is emerging[1–6]. The increasing molecular and cellular evidence suggests IPF is driven by an activation of the lung epithelium. In this model, the ectopic activation of the alveolar epithelial cells results in the production of chemokines, growth factors, and extracellular matrix proteins, promoting the migration, growth, and/or differentiation of fibroblasts, and macrophages and other immune cells. Furthermore, various life-style and environmental factors, and also genetic factors are reported to influence the onset, progression, and/or mortality.

While the clinical translation of these recent advancements in understanding the IPF mechanism within the lung tissue is expected, the onset and progression of human diseases involve multiple organs -i.e., the inter-organ mechanism[7–11]. The immune responses occur in a variety of diseases such as metabolic, neoplastic, cardiovascular diseases and also in aging[12–14]. The causative roles of gut microbiota are becoming recognized in an increasing number of diseases[15]. The nervous system influences metabolic states and vice versa[16]. Metabolic dysregulation is a risk factor for many diseases and they also accelerate aging influencing the longevity[17]. Exosomes are another type of systemic factors that are associated with many types of diseases[18]. The interactions of immune cells and lung cells are involved in the pathogenesis of IPF[1,2,4–6]. Furthermore, life-style and aging are critical influencers of IPF[1,5]. Hence, it is conceivable that the inter-organ crosstalk mediated by the immune cells, systemic factors, and/or neural system could be involved in the onset, progression, and/or complications of IPF. However, very little is studied on these possibilities.

Based on this background, we postulate that the cross-talk between the lung and the non-lung organs is a part of the mechanism in the onset and/or progression of IPF. The obvious choice of the approach to test this hypothesis is to examine molecular and cellular changes in non-lung organs that accompany, precede, or follow the pathological changes of the lung in IPF. However, this approach would be difficult as the availability of non-lung tissues from the IPF patients is limited, if any.

The availability of multi-modal omics data of multiple organs and diseases is growing in the public space[19–27]. Such data space, together with computational methods, could allow us to deduce what occurs in the non-lung organs of IPF-patients and to simulate how they are regulated.

Hence, we reasoned that such a large set of multi-modal omics data across many organs and diseases in the human body provides an uncharted biomedical space where an organ-to-organ interaction that causes and/or exacerbates IPF is embedded. To uncover such a latent inter-organ mechanism of IPF, we employ a generative computational approach (also referred to as "generative artificial intelligence (AI)")—a computational method that can produce various types of contents such as sentences, images, molecular structures, working-hypotheses(models), etc.[28].

Towards this goal, we designed a generative computational approach as follows:

1. To detect mechanistic relatedness of IPF to non-respiratory/non-pulmonary diseases.
2. To identify molecular features that characterize the relatedness detected in 1.
3. To identify ligand–receptor relationships across multiple organs that are linked to the features identified in 2.
4. To generate a map of the inter-organ mechanism of IPF with molecular and cellular resolution that explains the findings of 1–3.

Results

Detection of molecular relatedness of IPF to non-respiratory/non-pulmonary diseases

The molecular relatedness of IPF to non-respiratory/non-pulmonary diseases were identified by using the multi-modal generative topic modeling method that we developed and previously reported[29]. The overall design is summarized in Fig. 1, and it works as follows:

Two datasets are used for the multi-modal generative topic modeling: Datasets A and B. Dataset A consists of 6,954 human diseases excluding IPF, each of which is characterized by three disease omics modalities, AlteredExpression (Ae), Biomarker (Bm), and GeneticVariation (Gv), derived from DisGeNET v7.0[19,27]. "Ae" is the list of genes and proteins of which changes in expressions are associated with a corresponding disease(s). "Bm" is the list of biomarkers which are described for a corresponding disease(s). "Gv" is the list of genes of which mutations are reported for a corresponding disease(s). Dataset B consists of three types of IPF modality combination, each consisting of Bm/Gv (i.e., missing Ae), Gv/Ae (i.e., missing Bm), or Ae/Bm (i.e., missing Gv), also from DisGeNET v7.0[19,27]. The multi-modal generative topic modeling generates (i.e., predicts) the features of the missing modalities. The performance was evaluated by calculating the area under the receiver operating characteristic curve (AUC) values as previously described[29] and they were found to be above 0.8 for all three modalities (Supplementary Fig. S1). Next, from these computationally generated features, those derived from the modalities of IPF itself and those of obviously IPF-related diseases are removed. The diseases that are obviously related are those of which names contain "Pulmonary", "Lung", "Fibrosis", "Respir**", "Chest", "Pneumo**" (**could be any characters). The remaining features are now designated as "latent disease-omics features of IPF (also referred to as IPF-features)". Moreover, IPF and the diseases from which these IPF-features are derived in Dataset A establish "latent relatedness of IPF to other diseases".

Using this approach, we identified 83 latent IPF-features (Supplementary Table S1). The human-organ-expression analysis using THE HUMAN PROTEIN ATLAS v 21.1[23–25] (see also "Methods" section) found that their expression is most enriched in the liver (Fig. 2A, Supplementary Table S2). Additionally, we also detected the statistically significant (i.e., q-values < 0.05) enrichments in the immune system (bone marrow, lymphoid tissue, blood), the kidney, the thyroid gland, adipose tissue, the prostate, and the placenta. The cellular level analysis found the highest enrichment in the hepatocytes (Fig. 2B, Supplementary Table S2). In addition, we also detected the statistically significant (i.e., q-values < 0.05) enrichments in Kupffer cells and Hofbauer cells,

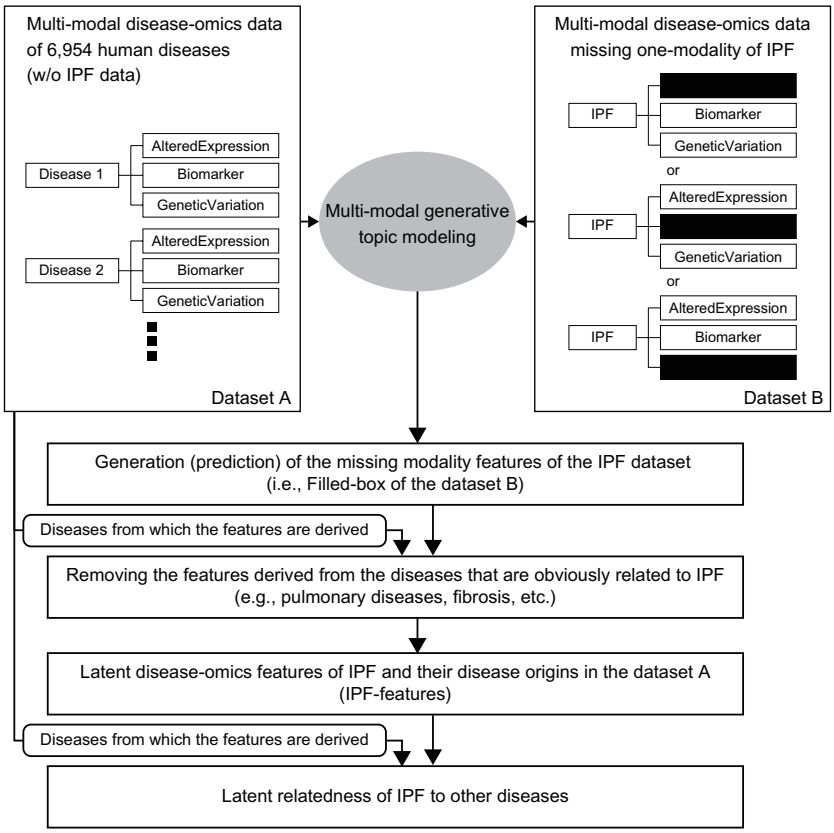


Figure 1. General overview of the multi-modal generative topic modeling approach for IPF. The previously developed method[29] is adapted to IPF.

two types of macrophages found in the liver and the placenta, respectively. These results suggest a possibility that the liver, in particular, hepatocytes and the intra-hepatic immune cells such as Kupffer cells, participate in the pathogenesis of IPF.

Latent relatedness of IPF to other diseases

Next, we determined the disease-label(s) of the 83 latent IPF-features to identify non-pulmonary/non-respiratory diseases to which IPF is related (Fig. 3, Supplementary Table S3) (see also "Methods" section). This analysis found these IPF-features are derived from neoplastic diseases. In addition, they are also labeled with autoimmune disorders, diabetes, Alzheimer's disease, rheumatoid arthritis, obesity, cardiovascular diseases (atherosclerosis, arteriosclerosis, hypertensive disease, etc.), systemic lupus erythematosus, and multiple sclerosis. The result suggests that these non-pulmonary/non-respiratory diseases are related to IPF at the molecular level.

Inter-organ mechanism of IPF

A putative inter-organ mechanism was computationally generated as described in Fig. 4. The step-by-step description (Steps 1–7) and the results from each step are as follows:

Step 1: To identify the ligands in the lung differentially expressed genes (DEgenes) (via CellChatDB, a human ligand–receptor combination database, as described in Fig. 4).

In IPF treatments, distinguishing IPF from the other non-IPF lung diseases is most critical for the better outcome[1–6]. Therefore, we analyzed DEgenes between IPF and non-IPF lung diseases subjects, rather than those between IPF and healthy subjects.

The DESeq2 analysis of the lung tissues obtained from 95 IPF and 204 non-IPF (unclassifiable interstitial pneumonia: UCIP, idiopathic nonspecific interstitial pneumonia: NSIP, idiopathic pleuroparenchymal fibroelastosis: PPFE, other idiopathic interstitial pneumonias: IIPs, hypersensitivity pneumonitis: HP, connective tissue diseases: CTD, and other interstitial lung disease) lung disease patients (see also "Methods" section) identified a total of 112 IPF-DEgenes (Supplementary Table S4).

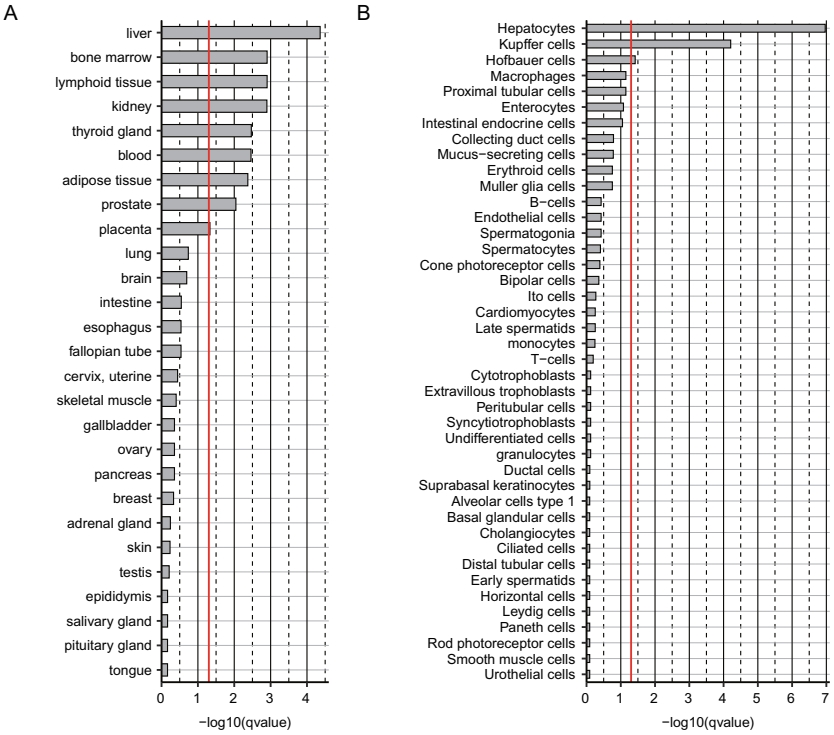


Figure 2. The organ and cell-enrichment analyses of the latent IPF-features. (**A**) The organ enrichment. (**B**) The cell-type enrichment. The enrichment level of the 83 IPF-features in each organ and each cell-type is shown as bar-graph of −log10(q-values) in the descending order. The q-value (qvalue) = 0.05 (the threshold for the statistical significance) is indicated as a red line in each graph. The raw data are available as Supplementary Table S2.

The CellChatDB analysis identified seven ligands (C–C motif chemokine ligand 18: CCL18, C–X–C motif chemokine ligand 9: CXCL9, C–X–C motif chemokine ligand 10: CXCL10, C–X–C motif chemokine ligand 11: CXCL11, interleukin 6: IL6, interferon gamma: IFNG, selectin E: SELE) in the 112 IPF-DEgenes (Table 1).

Step 2: To identify the receptors for the ligands in 1 (via CellChatDB as described in Fig. 4).

For the seven ligands identified in Step 1 above, CellChatDB identified eight receptors (corresponding to 10 genes) (atypical chemokine receptor 1: ACKR1, CXCR3, atypical chemokine receptor 3: ACKR3, interleukin 6 receptor/interleukin 6 cytokine family signal transducer: IL6R/IL6ST, interferon gamma receptor 1/interferon gamma receptor 2: IFNGR1/IFNGR2, CEA cell adhesion molecule 1: CEACAM1, cluster of differentiation 44: CD44, Golgi glycoprotein 1: GLG1), forming 13 ligand–receptor pairs (CCL18-ACKR1, CXCL9-ACKR1, CXCL10-ACKR1, CXCL11-ACKR1, CXCL9-CXCR3, CXCL10-CXCR3, CXCL11-CXCR3, CXCL11-ACKR3, IL6-IL6R/IL6ST, IFNG-IFNGR1/IFNGR2, SELE-CEACAM1, SELE-CD44, SELE-GLG1) (Table 1).

Step 3: To identify the receptors in the lung DEgenes (via CellChatDB as described in Fig. 4).

The CellChatDB analysis identified nine receptors (C–X–C motif chemokine receptor 3: CXCR3, C–X–C motif chemokine receptor 5: CXCR5, C–X–C motif chemokine receptor 6: CXCR6, bradykinin receptor B1: BDKRB1, cholinergic receptor nicotinic alpha 1 subunit: CHRNA1, transmembrane and immunoglobulin domain containing 3: TMIGD3, desmocollin 3: DSC3, programmed cell death 1: PDCD1, SELE) (Table 2).

Step 4: To identify the ligands for the receptors in 3 (via CellChatDB as described in Fig. 4).

For the nine receptors identified in Step 3 above, CellChatDB identified 16 ligands (platelet factor 4 variant 1: PF4V1, CXCL9, CXCL10, CXCL11, C–X–C motif chemokine ligand 13: CXCL13, platelet factor 4: PF4, C–X–C motif chemokine ligand 16: CXCL16, kininogen 1: KNG1, secreted LY6/PLAUR domain containing 1: SLURP1, secreted LY6/PLAUR domain containing 2: SLURP2, ectonucleoside triphosphate diphosphohydrolase 1: ENTPD1, desmoglein 1: DSG1, desmoglein 2: DSG2, cluster of differentiation 274: CD274, programmed cell death 1 ligand 2: PDCD1LG2, selectin P ligand: SELPLG), forming 17 ligand–receptor pairs (PF4V1–CXCR3, CXCL9–CXCR3, CXCL10–CXCR3, CXCL11–CXCR3, CXCL13–CXCR3, PF4–CXCR3, CXCL13–CXCR5, CXCL16–CXCR6, KNG1–BDKRB1, SLURP1–CHRNA1, SLURP2–CHRNA1, ENTPD1–TMIGD3, DSG1–DSC3, DSG2–DSC3, CD274–PDCD1, PDCD1LG2–PDCD1, SELPLG–SELE) (Table 2).

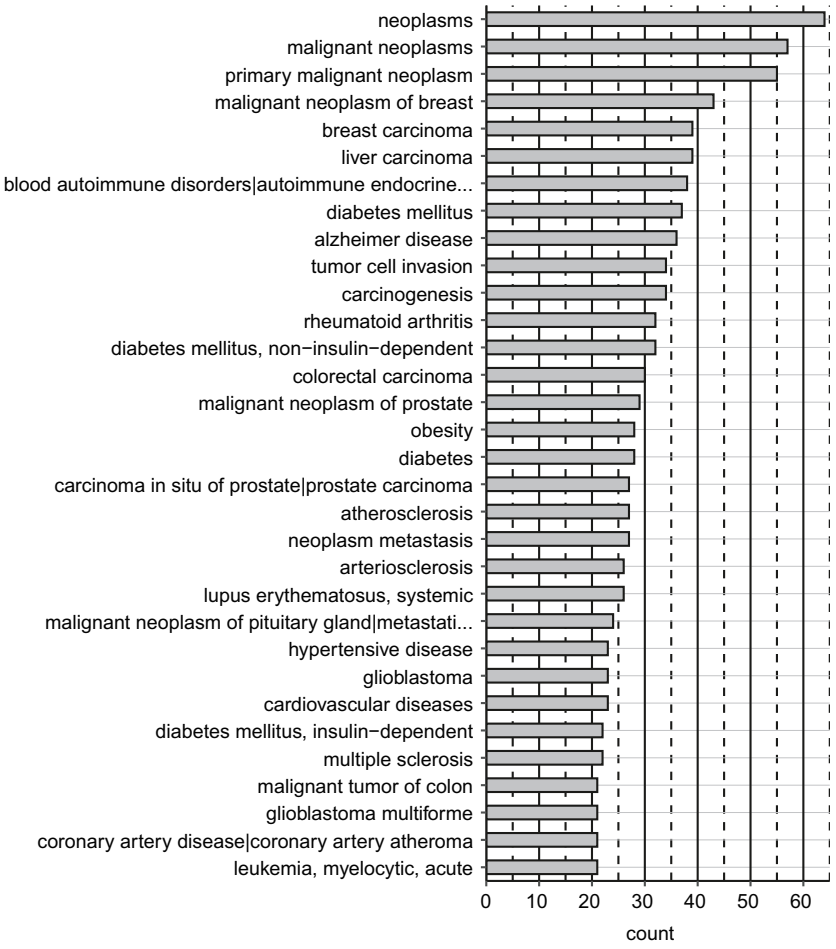


Figure 3. The latent diseases to which IPF is molecularly related. The frequency of the appearance of the 83 IPF-features in each disease is indicated as "count". Shown are the diseases of which counts are above 20 in the descending order. The long disease names are cut short and indicated as "..." at their ends. The raw data are available as Supplementary Table S3.

Step 5: To identify downstream and upstream targets of the ligand–receptor pairs found in Steps 2 and 4 by Kyoto encyclopedia genes and genomes (KEGG) pathway analysis and to select those that are among the 83 latent IPF-features identified by the multi-modal generative toping modeling.

We searched for the downstream and upstream signaling targets for these ligand–receptor pairs by the KEGG-mining (see "Methods" section). For the 13 ligand (lung)–receptor (non-lung) pairs (from Step 2, Table 1), we found six such targets in the latent IPF-features (Table 3)—engulfment and cell motility 1 (ELMO1) as the downstream target for CXCL9–CXCR3, CXCL10–CXCR3, and CXCL11–CXCR3 pairs, calcium/calmodulin dependent protein kinase IV (CAMK4) as the downstream target for the IFNG–IFNGR1 and IFNG–IFNGR2 pairs, and phosphoenolpyruvate carboxykinase 1 (PCK1) as the downstream target for the IL6–IL6R pair. In addition, calmodulin 1/calmodulin 2/calmodulin 3 (CALM1/CALM2/CALM3) were identified as upstream targets for the IL6-IL6R pair.

For the 17 ligand (non-lung)–receptor (lung) pairs (from Step 4, Table 2), we identified four targets (Table 4), ELMO1 as the downstream target for the PF4V1–CXCR3, CXCL9–CXCR3, CXCL10–CXCR3, CXCL11–CXCR3, CXCL13–CXCR3, CXCL13–CXCR5, PF4–CXCR3, and CXCL16–CXCR6 pairs, CALM1/CALM2/CALM3 as the downstream target for the KNG1–BDKRB1 pair.

Step 6: To identify cell-types where the ligands-receptors and their downstream and upstream singling targets found in Steps 1–5.

Our aim is to identify the inter-organ mechanism of IPF. The multi-modal generative topic-model found a possible involvement of the liver in this mechanism (Fig. 2). Hence, the ligand–receptor pair(s) that bridge the lung (the primary organ of IPF pathology) and the liver could be such a mechanism. Furthermore, to fulfill this

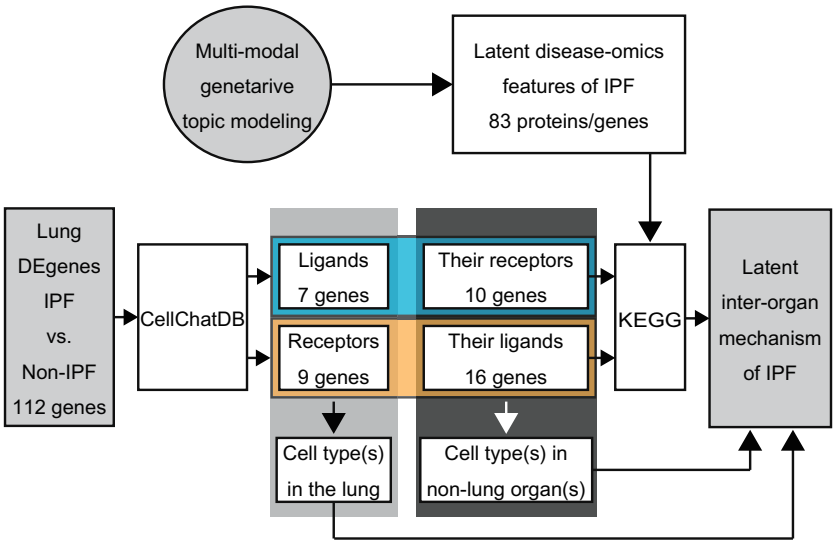


Figure 4. General overview of the computational framework to generate an inter-organ mechanism of IPF. See the "Methods" section for the detailed step-by-step description. The 83 latent IPF-features and 112 lung DEgenes (IPF vs. non-IPF) are found in Supplementary Tables S1 and S4, respectively.

Ligand (IPF-DEgene)	Receptor	Evidence
CCL18	ACKR1	PMID: 26740381
CXCL9	ACKR1	PMID: 26740381
CXCL10	ACKR1	PMID: 26740381
CXCL11	ACKR1	PMID: 26740381
CXCL9	CXCR3	KEGG: hsa04060
CXCL10	CXCR3	KEGG: hsa04060
CXCL11	CXCR3	KEGG: hsa04060
CXCL11	ACKR3	KEGG: hsa04060
IL6	IL6R / IL6ST	KEGG: hsa04060
IFNG	IFNGR1 / IFNGR2	KEGG: hsa04060
SELE	CEACAM1	PMID: 1378450
SELE	CD44	PMC4571854
SELE	GLG1	PMID: 11404363

Table 1. The ligands encoded by the IPF-DEgenes and their receptors. The ligands are IPF-DEgenes. The evidence for each ligand–receptor pair is indicated as described in CellChatDB. *ACKR1* atypical chemokine receptor 1, *ACKR3* atypical chemokine receptor 3, *CCL18* C–C motif chemokine ligand 18, *CD44* cluster of differentiation 44, *CEACAM1* CEA cell adhesion molecule 1, *CXCL9* C–X–C motif chemokine ligand 9, *CXCL10* C–X–C motif chemokine ligand 10, *CXCL11* C–X–C motif chemokine ligand 11, *CXCR3* C–X–C motif chemokine receptor 3, *GLG1* Golgi glycoprotein 1, *IFNG* interferon gamma, *IFNGR1* interferon gamma receptor 1, *IFNGR2* interferon gamma receptor 2, *IL6* interleukin 6, *IL6R* interleukin 6 receptor, *IL6ST* interleukin 6 cytokine family signal transducer, *SELE* selectin E.

mechanism, the expression of the non-lung component of the ligand–receptor pair should be enriched in the liver.

On the basis of this rationale, we examined the expression patterns of the non-lung components of the ligand–receptor pairs in the liver (Fig. 5, Supplementary Table S5). The analysis of the multi-organ human single-cell RNA sequencing (scRNA-seq) database, *Tabula Sapiens*, identified two pairs, KNG1 (the liver)–BDKRB1 (the lung) and IL6 (the lung)–IL6R/IL6ST (the liver), that could establish the lung-liver inter-organ mechanism. The expression of KNG1, the ligand for BDKRB1, is most enriched in hepatocytes, with lesser expression in the endothelial cells, fibroblasts, intrahepatic cholangiocytes, and T cells. The expression of IL6R/IL6ST, the receptor complex for IL6, is enriched in the endothelial cells of the hepatic sinusoid, intrahepatic cholangiocytes, and hepatocytes.

Ligand	Receptor (IPF-DEgene)	Evidence
PF4V1	CXCR3	KEGG: hsa04060
CXCL9	CXCR3	KEGG: hsa04060
CXCL10	CXCR3	KEGG: hsa04060
CXCL11	CXCR3	KEGG: hsa04060
CXCL13	CXCR3	KEGG: hsa04060
PF4	CXCR3	KEGG: hsa04060
CXCL13	CXCR5	KEGG: hsa04060
CXCL16	CXCR6	KEGG: hsa04060
KNG1	BDKRB1	KEGG: hsa04080
SLURP1	CHRNA1	KEGG: hsa04080
SLURP2	CHRNA1	KEGG: hsa04080
ENTPD1	TMIGD3	PMID: 21677139
DSG1	DSC3	PMID: 27298358
DSG2	DSC3	PMID: 27298358
CD274	PDCD1	PMID: 23954143
PDCD1LG2	PDCD1	PMID: 23954143
SELPLG	SELE	PMC4571854

Table 2. The receptors encoded by the IPF-DEgenes and their ligands. The receptors are IPF-DEgenes. The evidence for each ligand–receptor pair is indicated as described in CellChatDB. *BDKRB1* bradykinin receptor B1, *CD274* cluster of differentiation 274, *CHRNA1* cholinergic receptor nicotinic alpha 1 subunit, *CXCL9* C–X–C motif chemokine ligand 9, *CXCL10* C–X–C motif chemokine ligand 10, *CXCL11* C–X–C motif chemokine ligand 11, *CXCL13* C–X–C motif chemokine ligand 13, *CXCL16* C–X–C motif chemokine ligand 16, *CXCR3* C–X–C motif chemokine receptor 3, *CXCR5* C–X–C motif chemokine receptor 5, *CXCR6* C–X–C motif chemokine receptor 6, *DSG1* desmoglein 1, *DSG2* desmoglein 2, *DSC3* desmocollin 3, *ENTPD1* ectonucleoside triphosphate diphosphohydrolase 1, *KNG1* kininogen 1, *PF4* platelet factor 4, *PF4V1* platelet factor 4 variant 1, *PDCD1* programmed cell death 1, *PDCD1LG2* programmed cell death 1 ligand 2, *SELE* selectin E, *SELPLG* selectin P ligand, *SLURP1* secreted LY6/PLAUR domain containing 1, *SLURP2* secreted LY6/PLAUR domain containing 2, *TMIGD3* transmembrane and immunoglobulin domain containing 3.

Pathway	Ligand (IPF-DEgene)	Receptor	Latent disease-omics feature (IPF-feature)	Position
hsa04062	CXCL9	CXCR3	ELMO1	Downstream
hsa04062	CXCL10	CXCR3	ELMO1	Downstream
hsa04062	CXCL11	CXCR3	ELMO1	Downstream
hsa04380	IFNG	IFNGR1	CAMK4	Downstream
hsa04380	IFNG	IFNGR2	CAMK4	Downstream
hsa04151	IL6	IL6R	PCK1	Downstream
hsa05163	IL6	IL6R	CALM1	Upstream
hsa05163	IL6	IL6R	CALM2	Upstream
hsa05163	IL6	IL6R	CALM3	Upstream

Table 3. KEGG pathways of the ligand (IPF-DEgenes)-receptor pairs and their signaling molecules (IPF-features). The Pathways are from the KEGG pathways (human). The position indicates whether the corresponding latent disease-omics feature is the downstream or the upstream of the ligand–receptor pair in the corresponding KEGG pathway. *CALM1* calmodulin 1, *CALM2* calmodulin 2, *CALM3* calmodulin 3, *CAMK4* calcium/calmodulin dependent protein kinase IV, *CXCL9* C–X–C motif chemokine ligand 9, *CXCL10* C–X–C motif chemokine ligand 10, *CXCR3* C–X–C motif chemokine receptor 3, *ELMO1* engulfment and cell motility 1, *IFNG* interferon gamma, *IFNGR1* interferon gamma receptor 1, *IFNGR2* interferon gamma receptor 2, *IL6* interleukin 6, *IL6R* interleukin 6 receptor, *PCK1* phosphoenolpyruvate carboxykinase 1. KEGG pathways: hsa04062: Chemokine signaling pathway; hsa04380: Osteoclast differentiation; hsa04151: PI3K-Akt signaling pathway; hsa05163: Human cytomegalovirus infection.

Next, we examined the expression patterns of their partner components in the lung (Fig. 6, Supplementary Tables S6 and S7), which were originally found in the 112 IPF-lung DEgenes. The scRNA-seq analysis of IL6, the ligand for the IL6R/IL6ST receptor complex, in the lung using the *Tabula Sapiens* was performed. The result shows that the enrichment of IL6 expression in adventitial cells, endothelial cells, fibroblasts, mesothelial cells, respiratory mucous cells, and smooth muscle cells (Fig. 6A, Supplementary Table S6). We also examined whether the expression pattern of IL6 in the lung is altered in the IPF patients (Fig. 6B, Supplementary Table S7). The

Pathway	Ligand	Receptor (IPF-DEgenes)	Latent disease-omics feature (IPF-feature)	Position
hsa04062	PF4V1	CXCR3	ELMO1	Downstream
hsa04062	CXCL9	CXCR3	ELMO1	Downstream
hsa04062	CXCL10	CXCR3	ELMO1	Downstream
hsa04062	CXCL11	CXCR3	ELMO1	Downstream
hsa04062	CXCL13	CXCR3	ELMO1	Downstream
hsa04062	CXCL13	CXCR5	ELMO1	Downstream
hsa04062	PF4	CXCR3	ELMO1	Downstream
hsa04062	CXCL16	CXCR6	ELMO1	Downstream
hsa05200	KNG1	BDKRB1	CALM1	Downstream
hsa05200	KNG1	BDKRB1	CALM2	Downstream
hsa05200	KNG1	BDKRB1	CALM3	Downstream

Table 4. KEGG pathways of the ligand–receptor (IPF-DEgenes) pairs and their signaling molecules (IPF-features). The Pathways are from the KEGG pathways (human). The position indicates whether the corresponding latent disease-omics feature is the downstream or the upstream of the ligand–receptor pair in the corresponding KEGG pathway. *BDKRB1* bradykinin receptor B1, *CALM1* calmodulin 1, *CALM2* calmodulin 2, *CALM3* calmodulin 3, *CXCL9* C–X–C motif chemokine ligand 9, *CXCL10* C–X–C motif chemokine ligand 10, *CXCL11* C–X–C motif chemokine ligand 11, *CXCL13* C–X–C motif chemokine ligand 13, *CXCL16* C–X–C motif chemokine ligand 16, *CXCR3* C–X–C motif chemokine receptor 3, *CXCR5* C–X–C motif chemokine receptor 5, *CXCR6* C–X–C motif chemokine receptor 6, *ELMO1* engulfment and cell motility 1, *KNG1* kininogen 1, *PF4* platelet factor 4, *PF4V1* platelet factor 4 variant 1. KEGG pathways: hsa04062: Chemokine signaling pathway; hsa05200: Pathways in cancer.

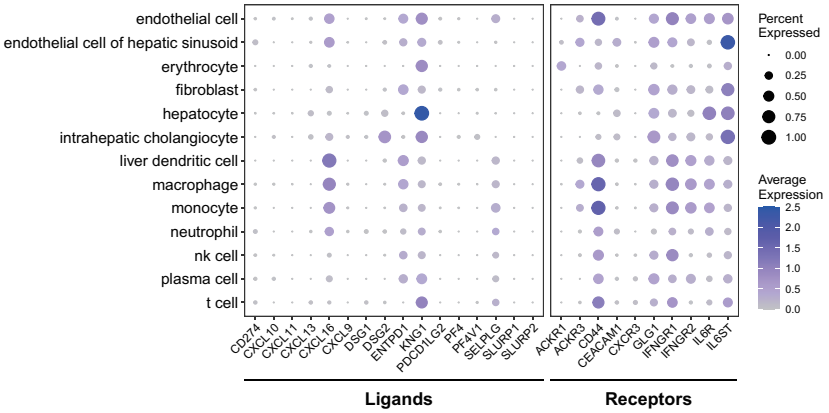


Figure 5. The hepatic expression of the ligands and receptors for the IPF pulmonary receptors and ligands. The level of each ligand and receptor in each cell-type in the liver is shown as dot. The size and the heat-intensity represent the ratio of cells expressing the gene in each cell-type cluster and the mean expression level of log-transformed counts [i.e., log(1 + count per 10,000)], respectively, as shown on the right side of the panel. The raw data are available as Supplementary Table S5. nk cell: natural killer cell.

result identified over twofold downregulation of the IL6 expression in endothelial cells and dendritic cells in the IPF lung. In addition, in the lung macrophage, its nearly twofold downregulation was also found. While the expression of BDKRB1 is detected more ubiquitously in the lung (Fig. 6A, Supplementary Table S6), it is nearly 1000-fold upregulated in the macrophages of the IPF lung, as compared to those of the healthy lung (Fig. 6B, Supplementary Table S7).

We also examined the expression patterns of their downstream and upstream signaling targets (Fig. 7). The most significant upregulation of CALM1/CALM2/CALM3, the downstream targets of the KNG1–BDKRB1 signaling and the upstream targets of the IL6–IL6R/IL6ST signaling, was detected in macrophages in the IPF lung (Fig. 7A, Supplementary Table S6). Lesser but statistically significant upregulation for one or more of these targets was also found in fibroblasts, dendritic cells, T/natural killer T (T/NKT) cells, ciliated cells, monocytes, mast cells, and alveolar type II cells (AT2) cells. Small but statistically significant downregulation was observed for CALM1 in alveolar type I cells (AT1) cells, AT2 cells, and club cells. Such downregulation was also detected for CALM3 in monocytes and dendritic cells.

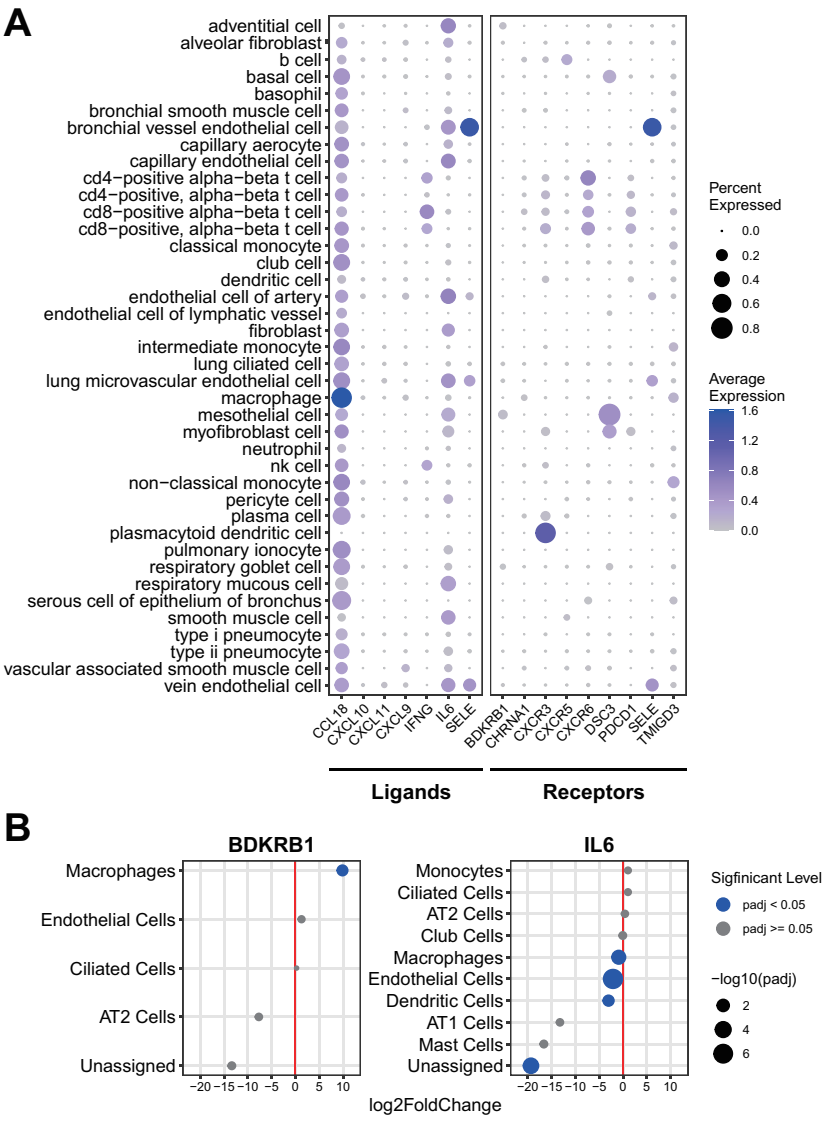


Figure 6. The expression of IL6 and BDKRB1 in the lung. (**A**) The level of each ligand and receptor (including IL6 and BDKRB1) in each cell-type in the lung of the healthy subjects (*Tabula Sapiens*) is shown as dot. The size and the heat-intensity represent the ratio of cells expressing the gene in each cell-type cluster and the mean expression level of log-transformed counts [i.e., log(1 + count per 10,000)], respectively, as shown on the right side of the panel. The raw data are available as Supplementary Table S6. nk cell: natural killer cell. (**B**) The differential expression of IL6 and BDKRB1 in each cell-type in the IPF-lung is shown as dot. The cell-types are indicated on the left. The differential expression of IPF vs. non-IPF is indicated as $\log_2$fold change ("log2FoldChange"). The dot size indicates the statistical significance of the differential expression as $-\log_{10}$p-adj ("− log10padj")—the larger size indicating more significant (i.e., less padj values). The blue and gray colors indicate padj < 0.05 and padj ≥ 0.05, respectively. The raw data are available as Supplementary Table S7. *padj* adjusted p-value, *AT1 cells* alveolar type I cells, *AT2 cells* alveolar type II cells.

The expression pattern of PCK1, the downstream target of IL6-IL6R/IL6ST signaling, was examined in the liver (Fig. 7B, Supplementary Table S6). The result shows its highest expression in hepatocytes. Its less abundant expression is detected in endothelial cells, erythrocytes, intrahepatic cholangiocytes, plasma cells, and T cells.

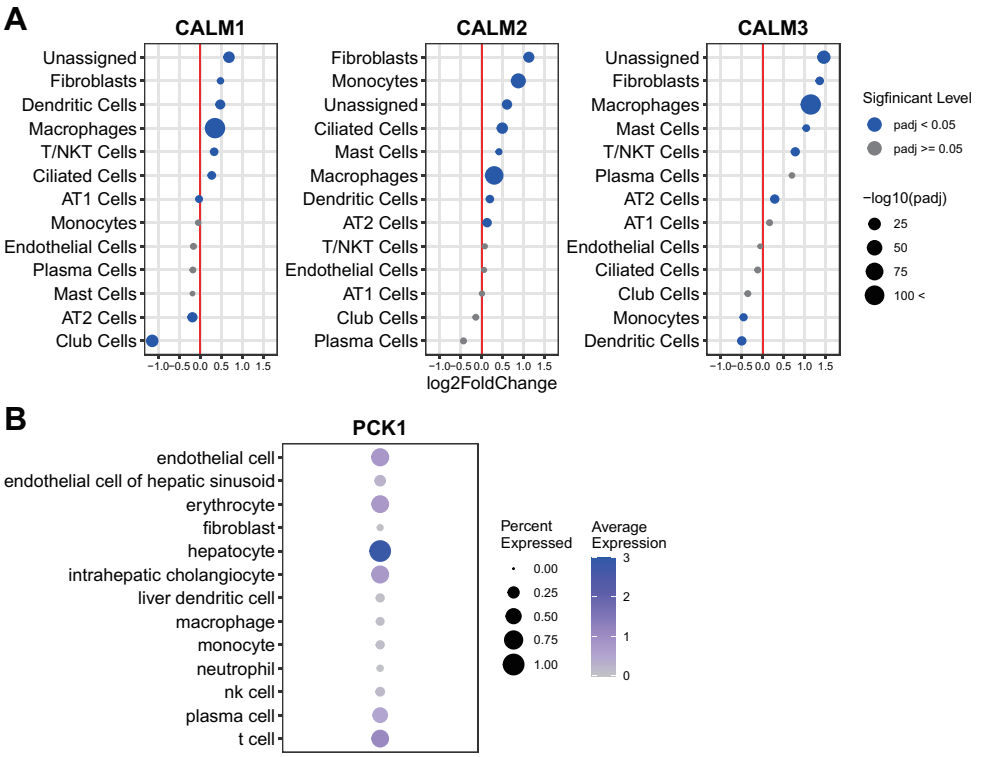


Figure 7. The expression of the signaling targets in the liver and the lung. (**A**) The differential expression of CALM1/CALM2/CALM3 in each cell-type in the IPF-lung is shown as dot. The cell-types are indicated on the left. The differential expression of IPF vs. non-IPF is indicated as $\log_2$fold change ("log2FoldChange"). The dot size indicates the statistical significance of the differential expression as $-\log_{10}$p-adj ("− log10padj")—the larger size indicating more significant (i.e., less padj values). The blue and gray colors indicate padj < 0.05 and padj ≥ 0.05, respectively. The raw data are available as Supplementary Table S7. padj: adjusted p-value; AT1 cells: alveolar type I cells; AT2 cells: alveolar type II cells. T/NKT cells: T/natural killer T cells. (**B**) The level of PCK1 in each cell-type in the liver is shown as dot. The size and the heat-intensity represent the ratio of cells expressing the gene in each cell-type cluster and the mean expression level of log-transformed counts [i.e., log(1 + count per 10,000)], respectively, as shown on the right side of the panel. The raw data are available as Supplementary Table S5. *nk cell* natural killer cell.

Step 7: To construct the inter-organ map on the basis of 1–6 results.

We put together the results obtained through Steps 1–6 and generated a landscape representing an inter-organ mechanism of IPF (Fig. 8). The logic is as follows: KNG1, expressed in the hepatic cells (Fig. 5), is the systemic ligand for its receptor, BDKRB1 (Table 2). BDKRB1 is also one of the 112 IPF-DEgenes expressed in the pulmonary cells (Table 1, Fig. 6B, Supplementary Table S1). Hence, the hepatic KNG1 directly interacts with pulmonary BDKRB1 across these organs (Fig. 8). CALM1/CALM2/CALM3, the latent IPF-features (Supplementary Table S1) are the known downstream targets of KNG1 (ligand)—BDKRB1 (receptor) signaling (KEGG: hsa05200, Pathways in cancer) (Table 4). In addition, CALM1/CALM2/CALM3 are also known upstream signaling components of the IL6 signaling (KEGG: hsa05163, Human cytomegalovirus infection pathway) (Table 3). CALM1/CALM2/CALM3 are expressed in the pulmonary cells and their expression is upregulated in the pulmonary macrophages and fibroblasts, etc. of the IPF lung (Fig. 7A). IL6 is one of the IPF-DEgenes (Fig. 6B, Supplementary Table S4) and is a systemic ligand for its receptor, IL6R/IL6ST (Table 1). IL6R/IL6ST complex is expressed in hepatic cells (Fig. 5). Hence, the signal from the liver is transduced to the lung via KNG1 (ligand)–BDKRB1 (receptor) interaction across these organs via CALM1/CALM2/CALM3 to the IL6 signal in the lung (Fig. 8). This pulmonary IL6 signal is transduced back to the liver via the IL6 (ligand)–IL6R/IL6ST (receptor) interaction in the liver (Fig. 8). PCK1, one of the IPF disease-omics features, is a known signaling molecule for the IL6 (ligand)–IL6R/IL6ST (receptor) interaction (KEGG: hsa04151, PI3K-Akt signaling pathway), and it is expressed in the hepatic cells (Fig. 7). Hence the IL6 signal from the IPF-lung is transduced in the liver via PCK1 signaling molecule (Fig. 8). With this logic, the mechanism described in Fig. 8 is generated. In this mechanism, the liver-derived KNG1 activates the CALM1/CALM2/CALM3 signaling pathway via BDKRB1 in the lung. This signal amplifies the expression and/or secretion of IL6 from the lung. The systemic IL6 activates the PCK1 signaling pathway via

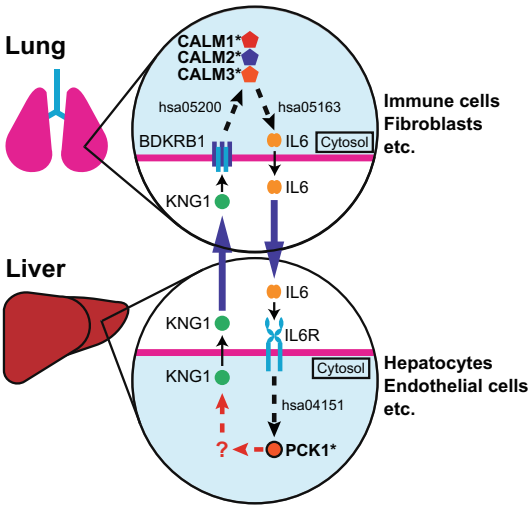


Figure 8. The predicted inter-organ mechanism of IPF. The solid arrows indicate the direct ligand–receptor interactions. The pathway connection (edge) is shown as dashed-arrows indicating the presence of one or more nodes (proteins: ligands, receptors, signaling targets) in between. The corresponding KEGG human pathways for each edge are indicated as hsa numbers. "?" indicates the lack of KEGG pathway connecting the nodes. *BDKRB1* bradykinin receptor B1, *CALM1/2/3* calmodulin 1/2/3, *IL6* interleukin 6, *IL6R* interleukin 6 receptor, *KNG1* kininogen 1, *PCK1* phosphoenolpyruvate carboxykinase 1. KEGG pathways: hsa04151: PI3K-Akt signaling pathway; hsa05163: Human cytomegalovirus infection pathway; hsa05200: Pathways in cancer.

IL6R/IL6ST in the liver. This feedforward and feedback mechanism across the liver and the lung triggers and/or exacerbates IPF pathogenesis.

Discussion

While the results shown in this study are computational, there are mounting pre-clinical and clinical evidence supporting our findings. They are as follows (below).

The latent relatedness of IPF to non-pulmonary diseases

By applying the multi-modal generative topic modeling to the multi-modal disease-omics data of 6,955 human diseases, we identified molecular and genetic relatedness of IPF to non-respiratory diseases such as various types of neoplasm, autoimmune disorders, diabetes, Alzheimer's disease, rheumatoid arthritis, obesity, cardiovascular diseases (atherosclerosis, arteriosclerosis, hypertensive disease, etc.), systemic lupus erythematosus, and multiple sclerosis (Fig. 3). A possible similarity of IPF to lung cancer is discussed in an editorial article[30]. In this article, rhodopsin guanine nucleotide exchange factors (RhoGEF) mediated epithelial cell transformation (ECT) 2 of AT2 cells in the lung could be a common mechanism between IPF and the lung cancer. Our topic modeling study shows the relatedness of IPF to diverse types of cancer (Fig. 3). As AT2 is a relatively specific resident cell-type in the lung, it is unlikely that the same mechanism is the basis of the relatedness of IPF to other non-lung cancers. However, a possibility of the ECT of other types of epithelial cells in non-lung tissues remains, which could explain the relatedness of IPF to other types of cancer as found in our study.

The relatedness of IPF to diabetes is another finding worth discussion. There are several clinical studies including clinical meta analyses suggesting an association between IPF and diabetes[31–34]. Our computational study also indicates molecular relatedness of IPF to diabetes (Fig. 3). Moreover, other IPF-related diseases found in this study such as Alzheimer's disease, obesity, cardiovascular diseases, systemic lupus erythematosus are linked to diabetes[35–39]. Furthermore, the two signaling nodes in the inter-organ mechanism of IPF proposed in our study (Fig. 8), CALM1/CALM2/CALM3 (in the lung) and PCK1 (in the liver), are both molecularly linked to diabetes (Supplementary Table S1). Taken together, IPF may share the same molecular underpinnings with diabetes and other related diseases (i.e., Alzheimer's disease, obesity, cardiovascular diseases, systemic lupus erythematosus).

The inter-organ mechanism of IPF

Our generative computational approach predicts a molecular crosstalk mechanism between the lung and the liver for IPF (Fig. 8). The possibility of the lung-liver interaction in IPF is further supported by a clinical observation of liver fibrosis in some IPF patients[40].

In the lung-liver interaction mechanism that we found, two secreted systemic factors, KNG1 and IL6, bridge the liver-lung crosstalk. Hence, based on this mechanism, the interference of KNG1–BDKRB1(a receptor for

KNG1) and/or IL6-IL6R (a receptor for IL6) interactions could bring a therapeutic benefit to IPF. In this regard, it is worth noting that the blocking IL-6 is shown to attenuate pulmonary fibrosis in mice[41].

In the same study, it is also shown that IPF patients exhibit the increased level of soluble interleukin 6 receptor subunit alpha (sIL-6Rα) in their lung tissues. However, the proposed therapeutic mechanism is the blocking of the intra-pulmonary interactions of sIL-6Rα and IL6. Furthermore, possible roles of interleukins in the pathogenesis of pulmonary fibrosis including IPF are recently discussed[42]. In these other studies, the IL6 inhibitory effects in IPF patients are discussed only in the context of heterologous cell–cell crosstalk within the lung tissue. However, as the IL6R is also present in hepatocytes, hepatic endothelial cells, and intrahepatic cholangiocytes (Fig. 5), such IL6–IL6R inhibitory effect could also occur in the liver. Hence, when the therapeutic inhibition of the IL6–IL6R interaction is effective, it is important to consider a possibility that such effects are also through the inhibition of this ligand–receptor interaction outside the lung tissue such as within the liver tissue.

In our inter-organ mechanism, we also propose that the calmodulin pathway (CALM1/CALM2/CALM3) is activated by the liver-derived KNG1 interaction with the lung BDKRB1, which then induces the IL6 pathway (Fig. 8). It is shown that a calmodulin inhibitor, trifluoperazine, exhibits an anti-inflammatory effect in a bleomycin-induced pulmonary fibrosis animal model[43]. This pre-clinical evidence supports our proposed mechanistic model.

The other signaling node in our inter-organ mechanism is PCK1 (Fig. 8). In our model, PCK1 pathway is activated by the IL6-IL6R interaction in the liver, which is feeds back to the lung pathogenesis of IPF via KNG1-BDKRB1 pathway. Recently, nintedanib, one of the two FDA-approved IPF therapeutics, is shown to attenuate experimental colitis via inhibiting the PCK1 pathway[44]. This study suggests that a part of the therapeutic effect of nintedanib on IPF is via the inhibition of the PCK1 pathway.

There are two pending questions in the proposed inter-organ mechanism of IPF. The ligands, receptors, and their signaling targets in this model are co-expressed in multiple cell types in their corresponding organs (Fig. 8). Hence, it remains unknown whether the KNG1–BDKRB1 and IL6–IL6R/IL6ST pathways function within the same cell-type or they interact in *trans* across different cell-types within the same organ.

Another question is whether the IL6–IL6R/IL6ST signal feeds back to KNG1 via PCK1 (Fig. 8). While the signaling of IL6–IL6R/IL6ST to PCK1 is established (hsa04151 KEGG pathway, PI3K-Akt signaling pathway in human), the link of PCK1 to KNG1 remains unknown. Upon the experimental validation of this link, the model becomes a closed feedforward and feedback "loop" across the liver and the lung.

These questions remain for the future studies and their results provide more detailed mechanistic description of IPF. Furthermore, they facilitate the designing of first-in-class therapeutic and/or diagnostic strategies for IPF.

In this study, we exploited a growing body of multi-modal disease-omics data and a generative computational power to predict an inter-organ mechanism of IPF with the molecular and cellular resolution. Furthermore, our retrospective reference-mining found multiple experimental and clinical evidence in support of the predicted mechanism as described above. Our proposed mechanism is detailed enough, providing a unique opportunity to design hypothesis-driven pre-clinical experiments and/or clinical studies to discover and evaluate first-in-class therapeutic and diagnostic targets for IPF. In addition, our study and results illustrate a computational framework to generate experimentally-testable mechanistic models for other diseases where very little mechanism is known.

Methods

Multi-modal generative topic modeling

The multi-modal generative topic modeling approach is as previously described[29]. This topic modeling approach exploits the similarities among diseases on the basis of their multi-modal omics features. In this study, we deleted IPF disease-omics data to identify latent IPF-features (see also Fig. 1 and the details in "Result" section).

Organ- and cell-type expression patterns of the latent IPF-features

The organs and cells where the IPF-features are expressed were identified by organ/cell enrichment analyses using THE HUMAN PROTEIN ATLAS v 21.1[23–25], as previously described[29]. Briefly, we generated a 2×2 contingency table showing the number of the genes of interest that are associated with the target organ(s)/cell(s), and performed chi-square test of independence by using the contingency table.

Latent relatedness of IPF to other diseases

We identified diseases to which IPF is related as previously described[29]. Briefly, the disease-labels of each latent IPF-feature were identified in the Dataset A (Fig. 1) and the frequency of each disease-label was counted. The disease-labels of the higher-frequency are determined as more related to IPF.

Lung RNA-seq data from patients

The studies with human subjects and data were approved by the Institutional Review Board of Advanced Telecommunications Research Institute International on behalf of Karydo TherapeutiX, Inc. (Approved Number: HK2101-2101, HK2101-2103, HK2101-2202) and of National Institutes of Biomedical Innovation, Health and Nutrition (Approved Number: 187) and of Kanagawa Cardiovascular and Respiratory Center (Approved Number: KCRC-19-0015). The informed consent was obtained from all subjects. All methods were performed in accordance with the relevant guidelines and regulations. The lung tissues were collected from 299 subjects. They consist of 173 idiopathic interstitial pneumonias (IIPs), 76 hypersensitivity pneumonitis (HP), 26 connective tissue diseases (CTD), 24 others (other interstitial lung diseases). The 173 IIPs are further composed of 95 IPF, 41 unclassifiable interstitial pneumonia (UCIP), 28 idiopathic nonspecific interstitial pneumonia (NSIP), 3 idiopathic pleuroparenchymal fibroelastosis (PPFE), and 6 other IIPs. RNA was purified from each sample and processed for RNA sequencing as follows: The lung tissues were sent to TAKARA BIO INC. (Shiga, Japan) for

sequencing. At TAKARA BIO INC., a total RNA was purified using NucleoSpin®RNA according to the provided protocol. Before RNA-sequencing, the total RNA for each specimen was checked for quality using Agilent 2000 TapeStation (Agilent Technologies, Santa Clara, CA, USA). The RNA Integrity Numbers (RINs) for all the specimens, as obtained by TapeStation, passed a score of 6.0 or greater. Upon this quality check, mRNA sequencing was performed using the Illumina Sequencer NovaSeq6000 with paired end reads of 150 bps. Read sequences obtained were mapped to genome sequences. Based on the positional information obtained from the mapping and the gene definition file, the gene units were mapped to the genome sequence. The expression level of each gene and transcript was calculated based on the positional information obtained by mapping and the gene definition file and an annotation information was added. The differentially expressed genes (DEgenes) in the lung tissues of IPF vs. all the other pulmonary diseases (UCIP, NSIP, PPFE, other IIPs (labeled as "IIP" in the raw count data), HP, CTD, other interstitial lung diseases (labeled as "Others" in the raw count data) were detected by using an R package, *DESeq2*[45], with the default parameter settings. The DEgenes were defined as the genes whose absolute values of log2FoldChange are ≥ 1 and also adjusted p-values are < 0.05. These IPF-DEgenes are referred to as "IPF-DEgenes" in this paper.

Identification of ligand–receptor pairs

From the IPF-DEgenes, those encoding ligand proteins or receptor proteins were identified using a human ligand–receptor combination database, CellChatDB[46]. Furthermore, we identified their corresponding receptor and ligand partners using the same database.

Generation of an inter-organ map of IPF

The overall design of the approach is described in Fig. 4 (see the "Results" section for its narrative description). The gene expression analyses in the organs and cells were performed as follows:

The single-cell gene expression across multiple healthy organs and that of the lung tissues of IPF patients were determined using two publicly available human single-cell RNA sequencing (scRNA-seq) datasets: (1) *Tabula Sapiens*, which is a database of multiple healthy tissues[22], and (2) GSE122960, which is a scRNA-seq dataset of the lung tissues of IPF patients/healthy donors[47]. For the *Tabula Sapiens*, the raw count data and the cell type annotation table were extracted using a Python package, *scanpy*[48], from the h5ad-formatted data at their FigShare deposit[49]. For the GSE122960, the raw count data were obtained from the Gene Expression Omnibus (GEO) deposit. However, the cell type annotation table for GSE122960 was not available. Therefore, we reproduced the cell type annotation from the downloaded count data using publicly available R program codes (https://github.com/NUPulmonary/Reyfman2018/tree/master)[47], which accompany the GSE122960 study. These codes were run in the *Seurat* package (v2.3.4)[50] as follows: (1) Cell quality control filtration by the number of genes and the percent of mitochondrial RNA; (2) Log-normalization with the scale factor 10,000; (3) Dection of highly variable genes; (4) Scaling for all genes; (5) Principal component analysis (PCA) and the elbow plot to determine the PCs used for the downstream analyses; (6) Cell clustering and the cell type determination by the marker genes. The parameter settings at each step and the cell type annotation were performed according to the obtained codes per each sample data. The DEgenes in each cell type of the lung derived from the IPF patients were determined by comparing the scRNA-seq data of the lung tissues derived from the IPF patients and the healthy-donors by *DESeq2*[45] following the developers' recommendations for single-cell analysis[51]. Briefly, we first set size-factors by 'computeSumFactors()' in the *scran* package[52]. The DESeq2 was performed by using the likelihood ratio test as significance testing, where we set the 'DESeq()' arguments to the following values: test = 'LRT', useT = TRUE, minmu = 1e−6, minReplicateForReplace = Inf. The genes were evaluated by the statistical significance level at 0.05 in adjusted p-value.

The KEGG-mining (steps 5 and 6 in the flow) was conducted as follows:

The KEGG-mining was performed to identify downstream and upstream targets of the ligand–receptor pairs and to determine which of the targets are the IPF-features. First, the gene symbols were converted into KEGG IDs, by first to Entrez IDs using the R function 'bitr()' of the R package *clusterProfiler*[53], and then to KEGG IDs from the Entrez IDs using KEGG API[54]. Next the KEGG pathways containing these KEGG IDs were extracted and their directed graphs were constructed using KGML[55]. In the graphs, each node was the attribute 'name' of the tag <entry>, and each edge started at the node corresponding to the attribute 'entry1' of the tag <relation> and ended at the node corresponding to the attribute 'entry2' of the same tag <relation>. Using these graphs, we identified the direct and indirect connections between the ligands/receptor and the latent IPF-features.

Ethics approval and consent to participate

The studies with human subjects and data were approved by the Institutional Review Board of Advanced Telecommunications Research Institute International on behalf of Karydo TherapeutiX, Inc. (Approved Number: HK2101-2101, HK2101-2103, HK2101-2202) and of National Institutes of Biomedical Innovation, Health and Nutrition (Approved Number: 187) and of Kanagawa Cardiovascular and Respiratory Center (Approved Number: KCRC-19-0015).

Data availability

The datasets generated and/or analyzed during the current study are available in GEO, GSE122960; figshare repository, https://figshare.com/projects/Tabula_Sapiens/100973; GitHub, https://github.com/skozawa170301ktx/IPF.

Code availability

All raw data of the results are available as supplementary tables of this paper and they are referenced within the manuscript accordingly. The code reported in this paper is available at: https://github.com/skozawa170301ktx/MultiModalDiseaseModeling.

Received: 28 April 2023; Accepted: 6 December 2023
Published online: 11 December 2023

References

1. Lederer, D. J. & Martinez, F. J. Idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **379**, 797–798. https://doi.org/10.1056/NEJMc1807508 (2018).
2. Martinez, F. J. *et al.* Idiopathic pulmonary fibrosis. *Nat. Rev. Dis. Primers* **3**, 17074. https://doi.org/10.1038/nrdp.2017.74 (2017).
3. Glass, D. S. *et al.* Idiopathic pulmonary fibrosis: Current and future treatment. *Clin. Respir. J.* **16**, 84–96. https://doi.org/10.1111/crj.13466 (2022).
4. Mei, Q., Liu, Z., Zuo, H., Yang, Z. & Qu, J. Idiopathic pulmonary fibrosis: An update on pathogenesis. *Front. Pharmacol.* **12**, 797292. https://doi.org/10.3389/fphar.2021.797292 (2021).
5. Barratt, S. L., Creamer, A., Hayton, C. & Chaudhuri, N. Idiopathic pulmonary fibrosis (IPF): An overview. *J. Clin. Med.* https://doi.org/10.3390/jcm7080201 (2018).
6. Yanagihara, T., Sato, S., Upagupta, C. & Kolb, M. What have we learned from basic science studies on idiopathic pulmonary fibrosis?. *Eur. Respir. Rev.* https://doi.org/10.1183/16000617.0029-2019 (2019).
7. Baracos, V. E., Martin, L., Korc, M., Guttridge, D. C. & Fearon, K. C. H. Cancer-associated cachexia. *Nat. Rev. Dis. Primers* **4**, 17105. https://doi.org/10.1038/nrdp.2017.105 (2018).
8. Bin, N. R. *et al.* An airway-to-brain sensory pathway mediates influenza-induced sickness. *Nature* https://doi.org/10.1038/s41586-023-05796-0 (2023).
9. Herrlich, A., Kefaloyianni, E. & Rose-John, S. Mechanisms of interorgan crosstalk in health and disease. *FEBS Lett.* **596**, 529–533. https://doi.org/10.1002/1873-3468.14313 (2022).
10. Osterhout, J. A. *et al.* A preoptic neuronal population controls fever and appetite during sickness. *Nature* **606**, 937–944. https://doi.org/10.1038/s41586-022-04793-z (2022).
11. Priest, C. & Tontonoz, P. Inter-organ cross-talk in metabolic syndrome. *Nat. Metab.* **1**, 1177–1188. https://doi.org/10.1038/s42255-019-0145-5 (2019).
12. Castelo-Branco, C. & Soveral, I. The immune system and aging: A review. *Gynecol. Endocrinol.* **30**, 16–22. https://doi.org/10.3109/09513590.2013.852531 (2014).
13. Hiam-Galvez, K. J., Allen, B. M. & Spitzer, M. H. Systemic immunity in cancer. *Nat. Rev. Cancer* **21**, 345–359. https://doi.org/10.1038/s41568-021-00347-z (2021).
14. Hotamisligil, G. S. Inflammation, metaflammation and immunometabolic disorders. *Nature* **542**, 177–185. https://doi.org/10.1038/nature21363 (2017).
15. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res.* **30**, 492–506. https://doi.org/10.1038/s41422-020-0332-7 (2020).
16. Rojas, M. *et al.* Metabolic syndrome: Is it time to add the central nervous system?. *Nutrients* https://doi.org/10.3390/nu13072254 (2021).
17. Pataky, M. W., Young, W. F. & Nair, K. S. Hormonal and metabolic changes of aging and the influence of lifestyle modifications. *Mayo Clin. Proc.* **96**, 788–814. https://doi.org/10.1016/j.mayocp.2020.07.033 (2021).
18. Kalluri, R. & LeBleu, V. S. The biology, function, and biomedical applications of exosomes. *Science* **367**, 6977. https://doi.org/10.1126/science.aau6977 (2020).
19. Pinero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839. https://doi.org/10.1093/nar/gkw943 (2017).
20. Regev, A. *et al.* The human cell atlas. *Elife* https://doi.org/10.7554/eLife.27041 (2017).
21. Regev, A. *et al.* Science forum: The human cell atlas. *Elife* https://doi.org/10.7554/eLife.27041 (2017).
22. Tabula Sapiens, C. *et al.* The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, 4896. https://doi.org/10.1126/science.abl4896 (2022).
23. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, 3321. https://doi.org/10.1126/science.aal3321 (2017).
24. Uhlen, M. *et al.* Proteomics: Tissue-based map of the human proteome. *Science* **347**, 1260419. https://doi.org/10.1126/science.1260419 (2015).
25. *The Human Protein Altas v.21.1*. https://www.proteinatlas.org.
26. Consortium, T. S. *Tabula Sapiens*. https://tabula-sapiens-portal.ds.czbiohub.org.
27. Group, I. B. I. *DisGeNET v7.0*. https://www.disgenet.org.
28. Dilmegani, C. *Generative AI in Life Sciences: Use Cases & Examples in 2023*, https://research.aimultiple.com/generative-ai-in-life-sciences/. (2023).
29. Kozawa, S. *et al.* Latent disease similarities and therapeutic repurposing possibilities uncovered by multi-modal generative topic modeling of human diseases. *Bioinform. Adv.* https://doi.org/10.1093/bioadv/vbad047 (2023).
30. Reyfman, P. A. & Gottardi, C. J. Idiopathic pulmonary fibrosis and lung cancer: Finding similarities within differences. *Am. J. Respir. Cell Mol. Biol.* **61**, 667–668. https://doi.org/10.1165/rcmb.2019-0172ED (2019).
31. Bai, L. *et al.* Idiopathic pulmonary fibrosis and diabetes mellitus: A meta-analysis and systematic review. *Respir. Res.* **22**, 175. https://doi.org/10.1186/s12931-021-01760-6 (2021).
32. Enomoto, T., Usuki, J., Azuma, A., Nakagawa, T. & Kudoh, S. Diabetes mellitus may increase risk for idiopathic pulmonary fibrosis. *Chest* **123**, 2007–2011. https://doi.org/10.1378/chest.123.6.2007 (2003).
33. Li, C. *et al.* Associations between diabetes and idiopathic pulmonary fibrosis: A study-level pooled analysis of 26 million people. *J. Clin. Endocrinol. Metab.* **106**, 3367–3380. https://doi.org/10.1210/clinem/dgab553 (2021).
34. Wang, D., Ma, Y., Tong, X., Zhang, Y. & Fan, H. Diabetes mellitus contributes to idiopathic pulmonary fibrosis: A review from clinical appearance to possible pathogenesis. *Front. Public Health* **8**, 196. https://doi.org/10.3389/fpubh.2020.00196 (2020).
35. Di Pietrantonio, N., Di Tomo, P., Mandatori, D., Formoso, G. & Pandolfi, A. Diabetes and its cardiovascular complications: Potential role of the acetyltransferase p300. *Cells* https://doi.org/10.3390/cells12030431 (2023).
36. Hamze, R. *et al.* Type 2 diabetes mellitus and alzheimer's disease: Shared molecular mechanisms and potential common therapeutic targets. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms232315287 (2022).
37. Hernandez-Negrin, H. *et al.* Obesity, diabetes, and cardiovascular risk burden in systemic lupus erythematosus: Current approaches and knowledge gaps-a rapid scoping review. *Int. J. Environ. Res. Public Health* https://doi.org/10.3390/ijerph192214768 (2022).
38. Kee, O. T. *et al.* Cardiovascular complications in a diabetes prediction model using machine learning: A systematic review. *Cardiovasc. Diabetol.* **22**, 13. https://doi.org/10.1186/s12933-023-01741-7 (2023).

39. Rosengren, A. & Dikaiou, P. Cardiovascular outcomes in type 1 and type 2 diabetes. *Diabetologia* **66**, 425–437. https://doi.org/10.1007/s00125-022-05857-5 (2023).
40. Cocconcelli, E. *et al.* Subclinical liver fibrosis in patients with idiopathic pulmonary fibrosis. *Intern. Emerg. Med.* **16**, 349–357. https://doi.org/10.1007/s11739-020-02376-2 (2021).
41. Le, T. T. *et al.* Blockade of IL-6 Trans signaling attenuates pulmonary fibrosis. *J. Immunol.* **193**, 3755–3768. https://doi.org/10.4049/jimmunol.1302470 (2014).
42. She, Y. X., Yu, Q. Y. & Tang, X. X. Role of interleukins in the pathogenesis of pulmonary fibrosis. *Cell Death Discov.* **7**, 52. https://doi.org/10.1038/s41420-021-00437-9 (2021).
43. Nakashima, J. M., Hyde, D. M. & Giri, S. N. Effects of a calmodulin inhibitor on bleomycin-induced lung inflammation in hamsters: Biochemical, morphometric, and bronchoalveolar lavage data. *Am. J. Pathol.* **124**, 528–536 (1986).
44. Li, H. *et al.* Nintedanib alleviates experimental colitis by inhibiting CEBPB/PCK1 and CEBPB/EFNA1 pathways. *Front. Pharmacol.* **13**, 904420. https://doi.org/10.3389/fphar.2022.904420 (2022).
45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. https://doi.org/10.1186/s13059-014-0550-8 (2014).
46. Jin, S. *et al.* Inference and analysis of cell-cell communication using Cell Chat. *Nat. Commun.* **12**, 1088. https://doi.org/10.1038/s41467-021-21246-9 (2021).
47. Reyfman, P. A. *et al.* Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536. https://doi.org/10.1164/rccm.201712-2410OC (2019).
48. Wolf, A., Ramirez, F. & Rybakov, S. *Scanpy Tutorials*. https://scanpy-tutorials.readthedocs.io/en/latest/index.html.
49. Pisco, A. *Figshare*. https://figshare.com/projects/Tabula_Sapiens/100973.
50. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420. https://doi.org/10.1038/NBT.4096 (2018).
51. Love, M. I., Anders, S. & Huber, W. *Analysing RNA-seq data with DESeq2*. http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html.
52. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000 Res.* **5**, 2122. https://doi.org/10.12688/f1000research.9501.2 (2016).
53. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics* **16**, 284–287. https://doi.org/10.1089/omi.2011.0118 (2012).
54. *KEGG API*. https://rest.kegg.jp/conv/hsa/ncbi-geneid.
55. Laboratories, K. *KGML (KEGG Markup Language)*. https://www.kegg.jp/kegg/xml/.

Acknowledgements

T.N.S., S.K., K.T., and S.T. thank K. Sugisaka, R. Takahashi, R. Ishikawa for their administrative assistance. T.N.S., S.K., K.T., and S.T. are also grateful to the members of Karydo TherapeutiX, Inc. for their supports, advice, and discussion throughout the course of this work. Y.N.-K., T.O., T.N., H.K., M.N.-I., and M.K. thank Yoichi Kurebayashi (Kobe University, Kobe, Japan), Naonori Ueda (RIKEN AIP, Kyoto, Japan), Masanori Shindo (NIBIOHN, Osaka, Japan), Yoshinori Nakamura (NIBIOHN, Osaka, Japan), Hiromitsu Kageyama (NIBIOHN, Osaka, Japan), Michiyo Kawai (NIBIOHN, Osaka, Japan), Akiko Fukagawa (NIBIOHN, Osaka, Japan), Hideyo Kamada (NIBIOHN, Osaka, Japan) for supporting PRISM project.

Author contributions

T.N.S. conceived the idea of the project, designed the study, and supervised the overall research project. S.K. designed and performed the multi-modal generative topic modeling analyses. K.T. performed the CellChat DB analyses and gene expression analysis. S.T. performed the KEGG-mining. Y.N.-K., T.O., H.K., T.N. designed the clinical sampling protocol, T.O., H.K., T.N. performed the clinical sampling, Y.N.-K. conducted RNA purification and sequencing, M.N.-I., M.K. performed RNA-seq data preparation. T.N.S., S.K., K.T., S.T., Y.N.-K. wrote the manuscript. All authors reviewed manuscript and approved the final manuscript.

Funding

This work was supported in part by MHLW Health, Labour and Welfare Sciences Research Grants Program Grant Number JPMH20AC5001(Y.N.-K.), Cabinet Office of Japan Government for the Public/Private R&D Investment Strategic Expansion Program (PRISM) (T.N.S., Y.N.-K., T.O.), Innovative Science and Technology Initiative for Security Grant Number JPJ004596 ATLA Japan (T.N.S.), and Nakatani Foundation (T.N.S.).

Competing interests

T.N.S., S.K., K.T., S.T. are employees of Karydo TherapeutiX, Inc. No other author possesses competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-49281-0.

Correspondence and requests for materials should be addressed to T.N.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

4) bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Cell type-specific weighting-factors to solve solid organs-specific limitations of single cell RNA-sequencing

Short title: Cell type-specific weighting-factors for solid organ scRNA-seq

Kengo Tejima[¶], Satoshi Kozawa[¶], Thomas N. Sato[*]

Karydo TherapeutiX, Inc., Kyoto, Japan; ERATO Sato Live Bio-Forecasting Project, Kyoto, Japan; The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan; V-iClinix Laboratory, Nara Medical University, Nara, Japan

[*]Corresponding Authors:

Thomas N. Sato

Karydo TherapeutiX, Inc.

2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan 619-0288

Email: island1005@gmail.com

[¶]These two authors contributed equally to this work

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

While single-cell RNA-sequencing (scRNA-seq) is a popular method to analyze gene expression and cellular composition at single-cell resolution, it harbors shortcomings: The failure to account for cell-to-cell variations of transcriptome-size (i.e., the total number of transcripts per cell) and also cell dissociation/processing-induced cryptic gene expression. This is particularly a problem when analyzing highly heterogeneous solid tissues/organs, which requires cell dissociation for the analysis. As a result, there exists a discrepancy between bulk RNA-seq result and virtually reconstituted bulk RNA-seq result using its composite scRNA-seq data. To fix this problem, we propose a computationally calculated coefficient, "cell type-specific weighting-factor (cWF)". Here, we introduce a concept and a method of its computation and report cWFs for 76 cell-types across 10 solid organs. Their fidelity is validated by more accurate reconstitution and deconvolution of bulk RNA-seq data of diverse solid organs using the scRNA-seq data and the cWFs of their composite cells. Furthermore, we also show that cWFs effectively predict aging-progression, implicating their diagnostic applications and also their association with aging mechanism. Our study provides an important method to solve critical limitations of scRNA-seq analysis of complex solid tissues/organs. Furthermore, our findings suggest a diagnostic utility and biological significance of cWFs.

Abbreviations: cWF, cell type-specific weighting factor; r-woRNAseq, real whole-organ RNA-seq; syn-woRNAseq, synthetic whole-organ RNA-seq; V-scRNAseq, virtual single cell RNA-seq.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Since the advent of single-cell RNA-sequencing (scRNA-seq) technologies early this century, this method is becoming one of the essential tools in biomedical fields [1]. While scRNA-seq provides useful information of cellular composition and gene expression at single cell resolution, it has limitations.

It is well-known that the total number of RNA molecules per cell, also known as transcriptome-size, varies from-cell-to-cell [2-4]. However, in scRNA-seq analyses, the transcripts are sequenced in individual cells and their transcript-counts are normalized, allowing the comparison of the relative abundance of each transcript-species across individual cells/cell-types [5-9]. This internal normalization process of the analyses cancels out the putative transcriptome size-variations among different cells, concealing potentially important functional differences among the cells.

This differential transcriptome-size can be determined by counting the library depth of the cells and/or by including normalizing spike-in RNA in the samples [2, 3, 10, 11]. Alternative, it can be quantified by PCR-based methods [12, 13]. However, these approaches are applicable only to culture cells, blood/immune cells, and cancer cells which are relatively uniform cell types and do not require cell dissociation. In contrast, they are not applicable to solid tissues/organs composed of highly heterogeneous cell types. Individual cells can be dissociated from solid tissues/organs and then analyzed by scRNA-seq. However, the cell-dissociation and subsequent harvesting and/or purification steps induce cryptic gene expression [14]. Due to such limitations, it is difficult to measure and compare the absolute transcript-counts of each gene in individual cells of the complex solid tissues/organs such as the heart, brain, kidney, etc.

Furthermore, the differential transcriptome-size has biological importance. Several studies indicate that the cell-size influences the overall cellular transcription [15-18]. It is also

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reported that c-myc amplifies the global transcription in tumor cells [2, 19], lymphocytes and embryonic stem cells [20]. Another study shows that the global transcription is repressed by the loss of MECP2, a methyl CpG binding protein, in embryonic stem cell-derived neurons [21]. The genome structure, such as ploidy/gene-dosage and nucleosome state, affects the overall transcription [12, 22, 23]. In additions, it is shown that the total mRNA-contents of tumor cells are associated with the extent of the cancer progression [24]. These results illustrate biological importance of accounting for the transcriptome size differences among cells.

In addition to the transcriptome-size, cell dissociation/processing-induced cryptic gene expression is another problem with scRNA-seq, in particular with solid tissues/organs [14]. The scRNA-seq analyses of the solid tissues and organs require enzymatic digesting and mechanical dissociation of individual cells. Spatial scRNA-seq requires tissue processing such as sectioning and laser-mediated harvesting [8, 9]. Such technical procedures induce cryptic gene expression [14]. The currently available scRNA-seq largely disregard this problem.

In this study, we address and overcome these limitations/problems associated with solid tissues/organs by developing a method to compute coefficients, cell type-specific weighting-factors (cWFs), to offset the transcriptome size- and cell dissociation/processing-induced problem. Furthermore, we also calculate cWFs for 76 cell-types across 10 solid organs and use them to illustrate their diagnostic and biological utilities.

Results

The cell type-specific weighting factor (cWF)

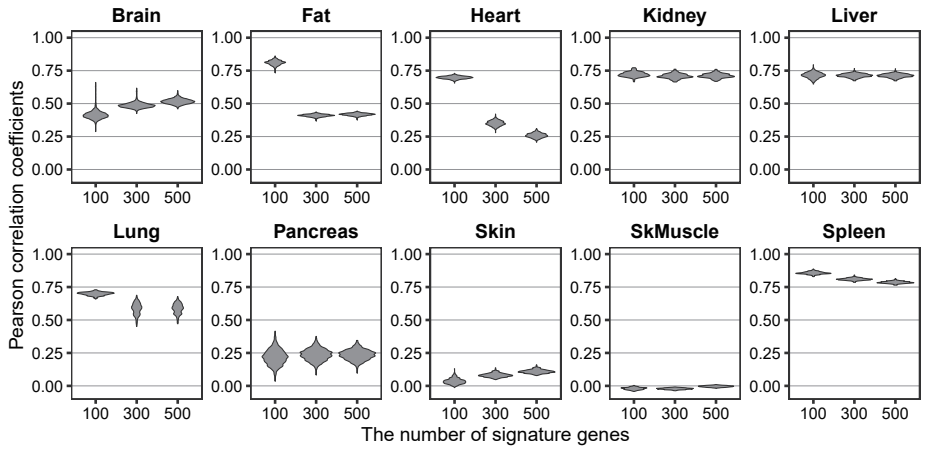
First, the major premise is follows:

If the transcriptome size differences and the cell dissociation/processing-induced cryptic gene expression (abbreviated as “cryptic gene expression” hereafter for convenience) are corrected

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

93 in the conventional scRNA-seq, the bulk (i.e., tissue/organ) RNA-seq should be the sum of
94 gene expression of each composite cell-type weighted by cell-type ratios.

95 We evaluated this premise. The synthetic whole-organ data were constructed from
96 single-cell transcript counts combined according to the experimentally determined fraction of
97 each cell-type in the organ (see Materials and methods). Briefly, the synthetic whole-organ
98 RNA-seq data are calculated as the sum of the normalized transcript counts weighted by the
99 known ratios of the composite cell-types of the organs, using the signature genes for
100 distinguishing the cell-types for each organ. The signature genes were determined by the
101 random forest classifier. These synthetic data for 10 organs (brain, fat, heart, kidney, liver, lung,
102 pancreas, skin, skeletal muscle, spleen) were then compared to the corresponding real whole-
103 organ RNA-seq data (Fig 1 and S1 Table). The comparisons were performed by calculating
104 their Pearson correlation coefficients.



105

106 **Fig 1. Incomplete reconstitution of the whole organ RNA-seq by the composite scRNA-seq.** The
107 similarity is shown as violin plot of the Pearson correlation coefficient for each number of the signature
108 genes (100, 300, 500) for each organ (indicated above each plot). The corresponding raw data are found in
109 S1 Table. SkMuscle: skeletal muscle.

110

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

111 The result shows low Pearson correlation coefficients (< 0.75) for most of the organs
112 (brain, heart, kidney, liver, lung, pancreas, skin, skeletal muscle) regardless of the number (100,
113 300, 500) of the top-ranked signature genes used. Particularly low coefficients are found with
114 the brain (< 0.6 for 100, 300, 500 signature genes), the fat (< 0.5 for 300, 500 signature genes),
115 the heart (< 0.4 for 300, 500 signature genes), the pancreas (< 0.3 for 100, 300, 500 signature
116 genes), the skin (< 0.2 for 100, 300, 500 signature genes), and the skeletal muscle (< 0.1 for
117 100, 300, 500 signature genes). These low Pearson correlation coefficients suggest that the
118 transcriptome-size of individual cell-types varies, resulting in the large gaps.
119 These results fail to satisfy the above-described major premise; hence, the differential
120 transcriptome size-variations and/or the cryptic gene expression are unrepresented in the
121 scRNA-seq data of the solid organs. To solve this problem, we developed an algorithm to
122 compute such unrepresented factors (Fig 2 and see Materials and methods). The concept of the
123 algorithm is schematically described in Fig 2 and as follows: In the real scRNA-seq data
124 analyses, the total transcript counts per cell (i.e., the transcriptome-size) of the real scRNA-seq
125 data are normalized (i.e., the total counts per cell are equal among all composite cell-types of
126 the organs). This results in the loss of a factor representing the putative differences of the
127 transcriptome-size and/or the cryptic gene expression across cell-types. This factor is computed
128 as the cell type-specific weighting-factor (cWF) (indicated as w_1, w_2, w_3, etc. in the Fig 2).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

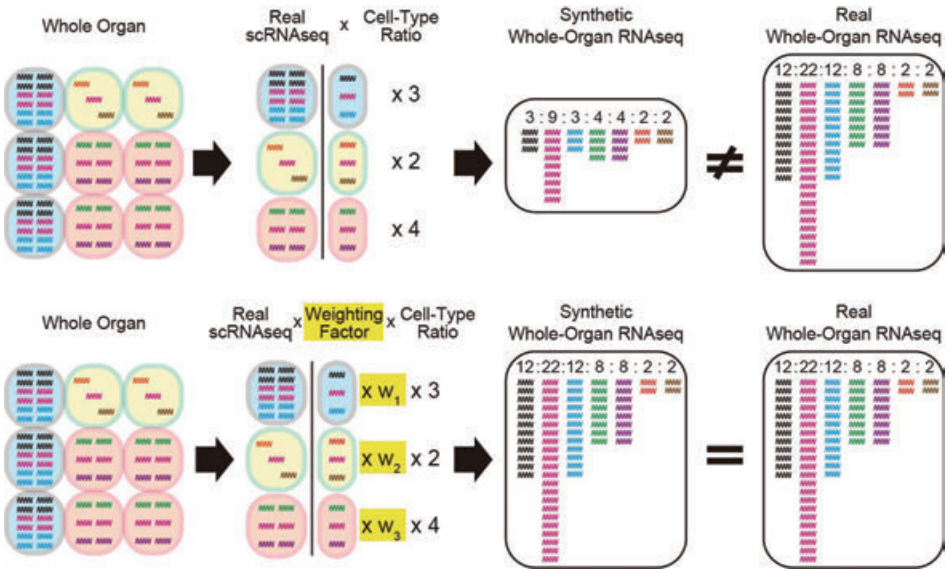


Fig 2. Graphical description of the overall concept of the cell type-specific weighting factors (cWFs). The top row describes the case without the cWFs, failing to reconstitute the real whole-organ RNA-seq data by the synthetic whole-organ RNA-seq data generated by assembling the composite scRNA-seq data without the cWFs. The bottom describes the case with the cWFs, accurately reconstitute the real whole-organ RNA-seq data by the synthetic whole-organ RNA-seq data generated by assembling the composite scRNA-seq data with the cWFs (w_1, w_2, w_3). The wiggling lines are cellular transcripts where different gene transcripts are in different colors. The ratios of the cellular transcript-counts in the synthetic and real whole-organ RNA-seq are indicated by the numbers at the top of the colored transcripts.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

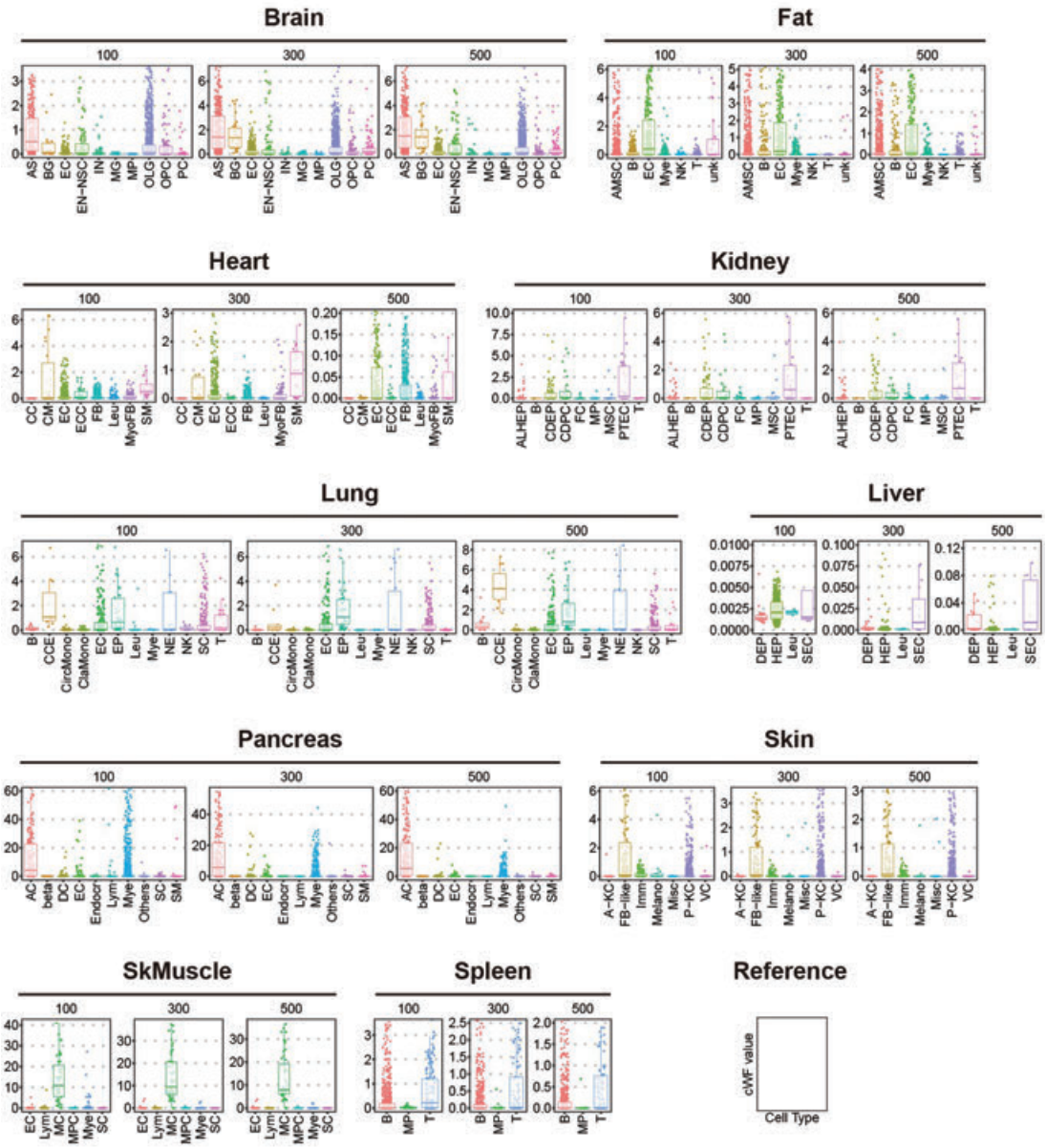


Fig 3. The cWFs for each cell-type in each organ. The values of the cWFs for each cell-type (bottom of each plot) for each number of the signature genes (indicated above each plot) for each organ (indicated at the top of the corresponding plots) are shown as box plots. The corresponding raw data are found in S2 Table. The cell-types are indicated as abbreviations and their full names are found in S2 Table. SkMuscle: skeletal muscle.

Using this algorithm, we computed cWFs for 76 cell-types across 10 organs (Fig 3 and S2 Table). The result shows significant variations of the cWFs among different cell-types in

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

147 many of the organs, uncovering the body-wide degree of such variations. In particular, the
148 following cell-types show significantly higher cWFs than the others within the same organ:
149 astrocyte (AS) in the brain, endothelial cell (EC) in the fat, muscle cell-types in the heart and
150 the skeletal muscle (CM in the heart, MC in the skeletal muscle), epithelial cell of proximal
151 tubule (PTEC) in the kidney, epithelial cell (EP) and neuroendocrine cell (NE) in the lung,
152 fibroblast-like cell (FB-like) in the skin, and T-cell (T) in the spleen. This result suggests that
153 these cell-types contain larger amounts of transcripts (i.e., the larger transcriptome-size) than
154 the others in the corresponding organs. Furthermore, the results show that such variable degrees
155 of differences in the transcriptome-size are widespread across diverse cell-types and organs.
156

157 # Validation of cWFs

158 We validate the computed cWFs by three methods: Reconstitution (Fig 4), deconvolution (Figs
159 5 and 6) of the whole-organ RNAseq data using their composite scRNA-seq data, and the
160 comparison of the cWFs to the experimentally measured transcript-contents (Fig 7).

161 First, we validated the cWFs by the reconstitution (Fig 4 and S3 Table, and see
162 Materials and methods). Using the mean value of cWFs for each cell-type, we weighted the
163 transcript counts for each cell-type. We then summed these weighted transcript counts for all
164 composite cell-types of each organ according to their ratios to construct the synthetic whole-
165 organ RNA-seq (syn-woRNAseq(+cWF)), data. Then, they are compared to the corresponding
166 real whole-organ RNA-seq (r-woRNAseq) data. The result shows the Pearson correlation
167 coefficients of 0.8 – 1.0 between the syn-woRNAseq(+cWF) and r-woRNAseq for all 10
168 organs with all or at least one of the 100, 300, 500 signature genes (Fig 4). These are significant
169 improvements from the 0 – 0.75 of the Pearson correlation coefficients without the cWFs (Fig
170 1). The heart shows the Pearson correlation coefficients of almost 1.0 with the cWFs (Fig 4),
171 as compared to those of 0.25 – 0.7 without the cWFs (Fig 1). Even for the spleen where the

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pearson correlation coefficients are 0.75 – 0.8 without the cWF (Fig 1), the cWF (i.e., syn-woRNAseq(+cWF)) improves the Pearson correlation coefficients up to 0.9 – 1.0 (Fig 4). The results demonstrate that cWFs correct the large gap between the synthetic whole-organ RNAseq without the cWFs (syn-woRNAseq(-cWF)) and r-woRNAseq, supporting the notion that cWFs reflect the transcriptome size-variations of the cells and their requirement for the accurate single-cell analyses.

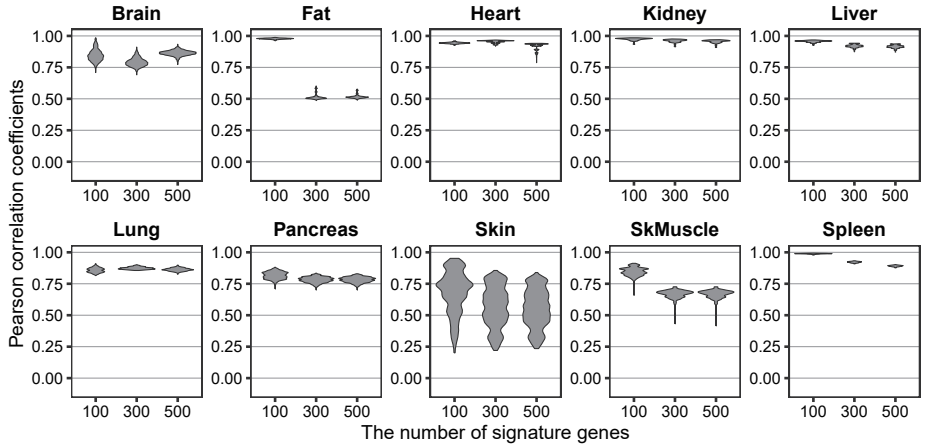


Fig 4. Accurate reconstitution of the whole organ RNA-seq by the composite scRNA-seq and cWFs. The similarity is shown as violin plot of the Pearson correlation coefficient for each number of the signature genes (100, 300, 500) for each organ (indicated above each plot). The corresponding raw data are found in S3 Table. SkMuscle: skeletal muscle.

Next, the validation by the deconvolution (Fig 5) was performed. We developed a deconvolution method that integrates cWFs in a Bayesian framework, as schematically described in Fig 5A.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

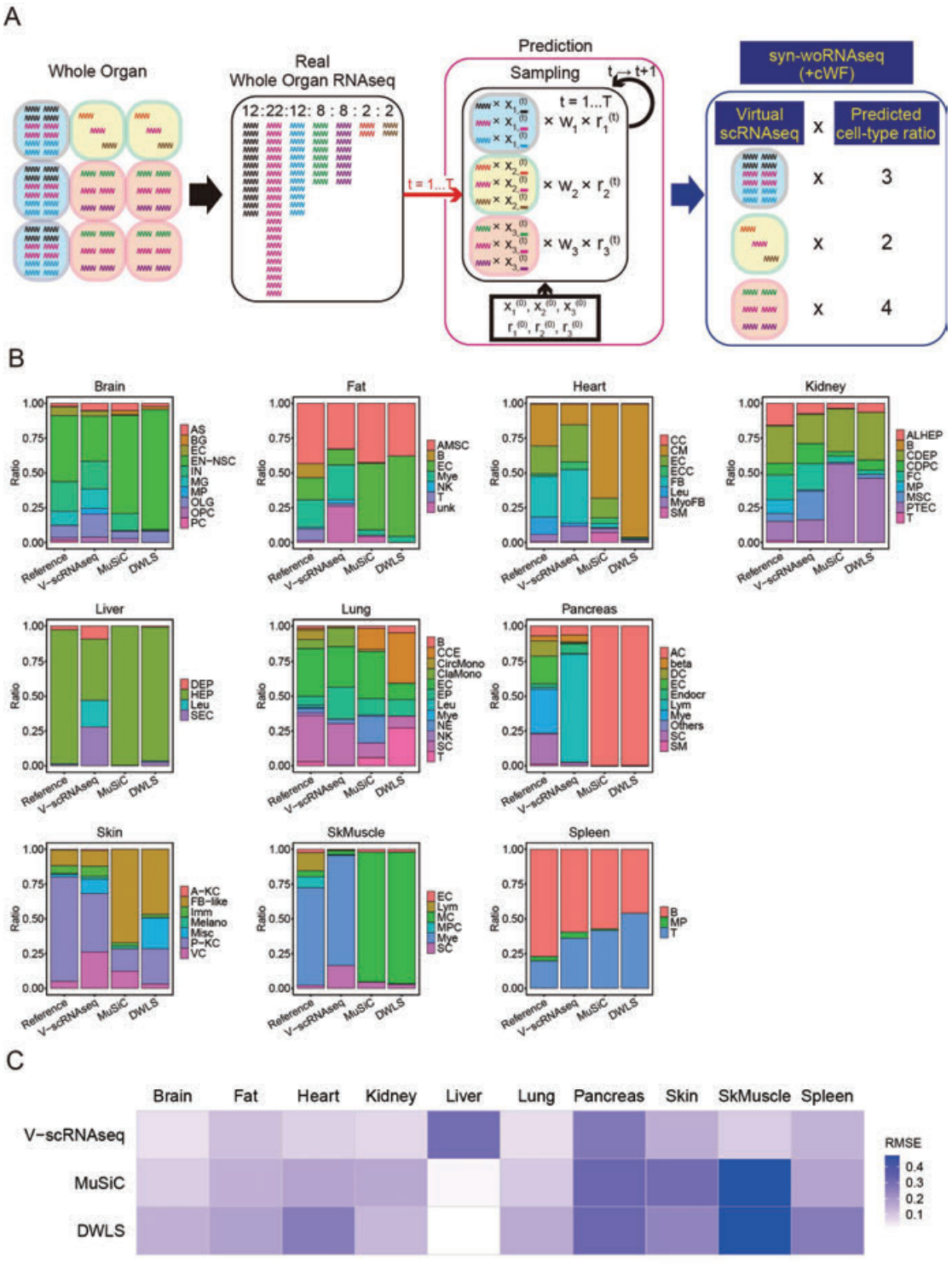


Fig 5. Accurate deconvolution of the whole organ RNA-seq by cWFs. (A) Graphical description of the overall concept of the deconvolution method with the cWFs. Shown is the case of an organ consisting of 3 cell-types (1, 2, 3) with 7 different genes (wiggling lines of different colors). The mean of the cWFs: w_1, w_2, w_3 for each cell-type 1, 2 and 3, respectively. The virtual transcript counts ($x_1^{(t)}$, $x_2^{(t)}$, $x_3^{(t)}$ where the different

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

192 genes are indicated by the wiggling lines of different colors for each cell-type) and the ratio ($r_1^{(t)}$, $r_2^{(t)}$, $r_3^{(t)}$)
193 of virtual cell-type 1, 2 and 3 at each iteration, t, are as indicated. Their initial conditions are $x_1^{(0)}$, $x_2^{(0)}$, $x_3^{(0)}$,
194 $r_1^{(0)}$, $r_2^{(0)}$, $r_3^{(0)}$. The iteration (t= 1, …, T) is performed until both **X** and **r** converge (see Materials and
195 methods). Upon completing the iterations, the sum of the virtual transcript counts weighted by the cWFs
196 (indicated as "Virtual scRNAseq") weighted by the virtual cell-type ratios (indicated as "Predicted cell-type
197 ratio) generates the synthetic whole-organ transcriptome (indicated as "syn-woRNAseq (+cWF)"), as shown
198 at the far right of this panel. **(B)** Bar graph showing the cell type-ratios computed by our deconvolution
199 method (V-scRNAseq) for each organ. The deconvolution was performed with the cWFs computed using
200 the optimal number of the signature genes for each organ (indicated in the accompanying S4 Table). The
201 reference cell-type ratios and the ratios estimated by the conventional deconvolution methods without cWFs,
202 MuSiC and DWLS, are shown as comparisons. The bar graphs are composed of the cell-types computed to
203 be present for each organ by our method. **(C)** The quantitative similarity of the computed ratios to the
204 reference ratios is shown by heatmap of RMSE for our method with cWFs (V-scRNAseq), MuSiC, and
205 DWLS. RMSE: Root Mean Squared Errors. The corresponding raw data are found in S4 Table. SkMuscle:
206 skeletal muscle.
207
208 Briefly, it works as follows (see also Materials and methods): The mean of the cWF
209 (w_1, w_2, w_3 in Fig 5A, the lower case 1, 2, 3 indicate cell-type 1, 2, 3) is used to compute virtual
210 transcript counts ("x" in Fig 5A) for each gene (indicated as wiggling lines of different colors
211 in Fig. 5A) and the ratio (indicated as "r" in Fig 5A) of each cell-type in each organ (painted
212 in different colors in Fig 5A). Using the signature-gene set, the computation is iteratively
213 performed by the sampling method of No-U-Turn Sampler (NUTS), in the Bayesian
214 framework with two hyperparameters, α and β, to account for the combinatorial influence of
215 the cell-type ratios and the gene expression patterns of each cell-type on the whole-organ-level
216 transcriptome, respectively. For the iteration (t = 1, …, T), the initial condition is set as $x_1^{(0)}$
217 and $r_1^{(0)}$ for the transcript counts and the ratio of cell-type 1 (see the diagram labels as
218 "Prediction" in Fig 5A). Those for cell-types 2 and 3 are indicated accordingly (see the diagram

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

219 labels as “Prediction” in Fig 5A). Such initial conditions are the cWF-weighted transcript count
220 for each gene and the reference cell-type ratio, respectively. The iteration is repeated until all
221 sampling variables converge with the signature-gene set. The counts of the non-signature gene
222 set are computed by an analytic approach in the Bayesian framework with a hyperparameter,
223 γ, which has the similar nature of β as above. The “r” is fixed as the above-estimated value,
224 and only “x” in the non-signature gene set is computed.

225 The deconvolution results for 10 organs using this method were compared to the real
226 cell-composition data of the corresponding organ (Fig 5, B and C). The results were also
227 compared to those of two conventional deconvolution methods without the cWF, MuSiC and
228 DWLS [25, 26]. These two were chosen as they have been applied to 1 – 4 organs and appear
229 to outperform the other published methods. The prediction of the cell-type ratios show that our
230 method (V-scRNAseq) outperforms the methods without the cWFs (MuSiC, DWLS) for 9
231 organs (brain, fat, heart, kidney, lung, pancreas, skin, skeletal muscle, spleen). For these organs,
232 the cell-type ratios predicted by our method are more similar to those of the real data than those
233 predicted by MuSiC or DWLS (Fig 5B and S4 Table). Notably, our method corrected the
234 abnormally large ratios of cardiac muscle and muscles in the heart and the skeletal muscle,
235 respectively, estimated by MuSiC and DWLS (Fig 5B and S4 Table). The lower root-mean-
236 squared error (RMSE) for these 9 organs with our method (Fig 5C and S4 Table) further
237 confirms the better performance of our method (Fig 5C and S4 Table). For the liver, our method
238 failed (Fig 5, B and C, and S4 Table), despite the high performance in the reconstitution (Fig
239 4). This may be due to the predominantly hepatocyte composition of the liver. Our method
240 assumes the symmetric Dirichlet distribution as prior distribution which influences the
241 posterior distribution – hence, the method is less suitable for the organ, such as the liver,
242 consisting of a single dominant cell-type.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In addition to the cell-type ratios, we also computed the virtual transcriptome of 23,131 (Brain, Fat, Heart, Lung, Spleen), 22,742 (Kidney), 23,104 (Liver), 15,682 (Pancreas), 21,233 (Skin) and 14,129 (the skeletal muscle) genes in each of the "virtual" cell-type and compared them to those of the real cell-type for the 10 organs (Fig 6 and see Materials and methods). Our approach computes a posterior distribution and its Expected *A Posteriori* (EAP) for counts of each transcript for each cell-type. Hence, we calculated Pearson correlation coefficients for each cell-type, using the EAPs of all genes in virtual and real cell-types for each organ (Fig 6 and S5 Table).

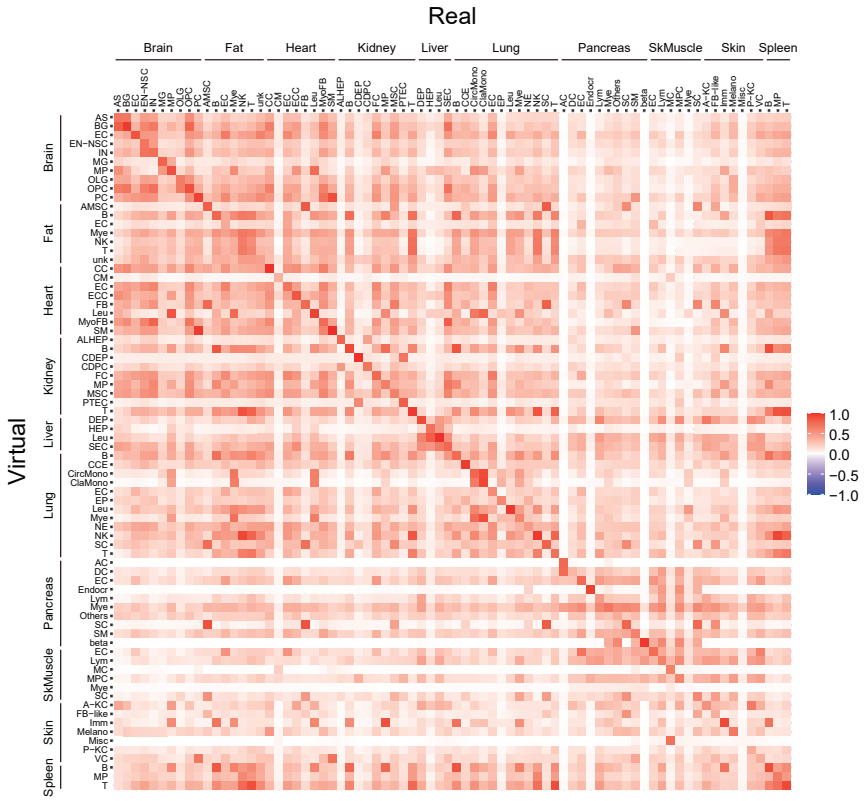


Fig 6. Computation of the complete transcriptome of the composite cell-types of 10 mouse organs. The heatmap of Pearson correlation coefficients for the transcriptome of all virtual and real cell-types across all 10 organs. The Pearson correlation coefficients are calculated for 23,131 (Brain, Fat, Heart, Lung, Spleen), 22,742 (Kidney), 23,104 (Liver), 15,682 (Pancreas), 21,233 (Skin) and 14,129 (the skeletal muscle) genes. The corresponding raw data are found in S5 Table. SkMuscle: skeletal muscle.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Their Pearson correlation coefficients indicate that the virtual transcriptomes are comparable to the real ones across all cell-types and organs. They also recapitulate the transcriptomics-relatedness of the same/related cell types across different organs (Fig 6). For example, the similarity of endothelial cells in multiple organ such as brain, fat, heart, liver and lung is recapitulated in the virtual transcriptomes among these organs (Fig 6). The relatedness of each of the immune system cells across multiple organs is also represented (Fig 6).

The reconstitution and deconvolution results demonstrate that cWFs enable the accurate representation of individual cells in the context of the whole-organs of highly heterogenous compositions, presumably by weighting the variations of the transcriptome-size of different cell-types. We further validate this account by directly comparing the cWFs variations to those of the experimentally-determined total mRNA-contents of the corresponding cell-types.

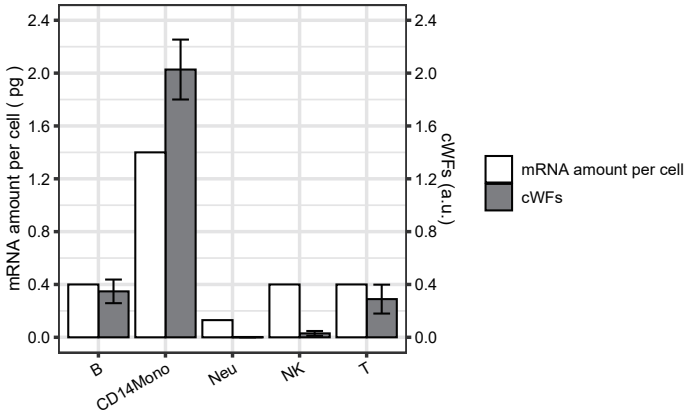


Fig 7. Comparison of the cWFs to the experimentally-determined total mRNA-contents of human PBMC cell-types. The comparison is shown as bar graphs of the experimentally-determined mRNA-amounts per cell (mRNA amount per cell: empty bar) and the computed cWFs (cWFs: filled-bar) for each cell-type. The bars are shown as mean ± S.E. (standard error). B: B-cell, CD14Mono: $CD14^{+}$ monocytes, Neu: Neutrophil, NK: Natural killer cell, T: T-cell. The validation of the cWFs by reconstitution and

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

276 deconvolution are found in S1 Fig (raw data are in S6 – S8 Tables). The corresponding raw data for Fig 7
277 are found in S9 Table.
278
279 The measurement of such variations for individual single-cells of solid organs are
280 difficult and the methods to measure such variations are limited to cultured cells and
281 systemic/circulating cells such as immune cells, and tumor cells [2, 10, 13, 24]. Therefore, we
282 used the experimentally-determined total mRNA-contents of the cell-types composing human
283 peripheral blood mononuclear cells (PBMCs) and compared them to the computed cWFs of
284 the corresponding cell-types (Fig 7).
285 We used bulk and single-cell RNA-seq data of human PBMCs [27-30] and computed
286 the cWFs of B-cells, $CD14^+$ monocytes, $CD16^+$ monocytes, dendritic cells, neutrophils, natural
287 killer (NK) cells, plasmablasts (PB), T-cells, and the others (e.g., red-blood-cells, platelets,
288 etc.) (see Materials and methods). The reconstitution and deconvolution analyses confirmed
289 that they correct the biased gene-expression in the scRNA-seq data caused by the normalization
290 of the transcriptome-size, as indicated by the higher Pearson correlation coefficients (S1A Fig,
291 and, S6 and S7 Tables) and lower RMSE/higher Person correlation coefficient (S1B Fig, and,
292 S8 Table) with the cWFs in the reconstitution and deconvolution analyses, respectively.
293 The total mRNA-amounts are experimentally measured and determined for B-cells,
294 $CD14^+$ monocytes, neutrophils, NK cells, and T-cells of PBMCs [11]. We made side-by-side
295 comparison of the computed cWFs and their experimentally-measured transcriptome-size of
296 the corresponding cell-types (Fig 7 and S9 Table). The result shows that the computed cWFs
297 accurately recapitulate the 3.5 - 10.8-fold higher transcriptome-size of the $CD14^+$ monocytes
298 compared to the B-, T-, NK cells, or neutrophils, respectively (i.e., $CD14^+$ monocytes: 1.4
299 pg/cell vs. B-/T-/NK-cells: 0.4 pg/cell, Neutrophils: 0.13 pg/cell). It also recapitulates the 3-
300 fold higher values of the B- and T-cells than that of neutrophils (i.e., 0.4 vs. 0.13 pg/cell). The
301 experimental values of the mRNA-amounts of B- and T-cells are approximately the same,

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

302 which is also recapitulated by virtually the same cWFs of these two cell-types. These side-by-
303 side comparisons provide another layer of evidence supporting the account that cWFs indeed
304 represent the transcriptome size-variations among different cell-types.
305

306 # Utilities of cWFs

307 Previously, biological and diagnostic utilities of differential transcriptome size are reported [2,
308 12, 15-24]. Hence, we explored such utilities for cWFs. We hypothesized that the
309 transcriptome-size of each cell-type differentially change over aging and that such a phenotype
310 could underlie the aging mechanism and could also be exploited as a new type of aging-
311 biomarker. To test this hypothesis, we examined whether the patterns of the cWFs change over
312 the course of aging. Furthermore, we examined whether such putative changes of the cWFs
313 actually indicate the progression of aging.

314 For this purpose, we used the *Tabula Muris Senis*, the atlas of the multiple-organ single-
315 cell transcriptomics across multiple aging stages of mouse [29, 31]. This atlas provides both
316 the whole-organ RNA-seq data and their corresponding composite scRNA-seq data. Using
317 these data, we computed cWFs of the composite cell-types of 5 organs (brain, heart, kidney,
318 lung, spleen) across 3 aging-stages (3, 18, and 24 months-old) of mice. The 3, 18, 24 months-
319 old (mos.) in mouse correspond to 25 - 26 years-old, 60 - 80 years-old, and ≥ 80 years-old in
320 human, respectively [32]. These 5 organs were chosen for the following two reasons: 1) they
321 performed well in both the reconstitution and deconvolution validations (Figs 4 and 5); 2) their
322 equivalent whole-organ/single-cell data are also present in our 10 organs-analysis described
323 above.

324 Using these datasets, we computed the cWFs of all composite cell-types for each organ.
325 The reconstitution (S2 Fig, and, S10 and S11 Tables) and deconvolution (S12 Table) analyses
326 confirmed that they correct the biased representations of the gene expression and cell-type

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

compositions, respectively, in the scRNA-seq data with one or more of the signature gene set(s) for each aging-stage. Based on these results, we selected the best performing numbers of the signature genes for each organ and aging-stage by omitting the conditions that failed to significantly improve the reconstitution (highlighted in pink in S12 Table) and selecting the ones with the lowest RMSE and the highest Pearson correlation coefficients in the deconvolution analysis (highlighted in light-green in S12 Table). This screening selected the cWFs computed with the following conditions: Brain (3 mos. – 500 signature genes, 18 mos. – 300 signature genes, 24 mos. – 300 signature genes), Heart (3 mos. – 100 signature genes, 18 mos. – 500 signature genes, 24 mos. – 300 signature genes), Kidney (3 mos. – 300 signature genes, 18 mos. – 500 signature genes, 24 mos. – 500 signature genes), Lung (3 mos. – 500 signature genes, 18 mos. – 300 signature genes, 24 mos. – 300 signature genes), Spleen (3 mos. – 300 signature genes, 18 mos. – 100 signature genes, 24 mos. – 500 signature genes).

With these selected cWFs, we examined whether their relative differences change over the course of aging. For this purpose, we compared the ratios of the cWFs of a pair of two distinct cell-types in each organ across three aging-stages (Fig 8 and S13 Table). The analysis detected their statistically significant ($p < 0.01$ by U-test) changes of many of the cell type-pairs in all organs over the course of aging.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

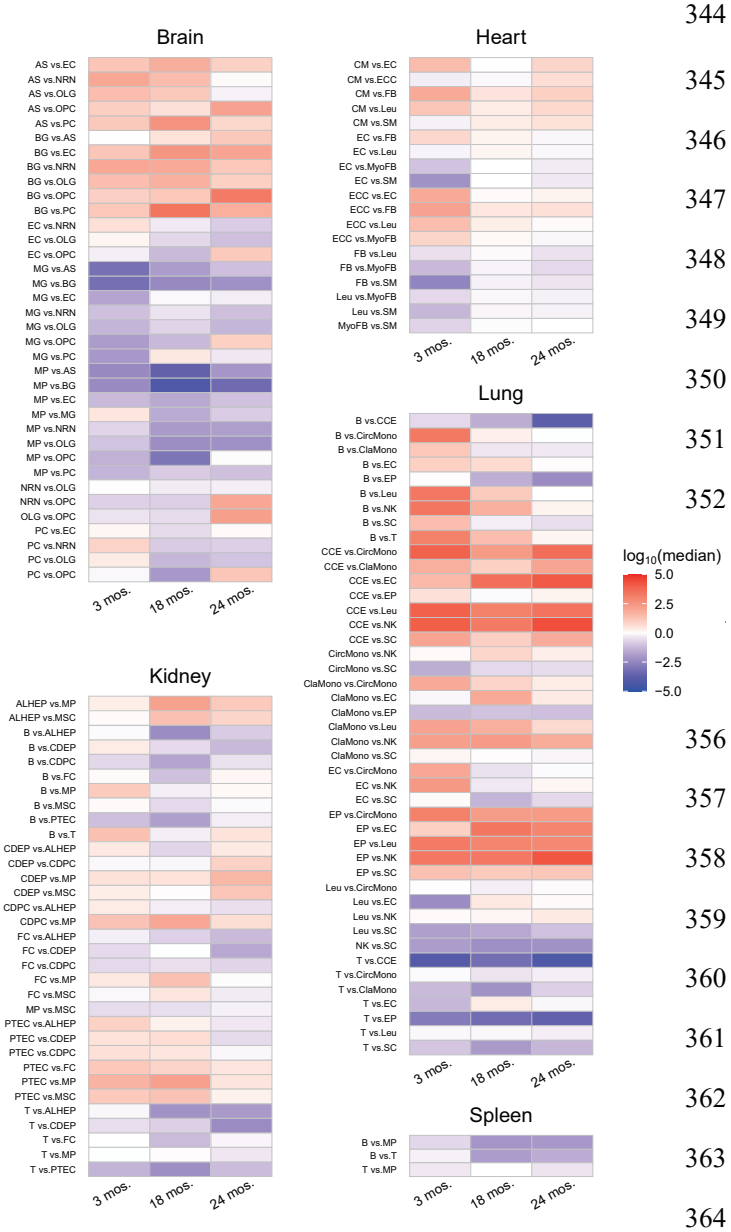


Fig 8. Differential cWFs during aging of mouse. The changes of the ratios of cWFs of cell type-pairs across three aging-stages (3 months-old: 3 mos., 18 months-old: 18 mos., 24 months-old: 24 mos.) are shown as heatmap for each organ (Brain, Heart, Kidney, Lung, Spleen). The cell type-pairs are indicated at the left of each heatmap. The ratios are the pairwise ratios of the corresponding cell-types. For example, AS vs. EC indicates the ratio of the cWF for AS (numerator) and EC (denominator). The ratios are indicated as $\log_{10}$ of the medians. Shown are the pairs resulted in $p < 0.01$ by U-test in one or more of the three pairwise aging stage-comparisons (18 mos. vs. 3 mos., 24 mos. vs. 3 mos., 24mos. vs. 18 mos.). The analysis was conducted with those showing the significant improvements in the reconstitution (S2 Fig, and, S10 and S11 Tables) and the best deconvolution results (S12 Table) for each organ. The full names of the cell type abbreviations and the raw data of the heatmap are found in S13 Table.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We next investigated whether such differential pairwise ratio-changes of the cWFs predict the progression of aging. For this purpose, we applied a machine learning algorithm, LightGBM [33], to the dataset of the pairwise cWF ratios between the cell type-pairs listed above (see Methods). First, each organ dataset was independently evaluated (Fig 9). The result shows high predictability with the pairwise cWF ratios between the brain, the lung, the heart, and the kidney cells, as indicated by their high prediction index scores (i.e., accuracy, macro-recall, macro-precision, macro-F1 score) of 0.88 – 0.98 (Fig 9A).

A

	Accuracy	macro−Precision	macro−Recall	macro−F1 Score
Brain	0.983	0.984	0.983	0.983
Heart	0.883	0.893	0.883	0.882
Kidney	0.883	0.883	0.883	0.881
Lung	0.967	0.967	0.967	0.967
Spleen	0.517	0.514	0.517	0.514

B

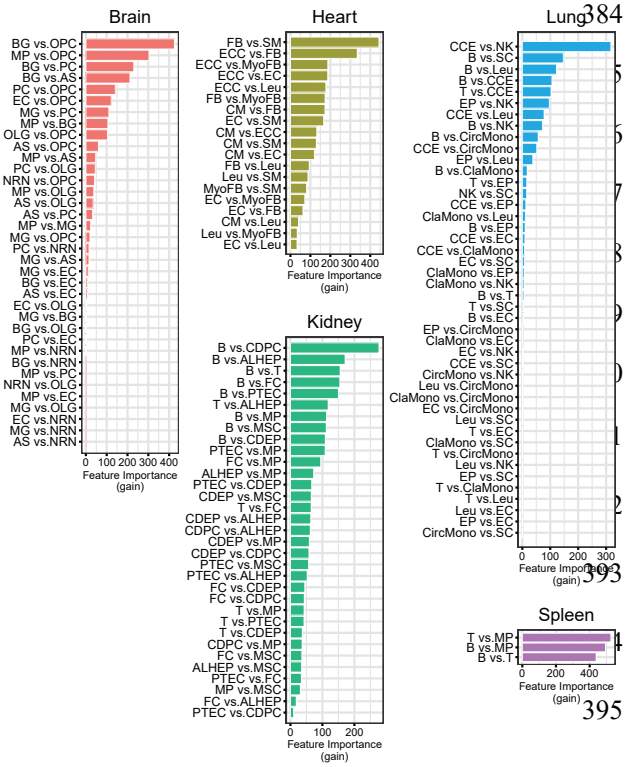


Fig 9. Prediction of the aging-stage by the differential cWFs of each organ. **(A)** The prediction scores (accuracy, macro-precision, macro-recall, macro-F1 score) for each organ (Brain, Heart, Kidney, Lung, Spleen) are shown. **(B)** The top-ranked features (i.e., pairwise cell-types) important for the prediction are shown as bar graphs for each organ. The feature importance is indicated by gain. The corresponding raw data are found in S14 Table.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The LightGBM prediction with all organs combined also shows high performance as indicated by its prediction index scores of 1 (Fig 10A). The feature importance analysis shows the top-ranked cell type-pairs are those of the lung and the brain, supporting the higher performance results with the cell-types of these organs when they are independently analyzed (Fig 9A). The cell-types indicated as important contributors to the prediction listed above for the individual organ analyses also ranked higher in this all-organ-combined analysis (Fig 10B and S14 Table). Taken together, the results demonstrate that the differential cWF ratios of the brain, the lung, the heart and the kidney accurately predict the progression of aging, suggesting a possible role of the differential transcriptome-size among the cells in these organs in the aging mechanism. The results also indicate a possibility of using these indices to determine the progression of aging.

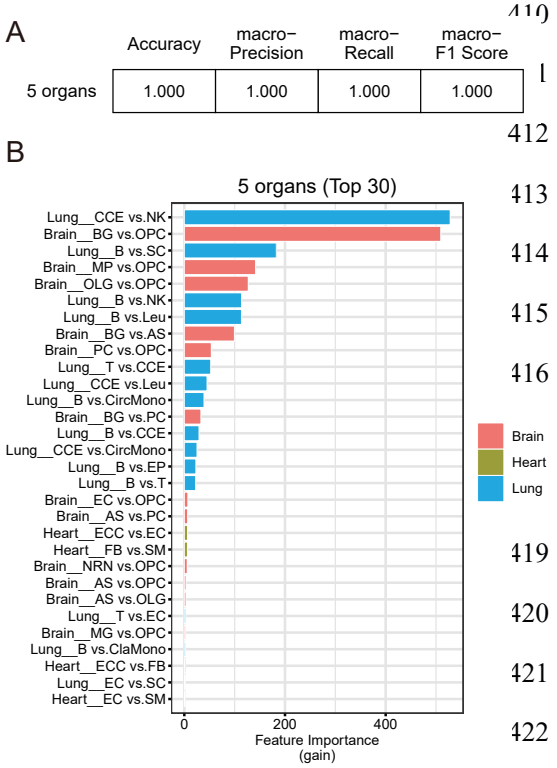


Fig 10. Prediction of the aging-stage by the differential cWFs of all 5 organs combined. (A) The prediction scores (accuracy, macro-precision, macro-recall, macro-F1) by combining the cWF of all pairwise cell-types across all organs are shown. **(B)** The top 30 important features (i.e., the pairwise ratios of the cWFs) for the prediction are shown as bar graphs for each organ. The feature importance is indicated by gain. The cell type-pairs and their organs are indicated as organ-name_cell-type pair (e.g., Lung_CCE vs. NK: the CCE vs. NK pair of the lung). The corresponding raw data are found in S14 Table.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

In this study, we report a computational method to calculate cWFs, coefficients to offset the lack of the representation of transcriptome size variations and differential cryptic gene expression among different cells of solid organs (Fig 2). Using this method, we computed and describe cWFs for 76 cell-types across 10 organs (Fig 3). We show that they indeed account for the relative variations of the transcriptome-size of the cells and their differential cryptic gene expression by demonstrating their requisite role for accurately reconstituting and deconvolving the whole-organ RNA-seq data using their composite scRNA-seq data weighted by the corresponding cWFs (Figs 4 – 6). We also show that they recapitulate the experimentally-determined total mRNA content-differences, using PBMC data (Fig 7).

Furthermore, we show that cWFs differentially change among diverse cell-types across various organs over aging and can be used to effectively predict aging progression in the aging mouse model (Figs 8 – 10). These results suggest a possible role of differential transcriptome size-change over lifetime in the aging mechanism. The results also indicate their possible diagnostic applications as biomarkers for aging progression and/or aging associated diseases.

Despite the usefulness of cWFs, some limitations must be noted. One limitation is the requirement of both bulk and single-cell RNA-seq data of the organs and the tissues of interest. Our method requires both of these datasets for the target organs/tissues to infer the cWFs of the composing cell-types. However, this may become less of a limiting factor as the availability of such data is exponentially growing. Once we determine the cWFs for the specific cell-types of organs, the same cWFs can be used for the same cell-types of the same organs regardless of their sources. Therefore, the cWFs reported in this study can be used for the same 76 cell-types and 10 organs for other studies. Another requirement is that the method needs at least a rough a priori knowledge about the cell-type compositions (i.e., the ratio of cell-types) of the organs and the tissue of interest -i.e., the method works only in a supervised framework. This limitation

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

450 could also become less limiting as a growing body of scRNA-seq data provides information on
451 the cell-type compositions of previously less characterized organs/tissues.
452

453 Materials and methods

454 Datasets

455 For the whole-organ RNA-seq data, two murine datasets are used: One from 11 weeks-old male
456 C57BL6/N Jcl mice and derived from 'iOrgans Atlas' (http://i-organs.atr.jp) [34, 35] and the
457 other from 3, 18, and 24 months-old male C57BL6/N mice and derived from the *Tabula Muris*
458 *Senis* [29]. For the scRNA-seq datasets, used are the brain, the fat, the heart, the liver, the lung,
459 and the spleen data from 11 weeks-old male C57BL6/JN mice described in *Tabula Muris* [36].
460 The kidney data are from 3 months-old male C57BL6/JN mice in *Tabula Muris Senis* [31].
461 The skeletal muscle and the pancreas data are from 8-10 weeks-old female C57BL6/J reported
462 in Mouse Cell Atlas [37]. The skin data are from 9 weeks-old female C57BL6/J mice [38]. The
463 aging mouse scRNA-seq datasets in the brain, the heart, the kidney, the lung and the spleen are
464 from 3, 18, 24 months-old male C57BL6/JN mice in *Tabula Muris Senis* [29, 31]. The human
465 PBMC datasets are GSM2871599 in GSE107572 (bulk RNA-seq) [43, 44] and GSM4557334
466 in GSE150728 (scRNA-seq) [30], both are from healthy donors.
467

468 Data preprocessing

469 Cell-type labels of the organs used in this study are as described in each corresponding original
470 single-cell study, except for those of the skeletal muscle, the pancreas, and the heart (the *Tabula*
471 *Muris Senis,* but not the *Tabula Muris,* data) and the human PBMC. For the skeletal muscle
472 and the pancreas, the multiple labels of the cell-types of the similar gene expression patterns in
473 the original study are combined into the single label as follows (*** indicating arbitrary

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

474 characters). For the skeletal muscle, 'B cell_***', 'Dendritic cell', and 'T cell' into 'Lymphatic
475 cell', 'Erythroblast_***', 'Granulocyte monocyte progenitor cell', 'Macrophage_***' and
476 'Neutrophil_***' into 'Myeloid cell', 'Muscle cell_***' into 'Muscle cell'. For the pancreas,
477 'B cell' and 'T cell' into 'Lymphatic cell', 'Dendritic cell', 'Erythroblast_***', 'Granulocyte',
478 and 'Macrophage_***' into 'Myeloid cell', 'Endothelial cell_***' into 'Endothelial cell',
479 'Smooth muscle cell_***' into 'Smooth muscle cell', 'Stromal cell_***' into 'Stromal cell',
480 'Dividing cell' and 'Glial cell' into 'Others'. For the heart dataset in *Tabula Muris Senis* [31],
481 the labels were replaced with the ones in *Tabula Muris* [20] using single-cell reference mapping
482 by '*Seurat*' v4 [39], as follows. First, the reference scRNA-seq data in *Tabula Muris* were
483 preprocessed by '*Seurat*' analyses with the log-normalization (where the scale factor is 10^6),
484 the highly variable genes selection with default parameters, the scaling for all genes, and
485 principal component analysis (PCA) acquiring 200 principal components (PCs). The number
486 of PCs used for the downstream analysis was determined at p-value below 0.0001 in JackStraw
487 analysis, resulting in 33 PCs. Using these 33 PCs reference data, the query heart scRNA-seq
488 datasets from *Tabula Muris Senis* [31] were annotated by the functions of
489 'FindTransferAnchors()' and 'TransferData()' in '*Seurat*'. For the human PBMC, we
490 combined cell-type labels as follows: 'Class-switched B' and 'B' into 'B', 'CD8m T', 'CD4m
491 T', 'CD4n T', 'CD4 T', 'gd T' and 'CD8eff T' into 'T', 'IgG PB' and 'IgA PB' into 'PB',
492 'pDC' and 'DC' into 'DC', 'RBC', 'Platelet', 'Activated Granulocyte' and 'SC & Eosinophil'
493 into 'Others'. The gene symbol-matching between the scRNA-seq and the whole-organ RNA-
494 seq data was conducted by using entrez gene IDs derived from 'org.Mm.egALIAS2EG' in an
495 R package, '*org.Mm.eg.db*'. ERCC-labelled genes were removed as they are spike-in genes. In
496 addition, the counts of the three genes (*Rn45s, Akap5, Lrrc17*) were removed as they are non-
497 mRNA artifacts significantly influencing the total counts. The normalization was performed

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

by adjusting the total counts of each cell in the scRNA-seq datasets to a million. The same normalization process was also applied to each of the whole-organ RNA-seq data.

Calculation of reference cell-type ratios

The reference cell-type ratios were as described in each original single-cell study, except for the heart and the brain. For the heart, the cell-type ratios described in *Tabula Muris* [36] were based on the separate analysis of the cardiac muscles and the non-muscle cells, resulting in a large under-estimation of the cardiac muscle ratio – 3.1% [36]. In contrast, the ratio determined by multiple other methods and considered as the gold-standard in the cardiovascular fields is 30 - 40% [40-46]. According to this gold-standard ratios, we made the following modifications: We set the ratio of the cardiac-muscle cells at 30% and the remaining at 70%. The latter, then, was divided by the non-muscle cell types by maintaining their ratios the same to those in the *Tabula Muris* data. For the brain, the cell-type ratio reported in the NIH database (https://www.nervenet.org/papers/brainrev99.html#Numbers) was used to modify those in the *Tabula Muris* scRNA-seq data [36], as follows. First, we divided the brain cell-type classes into four classes: 'neurons', 'glial cells', 'endothelial cells', 'others', and then set the ratio of these classes at 75 : 23 : 7 : 4, respectively, accordingly to their estimated ratios in the NIH murine brain database (https://www.nervenet.org/papers/brainrev99.html#Numbers). Second, we further divided 'neurons', 'glial cells', and 'others' into more cell-type classes according to *Tabula Muris* [36]. The 'neurons' were further divided into 'neuron-excitatory neurons and some neuronal stem cells' and 'neuron-inhibitory neurons'. The 'others' class was divided into 'brain pericyte-NA' and 'oligodendrocyte precursor cell-NA'. For the 'glial cells' class, we introduced three assumptions: 1) the 'glial cells' can be classified into four cell-types, 'microglial cell-NA', 'astrocyte-NA', 'Bergmann glial cells-NA', and 'oligodendrocyte-NA', according to *Tabula Muris* [36]; 2) the ratios of these four glial cell types follow those of

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tabula Muris [36] among them, and 3) the ratio of 'microglial cells-NA' in the whole brain is 0.1, as it is reported that the 'microglial cells' account for 10 – 15 % of the whole brain cells. On the basis of these assumptions, the ratios of each brain cell type was estimated as 'macrophage-NA' (ca. 0.2%), 'microglial cell-NA' (10.0%), 'astrocyte-NA' (ca. 2.2%), 'Bergmann glial cell-NA' (ca. 2.1%), 'brain pericyte-NA' (ca. 1.5%), 'endothelial cell-NA' (ca. 6.4%), 'neuron-excitatory neurons and some neuronal stem cells' (ca. 47.5%), 'neuron-inhibitory neurons' (ca. 21.3%), oligodendrocyte-NA' (ca. 8.7%) and 'oligodendrocyte precursor cell-NA' (ca. 1.9%) and used as the reference. For the human PBMC, the reference cell-type ratios are as described [27, 28], except for those of "$CD16^+$ monocytes" and "PB and Others". The ratios of these cell-types are derived as follows: The ratios of the cell-type label "the others" was divided to '$CD16^+$ monocytes' (ca. 7.21%), 'PB' (ca. 0.24%), 'Others' (ca. 3.77%), according to their RNA-seq data-based ratios.

Selection of signature gene sets

Random Forest (RF) was performed to extract signature genes for distinguishing cell-types from one another in each organ using the scRNA-seq count data derived from the cell-type reference datasets described in the above section. In this study, the '*randomForest*' package in R was used for tuning and producing a classifier by RF. The scRNA-seq data were first divided by 8:2 into the training and the test data. Using the training data, we determined two parameters, 'mtry' and 'ntree' for producing an RF-model. The parameter 'mtry' means the number of features used for an RF-model and was tuned by a function 'tuneRF()', while the parameter 'ntree' means the number of trees generated in an RF-model and was set as 500, which is sufficient to converge error rate in cell-type classification. The produced RF-model was validated by using the test data and F1-score calculated by a function 'F1_Score()' in an R package '*MLmetrics*'. Then we confirmed the validity of the model by F1-score over 0.8.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

548 Following the classifier production, the important features in the classifier were extracted as
549 the signature genes for each cell-type, where we used 'Mean Decrease in Gini' values as the
550 importance indicator for each gene.
551

552 ## Computation of the cell type-specific weighting-factors (cWFs)

553 The overall concept for computing the cWFs is graphically described in Fig 2 and is as follows:
554 The observed scRNA-seq counts are normalized (i.e., the total counts per cell is the same for
555 all individual cells). Hence, the putative differential transcriptome-size among various cell-
556 types is lost from the count data. The cWFs (w_1, w_2, w_3, etc.) offset such loss. In their
557 computation, we calculated at least 100 cWFs for each virtual-cell/cell-subject of each cell-
558 type and made them follow Gaussian distribution, instead of one cWF per cell-type. Based on
559 this concept, we developed a model as follows:

560 $$m\mathbf{y} = \sum_j w_j \mathbf{x}_j \ (j \in C^m) \quad (1)$$

561 where $\mathbf{y} \in \mathbb{R}^n$, $w_j \geq 0$ and $\mathbf{x}_j \in \mathbb{R}^n$ denote the normalized whole-organ RNA-seq counts
562 vector, the cWF for each cell subject j, and the normalized scRNA-seq counts vector for each
563 cell subject j, respectively. The n is the total number of the signature genes for the organs. In
564 this study, 100, 300, and 500 signature genes were selected according to the ranking on the
565 basis of 'Mean Decrease in Gini' values calculated by the RF analyses. The numbers, 100, 300,
566 and 500 were arbitrarily selected but the smaller number of signature genes is computationally
567 more cost-effective (see the details also in the section below). The combination of cell subjects,
568 C^m was selected at random while maintaining the reference cell-type ratio described above
569 with the total number of cell subjects, m, which is arbitrarily set to 100. In addition, we set C^m
570 containing at least 1 cell subject in each cell-type. With this model, w_j was calculated by
571 solving a quadratic problem under the constraint that the resulting value is non-negative as
572 follows:

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\widetilde{w}_j = \mathrm{argmin}_{w_j} \sum_i^S \left| m\mathbf{y}_i - \sum_j w_j \mathbf{x}_j \right|^2 \; s.t. \; w_j \geq 0 \quad (2)$$

Here, S is the number of whole-organ RNA-seq count data. This quadratic problem was solved by using '*osqp*' package in R. The process of both random-selection of the cell subject combination and $\widetilde{w}_j$ calculation was recursively performed until more than 100 w_j were generated for all selected cell subjects. In every recursion, Pearson correlation coefficients were calculated between the real whole-organ RNA-seq and the corresponding synthetic whole-organ RNA-seq generated with or without the weighting-factors. Then, by assuming that the distribution of the weighting-factors follows Gaussian distribution, we calculated the mean and variance of the weighted counts for each cell subject as follows:

$$\tilde{\mathbf{x}}_j \sim \mathcal{N}\left(\mu_{w_j}\mathbf{x}_j, \sigma_{w_j}^2 \mathrm{diag}\left(\mathbf{x}_j \odot \mathbf{x}_j\right)\right) \quad (3)$$

where, $\tilde{\mathbf{x}}_j$, μ_{w_j}, and $\sigma_{w_j}^2$ denote the weighted count vector for the cell subject j, the mean of the cWFs for the cell subject j, and the variance of the cWFs for the cell subject j, respectively. The operator $\odot$ represents the element-wise product between two vectors. On the basis of the computed mean and variance of the weighting-factors for each cell subjects, the weighted count vector of cell-type k was calculated by assuming that they follow Gaussian distribution as follows:

$$\bar{\mathbf{x}}_k \sim \mathcal{N}\left(\frac{1}{N_k}\sum_j \mu_{w_j}\mathbf{x}_j, \frac{1}{N_k}\sum_j \sigma_{w_j}^2 \mathrm{diag}\left(\mathbf{x}_j \odot \mathbf{x}_j\right)\right) \; (j \in C^k) \quad (4)$$

where, k, C^k, and N_k denote a cell-type, a set of subjects labelled with cell-type k, and the number of subjects in C^k, respectively.

Computing the cell-type ratios & gene expression patterns

The overall concept is graphically described as Fig 5A and the model is as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{r} \quad (5)$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where, $\mathbf{y}$, $\mathbf{X} = [\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_k, \ldots, \bar{\mathbf{x}}_K]$, and $\mathbf{r}$ are the whole-organ RNA-seq data vector, the matrix of the virtual transcript counts where the columns are the weighted-counts for each cell-type calculated as above, and the coefficient vector corresponding to the cell-type ratio, respectively. To compute $\mathbf{X}$ and $\mathbf{r}$, we employed Bayes' theorem. To apply Bayes' theorem, Gaussian noise $\mathcal{N}(\mathbf{0}, \beta\mathbf{I})$ was added to Equation (5) and a probabilistic model was developed as follows:

$$\mathrm{P}(\mathbf{y}|\mathbf{X}, \mathbf{r}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{r}, \beta\mathbf{I}) \quad (6)$$

where, β denotes a hyperparameter. According to the Bayes' theorem, the posterior distribution of $\mathbf{X}$ and $\mathbf{r}$ was obtained as

$$\mathrm{P}(\mathbf{X}, \mathbf{r}|\mathbf{y}) \propto \mathrm{P}(\mathbf{y}|\mathbf{X}, \mathbf{r})\mathrm{P}(\mathbf{X})\mathrm{P}(\mathbf{r}) \quad (7)$$

where, $\mathrm{P}(\mathbf{X})$ and $\mathrm{P}(\mathbf{r})$ denote prior distribution of $\mathbf{X}$ and $\mathbf{r}$, respectively. $\mathrm{P}(\mathbf{X})$ and $\mathrm{P}(\mathbf{r})$ are given as follows:

$$\mathrm{P}(\mathbf{X}) = \prod_{k=1}^{K} \mathrm{P}(\bar{\mathbf{x}}_k) = \prod_{k=1}^{K} \mathcal{N}(\bar{\mathbf{x}}_k|\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}_k^{-1}) \quad (8)$$

$$\mathrm{P}(\mathbf{r}) = Dirichlet(\mathbf{r}|\alpha\mathbf{1}_K) \quad (9)$$

where, α and $\mathbf{1}_K$ denote a hyperparameter and the K-dimensional all-ones vector. In prior distribution $\mathrm{P}(\mathbf{X})$, $\mathcal{N}(\bar{\mathbf{x}}_k|\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}_k^{-1})$ is the same as Equation (4). In other words, $\boldsymbol{\mu}'_k = \frac{1}{N_k}\sum_j \mu_{w_j}\mathbf{x}_j$ and $\boldsymbol{\Sigma}_k^{-1} = \frac{1}{N_k}\sum_j \sigma_{w_j}^2 \mathrm{diag}(\mathbf{x}_j \odot \mathbf{x}_j)$. To maximize the posterior distribution, we employed an extension of Hamiltonian Monte Carlo (HMC) sampler, the No-U-Turn Sampler (NUTS) by a software, 'stan', which was performed on R program by an R package, '*rstan*', and used only signature gene sets for sampling, as the running time is reduced. The probabilities of hyperparameters, α and β were set as zero-truncated standard normal distribution. In addition, we set zero for the negative values in the sampled elements of $\mathbf{X}$. The 'stan' sampling parameters, such as the number of iterations, max_treedepth, adapt_delta, thin, etc., were appropriately chosen to converge well the sampled distributions. The convergence was evaluated by R-hat values (Gelman-Rubin statistics) in all sampling variables below 1.1, which

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is a generally accepted value for convergence evaluation. For the expression pattern of non-signature genes, we set another noise parameter γ in Equation (6), instead of β and estimated $\bar{\mathbf{x}}_k$ (the expression patterns in non-signature genes) by the following probability equation:

$$\mathrm{P}(\bar{\mathbf{x}}_k|\mathbf{y}, \bar{\mathbf{r}}, \{\bar{\mathbf{x}}_l\}_{l\neq k}) \propto \mathrm{P}(\mathbf{y}|\mathbf{X}, \bar{\mathbf{r}})\mathrm{P}(\bar{\mathbf{x}}_k) \quad (10)$$

where, $\bar{\mathbf{r}}$ denotes the mean values of $\mathbf{r}$ determined by the sampled distribution above.

Then, $\mathrm{P}(\bar{\mathbf{x}}_k|\mathbf{y}, \bar{\mathbf{r}}, \{\bar{\mathbf{x}}_l\}_{l\neq k})$ follows Gaussian distribution and its mean and variance can be calculated by the following equations:

$$\mathbb{E}[\bar{\mathbf{x}}_k] = (\gamma \mathrm{r}_k^2\mathbf{I} + \boldsymbol{\Sigma}_k^{-1})^{-1}\left(\gamma \mathrm{r}_k\mathbf{y} + \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k' - \gamma\sum\nolimits_{l\neq k} \mathrm{r}_k\mathrm{r}_l\bar{\mathbf{x}}_l\right) \quad (11)$$

$$\mathrm{Var}[\bar{\mathbf{x}}_k] = \gamma \mathrm{r}_k^2\mathbf{I} + \boldsymbol{\Sigma}_k^{-1} \quad (12)$$

In the resulting $\mathbb{E}[\bar{\mathbf{x}}_k]$, all negative values were set to zero. The hyperparameter γ was set as either of $10^{-5}, 10^{-4}, \ldots, 10^{5}$. The estimating calculation was recursively performed until the root mean square error (RMSE) of $\bar{\mathbf{x}}$ between iterations gets below 1 or until at most 100 times enough to converge, where one iteration is the estimating calculations of Equation (11) and (12) for all k (i.e., all cell-types) and for each iteration, $\bar{\mathbf{x}}_l$ is from the prior iteration. The optimal value of the hyperparameter γ was determined by the least RMSE between the original whole-organ RNA-seq (i.e., $\mathbf{y}$) and the estimated one (i.e., $\bar{\mathbf{y}}$).

Selection of the optimal number of the signature genes

The optimal number (100, 300, or 500) of the signature genes were determined based on the three criteria as follows: 1) Improvement of the Pearson correlation coefficient between the r-woRNAseq and the syn-woRNAseq(+cWF), 2) Convergence in the deconvolution, 3) Higher similarity to the reference cell type composition (i.e., lower RMSE). The results with the selected optimal number of the signature genes were used in the comparisons with those

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

643 obtained by MuSiC and DWLS, and also for the analyses with the aging data and for the cWFs
644 analysis of the human PBMC data described below.
645

646 MuSiC and DWLS

647 MuSiC was conducted as described [26]. DWLS was performed as described [25], except that
648 a function, 'solve.QP()' (R package: '*quadprog*') was substituted by 'solve_osqp()' (R
649 package: '*osqp*') for quadratic problem solver.
650

651 Comparison of the real and virtual scRNA-seq counts of 10 organs

652 The mean counts of each cell-type in each organ were compared by Pearson correlation
653 coefficients. For the virtual scRNA-seq, the estimated mean count for each cell-type was
654 normalized to a million. For the real scRNA-seq, the mean count for each individual cell was
655 normalized to a million and averaged for each cell-type.
656

657 Analysis of differential cWFs in aging progression

658 The distribution of the pairwise cWF ratios was calculated for each aging-stage. The ratios of
659 the pairwise cWFs in each cell-type, instead of the cWFs themselves, were used as the cWFs
660 are relative values calculated from the normalized count datasets of the real whole-organ RNA-
661 seq and thus the cWFs values themselves can be used only for their comparison among the
662 cell-types within each organ at each aging-stage (i.e., the cWFs cannot be compared across the
663 aging-stages). Hence, their pairwise ratios were used for the comparisons across the aging-
664 stages. For each pairwise cell-type within an organ, the distribution of the cWF ratios was
665 calculated for all pair combinations of the consisting cells. For example, assuming that cell-
666 type X has 1,000 cells and cell-type Y has 500 cells with one cWF per cell, respectively, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

667 distribution of the ratio of cWFs of cell-type X vs. Y consisting of 500,000 values of the ratios
668 was calculated. When a cell-type contains zero value of cWF, we substituted the zero value for
669 the second minimum cWF value in the cell-type to avoid the division by zero. Next, for each
670 cell-types pair, we performed Mann-Whitney U-test to make statistical evaluations of the
671 differences of their ratio-distributions between the aging-stages (i.e., 18 mos. vs. 3 mos., 24
672 mos. vs. 3 mos., 24 mos. vs. 18 mos.). Zeros were generated if the resulting values were below
673 the lower bound of R program, 1e-308.
674

675 ## Machine learning prediction of the aging-stages by the differential
676 ## cWF ratios

677 For this analysis, we omitted the data of the cell-types pairs of which statistical analysis
678 described above resulted in $p \geq 0.01$ across all pairs of the aging-stages. Using the remaining
679 data, we performed LightGBM classifier analysis [33]. We first generated input data for
680 LightGBM classifier analysis using the cWF ratio distributions obtained as described in the
681 above section. The 100 cWF ratio distribution-values were randomly sampled for each cell-
682 types pair for each aging-stage. The sampled values were at random combined with each other,
683 resulting in the matrix of "the 100 cWFs ratio distribution values" x "the number of the cell-
684 type pairs" for each aging-stage. Using these matrices for all three aging-stages as the input
685 data, we generated a LightGBM classifier of aging-stages on Python, using the '*lightgbm*'
686 package [33]. The input data was divided by 8:2 into training data and test data. The training
687 data was used for tuning the parameters of LightGBM. The tuning was performed by the
688 '*optuna*' package [47], searching for the best parameters to maximize macro-F1 score.
689 Specifically, the searched parameters are max_depth, num_leaves, subsample, subsample_freq,
690 colsample_bytree, min_child_samples, ranging from 3 to 12, 2 to 256, 0.1 to 1.0, 1 to 7, 0.1 to
691 1.0, and 5 to 100, respectively. The model was generated using the best parameters and

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

692 evaluated using test data by accuracy, macro-Precision, macro-Recall, and macro-F1_score.
693 These evaluations were performed by a Python package, 's*cikit-learn*' [48]. In addition, we
694 also extracted feature importance based on 'gain' from the model and confirmed the high-
695 ranked contributors in the model (i.e., aging-stage classification).
696

697 Supporting information

698 **S1 Fig. Reconstitution and deconvolution of the bulk human PBMC RNA-seq by the composite**
699 **scRNA-seq.** (**A)** Reconstitution results with and without (no) cWFs are compared. The results with the
700 100, 300, 500 signature genes are shown. The similarity is shown as violin plot of the Pearson
701 correlation coefficients. The corresponding raw data are found in tables S6 and S7. **(B)** Bar graph
702 showing the cell type-ratios computed by the deconvolution method (V-scRNAseq) for each organ. The
703 deconvolution was performed with the cWFs computed using the optimal number of the signature genes
704 for each organ (indicated in the accompanying table S8). The bar graphs are composed of the cell-types
705 computed to be present for each organ by our method. The similarity scores (RMSE: Root Mean
706 Squared Errors, Pearson correlation coefficient) are indicated at the top of the bar. The corresponding
707 raw data of S1 Fig. are found in S6, S7, S8 Tables.
708

709 **S2 Fig. Reconstitution of the whole organ RNA-seq of the aging model mouse by the composite**
710 **scRNA-seq.** The results with and without (no) cWFs are compared for each aging-stage (3 mos., 18
711 mos. 24 mos.) for each number of the signature gens (100, 300, 500) and for each organ (Brain, Heart,
712 Kidney, Lung, Spleen). The similarity is shown as violin plots of the Pearson correlation coefficients.
713 The corresponding raw data are found in S10 Table (no cWFs) and S11 (with cWFs). The corresponding
714 raw data of S2 Fig. are found in S10 and S11 Tables. All supplementary tables below are available at
715 https://github.com/k-teji/VscRNAseq.
716

717 **S1 Table (Separate file)**
718 The raw data of Fig. 1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

719

720 **S2 Table (Separate file)**

721 The raw data of Fig. 3.

722

723 **S3 Table (Separate file)**

724 The raw data of Fig. 4.

725

726 **S4 Table (Separate file)**

727 The raw data of Fig. 5, B and C.

728

729 **S5 Table (Separate file)**

730 The raw data of Fig. 6.

731

732 **S6 Table (Separate file)**

733 The raw data of S1 Fig. (Pearson correlation coefficients for the result with no cWFs).

734

735 **S7 Table (Separate file)**

736 The raw data of S1A Fig. (Pearson correlation coefficients for the result with cWFs).

737

738 **S8 Table (Separate file)**

739 The raw data of S1B Fig. The best performing result (the lowest RMSE and the highest Pearson

740 correlation coefficient) shown in S1B Fig. is highlighted in light green.

741

742 **S9 Table (Separate file)**

743 The raw data of Fig. 7.

744

745 **S10 Table (Separate file)**

746 The raw data of S2 Fig. (Pearson correlation coefficients for the result with no cWFs).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

747

748 **S11 Table (Separate file)**

749 The raw data of S2 Fig. (Pearson correlation coefficients for the result with cWFs).

750

751 **S12 Table (Separate file)**

752 Deconvolution of the aging-mouse whole-organ RNA-seq data. The results omitted due to the lack of
753 the significant improvement by the cWFs in the reconstitution are highlighted in pink. The results used
754 in the analyses shown in Figs. 7, 8, 9 (the pairwise comparisons between the two aging-stages show
755 $p<0.01$ by U-test) are highlighted in light-green.

756

757 **S13 Table (Separate file)**

758 The raw data of Fig. 8.

759

760 **S14 Table (Separate file)**

761 The raw data of Fig. 9 (sheet 1) & Fig. 10 (sheet 2).

762

763 Acknowledgements

764 We thank K. Sugisaka, R. Takahashi, T. Ninomiya, R. Kitaura, R. Ishikawa, S. Taniyama for
765 their administrative and laboratory management assistance. We are also grateful to the
766 members of Karydo TherapeutiX, Inc. and Sato laboratory at ATR for advice and discussion
767 throughout the course of this work.

768

769 **Funding:** This work was funded by Innovative Science and Technology Initiative for Security
770 Grant Number JPJ004596 ATLA Japan (T.N.S.), JST ERATO Grant Number JPMJER1303
771 Japan (T.N.S.), Nakatani Foundation (T.N.S.).

772

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Author contributions: Conceptualization: TNS, Methodology: KT, SK, Investigation: TNS, KT, SK, Visualization: KT, Supervision: TNS, Writing-original draft: TNS, KT, Writing-review & editing: TNS, KT, SK.

Completing interests: T.N.S. is the inventor of the patents filed on the cell type-specific weighting-factors and holds shares of Karydo TherapeutiX, Inc., which owns the right to this patent. K.T. and S.K. are also employed by Karydo TherapeutiX, Inc. Correspondence and requests for materials should be addressed to T.N.S. (island1005@gmail.com).

Data and materials availability: The public datasets are as indicated and the raw data as supplementary tables and the code for our algorithm are available at https://github.com/k-teji/VscRNAseq.

References

1. Aldridge S, Teichmann SA. Single cell transcriptomics comes of age. Nat Commun. 2020;11(1):4307. Epub 20200827. doi: 10.1038/s41467-020-18158-5. PubMed PMID: 32855414; PubMed Central PMCID: PMCPMC7453005.

2. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting global gene expression analysis. Cell. 2012;151(3):476-82. doi: 10.1016/j.cell.2012.10.012. PubMed PMID: 23101621; PubMed Central PMCID: PMCPMC3505597.

3. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014;24(3):496-510. Epub 2013/12/05. doi: 10.1101/gr.161034.113. PubMed PMID: 24299736; PubMed Central PMCID: PMCPMC3941114.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

797 4. Coate JE, Doyle JJ. Variation in transcriptome size: are we getting the message?
798 Chromosoma. 2015;124(1):27-43. Epub 20141126. doi: 10.1007/s00412-014-0496-3. PubMed
799 PMID: 25421950.
800 5. Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-
801 sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75.
802 Epub 20170818. doi: 10.1186/s13073-017-0467-4. PubMed PMID: 28821273; PubMed
803 Central PMCID: PMCPMC5561556.
804 6. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and
805 bioinformatics pipelines. Exp Mol Med. 2018;50(8):1-14. Epub 20180807. doi:
806 10.1038/s12276-018-0071-8. PubMed PMID: 30089861; PubMed Central PMCID:
807 PMCPMC6082860.
808 7. Adil A, Kumar V, Jan AT, Asger M. Single-Cell Transcriptomics: Current Methods
809 and Challenges in Data Acquisition and Analysis. Front Neurosci. 2021;15:591122. Epub
810 20210422. doi: 10.3389/fnins.2021.591122. PubMed PMID: 33967674; PubMed Central
811 PMCID: PMCPMC8100238.
812 8. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to
813 spatial transcriptomics for biomedical research. Genome Med. 2022;14(1):68. Epub 20220627.
814 doi: 10.1186/s13073-022-01075-1. PubMed PMID: 35761361; PubMed Central PMCID:
815 PMCPMC9238181.
816 9. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-
817 cell and spatial multi-omics. Nature reviews Genetics. 2023:1-22. Epub 20230302. doi:
818 10.1038/s41576-023-00580-2. PubMed PMID: 36864178; PubMed Central PMCID:
819 PMCPMC9979144.
820 10. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of
821 cellular mixtures based on linearity of transcriptional signatures. Nat Commun.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

822 2019;10(1):2209. Epub 2019/05/19. doi: 10.1038/s41467-019-09990-5. PubMed PMID:
823 31101809; PubMed Central PMCID: PMCPMC6525259.
824 11. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous
825 enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife.
826 2017;6. Epub 20171113. doi: 10.7554/eLife.26476. PubMed PMID: 29130882; PubMed
827 Central PMCID: PMCPMC5718706.
828 12. Coate JE, Doyle JJ. Quantifying whole transcriptome size, a prerequisite for
829 understanding transcriptome evolution across species: an example from a plant allopolyploid.
830 Genome Biol Evol. 2010;2:534-46. Epub 20100705. doi: 10.1093/gbe/evq038. PubMed
831 PMID: 20671102; PubMed Central PMCID: PMCPMC2997557.
832 13. Jonasson E, Andersson L, Dolatabadi S, Ghannoum S, Aman P, Stahlberg A. Total
833 mRNA Quantification in Single Cells: Sarcoma Cell Heterogeneity. Cells. 2020;9(3). Epub
834 20200319. doi: 10.3390/cells9030759. PubMed PMID: 32204559; PubMed Central PMCID:
835 PMCPMC7140709.
836 14. Raj A, Rifkin SA, Andersen E, van Oudenaarden A. Variability in gene expression
837 underlies incomplete penetrance. Nature. 2010;463(7283):913-8. doi: 10.1038/nature08781.
838 PubMed PMID: 20164922; PubMed Central PMCID: PMC2836165.
839 15. Schmidt EE, Schibler U. Cell size regulation, a mechanism that controls cellular RNA
840 accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-
841 Y and the liver-enriched transcription factor DBP. The Journal of cell biology.
842 1995;128(4):467-83. Epub 1995/02/01. doi: 10.1083/jcb.128.4.467. PubMed PMID: 7532171;
843 PubMed Central PMCID: PMCPMC2199888.
844 16. Zhurinsky J, Leonhard K, Watt S, Marguerat S, Bahler J, Nurse P. A coordinated global
845 control over cellular transcription. Current biology : CB. 2010;20(22):2010-5. Epub 20101021.
846 doi: 10.1016/j.cub.2010.10.002. PubMed PMID: 20970341.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

847 17. Marguerat S, Bahler J. Coordinating genome expression with cell size. Trends in
848 genetics : TIG. 2012;28(11):560-5. Epub 20120802. doi: 10.1016/j.tig.2012.07.003. PubMed
849 PMID: 22863032.
850 18. Miettinen TP, Pessa HK, Caldez MJ, Fuhrer T, Diril MK, Sauer U, et al. Identification
851 of transcriptional and metabolic programs related to mammalian cell size. Current biology :
852 CB. 2014;24(6):598-608. Epub 20140306. doi: 10.1016/j.cub.2014.01.071. PubMed PMID:
853 24613310; PubMed Central PMCID: PMCPMC3991852.
854 19. Lin CY, Loven J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional
855 amplification in tumor cells with elevated c-Myc. Cell. 2012;151(1):56-67. doi:
856 10.1016/j.cell.2012.08.026. PubMed PMID: 23021215; PubMed Central PMCID:
857 PMCPMC3462372.
858 20. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc is a universal amplifier
859 of expressed genes in lymphocytes and embryonic stem cells. Cell. 2012;151(1):68-79. doi:
860 10.1016/j.cell.2012.08.033. PubMed PMID: 23021216; PubMed Central PMCID:
861 PMCPMC3471363.
862 21. Li Y, Wang H, Muffat J, Cheng AW, Orlando DA, Loven J, et al. Global transcriptional
863 and translational repression in human-embryonic-stem-cell-derived Rett syndrome neurons.
864 Cell stem cell. 2013;13(4):446-58. doi: 10.1016/j.stem.2013.09.001. PubMed PMID:
865 24094325; PubMed Central PMCID: PMCPMC4053296.
866 22. Chandler MG, Pritchard RH. The effect of gene concentration and relative gene dosage
867 on gene output in Escherichia coli. Mol Gen Genet. 1975;138(2):127-41. doi:
868 10.1007/BF02428117. PubMed PMID: 1105148.
869 23. Hu Z, Chen K, Xia Z, Chavez M, Pal S, Seol JH, et al. Nucleosome loss leads to global
870 transcriptional up-regulation and genomic instability during yeast aging. Genes Dev.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2014;28(4):396-408. doi: 10.1101/gad.233221.113. PubMed PMID: 24532716; PubMed Central PMCID: PMCPMC3937517.

24. Cao S, Wang JR, Ji S, Yang P, Dai Y, Guo S, et al. Estimation of tumor cell total mRNA expression in 15 cancer types predicts disease progression. Nature biotechnology. 2022;40(11):1624-33. Epub 20220613. doi: 10.1038/s41587-022-01342-x. PubMed PMID: 35697807; PubMed Central PMCID: PMCPMC9646498.

25. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan GC. Accurate estimation of cell-type composition from gene expression data. Nat Commun. 2019;10(1):2975. Epub 2019/07/07. doi: 10.1038/s41467-019-10802-z. PubMed PMID: 31278265; PubMed Central PMCID: PMCPMC6611906.

26. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1):380. Epub 2019/01/24. doi: 10.1038/s41467-018-08023-x. PubMed PMID: 30670690; PubMed Central PMCID: PMCPMC6342984.

27. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Correction to: Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):50. Epub 20190729. doi: 10.1186/s13073-019-0655-5. PubMed PMID: 31358023; PubMed Central PMCID: PMCPMC6661746.

28. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):34. Epub 20190524. doi: 10.1186/s13073-019-0638-6. PubMed PMID: 31126321; PubMed Central PMCID: PMCPMC6534875.

29. Schaum N, Lehallier B, Hahn O, Palovics R, Hosseinzadeh S, Lee SE, et al. Ageing hallmarks exhibit organ-specific temporal signatures. Nature. 2020;583(7817):596-602. Epub

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

896 20200715. doi: 10.1038/s41586-020-2499-y. PubMed PMID: 32669715; PubMed Central
897 PMCID: PMCPMC7757734.
898 30. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martinez-Colon GJ, McKechnie JL, et al. A
899 single-cell atlas of the peripheral immune response to severe COVID-19. medRxiv. 2020. Epub
900 20200423. doi: 10.1101/2020.04.17.20069930. PubMed PMID: 32511639; PubMed Central
901 PMCID: PMCPMC7276995.
902 31. Tabula Muris C. A single-cell transcriptomic atlas characterizes ageing tissues in the
903 mouse. Nature. 2020;583(7817):590-5. Epub 20200715. doi: 10.1038/s41586-020-2496-1.
904 PubMed PMID: 32669714; PubMed Central PMCID: PMCPMC8240505.
905 32. Dutta S, Sengupta P. Men and mice: Relating their ages. Life Sci. 2016;152:244-8.
906 Epub 20151024. doi: 10.1016/j.lfs.2015.10.025. PubMed PMID: 26596563.
907 33. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient
908 Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems.
909 2017;30.
910 34. Kozawa S, Sagawa F, Endo S, De Almeida GM, Mitsuishi Y, Sato TN. Predicting
911 Human Clinical Outcomes Using Mouse Multi-Organ Transcriptome. iScience.
912 2020;23(2):100791. Epub 20200109. doi: 10.1016/j.isci.2019.100791. PubMed PMID:
913 31928967; PubMed Central PMCID: PMCPMC7033637.
914 35. Kozawa S, Ueda R, Urayama K, Sagawa F, Endo S, Shiizaki K, et al. The Body-wide
915 Transcriptome Landscape of Disease Models. iScience. 2018;2:238-68. Epub 2018/11/15. doi:
916 10.1016/j.isci.2018.03.014. PubMed PMID: 30428375; PubMed Central PMCID:
917 PMCPMC6135982.
918 36. Tabula Muris C, Overall c, Logistical c, Organ c, processing, Library p, et al. Single-
919 cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727):367-

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

920 72. Epub 2018/10/05. doi: 10.1038/s41586-018-0590-4. PubMed PMID: 30283141; PubMed
921 Central PMCID: PMCPMC6642641.
922 37. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the Mouse Cell Atlas by
923 Microwell-Seq. Cell. 2018;172(5):1091-107 e17. Epub 2018/02/24. doi:
924 10.1016/j.cell.2018.02.001. PubMed PMID: 29474909.
925 38. Joost S, Annusver K, Jacob T, Sun X, Dalessandri T, Sivan U, et al. The Molecular
926 Anatomy of Mouse Skin during Hair Growth and Rest. Cell stem cell. 2020;26(3):441-57 e7.
927 Epub 20200227. doi: 10.1016/j.stem.2020.01.012. PubMed PMID: 32109378.
928 39. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al.
929 Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87 e29. Epub
930 20210531. doi: 10.1016/j.cell.2021.04.048. PubMed PMID: 34062119; PubMed Central
931 PMCID: PMCPMC8238499.
932 40. Banerjee I, Fuseler JW, Price RL, Borg TK, Baudino TA. Determination of cell types
933 and numbers during cardiac development in the neonatal and adult rat and mouse. American
934 journal of physiology Heart and circulatory physiology. 2007;293(3):H1883-91. Epub
935 2007/07/03. doi: 10.1152/ajpheart.00514.2007. PubMed PMID: 17604329.
936 41. Bergmann O, Zdunek S, Felker A, Salehpour M, Alkass K, Bernard S, et al. Dynamics
937 of Cell Generation and Turnover in the Human Heart. Cell. 2015;161(7):1566-75. Epub
938 2015/06/16. doi: 10.1016/j.cell.2015.05.026. PubMed PMID: 26073943.
939 42. Nag AC. Study of non-muscle cells of the adult mammalian heart: a fine structural
940 analysis and distribution. Cytobios. 1980;28(109):41-61. Epub 1980/01/01. PubMed PMID:
941 7428441.
942 43. Pinto AR, Ilinykh A, Ivey MJ, Kuwabara JT, D'Antoni ML, Debuque R, et al.
943 Revisiting Cardiac Cellular Composition. Circulation research. 2016;118(3):400-9. Epub

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.05.188276; this version posted December 25, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

944 2015/12/05. doi: 10.1161/CIRCRESAHA.115.307778. PubMed PMID: 26635390; PubMed
945 Central PMCID: PMCPMC4744092.
946 44. Raulf A, Horder H, Tarnawski L, Geisen C, Ottersbach A, Roll W, et al. Transgenic
947 systems for unequivocal identification of cardiac myocyte nuclei and analysis of
948 cardiomyocyte cell cycle status. Basic Res Cardiol. 2015;110(3):33. Epub 2015/05/01. doi:
949 10.1007/s00395-015-0489-2. PubMed PMID: 25925989; PubMed Central PMCID:
950 PMCPMC4414935.
951 45. Walsh S, Ponten A, Fleischmann BK, Jovinge S. Cardiomyocyte cell cycle control and
952 growth estimation in vivo--an analysis based on cardiomyocyte nuclei. Cardiovascular research.
953 2010;86(3):365-73. Epub 2010/01/15. doi: 10.1093/cvr/cvq005. PubMed PMID: 20071355.
954 46. Zhou P, Pu WT. Recounting Cardiac Cellular Composition. Circulation research.
955 2016;118(3):368-70. Epub 2016/02/06. doi: 10.1161/CIRCRESAHA.116.308139. PubMed
956 PMID: 26846633; PubMed Central PMCID: PMCPMC4755297.
957 47. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation
958 Hyperparameter Optimization Framework. Proceedings of the ACM SIGKDD International
959 Conference on Knowledge Discovery and Data Mining. 2019:2623-31. doi:
960 10.1145/3292500.3330701.
961 48. Pedregosa Fabianpedregosa F, Michel V, Grisel Oliviergrisel O, Blondel M,
962 Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. Journal of Machine
963 Learning Research. 2011;12(85):2825-30.
964

5-1-2　森准教授グループ

1）…… 143
2）…… 158
3）…… 174
4）…… 186
5）…… 198
6）…… 221

1）Matsubayashi M, Sakaguchi YM, Sahara Y, Nanaura H, Kikuchi S, Asghari A, Bui L, Kobashigawa S, **Nakanishi M**, Nagata R, Matsui TK, Kashino G, Hasegawa M, Takasawa S, Eriguchi M, Tsuruya K, Nagamori S, Sugie K, Nakagawa T, Takasato M, Umetani M, **Mori E**. 27-Hydroxycholesterol regulates human SLC22A12 gene expression through estrogen receptor action. ***FASEB J*** 2021; 35(1): e21262. doi: 10.1096/fj.202002077R.

2）Iwasa N, Matsui TK, Iguchi N, Kinugawa K, Morikawa N, Sakaguchi YM, Shiota T, Kobashigawa S, **Nakanishi M**, Matsubayashi M, Nagata R, Kikuchi S, Tanaka T, Eura N, Kiriyama T, Izumi T, Saito K, Kataoka H, Saito Y, Kimura W, Wanaka A, Nishimura Y, **Mori E**, Sugie K. Gene Expression Profiles of Human Cerebral Organoids Identify PPAR Pathway and PKM2 as Key Markers for Oxygen-Glucose Deprivation and Reoxygenation. ***Front Cell Neurosci*** 2021; 15: 605030. doi: 10.3389/fncel.2021.605030.

3）Nanaura H, Kawamukai H, Fujiwara A, Uehara T, Aiba Y, **Nakanishi M**, Shiota T, Hibino M, Wiriyasermkul P, Kikuchi S, Nagata R, Matsubayashi M, Shinkai Y, Niwa T, Mannen T, Morikawa N, Iguchi N, Kiriyama T, Morishima K, Inoue R, Sugiyama M, Oda T, Kodera N, Toma-Fukai S, Sato M, Taguchi H, Nagamori S, Shoji O, Ishimori K, Matsumura H, Sugie K, Saio T, Yoshizawa T, **Mori E**. C9orf72-derived arginine-rich poly-dipeptides impede phase modifiers. ***Nat Commun*** 2021; 12(1): 5301. doi: 10.1038/s41467-021-25560-0.

4）Tomo Shiota, Riko Nagata, Sotaro Kikuchi, Hitoki Nanaura, Masaya Matsubayashi, Mari Nakanishi, Shinko Kobashigawa, Noriyoshi Isozyme, Takao Kiriyama, Kazuaki Nagayama, Kazuma Sugie, Yoshito Yamashiro and Eiichiro Mori (2022) C9orf72-Derived Proline:Arginine Poly-Dipeptides Modulate Cytoskeleton and Mechanical Stress Response. *Frontiers in Cell and Developmental Biology, doi.org/10.3389/fcell.2022.750829*

5）Iguchi N, Isozumi N, Hattori Y, Imamura T, So M, Nanaura H, Kiriyama T, Eura N, Yamaoka M, **Nakanishi M**, Mori M, Ohki S, Kumeta H, Koga H, Watabe M, Mabuchi T, Kanemura S, Okumura M, Yoshizawa T, Ota I, Suzuki N, Aoki M, Yamashiro Y, Saio T, Sugie K, **Mori E**. Zinc finger domains bind low-complexity domain polymers. ***bioRxiv*** 2023.10.29.564599.

6）Sakaguchi YM, Wiriyasermkul P, Matsubayashi M, Miyasaka M, Sakaguchi N, Sahara Y, Takasato M, Kinugawa K, Sugie K, Eriguchi M, Tsuruya K, Kuniyasu H, Nagamori S, **Mori E**. Identification of three distinct cell populations for urate excretion in human kidney. ***bioRxiv*** 2023.06.29.545826.

1)

Received: 7 September 2020 | Revised: 11 November 2020 | Accepted: 25 November 2020

DOI: 10.1096/fj.202002077R

RESEARCH ARTICLE

27-Hydroxycholesterol regulates human *SLC22A12* gene expression through estrogen receptor action

Masaya Matsubayashi[1] | Yoshihiko M. Sakaguchi[1] | Yoshiki Sahara[2,3] | Hitoki Nanaura[1,4] | Sotaro Kikuchi[1] | Arvand Asghari[5] | Linh Bui[5] | Shinko Kobashigawa[1] | Mari Nakanishi[1] | Riko Nagata[1] | Takeshi K. Matsui[1,4] | Genro Kashino[6] | Masatoshi Hasegawa[7] | Shin Takasawa[8] | Masahiro Eriguchi[9] | Kazuhiko Tsuruya[9] | Shushi Nagamori[10] | Kazuma Sugie[4] | Takahiko Nakagawa[1] | Minoru Takasato[2,3] | Michihisa Umetani[5,11] | Eiichiro Mori[1,12]

[1]Department of Future Basic Medicine, Nara Medical University, Nara, Japan

[2]RIKEN Center for Biosystems Dynamics Research, Kobe, Japan

[3]Graduate School of Biostudies, Kyoto University, Kyoto, Japan

[4]Department of Neurology, Nara Medical University, Kashihara, Japan

[5]Center for Nuclear Receptors and Cell Signaling, Department of Biology and Biochemistry, University of Houston, Houston, TX, USA

[6]Radioisotope Research Center, Nara Medical University, Kashihara, Japan

[7]Department of Radiation Oncology, Nara Medical University, Kashihara, Japan

[8]Department of Biochemistry, Nara Medical University, Kashihara, Japan

[9]Department of Nephrology, Nara Medical University, Kashihara, Japan

[10]Department of Collaborative Research, Nara Medical University, Nara, Japan

[11]HEALTH Research Institute, University of Houston, Houston, TX, USA

[12]V-iCliniX Laboratory, Nara Medical University, Kashihara, Japan

Correspondence
Michihisa Umetani, Center for Nuclear Receptors and Cell Signaling, HEALTH Research Institute, University of Houston, Houston, TX 77204-5056, USA.
Email: mumetani@uh.edu

Eiichiro Mori, Department of Future Basic Medicine, Nara Medical University, Kashihara, Nara 634-8521, Japan.
Email: emori@naramed-u.ac.jp

Abstract

The excretion and reabsorption of uric acid both to and from urine are tightly regulated by uric acid transporters. Metabolic syndrome conditions, such as obesity, hypercholesterolemia, and insulin resistance, are believed to regulate the expression of uric acid transporters and decrease the excretion of uric acid. However, the mechanisms driving cholesterol impacts on uric acid transporters have been unknown. Here, we show that cholesterol metabolite 27-hydroxycholesterol (27HC) upregulates the uric acid reabsorption transporter URAT1 encoded by *SLC22A12* via estrogen receptors

Abbreviations: 27HC, 27-hydroxycholesterol; ABCG2, ATP-binding cassette sub-family G member 2; ER, estrogen receptor; ERE, estrogen response element; GLUT9, glucose transporter 9; iPS, induced pluripotent stem cell; OAT4, organic anion transporter 4; OAT10, organic anion transporter 10; PT, proximal tubule; SERM, selective estrogen receptor modulator; snRNA-seq, single nucleus RNA sequencing; URAT1, urate transporter 1.

Masaya Matsubayashi and Yoshihiko M. Sakaguchi contributed equally to this work.

[Correction added on December 24, 2020, after first Online publication: Copyright has been updated to Online Open with the legal statement.]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2020 The Authors. *The FASEB Journal* published by Wiley Periodicals LLC on behalf of Federation of American Societies for Experimental Biology

Funding information
JSPS KAKENHI, Grant/Award Number: JP17H07031, JP20H03199, JP16K19836, JP19K08150, JP18H06202, JP19K21306, JP20K16583, JP19K16925 and JP19K23976; AMED Brain/MINDS Beyond, Grant/Award Number: JP20dm0307032; National Institutes of Health, Grant/Award Number: HL127037; UH HEALTH Research Institute; Takeda Science Foundation; Kanzawa Medical Research Foundation; Uehara Memorial Foundation; Nakatomi Foundation; Konica Minolta Science and Technology Foundation; Naito Foundation; MSD Life Science Foundation; Mochida Memorial Foundation for Medical and Pharmaceutical Research; SENSHIN Medical Research Foundation; Terumo Foundation for Life Sciences and Arts; Nara Kidney Disease Research Foundation; Novartis Research; Nara Medical University; Sumitomo Dainippon Pharma Research; Tokyo Biochemical Research Foundation

(ER). Transcriptional motif analysis showed that the *SLC22A12* gene promoter has more estrogen response elements (EREs) than other uric acid reabsorption transporters such as *SLC22A11* and *SLC22A13*, and 27HC-activated *SLC22A12* gene promoter via ER through EREs. Furthermore, 27HC increased *SLC22A12* gene expression in human kidney organoids. Our results suggest that in hypercholesterolemic conditions, elevated levels of 27HC derived from cholesterol induce URAT1/*SLC22A12* expression to increase uric acid reabsorption, and thereby, could increase serum uric acid levels.

KEYWORDS
27-hydroxycholesterol, estrogen receptor, transcriptional regulatory element, uric acid

1 | INTRODUCTION

Elevated levels of serum uric acid have been reported as a risk factor for various diseases such as gout, renal diseases, and cardiovascular diseases.[1-5] Increased serum uric acid levels also trigger vascular endothelial dysfunction, which leads to metabolic syndrome and cardiovascular diseases.[6-11] Obesity, insulin resistance, aging, and hormonal imbalance are risk factors for metabolic diseases,[12-15] of which obesity is associated with hypercholesterolemia [16,17] and hyperuricemia.[18,19] Such evidence suggests the existence of a correlation between uric acid levels and cholesterol metabolism. However, the underlying mechanism of this phenomenon is unknown.

There are two causes of hyperuricemia: One is a hepatic overproduction of uric acid, and the other is an under-excretion of uric acid from the kidney and small intestine. Overproduction of uric acid is caused by activated hepatic xanthine oxidoreductase. The excretion of uric acid is regulated by several uric acid transporters, such as urate transporter 1 (URAT1)/*SLC22A12*, ATP-binding cassette sub-family G member 2 (ABCG2)/*ABCG2*, and glucose transporter 9 (GLUT9)/*SLC2A9*.[20-24] Filtered uric acid is reabsorbed into the proximal tubular cells from urine by reabsorption transporters including URAT1 and is secreted into the blood through GLUT9.[24] Uric acid reabsorption transporters play an important role in regulating serum uric acid levels. The expression of Urat1/*Slc22a12* protein/gene is increased in hyperuricemia or obesity in animal models.[25-27] Obesity has also been found to decrease the renal clearance of uric acid in humans.[19] These reports suggest that obesity and metabolic disorder affect the clearance of uric acid by changing URAT1/*SLC22A12* expression. Nevertheless, there is limited information on the mechanism(s) by which URAT1/*SLC22A12* expression is regulated in metabolic disorders conditions.

Previously, we found that the cholesterol metabolite 27-hydroxycholesterol (27HC) acts as an endogenous selective estrogen receptor (ER) modulator, or SERM, and promotes atherosclerotic lesion development in an ER-depending manner.[28] 27HC acts as an agonist or antagonist of ER in a tissue- and cell type-dependent manner.[29,30] In metabolic diseases such as hypercholesterolemia, the concentration of 27HC is increased.[31] Serum 27HC and cholesterol levels are positively correlated with each other, and serum 27HC levels are elevated in humans after 30 years of age.[30-33] Further, studies have observed that circulating 27HC levels are lower in premenopausal females than in males, but these levels are increased after menopause.[30,32] The profile of 27HC resembles that of serum uric acid. Serum uric acid levels are increased with aging and after menopause in females.[34-36] Males also tend to have higher uric acid levels than females.[37,38] These similarities hint at a potential correlation between 27HC and uric acid.

In this study, we investigated how the cholesterol metabolite 27HC impacts uric acid transporter URAT1/*SLC22A12* using evolutionarily conserved analysis, single-nucleus RNA-sequencing (snRNA-seq) analysis, cell culture assays, mouse kidney histology, and human kidney organoid assays.

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2 | MATERIALS AND METHODS

2.1 | Alignment of uric acid transporters

Amino acid sequences of human uric acid transporters (URAT1, OAT4, OAT10, ABCG2, NPT1, NPT4, GLUT9, OAT1, and OAT3) were obtained from the NCBI database. Multiple sequence alignment was performed by the MUSCLE program. Based on the alignment, an unrooted phylogenetic tree was generated by the Clustal program using the Neighbor-Joining (NJ) method[39,40] and plotted using DRAWTREE in the PHYLIP software suite.[41] For gene promoter regions of human uric acid transporters, the 1000 bases upstream sequences were identified from the UCSC Genome Browser (http://genome.ucsc.edu/). Multiple alignments of gene promoter sequences and the phylogenic tree were generated by the Clustal program using the NJ method. The alignment of gene promoter sequences was drawn by ES print 3 (http://espript.ibcp.fr/ESPript/ESPript/index.php). We identified sequence similarities depiction parameters as more than 0.75 at a global score.

2.2 | Transcription factor-binding site search and estrogen response element scores

Gene promoter sequences were obtained as described above. We searched potential ER-binding sites in each gene promoter sequence using the TFBIND website (http://tfbind.hgc.jp/). Based on the results corresponding to the consensus sequence, we searched the start and end position of each estrogen response element (ERE) sequence in the promoter sequence. The sequence Logo was made by Web Logo 3 (http://weblogo.threeplusone.com/). ERE score was defined as the calculated similarity score by TFBIND. ERE values were represented as a multiplication of the 10-fold ERE scores.

2.3 | Single-nucleus RNA-sequencing analysis

Single-nucleus transcriptome analysis was performed using a data set of single nucleus RNA-seq from a human adult kidney (GEO: GSE118184).[42] The digital gene expression matrix was processed for quality control, normalization, linear dimensional reduction, and unsupervised clustering of single-cell using Seurat (https://satijalab.org/seurat/) plug-in of R software (https://www.r-project.org/) according to the tutorial in the developer's website (https://satijalab.org/seurat/vignettes.html). Cells were filtered out as low-quality cells based on unique feature counts and mitochondrial counts. After the filtration, we applied the function "Sctransform" normalization to remove confounding sources of variation, mitochondrial mapping percentage, the total number of molecules, and cell cycle heterogeneity[43,44] in order to mitigate confounding effects. Principal component (PC) analysis was conducted and the most significant PCs of the data set were selected for two-dimensional Uniform Manifold Approximation and Projection (UMAP). Unsupervised clustering was performed by functions "FindClusters" and "RunUMAP" at a resolution level of 0.8. Single-cell RNA-seq measures transcripts from both cytoplasm and nucleus, whereas snRNA-seq measures only nuclear transcripts. Nuclei contain only a fraction of total cellular RNA. Although nuclear and cytoplasmic mRNAs are highly correlated,[45] some protein-coding mRNAs are retained in the nucleus.[46] Despite these differences, snRNA-seq data sets predict cell types comparably with high concordance.[47,48]

2.4 | Annotation of the clusters and cell-type identification

The cluster names used for the anatomical structures of the kidney are podocyte, PT: proximal tubule; LH (DL): the loop of Henle (descending loop); LH (AL): the loop of Henle (ascending loop); DCT: distal convoluted tubule; CNT: connecting tubule; PC: principal cell; ICA: intercalated cell type A; ICB: intercalated cell type B; EDC: endothelial cell; and U: undefined. They were identified based on the maker gene expression described in a previous report.[42]

2.5 | Co-expression analysis of three uric acid reabsorption transporters

Co-expression levels of three uric acid reabsorption transporters were shown in scatterplots of these transporters. The scatterplots of two gene expressions after extracting the cells that express either gene were also created using a Seurat function, "FeatureScatter," which is a plug-in of R software (https://www.r-project.org/). The axes represent the expression levels of each gene from count data after log-normalization.

2.6 | Analysis tissue *ESR1* and *ESR2* expression from RNA-seq database

We analyzed *ESR1* and *ESR2* gene expression in various tissues, as well as tested for sex differences in kidney *ESR1* and *ESR2* expression. The expression data referred to GTEx RNA-seq data (https://www.proteinatlas.org/ENSG00000091831-ESR1/tissue and https://www.proteinatlas.org/ENSG00000091831-ESR1/tissue/kidney). We statistically analyzed all male and female kidney sample data in GTEx RNA-seq.

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.7 | *SLC22A12* gene promoter cloning

All plasmid constructs were designed by SnapGene (GSL Biotech, Chicago, Illinois, USA). The human genomic DNA was isolated from U2OS cells using the QuickExtract™ DNA Extraction Solution (Lucigen, Wisconsin, USA). From the gDNA, the *SLC22A12* gene promoter was amplified using KOD One PCR Master Mix (TOYOBO, Osaka, Japan) or PrimeSTAR® Max DNA Polymerase (TAKARA BIO, Shiga, Japan) with primer F: gctcgctagcctcgaaatcagttccaaagagcctctaaagaag, and primer R: ccggattgccaagctagagaggcagctgctcca. After the PCR product was confirmed by the agarose gel electrophoresis, it was purified using QIAquick Gel Extraction Kit (Qiagen) and cloned into a pGL4.17 [luc2/Neo] Vector (Promega, Madison, Wisconsin, USA) using Xho I and Hind III restriction enzyme sites. The ERE sequences of *SLC22A12* gene promoter were amplified to get mutants (TGACC was changed to CAGTA) using PrimeSTAR Mutagenesis Basal Kit (TAKARA BIO, Shiga, Japan) with the following primers: ERE#1 primer F: aagcaggattcagtacaagggctcgcagtgcgt; primer R: cgagcccttgtactgaatcctgcttaaaagcca; ERE#2 primer F:ccagacactgcagtatgagaggccatagctgag; primer R: tggcctctcatactgcagtgtctggcctggaac; ERE#3 primer F: cagctgccagcagtacaagcccacacagagact; and primer R: tgtgggcttgtactgctggcagctgcccagccc. The constructs were amplified, and sequences were confirmed.

2.8 | Cell culture and generation of stable cell line

The human hepatocellular carcinoma cell line HepG2 was purchased from Cellular Engineering Technologies, Inc Cells were maintained in RPMI1640 medium (Nacalai Tesque, Kyoto, Japan) supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 μg/ml streptomycin. The cells were incubated under a humidified atmosphere containing 5% CO_2 at 37°C. To generate stably expressed cell lines, cells were seeded in six-well plates at 1.0×10^5 cells per well and introduced to linearized luciferase expression plasmid fused with *SLC22A12* gene promoter using Lipofectamine 2000 (Invitrogen, Waltham, Massachusetts, USA). At 24 hours after transfection, the medium was changed to G418 (800 μg/ml)-containing medium. We selected stably expressed cells by removing G418-sensitive cells with changing media every day for two weeks.

2.9 | Quantitation of promoter activity by luciferase assay

The luciferase activity was measured in HepG2 cells that stably express intact and ERE mutant *SLC22A12* gene promoter-regulated luciferase proteins. This was done using One-Glo EX Luciferase Assay System (Promega) and according to the manufacturer's instructions. Cells in ninety-six-well plates were treated with 17β-estradiol, estrone, 27HC, 3β-hydroxy-5-cholestenoic acid, or ICI 182,780 for 24 hours at indicated concentrations. Then, One-Glo EX reagent (80 μl/well) was added to the cells, and luciferase activity was measured with a SpectraMax. All experiments were performed in triplicate.

2.10 | Kidney immunohistochemistry

Kidney samples were collected from 12-month-old *Cyp7b1*$^{+/+}$ and *Cyp7b1*$^{-/-}$ mice on the C57BL/6 background. Collected samples were fixed in 4% formalin for 48 hours at 4°C, processed, paraffin-embedded, and cut into 6 μm-thick sections. After rehydration, the sections were treated in the citrate buffer (pH 6.0) at 95°C for 10 minutes for antigen retrieval, followed by 10% goat serum (Life Technologies) at room temperature for 1 hour. Then, the sections were blocked using an avidin/biotin blocking kit (Vector Laboratories, Burlingame, California, USA) for 15 minutes. Thereafter, an anti-Urat1 antibody (Millipore Sigma, HPA024575, 1:500) was added, and the sections were incubated overnight. After washing the sections, they were blocked again using 3% hydrogen peroxidase (ThermoFisher Scientific) and incubated with biotinylated goat-anti-rabbit antibody (Vector Laboratories, BA-1000, 1:200) for 1 hour. The sections were visualized using an avidin-biotin complex kit (ABC kit, Vector Laboratories) and DAB Quanto chromogen (ThermoFisher Scientific, Waltham, Massachusetts, USA), followed by counterstaining with Harris hematoxylin (ThermoFisher Scientific, Waltham, Massachusetts, USA). Image analysis was performed using Image J/Fiji. Proximal tubules were detected, and the maximum intensity of the brush border membrane in the proximal tubular cells was measured (Figure S5). About 230-240 points per image were measured in a blind manner. Statistical analysis was performed using the average of these intensities per image.

2.11 | Kidney organoid culture

Human kidney organoids were generated from induced pluripotent stem (iPS) cell line CRL1502 according to the procedures reported in our previous study[49] with some modification. Human iPS cells were treated with 8 μM CHIR99021 in a "basal medium," APEL2 (STEMCELL Technologies, Vancouver, Canada) that was supplemented with 1% Protein Free Hybridoma Medium II (PFHM II, GIBCO), for 5 days, and with FGF9 (200 ng/ml) and heparin (1 μg/ml) for an additional 2 days. Then, the cells were

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

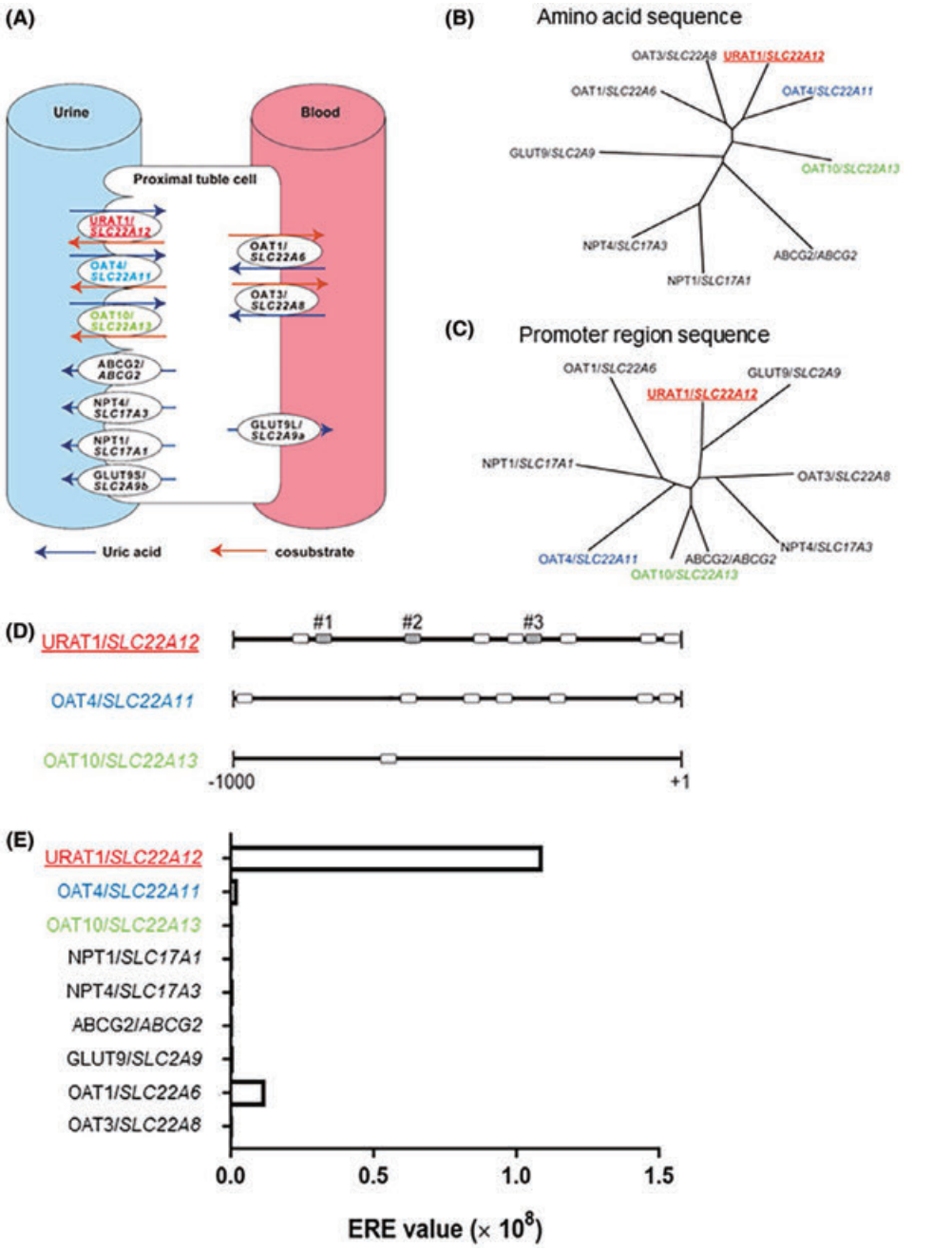


FIGURE 1 Phylogenetic trees of uric acid reabsorption transporters. A, Schema of uric acid transporters in kidney proximal tubular cells. B and C, Phylogenetic trees of the uric acid transporters (B) amino acid sequences and (C) gene promoter sequences. Three reabsorption transporters are indicated in color. D, A diagram of the gene promoter regions. White boxes show potential ERE sequences, and three grey boxes of *SLC22A12*, #1-#3, show the ERE sequences having ERE scores more than 0.8. E, A bar plot showing the ERE values of uric acid transporter gene promoter regions

collected, dissociated into 2.5 × 10^5 single cells, and spun down at 400 ×*g* for 2 minutes to form a pellet. The pellets were transferred onto a Transwell of 0.4 μm pore membrane (#3450, Corning, USA) to culture in liquid-air interfaces. The cells were cultured with the following factors in the basal media in the following order: 10 μM CHIR99021 for 1 hour, FGF9 (200 ng/ml) and heparin (1 μg/ml) for 5 days, and basal medium only for 13-15 days. Organoids were treated with 10 μM 27HC for the final 2 days before sampling (day 25-27 of differentiation) in order to examine

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the effect of 27HC on the *SLC22A12* expression but not on the organoid development.

2.12 | Quantitative RT-PCR

Total RNA was extracted from organoids using Nucleo Spin (MACHEREY NAGEL, Germany) and cDNA was synthesized from 500 ng of total RNA using PrimeScript™ RT Master Mix (TAKARA BIO, Shiga, Japan). Quantitative reverse-transcription PCR (qRT-PCR) was performed with SsoAdvanced Universal SYBR Green Supermix (Bio-Rad Laboratories, Hercules, California, USA) and specific primers for qRT-PCR. The reaction was run in triplicate, and the transcription of each gene was normalized to the mean values of *β-actin* and analyzed using the ΔΔCT method by StepOne Software Ver2.3. The primer sequences for qRT-PCR were designed by Bio-Rad (*SLC22A12*; qHsaCID0011483, *β-actin*; qHsaCED0036269). The data were collected from six independent experiments.

2.13 | Statistics and data analysis

Results of the luciferase assay and qRT-PCR were represented by the ratio against mean control values. All the values obtained represent means ± SEM. Student's *t* test, one-way ANOVA followed by Dunnett's test, or two-way ANOVA followed by Sidak's multiple comparisons test were performed to determine the significance among groups (GraphPad Prism7). A *P* value of less than .05 was considered statistically significant.

3 | RESULTS

3.1 | *SLC22A12* gene promoter is differentially regulated from *SLC22A11* and *SLC22A13* gene promoter through estrogen response elements

As shown in the model of transcellular uric acid transport in the proximal tubular (PT) cells (Figure 1A), several uric acid transporters regulate uric acid flow from/to the PT cell.[50] To investigate the molecular difference among uric acid transporters, we analyzed the amino acid sequences of these uric acid transporters with evolutionarily conserved analysis. The amino acid sequences of uric acid reabsorption transporters, URAT1 encoded by *SLC22A12*, OAT4 encoded by *SLC22A11*, and OAT10 encoded by *SLC22A13* showed a closer evolutionary relationship to each other (Figure 1B and S1A, B), suggesting that these transporters share a similar function.

Although the co-expression profile at single-cell resolutions and the difference in transcriptional regulation of uric acid transporters are important to understand the uric acid flow, these are largely unknown. To evaluate the expression of uric acid reabsorption transporters at single-cell resolutions, we analyzed publicly available snRNA-seq data of human adult kidneys. Each cell type was clearly separated (as shown in Figure S2A and gene markers of cell types shown in Figure S2B), and we focused on the distribution of uric acid transporters in the PT region in the snRNA-seq analysis (Figure S2A-C). The co-expression plots showed that *SLC22A12* was expressed independently of *SLC22A11* or *SLC22A13* genes (Figure S2D-F). In addition, the gene promoter sequence of *SLC22A12* shares a distant evolutionary relationship with *SLC22A11* or *SLC22A13* (Figure 1C and S3A). These analyses suggest that the *SLC22A12* gene is expressed independently of other uric acid reabsorption transporters under specific transcriptional regulation.

To elucidate the key factors for the unique transcriptional regulation of uric acid transporters, we further analyzed their gene promoter regions. Since the elevated uric acid level is associated with the metabolic condition and sex hormone status,[10,51] we focused on the binding motifs of metabolic or sex hormone-regulated transcription factors on the gene promoter regions. The *SLC22A12* gene promoter contains nine potential sequences of estrogen response elements (EREs) and has the highest ERE value among uric acid transporter genes (Figure 1D, E, and S3B-J). These analyses suggest that *SLC22A12* is most strongly regulated by ER among uric acid transporters.

Since the ERE activates the promoter activity through ER binding, we investigated the expression of ER in the human kidney. ERα/*ESR1* and ERβ/*ESR2* are expressed in the kidney, although the expression levels were ten to a hundred times lower than in tissues with the highest expression, such as the ovary and cervix (Figure S4A). In addition, there was no difference in the *ESR1* and *ESR2* expression between males and females (Figure S4B). These results suggest that the kidney ERs bind to EREs on the *SLC22A12* gene promoter regardless of sex and that the promoter activity is upregulated.

3.2 | Human *SLC22A12* gene promoter activity is upregulated by 27HC

To investigate whether ER regulates *SLC22A12* transcription activity, we created a plasmid construct that includes the *SLC22A12* gene promoter region fused to a luciferase expression vector (Figure 2A) and performed a luciferase assay using this construct. Our study aimed to examine the effect of ligands for ER in the kidney in a physiological condition where the ER expression is low (Figure S4A). This provided

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

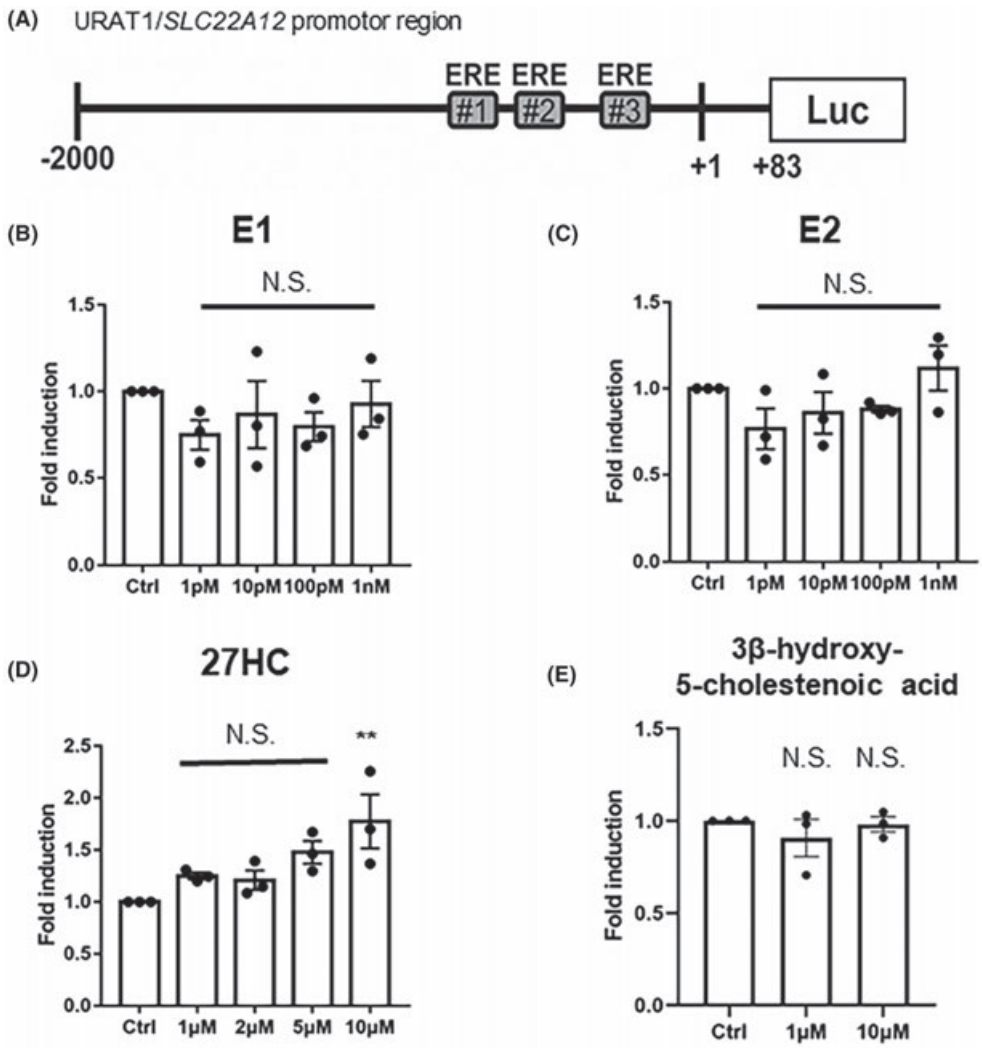


FIGURE 2 *SLC22A12* gene promoter activities. A, Schematic representation of a human URAT1/*SLC22A12* gene promoter-luciferase reporter construct. ERE#1, #2, and #3 represent the same site shown in Figure 1D. B-E, HepG2 cells with stable expression of *SLC22A12* gene promoter-driven luciferase proteins were treated with E1 (B) or E2 (C) at the indicated concentration for 24 hours. D, The effect of 27HC on the *SLC22A12* gene promoter activity in HepG2 cells. E, The effect of 3β-hydroxy-5-cholestenoic acid on the URAT1/*SLC22A12* promoter activity in HepG2 cells. The results were analyzed using one-way ANOVA followed by Dunnett's tests. ** $P < .01$ vs. control (ctrl). NS, nonsignificant. All experiments in triplicate were performed independently three times

the basis for our choice of HepG2 cells that have relatively similar levels of ER expression as in the kidney.

First, we examined whether estrogens, which are major ligands for ER, affected the gene promoter. There are different types of estrogens. 17β-estradiol (E2) is the most potent among the endogenous estrogens[52] and significantly decreases after menopause,[53] whereas estrone (E1) becomes the dominant circulating estrogen in postmenopausal females.[53] E1 and E2 did not change the *SLC22A12* gene promoter at physiological concentrations (E1: $P = .3013$, E2: $P = .0592$; Figure 2B,C). These results suggest that estrogens do not affect *SLC22A12* gene transcription given the physiological conditions in humans.

There is a correlation between serum uric acid levels and metabolic conditions.[12-15] Cholesterol metabolite 27HC acts as an endogenous ER ligand and regulates transcriptional activities through EREs.[33,54] Thus, to investigate whether metabolic disorders affect *SLC22A12* transcriptional activity, we examined the effect of 27HC on this gene promoter construct. Normal circulating levels of 27HC are typically measured in a range of 0.15-0.73 μM,[33,54] and the concentration of 27HC reaches millimolar levels in metabolic diseases such as hypercholesterolemia.[31] Gene promoter activity of SLC22A12 was increased by 27HC in a dose-dependent manner and induced significantly at 10 μM ($P = .007$; Figure 2D). Our result suggests that 27HC causes the elevated expression of

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

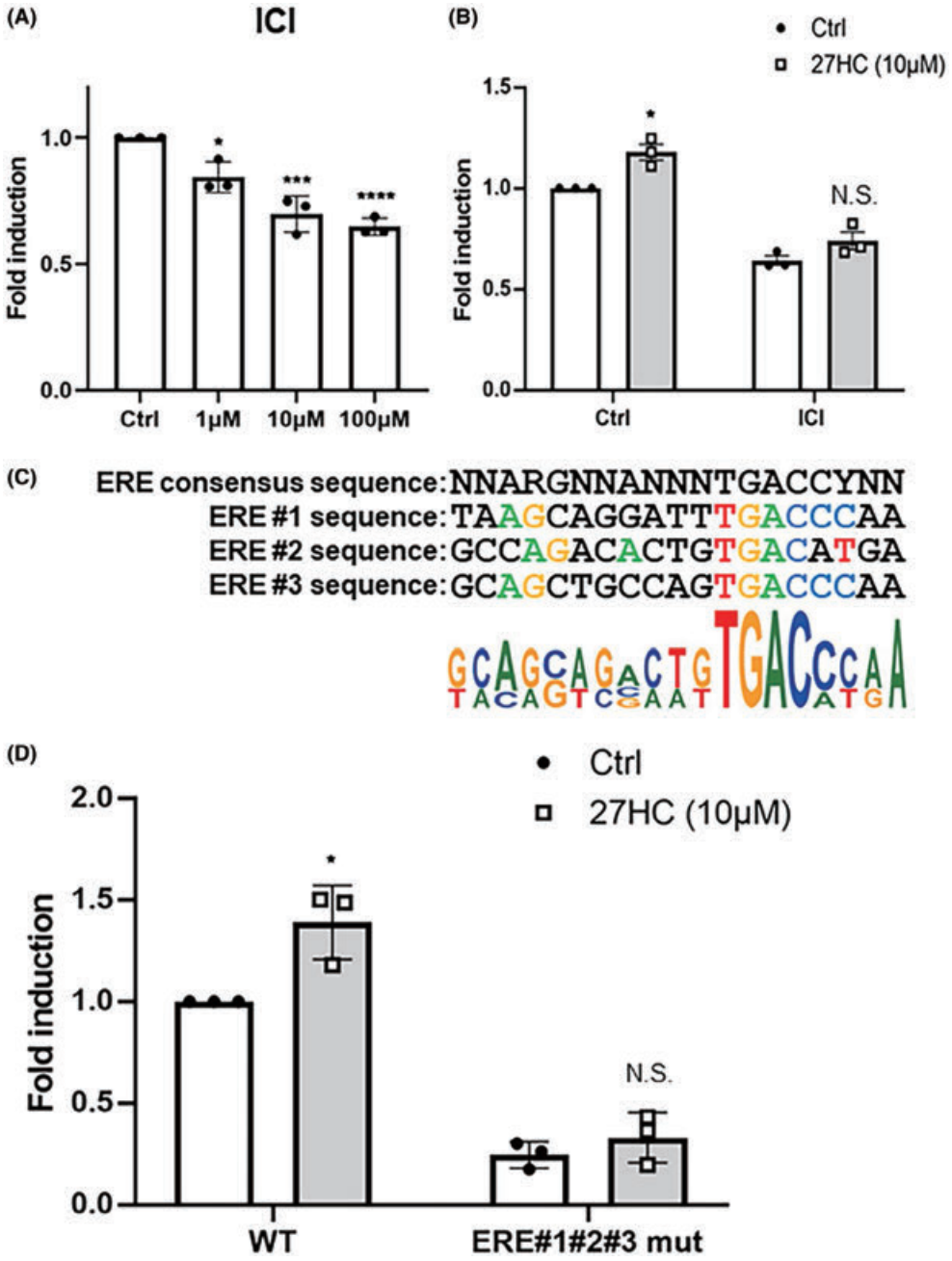


FIGURE 3 Involvement of ER and EREs in the *SLC22A12* gene promoter activity. A, HepG2 cells with stable expression of *SLC22A12* gene promoter-regulated luciferase proteins were treated with ICI 182,780. The results were analyzed using one-way ANOVA followed by Dunnett's multiple comparison tests. * $P < .05$, *** $P < .001$, **** $P < .0001$ vs. ctrl. B, HepG2 cells were treated with 27HC in the absence or presence of 100 μM ICI 182,780 for 24 hours. The results were analyzed using two-way ANOVA followed by Sidak's multiple comparisons tests. * $P < .05$ vs. vehicle control for 27HC. Open column; vehicle control, gray column; 27HC-treated. C, Consensus ERE sequence, potential ERE sequences on the SLC22A12 gene promoter, and sequence logo of the three EREs. The largest letter (TGAC) shows the conserved sequence among the sequences. D, The effect of 27HC on ERE mutant gene promoter activity. The two HepG2 cell lines that stably express intact or ERE mutant gene promoter-driven luciferase proteins were treated with 10 μM 27HC for 24 hours. The results were analyzed using two-way ANOVA followed by Sidak's multiple comparisons tests. * $P < .05$ compared with nontreatment of 27HC. Open column; vehicle control, gray column; 27HC-treated. All experiments in triplicate were performed independently three times

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the *SLC22A12* gene at the concentration in metabolic disease patients.

To examine whether the effect of 27HC was caused by 27HC itself or its metabolites, we further tested 3β-hydroxy-5-cholestenoic acid, a metabolite of 27HC. The treatment of 3β-hydroxy-5-cholestenoic acid did not increase *SLC22A12* gene promoter activity even in a high dose ($P = .5729$; Figure 2E). This result suggests that 27HC, rather than its metabolite, upregulates *SLC22A12* gene transcription.

3.3 | 27HC regulates human *SLC22A12* transcription via ER through EREs

To determine the impact of 27HC on *SLC22A12* gene promoter through ER, we treated cells with an ER antagonist, ICI 182,780.[55] ICI 182,780 alone decreased the *SLC22A12* gene promoter activity in a dose-dependent manner ($P < .0001$; Figure 3A), suggesting that this ER antagonist inhibits gene promoter basal activity through endogenous ER. Next, we treated cells with ICI 182,780 in the presence of 27HC. Since 27HC had a higher affinity for ERα than ICI 182,780,[54,56] we treated the higher concentration of ICI 182,780 (100 μM) rather than the concentration of 27HC (10 μM). The effect of 27HC was suppressed by a co-treatment with ICI 182,780 (Ctrl vs. 27HC, $P = .0217$, ICI vs. 27HC + ICI, $P = .3042$; Figure 3B). These results indicate that 27HC increases *SLC22A12* gene transcription via ER activity.

The *SLC22A12* promoter region has nine potential ERE sequences (Figure S3B), and we postulated that 27HC acted on the *SLC22A12* gene promoter through EREs. We created a mutant plasmid construct, in which three sequences with the highest ERE scores on the *SLC22A12* gene promoter region were mutated (Figure 3C). These ERE mutations significantly decreased the baseline of gene promoter activity ($P < .0001$; Figure 3D), indicating that ERE sequences play a major role in the *SLC22A12* transcription. ERE mutation also diminished the effect of 27HC on the *SLC22A12* gene promoter activity (ctrl vs. 27HC in intact ERE, $P = .0183$, ctrl vs. 27HC in ERE mutation, $P = .9495$, 27HC x ERE mutation interaction $P = .0492$; Figure 3D). These results suggest that 27HC regulates the *SLC22A12* gene promoter through ERE.

3.4 | 27HC increases Urat1 expression in the mouse kidneys

To evaluate the role of 27HC on Urat1 encoded by *Slc22a12* expression in the mouse kidneys, we applied immunohistochemistry using kidney samples from *Cyp7b1*$^{+/+}$ and *Cyp7b1*$^{-/-}$ mice. In this mouse model, the deficiency of the 27HC-catabolizing enzyme Cyp7b1 results in elevated serum and tissue 27HC levels without affecting cholesterol and bile acid metabolism.[57] Urat1 protein expression was observed in the apical membrane of the PT region (Figure S5A). The expression levels were observed to be increased in male *Cyp7b1*$^{-/-}$ mice kidneys compared with male *Cyp7b1*$^{+/+}$ mice (Figure 4A, B and S5B), suggesting that elevated 27HC levels induce Urat1 expression in mice.

Given the impact of 27HC on Urat1 expression via ER, there might be sex differences on the effect of 27HC. When we compared the results between sexes, *Cyp7b1*$^{-/-}$ male mice had significantly elevated expression of Urat1 ($P = .0057$, *t* test; Figure 4A,B and S5B), whereas females did not show a significant change of Urat1 expression levels between *Cyp7b1*$^{+/+}$ and *Cyp7b1*$^{-/-}$ genotypes ($P = .0737$, t test; Figure 4A,B and S5B). These results suggest sex differences might potentially influence Urat1 expression.

3.5 | Extracellular 27HC increases *SLC22A12* gene expression in human kidney organoids

From the analysis of promoter activity (Figure 2D) and immunohistochemical analysis in mice kidneys (Figure 4), we postulated that 27HC impacts URAT1/*SLC22A12* in human kidneys. The ERE value of the human *SLC22A12* gene promoter was higher than that of mice (Figure 5A and S6), suggesting that the effect of 27HC on the *SLC22A12* gene promoter activity in humans is greater than that in mice. To estimate the potential for the effect of 27HC on URAT1/*SLC22A12* in human kidneys, we performed kidney organoid culture from human iPS cells and examined the effect of 27HC on the *SLC22A12* expression (Figure 5B). 27HC significantly increased *SLC22A12* mRNA expression ($P = .043$) (Figure 5C). These results indicate that 27HC increases *SLC22A12* expression in human kidneys.

4 | DISCUSSION

In this study, we demonstrated that 27HC, an endogenous SERM, regulates the expression of URAT1/*SLC22A12*. In the basal condition, URAT1/*SLC22A12*, OAT4/*SLC22A11*, and OAT10/*SLC22A13* are involved in the reabsorption of uric acid into blood (Figure 6A). In hypercholesterolemic conditions such as cardiovascular and metabolic diseases, elevated levels of 27HC induce URAT1/*SLC22A12* expression, which can increase uric acid reabsorption and circulating uric acid levels (Figure 6B). Our findings provide new information on the relationship between hyperuricemia and metabolic syndromes.

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

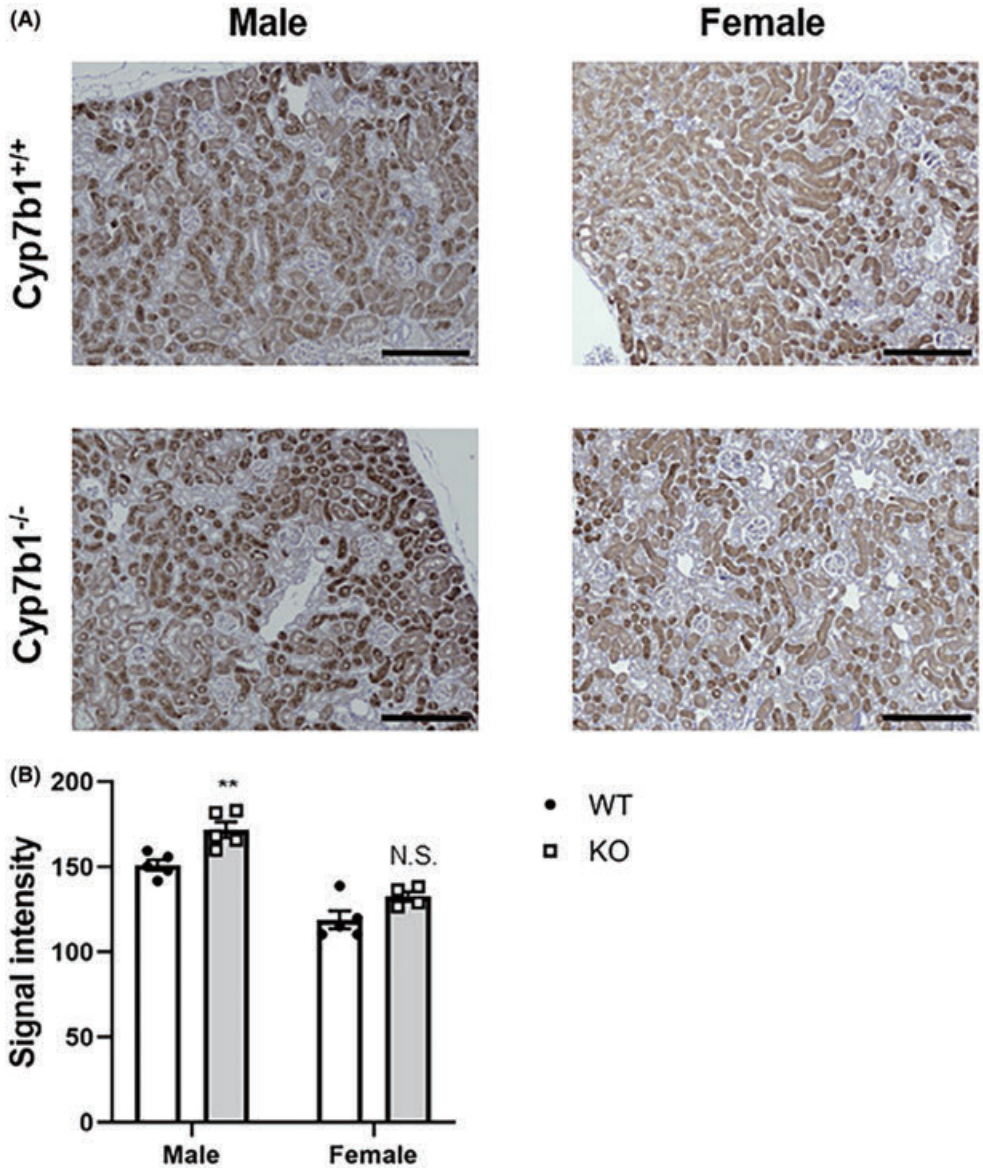


FIGURE 4 Immunohistochemical analysis of kidneys from *Cyp7b1*$^{+/+}$ and *Cyp7b1*$^{-/-}$ mice. A, Immunohistochemical analysis using an anti-Urat1 antibody. Upper panel: *Cyp7b1*$^{+/+}$ (WT), lower panel: *Cyp7b1*$^{-/-}$ (KO). Scale bar showed 200 μm. B, Image analysis on the signal intensities of Urat1 immunostaining in male and female PT cells. Open column; vehicle control, gray column; 27HC-treated. The results were analyzed using a two-tailed student's *t* test. $^{**}P < .01$ vs. WT. Open column; WT mice, gray column; KO mice

4.1 | Elevated 27HC in metabolic disorders increases *SLC22A12* gene promoter activity

Previous reports suggest that the expression levels of mouse Urat1/*Slc22a12* protein/gene are related to metabolic disorders, such as in insulin resistance and obesity mouse models.[25,27] However, it was unclear how metabolic conditions impact human URAT1/*SLC22A12* expression and function. 27HC is the most abundant oxysterol in human circulation,[30] and its circulating levels are elevated with hypercholesterolemia and metabolic or cardiovascular dysfunction such as atherosclerosis.[30] As the *SLC22A12* gene promoter contains EREs, we postulated that estrogens and SERM would affect *SLC22A12* expression. In our study, although estrogens and 27HC bind to ER,[54] estrogens (E1 and E2) did not alter *SLC22A12* gene promoter activity at physiological concentrations (Figure 2B,C), while 27HC induced *SLC22A12* gene expression in a dose-dependent manner (Figure 2D). We also observed the involvement of ER in the *SLC22A12* gene expression using ER antagonist ICI 182,780 and ERE mutants (Figure 3). Previously, we demonstrated that E2 and 27HC cause

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

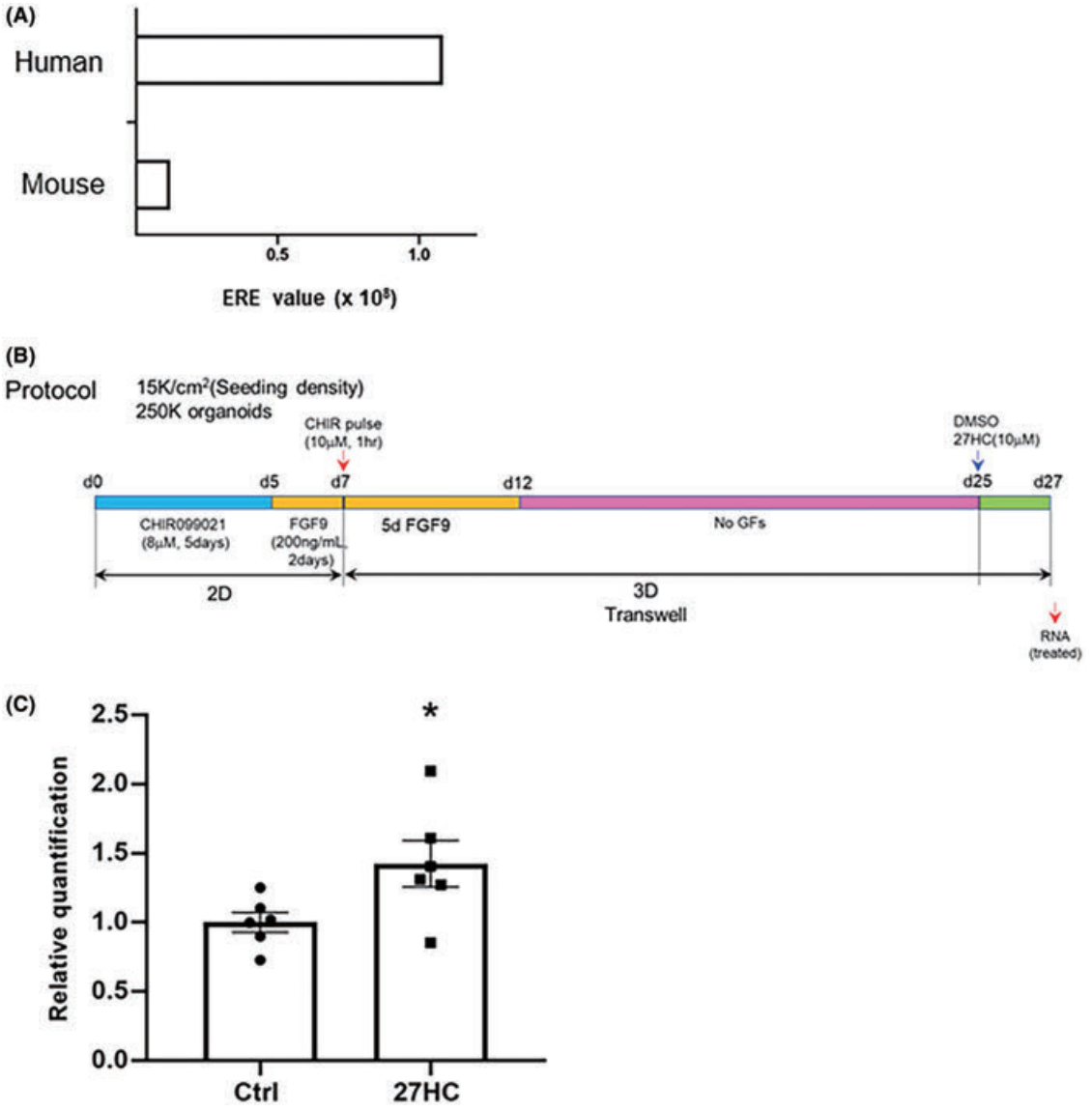


FIGURE 5 *SLC22A12* expression in human kidney organoids. A, A bar plot showing ERE values of human *SLC22A12* and mouse *Slc22a12* gene promoter regions. B, The experimental scheme of human kidney organoid culture with 27HC treatments. C, The effect of 27HC on *SLC22A12* mRNA expression in the human kidney organoids. The experiment was performed in six independent experiments. **P* < .05 vs. ctrl by two-tailed student's *t* test

unique conformational changes of ERα upon their bindings to ER.[54] Therefore, 27HC may upregulate *SLC22A12* gene promoter activity via ER, while estrogens do not, due to the difference in ERα conformational changes. Our current findings suggest that elevated 27HC levels in metabolic disorders induce *SLC22A12* transcription via ER through EREs.

As discussed in the Introduction, serum 27HC levels correlate well with serum cholesterol levels, and high cholesterol diet-feeding or metabolic dysfunction increases serum 27HC levels.[31] In addition, patients with spastic paraplegia type 5 (SPG5), who have mutations in the gene coding for CYP7B1, have elevated serum 27HC levels with 2-5 μM.[58-60] Also, the deficiency of Cyp7b1 increased 27HC levels in mice.[28,57] As elevated 27HC in *Cyp7b1*$^{-/-}$ mice increased Urat1 expression levels (Figure 4), it is plausible that these *CYP7B1* mutations also cause increased URAT1 expression in humans. The activity of 27HC on the *SLC22A12* gene promoter was mild at 5 μM ($P = .089$, Figure 2D). Thus, *CYP7B1* gene mutation alone may not have strong effects on *SLC22A12*; however, with high cholesterol diets and/or metabolic dysfunction, the serum 27HC levels can easily reach the levels that increase URAT expression enough to modulate serum uric acid levels. Further investigation into the relationship between human

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

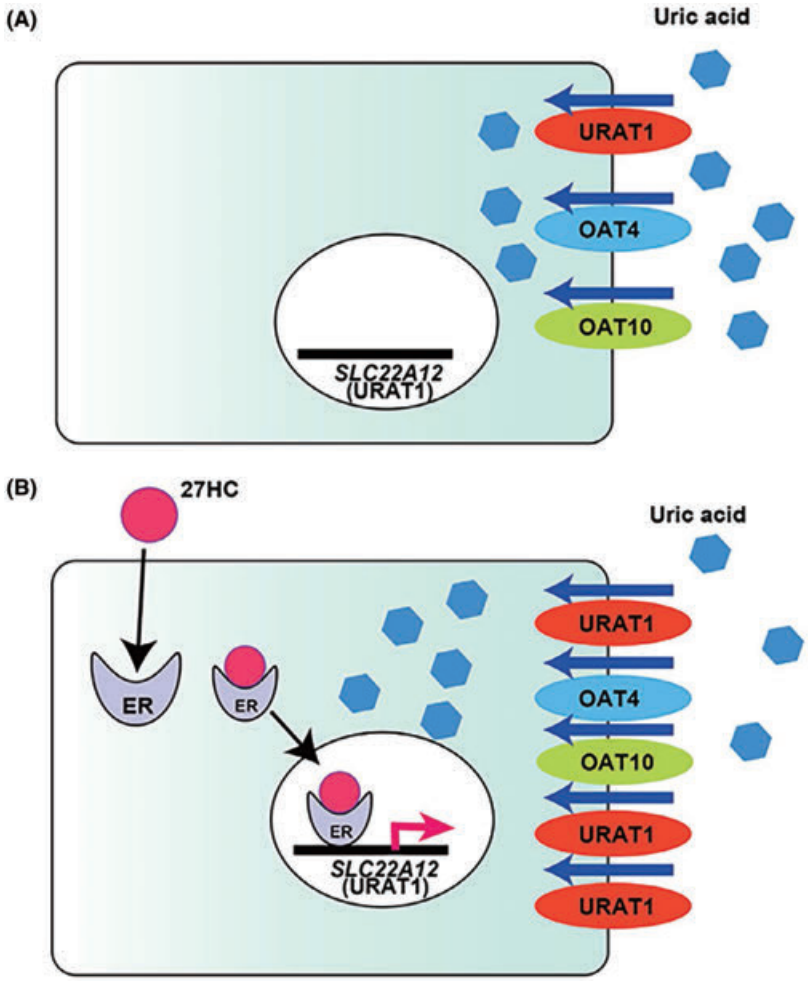


FIGURE 6 Graphical abstract. 27HC binds to ER and induces URAT1/*SLC22A12* expression. This results in an increase of URAT1 in the apical membrane, and more uric acid is transported into the cells, leading to increased serum uric acid levels. The red arrow demonstrates the upregulation of *SLC22A12* transcription. Blue arrows show the movement of uric acid. (A) basal and (B) metabolically dysfunctional condition

CYP7B1 gene mutation and URAT1/*SLC22A12* expression is required.

4.2 | Sex differences on the effect of 27HC on URAT1 expression

Given the impact of 27HC on *SLC22A12* gene promoter via ER, it is possible that sex differences on the effect of 27HC on *SLC22A12* activity are relevant. Our study showed that in *Cyp7b1*$^{-/-}$ mouse kidneys, there was a significant increase in Urat1 expression level in male mice, but not in female mice (Figure 4). This suggests the presence of sex differences in the effect of 27HC on Urat1 expression.

Although the mechanism underlying this sex difference is unclear, we deduce that the balance between estrogens and 27HC is important. For instance, 27HC and estradiol co-treatment halted the suppression of tyrosine hydroxylase expression by 27HC.[61] Estradiol did not change *Slc22a12* mRNA, but it suppressed Urat1 protein expression levels in mice.[62] In fact, Urat1 protein expression in wild-type male mice was greater than that in females (Figure 4 and previous report[63]). Given these facts, the effect of 27HC on Urat1 protein expression may be counteracted by the inhibitory effect of estradiol in female mice.

Despite clear sex differences in the effect of 27HC on mouse Urat1 expression, sex differences on the effect of 27HC on human URAT1 expression remain unclear. Existing studies have found that females, especially before menopause, are less likely to develop uric acid associated morbidities, such as cardiovascular and renal disorders, compared with males in human.[64,65] There is also evidence that E2 suppresses inflammatory responses, while 27HC increases them in the vasculature.[28] Hence, it is possible that the 27HC-induced upregulation of URAT1/*SLC22A12* expression in humans is indirectly antagonized with estrogens, and that it influenced the sex differences in the effect of 27HC. Further studies are required to reveal the relationship between 27HC and estrogens on URAT1/*SLC22A12* expression in humans.

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.3 | Relationship between serum 27HC levels and uric acid levels

As elevated URAT1 expression by 27HC potentially induces the reabsorption of uric acid, serum uric acid levels might be increased. However, in our study, we failed to detect differences in *Cyp7b1*$^{+/+}$ and *Cyp7b1*$^{-/-}$ mice (data not shown). Uric acid is metabolized to allantoin by uricase in rodents, but the activity of uricase had been lost in humans.[66] The basal serum uric acid levels in rodents are much lower than that in humans.[67] Also, the ERE score of the mouse *Slc22a12* gene promoter was much less than that of the corresponding human gene promoter (Figure 5A). In addition, human URAT1 has a fivefold higher affinity for uric acid than mouse Urat1, presumably a result of the evolutionary process for greater uric acid retention and more precise control of serum uric acid levels in humans than mice.[67] Thus, we speculated that it would be difficult to detect changes in serum uric acid levels by 27HC in mice.

Control of serum uric acid levels consists of the balance production and excretion of uric acid. As shown in Figure 1A, the excretion of uric acid is strictly regulated by several uric acid transporters, reabsorption, and secretion transporters in the kidney. Although our findings showed that 27HC increased the activity of the *SLC22A12* gene promoter, the effect of 27HC on other uric acid transporters remains elusive. The ERE value of the *SLC22A12* gene promoter is much higher than that of other uric acid transporter gene promoters (Figure S3), suggesting that URAT1/*SLC22A12* is the only transporter regulated by 27HC among uric acid transporters. Hence, the reabsorption of uric acid by URAT1 could be more dominant than secretion by secretion transporters, such as NPT1/4 and ABCG2, and excretion of uric acid to urine may be reduced.

The relationship between serum 27HC levels and uric acid levels in humans remains unknown. Given that rare variants of the human *SLC22A12* gene affect serum uric acid levels,[68] URAT1 might be important in serum uric acid control. It is quite likely that changes in wild-type URAT1/*SLC22A12* expression levels will impact serum uric acid levels. Further study is warranted to investigate the relationship between 27HC levels and serum uric acid levels.

5 | CONCLUSION

Based on our findings, it is plausible that elevated 27HC levels induce elevated URAT1 expression and potentially increase serum uric acid levels. People with high URAT1 expression levels are susceptible to diseases caused by hyperuricemia and hypercholesterolemia, which can be detrimental to their health. Our findings help to elucidate the correlation between hyperuricemia and metabolic syndromes and provide a potential therapeutic approach toward hyperuricemia-related diseases by modifying 27HC levels.

ACKNOWLEDGMENTS

The authors thank Keren-Happuch E Fan Fen for her critical reading of the manuscript and Aya Shimada for assisting our study. This work was supported by grants from JSPS KAKENHI [JP17H07031, JP20H03199 to EM, JP16K19836, JP19K08150 to S. Kobashigawa, JP18H06202, JP19K21306, JP20K16583 to HN, JP19K16925 to TKM, JP19K23976 to MN], AMED Brain/MINDS Beyond [JP20dm0307032 to EM], National Institutes of Health grant HL127037 and UH HEALTH Research Institute Pilot Funding for Basic Biomedical Science to MU, Takeda Science Foundation to EM and TKM, Kanzawa Medical Research Foundation to EM, Uehara Memorial Foundation to EM and S. Kikuchi, Nakatomi Foundation to EM, Konica Minolta Science and Technology Foundation to EM, Naito Foundation to EM, MSD Life Science Foundation to EM, Mochida Memorial Foundation for Medical and Pharmaceutical Research to EM, SENSHIN Medical Research Foundation to EM, Terumo Foundation for Life Sciences and Arts to EM, Nara Kidney Disease Research Foundation to EM, Novartis Research Grants to EM, KS, and HN, Nara Medical University Grant-in-Aid for Collaborative Research Projects to KS, Nara Medical University Grant-in-Aid for Young Scientists to TKM, Sumitomo Dainippon Pharma Research Grant to TKM, Tokyo Biochemical Research Foundation to S. Kikuchi, and by unrestricted funds provided to EM from Dr Taichi Noda (KTX Corp., Aichi, Japan) and Dr Yasuhiro Horii (Koseikai, Nara, Japan).

CONFLICT OF INTEREST

EM and TN received joint collaborative funding from Fujiyakuhin, Co. Ltd.

AUTHOR CONTRIBUTIONS

MM, YMS, MU, and EM designed research. MM, YMS, YS, HN, AA, LB, RN, and ME performed research. MM, YMS, YS, and S. Kikuchi analyzed data. A.A, LB, MM, MN, and MU performed immunohistochemical analysis. MM, YMS, MU, and EM wrote the paper. S. Kikuchi, MN, S. Kobashigawa, TKM, G.K, MH, ST, K.T, KS, MT, SN, MU, and TN helped to analyze and interpret the data, and critically revised the manuscript. MU and EM conceptualized the study, developed study design, supervised the authors throughout the study, and provided expertise in manuscript preparation. All authors read and approved the final manuscript.

ORCID

Masaya Matsubayashi https://orcid.org/0000-0003-3260-1735

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Michihisa Umetani https://orcid.org/0000-0001-8474-7088
Eiichiro Mori https://orcid.org/0000-0003-0945-196X

REFERENCES

1. Zhu Y, Pandya BJ, Choi HK. Prevalence of gout and hyperuricemia in the US general population: the National Health and Nutrition Examination Survey 2007–2008. *Arthritis Rheum*. 2011;63:3136-3141.
2. Yang T, Chu CH, Bai CH, et al. Uric acid level as a risk marker for metabolic syndrome: a Chinese cohort study. *Atherosclerosis*. 2012;220:525-531.
3. Zoccali C, Mallamaci F. Uric acid, hypertension, and cardiovascular and renal complications. *Curr Hypertens Rep*. 2013;15:531-537.
4. Ohno S, Kohjitani A, Miyata M, et al. Recovery of endothelial function after minor-to-moderate surgery is impaired by diabetes mellitus, obesity, hyperuricemia and sevoflurane-based anesthesia. *Int Heart J*. 2018;59:559-565.
5. Sumiyoshi H, Ohyama Y, Imai K, Kurabayashi M, Saito Y, Nakamura T. Association of uric acid with incident metabolic syndrome in a japanese general population. *Int Heart J*. 2019;60:830-835.
6. Cicero AFG, Fogacci F, Giovannini M, et al. Serum uric acid predicts incident metabolic syndrome in the elderly in an analysis of the Brisighella Heart Study. *Sci Rep*. 2018;8:11529.
7. Ren P, Gao M. Association between metabolic syndrome and the serum uric acid: a cohort study. *Clin Lab*. 2018;64:719-726.
8. Wang HJ, Shi LZ, Liu CF, Liu SM, Shi ST. Association Between Uric Acid and Metabolic Syndrome in Elderly Women. *Open Med (Wars)*. 2018;13:172-177.
9. Zhen H, Gui F. The role of hyperuricemia on vascular endothelium dysfunction. *Biomed Rep*. 2017;7:325-330.
10. Nakagawa T, Cirillo P, Sato W, et al. The conundrum of hyperuricemia, metabolic syndrome, and renal disease. *Intern Emerg Med*. 2008;3:313-318.
11. Grundy SM, Cleeman JI, Daniels SR, et al. Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. *Circulation*. 2005;112:2735-2752.
12. Reaven GM. Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes*. 1988;37:1595-1607.
13. Ferrannini E, Haffner SM, Mitchell BD, Stern MP. Hyperinsulinaemia: the key feature of a cardiovascular and metabolic syndrome. *Diabetologia*. 1991;34:416-422.
14. Ford ES, Giles WH, Dietz WH. Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey. *JAMA*. 2002;287:356-359.
15. Apridonidze T, Essah PA, Iuorno MJ, Nestler JE. Prevalence and characteristics of the metabolic syndrome in women with polycystic ovary syndrome. *J Clin Endocrinol Metab*. 2005;90:1929-1935.
16. Cooke JP, Oka RK. Does leptin cause vascular disease? *Circulation*. 2002;106:1904-1905.
17. Ishikawa-Takata K, Ohta T, Moritaki K, Gotou T, Inoue S. Obesity, weight change and risks for hypertension, diabetes and hypercholesterolemia in Japanese men. *Eur J Clin Nutr*. 2002;56:601-607.
18. Shirasawa T, Ochiai H, Yoshimoto T, et al. Cross-sectional study of associations between normal body weight with central obesity and hyperuricemia in Japan. *BMC Endocr Disord*. 2020;20:2.
19. Lee J, Sparrow D, Vokonas PS, Landsberg L, Weiss ST. Uric acid and coronary heart disease risk: evidence for a role of uric acid in the obesity-insulin resistance syndrome. The Normative Aging Study. *Am J Epidemiol*. 1995;142:288-294.
20. Nigam SK, Bhatnagar V. The systems biology of uric acid transporters: the role of remote sensing and signaling. *Curr Opin Nephrol Hypertens*. 2018;27:305-313.
21. Enomoto A, Kimura H, Chairoungdua A, et al. Molecular identification of a renal urate anion exchanger that regulates blood urate levels. *Nature*. 2002;417:447-452.
22. Matsuo H, Chiba T, Nagamori S, et al. Mutations in glucose transporter 9 gene SLC2A9 cause renal hypouricemia. *Am J Hum Genet*. 2008;83:744-751.
23. Matsuo H, Takada T, Ichida K, et al. Common defects of ABCG2, a high-capacity urate exporter, cause gout: a function-based genetic analysis in a Japanese population. *Sci Transl Med*. 2009;1:5ra11.
24. Anzai N, Ichida K, Jutabha P, et al. Plasma urate level is directly regulated by a voltage-driven urate efflux transporter URATv1 (SLC2A9) in humans. *J Biol Chem*. 2008;283:26834-26838.
25. Miao Z, Yan S, Wang J, et al. Insulin resistance acts as an independent risk factor exacerbating high-purine diet induced renal injury and knee joint gouty lesions. *Inflamm Res*. 2009;58:659-668.
26. Wu X, Zhang J, Liu T, et al. Uric acid crystal could inhibit Numb-induced URAT1 lysosome degradation in uric acid nephropathy. *J Physiol Biochem*. 2015;71:217-226.
27. Doshi M, Takiue Y, Saito H, Hosoyamada M. The increased protein level of URAT1 was observed in obesity/metabolic syndrome model mice. *Nucleosides Nucleotides Nucleic Acids*. 2011;30:1290-1294.
28. Umetani M, Ghosh P, Ishikawa T, et al. The cholesterol metabolite 27-hydroxycholesterol promotes atherosclerosis via proinflammatory processes mediated by estrogen receptor alpha. *Cell Metab*. 2014;20:172-182.
29. Umetani M. Re-adopting classical nuclear receptors by cholesterol metabolites. *J Steroid Biochem Mol Biol*. 2016;157:20-26.
30. Umetani M, Shaul PW. 27-Hydroxycholesterol: the first identified endogenous SERM. *Trends Endocrinol Metab*. 2011;22:130-135.
31. Brown AJ, Jessup W. Oxysterols and atherosclerosis. *Atherosclerosis*. 1999;142:1-28.
32. Burkard I, von Eckardstein A, Waeber G, Vollenweider P, Rentsch KM. Lipoprotein distribution and biological variation of 24S- and 27-hydroxycholesterol in healthy volunteers. *Atherosclerosis*. 2007;194:71-78.
33. Umetani M, Domoto H, Gormley AK, et al. 27-Hydroxycholesterol is an endogenous SERM that inhibits the cardiovascular effects of estrogen. *Nat Med*. 2007;13:1185-1192.
34. Kuzuya M, Ando F, Iguchi A, Shimokata H. Effect of aging on serum uric acid levels: longitudinal changes in a large Japanese population group. *J Gerontol A Biol Sci Med Sci*. 2002;57:M660-664.
35. Zhang X, Meng Q, Feng J, et al. The prevalence of hyperuricemia and its correlates in Ganzi Tibetan Autonomous Prefecture, Sichuan Province. *China. Lipids Health Dis*. 2018;17:235.
36. Ioannou GN, Boyko EJ. Effects of menopause and hormone replacement therapy on the associations of hyperuricemia with mortality. *Atherosclerosis*. 2013;226:220-227.
37. Chen-Xu M, Yokose C, Rai SK, Pillinger MH, Choi HK. Contemporary prevalence of gout and hyperuricemia in the United States and Decadal Trends: The National Health and Nutrition Examination Survey, 2007–2016. *Arthritis Rheumatol*. 2019;71:991-999.

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002078R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

38. Liu R, Han C, Wu D, et al. Prevalence of hyperuricemia and gout in mainland China from 2000 to 2014: A systematic review and meta-analysis. *Biomed Res Int*. 2015;204:1–12. https://doi.org/10.1155/2015/762820
39. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23:2947-2948.
40. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406-425.
41. Felsenstein J. PHYLIP-Phylogeny Inference Package (Version 3.2) Cladistics. 1989;5:164–166.
42. Wu H, Uchimura K, Donnelly EL, Kirita Y, Morris SA, Humphreys BD. Comparative analysis and refinement of human PSC-derived kidney organoid differentiation with single-cell transcriptomics. *Cell Stem Cell*. 2018;23:869-881.e868.
43. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36:411-420.
44. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell*. 2019;177:1888-1902.e1821.
45. Barthelson RA, Lambert GM, Vanier C, Lynch RM, Galbraith DW. Comparison of the contributions of the nuclear and cytoplasmic compartments to global gene expression in human cells. *BMC Genom*. 2007;8:340.
46. Bahar Halpern K, Caspi I, Lemze D, et al. Nuclear retention of mRNA in mammalian tissues. *Cell Rep*. 2015;13:2653-2662.
47. Habib N, Avraham-Davidi I, Basu A, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods*. 2017;14:955-958.
48. Lake BB, Codeluppi S, Yung YC, et al. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci Rep*. 2017;7:6031.
49. Takasato M, Er PX, Chiu HS, et al. Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. *Nature*. 2016;536:238.
50. Enomoto A, Endou H. Roles of organic anion transporters (OATs) and a urate transporter (URAT1) in the pathophysiology of human disease. *Clin Exp Nephrol*. 2005;9:195-205.
51. Yahyaoui R, Esteva I, Haro-Mora JJ, et al. Effect of long-term administration of cross-sex hormone therapy on serum and urinary uric acid in transsexual persons. *J Clin Endocrinol Metab*. 2008;93:2230-2233.
52. Thomas MP, Potter BV. The structural biology of oestrogen metabolism. *J Steroid Biochem Mol Biol*. 2013;137:27-49.
53. Onizuka Y, Nagai K, Ideno Y, et al. Association between FSH, E1, and E2 levels in urine and serum in premenopausal and postmenopausal women. *Clin Biochem*. 2019;73:105-108.
54. DuSell CD, Umetani M, Shaul PW, Mangelsdorf DJ, McDonnell DP. 27-hydroxycholesterol is an endogenous selective estrogen receptor modulator. *Mol Endocrinol*. 2008;22:65-77.
55. Wakeling AE, Dukes M, Bowler J. A potent specific pure antiestrogen with clinical potential. *Cancer Res*. 1991;51:3867-3873.
56. Dudley MW, Sheeler CQ, Wang H, Khan S. Activation of the human estrogen receptor by the antiestrogens ICI 182,780 and tamoxifen in yeast genetic systems: implications for their mechanism of action. *Proc Natl Acad Sci U S A*. 2000;97:3696-3701.
57. Li-Hawkins J, Lund EG, Turley SD, Russell DW. Disruption of the oxysterol 7alpha-hydroxylase gene in mice. *J Biol Chem*. 2000;275:16536-16542.
58. Schüle R, Siddique T, Deng HX, et al. Marked accumulation of 27-hydroxycholesterol in SPG5 patients with hereditary spastic paresis. *J Lipid Res*. 2010;51:819-823.
59. Marelli C, Lamari F, Rainteau D, et al. Plasma oxysterols: biomarkers for diagnosis and treatment in spastic paraplegia type 5. *Brain*. 2018;141:72-84.
60. Acimovic J, Lövgren-Sandblom A, Olin M, et al. Sulphatation does not appear to be a protective mechanism to prevent oxysterol accumulation in humans and mice. *PLoS One*. 2013;8:e68031.
61. Rantham Prabhakara JP, Feist G, Thomasson S, Thompson A, Schommer E, Ghribi O. Differential effects of 24-hydroxycholesterol and 27-hydroxycholesterol on tyrosine hydroxylase and alpha-synuclein in human neuroblastoma SH-SY5Y cells. *J Neurochem*. 2008;107:1722-1729.
62. Takiue Y, Hosoyamada M, Kimura M, Saito H. The effect of female hormones upon urate transport systems in the mouse kidney. *Nucleosides Nucleotides Nucleic Acids*. 2011;30:113-119.
63. Hosoyamada M, Ichida K, Enomoto A, Hosoya T, Endou H. Function and localization of urate transporter 1 in mouse kidney. *J Am Soc Nephrol*. 2004;15:261-268.
64. Cobo G, Hecking M, Port FK, et al. Sex and gender differences in chronic kidney disease: progression to end-stage renal disease and haemodialysis. *Clin Sci (Lond)*. 2016;130:1147-1163.
65. Reckelhoff JF. Gender differences in the regulation of blood pressure. *Hypertension*. 2001;37:1199-1208.
66. Álvarez-Lario B, Macarrón-Vicente J. Uric acid and evolution. *Rheumatology (Oxford)*. 2010;49:2010-2015.
67. Tan PK, Farrar JE, Gaucher EA, Miner JN. Coevolution of URAT1 and Uricase during Primate Evolution: Implications for Serum Urate Homeostasis and Gout. *Mol Biol Evol*. 2016;33:2193-2200.
68. Misawa K, Hasegawa T, Mishima E, et al. Contribution of rare variants of the SLC22A12 gene to the missing heritability of serum urate levels. *Genetics*. 2020;214:1079-1090.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Matsubayashi M, Sakaguchi YM, Sahara Y, et al. 27-Hydroxycholesterol regulates human *SLC22A12* gene expression through estrogen receptor action. *The FASEB Journal*. 2021;35:e21262. https://doi.org/10.1096/fj.202002077R

15306860, 2021, 1, Downloaded from https://faseb.onlinelibrary.wiley.com/doi/10.1096/fj.202002077R by Cochrane Japan, Wiley Online Library on [06/05/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2)

frontiers in Cellular Neuroscience

ORIGINAL RESEARCH
published: 08 June 2021
doi: 10.3389/fncel.2021.605030

Gene Expression Profiles of Human Cerebral Organoids Identify PPAR Pathway and *PKM2* as Key Markers for Oxygen-Glucose Deprivation and Reoxygenation

Naoki Iwasa[1†], Takeshi K. Matsui[1,2†], Naohiko Iguchi[1], Kaoru Kinugawa[1], Naritaka Morikawa[2], Yoshihiko M. Sakaguchi[2], Tomo Shiota[1], Shinko Kobashigawa[2], Mari Nakanishi[2], Masaya Matsubayashi[2], Riko Nagata[2], Sotaro Kikuchi[2], Tatsuhide Tanaka[3], Nobuyuki Eura[1], Takao Kiriyama[1], Tesseki Izumi[1], Kozue Saito[1], Hiroshi Kataoka[1], Yuichi Saito[4], Wataru Kimura[4], Akio Wanaka[3], Yuhei Nishimura[5], Eiichiro Mori[2,6] and Kazuma Sugie[1*]*

[1] Department of Neurology, Nara Medical University, Kashihara, Japan, [2] Department of Future Basic Medicine, Nara Medical University, Kashihara, Japan, [3] Department of Anatomy and Neuroscience, Nara Medical University, Kashihara, Japan, [4] Laboratory for Heart Regeneration, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan, [5] Department of Integrative Pharmacology, Mie University Graduate School of Medicine, Tsu, Japan, [6] V-iCliniX Laboratory, Nara Medical University, Kashihara, Japan

Ischemic stroke is one of the most common neurological diseases. However, the impact of ischemic stroke on human cerebral tissue remains largely unknown due to a lack of ischemic human brain samples. In this study, we applied cerebral organoids derived from human induced pluripotent stem cells to evaluate the effect of oxygen-glucose deprivation/reoxygenation (OGD/R). Pathway analysis showed the relationships between vitamin digestion and absorption, fat digestion and absorption, peroxisome proliferator-activated receptor (PPAR) signaling pathway, and complement and coagulation cascades. Combinational verification with transcriptome and gene expression analysis of different cell types revealed fatty acids-related PPAR signaling pathway and pyruvate kinase isoform M2 (*PKM2*) as key markers of neuronal cells in response to OGD/R. These findings suggest that, although there remain some limitations to be improved, our ischemic stroke model using human cerebral organoids would be a potentially useful tool when combined with other conventional two-dimensional (2D) mono-culture systems.

Keywords: cerebral organoid, human induced pluripotent stem cells, oxygen-glucose deprivation/reoxygenation, ischemic stroke, pyruvate kinase M2, peroxisome proliferator-activated receptor, ferritin light chain

OPEN ACCESS

Edited by:
Li Li,
Capital Medical University, China

Reviewed by:
Zhisong He,
ETH Zurich, Switzerland
Silvia Bolognin,
University of Luxembourg, Luxembourg

****Correspondence:***
Eiichiro Mori
emori@naramed-u.ac.jp
Kazuma Sugie
ksugie@naramed-u.ac.jp

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Cellular Neuropathology, a section of the journal Frontiers in Cellular Neuroscience

Received: *11 September 2020*
Accepted: *10 May 2021*
Published: *08 June 2021*

Citation:
Iwasa N, Matsui TK, Iguchi N, Kinugawa K, Morikawa N, Sakaguchi YM, Shiota T, Kobashigawa S, Nakanishi M, Matsubayashi M, Nagata R, Kikuchi S, Tanaka T, Eura N, Kiriyama T, Izumi T, Saito K, Kataoka H, Saito Y, Kimura W, Wanaka A, Nishimura Y, Mori E and Sugie K (2021) Gene Expression Profiles of Human Cerebral Organoids Identify PPAR Pathway and PKM2 as Key Markers for Oxygen-Glucose Deprivation and Reoxygenation. Front. Cell. Neurosci. 15:605030. doi: 10.3389/fncel.2021.605030

INTRODUCTION

Stroke, a cerebrovascular disease, is a common neurological disorder, and ischemic strokes are among the major causes of permanent morbidity and disability (Spescha et al., 2013; Knowland et al., 2014; Benjamin et al., 2019). Several therapeutic agents are available for acute ischemic stroke, including thrombolytics, anticoagulants, and antiplatelets. However, their efficacy is restricted and

their use is limited (Catanese et al., 2017). Mechanical thrombolytics and recombinant tissue plasminogen activator (rt-PA) have therapeutic time windows for hyperacute ischemic stroke, because the target of these therapies is to rescue the penumbra around the ischemic core (Fisher and Bastan, 2012; Dorado et al., 2014). Thus, elucidating the mechanism of ischemic stroke in human cerebral tissue is expected to be important for the development of appropriate therapies.

Thus far, two-dimensional (2D) neuron cultures have been utilized in ischemic models. However, there are limitations to these models as actual brain tissues are composed of multiple cell types and responses to ischemia take place from cell to cell. Recent progress in the development of organs in-a-dish (organoids) provides potential for the modeling of various diseases (Clevers, 2016). Organoids resemble the architecture of organs that are composed of multiple cell types, and three-dimensionally (3D) cultured cerebral organoids are expected to represent physiological environment (Pasca et al., 2015). In this study, we analyzed the gene expression profiles of human cerebral organoids after oxygen-glucose deprivation/reoxygenation (OGD/R).

MATERIALS AND METHODS

Human Cerebral Organoids Generation and Culture

Human induced pluripotent stem cells (iPSCs) which were Cellartis Human iPS Cell Line 12 (ChiPSC12, Y00285, Takara Bio, Kusatsu, Shiga, Japan) were cultivated in mTeSRTM1 medium (05851, Stemcell Technologies, Vancouver, British Columbia, Canada) and maintained in feeder-free condition with mTeSR-TM1 media. Feeder-free human iPSC line (XY) from Takara was obtained on six well plates (3506, Corning, New York, USA) coated with growth factors reduced Matrigel (356230; BD Biosciences, San Jose, CA, USA). At the time of passage, we added Rho-associated protein kinase (ROCK) inhibitor (final concentration 10 μM; Selleck Chemicals, Houston, Texas, USA), and maintained these cells with daily medium change without ROCK inhibitor until they reached about 70% confluency. Then, they were detached by versene solution (15040-066; Thermo Fisher Scientific, Waltham, MA, USA) and seeded by a 1:20 dilution ratio.

Cerebral organoids were differentiated according to previously published protocol (Lancaster et al., 2013). Human iPSCs were detached and subjected to embryoid body (EB) induction using the protocol. After 4 days, half of the media was replaced by human EB medium without ROCK inhibitor and basic fibroblast growth factor (bFGF). Two days later, the EBs were transferred into a neural induction media which consisted of DMEM-F12 with 1× N2 supplement (17502048; Thermo Fisher Scientific, Waltham, MA, USA), 1× Glutamax, 1× non-essential amino acids, and heparin (1 μg/ml) and then embedded in Matrigel after 5 days. Subsequently, the organoids were induced in organoid medium by using an orbital shaker. The total process took 6 weeks, from EB formation.

Immunohistochemical Analysis of Human Cerebral Organoids

Human cerebral organoids were fixed in 4% paraformaldehyde in Phosphate-Buffered Saline (PBS) overnight at 4°C, dehydrated with 30% sucrose in PBS and embedded in O.C.T. Compound (Thermo Fisher Scientific, Waltham, MA, USA). Cryostat sections (14 μm) were cut and mounted on slides (Thermo Fisher Scientific, Waltham, MA, USA). Mounted sections were incubated at room temperature for 1 h with blocking solution [3% normal goat serum + 0.3% Triton X-100 in tris-buffered saline (TBS)] and subsequently incubated with primary antibodies diluted in blocking solution overnight at 4°C. Antibodies specific for TUJ1 (1:400, T8660, Sigma-Aldrich, St. Louis, Missouri, USA) were used for immunostaining. After three washes with TBS, corresponding fluorescent dye Alexa Fluor 488-conjugated anti-mouse IgG (715-545-151, Jackson Immunoresearch, West Grove, PA, USA) secondary antibodies diluted in the blocking solution were added and samples were incubated at room temperature for 2 h and followed by 4′,6-diamidino-2-phenylindole (DAPI) (Vectashield Mounting Medium with DAPI, H-200; Vector Laboratories, Burlingame, CA, USA) staining. Finally, stained slides were rinsed with TBS three times, mounted, and analyzed using a FV3000 Confocal Microscope (Olympus, Shinjuku, Tokyo, Japan).

Oxygen-Glucose Deprivation/Reoxygenation and RNA Extraction of Human Cerebral Organoids

The organoids were plated in Dulbecco's Modified Eagle Medium without glucose [DMEM (No Glucose), 09891-25; NACALAI TESQUE, Kyoto, Kyoto, Japan], and 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan). Subsequently, the organoids were cultured in a hypoxic incubator (94% N_2, 5% CO_2, 1% O_2) at 37°C for 1 h. For reoxygenation, the medium was changed to Primary Neuron Basal Medium, then incubated (95% air, 5% CO_2) at 37°C for 1 h. RNA from human cerebral organoids was extracted according to the manufacturer's protocol supplied with TRIzol reagent (15596018; Thermo Fisher Scientific, Waltham, MA, USA). The concentration and purity of the RNA samples were measured using Spectrophotometer (Beckman Coulter, Brea, California, USA).

RNA Sequencing of Human Cerebral Organoids

Total RNA was isolated from the cells using the PureLink RNA Mini Kit (12183020; Thermo Fisher Scientific, Waltham, MA, USA) and according to the manufacturer's instructions. RNA concentration was analyzed by Qubit RNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) and the purity was assessed using the Qsep100 DNA Fragment Analyzer and RNA R1 Cartridge (BiOptic, New Taipei City, Taiwan). For RNA-sequencing (RNA-seq) analysis, each extracted RNA sample of non-treated organoid in equal concentration was adopted, and the same was adopted for OGD/R. Extracted RNA samples were either sent to a Bioengineering lab for RNA-sequencing analysis,

and or subjected to RT-PCR. Then, total RNA was converted to cDNA and used for Illumina sequencing library preparation, according to KAPA Stranded mRNA-Seq Kit protocols (KAPA Biosystems, Wilmington, MA, United States). DNA fragments were ligated by FastGene Adapter Kit (NIPPON Genetics, Bunkyo, Tokyo, Japan). After the purified cDNA library products were appreciated using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA), they were qualitatively evaluated using the Fragment Analyzer and dsDNA 915 Reagent Kit (Advanced Analytical Technologies) before it was finally sequenced (2 × 76 bp) on NextSeq 500 (Illumina, San Diego, CA, United States). Sickle (ver.1.33) is an error correction tool for quality trimming. It discards reads based upon the length threshold derived from quality score for FASTQ files. Using Sickle (ver.1.33), the threshold identified were quality values under 20 base and reads under 30 bp. Additionally, Hisat2 software was run with default parameters. The sequence alignment files generated by Sickle and Hisat2 were used to generate counts of mapped reads.

Differentially Expressed Genes Extraction of Human Cerebral Organoids

By comparing organoids under OGD/R condition to non-treated organoids and using R package edgeR (Robinson et al., 2010) in R software (version 3.5.3), differentially expressed genes (DEGs) were extracted. The MA plot in R is a visual tool for showing the total gene expression levels of DEGs. The X-axis represents log CPM—log counts per million—and are measures of gene expression level. The Y-axis indicates log FC. Log FC is the log fold-change, which in this case, is the log difference between cerebral organoids after OGD/R.

Gene Ontology and Pathway Analysis of Human Cerebral Organoids

Using the Database for Annotation, Visualization, and Integrated Discovery [DAVID, version 6.8 database (https://david.ncifcrf.gov/)], selected DEGs were analyzed. The DAVID shows the molecular function, biological process, and cellular component expressed in the gene profile. In this study, DAVID was applied to evaluate gene ontological (GO) annotation and Kyoto encyclopedia of genes and genomes (KEGG) pathways of DEGs. P-value < 0.05 was chosen as the threshold of KEGG pathway. Upregulation and downregulation of genes were analyzed separately.

Protein–Protein Interaction Analysis of Human Cerebral Organoids

Based on the Search Tool for the Retrieval of Interacting Genes (STRING, https://string-db.org/; Szklarczyk et al., 2019) online database, DEGs were inputted, and the protein–protein interaction (PPI) network was constructed. The STRING database shows the interaction of each protein from prediction or experiments. The threshold to construct the PPI network was confidence score >0.4.

Reverse Transcriptase-Polymerase Chain Reaction and Quantitative Polymerase Chain Reaction of Human Cerebral Organoids

For RT-PCR, the extracted RNAs were reverse transcribed using the protocol supplied with ReverTra Ace qPCR RT Master Mix (FSQ-201; TOYOBO, Osaka, Osaka, Japan). StepOne Plus Real-time PCR System (Thermo Fisher Scientific, Waltham, MA, USA) was used to amplify and quantify levels of target gene cDNA. Quantitative PCR (q-PCR) was conducted using SsoAdvanced Universal SYBR Green Supermix (172-5271; BioRad Laboratories, Hercules, CA, United States) and specific primers for q-PCR. Primers used in this study are listed in **Supplementary Table 1**. Reactions were run in triplicate. The expression of each gene was normalized to the geometric mean of β-actin (10025636; BioRad Laboratories, Hercules, CA, United States) as a housekeeping gene and analyzed using the ΔΔCT method. Statistical analysis was performed using the GraphPad Prism7 (GraphPad Software). A two-tailed unpaired Student's t-test was applied to the q-PCR data. P-value < 0.05 was defined as the threshold.

Tissue-Specific Gene Expression Analysis

Gene expression data for normal human tissues were analyzed from the genotype tissue expression project (GTEx) portal (http://www.gtexportal.org/home/) to compare the similarity of expression pattern of some genes.

Oxygen-Glucose Deprivation/Reoxygenation of Human Hepatocellular Carcinoma and Quantitative Polymerase Chain Reaction

The HepG2 which were human hepatocellular carcinoma cell line (Cellular Engineering Technologies, Inc., Coralville, IA, USA) were maintained in growth medium (RPMI1640; NACALAI TESQUE, Kyoto, Kyoto, Japan) supplemented with 10% fetal bovine serum (HyClone, SH30396.03, GE Healthcare Life Sciences, Logan, UT, USA), 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan). The cells were incubated at 37°C in 5% CO_2. Subsequently, cultured HepG2 cells were plated in DMEM without glucose (09891-25; NACALAI TESQUE, Kyoto, Kyoto, Japan), and 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan). Subsequently, the organoids were cultured in a hypoxic incubator (94% N_2, 5% CO_2, 1% O_2) at 37°C for 1 h. For reoxygenation, the medium was changed to growth medium, then incubated (95% air, 5% CO_2) at 37°C for 1 h. After cell culture, RNA extraction, RT-PCR and q-PCR were performed using the same procedure of human cerebral organoids. The sequences of primers used in this study are listed in **Supplementary Table 1**.

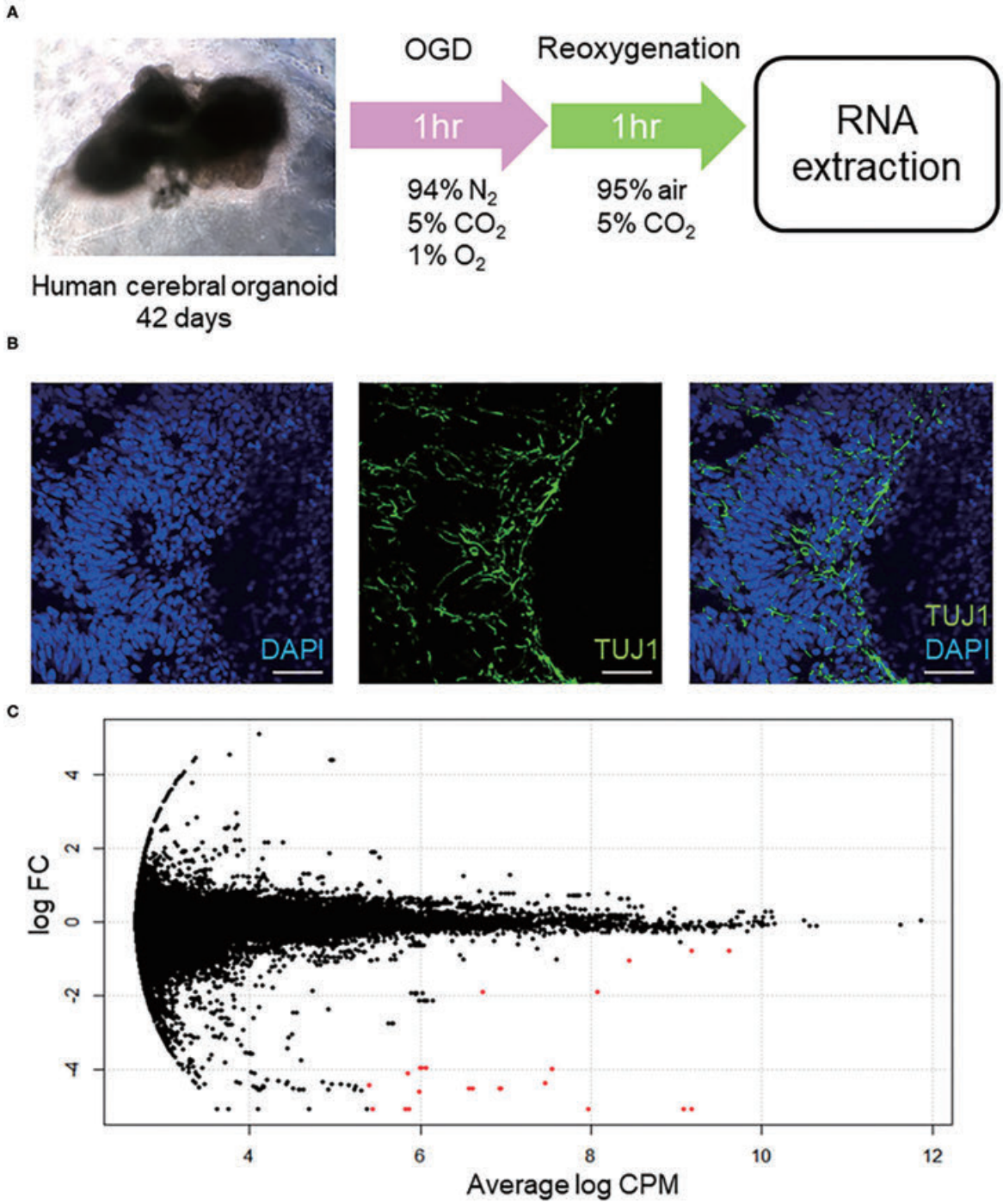


FIGURE 1 | (A) Schema of this experimental procedure. First, human cerebral organoids were generated and cultured for 42 days. Subsequently, the organoids were cultured in a hypoxic incubator (94% N_2, 5% CO_2, 1% O_2) at 37°C for 1 h. For reoxygenation, they were incubated (95% air, 5% CO_2) at 37°C for 1 h. After OGD/R treatment, RNA was extracted. **(B)** Immunohistochemical staining of non-treated human cerebral organoids. Immunohistochemical staining of neuronal cells' marker (TUJ1), and 49,6-diamidino-2phenylindole (DAPI). Fluorescent micrographs show coexisting of TUJ1 (green) and DAPI (blue) at human cerebral organoids. Bars = 100 μm. **(C)** MA plot of cerebral organoids. The X-axis is log CPM, which is a measure of gene expression level. The Y-axis indicates log FC, which is the log difference between cerebral organoids after OGD/R and controls. Red plots are significant DEGs. CPM, counts per million, FC, fold-change; DEGs, differentially expressed genes.

TABLE 1 | Profiling gene expression comparing organoids under OGD/R condition with non-treated organoids.

Gene symbol	logFC	*P*-value	FDR	Stage
AFP	−5.05202426	6.61E-89	1.19E-83	Down
TTR	−4.829589384	1.19E-37	7.09E-33	Down
APOA2	−4.375589058	1.38E-24	6.18E-20	Down
ALB	−3.992565039	4.23E-24	1.52E-19	Down
APOA1	−4.544446565	1.37E-17	3.52E-13	Down
RNA28SN4	−1.909649871	2.98E-14	5.94E-10	Down
APOC3	−4.616288539	1.08E-09	1.76E-05	Down
FTL	−0.803043639	2.06E-09	3.08E-05	Down
AHSG	−3.975861808	5.40E-09	7.45E-05	Down
FGG	−4.858272075	7.92E-09	0.0001015	Down
FABP1	−4.10691844	1.31E-07	0.00123622	Down
MIR3615	−1.070468786	1.86E-07	0.00167334	Down
FGB	−4.756570983	2.98E-06	0.02353793	Down
RNA45SN4	−1.899507421	3.02E-06	0.02353793	Down
FGA	−4.451593603	5.72E-06	0.04279685	Down
TF	−4.591735052	2.10E-05	0.14485088	Down
RBP4	−4.460754954	4.01E-05	0.2567811	Down
APOE	−2.151241972	4.23E-05	0.26168457	Down
SELENOP	−2.776594398	4.92E-05	0.29458499	Down
RN7SL2	1.271175267	0.0001831	0.88837764	Down
MIR4709	−1.029200022	0.0002927	1	Down
APOB	−4.704983376	0.0004883	1	Down
GPC3	−1.985098838	0.0005083	1	Down
ALDH1A1	−4.546504241	0.0005188	1	Down
YWHAE	4.380630428	0.0005188	1	Up
PTN	0.775148705	0.0005895	1	Up
s-rRNA	−0.5481881	0.0007606	1	Down
PMEL	−4.567564523	0.0009766	1	Down
NNAT	0.610458501	0.0026105	1	Up
RN7SL1	1.232302556	0.0038383	1	Up
GSTA1	−4.403571625	0.0039063	1	Down
APOA4	−4.361854192	0.0039063	1	Down
POSTN	0.705402287	0.0051026	1	Up
SERPINA1	−3.780463098	0.0063477	1	Down
DNAJB1	1.887700371	0.0066108	1	Up
RNA18SN5	−2.370210664	0.0073853	1	Down
MIR6132	1.736743211	0.0146333	1	Up
LOC105379506	5.83933038	0.015625	1	Up
RBP2	−4.698155314	0.015625	1	Down
DCT	−4.574216819	0.015625	1	Down
AMBP	−3.45106874	0.0214844	1	Down
VTN	−3.157212671	0.0214844	1	Down
H19	−3.053228094	0.0214844	1	Down
PEA15	0.753258601	0.0238198	1	Up
TUBA1A	0.250293775	0.025046	1	Up
MIR1244-1	−0.274612882	0.0268679	1	Down
SPINK1	−4.120454121	0.03125	1	Down
RPS6	−0.293592327	0.0351882	1	Down
MIR4442	0.351338228	0.0366267	1	Up
HSPA8	0.404708278	0.0390529	1	Up
MIR9-2	0.832371451	0.039617	1	Up
TMSB10	0.247345492	0.0457451	1	Up

Oxygen-Glucose Deprivation/Reoxygenation of Fibroblast, Glioblastoma and Murine Microglia, and Quantitative Polymerase Chain Reaction

Human telomerase reverse transcriptase immortalized fibroblasts (hTERT-BJ), human glioblastoma cell line (U251MG; obtained from Japanese collection of research bioresources) and murine microglial cell line (BV2 cells) were cultured in DMEM with 4.5 g/l glucose (08459-64; NACALAI TESQUE, Kyoto, Kyoto, Japan) supplemented with 10% fetal bovine serum (HyClone, SH30396.03, GE Healthcare Life Sciences, Logan, UT, USA), 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan). The cells were incubated at 37°C in 5% CO_2. Subsequently, cultured hTERT-BJ, U251MG, and BV2 cells were maintained or plated in DMEM without glucose (09891-25; NACALAI TESQUE, Kyoto, Kyoto, Japan) supplemented with 10% fetal bovine serum (HyClone, SH30396.03, GE Healthcare Life Sciences, Logan, UT, USA), 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan). Hypoxic experiments were performed in a CBS-90 hypoxia workstation (ASTEC, Kasuya, Fukuoka, Japan) at 37°C under severe hypoxic (1% oxygen) or mild conditions (10% oxygen). hTERT-BJ, U251MG, and BV2 cells plated in DMEM with 4.5 g/l glucose (08459-64; NACALAI TESQUE, Kyoto, Kyoto, Japan) or DMEM without glucose (09891-2; NACALAI TESQUE, Kyoto, Kyoto, Japan) were cultured in a hypoxic incubator (94% N_2, 5% CO_2, 1% O_2) at 37°C for 1 h. The mild hypoxic condition we used was 10% oxygen for 24 h. For reoxygenation, the medium was changed to DMEM with 4.5 g/l glucose (08459-64; NACALAI TESQUE, Kyoto, Kyoto, Japan) supplemented with 10% fetal bovine serum (HyClone, SH30396.03, GE Healthcare Life Sciences, Logan, UT, USA), 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan), then incubated (95% air, 5% CO_2) at 37°C for 1 h. After cell culture, RNA extraction was performed using PureLink RNA Mini Kit (12183020; Thermo Fisher Scientific, Waltham, MA, USA). For RT-PCR, the extracted RNAs were reverse transcribed using the protocol supplied with ReverTra Ace qPCR RT Master Mix (FSQ-201; TOYOBO, Osaka, Osaka, Japan). StepOne Plus Real-time PCR System (Thermo Fisher Scientific, Waltham, MA, USA) was used to amplify and quantify levels of target gene cDNA. Quantitative PCR (q-PCR) was conducted using SsoAdvanced Universal SYBR Green Supermix (172-5271; BioRad Laboratories, Hercules, CA, United States) and specific primers for q-PCR. The sequences of primers used in this study are listed in **Supplementary Table 1**. Reactions were run in triplicate. The expression of each gene was normalized to thegeometric mean of β-actin as a housekeeping gene and analyzed using the ΔΔCT method. Statistical analysis was performed using the GraphPad Prism7 (GraphPad Software). A two-tailed unpaired Student's *t*-test was applied to the q-PCR data. P-value < 0.05 was defined as the threshold.

Primary Cortical Neuron Culture of Mice

Cortical neurons were prepared from C57BL/6 mice on E16–17. Briefly, the mouse cerebral cortex was digested with 0.25

% trypsin and DNase for 15 min. Cells were passed through a 100 m nylon mesh. The resultant cell suspension was diluted with DMEM/F12 (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 10% fetal bovine serum (HyClone, SH30396.03, GE Healthcare Life Sciences, Logan, UT, USA) and 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan). Subsequently, the cells were plated on poly-D-lysine coated dishes and maintained at 37°C in 5% CO_2. The culture medium was replaced with a serum-free DMEM/F12 supplemented with B27 (Thermo Fisher Scientific, Waltham, MA, USA) 24 h after plating.

Immunocytochemistry of Primary Murine Cortical Neuron Culture

Cultured cortical neurons were fixed with 4% PFA for 1 h at room temperature and incubated with blocking solution containing 1% BSA and 0.1% Triton X-100 in PBS for 1 h. Primary antibodies, anti-TUJ1 (1:1000, 801201; BioLegend, San Diego, CA, USA) was applied overnight at 4°C. Fluorescent dye Alexa Fluor 488-conjugated anti-mouse IgG (715-545-151; Jackson Immunoresearch, West Grove, PA, USA) was used as the secondary antibody.

Oxygen-Glucose Deprivation/Reoxygenation of Primary Cortical Neuron and Quantitative Polymerase Chain Reaction

The cultured cortical neurons were plated in DMEM without glucose (09891-25; NACALAI TESQUE, Kyoto, Kyoto, Japan), and 1% penicillin-streptomycin (26253-84; NACALAI TESQUE, Kyoto, Kyoto, Japan). Subsequently, the organoids were cultured in a hypoxic incubator (94% N_2, 5% CO_2, 1% O_2) at 37°C for 1 h. For reoxygenation, the medium was changed to Primary Neuron Basal Medium, then incubated (95% air, 5% CO_2) at 37°C for 1 h. After cell culture, RNA extraction was performed using PureLink RNA Mini Kit (12183020; Thermo Fisher Scientific, Waltham, MA, USA). For RT-PCR, the extracted RNAs were reverse transcribed using the protocol supplied with ReverTra Ace qPCR RT Master Mix (FSQ-201; TOYOBO, Osaka, Osaka, Japan). StepOne Plus Real-time PCR System (Thermo Fisher Scientific, Waltham, MA, USA) was used to amplify and quantify levels of target gene cDNA. Quantitative PCR (q-PCR) was conducted using SsoAdvanced Universal SYBR Green Supermix (172-5271; BioRad Laboratories, Hercules, CA, United States) and specific primers for q-PCR. Primers used in this study are listed in **Supplementary Table 1**.

Cell culture and hypoxic condition were performed three times, and reactions were run in triplicate. The expression of each gene was normalized to the geometric mean of β-actin as a housekeeping gene and analyzed using the ΔΔCT method. Averaged relative quantification (RQ) was calculated from the average of three RQ values per three samples, every q-PCR. Therefore, one q-PCR demonstrated two averaged RQ values (Ctrl vs. OGD/R), and finally the data from three q-PCR demonstrated six averaged RQ values: three averaged RQ values of controlled samples vs. three averaged RQ after OGD/R. Statistical analysis was performed using the GraphPad Prism7 (GraphPad Software). A two-tailed unpaired Student's *t*-test was applied to the q-PCR data. *P*-value < 0.05 was defined as the threshold.

RESULTS

Transcriptome Analysis of Human Brain Ischemic Stroke Model in a Dish

We used six cerebral organoids from one batch and treated three organoids with OGD/R, which is a well-established method of mimicking the pathological processes of ischemia (Hossmann, 1998), to generate a model of cerebral ischemia/reperfusion *in vitro*, and three for normoxic condition. We first cultured the human cerebral organoids for 42 days. Subsequently, these organoids were placed under OGD condition for 1 h and reoxygenation for 1 h. After OGD/R treatment or normoxic condition, RNAs were extracted from each human cerebral organoid prior to RNA isolation and RNA sequencing, of which samples were from every organoid sample, whereupon gene ontology analysis was performed (**Figure 1A**). Immunohistochemical staining of neuronal cells' marker (TUJ1) and DAPI showed the existence of neuronal cells in our organoids model (**Figure 1B**). Significant DEGs are colored in red (**Figure 1C**). Using the threshold of $p < 0.05$ and $|\log FC| > 0.1$, a total of 52 DEGs were determined between intact cerebral organoids and cerebral organoids under OGD/R, including 14 upregulated and 38 downregulated genes (**Table 1**). When false discovery rate (FDR) adjusted *p*-value ≤ 0.05 was used as a threshold, 15 genes of *AFP, TTR, APOA2, ALB, APOA1, RNA28SN4, APOC3, FTL, AHSG, FGG, FABP1, MIR3615, FGB, RNA45SN4,* and *FGA* were detected, and all genes were found to be downregulated (**Table 1**). The top five significant downregulated genes were *AFP, TTR, APOA2, ALB,* and *APOA1*, and upregulated genes were *RN7SL2, YWHAE, PTN, NNAT, RN7SL1, POSTN, DNAJB1, MIR6132, LOC105379506,* and *PEA15* (**Table 1**). In the brain, YWHAE is involved in directing the movement of nerve cells by binding to other proteins (Toyo-oka et al., 2003). Pleiotrophin (*PTN*), which is mainly expressed in the cerebral cortex (Kido et al., 2014; Shen et al., 2017), plays important roles in cell growth and survival, cell migration, and the expression of inflammatory cytokines (Shen et al., 2017), and *POSTN* leads to decreased apoptosis during hypoxia (Aukkarasongsup et al., 2013). These results indicate that upregulated genes might be associated with neuroprotection.

Co-expression Network Construction

To further investigate pathways involved in ischemic response of human brain, several network analyses were applied. Gene ontological (GO) analysis in the biological process category showed that OGD/R on human cerebral organoids induced lipid metabolism and blood coagulation (**Table 2**). The pathway-based analysis, KEGG pathway, further highlighted the vitamin digestion and absorption, fat digestion and absorption, peroxisome proliferator-activated receptor (PPAR) signaling pathway, and complement and coagulation cascades (**Table 3**). Using the STRING online database, 52 DEGs

TABLE 2 | Go term analysis derived from biological process of human cerebral organoids after OGD/R (FDR < 0.05).

Term	Count	*P*-value	FDR
Retinoid metabolic process	11	5.91E-17	1.55E-13
Platelet degranulation	9	5.07E-11	7.27E-08
Lipoprotein metabolic process	7	1.96E-10	2.81E-07
Cholesterol efflux	6	2.02E-09	2.89E-06
Lipoprotein biosynthetic process	5	2.49E-09	3.58E-06
Phospholipid efflux	5	1.97E-08	2.82E-05
High-density lipoprotein particle remodeling	5	2.68E-08	3.84E-05
Reverse cholesterol transport	5	5.97E-08	8.56E-05
Triglyceride catabolic process	5	2.44E-07	3.50E-04
Cholesterol homeostasis	6	2.72E-07	3.90E-04
High-density lipoprotein particle assembly	4	5.47E-07	7.85E-04
Positive regulation of cholesterol esterification	4	8.20E-07	0.001176
Protein polymerization	4	2.77E-06	0.003979
Negative regulation of very-low-density lipoprotein particle remodeling	3	1.42E-05	0.020297
Cholesterol metabolic process	5	1.47E-05	0.021062
Lipid transport	5	2.28E-05	0.032765
Triglyceride homeostasis	4	2.47E-05	0.035457
Blood coagulation, fibrin clot formation	3	2.83E-05	0.040534

were filtered by the PPI network of DEGs including nodes and edges with parameters of minimum required interaction score >0.4 (medium confidence) (**Figure 2**). These analyses revealed potential pathways involved in cellular response to ischemic stroke.

AHSG, TTR, and FGG Are Downregulated After OGD/R in the Human Cerebral Organoids

Next, we searched for important genes involved in neurological diseases. Among the affected genes after OGD/R in human cerebral organoids, transthyretin (*TTR*), alpha-2-HS-glycoprotein (*AHSG*), and fibrinogen gamma chain (*FGG*) were found to be related to neurological disease. *AHSG* is associated with alopecia-mental retardation syndrome 1 (Reza Sailani et al., 2017), patients of familial amyloid polyneuropathy (*FAP*) carrying *TTR* Met30 mutation (Tanaka et al., 1994), and fibrinogen gamma chain (*FGG*) appear to be protected after ischemic stroke (Cheung et al., 2008). By quantitative PCR (q-PCR), we confirmed downregulation of these three genes; *TTR*, *AHSG,* and *FGG* were significantly downregulated (**Figure 3A**). Our results suggest that these three genes have potential to contribute to pathogenesis of ischemic stroke in the human brain.

To find the genes that play more important roles in the brain, we then verified these three genes with gene expression data of normal human tissues obtained from GTEx database. Though a vast majority of AHSG is secreted by the liver in adults (**Figure 3B**), AHSG is synthesized by various tissues at fetal stage (Stefan et al., 2006; Dabrowska et al., 2015). We demonstrated *AHSG* was downregulated after OGD/R in the human cerebral organoids (**Figure 3A**). Therefore, to validate *AHSG* expression after OGD/R in non-neural cells, human hepatocellular carcinoma (HepG2), human fibroblasts (hTERT-BJ), and human glioblastoma (U251MG), and murine microglia (BV2 cells) as neural cells were cultured and treated under OGD/R. *Ahsg* or *AHSG* expression of murine microglia, fibroblasts, and glioblastoma was unaltered after OGD/R (**Figure 3C**). *TTR* and *FGG* were not detected in BV2 cells and *TTR* was not detected in hTERT-BJ (**Figure 3C**). These results suggest that downregulated *AHSG* of human cerebral organoids might be specific to human fetal brain or human cerebral organoids. *AHSG* expression of hepatocellular carcinoma showed downward trend (1.43 fold; $P = 0.09$) after OGD/R (**Figure 3C**). This finding might indicate the existence of a cell type with similar gene expression to that of hepatic cells in human cerebral organoids.

Expression of *FTL* in Human Cerebral Organoids, Murine Cells, and Non-neural Cells

Ferritin light chain (*FTL*) is gene responsible for neurodegeneration with brain iron accumulation (Muhoberac and Vidal, 2019), and *FTL* was found to be abundantly expressed in the brain (**Figure 4A**). Although *FTL* was significantly downregulated after OGD/R by using RNA sequencing (**Table 1**), q-PCR analysis indicated a down-regulated trend of *FTL* (3.40 fold; $P = 0.18$) and no change to the ferritin heavy chain (*FTH*) in the cerebral organoids after OGD/R (**Figure 4B**). Immunohistochemical staining of neuronal cells' marker (TUJ1) showed the existence of neuronal cells in primary cultured cerebral cortical neurons of mice (**Figure 4C**). To examine the expression level of *FTL* in mice, the expression level of *Ftl* was tested and unaltered, using primary cultured murine cortical neurons and murine microglia (BV2 cells) treated with OGD/R (**Figure 4D**). *FTL* expression of OGD/R in non-neural cells, hepatocellular carcinoma (HepG2), telomerase reverse transcriptase immortalized fibroblasts (hTERT-BJ), and glioblastoma (U251MG), were 2D cultured and challenged by OGD/R. While the expression level of *FTL* was unaltered in hTERT-BJ and U251MG, it was downregulated by OGD/R treatment in HepG2 (**Figure 4D**). Additionally, to analyze the influence of different hypoxic conditions in non-neural cells (hTERT-BJ and U251MG), hTERT-BJ and U251MG cells were 2D cultured and treated by mild hypoxia (i.e., approximately 10% O_2 for 24 h). The expression level of *FTL* was upregulated by mild hypoxia treatment in both cell lines (**Supplementary Figure 1B**). In murine primary cultured neurons, fibroblast, and glioblastoma cells, *FTL* expression was not affected by OGD/R as severe hypoxia and might be a reaction specific to human cerebral organoids.

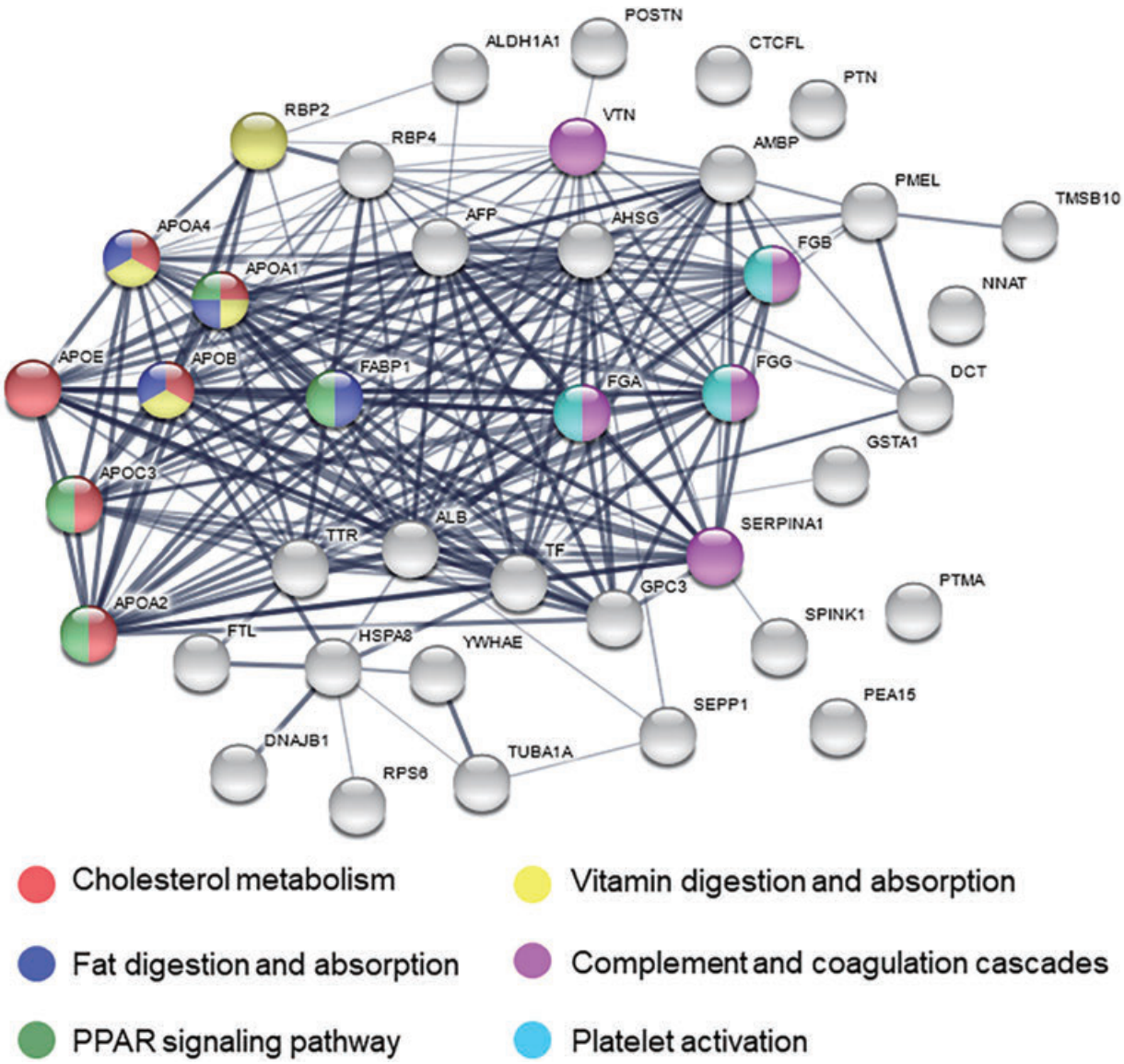


FIGURE 2 | Protein–protein interaction network of differentially expressed genes. Nodes are proteins and the thickness of lines indicates strong interaction between proteins. Colored nodes mean related pathway.

PKM2 Is Upregulated by OGD/R in Human Cerebral Organoids and Mouse Cortical Neurons

We validated some genes which could be indicators of hypoxia and ischemic markers for cerebral organoids and other cell types. Pyruvate kinase M2 (PKM2), which is associated with glycolytic pathway, converting glucose to lactate and generating ATP (Yang and Lu, 2015), is upregulated after OGD in murine dendritic cells (Jiang et al., 2018). Thus, we searched for important gene associated with hypoxia, and *PKM2* were significantly upregulated after OGD/R in the human cerebral organoids (**Figure 5A**). Vascular endothelial growth factor (*VEGF*) was relatively upregulated (1.83 fold; $P = 0.15$) and hypoxia inducible factor 1A (*HIF1A*) were not clearly changed after OGD/R in human cerebral organoids (**Figure 5A**). BV2 cells, hTERT-BJ, U251MG, and HepG2 were cultured and treated under OGD/R. *Pkm2* and *PKM2* did not show significant change after OGD/R (**Figure 5B**). These results suggest that upregulation of *PKM2* of human cerebral organoids might be specific to humans. *VEGFA* upregulation was significantly verified in hTERT-BJ and U251MG, *Vegfa* was upregulated in BV2 cells, and *Vegfa* and *VEGFA* did not change in murine primary cultured neurons

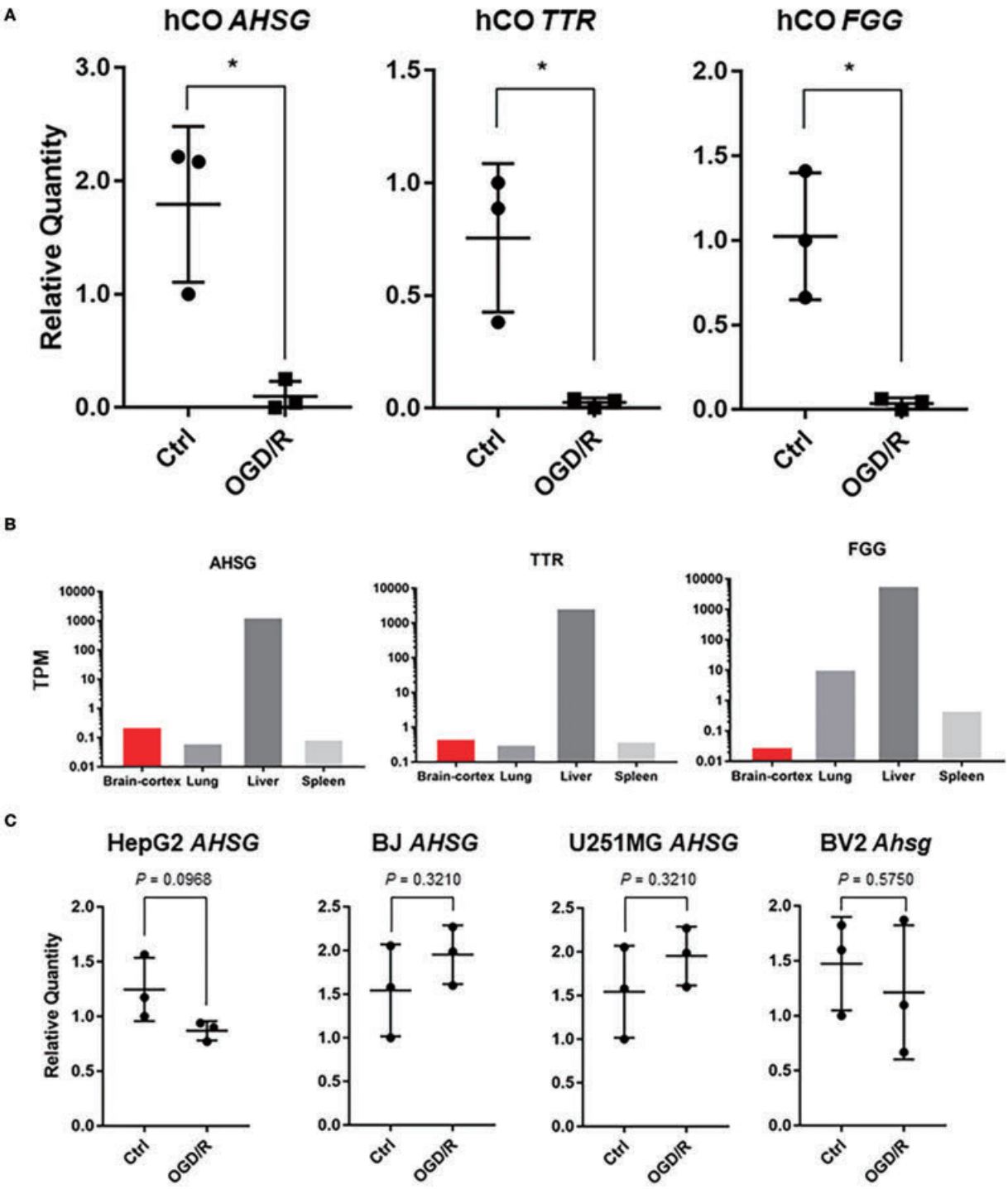


FIGURE 3 | (A) qPCR analysis of cerebral organoids [control vs. OGD/R condition]. Gene expression analysis of *AHSG*, *TTR,* and *FGG* on non-treated organoids (Ctrl) vs. organoids after OGD/R (OGD/R). The *P*-values were 0.0138 (*AHSG*), 0.0185 (*TTR*), and 0.0104 (*FGG*). $^*P < 0.05$ vs. ctrl. **(B)** Tissue-specific gene expression levels using GTEx database. *AHSG*, *TTR,* and *FGG.* The Y-axis indicates transcripts per million (TPM). **(C)** After OGD/R, expression level of *AHSG*, validated using hepatocellular carcinoma (HepG2), telomerase reverse transcriptase immortalized fibroblasts (hTERT-BJ), glioblastoma (U251MG), and murine microglia (BV2). *TTR* and *FGG* were not detected in BV2 cells and *TTR* was not detected in hTERT-BJ.

TABLE 3 | KEGG pathway of differentially expressed genes.

Term	Count	Genes	FDR	*P*-value
Vitamin digestion and absorption	4	APOA4, APOB, APOA1, RBP2	0.058366266	6.24E-05
Fat digestion and absorption	4	APOA4, APOB, APOA1, FABP1	0.332117918	3.56E-04
PPAR signaling pathway	4	APOA2, APOA1, APOC3, FABP1	1.618503332	0.0017426
Complement and coagulation cascades	4	FGG, FGA, FGB, SERPINA1	1.760575716	0.0018968
Platelet activation	3	FGG, FGA, FGB	54.24273463	0.0801693

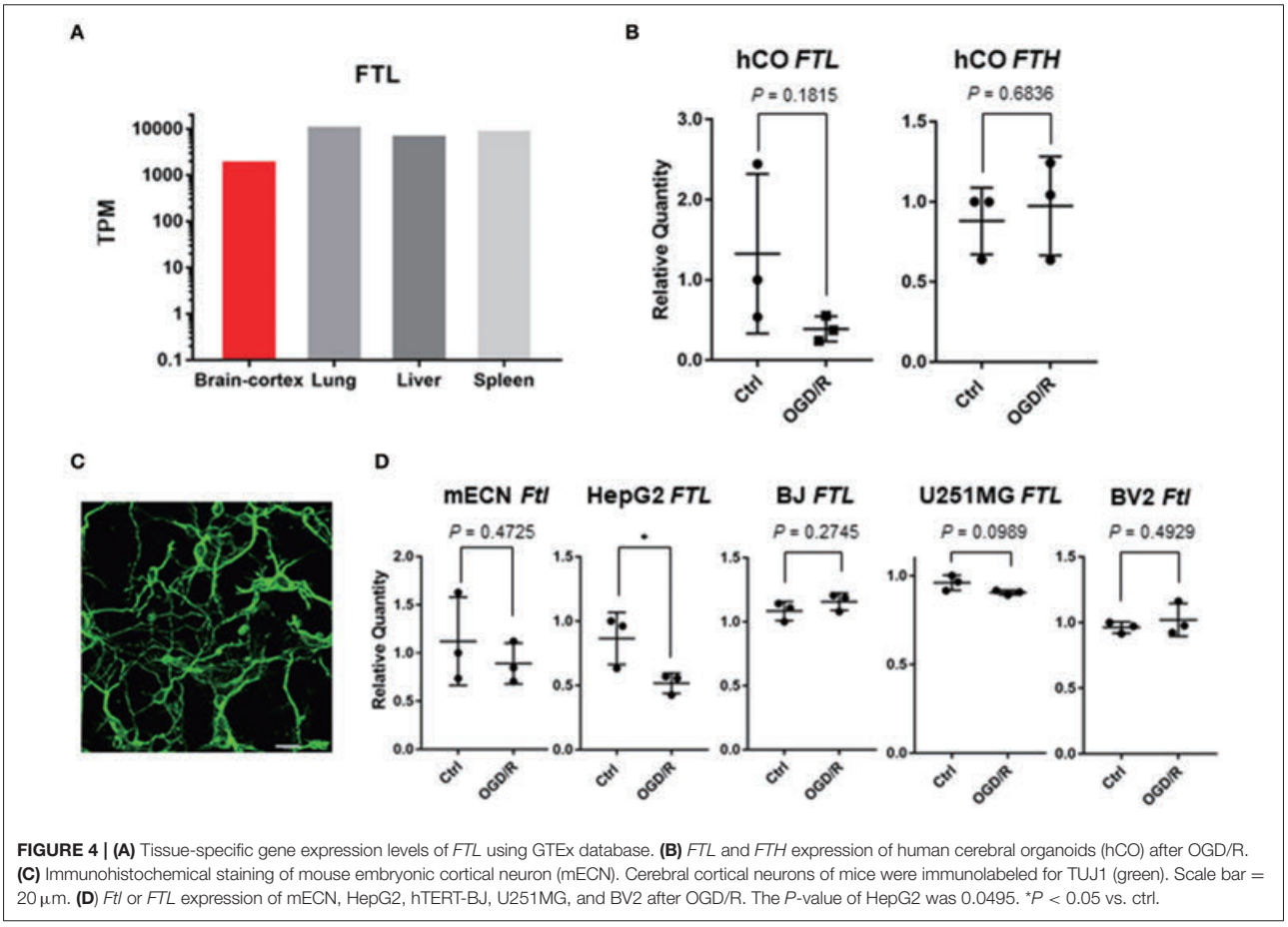


FIGURE 4 | (A) Tissue-specific gene expression levels of *FTL* using GTEx database. **(B)** *FTL* and *FTH* expression of human cerebral organoids (hCO) after OGD/R. **(C)** Immunohistochemical staining of mouse embryonic cortical neuron (mECN). Cerebral cortical neurons of mice were immunolabeled for TUJ1 (green). Scale bar = 20 μm. **(D)** *Ftl* or *FTL* expression of mECN, HepG2, hTERT-BJ, U251MG, and BV2 after OGD/R. The *P*-value of HepG2 was 0.0495. $^*P < 0.05$ vs. ctrl.

and HepG2 (**Figure 5C**). We found that the change of hypoxic marker genes independently showed various patterns in human cerebral organoids. We focused on some genes associated with transcriptional factors. Because prolyl hydroxylase 3 (PHD3, encoded by *EGLN3* gene) works as a co-activator of PKM2 and makes the connectivity of HIF1A and PKM2 strong (Luo et al., 2011), we validated expression of *PHD3/Egln3*. *PHD3* was relatively upregulated in organoids (1.74 fold; $P = 0.05$) (**Figure 6A**). According to pathway analysis of PPAR, we validated relatively downregulated apolipoprotein A1 *(APOA1)* (3.86 fold; $P = 0.35$), apolipoprotein A2 (*APOA2*) (33.28 fold; $P = 0.06$), apolipoprotein C3 (*APOC3*) (20.12 fold; $P = 0.05$), and fatty acid binding protein 1 (*FABP1*) (25.20 fold; $P = 0.08$) (**Figure 6B**). The expression level of *PPARA* as a nuclear transcription factor was upregulated in BV2 cells and was not altered in other cell types (**Figure 6C**). Because the RE1-silencing transcription factor (*REST*) as another nuclear transcription factor works against the oxidative stress of aging brain and Alzheimer's Disease using immunostaining of human brain (Lu et al., 2014), we tested the expression level of *REST* and it was not clearly changed (**Figure 6D**).

DISCUSSION

In vivo models reflect many factors including central nervous system, vascular system, and immune systems, and the interaction and physiological systems make interpretations of the

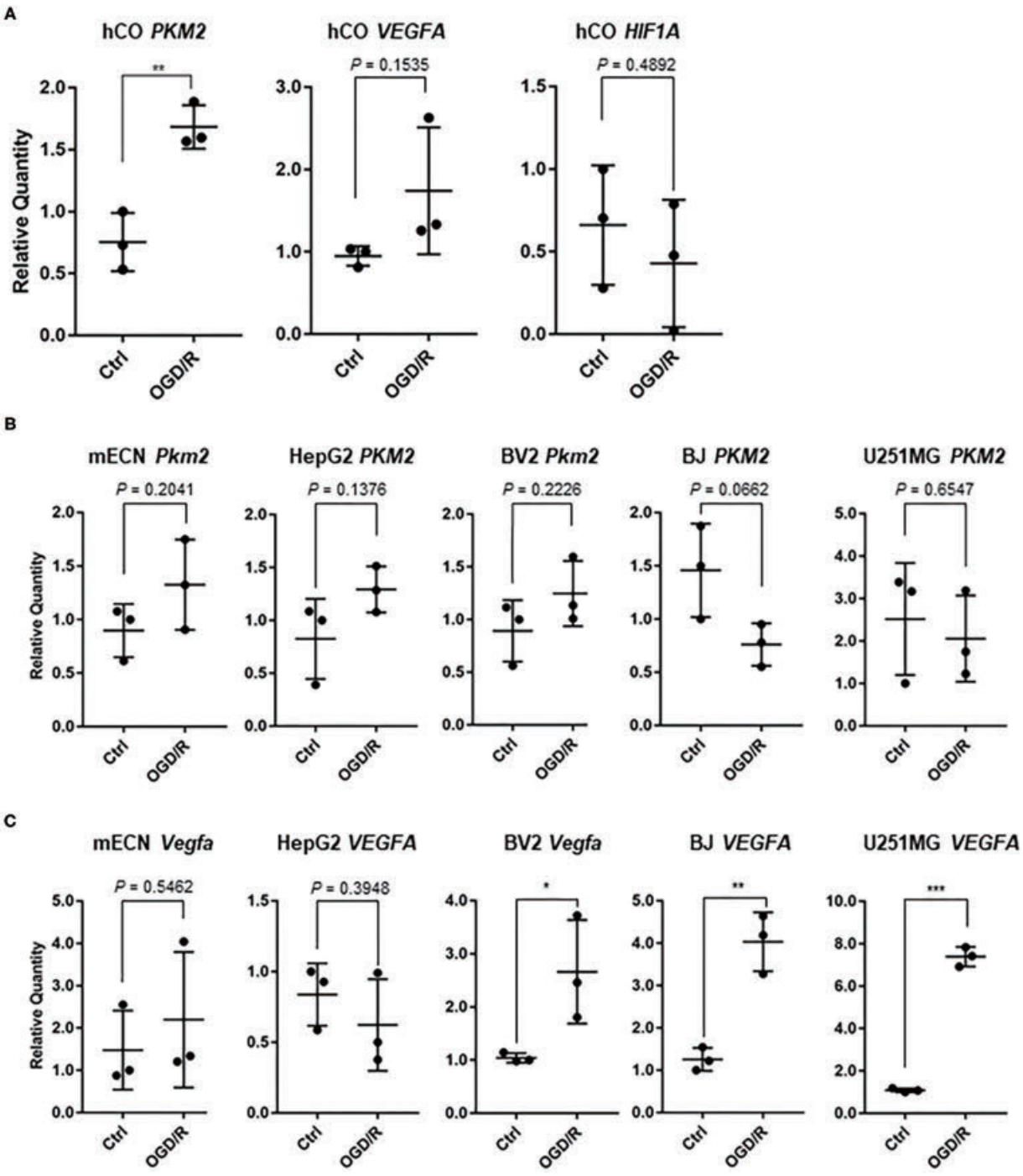


FIGURE 5 | (A) In human cerebral organoids (hCO), expression of hypoxic marker genes; *PKM2*, *VEGFA,* and *HIF1A* after OGD/R. The *P*-value of *PKM2* was 0.0054. $^{**}P < 0.001$ vs. ctrl. **(B)** After OGD/R, expression level of *PKM2* and *Pkm2* were validated using human cerebral organoids, mouse embryonic cortical neuron (mECN), hepatocellular carcinoma (HepG2), telomerase reverse transcriptase immortalized fibroblasts (hTERT-BJ), and glioblastoma (U251MG). **(C)** After OGD/R, expression level of *VEGFA* and *Vegfa* were validated using human cerebral organoid, mECN, HepG2, hTERT-BJ, and U251MG. The *P*-values were 0.0457 (BV2), 0.0030 (BJ), and < 0.0001 (U251MG). $^{*}P < 0.05$, $^{**}P < 0.001$, $^{***}P < 0.0001$ vs. ctrl.

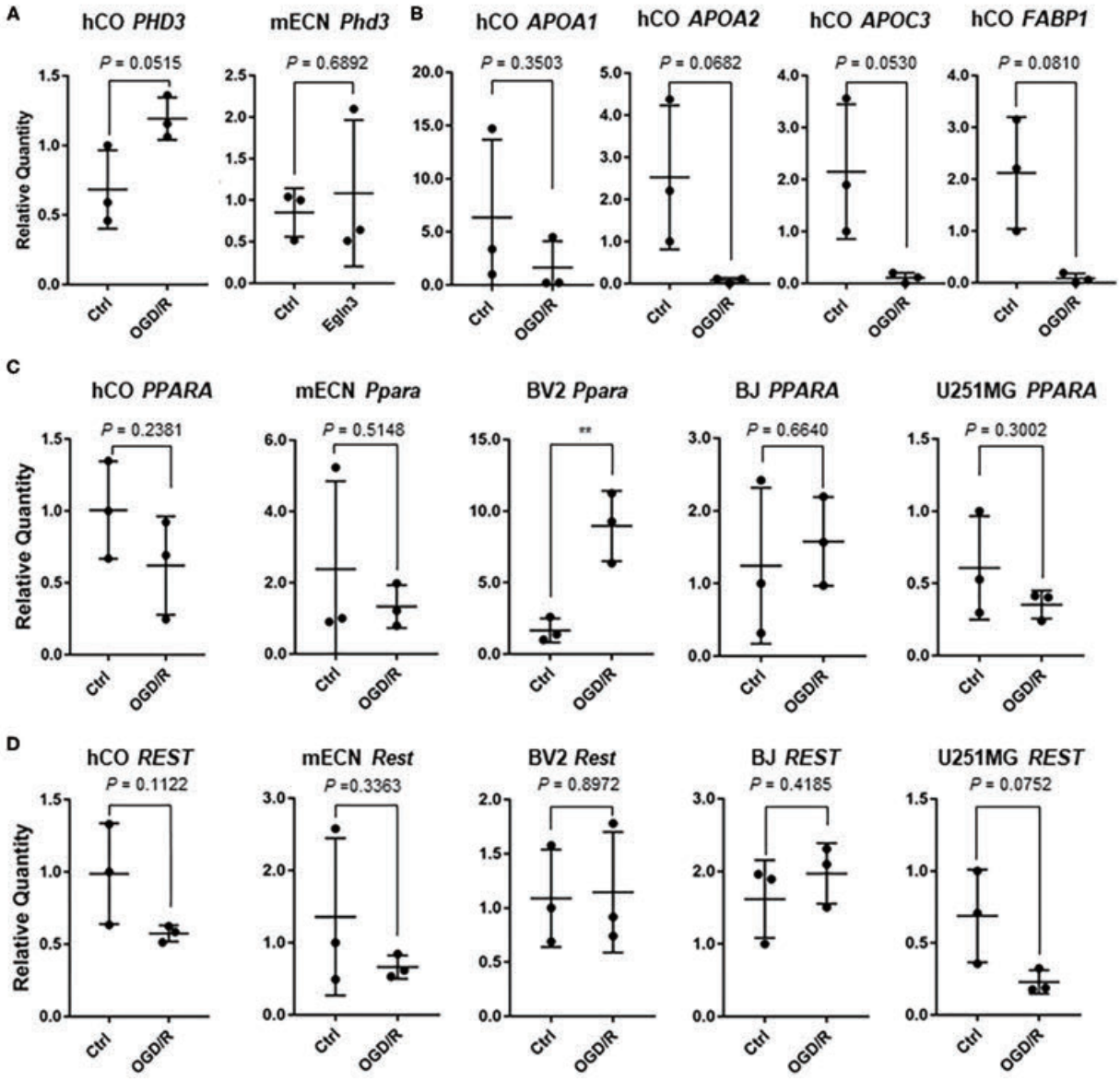


FIGURE 6 | (A) Expression of *PHD3* in human cerebral organoid (hCO) and *Egln3* in mECN after OGD/R. **(B)** After OGD/R, expression level of *APOA1, APOA2, APOC3,* and *FABP1* were validated using hCO. **(C)** After OGD/R, expression level of *PPARA* and *Ppara* were validated using hCO, mECN, HepG2, hTERT-BJ, and U251MG. The *P*-value of BV2 was 0.0082. $^{**}P < 0.001$ vs. ctrl. **(D)** After OGD/R, expression level of *REST* and *Rest* were validated using hCO, mECN, HepG2, hTERT-BJ, and U251MG.

results complex. On the contrary, *in vitro* 2D models consist of homotypic cell populations without cellular diversity and are unlikely to represent tissue architecture of actual tissues. Therefore, development and application of *in vitro* 3D model organoids are required to interpret the actual organ systems and cellular reactions. Additionally, *in vitro* maturation of human iPSC-derived 3D organoids could be a useful model for studying the pathophysiology of disease. Some genes of 3D cultured chondrocytes are highly increased and other genes are decreased compared with 2D models (Caron et al., 2012), and kidney organoids are more sensitive to nephrotoxicity than 2D cultured cell lines (Astashkina et al., 2012). Although human cerebral organoids lack fully matured neuronal tissue, our model still provides a biologically plausible means to study ischemic injury response in the human brain tissue. By using human cerebral organoids combined with series of transcriptome analysis, we revealed that *TTR*, *AHSG,* and *FGG* were downregulated after OGD/R. Although we were able to detect a downregulation of

FTL in human cerebral organoids after OGD/R by using RNA-seq, we found no significant change in q-PCR. Simultaneously we validated unaltered *FTL* in non-neural cells (fibroblasts and glioblastoma), murine microglia, and primary cultured murine cortical neurons after OGD/R. A larger sample size of cerebral organoids is preferable for validating *FTL* expression level because there are differences in maturity among individual organoids. Although *FTL* was more expressed than *AHSG*, *TTR,* and *FGG* using gene expression data for normal human tissues (**Figures 3B**, **4A**), it remains debatable whether iron itself and iron accumulation were beneficial or harmful after experimental stroke. Some groups have shown that brain damage after ischemia is correlated with overloaded iron (Castellanos et al., 2002; Mehta et al., 2004), while others suggest that iron storage is not evil (Christensen et al., 2002; Millerot et al., 2005). Therefore, further investigations are required to clarify the role of iron and iron storage after ischemic stroke.

We revealed that *FTL*, *TTR*, *AHSG,* and *FGG* are downregulated after OGD/R, but the distribution of FTL, TTR, AHSG, and FGG in human tissue is independent (**Figures 3B**, **4A**). While FTL, TTR, AHSG, and FGG were abundantly expressed in liver, TTR, AHSG, and FGG were not sufficient in brain (**Figures 3B**, **4A**). To validate the reaction of hepatocytes, we used HepG2 cells and *FTL* and *AHSG* tended to downregulate after OGD/R (**Figures 3C**, **4D**). It could be possible that hepatocytes or hepatocyte-like cells co-exist in the human cerebral organoids. Although cell types of the human cerebral organoids are derived from ectodermal origin (Quadrato et al., 2017), hepatic cells are derived from endoderm origin. There might be co-existing hepatic cells or hepatocyte-like cells in the human cerebral organoids we used in this study, and this might be potential limitations of the current study.

We revealed the expression level of *PKM2* of the human cerebral organoids after OGD/R was predominantly elevated. On the contrary the expression level of *PKM2* of other cells after OGD/R was not evidently changed. *Pkm2*, both mRNA and protein, is upregulated after OGD in murine dendritic cells, which are immune-modulatory cells that do not exist in our human cerebral organoids (Jiang et al., 2018). These results suggest that the upregulation of *PKM2* might be useful as a hypoxic marker in human cerebral organoids, while an expression of *PKM2* might be controversial in other cell types. Although PKM2 plays protective role in acute phase of focal ischemic stroke model of mice (Chen et al., 2018), loss of Pkm2 using *Pkm2* knockout mice also protects ischemic tissues by maintaining mitochondrial biogenesis (Hauck et al., 2020). Thus, the role of *PKM2* upon OGD/R insults is still controversial. *PKM2* is a target gene of HIF-1α, regulates gene expression of proteins associated with the glycolytic pathway and can accelerate metabolic reprogramming (Luo et al., 2011). Furthermore, *PHD3/Egln3* induced by hypoxia works as a co-activator of *PKM2* by HIF pathway (Pescador et al., 2005). *PHD3/Egln3* is relatively upregulated after OGD/R. This might accelerate metabolic switch and lead to decreased high-density lipoproteins (HDLs) including *APOA1*, *APOA2,* and *APOC3*, all of which showed a trend of downregulation upon OGD/R (**Figure 6B**). Therefore, PKM2 may act synergistically with PHD/Hif-1α in our OGD/R model. On the other hand, although some genes about fatty acid were relatively downregulated (**Figure 6B**), *PPARA* downregulation was validated only in murine microglia (**Figure 6C**). The reason why *PPARA* was unchanged in cerebral organoids might be due to a lack of microglia derived from mesoderm. Novel system of cerebral organoids including microglia will resolve this limitation. REST is a transcriptional factor as a silencer, concerning neurogenesis, neuronal differentiation, and growth (Paquette et al., 2000). We validated relatively downregulated *REST* in cerebral organoids, not murine primary cultured neurons (**Figure 6D**). Because some genes about neuronal movement, growth, and survival were upregulated after OGD/R using RNA-seq data set, *REST* downregulation might lead to neurogenesis and neuronal differentiation. These results could be specific for 3D embryonic culture system. PKM2 is inhibited by PPARγ in two breast cancer cell lines, and this inhibition might decrease ATP levels and avoid apoptosis (Shashni et al., 2013). We showed PPAR signaling pathway is associated (**Figure 2**) and *PKM2* was upregulated after OGD/R in human cerebral organoids (**Figure 5A**). This could suggest that PPARγ is downregulated and *PKM2* is upregulated subsequently after ischemia. The PPARγ agonist conducts neuroprotection against focal ischemia in the rat brain (Zhao et al., 2005). Dual PPARα/γ agonist improves stroke outcome after transient cerebral ischemia in mice (Boujon et al., 2019). Therefore, therapeutic agents related lipid metabolism, including PPAR agonist, may be effective for ischemia.

Our study did not clarify either the change of common hypoxic marker genes (*VEGF* and *HIF1A*) of human cerebral organoids occurs after OGD/R (**Figures 5A,C**) or a change in other hypoxic marker genes (*LDHA*, *CA9,* and *PGK1*) (**Supplementary Figure 1C**). Lacking vascular system, brain organoids necrotize in their core as they grow and exist around the core healthy tissues *in vitro* (Lancaster, 2018). To resolve the problem of central core necrosis of brain organoids, they are transplanted into the mouse brain as vascularized model of brain organoids *in vivo* (Mansour et al., 2018). *In vitro* organotypic slice culture improved long-term survival (Giandomenico et al., 2021), and air–liquid interface culture for cerebral organoids improved survival, axon outgrowth, and decreased apoptosis rate (Giandomenico et al., 2019). It might be the reason why *VEGFA* and *HIF1A* did not significantly change, and the lack of vascular system was one of the limitation of cerebral organoids. Although *VEGF* and *HIF1A* were not markedly upregulated, we validated upregulated *HSPA8*, stress-induced protein: heat-shock cognate protein 70 (HSC70) coding gene, using RNA-seq (**Table 1**). Because HSC70 assembles after cerebral ischemia of rats (Hu et al., 1998) and HSC70 involves diverse proteostasis mechanisms including chaperone mediated autophagy which contributes to oxidative stress response (Kiffin et al., 2004; Loeffler et al., 2016), upregulated *HSPA8* coding HSC70 indicated the oxidative stress of our model was existing.

There are a number of groups trying to apply organoid system to study isolated organ model in a dish. For example, James Hudson and colleagues showed that mild-hypoxia and/or metabolic changes are critical environmental triggers

for maturation of cardiac cells using human cardiac organoids (Mills et al., 2019). In their experiments, they treated the human cardiac organoids by changing the serum free media, lowering glucose concentration, and increasing the concentration of fatty acid. Palmitate was used as a fatty acid substrate, as it is one of the most abundant fatty acids circulating during the neonatal period (Bougnères et al., 1982). Importantly, Hesham Sadek and colleagues discovered mild-hypoxia to be a stimulus for the metabolic switch from carbohydrates to fatty acids in a neonatal mouse cardiac tissue (Puente et al., 2014) and adult mouse cardiac tissue (Nakada et al., 2017), and therefore human cardiac organoids could model the postnatal switch in metabolic substrates from carbohydrates to fatty acids. In our experiment, we removed glucose and oxygen without substituting fatty acid. Thus, the human brain organoids do not get substrates even if the gene expression may alter metabolic switch from carbohydrates to fatty acids. The difference of *FTL* expression between mild-hypoxia and severe-hypoxia **Figure 4D**, **Supplementary Figure 1B** might suggest mild-hypoxia tends to metabolize fatty acid, compared with severe-hypoxia. Taken together, our experimental condition is not comparable to environmental changes that induce postnatal metabolic shift; instead, it mimics ischemic tissue. More detailed time course about hypoxia will be needed for further investigation in the future. It is possible that neural cells in the brain and cardiac muscle cells in the heart may respond differentially to severe hypoxia. Therefore, in order to adapt their findings from the cardiac organoids (Mills et al., 2017) to our brain organoids, neural cells would need to have similar capability to respond to same environmental stimuli. Further investigations are required to understand how neural and non-neural cells in the brain respond to pathogenic severe hypoxia.

Our goal here is to extract the direct effects on neural tissues out of spatio-temporally complex system organized by various cell types. Our cerebral organoids are free from the effect of blood vessels and blood-derived inflammatory cells invading through blood vessels. Therefore, our experimental model enabled us to evaluate the direct effect of deprivation of both oxygen and glucose followed by reoxygenation on brain tissue. Although we are aware of the limitations of applying the human cerebral organoid as a model for human ischemic stroke, there are expectations for the potential development and advancement for further applications of our model.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ddbj.nig.ac.jp/, DRA010225.

ETHICS STATEMENT

The animal study was reviewed and approved by Animal Care Committee of Nara Medical University.

AUTHOR CONTRIBUTIONS

NIw, TM, KSu, and EM designed the study. NIw, TM, NIg, KK, NM, YMS, TS, SKo, TT, MN, and RN conducted the research. NIw, TM, NIg, KK, YMS, MM, NE, TK, TI, KSa, HK, YS, WK, AW, YN, EM, and KSu analyzed the data. NIw, TM, TT, KSu, and EM wrote the paper. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This work was supported by grants from JSPS KAKENHI (JP17H07031, JP20H03199 to EM, JP19K16925 to TM, JP19K07978 to TK), AMED Brain/MINDS Beyond (JP20dm0307032 to EM), AMED The Program for Technological Innovation of Regenerative Medicine (JP20bm0704039 to TM), AMED Osaka University Seeds (A) to TM, Takeda Science Foundation to EM and TM, Kanzawa Medical Research Foundation to EM, Uehara Memorial Foundation to EM, Nakatomi Foundation to EM, Konica Minolta Science and Technology Foundation to EM, Naito Foundation to EM, MSD Life Science Foundation to EM, Mochida Memorial Foundation for Medical and Pharmaceutical Research to EM, SENSHIN Medical Research Foundation to EM, Terumo Foundation for Life Sciences and Arts to EM, Nara Kidney Disease Research Foundation to EM, Novartis Research Grants to EM and KSu, Sumitomo Dainippon Pharma Research Grant to TM, Nara Medical University Grant-in-Aid for Collaborative Research Projects to KSu, Nara Medical University Grant-in-Aid for Young Scientists to TM, and by unrestricted funds provided to EM from Dr. Taichi Noda (KTX Corp., Aichi, Japan) and Dr. Yasuhiro Horii (Koseikai, Nara, Japan). The authors declare that this study received funding from parties mentioned above. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or decision to submit the article for publication.

ACKNOWLEDGMENTS

The authors thank Keren-Happuch E. Fan Fen for her critical reading of the manuscript. hTERT-BJ were kindly gifted by Keiji Suzuki (Department of Radiation Medical Sciences, Atomic Bomb Disease Institute, Nagasaki University, Nagasaki, Japan).

This manuscript has been released as a Pre-Print at bioRxiv (Iwasa et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncel.2021.605030/full#supplementary-material

Supplementary Figure 1 | (A) Immunohistochemical staining of mouse embryonic cortical neuron (mECN). Cerebral cortical neurons of mice were immunolabeled for TUJ1 (green). Scale bar = 100 μm. **(B)** Expression level of *Ftl* and *FTL* under mild hypoxia (i.e., O_2 10% for 24 h). **(C)** In hCO, expression of other hypoxic marker genes; *LDHA*, *CA9*, and *PGK1* after OGD/R.

Supplementary Table 1 | Primer pairs for qPCRs used in this study.

REFERENCES

Astashkina, A. I., Mann, B. K., Prestwich, G. D., and Grainger, D. W. (2012). Comparing predictive drug nephrotoxicity biomarkers in kidney 3-D primary organoid culture and immortalized cell lines. *Biomaterials* 33, 4712–4721. doi: 10.1016/j.biomaterials.2012.03.001

Aukkarasongsup, P., Haruyama, N., Matsumoto, T., Shiga, M., and Moriyama, K. (2013). Periostin inhibits hypoxia-induced apoptosis in human periodontal ligament cells via TGF-beta signaling. *Biochem. Biophys. Res. Commun.* 441, 126–132. doi: 10.1016/j.bbrc.2013.10.027

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., et al. (2019). Heart disease and stroke statistics-2019 update: a report from the american heart association. *Circulation* 139, e56–e528. doi: 10.1161/cir.0000000000000659

Bougnères, P. F., Karl, I. E., Hillman, L. S., and Bier, D. M. (1982). Lipid transport in the human newborn. Palmitate and glycerol turnover and the contribution of glycerol to neonatal hepatic glucose output. *J. Clin. Invest.* 70, 262–270. doi: 10.1172/jci110613

Boujon, V., Uhlemann, R., Wegner, S., Wright, M. B., Laufs, U., Endres, M., et al. (2019). Dual PPARα/γ agonist aleglitazar confers stroke protection in a model of mild focal brain ischemia in mice. *J. Mol. Med. (Berl).* 97, 1127–1138. doi: 10.1007/s00109-019-01801-0

Caron, M. M., Emans, P. J., Coolsen, M. M., Voss, L., Surtel, D. A., Cremers, A., et al. (2012). Redifferentiation of dedifferentiated human articular chondrocytes: comparison of 2D and 3D cultures. *Osteoarthritis Cartilage* 20, 1170–1178. doi: 10.1016/j.joca.2012.06.016

Castellanos, M., Puig, N., Carbonell, T., Castillo, J., Martinez, J., Rama, R., et al. (2002). Iron intake increases infarct volume after permanent middle cerebral artery occlusion in rats. *Brain Res.* 952, 1–6. doi: 10.1016/s0006-8993(02)03179-7

Catanese, L., Tarsia, J., and Fisher, M. (2017). Acute ischemic stroke therapy overview. *Circ. Res.* 120, 541–558. doi: 10.1161/circresaha.116.309278

Chen, D., Wei, L., Liu, Z. R., Yang, J. J., Gu, X., Wei, Z. Z., et al. (2018). Pyruvate kinase M2 increases angiogenesis, neurogenesis, and functional recovery mediated by upregulation of STAT3 and focal adhesion kinase activities after ischemic stroke in adult mice. *Neurotherapeutics* 15, 770–784. doi: 10.1007/s13311-018-0635-2

Cheung, E. Y., Uitte de Willige, S., Vos, H. L., Leebeek, F. W., Dippel, D. W., Bertina, R. M., et al. (2008). Fibrinogen gamma' in ischemic stroke: a case-control study. *Stroke* 39, 1033–1035. doi: 10.1161/strokeaha.107.495499

Christensen, H., Boysen, G., Johannesen, H. H., Christensen, E., and Bendtzen, K. (2002). Deteriorating ischaemic stroke. cytokines, soluble cytokine receptors, ferritin, systemic blood pressure, body temperature, blood glucose, diabetes, stroke severity, and CT infarction-volume as predictors of deteriorating ischaemic stroke. *J. Neurol. Sci.* 201, 1–7. doi: 10.1016/s0022-510x(02)00160-0

Clevers, H. (2016). Modeling development and disease with organoids. *Cell* 165, 1586–1597. doi: 10.1016/j.cell.2016.05.082

Dabrowska, A. M., Tarach, J. S., Wojtysiak-Duma, B., and Duma, D. (2015). Fetuin-A (AHSG) and its usefulness in clinical practice. Review of the literature. *Biomed. Pap. Med. Fac. Univ. Palacky. Olomouc. Czech. Repub.* 159, 352–359. doi: 10.5507/bp.2015.018

Dorado, L., Millan, M., and Davalos, A. (2014). Reperfusion therapies for acute ischemic stroke: an update. *Curr. Cardiol. Rev.* 10, 327–335. doi: 10.2174/1573403x10666140320144637

Fisher, M., and Bastan, B. (2012). Identifying and utilizing the ischemic penumbra. *Neurology* 79(13 Suppl 1), S79–S85. doi: 10.1212/WNL.0b013e3182695814

Giandomenico, S. L., Mierau, S. B., Gibbons, G. M., Wenger, L. M. D., Masullo, L., Sit, T., et al. (2019). Cerebral organoids at the air-liquid interface generate diverse nerve tracts with functional output. *Nat. Neurosci.* 22, 669–679. doi: 10.1038/s41593-019-0350-2

Giandomenico, S. L., Sutcliffe, M., and Lancaster, M. A. (2021). Generation and long-term culture of advanced cerebral organoids for studying later stages of neural development. *Nat. Protoc.* 16, 579–602. doi: 10.1038/s41596-020-00433-w

Hauck, L., Dadson, K., Chauhan, S., Grothe, D., and Billia, F. (2020). Inhibiting the Pkm2/b-catenin axis drives *in vivo* replication of adult cardiomyocytes following experimental MI. *Cell Death Differ.* 28, 1398–1417. doi: 10.1038/s41418-020-00669-9

Hossmann, K. A. (1998). Experimental models for the investigation of brain ischemia. *Cardiovasc. Res.* 39, 106–120. doi: 10.1016/s0008-6363(98)00075-3

Hu, B. R., Park, M., Martone, M. E., Fischer, W. H., Ellisman, M. H., and Zivin, J. A. (1998). Assembly of proteins to postsynaptic densities after transient cerebral ischemia. *J. Neurosci.* 18, 625–633. doi: 10.1523/jneurosci.18-02-00625.1998

Iwasa, N., Matsui, T. K., Morikawa, N., Sakaguchi, Y. M., Shiota, T., Iguchi, N., et al. (2020). Oxygen-glucose deprivation and reoxygenation on human cerebral organoids alters expression related to lipid metabolism. *bioRxiv [Preprint].* doi: 10.1101/2020.03.31.017335

Jiang, H., Zhang, S., Song, T., Guan, X., Zhang, R., and Chen, X. (2018). Trichostatin a protects dendritic cells against oxygen-glucose deprivation via the SRSF3/PKM2/glycolytic pathway. *Front. Pharmacol.* 9:612. doi: 10.3389/fphar.2018.00612

Kido, M., Nakamura, Y., Nemoto, K., Takahashi, T., Aleksic, B., Furuichi, A., et al. (2014). The polymorphism of *YWHAE*, a gene encoding 14-3-3epsilon, and brain morphology in schizophrenia: a voxel-based morphometric study. *PLoS ONE* 9:e103571. doi: 10.1371/journal.pone.0103571

Kiffin, R., Christian, C., Knecht, E., and Cuervo, A. M. (2004). Activation of chaperone-mediated autophagy during oxidative stress. *Mol. Biol. Cell* 15, 4829–4840. doi: 10.1091/mbc.e04-06-0477

Knowland, D., Arac, A., Sekiguchi, K. J., Hsu, M., Lutz, S. E., Perrino, J., et al. (2014). Stepwise recruitment of transcellular and paracellular pathways underlies blood-brain barrier breakdown in stroke. *Neuron* 82, 603–617. doi: 10.1016/j.neuron.2014.03.003

Lancaster, M. A. (2018). Brain organoids get vascularized. *Nat. Biotechnol.* 36, 407–408. doi: 10.1038/nbt.4133

Lancaster, M. A., Renner, M., Martin, C. A., Wenzel, D., Bicknell, L. S., Hurles, M. E., et al. (2013). Cerebral organoids model human brain development and microcephaly. *Nature* 501, 373–379. doi: 10.1038/nature12517

Loeffler, D. A., Klaver, A. C., Coffey, M. P., Aasly, J. O., and LeWitt, P. A. (2016). Age-related decrease in heat shock 70-kDa protein 8 in cerebrospinal fluid is associated with increased oxidative stress. *Front. Aging Neurosci.* 8:178. doi: 10.3389/fnagi.2016.00178

Lu, T., Aron, L., Zullo, J., Pan, Y., Kim, H., Chen, Y., et al. (2014). REST and stress resistance in ageing and Alzheimer's disease. *Nature* 507, 448–454. doi: 10.1038/nature13163

Luo, W., Hu, H., Chang, R., Zhong, J., Knabel, M., O'Meally, R., et al. (2011). Pyruvate kinase M2 is a PHD3-stimulated coactivator for hypoxia-inducible factor 1. *Cell* 145, 732–744. doi: 10.1016/j.cell.2011.03.054

Mansour, A. A., Gonçalves, J. T., Bloyd, C. W., Li, H., Fernandes, S., Quang, D., et al. (2018). An *in vivo* model of functional and vascularized human brain organoids. *Nat. Biotechnol.* 36, 432–441. doi: 10.1038/nbt.4127

Mehta, S. H., Webb, R. C., Ergul, A., Tawfik, A., and Dorrance, A. M. (2004). Neuroprotection by tempol in a model of iron-induced oxidative stress in acute ischemic stroke. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 286, R283–R288. doi: 10.1152/ajpregu.00446.2002

Millerot, E., Prigent-Tessier, A. S., Bertrand, N. M., Faure, P. J., Mossiat, C. M., Giroud, M. E., et al. (2005). Serum ferritin in stroke: a marker of increased body iron stores or stroke severity? *J. Cereb. Blood Flow Metab.* 25, 1386–1393. doi: 10.1038/sj.jcbfm.9600140

Mills, R. J., Parker, B. L., Quaife-Ryan, G. A., Voges, H. K., Needham, E. J., Bornot, A., et al. (2019). Drug screening in human PSC-cardiac organoids identifies pro-proliferative compounds acting via the mevalonate pathway. *Cell Stem Cell* 24, 895.e896–907.e896. doi: 10.1016/j.stem.2019.03.009

Mills, R. J., Titmarsh, D. M., Koenig, X., Parker, B. L., Ryall, J. G., Quaife-Ryan, G. A., et al. (2017). Functional screening in human cardiac organoids reveals a metabolic mechanism for cardiomyocyte cell cycle arrest. *Proc. Natl. Acad. Sci. U.S.A.* 114, E8372–E8381. doi: 10.1073/pnas.1707316114

Muhoberac, B. B., and Vidal, R. (2019). Iron, ferritin, hereditary ferritinopathy, and neurodegeneration. *Front. Neurosci.* 13:1195. doi: 10.3389/fnins.2019.01195

Nakada, Y., Canseco, D. C., Thet, S., Abdisalaam, S., Asaithamby, A., Santos, C. X., et al. (2017). Hypoxia induces heart regeneration in adult mice. *Nature* 541, 222–227. doi: 10.1038/nature20173

Paquette, A. J., Perez, S. E., and Anderson, D. J. (2000). Constitutive expression of the neuron-restrictive silencer factor (NRSF)/REST in differentiating neurons disrupts neuronal gene expression and causes axon pathfinding errors *in vivo*. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12318–12323. doi: 10.1073/pnas.97.22.12318

Pasca, A. M., Sloan, S. A., Clarke, L. E., Tian, Y., Makinson, C. D., Huber, N., et al. (2015). Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nat. Methods* 12, 671–678. doi: 10.1038/nmeth.3415

Pescador, N., Cuevas, Y., Naranjo, S., Alcaide, M., Villar, D., Landázuri, M. O., et al. (2005). Identification of a functional hypoxia-responsive element that regulates the expression of the egl nine homologue 3 (egln3/phd3) gene. *Biochem. J.* 390(Pt 1), 189–197. doi: 10.1042/bj20042121

Puente, B. N., Kimura, W., Muralidhar, S. A., Moon, J., Amatruda, J. F., Phelps, K. L., et al. (2014). The oxygen-rich postnatal environment induces cardiomyocyte cell-cycle arrest through DNA damage response. *Cell* 157, 565–579. doi: 10.1016/j.cell.2014.03.032

Quadrato, G., Nguyen, T., Macosko, E. Z., Sherwood, J. L., Min Yang, S., Berger, D. R., et al. (2017). Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* 545, 48–53. doi: 10.1038/nature22047

Reza Sailani, M., Jahanbani, F., Nasiri, J., Behnam, M., Salehi, M., Sedghi, M., et al. (2017). Association of AHSG with alopecia and mental retardation (APMR) syndrome. *Hum. Genet.* 136, 287–296. doi: 10.1007/s00439-016-1756-5

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Shashni, B., Sakharkar, K. R., Nagasaki, Y., and Sakharkar, M. K. (2013). Glycolytic enzymes PGK1 and PKM2 as novel transcriptional targets of PPARγ in breast cancer pathophysiology. *J. Drug Target* 21, 161–174. doi: 10.3109/1061186x.2012.736998

Shen, D., Podolnikova, N. P., Yakubenko, V. P., Ardell, C. L., Balabiyev, A., Ugarova, T. P., et al. (2017). Pleiotrophin, a multifunctional cytokine and growth factor, induces leukocyte responses through the integrin Mac-1. *J. Biol. Chem.* 292, 18848–18861. doi: 10.1074/jbc.M116.773713

Spescha, R. D., Shi, Y., Wegener, S., Keller, S., Weber, B., Wyss, M. M., et al. (2013). Deletion of the ageing gene p66(Shc) reduces early stroke size following ischaemia/reperfusion brain injury. *Eur. Heart J.* 34, 96–103. doi: 10.1093/eurheartj/ehs331

Stefan, N., Hennige, A. M., Staiger, H., Machann, J., Schick, F., Kröber, S. M., et al. (2006). Alpha2-Heremans-Schmid glycoprotein/fetuin-A is associated with insulin resistance and fat accumulation in the liver in humans. *Diabetes Care* 29, 853–857. doi: 10.2337/diacare.29.04.06.dc05-1938

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tanaka, Y., Ando, Y., Kumamoto, T., Miyazaki, A., Nakamura, M., Nakayama, M., et al. (1994). Changed affinity of apolipoprotein AII to high density lipoprotein (HDL) in patients with familial amyloidotic polyneuropathy (FAP) type I. *Biochim. Biophys. Acta* 1225, 311–316. doi: 10.1016/0925-4439(94)90012-4

Toyo-oka, K., Shionoya, A., Gambello, M. J., Cardoso, C., Leventer, R., Ward, H. L., et al. (2003). 14-3-3epsilon is important for neuronal migration by binding to NUDEL: a molecular explanation for Miller-Dieker syndrome. *Nat. Genet.* 34, 274–285. doi: 10.1038/ng1169

Yang, W., and Lu, Z. (2015). Pyruvate kinase M2 at a glance. *J. Cell Sci.* 128, 1655–1660. doi: 10.1242/jcs.166629

Zhao, Y., Patzer, A., Gohlke, P., Herdegen, T., and Culman, J. (2005). The intracerebral application of the PPARgamma-ligand pioglitazone confers neuroprotection against focal ischaemia in the rat brain. *Eur. J. Neurosci.* 22, 278–282. doi: 10.1111/j.1460-9568.2005.04200.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Iwasa, Matsui, Iguchi, Kinugawa, Morikawa, Sakaguchi, Shiota, Kobashigawa, Nakanishi, Matsubayashi, Nagata, Kikuchi, Tanaka, Eura, Kiriyama, Izumi, Saito, Kataoka, Saito, Kimura, Wanaka, Nishimura, Mori and Sugie. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

3)

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-25560-0 OPEN

C9orf72-derived arginine-rich poly-dipeptides impede phase modifiers

Hitoki Nanaura[1,2,19], Honoka Kawamukai[3,4,19], Ayano Fujiwara[5,19], Takeru Uehara[5], Yuichiro Aiba[6], Mari Nakanishi[2], Tomo Shiota[1,2], Masaki Hibino[6], Pattama Wiriyasermkul[7,18], Sotaro Kikuchi[2], Riko Nagata[2], Masaya Matsubayashi[2], Yoichi Shinkai[8], Tatsuya Niwa[9], Taro Mannen[5], Naritaka Morikawa[2], Naohiko Iguchi[1], Takao Kiriyama[1], Ken Morishima[10], Rintaro Inoue[10], Masaaki Sugiyama[10], Takashi Oda[11,12], Noriyuki Kodera[13], Sachiko Toma-Fukai[14], Mamoru Sato[11], Hideki Taguchi[9], Shushi Nagamori[7,18], Osami Shoji[6], Koichiro Ishimori[3,15], Hiroyoshi Matsumura[5], Kazuma Sugie[1], Tomohide Saio[4,16]✉, Takuya Yoshizawa[5]✉ & Eiichiro Mori[2,17]✉

Nuclear import receptors (NIRs) not only transport RNA-binding proteins (RBPs) but also modify phase transitions of RBPs by recognizing nuclear localization signals (NLSs). Toxic arginine-rich poly-dipeptides from *C9orf72* interact with NIRs and cause nucleocytoplasmic transport deficit. However, the molecular basis for the toxicity of arginine-rich poly-dipeptides toward NIRs function as phase modifiers of RBPs remains unidentified. Here we show that arginine-rich poly-dipeptides impede the ability of NIRs to modify phase transitions of RBPs. Isothermal titration calorimetry and size-exclusion chromatography revealed that proline:arginine (PR) poly-dipeptides tightly bind karyopherin-β2 (Kapβ2) at 1:1 ratio. The nuclear magnetic resonances of Kapβ2 perturbed by PR poly-dipeptides partially overlapped with those perturbed by the designed NLS peptide, suggesting that PR poly-dipeptides target the NLS binding site of Kapβ2. The findings offer mechanistic insights into how phase transitions of RBPs are disabled in *C9orf72*-related neurodegeneration.

[1] Department of Neurology, Nara Medical University, Kashihara, Nara, Japan. [2] Department of Future Basic Medicine, Nara Medical University, Kashihara, Nara, Japan. [3] Graduate School of Chemical Sciences and Engineering, Hokkaido University, Sapporo, Hokkaido, Japan. [4] Graduate School of Medical Sciences, Tokushima University, Tokushima, Japan. [5] College of Life Sciences, Ritsumeikan University, Kusatsu, Shiga, Japan. [6] Department of Chemistry, Graduate School of Science, Nagoya University, Nagoya, Aichi, Japan. [7] Laboratory of Bio-Molecular Dynamics, Department of Collaborative Research, Nara Medical University, Kashihara, Nara, Japan. [8] Molecular Neurobiology Research Group, Biomedical Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan. [9] Cell Biology Center, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan. [10] Institute for Integrated Radiation and Nuclear Science, Kyoto University, Kyoto, Japan. [11] Graduate School of Medical Life Science, Yokohama City University, Yokohama, Japan. [12] Department of Life Science, Rikkyo University, Toshima ku, Tokyo, Japan. [13] Nano Life Science Institute (WPI-NanoLSI), Kanazawa University, Kanazawa, Japan. [14] Division of Materials Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara, Japan. [15] Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo, Hokkaido, Japan. [16] Institute of Advanced Medical Sciences, Tokushima University, Tokushima, Japan. [17] V-iCliniX Laboratory, Nara Medical University, Kashihara, Nara, Japan. [18] Present address: Department of Laboratory Medicine, The Jikei University School of Medicine, Tokyo, Japan. [19] These authors contributed equally: Hitoki Nanaura, Honoka Kawamukai, Ayano Fujiwara. ✉email: saio@tokushima-u.ac.jp; t-yosh@fc.ritsumei.ac.jp; emori@naramed-u.ac.jp

Low-complexity protein sequences (LC-domains), regions with little diversity in the amino acid composition, are often found in RNA-binding proteins (RBPs) and are prone to self-associate, to drive phase transitions into liquid-like or gel-like states[1–4]. Regulating self-association of RBPs, including fused in sarcoma (FUS), TAR DNA-binding protein of 43 kDa (TDP43), and other heterogeneous nuclear ribonucleoproteins (hnRNPs), to suppress the formation of pathogenic fibrils, is crucial in the prevention of neurodegenerative diseases[5,6]. Nuclear import receptor (NIR) karyopherin-β2 (Kapβ2) not only controls the nucleocytoplasmic distribution of RBPs but also acts as a phase modifier to regulate self-association of FUS by recognizing proline–tyrosine nuclear localization signal (NLS)[7–10]. NIRs Importinα/Importinβ1 complex (Impα/β1) also modify self-association of TDP43[8]. Decreased affinity of Kapβ2 and FUS leads to aberrant cytoplasmic localization of FUS[11], which is observed in amyotrophic lateral sclerosis (ALS) pathology[12,13].

A hexanucleotide repeat expansion in *C9orf72* is the most prevalent form of familial ALS and frontotemporal dementia (C9-ALS/FTD)[14]. Similar to other forms of ALS, mislocalization and aggregation of RBPs are observed in C9-ALS/FTD[12,13]. The main mechanisms of C9-ALS/FTD have been proposed as follows: *C9orf72* haploinsufficiency, toxic gain-of-function from repeat RNA or poly-dipeptides, or some combination of the above[15]. Repeat expansion in *C9orf72* produces five different poly-dipeptides, proline:arginine (PR), glycine:arginine (GR), proline:alanine (PA), glycine:proline (GP), and glycine:alanine (GA), through repeat-associated non-AUG (RAN) translation[16,17].

Arginine-rich, PR and GR, poly-dipeptides show similarly high toxic effect on membraneless organelles, proteins with LC-domains, and nucleocytoplasmic transport (NCT)[14,18,19]. Arginine-rich poly-dipeptides bind proteins with LC-domains and stabilize polymers formed by protein–protein interaction through LC-domains[2,4]. Existing genetic studies reveal that the repeat expansion in *C9orf72* gene disrupts NCT[20–22]. One mechanism of the disruption of NCT is that PR poly-dipeptides bind and stabilize phenylalanine:glycine-rich domains of nuclear pore proteins[3], although controversial results have also been reported[23].

Another proposed mechanism of the impairment of NCT is the toxicity of arginine-rich poly-dipeptides towards NIRs. Proteomic studies have indicated that NIRs are potential interactors of arginine-rich poly-dipeptides[2,4]. More recently, arginine-rich poly-dipeptides have been found to interact with NIRs and cause NCT deficit[24,25], although the molecular basis for the way in which arginine-rich poly-dipeptides affect NIRs function as phase modifiers remains elusive.

In this study, we investigated the effect of arginine-rich poly-dipeptides on phase modifiers through the use of multiple biochemical and biophysical methods. The approaches we utilized in this study include the following: (i) an assessment of phase transitions by droplet formation and hydrogel binding assay, (ii) an interaction analysis in a cell and a test tube, and (iii) a detailed solution nuclear magnetic resonance (NMR) analysis verified by molecular dynamics (MD) simulation. We describe here how arginine-rich poly-dipeptides impede phase modifiers, providing mechanistic insights into *C9orf72*-related neurodegeneration.

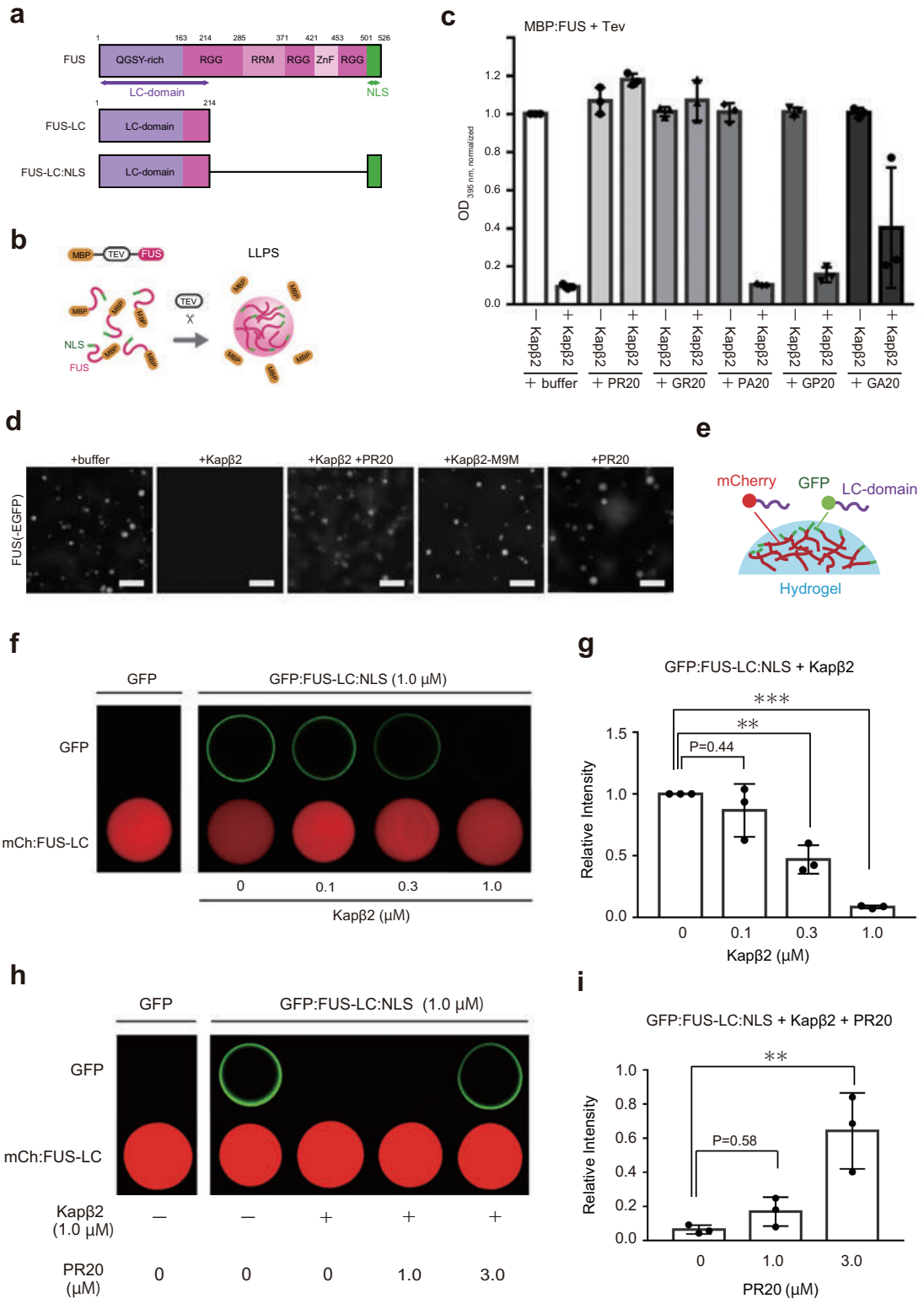
Results

Arginine-rich poly-dipeptides impede NIRs function of modifying RBP phase transitions. First, we examined the effect of five different poly-dipeptides on Kapβ2 in terms of melting full-length FUS droplets (Fig. 1a, b). Turbidity assessment showed that Kapβ2 kept FUS from changing into liquid-like droplets and the addition of equimolar of PR/GR poly-dipeptides inhibited the ability of Kapβ2, whereas that of PA/GP/GA poly-dipeptides did not (Fig. 1c). This suggests that arginine-rich poly-dipeptides disable the Kapβ2 function of melting FUS droplets. To further investigate, we focused on PR poly-dipeptides and evaluated their effect on Kapβ2 activity. We observed with fluorescence microscopy that FUS forms liquid-like droplets and these FUS droplets are dissolved by Kapβ2 (Fig. 1d). By contrast, Kapβ2 loses the ability to suppress the phase transition of FUS in the presence of PR poly-dipeptides (Fig. 1d). As previously reported[7], the designed NLS peptide—M9M, an inhibitor for Kapβ2—abolished the suppression of the phase transition of FUS (Fig. 1d and Supplementary Fig. 1a). These data suggest that PR poly-dipeptides disable the Kapβ2 function of dissolving FUS droplets.

Next, we performed hydrogel binding assay to test the effect of PR poly-dipeptides on the ability of Kapβ2 in preventing polymerization of the LC-domain (Fig. 1e). Hydrogels of mCherry fusion LC-domain of FUS (mCh:FUS-LC) were incubated with green fluorescent protein (GFP) fusion FUS-LC (GFP:FUS-LC) or GFP:FUS-LC fusion FUS-NLS (501–526, GFP:FUS-LC:NLS; Fig. 1a). Hydrogel binding of GFP:FUS-LC was not blocked (Supplementary Fig. 1b), whereas that of GFP:FUS-LC:NLS was blocked by Kapβ2 (Fig. 1f, g), suggesting that Kapβ2 recognizes NLS of FUS and keeps FUS-LC monomeric. In addition, we observed that PR poly-dipeptides inhibited the ability of Kapβ2 from blocking the hydrogel binding of GFP:FUS-LC:NLS (Fig. 1h, i) and hnRNPA2-LC, another RBP interacting with Kapβ2 (Supplementary Fig. 1c, d), suggesting that PR poly-dipeptides impede Kapβ2 function of regulating RBPs phase transitions. We also observed that PR poly-dipeptides inhibited the ability of other NIRs Impα/β1 in modifying the phase transition of LC-domain of TDP43 (TDP43-LC) (Supplementary Fig. 1e, f). Together, these results indicate that PR poly-dipeptides impede NIRs from modifying RBPs phase transitions.

PR poly-dipeptides directly bind Kapβ2. To confirm the interaction between Kapβ2 and PR poly-dipeptides in a cellular environment, we performed immunoprecipitation (Fig. 2a and Supplementary Fig. 2a, b). We observed that endogenous Kapβ2 was co-immunoprecipitated with PR poly-dipeptides expressed in HeLa cells (Fig. 2a). Next, we performed on-column binding assay with purified recombinant proteins, to examine the direct interaction between PR poly-dipeptides and Kapβ2. We found that both MBP:PR18 (18 repeats of Pro-Arg) and Glutathione S-transferase (GST):PR18 bound Kapβ2 (Fig. 2b and Supplementary Fig. 2c), whereas MBP:PR8 (8 repeats of Pro-Arg) did not (Fig. 2b). We also found that GST:PR18 did not bind to control proteins with variety pI values (Supplementary Fig. 2d). We observed interactions between MBP:PR18 with other NIRs Impβ1 or its adapter protein Impα by on-column binding assay (Supplementary Fig. 2e). These results suggest that PR poly-dipeptides target NIR family proteins. In addition, we used isothermal titration calorimetry (ITC) and size-exclusion chromatography with multi-angle light scattering (SEC-MALS) to analyze the interaction quantitatively. ITC revealed that MBP:PR18 bound to Kapβ2 at a K_d value of 81.3 nM (Fig. 2c). MBP:PR18 bound Kapβ2 more strongly than full-length FUS does (Supplementary Table 1). SEC-MALS showed that MBP:PR18 bound Kapβ2 in a 1 : 1 ratio (Fig. 2d). We also confirmed the complex formation of MBP:PR18 and Kapβ2 by analytical ultracentrifugation (AUC) (Supplementary Fig. 2f). These results indicate that repeat extended PR poly-dipeptides directly interact with Kapβ2.

PR poly-dipeptides target the NLS-binding site of Kapβ2. We further investigated the binding sites of PR poly-dipeptides on



Kapβ2 using solution NMR. Despite the large size of Kapβ2 (100 kDa)[11], the use of advanced NMR techniques, including methyl-selective isotope labeling (Fig. 3a) and methyl-transverse relaxation-optimized spectroscopy[26,27], achieved high-quality NMR spectra (Fig. 3b and Supplementary Fig. 3). NMR spectra of the isotopically labeled Kapβ2 were acquired in the absence and presence of PR20 (Fig. 3c). The addition of PR20 induced significant perturbations to several resonances of Kapβ2 (Fig. 3c), indicating that PR poly-dipeptides bind to specific regions of Kapβ2.

In order to obtain information about the binding site of PR poly-dipeptides, we performed reference NMR experiments. The

Fig. 1 Arginine-rich poly-dipeptides impede the Kapβ2 function of modifying FUS phase transitions. a Domain architecture of FUS, LC-domain of FUS (FUS-LC), and FUS-LC fusion NLS (FUS-LC:NLS). **b** Graphical representation of FUS droplet formation for microscope observation and turbidity assay. **c** Turbidity of 8 μM MBP:FUS in the presence of buffer, ±8 μM Kapβ2, and ±8 μM PR20/GR20/PA20/GP20/GA20. OD 395 nm is normalized to measurement of MBP:FUS + buffer + Tev. Mean of three technical replicates, ±SD. **d** Microscopic images of FUS droplets in the absence and presence of Kapβ2 and/or PR20 show that FUS droplets were dissolved by Kapβ2 and not melted in the presence of PR20. Mixture of 7.6 μM MBP:FUS and 0.4 μM MBP:FUS:EGFP were treated with TEV for an hour in the presence or absence of 16 μM Kapβ2, 16 μM Kapβ2–M9M complex, and 50 μM PR20. The experiment was independently repeated three times with similar results. Here, 10 μm scale bars are shown in the image. **e** Graphical description of hydrogel binding assay. Hydrogels of mCh:LC-domain accumulate GFP:LC-domain as it co-polymerizes. **f** Hydrogel binding assay for FUS-LC in the absence and presence of Kapβ2. mCh:FUS-LC (lower images) were incubated with 1.0 μM of GFP (left panel) or 1.0 μM of GFP:FUS-LC:NLS (right panel) in the presence of different concentrations of Kapβ2 (left to right: 0.1, 0.3, and 1.0 μM). **g** Quantitative analysis of Fig. 1f. Relative intensity of GFP signals is shown as the mean of three independent experiments ± SD, analyzed by one-way ANOVA followed by Dunnett's multiple comparison test (**$P < 0.01$, ***$P < 0.001$). **h** Hydrogel binding assay for FUS-LC in the absence and presence of Kapβ2 and/or PR20. Hydrogel droplets of mCh:FUS-LC (lower images) were incubated with 1.0 μM of GFP (left panel) or GFP:FUS-LC:NLS (right panel). GFP:FUS-LC:NLS containing Kapβ2 (1.0 μM) was challenged for homotypic polymer extension in the absence or presence of different concentration of PR20. **i** Quantitative analysis of Fig. 1h. Relative intensity of GFP signals is shown as the mean of three independent experiments ± SD, analyzed by one-way ANOVA followed by Dunnett's multiple comparison test (**$P < 0.01$). Source data are provided as a Source Data file.

perturbations induced by PR20 were compared with those induced by M9M, as the latter has previously been shown to bind to the NLS-binding site of Kapβ2[28]. As represented by peaks #63 and #69, several resonances show significant perturbations induced only by PR20, suggesting that these resonances are derived from regions dedicated for the binding of PR20 (Fig. 4). Likewise, we found several other resonances that were perturbed only by M9M (e.g., Peak #2 and #19), which can be derived from the regions dedicated to the binding of NLS (Fig. 4 and Supplementary Fig. 4a). Interestingly, several of the perturbed resonances (e.g., Peak #40 and #150) induced by the addition of PR20 were also perturbed by the binding of M9M, which suggest that these resonances were derived from regions responsible for the binding of both PR20 and NLS (Fig. 4 and Supplementary Fig. 4a). Although the perturbations could have been caused by allosteric effect, the shared perturbations indicate that PR20 partially occupy the NLS-binding site of Kapβ2. The additional reference experiment using FUS (1-500, FUS-ΔNLS) show that perturbed resonances induced by FUS-ΔNLS (e.g., Peak #37 and #173) were not perturbed by PR20, indicating that the binding site for FUS-ΔNLS does not overlap with that for PR20 (Fig. 4 and Supplementary Fig. 4b). Taken together, the NMR data suggest that the binding site of PR poly-dipeptides on Kapβ2 partially overlaps with the Kapβ2 region responsible for the recognition of NLS.

As seen in the crystal structure[11], NLS is recognized by the cavity of Kapβ2 in which several methyl-baring residues, including isoleucine residues (I457, I540, I642, I722, I773, I804), leucine residues (L419, L539, L767), valine residues (V643, V724), and methionine residue (M308), are located (Supplementary Fig. 5). Given that M9M binds to the NLS-binding site of Kapβ2, these residues can be attributed to methyl resonances that were perturbed by the binding of M9M (Fig. 4 and Supplementary Fig. 4a). Among these residues, L539, I540, I642, and V643 were found to be located at the negatively charged cavity of Kapβ2 (Fig. 4k and Supplementary Fig. 5). Considering the positive charge of PR poly-dipeptides, these four residues located at the negatively charged cavity of Kapβ2 may be important for its interaction with PR poly-dipeptides.

PR poly-dipeptides compete with FUS-NLS for Kapβ2. Strikingly, an MD simulation demonstrated that PR poly-dipeptides were recognized by the negatively charged cavity. The input structure for MD simulation was prepared by taking chain A of the crystallographic dimer from the crystal structure of Kapβ2 in complex with FUS (PDB: 5YVG) and mutating side chains of chain X to repeated PR sequence. L539, I540, I642, V643, I722, L767, and I804 were shown to be located in the NLS-binding site of Kapβ2 (Fig. 5a, b and Supplementary Figs. 5 and 6), an observation that is highly consistent with data from NMR. Biophysical experiments corroborated by MD simulation indicate that PR poly-dipeptides partially bind to the NLS-binding site of Kapβ2.

To test whether PR poly-dipeptides compete with NLS of FUS for the NLS-binding site of Kapβ2, we performed a pull-down binding assay. FUS-NLS (501–526) fused to maltose-binding protein (MBP) (MBP:FUS-NLS) was subjected to GST:Kapβ2 to form a complex at a 1 : 1 ratio. We observed the release of MBP:FUS-NLS from Kapβ2 in the presence of PR poly-dipeptides (Fig. 5c), suggesting that PR poly-dipeptides compete with NLS of FUS for Kapβ2.

A series of interaction analyses using ITC, SEC-MALS, and NMR revealed that PR poly-dipeptides tightly bound Kapβ2 at a 1 : 1 ratio. The binding sites on Kapβ2 partially overlapped with sites used for the recognition of NLS, but they did not overlap with those used for the recognition of FUS-ΔNLS. Based on MD simulation, the amino acids close to the PR poly-dipeptide were in good agreement with those indicated by NMR experiments. These data imply the presence of a potential mechanism—that of PR poly-dipeptides toxicity toward Kapβ2 function in modifying phase transitions, in which PR poly-dipeptides interfere with the interaction between Kapβ2 and NLS of FUS (Fig. 6).

Discussion

We showed that PR/GR poly-dipeptides inhibited the Kapβ2 function as a modifier of RBPs phase transitions, whereas PA/GP/GA poly-dipeptides did not, and that PR poly-dipeptides target the NLS-binding site of Kapβ2, which contains a negatively charged region. Recent studies on cellular systems have revealed that arginine-rich poly-dipeptides interact with NIRs resulting in NCT deficit[24,25] and PR poly-dipeptides have selective effect on NIRs at lower concentrations than GR poly-dipeptides[24]. We showed by a series of experiments that PR poly-dipeptides compete with NLS for binding to Kapβ2, which may provide a molecular basis for the toxicity of positively charged arginine-rich poly-dipeptides toward NIRs.

Through our experiments, we found that PR poly-dipeptides target the NLS-binding site of Kapβ2, which may lead to dysregulation of phase transitions of RBPs containing proline–tyrosine NLS. Besides FUS, many hnRNPs have proline–tyrosine NLS[29] and some of them, including hnRNPA1 and hnRNPA2, are relevant to ALS[13] and interact with TDP43[30,31]. As demonstrated in Supplementary Fig. 1, PR poly-dipeptides impede Kapβ2 function to modify the phase transition of hnRNPA2 containing

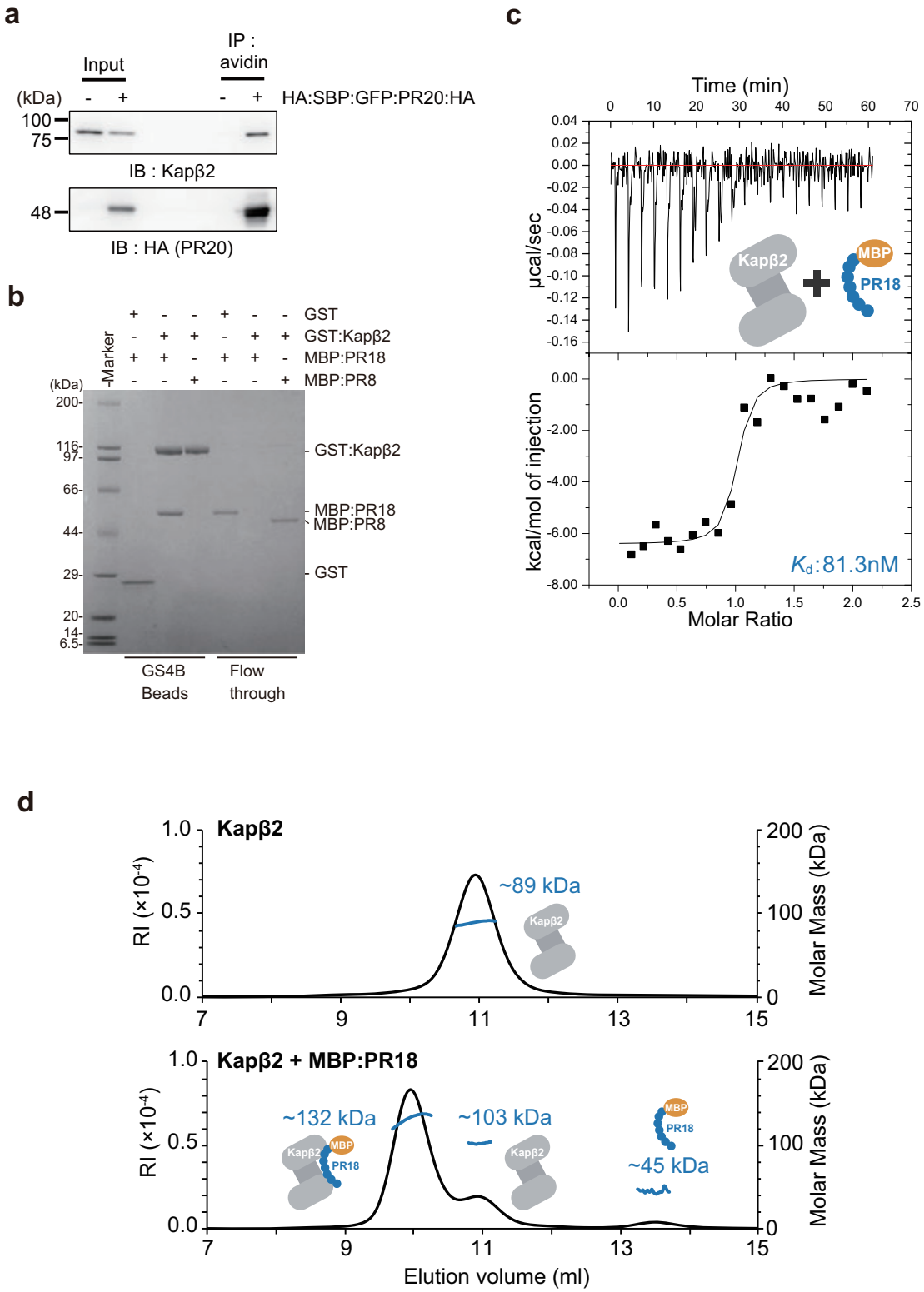


Fig. 2 Interaction between Kapβ2 and PR poly-dipeptides. a Immunoprecipitation showing interaction between endogenous Kapβ2 and HA:SBP:GFP:PR20:HA expressed in HeLa cells. **b** Pull-down binding assay showing interaction between GST:Kapβ2 and MBP:PR18/MBP:PR8. **c** Dissociation constant (K_d) measured by ITC of Kapβ2 (ΔLoop) binding to MBP:PR18. **d** SEC-MALS of Kapβ2 in the absence and presence of MBP:PR18, showing a 1 : 1 complex formation between Kapβ2 and MBP:PR18. The experiments (**a**, **b**) were independently repeated three times or more with similar results. Source data are provided as a Source Data file.

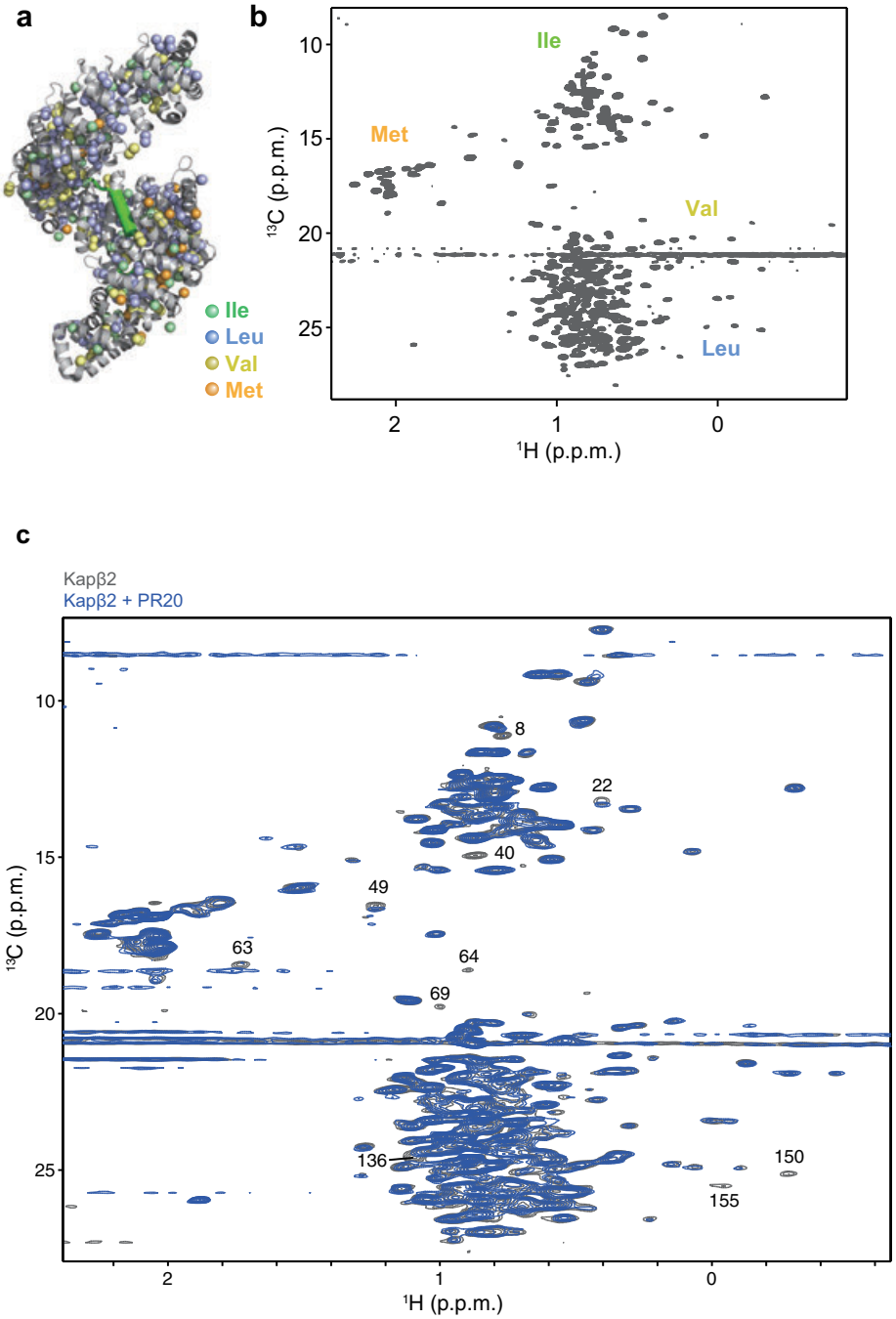
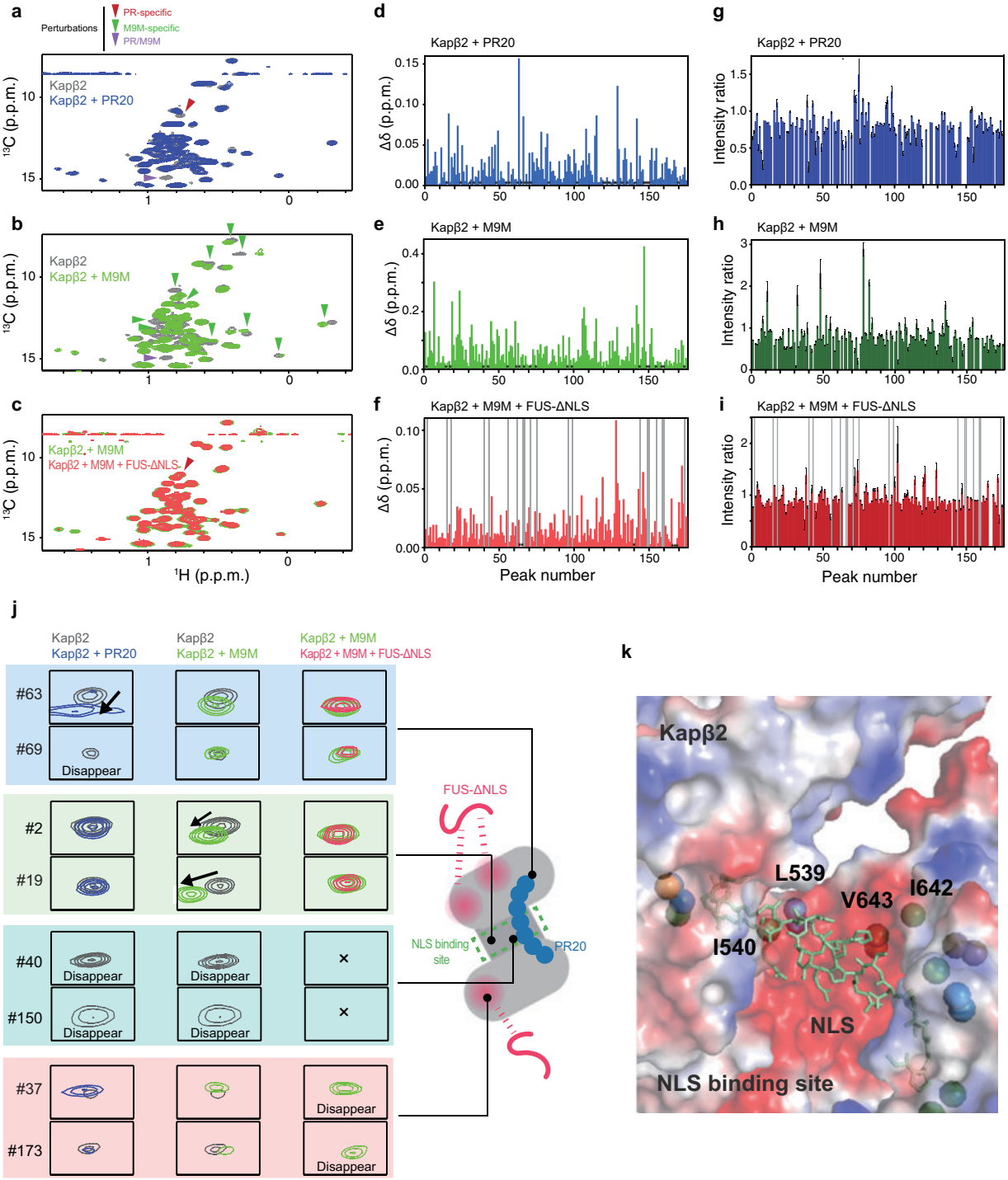


Fig. 3 Solution NMR of Kapβ2 and interaction between Kapβ2 and PR20. a The crystal structure of Kapβ2 in complex with NLS of FUS. NLS of FUS is presented as a green cylinder and Kapβ2 is shown as a gray ribbon, with spheres representing methyl groups of Ala (pink), Ile (green), Leu (blue), Val (yellow), and Met (orange). Kapβ2 is enriched in methyl-baring residues, indicating that sufficient information can be collected from the ^{1}H-^{13}C-correlated methyl NMR spectra. **b** ^{1}H-^{13}C-correlated methyl NMR spectra of [U-^{2}H; Ile-δ1-$^{13}CH_3$; Leu, Val-$^{13}CH_3$/$^{12}CH_3$]-labeled Kapβ2. The spectral regions for the resonances of Ile, Met, Val, and Leu methyl groups are labeled. **c** Interaction between Kapβ2 and PR poly-dipeptide investigated by solution NMR. ^{1}H-^{13}C-correlated methyl NMR spectra of [U-^{2}H; Ile-δ1-$^{13}CH_3$; Leu, Val-$^{13}CH_3$/$^{12}CH_3$]-labeled Kapβ2 in the absence (gray) and presence (blue) of PR20. Significant perturbations were observed for several resonances, indicating that PR poly-dipeptides bind to specific region(s) of Kapβ2. The perturbed resonances are indicated by peak numbers. The peak numbering corresponds to those in Supplementary Fig. 3.

proline–tyrosine NLS, which might correlate with pathophysiology. Moreover, it has been shown that decreased nuclear hnRNPA3, which also has proline–tyrosine NLS, leads to an accumulation of repeat RNA and poly-dipeptides in C9-ALS/FTD[32]. If dysfunction of Kapβ2 due to PR poly-dipeptides leads to mislocalization of hnRNPA3, toxic repeat RNA and PR poly-dipeptides, products of RAN-translation, would increase, resulting in an exacerbation of C9-ALS/FTD pathophysiology.

Given existing evidence[33,34], the binding of PR poly-dipeptides to Kapβ2 might also affect the interaction between Kapβ2 and



proteins with Arg/Gly/Gly (RGG) regions. A previous NMR study reported that RGG regions of the cold-inducible RBP (CIRBP-RGG) is recognized by Kapβ2[33]. The NMRs of Kapβ2 perturbed by CIRBP-RGG partially overlapped with those perturbed by PR poly-dipeptides (Supplementary Fig. 7), indicating that PR poly-dipeptides interact with the region of Kapβ2 responsible for recognizing RGG regions. The data imply that the binding of PR poly-dipeptides to Kapβ2 may also dysregulate the localization of CIRBP and thus affect the stress response system. A recent study reported that RGG regions of FUS are capable of binding the NLS-binding site of Kapβ2[34]. The binding affinities are weaker than the interaction between PR18 and Kapβ2 (Supplementary Table 1). Thus, PR poly-dipeptides likely inhibit the interaction between RGG regions and Kapβ2 in all respects.

We showed that PR poly-dipeptides disrupt the chaperone function of NIRs for FUS/TDP43. In addition to neuronal inclusions of poly-dipeptides, aberrant cytoplasmic localization of FUS and TDP43 are observed in C9-ALS/FTD pathology[12,13]. Cytoplasmic aggregations of FUS are observed less frequently than those of TDP43 in post-mortem examinations[13]. The

Fig. 4 Interaction between Kapβ2 and PR20/M9M/FUS-ΔNLS investigated by solution NMR. ^{1}H-^{13}C-correlated methyl NMR spectra of [U-^{2}H; Ile-δ1-$^{13}CH_3$; Leu,Val-$^{13}CH_3/C^2H_3$]-labeled Kapβ2 in the absence (gray) and presence (blue) of PR20 (**a**), in the absence (gray) and presence (green) of M9M (**b**), or in the presence of M9M (green) and the presence of M9M and FUS-ΔNLS (magenta) (**c**). Significant representative perturbations are indicated by arrow heads. Perturbations only seen for PR20 are indicated by red arrow heads. Perturbations only seen for M9M are indicated by green arrow heads. Perturbations common to PR20 and M9M are indicated by purple arrow heads. For clarity, only the region of the Ile methyl resonances is shown. The full-range spectra are shown in Supplementary Fig. 4 (for **a**) and Supplementary Fig. 5 (for **b** and **c**). Chemical shift differences (**d**–**f**) and intensity ratios (**g**–**i**) of the methyl resonances of Kapβ2 by the interaction with PR20 (**d**, **g**), M9M (**e**, **h**), and FUS-ΔNLS (**f**, **i**) are shown. In **d**–**f**, resonances that disappeared with the addition of a binding partner are indicated by asterisks. In **f** and **i**, resonances that disappeared with the complex formation with M9M are indicated by gray bars. In order to compare the perturbations, peaks on the spectra are labeled in peak numbers as shown in Supplementary Fig. 3. Less significant perturbations with the addition of PR poly-dipeptides than the addition of M9M can be explained by lower binding affinity of PR poly-dipeptides for Kapβ2 (Supplementary Table 1). **j** Selected views of the interaction between Kapβ2 and PR20/M9M/FUS-ΔNLS. Selected views of the representative resonance from ^{1}H-^{13}C-correlated methyl NMR spectra of [U-^{2}H; Ile-δ1-$^{13}CH_3$; Leu, Val-$^{13}CH_3/C^2H_3$]-labeled Kapβ2 in the absence (gray) and presence of PR20 (blue), M9M (green), and M9M and FUS-ΔNLS (magenta). Graphical representation corresponds to the interaction between Kapβ2 and PR20/M9M/FUS-ΔNLS. **k** An expanded view of the electrostatic surface potential of Kapβ2. Positive and negative surface potentials are drawn in blue and red, respectively. NLS of FUS, which is compatible with M9M, is represented as a stick model, colored pale green.

difference in tendency for aggregation between FUS and TDP43 might be derived from several factors, such as the core structures of LC-domain[35] and posttranslational modifications[9,10,35,36]. A solid-state NMR study[35] revealed that there is no hydrophobic interaction within the labile fibril core of FUS-LC and no polymorphism unlike TDP43-LC[37] or other pathogenic amyloid fibrils such as β-amyloid[38]. A recent study reported that in sporadic ALS, nuclear-cytoplasmic mislocalization of FUS was present, although cytoplasmic FUS aggregates were absent, and localization of FUS in the cytoplasm was observed prior to the appearance of TDP43 aggregates in the mouse model[39]. Hence, FUS might have a more significant role in ALS pathophysiology than previously considered. It has been reported that aggregates of TDP43 protect cells and toxic liquid phase of TDP43 harm cells[40]; therefore, further investigations are required to conclude which phase of TDP43 affects ALS pathophysiology.

Several posttranslational modifications of FUS are known to alter FUS phase transitions, including methylation, phosphorylation, and acetylation[9,10,35,36,41]. Among them, the methylation of arginine and the phosphorylation of serine are known to reduce FUS phase transitions[9,10,35,36]. RGG-rich regions of FUS are extensively methylated by protein arginine methyl transferases (PRMTs)[42]. Arginine methylation of FUS harboring disease-linked mutations decreases binding affinity to Kapβ2 and impairs nuclear transporting activity of Kapβ2[43]. Controversial results have been also reported; inclusions in ALS-FUS patients contain methylated FUS, but no methylated FUS is observed in inclusions of FTD-FUS patients, suggesting the possibility that there remain unknown pathways generating disease-related inclusions[9,43]. We revealed that PR and GR poly-dipeptides impede the chaperone activity of Kapβ2. Gly/Arg-rich sequence tends to be methylated by PRMTs[44], leading to an alternation of charge distribution. A recent study revealed that the symmetric dimethylation of GR poly-dipeptides reduces toxicity and correlates with disease duration in C9-ALS/FTD[45]. Therefore, PR poly-dipeptides might have a stronger inhibitory effect toward Kapβ2 than GR poly-dipeptides due to the evasion of methylation, although it has yet to be shown whether PR poly-dipeptides are methylated in C9-ALS/FTD.

We also demonstrated that PR poly-dipeptides tightly bind Kapβ2 in a cellular system, in addition to experiments using purified proteins. Although NIRs are abundant in a cellular environment, increased NIRs suppress the pathological protein interaction by arginine-rich poly-dipeptides[25]. A previous study revealed that overexpression of KAP104, the yeast homolog of Kapβ2, strongly suppressed the toxicity of PR poly-dipeptides[22]. These results suggest that the relation between the quantities of NIRs and those of PR poly-dipeptides might be crucial in a cellular environment. A pathological study reported that in C9-ALS/FTD, poly-dipeptides accumulate prior to TDP43 aggregates[46]. Hence, poly-dipeptide pathology might reflect the upstream of the cascade of C9-ALS/FTD pathophysiology. Given that increased NIRs reduce the toxicity of arginine-rich poly-dipeptides and modify aberrant RBP phase transitions, our study highlights the molecular mechanism for potential C9-ALS/FTD therapeutic targets.

In summary, we showed how PR poly-dipeptides affect Kapβ2 function as a phase modifier using multiple biochemical and biophysical methods. In addition to the ability of PR poly-dipeptides to stabilize self-association of LC-domains, PR poly-dipeptides are found to bind NIRs and impede their function as phase modifiers. Our current study offers additional mechanistic insights on *C9orf72*-related neurodegeneration.

Methods

Constructs, protein expression, and purification. Kapβ2 and M9M were expressed from GST fusion constructs using pGEX6P-1 and pGEX-TEV vectors, respectively. FUS and PRn proteins were expressed from MBP-fusion constructs using the pMAL-TEV vector. All recombinant proteins were expressed individually in BL21(DE3). *Escherichia coli* cells were induced with 0.5 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG) for 12 h at 20 °C. Bacteria expressing Kapβ2 was lysed with a sonicator in buffer containing 50 mM Tris pH 7.5, 200 mM NaCl, 20% (v/v) glycerol, and 2 mM dithiothreitol (DTT). Kapβ2 was purified using GSH Sepharose beads (GS4B, GE Healthcare), cleaved with HRV3C protease, anion exchange chromatography (HiTrap Q HP, GE Healthcare), and gel filtration chromatography (Superdex200 16/60, GE Healthcare) in a buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 2 mM DTT, 2 mM $Mg(OAc)_2$, and 10% glycerol. GST:M9M was purified using GSH Sepharose beads (GS4B, GE Healthcare) and gel filtration chromatography (Superdex200 16/60, GE Healthcare). To assemble the Kapβ2–M9M complex, purified Kapβ2 and GST:M9M were mixed, and GST tag was cleaved with Tobacco Etch Virus (TEV) protease. Kapβ2–M9M complex was purified by gel filtration chromatography (Superdex200 16/60, GE Healthcare) and the remaining GST was removed by GSH Sepharose beads (GS4B, GE Healthcare). MBP:FUS and MBP:PRn proteins were lysed in 50 mM Tris pH 7.5, 1.5 M NaCl, 10% glycerol, and 2 mM DTT. MBP-fusion proteins were purified by affinity chromatography, using amylose resin eluted with buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, 10% glycerol, 2 mM DTT, and 20 mM Maltose. It was then further purified for MBP:PRn by cation-exchange chromatography (HiTrap SP HP, GE Healthcare) and gel filtration chromatography (Superdex200 16/60, GE Healthcare). Twenty repeats of PR/GR/PA/GP/GA with an HA epitope tag at the C terminus (PR/GR/PA/GP/GA20:HA) peptides were synthesized by SCRUM, Inc. (Japan). Primers for plasmid construction are listed in Supplementary Table 2.

Turbidity assay and imaging of turbid solution. Prior to adding TEV protease, we mixed 8 μM MBP:FUS, ±8 μM Kapβ2, and ±PR/GR/PA/GP/GA20:HA in a buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 2 mM $Mg(OAc)_2$, 20 μM $Zn(OAc)_2$, and 2 mM DTT to reaction volumes of 100 μL. TEV protease was added to the premixture to a final concentration of 40 μg/mL, then incubated at 30 °C for 60 min for all the MBP-fusion protein to be digested. The solution was left to cool down to 20 °C before the measurement of OD 395 nm using a plate reader. For the imaging experiment and prior to adding TEV protease, 8 μM MBP:FUS including 0.4 μM MBP:FUS:EGFP, ±16 μM Kapβ2 or 16 μM

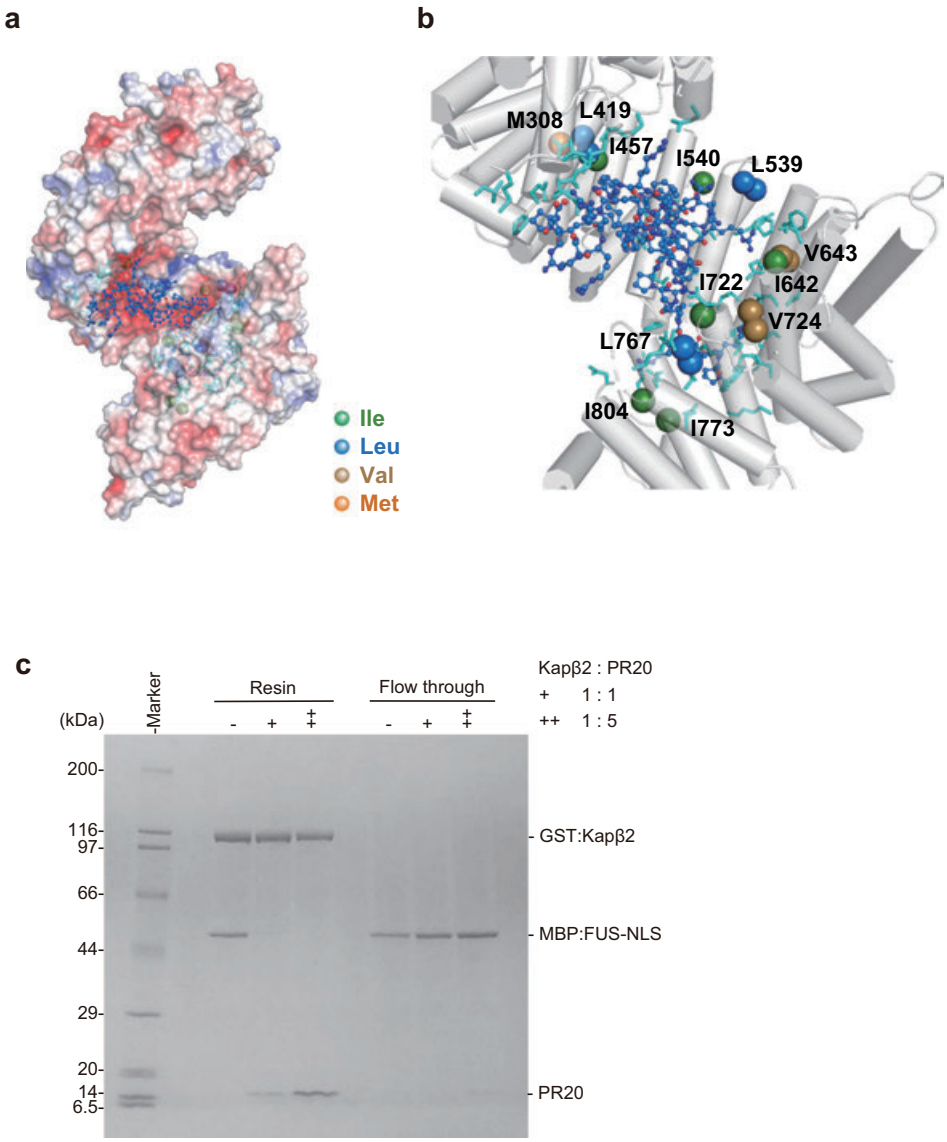


Fig. 5 The model of interaction between PR poly-dipeptides and Kapβ2. a Electrostatic potential of Kapβ2 and MD calculated model of PR poly-dipeptide inside of Kapβ2 cavity are shown. Blue and red colors show the positively charged and negatively charged regions, respectively. PR poly-dipeptide is colored blue. **b** Close-up view of Kapβ2 cavity. Methyl groups located close to the Kapβ2 cavity, isoleucine, leucine, valine, and methionine, are represented as spheres colored green, blue, yellow, and orange, respectively. Amino acids within 5 Å around the PR poly-dipeptide are colored cyan. **c** Pull-down binding assay showing the competition of PR20 and NLS of FUS to GST:Kapβ2. The experiment was independently repeated twice with similar results. Source data are provided as a Source Data file.

Kapβ2–M9M complex, and ±50 μM PR20:HA were mixed in a buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 2 mM $Mg(OAc)_2$, 20 μM $Zn(OAc)_2$, and 2 mM DTT to reaction volumes of 20 μL. TEV protease was added to the premixture to a final concentration of 40 μg/mL, then incubated at room temperature for 3 h. MBP:FUS:EGFP in the solutions was imaged using a fluorescent microscope (Nikon Eclipse Ti2) equipped with a ×40, 0.6 na objective lens (Nikon) and a CMOS camera (ORCA-spark, Hamamatsu). The system was operated using the NIS-Elements software (Nikon).

Hydrogel binding assay. Expression plasmids for recombinant proteins (GFP/mCherry fusion LC-domain of FUS (residue 2–214, GFP/mCh:FUS-LC) and GFP/mCherry fusion LC-domain of hnRNPA2 (residue 181–341, GFP/mCh:hnRNPA2-LC)) were obtained from the Steven L. McKnight Laboratory. Expression plasmid for GFP:FUS-LC fusion NLS (residue 501–526, GFP:FUS-LC:NLS), mCh fusion LC-domain of TDP43 (residue 262–414, mCh:TDP43-LC), and GFP-tagged cNLS (residue 78–99) fusion TDP43-LC (GFP:cNLS:TDP43-LC) were constructed using In-Fusion HD Cloning Kit (Takara Bio, Inc.). GFP/mCh:FUS-LC, GFP:FUS-LC:NLS, and GFP/mCh:hnRNPA2-LC were expressed in *E. coli* BL21(DE3) cells with 0.5 mM IPTG at 20 °C overnight and purified by Ni-NTA Agarose (FUJIFILM Wako Pure Chemical Corporation), as described in a previous study[47]. GFP:cNLS:TDP43-LC were expressed with 1.0 mM IPTG at 37 °C for 3 h and mCh:TDP43-LC were expressed with 0.5 mM IPTG at 16 °C overnight. GFP:cNLS:TDP43-LC and mCh:TDP43-LC were purified by Ni-NTA as described in a previous study[48] with some modifications. For the purification of mCh:TDP43-LC/GFP:cNLS:TDP43-LC, cells were lysed in buffer containing 25 mM Tris-HCl pH 7.5, 200 mM NaCl, 2 M/4 M urea, 10 mM β-mercaptoethanol, and Protease Inhibitor Cocktail Tablets (Sigma), washed with a buffer (25 mM Tris-HCl pH 7.5, 200 mM NaCl, 2 M Urea, 10 mM β-mercaptoethanol, and 20 mM imidazole), and eluted from the resin with an elution buffer (25 mM Tris-HCl pH 7.5, 200 mM NaCl, 2 M Urea, 10 mM β-mercaptoethanol, and 300 mM imidazole). Hydrogel droplets of mCh:FUS-LC, mCh:hnRNPA2-LC, and mCh:TDP43-LC were prepared as reported in a previous study[47]. For hydrogel binding assays, purified GFP-fused proteins were diluted to 1 μM in the buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 20 mM β-mercaptoethanol, 0.1 mM phenylmethylsulfonyl fluoride, and 0.5 mM EDTA) and pipetted onto a hydrogel dish. After overnight

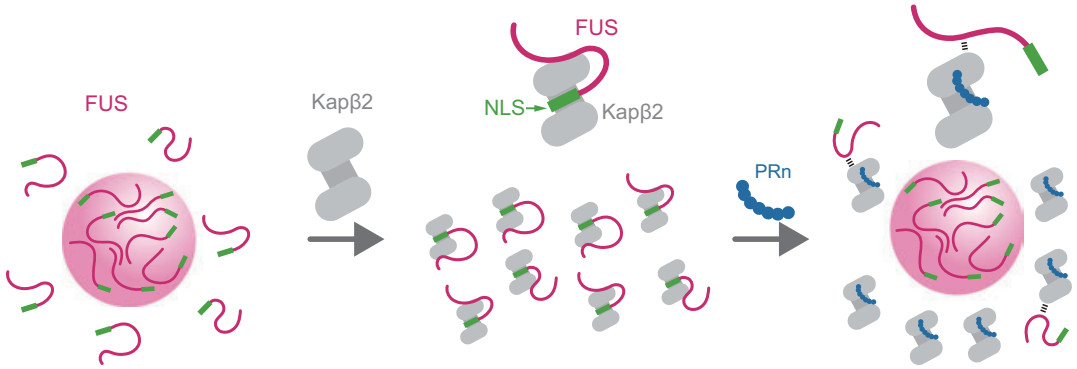


Fig. 6 Graphical abstract for the model of interaction between PR poly-dipeptides and Kapβ2. FUS is prone to self-associate (left). Kapβ2 modifies a phase transition of FUS by recognizing NLS (middle). PR poly-dipeptides partially bind to the NLS-binding site of Kapβ2 and impede its ability (right).

incubation, horizontal sections of the droplets were scanned with excitation wavelengths on a confocal microscope (FLUOVIEW FV3000, OLYMPUS). Relative intensity of GFP signals across the hydrogel droplets were measured in triplicate by using the profile plot mode in ImageJ. The values are shown as mean ± SD. Statistical analyses were performed using Graphpad Prism version 7.

Immunoprecipitation. The pcDNA5/FRT/TO-HA:SBP plasmid was generated by inserting the PCR-amplified HA:SBP sequence into the pcDNA5/FRT/TO vector (Thermo Fisher Scientific) between the HindIII and BamHI sites. The HA:SBP:GFP:PR20:HA plasmid was constructed by ligating the GFP:PR20:HA sequence into the pcDNA5/FRT/TO-HA-SBP between the EcoRV and XhoI sites. The HA:SBP:Kapβ2 plasmid was constructed by ligating the Kapβ2 sequence into the pcDNA5/FRT/TO-HA:SBP between the BamHI and NotI sites. HeLa cells were cultured in Dulbecco's modified Eagle medium/10% fetal bovine serum at 37 °C with 5% CO_2. Cells were transfected with plasmids using Lipofectamine 3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. HeLa cytosol extracts were prepared by Cytoplasmic Extraction buffer (CE; 10 mM HEPES pH 7.9, 60 mM KCl, 1 mM EDTA, 0.075% NP40, 1 mM DTT, and cOmplete™ EDTA-free protease inhibitor [Sigma-Aldrich]) for 10 min on ice and then separated from nuclear by centrifugation at 500 × *g* for 10 min. The resultant extracts were cleared by centrifugation at 20,400 × *g* for 10 min. The supernatant containing the HeLa cytosol extracts were mixed with the Streptavidin Mag Sepharose (GE Healthcare) and rotated at 4 °C overnight. The beads were finally washed five times with CE buffer. The anti-Kapβ2 (sc-166127, 1 : 1000) antibody was obtained from Santa Cruz Biotechnology. The anti-HA (561, 1 : 10,000) antibodies were obtained from MBL. The anti-Mouse IgG H&L (Horseradish peroxidase, HRP) (ab6789, 1 : 10,000) and anti-Rabbit IgG H&L (HRP) (ab6721, 1 : 10,000) antibodies were purchased from Abcam. HeLa cells were purchased from ATCC.

Pull-down binding assay. In-vitro pull-down binding assays showing the interaction between PR poly-dipeptides and Kapβ2 were performed using GST:Kapβ2, GST:PR18, GST:Importinα, or GST:Importinβ1 immobilized on GSH sepharose beads (GE Healthcare). Four micrograms of GST proteins were immobilized on 30 µL of beads. GST protein beads were incubated with equal molar protein, either MBP:PR18, MBP:PR8, Kapβ2, BSA, MBP, EGFP, or Lysozyme for 20–30 min and washed three times with a buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 2 mM $Mg(OAc)_2$, and 2 mM DTT. For the competitive binding assay, 16 µg GST:Kapβ2 were immobilized on 160 µL beads and washed three times with the buffer before the beads were aliquoted into four tubes and incubated with equal molar MBP:FUS-NLS (residue 501–526) ± PR20:HA for 20–30 min. Unbound proteins were washed three times with the buffer. Bound proteins and unbound proteins were separated by SDS-polyacrylamide gel electrophoresis and stained with Coomassie Brilliant Blue.

Isothermal titration calorimetry. ITC experiments were performed with a Malvern iTC200 calorimeter (Malvern Instruments). Proteins were dialyzed overnight against buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, and 2 mM β-mercaptoethanol. MBP:PR18 (200 µM) were titrated into the sample cell containing 20 µM Kapβ2 (ΔLoop, residues 321–371 were replaced with GGSGGSGS linker). ITC experiments were performed at 25 °C with 19 rounds of 2 µL injections. Data were analyzed using Origin software.

Size-exclusion chromatography with multi-angle light scattering. SEC-MALS was measured using DAWN HELEOS8+ (Wyatt Technology Corporation) downstream of a Shimadzu liquid chromatography system connected to Superdex200 10/300 GL (GE Healthcare) gel filtration column. The differential refractive index (Shimadzu Corporation) downstream of MALS was used to obtain protein concentration. The column was equilibrated with a running buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 2 mM $MgCl_2$, 10% glycerol, and 2 mM β-mercaptoethanol. Flow rate was set to 0.5 mL min^{-1} and 100 µL of the sample was injected. Kapβ2 at a concentration of 28 µM was injected in the absence and presence of 55 µM of MBP:PR18. The data were analyzed with ASTRA version 7.0.1 (Wyatt Technology Corporation).

Analytical ultracentrifugation (AUC). AUC measurements were conducted with ProteomeLab XL-I (Beckman Coulter). Three solutions, 28 µM Kapβ2, 28 µM MBP:PR18, and the mixture of 28 µM Kapβ2 + 28 µM MBP:PR18, were prepared in the buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 2 mM $MgCl_2$, 2 mM DTT, and 10% glycerol. The solutions were filled in 1.5 mm path-length titanium cells (Nanolytics). The measurements were performed with Rayleigh-interference optics at rotor speed of 125,000 × *g* (at 70 mm radius). The temperature was set at 25 °C. The weight concentration distribution as a function of sedimentation coefficient $c(s_{20,w})$ and frictional ratio f/f_0 were computed with SEDFIT[49] and Igor Pro software. The sedimentation coefficient was normalized to be the value at 20 °C in pure water $s_{20,w}$. The molecular weight M was calculated with the corresponding peak value of $s_{20,w}$ and f/f_0[49].

Expression and purification of isotopically labeled Kapβ2. The Kapβ2 expression plasmid was transformed into BL21(DE3) cells. The protein samples with $^{1}H,^{13}C$-labeled methyl groups in deuterium background were prepared as described in a previous study[26]. The cells were grown in medium with $^{15}NH_4Cl$ (2 gL^{-1}, CIL) and $^{2}H_7$-glucose (2 g L^{-1}, CIL) in 99.9% $^{2}H_2O$ (Isotec). The precursors for methyl groups of Ile, Val, and Leu, α-ketobutyric acid (50 mg L^{-1}) and α-ketoisovaleric acid (80 mg L^{-1}), and [$^{13}CH_3$] methionine (50 mg L^{-1}) were added to the culture 1 h before the addition of IPTG. Protein expression was induced by the addition of 0.5 mM IPTG at OD_{600} ~ 0.6, followed by ~16 h of incubation at 25 °C. Cells were collected and re-suspended in the lysis buffer containing 50 mM HEPES pH 7.4, 150 mM NaCl, 20% glycerol, 2 mM DTT, and 2 mM EDTA. Cells were disrupted by a sonicator and centrifuged at 38,900 × *g* for 45 min. Proteins were purified using GS4B resin (GE Healthcare) and eluted with the buffer containing 20 mM HEPES pH 7.4, 20 mM NaCl, 2 mM EDTA, 10% glycerol, and 30 mM GSH. The GST tag was removed by HRV3C protease at 4 °C for ~16 h. The protein was further purified by anion exchange using HiTrap Q FF (GE Healthcare) with the buffer containing 20 mM Imidazole pH 6.5, 20–1000 mM NaCl, 2 mM EDTA, 2 mM DTT, and 20% glycerol, followed by gel filtration using Superdex200 16/60 (GE Healthcare) with the buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 2 mM $MgCl_2$, 2 mM DTT, and 2% glycerol. Protein concentration was determined spectrophotometrically at 280 nm using a corresponding extinction coefficient.

NMR experiments. The isotopically labeled Kapβ2 was prepared in the NMR buffer containing 20 mM ^{2}H-Tris pH 7.4, 150 mM NaCl, 2 mM $MgCl_2$, 2 mM DTT, and 2% glycerol, concentrated to 20 µM. NMR experiments were performed on Bruker 600 MHz and 800 MHz NMR operated with Topspin software at 20 °C. Perturbation of the side chain methyl resonances were each monitored using ^{1}H-^{13}C heteronuclear multiple quantum coherence. Spectra were processed using the NMRPipe software[50]. Data analyses were performed on Olivia software (https://github.com/yokochi47/Olivia). The perturbations of the resonances were evaluated based on chemical shift change or intensity change. Chemical shift perturbations for the methyl groups were calculated based on the following equation:

$$\Delta\delta = \sqrt{\left(\frac{\Delta\delta_H}{\alpha}\right)^2 + \left(\frac{\Delta\delta_C}{\beta}\right)^2} \quad (1)$$

where $\Delta\delta_H$ and $\Delta\delta_C$ are the respective chemical shift changes of ^1H and ^{13}C by the addition of the ligand, and α and β are the respective chemical shift distributions of ^1H and ^{13}C of methyl groups as reported in the Biological Resonance Data Bank (http://www.bmrb.wisc.edu)[51]. Errors for the intensity ratio were estimated based on the peak intensity and noise level.

In order to investigate the interaction between Kapβ2 and PR poly-dipeptide by NMR, the NMR spectra of the isotopically labeled Kapβ2 in the absence and presence of the synthetic peptide of PR20:HA were measured. To investigate the interaction between Kapβ2 and M9M, the Kapβ2–M9M complex was prepared in the following manner. Purified GST:M9M protein was added to the isotopically labeled Kapβ2, followed by the removal of the GST tag by TEV protease at room temperature for 2 h and then at 4 °C for ~16 h before it was buffer-exchanged into the NMR buffer using Amicon Ultra-4 50k (Merck). To investigate the interaction between Kapβ2 and FUS-ΔNLS by NMR, the NMR spectra of isotopically labeled Kapβ2 in complex with M9M were measured in the absence and presence of FUS-ΔNLS.

MD simulation. MD simulations were performed by Maestro version 11.0.014 (Schrödinger) with Desmond program (D. E. Shaw Research)[52]. Protonation states of titratable residues at pH 7.0 were determined by the PROPKA program[53,54]. The protein and PR poly-dipeptide were solvated with a rectangular box of TIP3P[55] water molecules and neutralized by adding three sodium ions. The simulation system was equilibrated with the default protocol of the Desmond program. The protein backbone (except for the PR poly-dipeptide) was restrained by a force constant of 100 kcal $(mol\ Å)^{-1}$ to keep the original Kapβ2 structure during simulations. After a 5 ns relaxation simulation, a 100 ns MD simulation was performed under the conditions where equilibration of simulation was performed using the isothermal–isobaric ensemble (NPT). The force field used was OPLS-2005[56], the van der Waals interaction between atoms separated by over 10 Å were cut off, and long-range interactions were computed with the particle mesh Ewald method. The thermostat and barostat used were Nosé–Hoover chain[57] at 300 K and Martyna–Tobias–Klein[58] at 1 atm with relaxation of 1 and 2 ps. The RESPA integrator[59] was used with Fourier-space electrostatics computed every 6 fs and all remaining interactions computed every 2 fs. Surface electrostatic potential was calculated using Adaptive Poisson–Boltzmann Solver (version 2.1). The structure was drawn with PyMol program (version 2.3.0.). The electrostatic potential of Kapβ2 was calculated by PyMOL APBS tool at 0.15 M of monovalent salt (±1) and represented within a ± 5 kT/e range.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data supporting the findings of this study are available from the corresponding authors upon reasonable request. PDB ID referred to in this study is as follows: 5YVG and 4FDD. Source data are provided with this paper.

Received: 28 August 2020; Accepted: 16 August 2021;
Published online: 06 September 2021

References

1. Kato, M. et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012).
2. Lin, Y. et al. Toxic PR poly-dipeptides encoded by the C9orf72 repeat expansion target LC domain polymers. *Cell* **167**, 789–802.e712 (2016).
3. Shi, K. Y. et al. Toxic PRn poly-dipeptides encoded by the C9orf72 repeat expansion block nuclear import and export. *Proc. Natl Acad. Sci. USA* **114**, E1111–e1117 (2017).
4. Lee, K. H. et al. C9orf72 dipeptide repeats impair the assembly, dynamics, and function of membrane-less organelles. *Cell* **167**, 774–788.e717 (2016).
5. Kato, M. & McKnight, S. L. A solid-state conceptualization of information transfer from gene to message to protein. *Annu. Rev. Biochem.* **87**, 351–390 (2018).
6. Yoshizawa, T., Nozawa, R.-S., Jia, T. Z., Saio, T. & Mori, E. Biological phase separation: cell biology meets biophysics. *Biophys. Rev.* **12**, 519–539 (2020).
7. Yoshizawa, T. et al. Nuclear import receptor inhibits phase separation of FUS through binding to multiple sites. *Cell* **173**, 693–705.e622 (2018).
8. Guo, L. et al. Nuclear-import receptors reverse aberrant phase transitions of RNA-binding proteins with prion-like domains. *Cell* **173**, 677–692.e620 (2018).
9. Hofweber, M. et al. Phase separation of FUS is suppressed by its nuclear import receptor and arginine methylation. *Cell* **173**, 706–719.e713 (2018).
10. Qamar, S. et al. FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation-pi interactions. *Cell* **173**, 720–734.e715 (2018).
11. Zhang, Z. C. & Chook, Y. M. Structural and energetic basis of ALS-causing mutations in the atypical proline–tyrosine nuclear localization signal of the Fused in Sarcoma protein (FUS). *Proc. Natl Acad. Sci. USA* **109**, 12017–12021 (2012).
12. Keller, B. A. et al. Co-aggregation of RNA binding proteins in ALS spinal motor neurons: evidence of a common pathogenic mechanism. *Acta Neuropathol.* **124**, 733–747 (2012).
13. Boeynaems, S., Bogaert, E., Van Damme, P. & Van Den Bosch, L. Inside out: the role of nucleocytoplasmic transport in ALS and FTLD. *Acta Neuropathol.* **132**, 159–173 (2016).
14. Odeh, H. M. & Shorter, J. Arginine-rich dipeptide-repeat proteins as phase disruptors in C9-ALS/FTD. *Emerg. Top. Life Sci.* **4**, 293–305 (2020).
15. Balendra, R. & Isaacs, A. M. *C9orf72*-mediated ALS and FTD: multiple pathways to disease. *Nat. Rev. Neurol.* **14**, 544–558 (2018).
16. Mori, K. et al. Bidirectional transcripts of the expanded C9orf72 hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. *Acta Neuropathol.* **126**, 881–893 (2013).
17. Zu, T. et al. RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. *Proc. Natl Acad. Sci. USA* **110**, E4968–E4977 (2013).
18. Kwon, I. et al. Poly-dipeptides encoded by the C9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. *Science* **345**, 1139–1145 (2014).
19. Mizielinska, S. et al. C9orf72 repeat expansions cause neurodegeneration in *Drosophila* through arginine-rich proteins. *Science* **345**, 1192–1194 (2014).
20. Zhang, K. et al. The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. *Nature* **525**, 56–61 (2015).
21. Freibaum, B. D. et al. GGGGCC repeat expansion in C9orf72 compromises nucleocytoplasmic transport. *Nature* **525**, 129–133 (2015).
22. Jovicic, A. et al. Modifiers of C9orf72 dipeptide repeat toxicity connect nucleocytoplasmic transport defects to FTD/ALS. *Nat. Neurosci.* **18**, 1226–1229 (2015).
23. Vanneste, J. et al. C9orf72-generated poly-GR and poly-PR do not directly interfere with nucleocytoplasmic transport. *Sci. Rep.* **9**, 15728 (2019).
24. Hayes, L. R., Duan, L., Bowen, K., Kalab, P. & Rothstein, J. D. C9orf72 arginine-rich dipeptide repeat proteins disrupt karyopherin-mediated nuclear import. *eLife* **9**, e51685 (2020).
25. Hutten, S. et al. Nuclear import receptors directly bind to arginine-rich dipeptide repeat proteins and suppress their pathological interactions. *Cell Rep.* **33**, 108538 (2020).
26. Saio, T., Guan, X., Rossi, P., Economou, A. & Kalodimos, C. G. Structural basis for protein antiaggregation activity of the trigger factor chaperone. *Science* **344**, 1250494–1250494 (2014).
27. Saio, T., Kawagoe, S., Ishimori, K. & Kalodimos, C. G. Oligomerization of a molecular chaperone modulates its activity. *eLife* **7**, e35731 (2018).
28. Cansizoglu, A. E., Lee, B. J., Zhang, Z. C., Fontoura, B. M. A. & Chook, Y. M. Structure-based design of a pathway-specific nuclear import inhibitor. *Nat. Struct. Mol. Biol.* **14**, 452–454 (2007).
29. Kimura, M. et al. Extensive cargo identification reveals distinct biological roles of the 12 importin pathways. *eLife* **6**, e21184 (2017).
30. Ryan, V. H. et al. Mechanistic view of hnRNPA2 low-complexity domain structure, interactions, and phase separation altered by mutation and arginine methylation. *Mol. Cell* **69**, 465–479.e467 (2018).
31. Buratti, E. et al. TDP-43 binds heterogeneous nuclear ribonucleoprotein A/B through its C-terminal tail: an important region for the inhibition of cystic fibrosis transmembrane conductance regulator exon 9 splicing. *J. Biol. Chem.* **280**, 37572–37584 (2005).
32. Mori, K. et al. Reduced hnRNPA3 increases C9orf72 repeat RNA levels and dipeptide-repeat protein deposition. *EMBO Rep.* **17**, 1314–1325 (2016).
33. Bourgeois, B. et al. Nonclassical nuclear localization signals mediate nuclear import of CIRBP. *Proc. Natl Acad. Sci. USA* **117**, 8503–8514 (2020).
34. Gonzalez, A. et al. Mechanism of karyopherin-β2 binding and nuclear import of ALS variants FUS(P525L) and FUS(R495X). *Sci. Rep.* **11**, 3754 (2021).
35. Murray, D. T. et al. Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains. *Cell* **171**, 615–627.e616 (2017).
36. Monahan, Z. et al. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J.* **36**, 2951–2967 (2017).
37. Cao, Q., Boyer, D. R., Sawaya, M. R., Ge, P. & Eisenberg, D. S. Cryo-EM structures of four polymorphic TDP-43 amyloid cores. *Nat. Struct. Mol. Biol.* **26**, 619–627 (2019).
38. Gremer, L. et al. Fibril structure of amyloid-β(1-42) by cryo-electron microscopy. *Science* **358**, 116–119 (2017).
39. Tyzack, G. E. et al. Widespread FUS mislocalization is a molecular hallmark of amyotrophic lateral sclerosis. *Brain* **142**, 2572–2580 (2019).
40. Bolognesi, B. et al. The mutational landscape of a prion-like domain. *Nat. Commun.* **10**, 4162 (2019).

41. Bock, A. S. et al. N-terminal acetylation modestly enhances phase separation and reduces aggregation of the low-complexity domain of RNA-binding protein fused in sarcoma. *Protein Sci.* **30**, 1337–1349 (2021).
42. Ong, S. E., Mittler, G. & Mann, M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nat. Methods* **1**, 119–126 (2004).
43. Dormann, D. et al. Arginine methylation next to the PY-NLS modulates Transportin binding and nuclear import of FUS. *EMBO J.* **31**, 4258–4275 (2012).
44. Najbauer, J., Johnson, B. A., Young, A. L. & Aswad, D. W. Peptides with sequences similar to glycine, arginine-rich motifs in proteins interacting with RNA are efficiently recognized by methyltransferase(s) modifying arginine in numerous proteins. *J. Biol. Chem.* **268**, 10501–10509 (1993).
45. Gittings, L. M. et al. Symmetric dimethylation of poly-GR correlates with disease duration in C9orf72 FTLD and ALS and reduces poly-GR phase separation and toxicity. *Acta Neuropathol.* **139**, 407–410 (2020).
46. Baborie, A. et al. Accumulation of dipeptide repeat proteins predates that of TDP-43 in frontotemporal lobar degeneration associated with hexanucleotide repeat expansions in C9ORF72 gene. *Neuropathol. Appl. Neurobiol.* **41**, 601–612 (2015).
47. Kato, M., Lin, Y. & McKnight, S. L. Cross-beta polymerization and hydrogel formation by low-complexity sequence proteins. *Methods* **126**, 3–11 (2017).
48. Lin, Y. et al. Redox-mediated regulation of an evolutionarily conserved cross-β structure formed by the TDP43 low complexity domain. *Proc. Natl Acad. Sci. USA* **117**, 28727–28734 (2020).
49. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78**, 1606–1619 (2000).
50. Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
51. Rosenzweig, R. et al. ClpB N-terminal domain plays a regulatory role in protein disaggregation. *Proc. Natl Acad. Sci. USA* **112**, E6872–E6881 (2015).
52. Bowers, et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC '06: Proc. 2006 ACM/IEEE Conference on Supercomputing* 9451996 (IEEE, Tampa, 2006).
53. Li, H., Robertson, A. D. & Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **61**, 704–721 (2005).
54. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **73**, 765–783 (2008).
55. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
56. Banks, J. L. et al. Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* **26**, 1752–1780 (2005).
57. Martyna, G. J., Klein, M. L. & Tuckerman, M. Nosé–Hoover chains: the canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635–2643 (1992).
58. Martyna, G. J., Tuckerman, M. E., Tobias, D. J. & Klein, M. L. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.* **87**, 1117–1157 (1996).
59. Tuckerman, M., Berne, B. J. & Martyna, G. J. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* **97**, 1990–2001 (1992).

Acknowledgements

This work was supported by grants from AMED [JP20dm0307032 and JP21wm0425004 to E.M.; JP21ek0109437 to T. Saio; and JP21ek0109558 to T.Y.], JSPS KAKENHI [JP17H07031 and JP20H03199 to E.M.; JP18H06202, JP19K21306, and JP20K16583 to H.N.; JP19H04945, JP19K06504, JP18H05229, JP17H05657, JP17H05867, JP20KK0156, and JP20H03199 to T. Saio; JP19H05769 to K.I.; JP19K16060 to T.Y.; JP20K06493 to T.M.; JP19K16088 and JP21K15051 to K.M.; JP19KK0071 and JP20K06579 to R.I.; JP18H05229, JP18H05534, and JP18H03681 to M. Sugiyama; JP20K06527 to T.O.; JP20H00327 to N.K.; JP21K06031 to S.T.F.; JP18H02391 to M. Sato; JP20H05925 and JP21H04763 to H.T.; JP18K06094 to H.M.; JP19K23976 to M.N.; JP19K17043 to T. Shiota; JP21K15032 to S.K.; JP19K17044 to N.I.; and JP19K07978 to T.K.], JST FOREST Program JPMJFR204W to T. Saio; JST CREST Program JPMJCR1762 to N.K.; the Cooperative Research Project Program of Life Science Center for Survival Dynamics, Tsukuba Advanced Research Alliance (TARA Center), University of Tsukuba, Japan (Grant number: 181107, 201904, 202007) to S.T.F.; Joint Research Programs, the Institute of Advanced Medical Sciences, Tokushima University to E.M., T. Saio, and K.I.; Takeda Science Foundation to E.M. and T. Saio; Kanzawa Medical Research Foundation to E.M.; Uehara Memorial Foundation to E.M., T.Y., and S.K.; Nakatomi Foundation to E.M.; Konica Minolta Science and Technology Foundation to E.M.; Naito Foundation to E.M.; MSD Life Science Foundation to E.M.; Mochida Memorial Foundation for Medical and Pharmaceutical Research to E.M.; SENSHIN Medical Research Foundation to E.M.; Terumo Foundation for Life Sciences and Arts to E.M.; Nara Kidney Disease Research Foundation to E.M.; Novartis Research Grants to E.M., H.N., and K.S.; Nara Medical University Grant-in-Aid for Collaborative Research Projects to K.S.; Akiyama Life Science Foundation to T. Saio; Northern Advancement Center for Science and Technology to T. Saio; The Sumitomo Foundation to T. Saio; Astellas Foundation for Research on Metabolic Disorders to T. Saio; Senri Life Science Foundation to T. Saio; The Nakajima Foundation to T. Saio; The Asahi Glass Foundation to T. Saio; Daiichi Sankyo Foundation of Life Science to T.Y.; The Nakabayashi Trust For ALS Research Grants-in-Aid to T.Y.; Integrated Research Consortium on Chemical Sciences to Y.A.; Izumi Science and Technology Foundation to Y.A.; Tokyo Biochemical Research Foundation to S.K.; and by unrestricted funds provided to E.M. from Dr. Taichi Noda (KTX Corp., Aichi, Japan) and Dr. Yasuhiro Horii (Koseikai, Nara, Japan). This work was also partially supported by the project for Construction of the basis for the advanced materials science and analytical study by the innovative use of quantum beam and nuclear sciences in KURNS. We thank Steven L. McKnight, Yuh-Min Chook, Charalampos Kalodimos, Masato Kato, Benjamin Tu, Kentaro Shiraki, and Keren-Happuch E. for their critical reading of the manuscript. We thank Fumio Takahashi for his help with microscopic imaging and Yoichi Takeda for his help with ITC data collection. We thank Hiroyuki Kumeta for his help in setting up NMR experiments. The NMR experiments were performed at Hokkaido University Advanced NMR Facility, a member of NMR Platform.

Author contributions

H.N., H.K., A.F., T. Saio, T.Y., and E.M. designed the research. H.N., H.K., A.F., T.U., Y.A., M.N., M.H., S.K., M.M., Y.S., T.N., T.M., N.M., K.M., R.I., T. Saio, and T.Y. performed research. H.N., H.K., A.F., S.K., Y.A., T. Saio, and T.Y. analyzed the data. H.N., Y.A., T. Saio, T.Y., and E.M. wrote the paper. T. Shiota, P.W., R.N., I.N., T.K., M. Sugiyama, T.O., N.K., S.T.F., M. Sato, H.T., S.N, O.S., K.I., H.M., and K.S. helped to analyze and interpret the data, and critically revise the manuscript. T. Saio, T.Y., and E.M. conceptualized the study, developed the study design, supervised the authors throughout the study, and provided expertise in manuscript preparation. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25560-0.

Correspondence and requests for materials should be addressed to T.S., T.Y. or E.M.

Peer review information *Nature Communications* thanks Erik Martin and other, anonymous, reviewers for their contributions to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

4)

frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 23 March 2022
doi: 10.3389/fcell.2022.750829

C9orf72-Derived Proline:Arginine Poly-Dipeptides Modulate Cytoskeleton and Mechanical Stress Response

Tomo Shiota[1†], Riko Nagata[2†], Sotaro Kikuchi[2], Hitoki Nanaura[1], Masaya Matsubayashi[2], Mari Nakanishi[2], Shinko Kobashigawa[2], Noriyoshi Isozumi[2], Takao Kiriyama[1], Kazuaki Nagayama[3], Kazuma Sugie[1], Yoshito Yamashiro[4] and Eiichiro Mori[2,5]**

[1]Department of Neurology, Nara Medical University, Kashihara, Japan, [2]Department of Future Basic Medicine, Nara Medical University, Kashihara, Japan, [3]Micro-Nano Biomechanics Laboratory, Department of Mechanical Systems Engineering, Ibaraki University, Hitachi, Japan, [4]Life Science Center for Survival Dynamics, Tsukuba Advanced Research Alliance, Tsukuba, Japan, [5]V-iCliniX Laboratory, Nara Medical University, Kashihara, Japan

OPEN ACCESS

Edited by:
Rafael Fridman,
Wayne State University, United States

Reviewed by:
Fernando G. Vieira,
Amyotrophic Lateral Sclerosis Therapy Development Institute, United States
Kenichi Funamoto,
Tohoku University, Japan

***Correspondence:**
Yoshito Yamashiro
yamayoshito@tara.tsukuba.ac.jp
Eiichiro Mori
emori@naramed-u.ac.jp

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Signaling, a section of the journal Frontiers in Cell and Developmental Biology

Received: 31 July 2021
Accepted: 07 March 2022
Published: 23 March 2022

Citation:
Shiota T, Nagata R, Kikuchi S, Nanaura H, Matsubayashi M, Nakanishi M, Kobashigawa S, Isozumi N, Kiriyama T, Nagayama K, Sugie K, Yamashiro Y and Mori E (2022) C9orf72-Derived Proline: Arginine Poly-Dipeptides Modulate Cytoskeleton and Mechanical Stress Response. Front. Cell Dev. Biol. 10:750829. doi: 10.3389/fcell.2022.750829

Proline:arginine (PR) poly-dipeptides from the GGGGCC repeat expansion in *C9orf72* have cytotoxicity and bind intermediate filaments (IFs). However, it remains unknown how PR poly-dipeptides affect cytoskeletal organization and focal adhesion (FA) formation. Here, we show that changes to the cytoskeleton and FA by PR poly-dipeptides result in the alteration of cell stiffness and mechanical stress response. PR poly-dipeptides increased the junctions and branches of the IF network and increased cell stiffness. They also changed the distribution of actin filaments and increased the size of FA and intracellular calcium concentration. PR poly-dipeptides or an inhibitor of IF organization prevented cell detachment. Furthermore, PR poly-dipeptides induced upregulation of mechanical stress response factors and led to a maladaptive response to cyclic stretch. These results suggest that the effects of PR poly-dipeptides on mechanical properties and mechanical stress response may serve as a pathogenesis of *C9orf72*-related neurodegeneration.

Keywords: *C9orf72*, PR poly-dipeptides, cytoskeleton, Focal Adhesion (FA), actin

INTRODUCTION

A hexanucleotide repeat expansion in *C9orf72* is common in familial amyotrophic lateral sclerosis (ALS) with frontotemporal dementia (FTD) (C9-ALS/FTD) (DeJesus-Hernandez et al., 2011), which leads to the production of proline:arginine (PR), glycine:arginine (GR), glycine:alanine (GA), proline:alanine (PA), and glycine:proline (GP) poly-dipeptides (Ash et al., 2013; Kwon et al., 2014). Arginine-rich poly-dipeptides show cytotoxicity (Kwon et al., 2014; Jovičić et al., 2015; Hao et al., 2019) and cause cell deformation and degeneration (Kwon et al., 2014; Mizielinska et al., 2014), and intermediate filaments (IFs) were found to be the binding targets of arginine-rich poly-dipeptides (Lee et al., 2016; Lin et al., 2016).

IFs regulate the focal adhesion (FA) formation and cellular mechanical properties (Kim et al., 2016; Hu et al., 2019; Laly et al., 2021). FA is responsible for adhesion to the extracellular matrix, which is formed by protein complexes such as integrin, vinculin, paxillin, focal adhesion kinase (FAK), talin, zyxin, α-Actinin, vasodilator-stimulated phosphoprotein (VASP) (Abercrombie and Dunn, 1975). Cytoskeleton and FA convert intracellular and extracellular mechanical forces into intracellular signals that induce mechanical stress responses (Hurtley, 1998; Fletcher and Mullins,

2010; López-Colomé et al., 2017). The cytoskeletal dysfunction was found as a feature of neurodegenerative diseases, for example Parkinson's disease (McLean et al., 2000; Sweers et al., 2011; Häbig et al., 2013; Lu et al., 2017), Alzheimer's disease (Bamburg and Bloom, 2009; Levy Nogueira et al., 2016), and ALS (Wu et al., 2012; Smith et al., 2014; Nicolas et al., 2018; Zhang et al., 2022). However, it is unclear whether the mechanical stress response is associated with the pathogenesis of neurodegeneration. Therefore, it is crucial to evaluate the cytoskeletal organization, FA formation and the mechanical stress response induced by arginine-rich poly-dipeptides.

In this study, we evaluated the effects of PR poly-dipeptides, the most toxic among five different poly-dipeptides (Kwon et al., 2014), on the cytoskeletal organization, cell stiffness, and FA formation. PR poly-dipeptides induced a high-density network of IFs and increased cell stiffness in conjunction with the abnormal remodeling of actin filaments and FA. PR poly-dipeptides prevented cell detachment. Furthermore, they induced the expression of mechanical stress response factors, leading to a maladaptive mechanical stress response. These results suggest that PR poly-dipeptides alter the stability of IFs, resulting in changes to mechanical stress responses.

MATERIALS AND METHODS

Peptide Synthesis

A synthetic peptide consisting of twenty repeats of the PR poly-dipeptide (PR_{20}) with an HA tag at the carboxyl terminus was synthesized (SCRUM Inc., Tokyo, JAPAN).

Cell Culture

Human osteosarcoma cells (U2OS) and fibroblast (BJ-hTERT) cells were cultured in Dulbecco's modified Eagle medium (DMEM) high glucose with 10% fetal bovine serum (FBS, Hyclone) and 1% penicillin-streptomycin at 37°C in 5% CO_2. The cells were used for experiments after a one-hour treatment at 37°C with a synthetic peptide consisting of PR_{20} (final concentration 10 μM). Rat vascular smooth muscle cells (SMCs; Lonza, R-ASM-580) were grown in DMEM with 20% FBS and 1x Antibiotic-Antimyotic (Thermo Fisher Scientific).

Immunofluorescence

The U2OS cells, fibroblast BJ-hTERT cells or rat vascular SMCs were fixed in 4% paraformaldehyde in phosphate-buffered saline (PBS) at room temperature for 15 min and permeabilized with 0.1% Triton X-100 in PBS for 10 min. The fixed cells were incubated with a blocking solution (5% bovine serum albumin in PBS with 0.1% Tween20) at room temperature for 1 h. The cells were incubated with primary antibodies, vimentin (Santa Cruz Biotechnology, sc6260, 1:100), cytokeratin 14 (Abcam, ab7800, 1:250), vinculin (Abcam, ab129002, 1:500), phospho-Paxillin (Cell Signaling, 2541, 1:50), phospho-ERM (Cell Signaling, 3726, 1:200), or phospho-Cofilin (Cell Signaling, 3313, 1:100) in the blocking solution at 4°C overnight. Secondary antibodies (Thermo Fisher Scientific, A-21422, A-21429, 1:2000) were incubated at room temperature for 1 h in the blocking solution. Alexa Fluor488-phalloidin (Thermo Fisher Scientific, A12379, 1:1000) and Alexa Fluor555-β-tubulin (Abcam, ab206627, 1:1000) were also incubated at room temperature for an hour in the blocking solution. Images were captured using the confocal microscope FV3000 (Olympus, Tokyo) or LSM 710 (ZEISS). Captured images were analyzed by ImageJ (version 1.53d, National Institutes of Health (NIH), United States).

Imaging Analysis

Analysis of IF organization was performed by quantifying the junctions, branches, and alignments. The images were converted to 8-bit and binarized using fixed contrast and brightness thresholds. For junction and branch measurements, 100 μm^2 regions of interest (ROI) in the cytoplasm, excluding the nucleus, were selected. At least one ROI was selected per cell. The number of junctions and branches was measured in the skeletonized images using the skeleton plugin (https://imagej.net/AnalyzeSkeleton, last accessed 1 March 2022). The bundle alignments of IFs were measured by the dispersion of the IF angles in whole cells (Laly et al., 2021), using the OrientationJ plugin (http://bigwww.epfl.ch/demo/orientation/, last accessed 1 March 2022). Fluorescence intensity of IFs in perinuclear regions was measured within 1 μm from the nucleus. Fluorescence intensity of actin filaments was evaluated by the Quimp plugin, which outlines a cell using their original algorithm (Baniukiewicz et al., 2018). Based on the outline, the plugin measures the fluorescence intensity of the whole cell and around the cell cortex. As the fluorescence intensity around the cell cortex is affected by the size of the cell, the ratio of fluorescence intensity around the cell cortex to the fluorescence intensity of the whole cell was calculated. To measure the fluorescence intensity of pERM, a region of cells was defined in the merged image. The average fluorescence intensity of this region was measured on the pERM image. The average fluorescence intensity was divided by the number of cells in the ROI to obtain the fluorescence intensity of pERM per cell.

Analysis of FA was performed by quantifying the size and the number of FA. The images were converted to 8-bit and binarized using fixed contrast and brightness thresholds. The average size and the number of FAs per cell were quantified from binary images using the "Analyze Particle" function in ImageJ (https://imagej.net/imaging/particle-analysis, last accessed 1 March 2022) with the following settings: size (pixel̈2), 0.5-infinity; circularity, 0.00–1.00. The polymerized β-tubulin was measured using the Tubeness plugin (https://imagej.net/plugins/tubeness, last accessed 1 March 2022), which identified tube-like structures. The ratio of the polymerized β-tubulin to cell area was calculated.

Atomic Force Microscopy

Atomic force microscopy (AFM) measurements were performed using a NanoWizard IV AFM (JPK Instruments-AG, Germany) mounted on top of an inverted optical microscope (IX73, Olympus, Japan) equipped with a digital CMOS camera (Zyla, Andor) as described in a previous study (Nagayama et al., 2019). Prior to AFM imaging of the surface topography and mechanical properties of U2OS cells in PR_{20}-treated cells, the cells were adapted to a CO_2-independent medium (Invitrogen) for 30 min at room temperature (25 °C). AFM quantitative imaging (QI)

mode was used to obtain a force–displacement curve at a resolution of 128 × 128 pixels (100 μm × 100 μm of measured area) by a precisely controlled high-speed indentation test using rectangular-shaped silicon nitride cantilevers with a cone probe (BioLever-mini, BL-AC40TS-C2, Olympus, Japan). The test was performed at a spring constant of 0.08–0.10 N/m and a nominal tip radius of 10 nm. The QI mode measurements were performed within an hour after the transfer of the specimen to the AFM. These high-speed indentations were performed until a preset force of 1 nN was reached. This typically corresponded to cell indentation depths of 300–400 nm. Cell elasticity was calculated from the obtained force–displacement curves by applying the Hertzian model (Hertz, 1881), which approximates the sample to be isotropic and linearly elastic. Young's (elastic) modulus is extracted by fitting all force–displacement curves with the following Hertzian model approximation:

$$F = \frac{2E \cdot \tan\alpha}{\pi(1-\nu^2)}\delta^2$$

where F is the applied force, E is the elastic modulus, ν is the Poisson's ratio (0.5 for a non-compressible biological sample), α is the opening angle of the cone of the cantilever tip, and δ is the indentation depth of the sample recorded in the force–displacement curves. Using the results of the Hertzian model approximation, we identified the Z contact points (specimen surface) and the elastic modulus of the specimens at each pixel and produced a surface topography map and elastic modulus map of the specimens.

Intracellular Calcium Imaging

The intracellular calcium concentration was analyzed using the Calbryte590™ assay kit (AAT Bioquest, United States). The U2OS cells were seeded in 96-well glass bottom microplates and cultured for 24 h. The cells were incubated for an hour at 37°C with Calbryte590™. After removal of the Calbryte590™ solution, the cells were administrated to 10 μM of PR_{20} for an hour and analyzed by a confocal microscopy and SpectraMax microplate reader (Molecular Devices, United States).

Cell Detachment Experiment

The U2OS cells were transfected with LifeAct-GFP plasmids (60101, ibidi GmbH) using FuGENE-HD transfection reagent (E2311, Promega). For the cell detachment experiment, transfected cells were seeded in glass bottom plates and cultured for 24 h. The cells were administered to 10 μM of PR_{20} for an hour, 50 μM of (-)-Blebbistatin (B0560-1MG, Sigma-Aldrich) for an hour (Korobova et al., 2014), 10 μM of Nocodazole (M1404-10MG, Sigma-Aldrich) for 30 min (Baum et al., 2014), 2 μM of Withaferin-A (ab120644, Abcam) for 3 hours (Grin et al., 2012), and observed after 30 min incubation with 5 μM of EDTA. EDTA reduces integrin-mediated adhesion, which is a Ca^{2+}-dependent FA molecule by chelating Ca^{2+} and Mg^{2+}, causing cell detachment (Mould et al., 1995). Equivalent amounts of DMSO (D8418, Sigma–Aldrich) were added as vehicle control, as Blebbistatin, Nocodazole and Withaferin-A were diluted with DMSO. Blebbistatin, Nocodazole and Withaferin-A are used as cytoskeleton inhibitors. Blebbistatin is an inhibitor of myosinII and inhibits the formation of actomyosin and suppresses stress fiber formation (Kovács et al., 2004). Nocodazole binds to the growing end of β-tubulin and prevents microtubule polymerization (Chang et al., 2008). Withaferin-A binds to the vimentin rod domain and disrupts both Vim-IF and Keratin-IF network (Bargagna-Mohan et al., 2007; Grin et al., 2012). Images were captured using a confocal microscope FV3000 (Olympus, Tokyo). Captured images were analyzed by ImageJ (version 1.53d, NIH, United States). As detached cells with a round shape showed high circularity (Fuhrmann et al., 2014), degrees of cell detachment were analyzed by measuring the circularity of cells.

Cyclic Stretch Experiment

Cyclic stretch was performed using a uniaxial cell stretch system (Central Workshop Tsukuba University) as described in a previous study (Yamashiro et al., 2020). We employed rat vascular smooth muscle cells (SMCs) or U2OS cells. The cells were plated on silicon elastomer bottomed culture plates (SC4Ha, Menicon Life Science) coated with a cell attachment factor containing gelatin (Thermo Fisher Scientific, S006100) and subjected to cyclic stretch with a frequency of 1.0 Hz (60 cycles/min) and 20% strain for 6 h.

Western Blotting

Total proteins are isolated from rat vascular SMCs treated with 10 μM of PR_{20} using RIPA lysis buffer. After sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE), proteins were transferred to polyvinylidene difluoride membranes (PVDF). Membranes were blocked with and incubated overnight at 4°C with the following primary antibodies: Thbs1 (NeoMarkers, MS-421), Egr1 (Cell Signaling, 4154), phospho-FAK(Y397) (Cell Signaling, 3283), FAK (Cell Signaling, 3285), phospho-ERK (Cell Signaling, 4376), ERK (Cell Signaling, 9102), phospho-ERM (Cell Signaling, 3141), ERM (Cell Signaling, 3142), GAPDH (Cell Signaling, 2118). Then, membranes were incubated with second antibodies anti-mouse or anti-rabbit HRP-conjugated antibody (1: 1000, Bio-Rad). Blots were visualized using a chemiluminescence kit (Santa Cruz Biotechnology) or SuperSignal West Femto Maximum Sensitivity Substrate (Thermo Fisher Scientific).

Statistical Analysis

All experiments are presented as means ± SD (except for **Figure 6B**). Statistical analysis was performed using Prism 8 (GraphPad Software, California, United States). A Shapiro-Wilk test was used to test for normality. The Mann–Whitney U test, a nonparametric test, or unpaired *t* test was conducted. $p < 0.05$ denotes statistical significance.

RESULTS

PR Poly-Dipeptides Increase the Junctions and Branches of the Intermediate Filament Network

To investigate the effect of PR poly-dipeptides on cytoskeleton, we first examined the changes of Vimentin (Vim) and Keratin

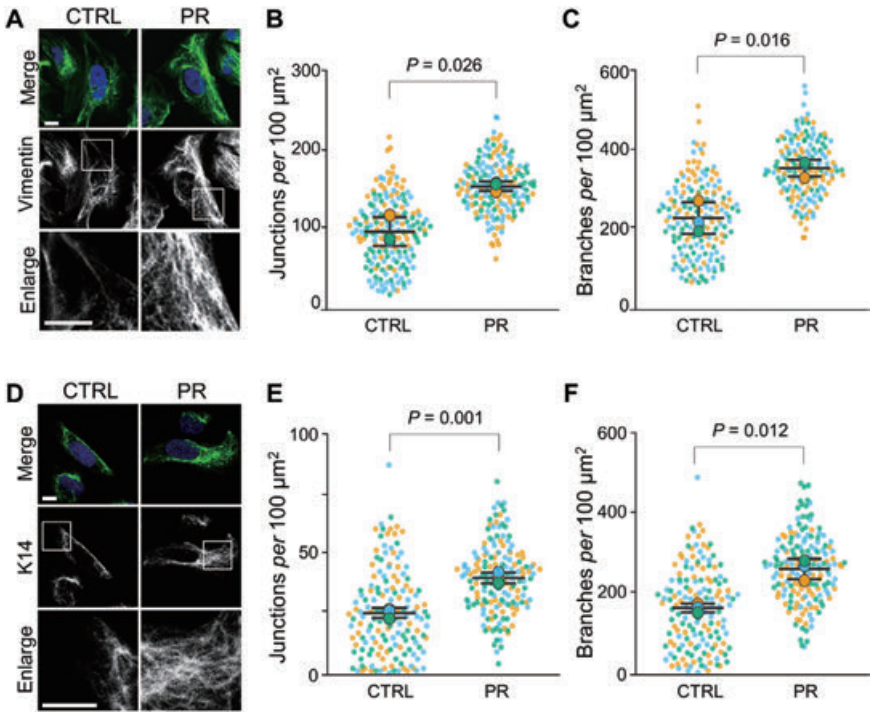


FIGURE 1 | PR poly-dipeptides (PR_{20}) change the organization of intermediate filaments (IF) of U2OS cells. **(A)** Fluorescent images of Vim-IFs in untreated control (CTRL) and PR_{20}-treated (PR) cells. Vim-IFs (green) and DAPI (blue) are shown. Scale bars are 10 µm. **(B,C)** Quantification of Vim-IF junctions (in B) and branches (in C) per 100 μm^2 in the cytoplasm in CTRL and PR cells. Bars are means ± SD. In each experiment, 50 to 60 cells were evaluated ($n = 3$). P value shows Welch's *t*-test. **(D)** Fluorescent images of cytokeratin14 (K14) in CTRL and PR cells. K14 (green) and DAPI (blue) are shown. Scale bars are 10 µm. **(E,F)** Quantification of Keratin-IF junctions (in E) and branches (in F) per 100 μm^2 in the cytoplasm in CTRL and PR cells. Bars are means ± SD. In each experiment, 50 to 58 cells were evaluated ($n = 3$). *p* value shows Welch's *t*-test. In quantification graphs, small dots are individual datapoints, and large dots indicate the averages of each experiment in orange (first; $n = 1$), blue (second; $n = 2$) and green (third; $n = 3$), respectively.

intermediate filament organization in U2OS cells. Vim-IFs were predominantly localized to the perinuclear region in the control cells (CTRL; vehicle control) (**Figure 1A**). After 1 h of exposure to 10 µM of PR_{20}, the fluorescence intensity of Vim-IFs in the perinuclear region decreased ($p = 0.028$; 19.8 ± 2.8 a.u. in CTRL, 12.6 ± 1.5 a.u. in PR_{20}-treated; **Supplementary Figure S1A**) and Vim-IFs formed a mesh network in the cytoplasm (**Figure 1A**). The IF organization was analyzed by the alignments and the number of branches and junctions in skeletonized images (**Supplementary Figure S1B**). Vim-IF alignments were comparable between CTRL and PR_{20}-treated cells ($p = 0.678$; 21.0 ± 2.2 in CTRL, 21.9 ± 2.5 in PR_{20}-treated; **Supplementary Figure S1C**). There were more junctions of Vim-IFs in PR_{20}-treated cells than in CTRL ($p = 0.026$; 97.3 ± 18.1 in CTRL, 154.5 ± 5.9 in PR_{20}-treated; **Figure 1B**). There were also more branches of Vim-IFs in PR_{20}-treated cells than in CTRL ($p = 0.016$; 232.9 ± 39.5 in CTRL, 358.2 ± 21.0 in PR_{20}-treated; **Figure 1C**).

Next, Keratin-IFs were analyzed using cells stained with anti-cytokeratin14 (K14) antibody. Keratin-IFs formed a mesh network in the cytoplasm of PR_{20}-treated cells (**Figure 1D**). The fluorescence intensity of Keratin-IFs in the perinuclear region was comparable between CTRL and PR_{20}-treated cells ($p = 0.428$; 12.7 ± 1.3 in CTRL, 13.9 ± 2.2 in PR_{20}-treated; **Supplementary Figure S1D**). Keratin-IF alignments were also comparable between CTRL and PR_{20}-treated cells ($p = 0.919$; 44.7 ± 3.6 in CTRL, 45.1 ± 5.2 in PR_{20}-treated; **Supplementary Figure S1E**). There were more junctions ($p = 0.001$; 25.5 ± 2.1 in CTRL, 40.3 ± 2.3 in PR_{20}-treated) and branches ($p = 0.012$; 165.9 ± 11.0 in CTRL, 264.3 ± 26.0 in PR_{20}-treated) of Keratin-IFs in PR_{20}-treated cells than in CTRL (**Figures 1E,F**). Together, these results indicate that PR poly-dipeptides induce a high-density network of IFs in the cytoplasm.

PR Poly-Dipeptides Increase Cell Stiffness

To evaluate the effects of PR poly-dipeptides on cellular mechanics, we analyzed the cell stiffness and cell height by atomic force microscopy (AFM) in U2OS cells (**Figure 2A**). The number of filamentous actin (F-actin) stress fibers across the whole cell was reduced, and F-actin stress fibers around the nucleus were thick and short in PR_{20}-treated cells (**Figure 2B**, middle, arrows). We measured the average elastic modulus at the cell central region. Cell elasticity at the cell central region increased in PR_{20}-treated cells than in CTRL ($p = 0.032$; 8.6 ± 5.1 kPa in CTRL, 12.9 ± 8.2 kPa in PR_{20}-treated; **Figure 2C**). There were no significant differences in cell height between CTRL

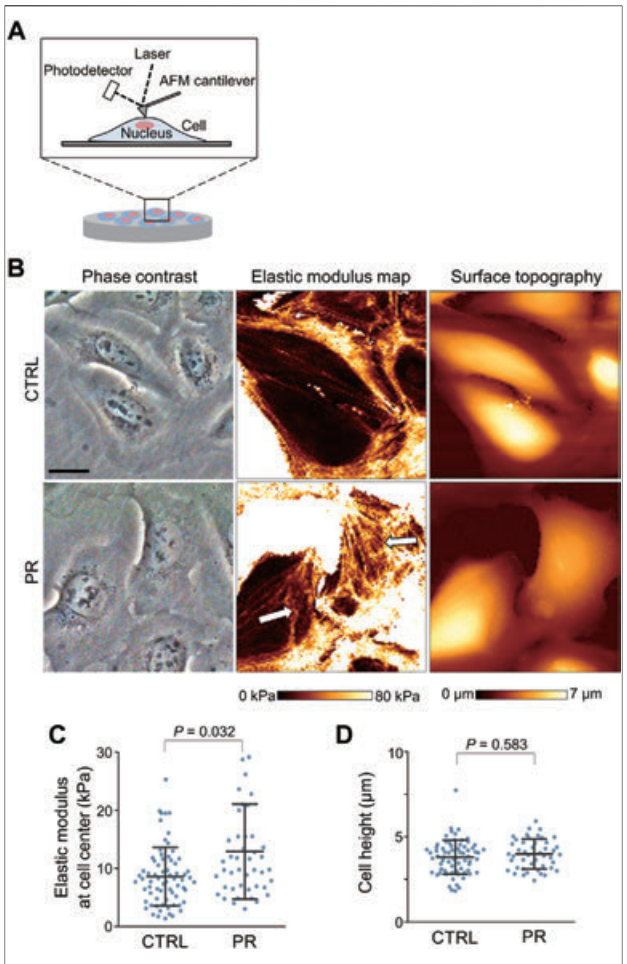


FIGURE 2 | Atomic force microscopy (AFM) images of the surface of U2OS cells. **(A)** Schematic view of AFM measurement of cell surface, including cell stiffness and cell height. **(B)** Representative images of phase contrast images (left), elastic modulus maps (middle), and surface topographic images (right) of U2OS cells measured by AFM. Arrows in the middle image show thick- and short-stress fibers in PR_{20}-treated (PR) cells. Scale bars are 20 μm. **(C,D)** Quantification graph shows elastic modulus at cell central region (in C) and cell height (in D) in CTRL and PR cells. Bars are means ± SD. 45 to 69. Cells were evaluated per condition. P value shows Mann–Whitney U test.

and PR_{20}-treated cells (p = 0.583; 3.9 ± 1.0 μm in CTRL, 4.0 ± 0.9 μm in PR_{20}-treated; **Figure 2D**). These results suggest that PR poly-dipeptides increase cell stiffness by thick and short F-actin stress fibers.

PR Poly-Dipeptides Change the Distribution of F-Actin

To investigate the effects of PR poly-dipeptides on other cytoskeletal elements, we analyzed the distribution of F-actin by fluorescence images (**Supplementary Figure S2A**). In CTRL, F-actin extended across the whole cell and equally distributed in the cytoplasm, whereas after PR_{20}-treatment, F-actin was not distributed equally in the cytoplasm but mainly localized around the cell cortex in U2OS cells and BJhTERT cells (**Figure 3A**). PR_{20}-treated cells showed a higher fluorescence intensity of F-actin around the cell cortex than CTRL in both cell lines (U2OS; p = 0.007; 8.4 ± 0.6%, in CTRL, 11.0 ± 0.7% in PR_{20}-treated cells. BJhTERT; p = 0.032; 9.1 ± 0.2%, in CTRL, 10.8 ± 0.6% in PR_{20}-treated cells; **Figures 3B,C**). Similar to U2OS cells and BJhTERT cells (**Figure 3A**), PR_{20} changed F-actin organization in rat vascular smooth muscle cells (SMCs) (**Supplementary Figure S2B**). PR_{20} also caused nuclear deformation and the unclear boundary between the nucleus and the endoplasmic reticulum (ER) in SMCs (**Supplementary Figure S2C**). These results suggest that PR poly-dipeptides change the distribution of F-actin.

To evaluate whether the redistribution of F-actin by PR poly-dipeptides result from the F-actin reorganization, we analyzed Ezrin/Radixin/Moesin (ERM) proteins and Cofilin. As cortical F-actin is formed by ERM proteins (Bretscher et al., 2002), we analyzed cortical F-actin by ERM phosphorylation in U2OS cells (**Supplementary Figure S2D**). Phospho-ERM (pERM) localized to protrusive structures, such as filopodia, both in CTRL and PR_{20}-treated cells (**Figure 3D**). The fluorescence intensity of pERM per cell was increased in PR_{20}-treated cells (p = 0.048; 0.1 ± 0.0 a.u. in CTRL, 0.4 ± 0.1 a.u. in PR_{20}-treated; **Figure 3E**). As F-actin is depolymerized by phosphorylated Cofilin (pCofilin) (Sumi et al., 1999), we evaluated pCofilin. The level of pCofilin was comparable between CTRL and PR_{20}-treated cells in U2OS cells and SMCs (**Supplementary Figures S2E,F**). These results suggest that PR poly-dipeptides enhance the reorganization of cortical F-actin through pERM.

To evaluate the effects of PR poly-dipeptides on microtubules, we analyzed the ratio of polymerized β-tubulin. In CTRL, β-tubulin showed a filamentous network in the cytoplasm, but it appeared as dots with a diffused network in PR_{20}-treated cells (**Figure 3F**). Compared to CTRL, PR_{20}-treatment reduced the ratio of polymerized β-tubulin (p = 0.035; 13.9 ± 1.0% in CTRL, 11.3 ± 0.3% in PR_{20}-treated cells; **Figure 3G**). These results suggest that PR poly-dipeptides disrupt the stability of the microtubule network.

PR Poly-Dipeptides Increase FA Size and Intracellular Calcium Concentration

To investigate the effects of PR poly-dipeptides on FA formation, we analyzed the localization, the number and size of Vinculin and pPaxillin in U2OS cells (**Supplementary Figure S3A**). In CTRL, we found Vinculin and pPaxillin to be equally distributed in the cytoplasm (**Supplementary Figure S3B**; **Figure 4A**). On the other hand, in PR_{20}-treated cells, Vinculin and pPaxillin were reduced in the cell central region and mainly observed in the cell periphery (**Supplementary Figure S3B**; **Figure 4A**). Although the number of Vinculin decreased in PR_{20}-treated cells (p = 0.003; 68.9 ± 3.6 in CTRL, 36.6 ± 6.2 in PR_{20}-treated; **Supplementary Figure S3C**), the size of Vinculin was comparable between CTRL and PR_{20}-treated cells (p = 0.506; 1.6 ± 0.1 μm^2 in CTRL, 1.7 ± 0.3 μm^2 in PR_{20}-treated; **Supplementary Figure S3D**). There was no significant difference in the fluorescence intensity of pPaxillin

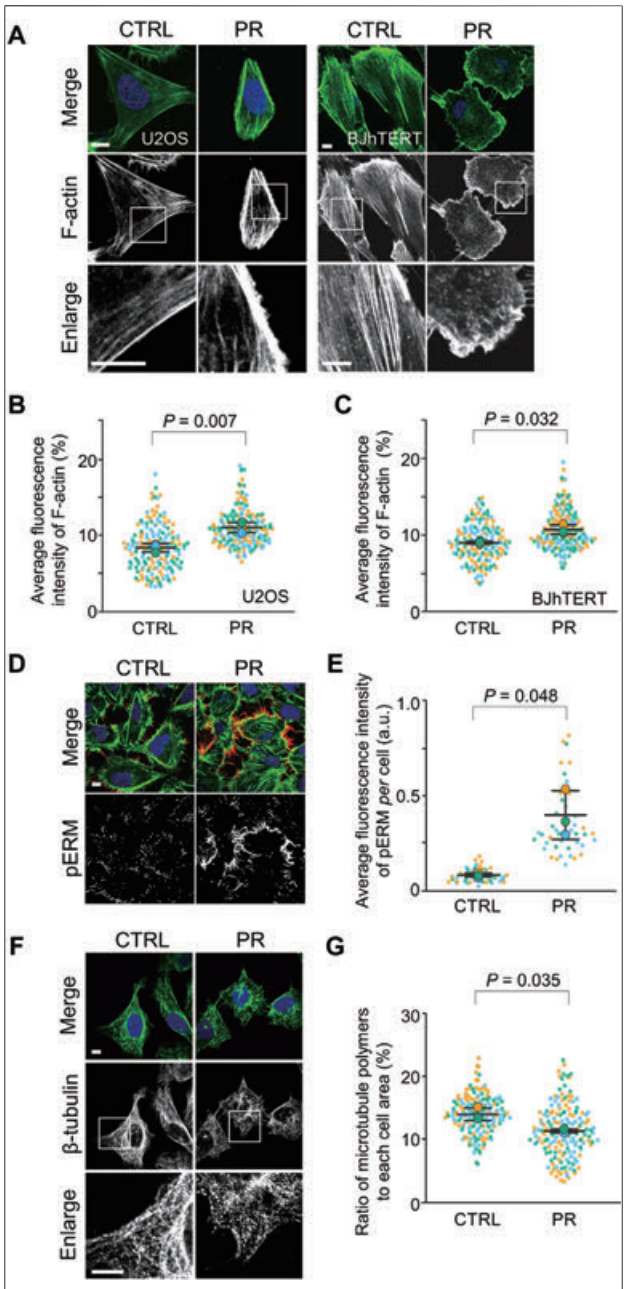


FIGURE 3 | PR_{20} changes the organization of F-actin and β-tubulin and promotes the phosphorylation of ERM (pERM). **(A)** Fluorescent images of phalloidin in U2OS and BJhTERT cells after treatment with 10 μM of PR_{20} or CTRL. phalloidin (green) and DAPI (blue) are shown. Scale bars are 10 μm. **(B,C)** Quantification of average fluorescence intensity of F-actin in U2OS cells (in B) and BJhTERT cells (in C). Bars are means ± SD. In each experiment, 50 to 55 cells were evaluated (n = 3). p value shows Welch's t-test. **(D)** Fluorescent images of pERM (red), phalloidin (green) and DAPI (blue) after treatment with or without 10 μM of PR_{20} in U2OS cells. A bar is 10 μm. **(E)** Average fluorescence intensity of pERM per cell after treatment with 10 μM of PR_{20} or CTRL. Bars are means ± SD. 15 regions were evaluated (n = 3). p value shows Welch's t-test. **(F)** Fluorescent images of β-tubulin in U2OS cells after treatment with 10 μM of PR_{20} or CTRL. Bars are 10 μm. **(G)** The ratio of

(*Continued*)

FIGURE 3 | microtubule polymers in CTRL and PR_{20}-treated (PR) cells. Bars are means ± SD. In each experiment, 50 to 57 cells were evaluated (n = 3). p value shows Welch's t-test. In quantification graphs, small dots are individual datapoints, and large dots indicate the averages of each experiment in orange (first; n = 1), blue (second; n = 2) and green (third; n = 3), respectively.

between CTRL and PR_{20}-treated cells (p = 0.131; 3.3 ± 1.0 a.u. in CTRL, 6.2 ± 2.2 a.u. in PR_{20}-treated; **Figure 4B**). However, the size of pPaxillin was larger in PR_{20}-treated cells than in CTRL (p = 0.034; 1.4 ± 0.0 μm^2 in CTRL, 1.6 ± 0.1 μm^2 in PR_{20}-treated; **Figure 4C**). These results suggest that PR poly-dipeptides induce the maturation of FA.

Since FA size is regulated by intracellular calcium (Huang et al., 2018), we measured the intracellular calcium concentration using calcium indicator calbryte590™ (**Supplementary Figure S3E**). The fluorescent images of PR_{20}-treated cells showed a stronger fluorescence signal compared to CTRL cells (**Figure 4D**). The calbryte590™ fluorescence intensity using a microplate reader was higher in PR_{20}-treated cells than in CTRL (p = 0.038; 0.7 ± 0.2 in CTRL, 1.3 ± 0.2 a.u. in PR_{20}-treated; **Figure 4E**). These results suggest that PR poly-dipeptides increase intracellular calcium concentration.

PR Poly-Dipeptides Prevent Cell Detachment

To investigate the impact of cytoskeletal organization on adhesive strength, U2OS cells were transfected with Lifeact-GFP, which enables live imaging of F-actin, and challenged by cell detachment using EDTA after PR poly-dipeptides and cytoskeleton inhibitors, Withaferin-A, Blebbistatin and Nocodazole. Before EDTA treatment, F-actin stress fibers extended across the whole cell in CTRL. PR_{20}-treated cells enhanced cortical F-actin compared to CTRL (**Figure 5A**), while Withaferin-A-treated cells were comparable to CTRL (**Figure 5A**). Blebbistatin-treated cells showed fewer F-actin stress fibers than CTRL (**Figure 5A**). Nocodazole-treated cells showed more F-actin stress fibers than CTRL (**Figure 5A**). After EDTA-treatment, CTRL and Nocodazole-treated cells showed a rounded shape and F-actin was localized in the cell cortex, whereas PR_{20}, Withaferin-A and Blebbistatin-treated cells maintained a spread shape and formed protrusions (**Figure 5A**).

As the detached cells showed a rounded shape and high circularity (Fuhrmann et al., 2014), degrees of cell detachment were analyzed by measuring the circularity of cells. The circularity of PR_{20}-treated cells was lower than CTRL (p < 0.001; 0.9 ± 0.1 in CTRL, 0.7 ± 0.2 in PR_{20}; **Figure 5B**; **Supplementary Figure S4**). Withaferin-A-treated cells (p < 0.001; 0.8 ± 0.0) and Blebbistatin-treated cells (p < 0.001; 0.7 ± 0.2) also decreased circularity after EDTA-treatment (**Figure 5B**; **Supplementary Figure S4**). Nocodazole-treated cells showed no significant difference in the circularity compared to CTRL after EDTA-treatment (p = 0.170; 0.9 ± 0.0 in Nocodazole; **Figure 5B**; **Supplementary Figure S4**). These results indicate that PR poly-dipeptides prevent cell detachment via the reorganization of IFs and F-actin.

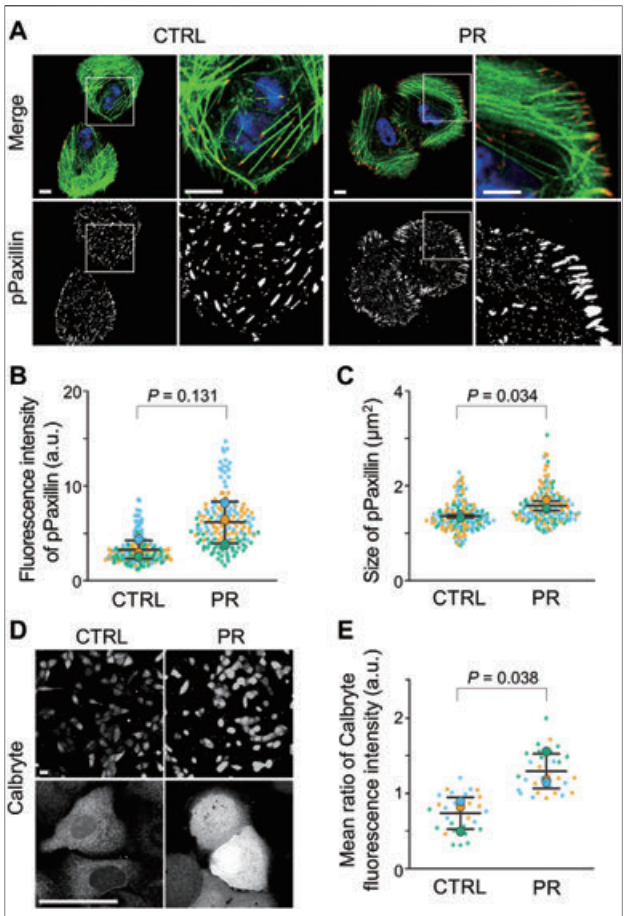


FIGURE 4 | PR_{20} increases the size of FA and intracellular calcium concentration. **(A)** Fluorescent images of pPaxillin (red), phalloidin (green) and DAPI (blue) are shown. Scale bars are 10 μm. **(B,C)** Average fluorescence intensity of pPaxillin (in B) and size of pPaxillin (in C). Bars are means ± SD. In each experiment, 41 to 57 cells were evaluated ($n = 3$). p value shows Welch's t-test. **(D,E)** Fluorescence images of Calbryte590™ in CTRL and PR_{20}-treated (PR) cells. Scale bars are 50 μm. Quantification shows in E. Bars are means ± SD. In each experiment, 10 cells were evaluated ($n = 3$). p value shows Welch's t-test. In quantification graphs, small dots are individual datapoints, and large dots indicate the averages of each experiment by orange (first; $n = 1$), blue (second; $n = 2$), and green (third; $n = 3$), respectively.

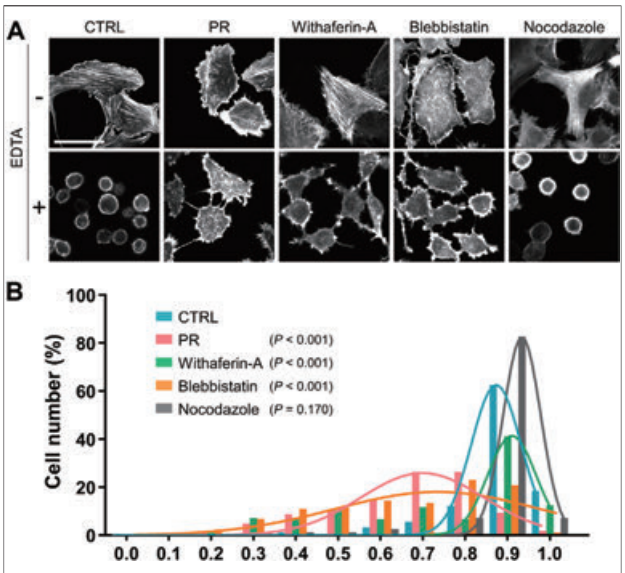


FIGURE 5 | PR_{20} inhibits cell detachment by EDTA. **(A)** Representative LifeAct-GFP images of U2OS cells in CTRL and PR_{20}-treated (PR) and the cytoskeletal inhibitors, Withaferin-A, Blebbistatin and Nocodazole treatment, following detachment by 5 μM EDTA. A bar is 50 μm. **(B)** A histogram of the cell circularity of U2OS cells after detachment by 5 μM EDTA in CTRL (n = 179, blue), PR_{20} (n = 158, red), Withaferin-A (n = 243, green), Blebbistatin (n = 316, orange), and Nocodazole (n = 242, gray). A circularity of 1.0 indicates a complete circle. As it approaches 0.0, it indicates an elongated polygon. p value shows Welch's t-test.

PR Poly-Dipeptides Attenuate the Cyclic Stretch-Induced Reorientation of F-Actin Stress Fibers

Thus far, we have observed that PR poly-dipeptides induced changes in cytoskeletal organization and FA formation, leading to the increase in cell stiffness and the prevention of cell detachment. Therefore, we assume that PR poly-dipeptides might alter the mechanical stress response. To test this hypothesis, we used rat SMCs and U2OS cells and performed uniaxial stretch experiment. Mechanical stretch responses were evaluated by the reorientation angle and the expression of mechanical stress response factor, such as early growth response 1 (Egr1), Thbs1, phosphorylation of extracellular signal-regulated kinase (ERK) and FAK as previously reported (Yamashiro et al., 2015; Yamashiro et al., 2018; Yamashiro et al., 2020). Interestingly, in rat SMCs, PR_{20}-treatment activated FA and upregulated the expression of mechanical stress response factors prior to stretch-stimuli (**Figures 6A,B**). Although CTRL cells were aligned perpendicularly to stretch direction, PR_{20}-treated cells failed to align correctly and decreased number of cells after stretching (**Figure 6C**). A histogram of the percentage of the orientation angle (θ) showed that PR_{20} suppressed the cyclic stretch-induced reorientation of F-actin stress fibers ($p < 0.001$, 75.3 ± 10.5° in CTRL, 59.4 ± 23.2° in PR_{20}; **Figure 6D**). U2OS cells showed similar results as rat SMCs ($p < 0.001$, 73.1 ± 10.7° in CTRL, 65.8 ± 8.8° in PR_{20}; **Figures 6E,F**). Consequently, these results suggest that PR poly-dipeptides promote the activation of FA and induce upregulation of mechanical stress response factors, leading to sensitive and maladaptive responses to mechanical stimuli.

DISCUSSION

Here we show that PR poly-dipeptides increased the junctions and branches of the IF network and cell stiffness. This is in conjunction with changes to the distribution of F-actin and the enhancement of FA size. We also demonstrated that cytoskeletal reorganization induced by PR poly-dipeptides prevented cell detachment and led to maladaptive responses to cyclic stretch.

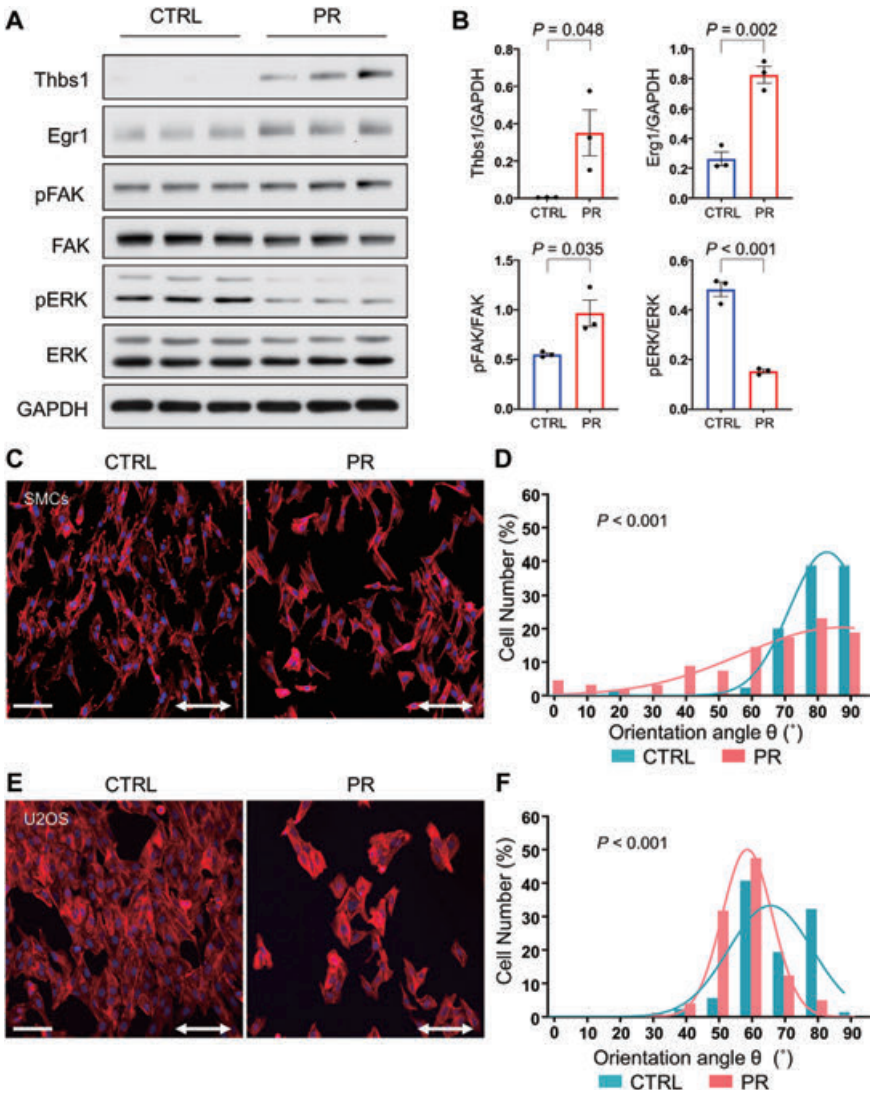


FIGURE 6 | PR_{20} prohibits the cyclic stretch-induced reorientation of F-actin stress fibers. **(A)** Western blotting shows Thbs1, Egr1, pFAK, FAK, pERK and ERK levels in CTRL and PR_{20}-treated (PR) cells ($n = 3$). GAPDH was used as a loading control. **(B)** Quantification graphs are shown. Bars are means ± SEM. p value shows unpaired t-test. **(C,E)** Rat vascular SMCs (in C) and U2OS cells (in E) with or without 10 μM of PR_{20} were subjected to cyclic stretch (20% strain, 1.0 Hz (60 cycles/min) for 6 hours. Phalloidin (red) and DAPI (blue) are shown. Scale bars are 100 μm. The two-way arrows indicate stretch direction. **(D,F)** Histograms of the percentage of the orientation angle (θ) for each cell in rat vascular SMCs (in D) and U2OS cells (in F). The orientation of each cell was analyzed by measuring the orientation angle (θ) of the long axis of the ellipse relative to the stretch axis in CTRL (blue) and PR cells (red). 70 to 104 cells were evaluated in each condition. p value shows Mann–Whitney U test.

PR poly-dipeptides bind proteins with low-complexity domains and nuclear import receptors (Lee et al., 2016; Lin et al., 2016; Hayes et al., 2020; Hutten et al., 2020; Nanaura et al., 2021). Head domains of IF proteins self-associate *via* the formation of labile but structurally specific cross-β interactions (Zhou et al., 2021), and PR poly-dipeptides target these polymeric forms of the IF head domains (Lin et al., 2016). In this study, we show that PR poly-dipeptides increased the junctions and branches of the IF network (**Figure 1**). We speculate that PR poly-dipeptides stabilize these dynamic cross-β interaction of the IF head domains, and lead to the malfunction and dysregulation of IF network formation (**Figure 7**). However, the detailed mechanisms of how PR poly-dipeptides interfere with cytoskeleton in cells remain unknown.

As the IF network contributes to cell stiffness (Laly et al., 2021), the high-density network of IFs induced by PR poly-dipeptides might increase cell stiffness. PR poly-dipeptides also reduced the fluorescence intensity of Vim-IFs in the perinuclear region but not Keratin-IFs (**Figures 1A,D**). These data might result from a difference in IF localization—Vim-IFs diffuse in the

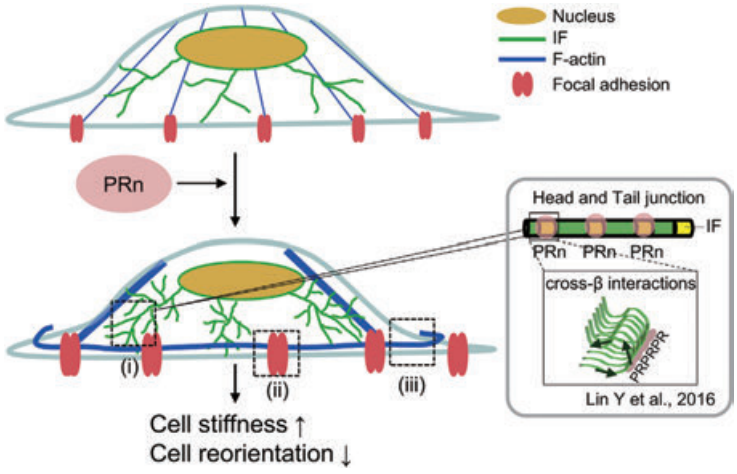


FIGURE 7 | Graphical abstract. PR poly-dipeptides bind to IF head domain and induced the alterations of cellular mechanics. PR poly-dipeptides (pink) bind to the IF head domain (yellow) and dysregulate the IF dynamics. This results in an increase of IF branches and IF mesh network (green) (i). The IF network increases the size of FA (red) (ii) and thickness of F-actin stress fibers (blue) (iii), leading to cell stiffness. Further, the alterations of cytoskeleton and FA disrupt cell reorientation after cyclic stretch. IF: intermediate filament, PRn: PR poly-dipeptides.

perinuclear region and the cytoplasm, whereas Keratin-IFs mainly distribute in the cytoplasm (Grin et al., 2012; Li et al., 2019).

PR poly-dipeptides induced the formation of cortical F-actin and the maturation of FA (**Figures 3**, **4**). The formation of cortical F-actin and the maturation of FA are regulated by IF organization (Helfand et al., 2011). The IF organization by PR poly-dipeptides might alter the formation of cortical F-actin and the maturation of FA. Cortical F-actin is formed by ERM proteins (Bretscher et al., 2002). ERM proteins also induce FAK activation via signaling pathways independent of FA (Poullet et al., 2001). PR poly-dipeptides might induce the maturation of FA via ERM phosphorylation. Nonetheless, it is still unclear how PR poly-dipeptides affect the formation of cortical F-actin and the maturation of FA.

The dysregulation of intracellular calcium is reported as a cause of cell death in C9-ALS/FTD (Li et al., 2020; Choi et al., 2019; Dafinca et al., 2016; Dafinca et al., 2020; Baughman et al., 2011). PR poly-dipeptides increased the intracellular calcium concentration (**Figures 4D,E**). The increase in intracellular calcium concentration promotes cytoskeletal reorganization and FA maturation (Price et al., 2003; Hartzell et al., 2016; O'Brien et al., 1997). Cytoskeletal reorganization and FA maturation by PR poly-dipeptides might be related to the increase in intracellular calcium. Further investigation is required to understand how the increase in intracellular calcium by PR poly-dipeptides is related to cytoskeletal reorganization and FA maturation.

PR poly-dipeptides, Blebbistatin, and Withaferin-A prevented cell detachment (**Figure 5**). Cell detachment depends on the adhesive strength and cell stiffness (Sen and Kumar, 2009; Tee et al., 2011). While Blebbistatin decreases cell stiffness by myosin II inhibition (Sen and Kumar, 2009), Withaferin-A increases cell stiffness by inhibition of IF organization (Bargagna-Mohan et al., 2007). Taken together, the alteration of IF organization by PR poly-dipeptides might prevent cell detachment through increasing cell stiffness.

Further, we found that PR poly-dipeptides induced upregulation of Thbs1, a mechanical stress response factor (**Figures 6A,B**). Thbs1 is induced by cyclic stretch and is related to mechanical stress response (Yamashiro et al., 2020; Laly et al., 2021), suggesting that PR poly-dipeptides induce a maladaptive mechanical stress response in a Thbs1-dependent manner. For the adaptive mechanical stress response, the cells first reduce the adhesive strength of FA, and then they reorient to a perpendicular position and reorganize F-actin stress fibers and FA (Nagayama et al., 2012; Livne et al., 2014). The maturation of FA by PR poly-dipeptides increases adhesive strength and might consequently suppress cell reorientation after cyclic stretch. The mechanical stress response regulates a wide range of cellular functions involved in cell death (King et al., 2011). Hence, further research is required to investigate the effect of maladaptive mechanical stress responses induced by PR poly-dipeptides on intracellular signaling.

PR poly-dipeptides reduced the number of cells after cyclic stretch (**Figures 6C,E**). PR poly-dipeptides cause cell death in a time-dependent manner (Kwon et al., 2014), resulting in the reduction of the number of cells. In addition, by FA maturation, cells treated with PR poly-dipeptides are expected to be exposed to higher continuous mechanical stress during cyclic stretch, resulting in the reduction of number of cells.

This study revealed that PR poly-dipeptides change the IF organization, leading to an alteration of the mechanical stress response. These results suggest that the pathogenesis by arginine-rich poly-dipeptides might be partially explained by the changes in cellular mechanical properties and mechanical stress responses.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

TS, RN, YY, and EM designed the research. TS, RN, KN, and YY performed research. TS, KN, and YY analyzed the data. TS, YY, and EM wrote the paper. SKi, HN, MM, MN, SKo, NI, TK, KN, and KS helped to analyze and interpret the data, and critically revise the manuscript. YY and EM conceptualized the study, developed the study design, supervised the authors throughout the study, and provided expertise in manuscript preparation. All authors read and approved the final manuscript.

FUNDING

This work was supported by the grants from AMED (JP20dm0307032 and JP21wm0425004 to EM, JP21ek0210158 to YY, JP20gm0810005 to KN), JSPS KAKENHI (JP20H03199 to EM, JP21H02677 and JP21K19881 to YY, JP19K17043 to TS, JP21K15032 to SOK, JP19K21306, JP20K16583 to HN, JP19K23976 to MN, JP19K08150 to SHK, JP19K07978 to TK, JP21H03804 to KN), Uehara Memorial Foundation to EM and SOK, Naito Foundation to EM, MSD Life Science Foundation to EM and YY, Tokyo Biochemical Research Foundation to SOK and by unrestricted funds provided to EM from Dr. Taichi Noda (KTX Corp., Aichi, Japan) and Dr. Yasuhiro Horii (Koseikai, Nara, Japan).

ACKNOWLEDGMENTS

The authors thank Dr. Akira Nagafuchi for his advice on experimental design and useful discussions. The authors thank Keren-Happuch E Fan Fen for her critical reading of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2022.750829/full#supplementary-material

REFERENCES

Abercrombie, M., and Dunn, G. A. (1975). Adhesions of Fibroblasts to Substratum during Contact Inhibition Observed by Interference Reflection Microscopy. *Exp. Cel Res.* 92 (1), 57–62. doi:10.1016/0014-4827(75)90636-9

Ash, P. E. A., Bieniek, K. F., Gendron, T. F., Caulfield, T., Lin, W.-L., Dejesus-Hernandez, M., et al. (2013). Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS. *Neuron* 77 (4), 639–646. doi:10.1016/j.neuron.2013.02.004

Bamburg, J. R., and Bloom, G. S. (2009). Cytoskeletal Pathologies of Alzheimer Disease. *Cell Motil. Cytoskeleton* 66 (8), 635–649. doi:10.1002/cm.20388

Baniukiewicz, P., Collier, S., and Bretschneider, T. (2018). QuimP: Analyzing Transmembrane Signalling in Highly Deformable Cells. *Bioinformatics* 34 (15), 2695–2697. doi:10.1093/bioinformatics/bty169

Bargagna-Mohan, P., Hamza, A., Kim, Y.-e., Khuan Ho, Y., Mor-Vaknin, N., Wendschlag, N., et al. (2007). The Tumor Inhibitor and Antiangiogenic Agent Withaferin A Targets the Intermediate Filament Protein Vimentin. *Chem. Biol.* 14 (6), 623–634. doi:10.1016/j.chembiol.2007.04.010

Baughman, J. M., Perocchi, F., Girgis, H. S., Plovanich, M., Belcher-Timme, C. A., Sancak, Y., et al. (2011). Integrative Genomics Identifies MCU as an Essential Component of the Mitochondrial Calcium Uniporter. *Nature* 476 (7360), 341–345. doi:10.1038/nature10234

Baum, M., Erdel, F., Wachsmuth, M., and Rippe, K. (2014). Retrieving the Intracellular Topology from Multi-Scale Protein Mobility Mapping in Living Cells. *Nat. Commun.* 5, 4494. doi:10.1038/ncomms5494

Bretscher, A., Edwards, K., and Fehon, R. G. (2002). ERM Proteins and merlin: Integrators at the Cell Cortex. *Nat. Rev. Mol. Cel Biol* 3, 586–599. doi:10.1038/nrm882

Chang, Y.-C., Nalbant, P., Birkenfeld, J., Chang, Z.-F., and Bokoch, G. M. (2008). GEF-H1 Couples Nocodazole-Induced Microtubule Disassembly to Cell Contractility via RhoA. *MBoC* 19 (5), 2147–2153. doi:10.1091/mbc.e07-12-1269

Choi, S. Y., Lopez-Gonzalez, R., Krishnan, G., Phillips, H. L., Li, A. N., Seeley, W. W., et al. (2019). C9ORF72-ALS/FTD-associated Poly(GR) Binds Atp5a1 and Compromises Mitochondrial Function *In Vivo*. *Nat. Neurosci.* 22 (6), 851–862. doi:10.1038/s41593-019-0397-0

Dafinca, R., Barbagallo, P., Farrimond, L., Candalija, A., Scaber, J., Ababneh, N. a. A., et al. (2020). Impairment of Mitochondrial Calcium Buffering Links Mutations in C9ORF72 and TARDBP in iPS-Derived Motor Neurons from Patients with ALS/FTD. *Stem Cel Rep.* 14 (5), 892–908. doi:10.1016/j.stemcr.2020.03.023

Dafinca, R., Scaber, J., Ababneh, N. a., Lalic, T., Weir, G., Christian, H., et al. (2016). C9orf72 Hexanucleotide Expansions Are Associated with Altered Endoplasmic Reticulum Calcium Homeostasis and Stress Granule Formation in Induced Pluripotent Stem Cell-Derived Neurons from Patients with Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Stem Cells* 34 (8), 2063–2078. doi:10.1002/stem.2388

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., et al. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron* 72 (2), 245–256. doi:10.1016/j.neuron.2011.09.011

Fletcher, D. A., and Mullins, R. D. (2010). Cell Mechanics and the Cytoskeleton. *Nature* 463 (7280), 485–492. doi:10.1038/nature08908

Fuhrmann, A., Li, J., Chien, S., and Engler, A. J. (2014). Cation Type Specific Cell Remodeling Regulates Attachment Strength. *PLoS One* 9 (7), e102424. doi:10.1371/journal.pone.0102424

Grin, B., Mahammad, S., Wedig, T., Cleland, M. M., Tsai, L., Herrmann, H., et al. (2012). Withaferin a Alters Intermediate Filament Organization, Cell Shape and Behavior. *PLoS One* 7 (6), e39065. doi:10.1371/journal.pone.0039065

Häbig, K., Gellhaar, S., Heim, B., Djuric, V., Giesert, F., Wurst, W., et al. (2013). LRRK2 Guides the Actin Cytoskeleton at Growth Cones Together with ARHGEF7 and Tropomyosin 4. *Biochim. Biophys. Acta* 1832, 2352–2367. doi:10.1016/j.bbadis.2013.09.009

Hao, Z., Liu, L., Tao, Z., Wang, R., Ren, H., Sun, H., et al. (2019). Motor Dysfunction and Neurodegeneration in a C9orf72 Mouse Line Expressing Poly-PR. *Nat. Commun.* 10 (1), 2906. doi:10.1038/s41467-019-10956-w

Hartzell, C. A., Jankowska, K. I., Burkhardt, J. K., and Lewis, R. S. (2016). Calcium Influx through CRAC Channels Controls Actin Organization and Dynamics at the Immune Synapse. *Elife* 5. doi:10.7554/eLife.14850

Hayes, L. R., Duan, L., Bowen, K., Kalab, P., and Rothstein, J. D. (2020). C9orf72 Arginine-Rich Dipeptide Repeat Proteins Disrupt Karyopherin-Mediated Nuclear Import. *Elife* 9. doi:10.7554/eLife.51685

Helfand, B. T., Mendez, M. G., Murthy, S. N. P., Shumaker, D. K., Grin, B., Mahammad, S., et al. (2011). Vimentin Organization Modulates the Formation of Lamellipodia. *MBoC* 22 (8), 1274–1289. doi:10.1091/mbc.e10-08-0699

Hu, J., Li, Y., Hao, Y., Zheng, T., Gupta, S. K., Parada, G. A., et al. (2019). High Stretchability, Strength, and Toughness of Living Cells Enabled by Hyperelastic Vimentin Intermediate Filaments. *Proc. Natl. Acad. Sci. U.S.A.* 116 (35), 17175–17180. doi:10.1073/pnas.1903890116

Huang, H., Sun, Z., Hill, M. A., and Meininger, G. A. (2018). A Calcium Mediated Mechanism Coordinating Vascular Smooth Muscle Cell Adhesion during KCl Activation. *Front. Physiol.* 9, 1810. doi:10.3389/fphys.2018.01810

Hurtley, S. M. (1998). Cell Biology of the Cytoskeleton. *Science* 279 (5350), 459. doi:10.1126/science.279.5350.459a

Hutten, S., Usluer, S., Bourgeois, B., Simonetti, F., Odeh, H. M., Fare, C. M., et al. (2020). Nuclear Import Receptors Directly Bind to Arginine-Rich Dipeptide Repeat Proteins and Suppress Their Pathological Interactions. *Cel Rep.* 33 (12), 108538. doi:10.1016/j.celrep.2020.108538

Jovičić, A., Mertens, J., Boeynaems, S., Bogaert, E., Chai, N., Yamada, S. B., et al. (2015). Modifiers of C9orf72 Dipeptide Repeat Toxicity Connect Nucleocytoplasmic Transport Defects to FTD/ALS. *Nat. Neurosci.* 18 (9), 1226–1229. doi:10.1038/nn.4085

Kim, J., Jang, J., Yang, C., Kim, E. J., Jung, H., and Kim, C. (2016). Vimentin Filament Controls Integrin α5β1-mediated Cell Adhesion by Binding to Integrin through its Ser38 Residue. *FEBS Lett.* 590 (20), 3517–3525. doi:10.1002/1873-3468.12430

King, J. S., Veltman, D. M., and Insall, R. H. (2011). The Induction of Autophagy by Mechanical Stress. *Autophagy* 7 (12), 1490–1499. doi:10.4161/auto.7.12.17924

Korobova, F., Gauvin, T. J., and Higgs, H. N. (2014). A Role for Myosin II in Mammalian Mitochondrial Fission. *Curr. Biol.* 24 (4), 409–414. doi:10.1016/j.cub.2013.12.032

Kovács, M., Tóth, J., Hetényi, C., Málnási-Csizmadia, A., and Sellers, J. R. (2004). Mechanism of Blebbistatin Inhibition of Myosin II. *J. Biol. Chem.* 279, 35557–35563. doi:10.1074/jbc.M405319200

Kwon, I., Xiang, S., Kato, M., Wu, L., Theodoropoulos, P., Wang, T., et al. (2014). Poly-dipeptides Encoded by the C9orf72 Repeats Bind Nucleoli, Impede RNA Biogenesis, and Kill Cells. *Science* 345 (6201), 1139–1145. doi:10.1126/science.1254917

Laly, A. C., Sliogeryte, K., Pundel, O. J., Ross, R., Keeling, M. C., Avisetti, D., et al. (2021). The Keratin Network of Intermediate Filaments Regulates Keratinocyte Rigidity Sensing and Nuclear Mechanotransduction. *Sci. Adv.* 7 (5). doi:10.1126/sciadv.abd6187

Lee, K.-H., Zhang, P., Kim, H. J., Mitrea, D. M., Sarkar, M., Freibaum, B. D., et al. (2016). C9orf72 Dipeptide Repeats Impair the Assembly, Dynamics, and Function of Membrane-Less Organelles. *Cell* 167 (3), 774–788. doi:10.1016/j.cell.2016.10.002

Levy Nogueira, M., Epelbaum, S., Steyaert, J. M., Dubois, B., and Schwartz, L. (2016). Mechanical Stress Models of Alzheimer's Disease Pathology. *Alzheimer's Demen.* 12, 324–333. doi:10.1016/j.jalz.2015.10.005

Li, J., Gao, W., Zhang, Y., Cheng, F., Eriksson, J. E., Etienne-Manneville, S., et al. (2019). Engagement of Vimentin Intermediate Filaments in Hypotonic Stress. *J. Cel Biochem* 120 (8), 13168–13176. doi:10.1002/jcb.28591

Li, S., Wu, Z., Li, Y., Tantray, I., De Stefani, D., Mattarei, A., et al. (2020). Altered MICOS Morphology and Mitochondrial Ion Homeostasis Contribute to Poly(GR) Toxicity Associated with C9-ALS/FTD. *Cel Rep.* 32 (5), 107989. doi:10.1016/j.celrep.2020.107989

Lin, Y., Mori, E., Kato, M., Xiang, S., Wu, L., Kwon, I., et al. (2016). Toxic PR Poly-Dipeptides Encoded by the C9orf72 Repeat Expansion Target LC Domain Polymers. *Cell* 167 (3), 789–802. doi:10.1016/j.cell.2016.10.003

Livne, A., Bouchbinder, E., and Geiger, B. (2014). Cell Reorientation under Cyclic Stretching. *Nat. Commun.* 5, 3938. doi:10.1038/ncomms4938

López-Colomé, A. M., Lee-Rivera, I., Benavides-Hidalgo, R., and López, E. (2017). Paxillin: a Crossroad in Pathological Cell Migration. *J. Hematol. Oncol.* 10 (1), 50. doi:10.1186/s13045-017-0418-y

Lu, Z., Li, H., Hou, C., Peng, Y., Long, J., and Liu, J. (2017). Endogenously Generated Amyloid-β Increases Stiffness in Human Neuroblastoma Cells. *Eur. Biophys. J.* 46, 415–424. doi:10.1007/s00249-016-1185-3

McLean, P. J., Kawamata, H., Ribich, S., and Hyman, B. T. (2000). Membrane Association and Protein Conformation of Alpha-Synuclein in Intact Neurons. Effect of Parkinson's Disease-Linked Mutations. *J. Biol. Chem.* 275, 8812–8816. doi:10.1074/jbc.275.12.8812

Mizielinska, S., Grönke, S., Niccoli, T., Ridler, C. E., Clayton, E. L., Devoy, A., et al. (2014). C9orf72 Repeat Expansions Cause Neurodegeneration in Drosophila through Arginine-Rich Proteins. *Science* 345 (6201), 1192–1194. doi:10.1126/science.1256800

Mould, A. P., Akiyama, S. K., and Humphries, M. J. (1995). Regulation of Integrin α5β1-Fibronectin Interactions by Divalent Cations. *J. Biol. Chem.* 270 (44), 26270–26277. doi:10.1074/jbc.270.44.26270

Nagayama, K., Kimura, Y., Makino, N., and Matsumoto, T. (2012). Strain Waveform Dependence of Stress Fiber Reorientation in Cyclically Stretched Osteoblastic Cells: Effects of Viscoelastic Compression of Stress Fibers. *Am. J. Physiol. Cel Physiol* 302, C1469–C1478. doi:10.1152/ajpcell.00155.2011

Nagayama, K., Uchida, K., and Sato, A. (2019). A Novel Micro-grooved Collagen Substrate for Inducing Vascular Smooth Muscle Differentiation through Cell Tissue Arrangement and Nucleus Remodeling. *J. Mech. Behav. Biomed. Mater.* 90, 295–305. doi:10.1016/j.jmbbm.2018.10.005

Nanaura, H., Kawamukai, H., Fujiwara, A., Uehara, T., Aiba, Y., Nakanishi, M., et al. (2021). C9orf72-derived Arginine-Rich Poly-Dipeptides Impede Phase Modifiers. *Nat. Commun.* 12 (1), 5301. doi:10.1038/s41467-021-25560-0

Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., et al. (2018). Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. *Neuron* 97 (6), 1268–e6. doi:10.1016/j.neuron.2018.02.027

O'Brien, E. T., Salmon, E. D., and Erickson, H. P. (1997). How Calcium Causes Microtubule Depolymerization. *Cell Motil Cytoskeleton* 36, 125–135. doi:10.1002/(SICI)1097-0169(1997)36:2<125::AID-CM3>3.0.CO;2-8

Poullet, P., Gautreau, A., Kadaré, G., Girault, J. A., Louvard, D., and Arpin, M. (2001). Ezrin Interacts with Focal Adhesion Kinase and Induces its Activation Independently of Cell-Matrix Adhesion. *J. Biol. Chem.* 276, 37686–37691. doi:10.1074/jbc.M106175200

Price, L. S., Langeslag, M., ten Klooster, J. P., Hordijk, P. L., Jalink, K., and Collard, J. G. (2003). Calcium Signaling Regulates Translocation and Activation of Rac. *J. Biol. Chem.* 278, 39413–39421. doi:10.1074/jbc.M302083200

Sen, S., and Kumar, S. (2009). Cell-Matrix De-adhesion Dynamics Reflect Contractile Mechanics. *Cel. Mol. Bioeng.* 2 (2), 218–230. doi:10.1007/s12195-009-0057-7

Smith, B. N., Ticozzi, N., Fallini, C., Gkazi, A. S., Topp, S., Kenna, K. P., et al. (2014). Exome-wide Rare Variant Analysis Identifies TUBA4A Mutations Associated with Familial ALS. *Neuron* 84 (2), 324–331. doi:10.1016/j.neuron.2014.09.027

Sumi, T., Matsumoto, K., Takai, Y., and Nakamura, T. (1999). Cofilin Phosphorylation and Actin Cytoskeletal Dynamics Regulated by Rho- and Cdc42-Activated LIM-Kinase 2. *J. Cel Biol* 147 (7), 1519–1532. doi:10.1083/jcb.147.7.1519

Sweers, K., van der Werf, K., Bennink, M., and Subramaniam, V. (2011). Nanomechanical Properties of α-synuclein Amyloid Fibrils: a Comparative Study by Nanoindentation, Harmonic Force Microscopy, and Peakforce QNM. *Nanoscale Res. Lett.* 6 (1), 270. doi:10.1186/1556-276x-6-270

Tee, S.-Y., Fu, J., Chen, C. S., and Janmey, P. A. (2011). Cell Shape and Substrate Rigidity Both Regulate Cell Stiffness. *Biophysical J.* 100 (5), L25–L27. doi:10.1016/j.bpj.2010.12.3744

Wu, C.-H., Fallini, C., Ticozzi, N., Keagle, P. J., Sapp, P. C., Piotrowska, K., et al. (2012). Mutations in the Profilin 1 Gene Cause Familial Amyotrophic Lateral Sclerosis. *Nature* 488 (7412), 499–503. doi:10.1038/nature11280

Yamashiro, Y., Papke, C. L., Kim, J., Ringuette, L. J., Zhang, Q. J., Liu, Z. P., et al. (2015). Abnormal Mechanosensing and Cofilin Activation Promote the Progression of Ascending Aortic Aneurysms in Mice. *Sci. Signal.* 8 (399), ra105. doi:10.1126/scisignal.aab3141

Yamashiro, Y., Thang, B. Q., Ramirez, K., Shin, S. J., Kohata, T., Ohata, S., et al. (2020). Matrix Mechanotransduction Mediated by Thrombospondin-1/integrin/YAP in the Vascular Remodeling. *Proc. Natl. Acad. Sci. U.S.A.* 117 (18), 9896–9905. doi:10.1073/pnas.1919702117

Yamashiro, Y., Thang, B. Q., Shin, S. J., Lino, C. A., Nakamura, T., Kim, J., et al. (2018). Role of Thrombospondin-1 in Mechanotransduction and Development of Thoracic Aortic Aneurysm in Mouse and Humans. *Circ. Res.* 123 (6), 660–672. doi:10.1161/circresaha.118.313105

Zhang, S., Cooper-Knock, J., Weimer, A. K., Shi, M., Moll, T., Marshall, J. N. G., et al. (2022). , S0896-6273, 01036–01039. doi:10.1016/j.neuron.2021.12.

019Genome-wide Identification of the Genetic Basis of Amyotrophic Lateral Sclerosis*Neuron*

Zhou, X., Lin, Y., Kato, M., Mori, E., Liszczak, G., Sutherland, L., et al. (2021). Transiently Structured Head Domains Control Intermediate Filament Assembly. *Proc. Natl. Acad. Sci. U S A.* 118 (8). doi:10.1073/pnas.2022121118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Shiota, Nagata, Kikuchi, Nanaura, Matsubayashi, Nakanishi, Kobashigawa, Isozumi, Kiriyama, Nagayama, Sugie, Yamashiro and Mori. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Zinc finger domains bind low-complexity domain polymers

Naohiko Iguchi[1,17], Noriyoshi Isozumi[2,17], Yoshikazu Hattori[3], Tomohiro Imamura[4], Masatomo So[2], Hitoki Nanaura[1], Takao Kiriyama[1], Nobuyuki Eura[1], Minako Yamaoka[1], Mari Nakanishi[5], Masashi Mori[4], Shinya Ohki[6], Hiroyuki Kumeta[7], Hironori Koga[8], Mai Watabe[9], Takuya Mabuchi[9,10], Shingo Kanemura[9,11], Masaki Okumura[9], Takuya Yoshizawa[12], Ichiro Ota[13], Naoki Suzuki[14], Masashi Aoki[14], Yoshito Yamashiro[15], Tomohide Saio[3,9,16], Kazuma Sugie[1]*, Eiichiro Mori[2,5]*

[1]Department of Neurology, Nara Medical University, Kashihara, Nara, Japan.
[2]Department of Future Basic Medicine, Nara Medical University, Kashihara, Nara, Japan.
[3]Division of Molecular Life Science, Institute of Advanced Medical Sciences, Tokushima University, Tokushima, Japan.
[4]Research Institute for Bioresources and Biotechnology, Ishikawa Prefectural University, Nonoichi, Ishikawa, Japan.
[5]V-iCliniX Laboratory, Nara Medical University, Kashihara, Nara, Japan.
[6]Center for Nano Materials and Technology (CNMT), Japan Advanced Institute of Science and Technology (JAIST), Nomi, Ishikawa, Japan.
[7]Faculty of Advanced Life Science, Hokkaido University, Sapporo, Hokkaido, Japan.
[8]Department of Bioproduction Science, Ishikawa Prefectural University, Nonoichi, Ishikawa, Japan.
[9]Frontier Research Institute for Interdisciplinary Sciences, Tohoku University, Sendai, Miyagi, Japan.
[10]Institute of Fluid Science, Tohoku University, Sendai, Miyagi, Japan.
[11]School of Science, Kwansei Gakuin University, Sanda, Hyogo, Japan.
[12]College of Life Sciences, Ritsumeikan University, Kusatsu, Shiga, Japan.
[13]Department of Otolaryngology-Head and Neck Surgery, Kindai University Nara Hospital, Ikoma, Nara, Japan.
[14]Department of Neurology, Tohoku University Graduate School of Medicine, Aoba-ku, Sendai, Japan.
[15]Department of Advanced Medical Technologies, National Cerebral and Cardiovascular Center Research Institute, Suita, Osaka, Japan.
[16]Institute for Integrated Radiation and Nuclear Science, Kyoto University, Sennan-gun, Osaka, Japan.

[17]These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

*Corresponding authors.
Email: ksugie@naramed-u.ac.jp (KS); emori@naramed-u.ac.jp (EM).

Abstract

Self-association of low-complexity protein sequences (LC domains) is important for polymer formation. Several molecular chaperones are involved in the regulation of LC domain polymer formation. However, the mechanisms underlying cell recognition of LC domain polymers remain unclear. Here we show that zinc finger domains (ZnFs) bind LC domains of RNA-binding proteins in a cross-β polymer-dependent manner. ZnFs bound to LC domain hydrogels and suppressed LC domain polymer formation. Moreover, ZnFs preferentially recognize LC domains in the polymeric state. These findings suggest that ZnFs act as physiological regulators of LC domain polymer formation.

Main

Low-complexity protein sequences (LC domains) are disordered regions of proteins with a limited number of amino acids that drive phase transitions into gel-like or liquid-like states[1-3], which in turn form membrane-less organelles and play a key role within living cells[4-6]. Dysregulating phase separation of RNA binding proteins (RBPs) with LC domains, including heterogeneous nuclear ribonucleoprotein A2 (hnRNPA2)[7], fused in sarcoma (FUS)[8], and TAR DNA-binding protein of 43 kDa (TDP43)[9], leads to the formation of pathogenic fibrils and may cause amyotrophic lateral sclerosis (ALS) and other neurodegenerative diseases[10,11].

There are several known influences on LC domains. For example, post-translational modifications control chemical properties of LC domains[12]. Karyopherins[13-16] and other molecular chaperones[17-19] have been found to regulate the phase separation of LC domains. A recent study also demonstrated how de novo designed proteins have the potential to inhibit the polymer formation of amyloid fibrils[20]. However, However, it remains unclear how the length of LC domain polymers are determined and subsequently kept from becoming too long.

In this study, we show zinc finger domains (ZnFs) as candidate molecules in modifying ALS pathogenesis from three bioinformatic approaches. We investigated the role of ZnFs in phase separation, with a focus on the interaction between ZnFs and LC domains. The approaches we utilized in this study include the following: i) an assessment of interactions by hydrogel binding assay; ii) monitoring the polymer formation; and iii)

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a detailed solution nuclear magnetic resonance (NMR) analysis. We describe here how ZnFs interact with LC domains, providing additional mechanistic insights on ZnFs.

Genes encoding Zn-binding and ZnF are enriched in ALS with FUS-mutations.

Cell-fate transition is involved in cancer, cardiovascular diseases, and neurodegenerative diseases[21,22]. To examine the molecules commonly altered, we analyzed two datasets: RNA-seq data of motor neurons (MNs) derived from *FUS* mutant human-induced pluripotent stem cells (hiPSCs)[23] and ChIP-seq (H3K27ac) data of cell-fate transition induced by hypoxia mimic condition[24]. The genes common to both datasets included Krüppel-like factor 4 (KLF4) and other transcription factors (Fig. 1a). KLF4 is a transcription factor with ZnFs that regulates diverse cellular processes and one of the stem cell factors required for the induction of iPSCs[25]. Gene ontology (GO) enrichment analysis of two datasets revealed significant differences in the expression of genes related to transcription factors and metal binding as with genes related to cell adhesion, cytoskeleton, and cell migration. In addition, protein domain enrichment analysis in MNs derived from *FUS* mutant hiPSCs showed differences in the expression of genes with ZnFs (Extended Data Fig. 1). The genes encoding zinc binding and ZnF were enriched compared to the human proteome (Fig. 1b). These results showed that genes encoding Zn-binding and ZnF are enriched in ALS with FUS-mutations.

ZnFs of transcription factors interact with LC domains of RBPs.

KLF4 forms round droplets in the nucleus and has been proposed to recruit other transcription factors[26]. To test if KLF4 interacts with nuclear proteins, we performed the hydrogel binding assay. In this assay, we employed green fluorescent protein (GFP) tagged KLF4 fragments (Fig. 1c). Hydrogels of mCherry fusion LC domain of hnRNPA2 (mCh:A2-LC) were incubated with GFP-fused KLF4 fragments. Only GFP-fused ZnFs of KLF4 (GFP:KLF4-ZnF1-3) bound to the mCh:A2-LC hydrogel; other KLF4 domains did not. In addition, the DNA binding domain of other representative stem cell transcription factors, including octamer-binding transcription factor 4 (OCT4), SRY-box transcription factor 2 (SOX2) and myc proto-oncogene protein (MYC), did not bind to A2-LC (Extended Data Fig. 2a).

Zinc finger E-box binding homeobox 1 (ZEB1) and ZEB2, zinc finger protein 512B (ZNF512B) are major transcription factors and related to neurodegenerative diseases[28-29]. GFP fusion ZnFs of these transcription factors showed binding to hydrogels of mCh:A2-LC in a similar manner as seen in GFP:KLF4-ZnF1-3 (Fig. 1d). In addition, we observed that ZnFs bind to other LC domains of FUS (FUS-LC) and TDP43 (TDP43-

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

LC) as with A2-LC (Fig. 1e). Among domains of ZEB2, ZnFs of ZEB2 bound to mCh:A2-LC hydrogels (Fig. 1f). The larger the number of ZnFs was, the stronger ZnFs bound to the A2-LC hydrogels (Extended Data Fig. 2b and 2c). ZnFs of transcription factors other than KLF4 also play an important role in binding to LC domains, and the number of ZnFs affected the recognition of LC domains.

ZnF binds to various LC domains in a common binding manner.

To investigate the interaction between ZnF and LC domains, we performed solution NMR experiments. ZnF8 of ZEB2 was used as a representative for the NMR experiments. The ^{1}H-^{15}N NMR spectrum of ^{15}N-labeled ZnF8 showed negligible perturbations by the addition of mCherry, indicating that ZnF8 does not interact with mCherry (Fig. 2a). In contrast, the addition of mCh:A2-LC reduced the signal intensities of ZnF8, indicating that ZnF8 interacts with the LC domain of hnRNPA2 (Fig. 2a). We then assigned NMR signals of ZnF8 to identify residues of ZnF8 that interact with the LC domain of hnRNPA2 (Extended Data Fig. 3a). The addition of mCh:A2-LC eliminated ZnF8 signals, but did not change ZnF8 chemical shift values. For this reason, we used intensity ratios of ZnF8 signals to identify the binding sites of ZnF8 to mCh:A2-LC (Fig. 2b). With the addition of mCh:A2-LC, signal intensities were most significantly reduced in D1058, K1059, G1061, F1064, Q1072, H1073, and N1075.

The ZnF8 structure predicted by AlphaFold 2 has one antiparallel β-sheet and one α-helix (Fig. 2d). In this ZnF8 structure, the side chain of H1076 is not oriented to allow binding to a zinc ion. In fact, the chemical shift of H1076 side chain indicated a non-binding state with a zinc ion (Extended Data Fig. 3b). Thus, ZnF8 is likely bound to a zinc ion by three residues, C1057, C1060, and H1073, as predicted by AlphaFold 2. The interaction sites of ZnF8 with mCh:A2-LC were mapped to a loop connecting two β-strands and part of an α-helix on the predicted structure (Fig. 2d). Since hydrogel binding assays in Fig. 1e showed that ZnF can bind to various LC domains, we also analyzed the interaction of ZnF8 with LC domains of FUS and TDP43. The results showed that ZnF8 interacted with mCh:FUS-LC and mCh:TDP43-LC in a similar way with mCh:A2-LC (Fig.2b), suggesting that ZnF uses a common region for binding to various LC domains.

ZnF is known as a typical DNA-binding domain[30]. To investigate the difference between the binding of ZnF to LC domains and to DNA, the binding of ZnF8 to DNA was analyzed using NMR. The ^{1}H-^{15}N NMR spectrum of ZnF8 with DNA was clearly different from the ZnF8 spectrum with the LC domains (Fig. 2a). We analyzed the binding of ZnF8 to DNA using chemical shift change and revealed that Y1055, F1064, H1066, and M1074 of ZnF8 were strongly involved in DNA binding (Fig. 2c). These residues

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were mapped to the loop region between the β-sheet and the α-helix and part of the α-helix on the ZnF8 structure (Fig. 2d). Together, these results demonstrate that the binding region of ZnF8 to DNA is different from that to the LC domains.

ZnFs suppressed LC domain polymer formation.

We further investigated the effect of ZnF to LC domain polymer formation. First, we performed uptake experiments of GFP-fused ZnF into FUS droplets (Fig. 3a). GFP-fused ZnF8 was incorporated into the droplets more than GFP alone. We then measured the refractive index (RI) inside FUS droplets to confirm the effect of ZnF (Fig. 3b). The size of the FUS droplets was almost the same for all four conditions: no peptide, ZnF8, ZnF7-8, and ZnF6-8 (Extended Data Fig. 4). ZnF6-8 decreased RI of FUS droplets, but there was no significant difference in the RI of FUS droplets with ZnF8 and ZnF7-8 (Fig. 3c). This result suggests that the addition of ZnF6-8 makes the inside of the FUS droplets sparse. Further, we monitored the polymer formation of A2-LC using the thioflavin T (ThT), and ZnF6-8 suppressed A2-LC polymer formation (Fig. 3d).

ZnF preferentially recognizes LC polymers over LC monomers.

Our MD simulation predicted a binding model of ZnF and the LC polymers (Extended Data Fig. 5). To clarify the mechanism by which ZnF recognizes the LC domain, we tried to demonstrate whether the polymeric state of LC domain is related to ZnF binding. The F/Y to S mutation in A2-LC has been reported to suppress polymer formation[31]. In addition, we prepared a A2-LC mutant (A2-LC_MM) in which multiple F/Y residues were mutated to S (Y257S, Y264S, Y278S, F291S, F309S, and Y319S). The interaction of these mutants with ZnF8 was monitored by NMR spectra (Fig. 4a). With the addition of mCh:A2-LC and mCh:A2-LC_MM, K1059 and G1061 showed significant changes in ZnF8 signals, while Y1070 showed little change (Fig. 4b). Overall, the intensity ratio of ZnF8 signals was smaller in the order of mCh:A2-LC, mCh:A2-LC_F291S, and mCh:A2-LC_MM (Fig. 4b). These results suggest that ZnF preferentially recognize LC domains in the polymeric state. We offer a model that ZnFs act as physiological regulators of LC domain polymer formation (Fig. 4c)

Discussion

ZnF is a widespread domain structure. ZnFs with two cysteins and two histidines (C2H2-type) are mostly found in transcription factors and recognize DNA[32]. Some C2H2-type ZnFs interact with other proteins and play an important role in their communication with other proteins that are attracted to the multimolecule transcriptional complex formed on

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DNA[33]. Through our experiments, we found that ZnFs preferentially recognize LC domains in the polymeric state, and the interaction sites of ZnF with LC domains were mapped to a loop connecting two β-strands. De novo designed proteins binding amyloid fibrils contain one α-helix and two or three β-strands, similar to the structure of C2H2-type ZnF, and the primary interaction between these designed proteins and amyloid fibrils is the stacking of a β-strand of the designed proteins onto the β-strand of the amyloid fibrils[20]. ZnF structure with one antiparallel β-sheet and one α-helix surrounding zinc ion may be important in the identification of LC domain polymers.

We found that FUS droplets absorbed ZnF and began to disappear, and that the polymer formation of LC domains was suppressed in the presence of ZnFs in ThT assay. Molecular chaperones, including karyopherin-β2[13,15], cyclophilin A[31], and heat shock proteins[17-19] interact with LC domains or other segments of RBPs and regulate phase separation. ZnFs are new factors in phase separation with the specific function recognizing LC domain polymers and regulating phase separation suppressively.

Transcription factors with ZnFs, such as KLF4, ZEB1 and ZEB2, not only regulate the specific expression of human genome in the brain[34] but also are involved in neural development and neurodegenerative diseases[35-38]. Proteins with ZnFs or zinc binding lesions have been suggested to be involved in phase separation[39-41]. Further elucidation of the function of ZnFs, including interaction with LC domains, may lead to an understanding of the pathology of neurodegenerative diseases and the invention of therapeutic methods.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Bioinformatic analysis

RNA-seq of MNs derived from *FUS*$^{H517D/H517D}$ mutant hiPSCs was used for bioinformatic analyses[23]. Among the differential expressed genes (DEGs) in MNs (SDs and Axon), those with a statistically significant difference ($P < 0.1$) were selected. Our previous ChIP-seq (H3K27ac) data (GSE189071) of cobalt chroride ($CoCl_2$)-treated (hypoxia mimic) human aortic endothelial cells (HAEC) was also used for bioinformatic analyses[24]. RNAs that did not express genes such as antisense RNAs, microRNAs, non-cording RNAs were removed from all data. For gene ontology (GO) analysis and domain enrichment analysis (InterPro), the DAVID server was used. Due to the limited number of input genes on the DAVID server, the ChIP-seq data used the top 6,000 genes in the score. For zinc binding and ZnF enrichment analysis, proteins with zinc binding or ZnF domain were annotated using the UniProt database (keyword: KW-0862 and KW-0863). The human proteome dataset for gene counts was obtained from UniProt sever (Proteome ID: UP000005640; release 2022_04).

Expression and purification of recombinant proteins

Expression plasmids for His-mCherry fusion LC domains of hnRNPA2 (residues 181–341, mCh:hnRNPA2-LC) and FUS (residues 2–214, mCh:FUS-LC) were obtained from the Steven L. McKnight Laboratory. Expression plasmids for other His-mCherry or His-GFP fusion proteins were constructed using in-Fusion HD Cloning Kit (Takara Bio, Inc.). All mCherry/GFP fusion proteins were expressed in *E. coli* BL21(DE3) cells with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). MCh:hnRNPA2-LC (residues 181–341), hnRNPA2-LC mutants (mCh:hnRNPA2-LC_F291S and mCh:hnRNPA2-LC_MM), and mCh:FUS-LC (residues 2–214) were expressed at 20°C overnight, and mCh:TDP43-LC (residues 262–414) was expressed at 16°C overnight. All mCherry fusion proteins were purified by Ni-NTA agarose (FUJIFILM Wako Pure Chemical Corporation) according to steps described in our previous study[42]. GFP:KLF4-AD (residues 1-157), GFP:KLF4-RD (residues 158-385), GFP:KLF4-NLS (residues 430-512), GFP:KLF4-ZnF1-3 (residues 430-512), GFP:SOX2-HMG (residues 41-109), GFP:MYC-bHLH-LZ (residues 368-448), GFP:OCT4-POU (residues 138-289), GFP:ZEB1-ZnF5-7 (residues 904-981), GFP:ZEB2-ZnF6-8 (residues 999-1076), GFP:ZNF512B-ZnF345 (residues 540-653), GFP:OCT4-POU (residues 138-289), GFP:ZEB2-G-rich1 (residues 37-196), GFP:ZEB2-ZnF1-4 (residues 211-334), GFP:ZEB2-SMAD (residues 335-580), GFP:ZEB2-ZnF5 (residues 581-605), GFP:ZEB2-homeobox (residues 606-703), GFP:ZEB2-LC (residues 704-998), GFP:ZEB2-G-rich2 (residues 1084-1215),

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GFP:ZEB2-ZnF6-7 (residues 999-1049), GFP:ZEB2-ZnF7-8 (residues 1027-1076), GFP:ZEB2-ZnF6 (residues 999-1021), GFP:ZEB2-ZnF7 (residues 1027-1049), and GFP:ZEB2-ZnF8 (residues 1055-1076) were expressed at 20°C overnight and purified using the same method as mCherry fusion proteins. Briefly, cells were lysed in a buffer containing 25mM Tris-HCl pH 7.5, 200mM NaCl, 10mM β-ME, and Protease Inhibitor Cocktail Tablets. GFP fusion proteins were purified using Ni-NTA agarose with a wash buffer (25mM Tris-HCl pH 7.5, 200mM NaCl, 10mM β-ME, and 20mM imidazole) and elution buffer (25mM Tris-HCl pH 7.5, 200mM NaCl, 10mM β-ME, and 300 mM imidazole). Where necessary, 2 M Urea was added to the buffers.

For His-hnRNPA2, the expression plasmid was constructed using in-Fusion HD Cloning Kit with pHis-parallel1 vector as the template. His-hnRNPA2-LC was expressed in *E. coli* BL21(DE3) cells with 0.5 mM IPTG at 23°C overnight. Cells were lysed in buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 10mM β-ME, Protease Inhibitor) and centrifuged at 12,000×g for 30 minutes. After centrifugation, the pellet was collected and dissolved in solubilization buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 10mM β-ME, and 6 M guanidine) followed by Ni-NTA purification.

Hydrogel binding assays

Hydrogel droplets of mCh:hnRNPA2-LC, mCh:FUS-LC and mCh:TDP43-LC were prepared according to procedures reported in a previous study[43]. For hydrogel binding assays, purified GFP fusion proteins were diluted to 1 μM in the buffer (20mM Tris-HCl pH 7.5, 150mM NaCl, 20mM β-ME, 0.1mM phenylmethylsulfonyl fluoride (PMSF), and 0.5mM EDTA) and pipetted onto a hydrogel dish. Horizontal sections of the droplets were scanned with excitation wavelengths on a confocal microscope (FLUOVIEW FV3000, OLYMPUS) after overnight incubation. We used the profile plot mode in ImageJ and measured relative intensity of GFP signals across the hydrogel droplets in triplicate. The values are presented as mean ± SD. Graphpad Prism version 7 was used for statistical analysis.

ZnF peptides

The GST:ZnF expression plasmid was transformed into *E. coli* BL21(DE3) cells. *E. coli* cells with the plasmid were cultured in LB/Kan medium at 37°C, and protein expression was introduced by the addition of 0.5 mM IPTG. After 3 h of incubation at 37°C, cells were collected and suspended in PBS containing 0.1 mM EDTA and protease inhibitor. For ^{15}N-labeling or ^{13}C,^{15}N-double labeling, *E. coli* cells were respectively cultured in M9/Kan medium with $^{15}NH_4Cl$, or $^{15}NH_4Cl$ and ^{13}C-glucose at 37°C. The ^{15}N-labeled

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

protein was expressed by the addition of 0.5 mM IPTG followed by a 4 h incubation at 37°C. Cells were sonicated and centrifugated at 12,000×g for 30 min at 4°C. GST-fused proteins were purified using Glutathione Sepharose 4B beads (Cytiva) and GST-tag was cleaved with PreScission Protease (Cytiva) at 4°C for 16 h. ZnF peptides were purified using gel filtration (Superdex 30 pg16/600, Cytiva) with a buffer (20 mM Tris-HCl pH 8.0, 200 mM NaCl, and 1 mM DTT), and desalted using Sep-Pak C18 (Waters). The desalted peptides containing acetonitrile were dried under vacuum conditions.

ZnF-binding DNA

Two complementary single strand oligonucleotides (5'-GTAATCTGGGCCACCTGCCTGGGAGGA-3' and 5'-TCCTCCCAGGCAGGTGGCCCAGATTAC-3') corresponding E2 box[44] were purchased from Sigma-Aldrich. An equal molar ratio of single strand oligonucleotides was mixed and incubated at 95°C for 5 min. The double-stranded DNA obtained by this annealing was used for NMR experiments.

NMR experiments

For titration experiments, the ^{15}N-labeled sample at 0.1 mM concentration were dissolved in a solution containing 20 mM Tris-HCl pH 7.5, 200 mM NaCl, 20 mM β-ME, 0.1 mM PMSF, 0.12 mM $ZnCl_2$, and 5% ^2H_2O. NMR spectra were measured on the Bruker AVANCE 500 MHz spectrometer. The sample temperature was set at 15°C. ^{15}N-SOFAST HMQC pulse sequence was employed to obtain backbone ^{1}H-^{15}N correlation spectrum. The following samples were used for the titration experiments; mCherry, mCherry:A2-LC, mCherry:FUS-LC, mCherry:TDP43-LC, and ZnF-binding DNA. The proteins and DNA were prepared at 1 mM and titrated into ZnF8 at the molar ratios of 0.2, 0.5, 1, and 2.

For backbone resonance assignment experiments, the ^{13}C,^{15}N-labeled sample at 0.9 mM concentration were dissolved in a solution containing 20 mM Tris-HCl pH 7.5, 20 mM β-ME, 0.1 mM PMSF, 1 mM $ZnCl_2$, and 5% ^2H_2O. NMR spectra were measured on Bruker AVANCE III 600 MHz equipped with a TBI probe and AVANCE NEO 800 MHz spectrometers equipped with a CPTCI probe. The sample temperature was set at 10°C. Backbone resonance assignment was carried out using the following triple resonance spectra: HNCO, HNCA, HN(CO)CA, HNCACB, and CBCA(CO)NH. The assignment was transferred to titration experiment conditions by tracing the change of sample concentration. Data were processed using NMRPipe[45], and the spectra were analyzed using NMRFAM-Sparky[46] and MagRO-NMRView[47]. The standard deviation

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(*SD*) of intensity ratio was estimated from the signal to noise ratio (*SNR*) using the following equation:

$$SD = \sqrt{SNR_{ref}^{-2} + SNR_{tit}^{-2}}$$

where SNR_{ref}^{-2} and SNR_{tit}^{-2} are the signal to noise ratios of the peak on the reference spectrum and titrated spectrum, respectively. The chemical shift perturbation (*CSP*) of the backbone amide group was averaged using the following equation:

$$CSP = \sqrt{\Delta\delta_{\mathrm{H}}^2 + (\frac{\Delta\delta_{\mathrm{N}}}{5})^2}$$

where $\Delta\delta_{\mathrm{H}}$ and $\Delta\delta_{\mathrm{N}}$ are the chemical shift changes of proton and nitrogen, respectively.

Refractive index measurement inside FUS droplet using the holotomography microscope

The LLPS-droplet formation was initiated by mixing 10 μM of MBP:FUS with 8% polyethylene glycol 8000 in a buffer containing 20 mM Tris-HCl, pH 7.5, 200 mM NaCl, 10% glycerol, and 20 mM β-ME. After 1 min, 10 μM of ZnF8, ZnF7-8, or ZnF6-8 was added to the mixture and incubated for 5 min. For the GFP-fluorescence imaging, 10 μM of GFP:ZnF8, GFP:ZnF7-8, or GFP:ZnF6-8 was added to the mixture and incubated for 5 min. The droplets were observed using a holotomography microscope with laser-induced fluorescence system (Tomocube, Inc.)[48]. The refractive index (RI) and fluorescence image show the XY cross section of the droplet center. Average RI inside the droplet and their radius were calculated by enclosing the RI image in a circle using TomoStudio version 3.2.8 software (Tomocube, Inc.) and plotted by the Igor Pro version 6.36 software (WaveMetrics). The statistical significance of differences was examined by a one-way analysis of variance with Tukey's honest significant difference (HSD) post hoc testing. All statistical tests were performed using KaleidaGraph version 4.5.1 software (Synergy Software) at a significance level of $\alpha = 0.05$.

ThT fluorescence assays

His-tagged hnRNPA2-LC was dissolved in stock buffer containing 20 mM sodium phosphate, pH 7.4, 500 mM NaCl, and 8 M Urea. The sample was diluted to 20 μM with

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

an assay buffer containing 20 mM Tris-HCl, pH 7.5, 200 mM NaCl, 20 mM β-ME, 0.1 mM PMSF, 1 mM $ZnCl_2$, and 20 μM ThT. In the condition with ZnF, a sample containing 20 μM ZnF6-8 was prepared. ThT fluorescence signal was monitored with a Varioskan LUX plate reader (Thermo Fisher Scientific) using excitation and emission wavelengths of 440 and 484 nm, respectively. Fluorescence data were analyzed using Skanlt software (Thermo Fisher Scientific).

MD simulation

Molecular docking simulations were performed using the AutoDock Vina 1.1.2 software[49] on the ZnF8 and hnRNPA2 complex. The ZnF8 domain between residues Glu1052 and Ser1079 was extracted from the molecular structures of the ZEB2 estimated by the AlphaFold 2[50]. The cryo-electron microscopy structure of the low-complexity (LC) domain of hnRNPA2 (PDB ID: 6WQK)[51] was used. The size of the grid boxes was chosen based on its ability to occupy the whole ZnF8 and hnRNPA2 complex. The molecular dynamics (MD) simulations of ZnF8 bound to the LC domain of hnRNPA2 were then performed with the CHARMM36m force field[52] using LAMMPS software[53]. The selected docked poses provided in the docking simulations were used as starting structures, placed in a periodic simulation box and solvated with water. The salt concentration was set to 150 mM NaCl, and corresponding numbers of Na+ and Cl− ions were added. After the steepest-descent energy minimization, production runs were performed for 100 ns in the NPT ensemble at 300 K and 1 atm, and trajectory data were collected every 10 ps. During these MD simulations, C' atoms of terminal residues and Cα atoms at the beginning and end of β sheet were fixed to maintain the secondary structures of the LC domain of hnRNPA2. The particle−particle particle−mesh (PPPM) method[54] was used to calculate long-range electrostatic interactions. Equations of motion were integrated using the Verlet algorithm[55] with a time step of 2 fs, along with the SHAKE algorithm[56] to constrain the lengths of bonds to hydrogen. The binding free energy between ZnF8 and hnRNPA2 was calculated by the MM-GBSA method[57] using the MD simulation trajectories over the last 50 ns. A model that has the lowest binding free energy among the estimated binding sites is used as the most plausible binding model.

Acknowledgements

This work was supported by grants from AMED [JP23wm0425004 to E.M.; JP23ek0109642 to T.S.; JP23ek0109558 to T.Y. and H.N.; JP21ek0210158 to Y.Y.], JSPS KAKENHI [JP20H03199 to E.M.; JP19K17044 to N. Iguchi; JP20KK0156, JP22K18361,

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and JP22H02560 to T.S.; JP23K05657 to Y.H.; JP22K15734 to H.N.; JP21H02677 to Y.Y.; JP22K15691 to M.N.; JP21H05095, JP22H02205, JP23KK0105 to M.O.], MEXT Grant-in-Aid for Transformative Research Areas (B) [JP21H05094 and JP21H05093 to T.S.], JST FOREST Program [JPMJFR204W to T.S.; JPMJFR201F to M.O.]; Takeda Science Foundation to E.M. and T.S.; Kato Memorial Bioscience Foundation to H.N.; Uehara Memorial Foundation to M.O. and S.K. This work was partially supported by Joint Usage and Joint Research Programs, Institute of Advanced Medical Sciences, Tokushima University. This work was supported in part by "Advanced Research Infrastructure for Materials and Nanotechnology in Japan (ARIM)" of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) under Grant Number JPMXP12-22JI0044. We thank Shunsuke Tomita (AIST), Shinsuke Niwa (Tohoku University), Yoichi Shinkai (AIST), Wataru Iwasaki (Tokyo University), and Noriyuki Kodera (Kanazawa University) for technical support. NMR experiments were performed at the Hokkaido University Advanced NMR Facility, a member of the NMR Platform.

Author contributions

N. Iguchi, N. Isozumi, K. Sugie, and E.M. designed the research. N. Iguchi, N. Isozumi, Y.H., T.I., M.S., H.N., M.M., H. Kumeta., H. Koga, M.W., T.M., S.K., M.O., T.Y., Y.Y. and T. Saio performed research. N. Iguchi, N. Isozumi, Y.H., T.M., M.S., M.W., S.K., M.O., Y.Y., and T. Saio analyzed the data. N. Iguchi, N. Isozumi, Y.Y., T. Saio, and E.M. wrote the paper. T.K., N.E., M.Y., M.N., S.O., I.O., N.S., M.A., and K.S. helped to analyze and interpret the data, and critically revise the manuscript. K. Sugie, and E.M. conceptualized the study, developed the study design, supervised the authors throughout the study, and provided expertise in manuscript preparation. All authors read and approved the final manuscript.

Competing interests

EM is a Founder CEO of molmir, Inc..

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Kato M, et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell.* **149**, 753-767 (2012)

2. Forman-Kay JD, & Mittag T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure.* **21**, 1492-1499. (2013)

3. Nott TJ, et al. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell.* **57**, 936-947 (2015)

4. Banani SF, Lee HO, Hyman AA, & Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol.* **18**, 285-298 (2017)

5. Shin Y, & Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science.* **357**, eaaf4382 (2017)

6. Wang B, et al. Liquid-liquid phase separation in human health and diseases. *Signal Transduct Target Ther.* **6**, 290 (2021)

7. Ryan VH, et al. Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation. *Mol Cell.* **69**, 465-479 (2018)

8. Burke KA, et al. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Mol Cell.* **60**, 231-241 (2015)

9. Lin Y, et al. Redox-mediated regulation of an evolutionarily conserved cross-β structure formed by the TDP43 low complexity domain. *Proc Natl Acad Sci U S A.* **117**, 28727-28734 (2020)

10. Harrison AF & Shorter J. RNA-binding proteins with prion-like domains in health and disease. *Biochem J.* **474**, 1417-1438 (2017)

11. Carey JL & Guo L. Liquid-Liquid Phase Separation of TDP-43 and FUS in Physiology and Pathology of Neurodegenerative Diseases. *Front Mol Biosci.* **9**, 826719 (2022)

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

12. Murray DT, et al. Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell.* **171**, 615-627 (2017)

13. Yoshizawa T, et al. Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites. *Cell.* **173**, 693-705 (2018)

14. Qamar S, et al. FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions. *Cell.* **173**, 720-734 (2018)

15. Guo L, et al. Nuclear-Import Receptors Reverse Aberrant Phase Transitions of RNA-Binding Proteins with Prion-like Domains. *Cell.* **173**, 677-692 (2018)

16. Hofweber M, et al. Phase Separation of FUS Is Suppressed by Its Nuclear Import Receptor and Arginine Methylation. *Cell*. **173**, 706-719 (2018)

17. Liu Z, et al. Hsp27 chaperones FUS phase separation under the modulation of stress-induced phosphorylation. *Nat Struct Mol Biol.* **27**, 363-372 (2020)

18. Gu J, et al. Hsp40 proteins phase separate to chaperone the assembly and maintenance of membraneless organelles. *Proc Natl Acad Sci U S A.* **117**, 31123-31133 (2020)

19. Frottin F, et al. The nucleolus functions as a phase-separated protein quality control compartment. *Science.* **365**, 342-347 (2019)

20. Murray KA, et al. De novo designed protein inhibitors of amyloid aggregation and seeding. Proc Natl Acad Sci U S A. **119**, 34 (2022)

21. Kovacic JC, et al. Endothelial to Mesenchymal Transition in Cardiovascular Disease: JACC State-of-the-Art Review. *J Am Coll Cardiol.* **73**, 190-209 (2019)

22. Na TY, Schecterson L, Mendonsa AM, & Gumbiner BM. The functional activity of E-cadherin controls tumor cell metastasis at multiple steps. *Proc Natl Acad Sci U S A.* **117**, 5931-5937 (2020)

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

488 23. Akiyama T, et al. Aberrant axon branching via Fos-B dysregulation in FUS-ALS
489 motor neurons. *EBioMedicine.* **45**, 362-378 (2019)
490
491 24. Yamashiro Y, et al. Partial endothelial-to-mesenchymal transition mediated by HIF-
492 induced CD45 in neointima formation upon carotid artery ligation. *Cardiovasc Res.*
493 **119**, 1606-1618 (2023)
494
495 25. Takahashi K, & Yamanaka S. Induction of pluripotent stem cells from mouse
496 embryonic and adult fibroblast cultures by defined factors. *Cell*. **126**, 663-676 (2006)
497
498 26. Sharma R, et al. Liquid condensation of reprogramming factor KLF4 with DNA
499 provides a mechanism for chromatin organization. *Nat Commun.* **12**, 5579 (2021)
500
501 27. Cassandri M, et al. Zinc-finger proteins in health and disease. *Cell Death Discov.* **3**,
502 17071 (2017)
503
504 28. Al-Naama N, Mackeh R, Kino T. C2H2-Type Zinc Finger Proteins in Brain
505 Development, Neurodevelopmental, and Other Neuropsychiatric Disorders:
506 Systematic Literature-Based Analysis. *Front Neurol.* **11**, 32 (2020)
507
508 29. Iida A, et al. A functional variant in ZNF512B is associated with susceptibility to
509 amyotrophic lateral sclerosis in Japanese. *Hum Mol Genet.* **20**, 3684-3692 (2011)
510
511 30. Laity JH, Lee BM, & Wright PE. Zinc finger proteins: new insights into structural and
512 functional diversity. *Curr Opin Struct Biol.* **11**, 39-46 (2001)
513
514 31. Xiang S, et al. The LC Domain of hnRNPA2 Adopts Similar Conformations in
515 Hydrogel Polymers, Liquid-like Droplets, and Nuclei. *Cell.* **163**, 829-839 (2015)
516
517 32. Yesudhas D, Batool M, Anwar MA, Panneerselvam S, & Choi S. Proteins
518 Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains
519 in Stem Cell Transcription Factors. *Genes (Basel).* **8**, 192 (2017)
520
521 33. Mackeh R, Marr AK, Fadda A, & Kino T. C2H2-Type Zinc Finger Proteins:
522 Evolutionarily Old and New Partners of the Nuclear Hormone Receptors. *Nucl Recept*
523 *Signal.* **15**, 1550762918801071 (2018)

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

34. Farmiloe G, Lodewijk GA, Robben SF, van Bree EJ, & Jacobs FMJ. Widespread correlation of KRAB zinc finger protein binding with brain-developmental gene expression patterns. *Philos Trans R Soc Lond B Biol Sci.* **375**, 20190333 (2020)

35. Bu S, Lv Y, Liu Y, Qiao S, & Wang H. Zinc Finger Proteins in Neuro-Related Diseases Progression. *Front Neurosci.* **15**, 760567 (2021)

36. Moore DL, et al. KLF family members regulate intrinsic axon regeneration ability. *Science.* **326**, 298-301 (2009)

37. Quintes S, et al. Zeb2 is essential for Schwann cell differentiation, myelination and nerve repair. *Nat Neurosci.* **19**, 1050-1059 (2016)

38. Hegarty SV, et al. Zeb2 is a negative regulator of midbrain dopaminergic axon growth and target innervation. *Sci Rep.* **7**, 8568 (2017)

39. Gu S, Xu M, Chen L, Shi X, & Luo SZ. A liquid-to-solid phase transition of Cu/Zn superoxide dismutase 1 initiated by oxidation and disease mutation. *J Biol Chem.* **299**, 102857 (2023)

40. Das B, et al. A Zn-dependent structural transition of SOD1 modulates its ability to undergo phase separation. EMBO J. 42, e111185 (2023)

41. Monette A, et al. Pan-retroviral Nucleocapsid-Mediated Phase Separation Regulates Genomic RNA Positioning and Trafficking. *Cell Rep.* **31**, 107520 (2020)

42. Nanaura H, et al. C9orf72-derived arginine-rich poly-dipeptides impede phase modifiers. *Nat Commun.* **12**, 5301 (2021)

43. Kato, M., Lin, Y. & McKnight, S. L. Cross-beta polymerization and hydrogel formation by low-complexity sequence proteins. *Methods.* **126**, 3–11 (2017).

44. Verschueren K, et al. SIP1, a novel zinc finger/homeodomain repressor, interacts with Smad proteins and binds to 5'-CACCT sequences in candidate target genes. *J Biol Chem.* **274**, 20489-20498 (1999)

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

45. Delaglio F, et al. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR.* **6**, 277-293 (1995)

46. Lee W, Tonelli M, & Markley JL. NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics.* **31**, 1325-1327 (2015)

47. Kobayashi N, et al. Noise peak filtering in multi-dimensional NMR spectra using convolutional neural networks. *Bioinformatics.* **34**, 4300-4301 (2018)

48. Baek, Y. & Park, Y. Intensity-based holographic imaging via space-domain Kramers–Kronig relations. Nat. Photonics **15**, 354–360 (2021).

49. Trott O, & Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. **31**, 455-461 (2010)

50. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* **596**, 583-589 (2021).

51. Lu, J. et al. CryoEM structure of the low-complexity domain of hnRNPA2 and its conversion to pathogenic amyloid. *Nat. Commun.* **11**, 4090 (2020).

52. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods.* **14**, 71-73 (2017).

53. Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comp. Phys.* **117**, 1-19 (1995).

54. Hockney, R.W. & Eastwood, J.W. Computer simulation using particles. *Hilger, Bristol.* (1989).

55. Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review.* **159**, 98-103 (1967).

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

56. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comp. Phys.* **23**, 327-341 (1977).

57. Miller, B.R. et al. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J. Chem. *Theory Comput.* **8**, 3314-3321 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

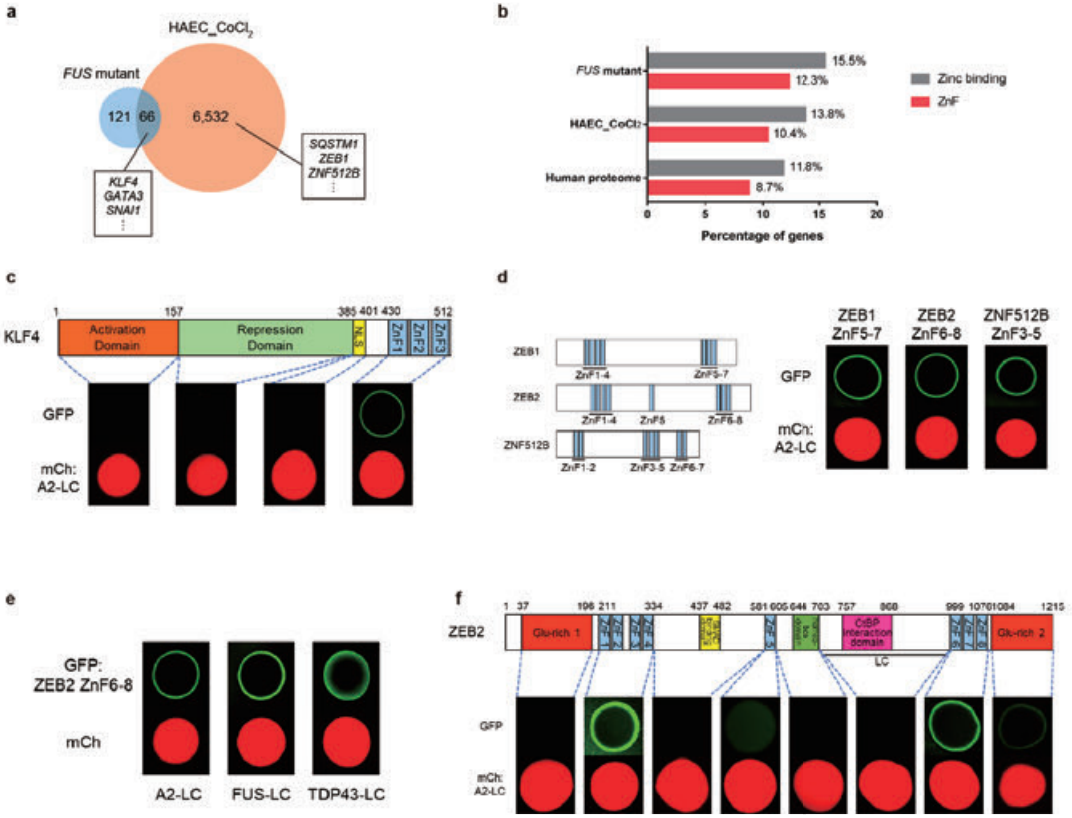


Fig. 1: ZnFs interact with LC domains. a. Venn diagram showing gene overlap in two comparisons, RNA-seq data of motor neurons derived from *FUS* mutant hiPSCs (*FUS* mutant) and ChIP-seq data of $CoCl_2$-treated Human aortic endothelial cell (HAEC_$CoCl_2$). **b.** Percentages of genes encoding zinc binding and ZnF domains in the three datasets of FUS mutant, HAEC_$CoCl_2$, and human proteome. **c**. Binding assays of GFP fusion KLF4 fragments to hydrogels derived from the mCherry fusion hnRNPA2 LC domain (mCh:A2-LC). **d**. Binding assays of various GFP fusion ZnF (GFP:ZEB1 ZnF5-7, ZEB2 ZnF6-8, and ZNF512B ZnF3-5) to mCh:A2-LC. **e**. Binding assays of GFP fusion ZEB2 ZnF6-8 (GFP:ZEB2 ZnF6-8) to various mCherry fusion LC domains (mCh:A2-LC, mCh:FUS-LC, and mCh:TDP43-LC). **f.** Binding assays of GFP fusion ZEB2 fragments to mCh:A2-LC hydrogels.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

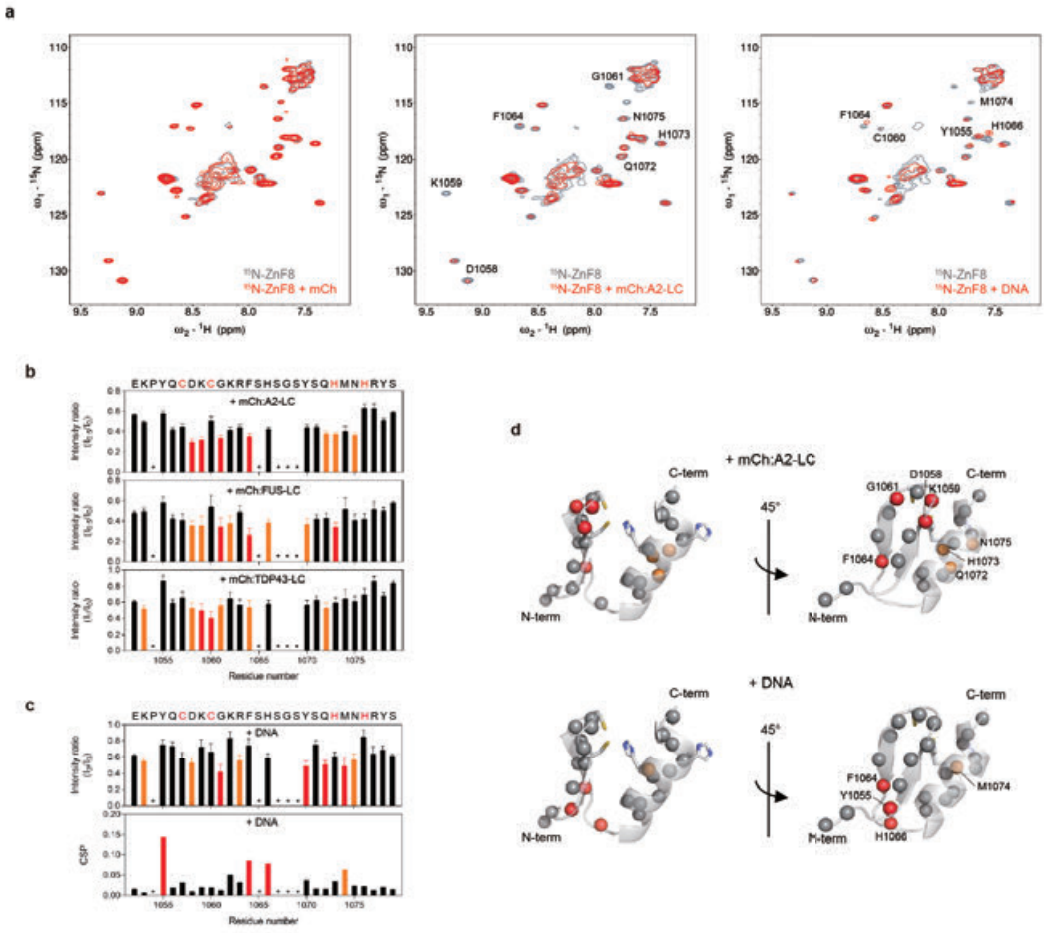
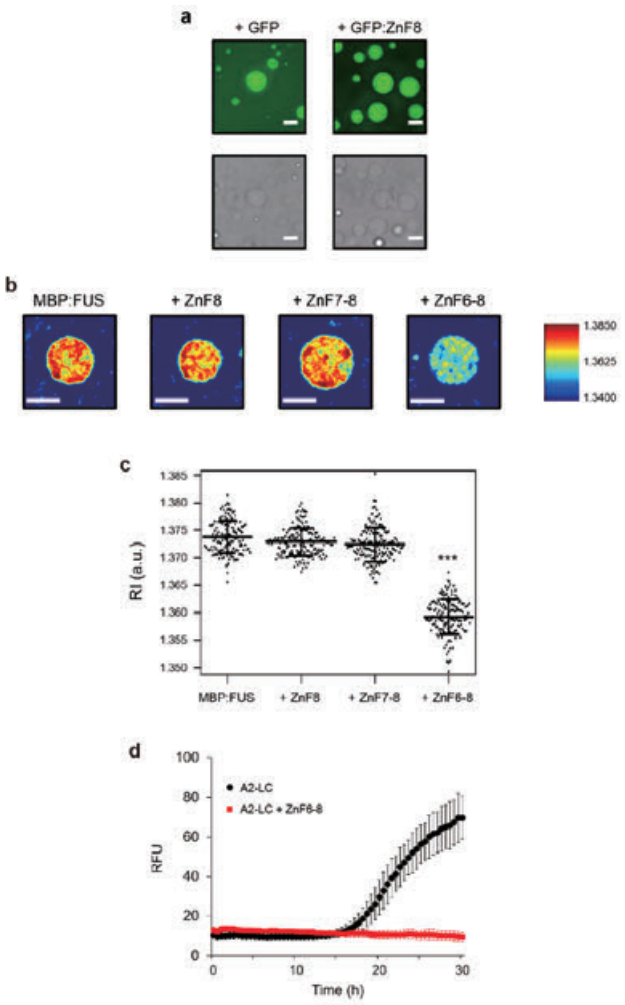


Fig. 2: The binding mode of ZnF to LC domains is different from that to DNA. a. ^{1}H-^{15}N NMR spectra of ^{15}N-labeled ZEB2-ZnF8 (^{15}N-ZnF8) with mCherry (mCh; left), mCherry-fused hnRNPA2 LC domain (mCh:A2-LC; middle), ZnF-binding DNA (right). Signals labeled by amino acid type and residue number showed significant changes. **b**. Change in signal intensity of ^{15}N-ZnF8 upon LC-binding. The intensity ratio is the signal intensity of LC-bound ZnF8 divided by the signal intensity of LC-free ZnF8. Significantly changed residues (intensity ratios less than mean−0.5SD and mean−SD) are indicated by orange and red bars, respectively. Residues for which no signal was detected in the absence of LC domains are indicated by asterisks. **c**. Changes in signal intensity (upper) and chemical shift (lower) of ^{15}N-ZnF8 upon DNA-binding. In the graph of intensity ratios, significantly changed residues (intensity ratios less than mean−0.5SD and mean−SD) are indicated by orange and red bars, respectively. In the graph of chemical

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

shift perturbations (CSP), significantly changed residues (CSP greater than mean+0.5SD and mean+SD) are indicated by orange and red bars, respectively. Residues for which no signal was detected in the absence of LC domains are indicated by asterisks. **d.** Ribbon model of ZEB2-ZnF8 predicted by AlphaFold2. In the upper panel, residues that were significantly changed by binding to mCh:A2-LC (intensity ratios less than mean−0.5SD and mean−SD) are shown in orange and red, respectively. In the lower panel, residues that were significantly changed by DNA-binding (CSP greater than mean+0.5SD and mean+SD) are indicated by orange and red bars, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



677
678

679 **Fig. 3: The binding of ZnF affects the phase separation of LC domains. a.**
680 Microscopic images of MBP:FUS droplets in the presence of GFP or GFP fusion ZEB2
681 ZnF8 (GFP:ZnF8). Scale bars are 5 μm. **b.** Refractive index (RI) images of MBP:FUS
682 droplets in the absence (MBP:FUS) and presence of ZEB2 ZnF peptides (ZnF8, ZnF7-8,
683 and ZnF6-8). Scale bars are 5 μm. **c.** Quantitative analysis of Fig. 3b. Statistical
684 significance was examined by one-way analysis of variance with Tukey's honest
685 significant difference post hoc test (***$p < 0.001$). **d**. Thioflavin T assays of hnRNPA2
686 LC domain (A2-LC) in the absence and presence of ZEB2 ZnF6-8 peptide (ZnF6-8). Data
687 represent means ± SD ($n = 4$).

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.29.564599; this version posted October 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

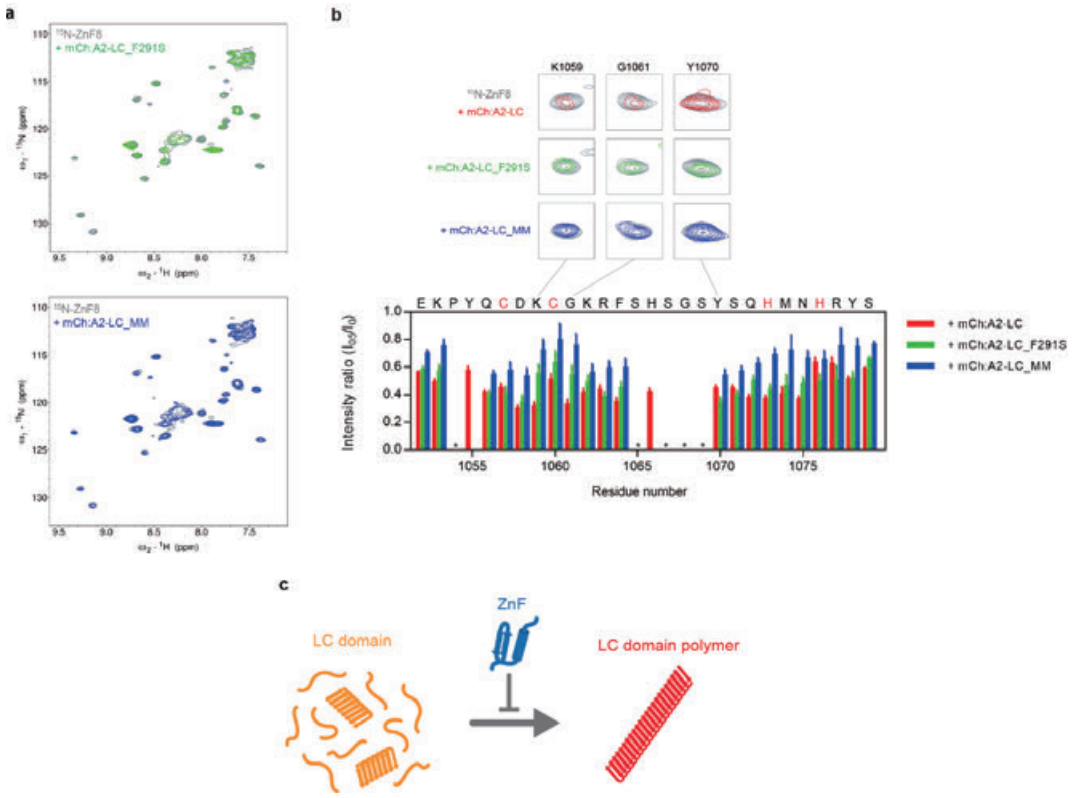


Fig. 4: ZnF preferentially recognizes LC polymers over monomeric LC domains. a. ^{1}H-^{15}N NMR spectra of ^{15}N-labeled ZEB2 ZnF8 (^{15}N-ZnF8) with mCherry fusion hnRNPA2 LC domain mutants (mCh:A2-LC_F291S and mCh:A2-LC_MM). **b.** Change in signal intensity of ^{15}N-ZnF8 in the addition of mCh:A2-LC (red), mCh:A2-LC_F291S (green), and mCh:A2-LC_MM (blue). The intensity ratio is the signal intensity of A2-LC-bound ZnF8 divided by the signal intensity of A2-LC-free ZnF8. Residues for which no signal was detected in the absence of LC domains are indicated by asterisks. In the upper panel, the NMR signals corresponding to the three residues K1059, G1061, and Y1070 are shown. **c.** a model that ZnFs play as physiological regulators of LC domain polymer formation.

6)

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identification of three distinct cell populations for urate excretion in human kidney

Yoshihiko M. Sakaguchi [1,2,3,#], Pattama Wiriyasermkul [2,3,#], Masaya Matsubayashi [4,#], Masaki Miyasaka[2,3], Nau Sakaguchi [1], Yoshiki Sahara [5,6], Minoru Takasato [5,6], Kaoru Kinugawa [1,7], Kazuma Sugie [7], Masahiro Eriguchi [8], Kazuhiko Tsuruya [8], Hiroki Kuniyasu [9], Shushi Nagamori [2,3,10,*], Eiichiro Mori [1,11,]**

[1] Department of Future Basic Medicine, Nara Medical University, Kashihara, Nara, Japan

[2] Center for SI Medical Research, [3] Department of Laboratory Medicine, The Jikei University School of Medicine, Tokyo, Japan

[4] Biological Research Department, Research Institute, Fuji Yakuhin Co., Ltd., Saitama, Japan.

[5] RIKEN Center for Biosystems Dynamics Research, Kobe, Hyogo, Japan

[6] Graduate School of Biostudies, Kyoto University, Kyoto, Japan

[7] Department of Neurology, Nara Medical University, Kashihara, Nara, Japan

[8] Department of Nephrology, Nara Medical University, Kashihara, Nara, Japan

[9] Department of Molecular Pathology, Nara Medical University, Kashihara, Nara, Japan

[10] Department of Collaborative Research for Bio-Molecular Dynamics, Nara Medical University, Kashihara, Nara, Japan

[11] V-iCliniX Laboratory, Nara Medical University, Kashihara, Nara, Japan

[#] Y.M.S., P.W. and M.M. contributed equally to this work

*** Correspondence:**

Shushi Nagamori

E-mail: snagamori@nagamori-lab.jp

**** Correspondence:**

Eiichiro Mori

E-mail: emori@naramed-u.ac.jp

Keywords: transporter, single nucleus RNA-sequencing, uric acid, human kidney, gout, solute carriers, purine

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Highlight/Key points

- We identified three distinct cell populations within the human renal anatomy that predict putative cellular transport mode, and our findings indicate cellular inhomogeneity with distinct roles such as urate secretion and reabsorption.
- Our model of physiological urate handling demonstrates the excretion dynamics in human kidney in terms of single cell-units.
- Our cellular urate transport analyses suggest the reversibility of some urate transporters even in certain physiological conditions.
- The physiological function of SLC2A9 is not limited to urate reabsorption; it is also involved in urate secretion restriction.
- This methodology can be applied to investigations of transport mechanisms in general, regardless of epithelial cell types, species, and substrates.

Abstract

In humans, uric acid is an end-product of purine metabolism. Urate excretion from human kidney is tightly regulated by reabsorption and secretion. At least eleven genes have been identified as human renal urate transporters. However, it remains unclear whether all renal tubular cells express the same set of urate transporters. Here we show that renal tubular cells are divided into three distinct cell populations for urate handling. Analysis of healthy human kidneys at single-cell resolution revealed that not all renal tubular cells expressed the same set of urate transporters. Only 32% of renal tubular cells were related to both reabsorption and secretion, while the remaining renal tubular cells were related to either reabsorption or secretion, at 5% and 63% respectively. These results provide physiological insight into the molecular function of the transporters and renal urate handling on cell-units. Our findings also suggest that three different tubular cell populations cooperate to regulate urate excretion from human kidney.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

58

59 ## Introduction

60 Excretion, the process of biological waste removal, is a vital homeostatic mechanism in all
61 organisms. Humans have turnovers of 3.3×10^{11} cells per day (Sender & Milo, 2021), and wastes of
62 nucleobases, the basic components of DNA and RNA, are removed with each turnover. Nucleobases
63 are classified into pyrimidine and purine bases. Pyrimidine bases are catabolized to water, carbon
64 dioxide, and ammonia, and purine bases are catabolized to uric acid as a final metabolic waste.
65 Physiologically, more than 90% of serum uric acid exists as monosodium urate, which is excreted by
66 urate transporters to maintain nucleic acid homeostasis. In other mammals, uric acid is hydrolyzed by
67 uricase (urate oxidase). However, the loss of uricase during primate evolution (Kratzer *et al.*, 2014)
68 resulted in urate excretion becoming a more important nucleic acid homeostatic mechanism in humans
69 than in other mammals.

70 Approximately two-thirds of urate excretion in humans is renal excretion, and the remaining one-
71 third is extra-renal excretion, such as intestinal excretion (Sorensen, 1965). The renal excretion in
72 humans is a reabsorption-dominant system composed of reabsorption and secretion (Steele &
73 Rieselbach, 1967). Identification of urate post-secretory reabsorption site (Diamond & Paolino, 1973)
74 led to these physiological urate excretion dynamics being proposed as a “four-component model”
75 (Levinson & Sorensen, 1980). These dynamics are tightly regulated by functional urate transport unit
76 (*urate transportome*), which consists of urate transporters and scaffold proteins (Anzai *et al.*, 2004;
77 Srivastava *et al.*, 2019).

78 At least eleven genes have been identified as transporters in human kidney. The ATP-binding
79 cassette (ABC) transporter or solute carrier (SLC) transporter (Enomoto *et al.*, 2002; Ichida *et al.*,
80 2003; Matsuo *et al.*, 2008; Iharada *et al.*, 2010; Chiba *et al.*, 2015; Hyndman *et al.*, 2016), and specific
81 counterpart efflux transporters to influx transporters are involved in urate reabsorption and secretion,
82 respectively (Bobulescu & Moe, 2012; Hyndman *et al.*, 2016; Nigam & Bhatnagar, 2018). Recently,
83 some urate reabsorption transporters (SLC22A11, SLC22A12, and SLC22A13) are found not to be co-
84 expressed (Matsubayashi *et al.*, 2021). However, it is still unclear whether all renal tubular cells express
85 the same set of urate transporters, and how the distribution of their cells in human kidney result in
86 physiological urate excretion.

87 To clarify whether all renal tubular cells express the same set of urate transporters, we analyzed
88 healthy human kidneys at single-cell resolution. We found that not all the renal tubular cells expressed

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the same set of urate transporters, and identified three distinct cell populations for urate handling. Our findings revealed that the molecular function of the transporters is variable and differs across the expression patterns of *transportome* components on single cell-units. Based on these results, we developed a physiological model for urate handling on single cell-units to visualize the dynamics of urate excretion in human kidney with higher resolution.

Methods

snRNA-seq datasets

We made no distinction between single-nucleus RNA sequencing (snRNA-seq) or scRNA-seq and treat them as scRNA-seq datasets (Matsubayashi *et al.*, 2021). We sought for scRNA-seq datasets based on two criteria. First, the datasets must be from kidneys of adult human males. Secondly, each dataset was used as a control in previous studies. The kidney in GSE118184 was validated as healthy kidney with a serum creatinine measurement of 1.03 mg/dL (Wu *et al.*, 2018). Two kidneys in GSE131882 were validated as healthy kidneys via H&E images in which no evidence of glomerulosclerosis, interstitial fibrosis, or immune cell infiltrate was found (Wilson *et al.*, 2019). Consequently, we selected and downloaded these three datasets from a public functional genomics data repository (Supplementary Table 1). As shown in other studies, we confirmed that we had an appropriate number of datasets for our study (Wilson *et al.*, 2019; Liao *et al.*, 2020). The gene name in the dataset "GSE118184" and "GSE131882" used HGNC gene symbol and Ensemble gene ID, respectively. We unified the gene names in all datasets as HGNC gene symbols by biomaRt (Durinck *et al.*, 2005, 2009).

Quality control of the datasets

Quality control was processed using the Seurat plug-in of R software (Butler *et al.*, 2018; Stuart *et al.*, 2019). Workflow was built according to the tutorial, "Guided tutorial – 2,700 PBMCs" listed in the Seurat website (https://satijalab.org/seurat/vignettes.html). To exclude low-quality and dying cells, we filtered cells out if their mitochondrial gene content was > 5% because high mitochondrial gene content is reported to be related to low-quality or dying cells. To exclude empty droplets and cell doublets or multiplets, we filtered cells out if their gene counts detected were less than 200 or more than each threshold. The thresholds of the three datasets were set as 6000, 6000, and 3000, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Finally, the datasets used in this study consist of three males and a total of 10,080 cells (Supplementary Table 1).

Multiple dataset integration and data pre-processing of cell clustering

Data integration and data pre-processing of cell clustering were performed using the Seurat plug-in of R software. The workflow was set following the tutorials, "SCTransform", "Cell Cycle Regression", and "Integration and Label Transfer - SCTransform" listed in the Seurat website (https://satijalab.org/seurat/vignettes.html). We selected sctransform as a normalization method; it uses regularized negative binomial regression to normalize UMI count data (Hafemeister & Satija, 2019). Variations in technical factors of scRNA-seq data tend to confound the actual biological variations (Stegle *et al.*, 2015; Vallejos *et al.*, 2017). We utilized the sctransform normalization and data integration by Seurat in the pre-processing stage to reduce more technical factor variations and preserve real biological variations compared to the standard Seurat workflow (Butler *et al.*, 2018; Hafemeister & Satija, 2019). We performed the sctransform normalization twice. The first sctransform normalization was performed with the regression of nFeature_RNA and nCount_RNA.

Cell cycle regression is a method to lessen the effects of cell cycle heterogeneity in scRNA-seq data by calculating cell cycle phase scores based on canonical markers (Nestorowa *et al.*, 2016). The calculation utilized the values obtained after the first sctransform. After the calculation of cell cycle scores, a second sctransform normalization was performed again with the regression of nFeature_RNA, nCount_RNA, and cell cycle scores. Finally, we integrated the three datasets according to the tutorial "Integration and Label Transfer - SCTransform" (Butler *et al.*, 2018).

Cell clustering

To visualize the cell types in the three datasets, we performed cell clustering. The process of cell clustering was also performed using the Seurat plug-in of R software. The workflow was built according to the tutorial "Integration and Label Transfer - SCTransform" in the Seurat website (https://satijalab.org/seurat/vignettes.html). After an integration of the three datasets, we projected these datasets onto two dimensions with uniform manifold approximation and projection (UMAP), a dimension reduction technique. All single nuclei were divided into a total of 16 clusters. The clusters were named by the renal anatomical, which were annotated by marker genes (Supplementary Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The marker genes were selected based on criteria listed in previous studies (Wu *et al.*, 2018; Wilson *et al.*, 2019).

Annotation of segments in proximal tubule (PT) clusters

Three PT clusters were divided into S1: S1 segment of proximal tubule, S2: S2 segment of proximal tubule, and S3: S3 segment of proximal tubule. These clusters were annotated by marker genes previously reported (Supplementary Table 3). Selection of the marker genes for S1 and S3 segments were based on previous studies (Ba *et al.*, 2003; Mather & Pollock, 2011). As no specific marker genes for S2 segments were reported, we annotated the S2 segments from the cluster which was not marked as S1 or S3 segments in three PT clusters.

Gene positivity and gene negativity

Gene expression data were normalized by the "LogNormalize" method on Seurat. First, gene counts for each cell were divided by the total counts for that cell and the values multiplied by 10,000 were set as the default scale factor on Seurat. Then, these were natural-log transformed. After the log transformation, we defined "gene X"-positivity (the presence of gene X) as cells that expressed "gene X" above the cutoff value. By contrast, "gene X"-negativity (the absence of gene X) was defined as cells that expressed "gene X" less than the cutoff value. A value '0.5' was set as the cutoff value because the expression levels had a neckline around the value shown in a violin plot. The gene-positive cell numbers were counted after the translation from the quantitative expression data to binary expression data (gene-positive or negative).

All urate transporter genes were classified into one of four types, Apical Influx (AI) transporter, Apical Efflux (AE) transporter, Basolateral Influx (BI) transporter, or Basolateral Efflux (BE) transporter, based on a previous study (Hyndman *et al.*, 2016). The positivity or negativity of each gene was then examined as described above. After that, their expression profiles were constructed into cell populations based on the positivity/negativity of AI and BI transporters. We classified cell populations into one of four cell populations (Apical Influx Transporter (AIP) cell population, Basolateral Influx Transporter (BIP) cell population, Double Positive (DIP) cell population, or Double Negative (DIN) cell population). The AIP cell population had the positivity of at least one AI transporter (*SLC22A11* and/or *SLC22A12*) and the negativity of BI transporters (*SLC22A6*, *SLC22A7*, and *SLC22A8*). The BIP cell population had the positivity of at least one BI transporter (*SLC22A6*, *SLC22A7*, and/or *SLC22A8*) and the negativity of AI transporters (*SLC22A11* and *SLC22A12*). The

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DIP cell population had the positivity of both AI transporters (*SLC22A11* and/or *SLC22A12*) and BI transporters (*SLC22A6*, *SLC22A7,* and/or *SLC22A8*). The DIN cell population had the negativity of both AI and BI transporters. After that, we subcategorized each cell populations by adding the positivity/negativity of AE and BE transporters in the same way. Finally, we calculated the cell number presented in Table 2 as the numbers of cells which showed the positivity of indicated transporters.

Visualization of the expression data – dot plots, violin plot, and bar plot

The visualization of the gene expression data was performed using Seurat plug-in of R software (Butler *et al.*, 2018; Stuart *et al.*, 2019), ggplot2 plug-in of R software (Wickham, 2010), and matplotlib.plt library of python software (Hunter, 2007). All expression levels used in data visualization were the scaled data after the log transformation as described in the previous section. Dot and violin plots of gene expression in each cluster and scatter plots of two genes expression were created by Seurat function, *"Dotplot"*, *"VlnPlot"*, plug-in of R software. We created bar plots using matplotlib, which is the Python library for producing plots and data visualizations. Plots reside within a *Figure* object that was created by a function named *"Figure"* in matplotlib.plt library, and the bar plots were made by the *show* method in *Figure* object.

Data availability

R and Python codes for data analysis and visualization is available in GitHub repository at https://github.com/yoshi-sci/urate-transporter-analysis.

Results

1 Urate handling transporter identification and localization

To examine the localization of the transporters in human kidney at single-cell resolution, we first verified the renal anatomical regions of all single nuclei in the three datasets. After unsupervised clustering of all single nuclei in the datasets, all clusters were annotated by the marker genes (Figure 1A, Supplementary Figure 1A). The annotation was based on the anatomical regions, and cells in the clusters were calculated as percentages in the regions (Figure 1B, Supplementary Figure 1B). Proximal tubule (PT) is the region for urate reabsorption and secretion, and it was divided into three segments based on anatomical localization. Using the PT region-specific marker genes, we were able to annotate

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

three PT segments: S1, S2, and S3 segments (Figure 1C). The three analyzed datasets covered all cell types of renal tubules including three PT segments (Supplementary Figure 1C), which allowed us to evaluate the localization of urate transporters.

At least eleven genes have been identified as urate transporters in human kidney. To confirm the transporters and the localization for urate handling in human kidneys, we focused on the expression of the eleven known transporters across the regions (Supplementary Figure 1D). Except for *ABCG2*, *ABCC4,* and *SLC22A13*, the remaining eight transporters were specifically expressed in the PT and the descending loop of Henle (DL). *ABCG2* and *SLC22A13* were rarely expressed in all clusters, while *ABCC4* was the only transporter highly expressed in the intercalated cell type A region (Figure 1D). The results clarified that transport via these eight transporters existed mainly in the PT, and to some degree in the DL region. Consequently, our subsequent analyses focused on 1) the four regions (S1, S2, S3, and DL), and 2) the eight urate transporters specifically expressed in the clusters (SLC2A9, SLC17A1, SLC17A3, SLC22A6, SLC22A7, SLC22A8, SLC22A11, and SLC22A12) (Figure 1E).

The eight transporters have unique characteristics for cellular urate transport. In physiological conditions, the transporters are proposed to be unidirectional based on the transport affinity and substrate availability both outside and inside the cells (Bobulescu & Moe, 2012; Hyndman *et al.*, 2016). Table 1 summarizes the eight transporters in terms of their transport systems, substrates, affinities (K_m) for urate, and average expressions in the clusters of PT and DL (Koepsell, 2013; Mueckler & Thorens, 2013; Reimer, 2013), demonstrating that these transporters can be classified based on their cellular function for urate transport.

2 Classification of four transporter types and four cell populations

To classify the eight transporters, we considered polarized localization (apical or basolateral membranes) and urate flux (efflux or influx). We categorized the eight transporters into one of four types (Figure 2A): Apical Influx urate transporters (AI transporters), Apical Efflux urate transporters (AE transporters), Basolateral Influx urate transporters (BI transporters), and Basolateral Efflux urate transporters (BE transporters). The cooperation of AI and BE transporters leads renal tubular cells to function as reabsorption, whereas the cooperation of BI and AE transporters leads their cells to function as secretion. This method of classification helped us to predict the cellular urate transport function.

Based on the cellular expression profiles of the transporter types, we hypothesized that cell populations have distinct cellular functions for transport. Based on the gene-positivity of the influx

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

transporters, PT and DL cells were categorized into one of four cell populations (Figure 2B): 1) AI transporter—positive cell population expressing AI transporters but not BI transporters (AIP cell population); 2) BI transporter—positive cell populations expressing BI transporter but not AI transporters (BIP cell population); 3) Double Influx transporters—positive cell populations expressing both AI and BI transporters (DIP cell population); and 4) Double Influx transporters—negative cells lacking the expression of both AI and BI transporters (DIN cell population). For example, the AIP cell population expressed *SLC22A11*, *SLC22A12*, or both, but had no expression of *SLC22A6*, *SLC22A7,* or *SLC22A8*. The positivity of AI (BI) transporters was defined as the expression of at least one AI (BI) transporter in each cell.

To elucidate the expression profiles of the influx transporters in the cell populations, we first examined how frequently the same types of influx transporters were co-expressed. The most frequent AI transporter was *SLC22A11* across the regions (Supplementary Figure 2A, B), while the most frequent BI transporter was *SLC22A6* (Supplementary Figure 2C-E). We found that for AI transporters, all tubular regions showed that about 70 – 80% of the DIP cell population expressed *SLC22A11* but not *SLC22A12*, and less than 10% of the DIP cell population expressed both *SLC22A12* and *SLC22A11* (Figure 2C). For BI transporters, approximately 75% of the DIP cell population in the S1 and S3 regions showed co-expression of two or more different BI transporters, although the proportion decreased in the S2 and DL regions (Figure 2D). These frequency trends of influx transporters in the AIP (BIP) cell population were also found to be common in the DIP cell population (Supplementary Figure 2F, G). These results suggest that only one kind of AI transporter, mostly SLC22A11, is expressed, while other kinds of BI transporters might be co-expressed in addition to SLC22A6 (Figure 2E).

3 Expression of the efflux transporters indispensable for predicting cellular urate transport

We made the following two assumptions when classifying the three cellular urate transport directions, reabsorption mode, secretion mode, and bidirectional mode. First, we assume that transport direction was based on the availability of urate in both the lumen and the blood, and the efflux of urate would not take place without the influx of urate into the cells. Thus, the DIN cell population was omitted due to the lack of influx transporters (Supplementary Figure 3A). Secondly, we assumed that a complete transport direction mode required the co-expression of an influx transporter on one side of the polarized membrane (either AI or BI transporter) and an efflux transporter on the other side of the polarized membrane in the same cell. In other words, the AI and BI transporters were assigned as the primitive transporters, while the AE and BE transporters accomplished such transport modes as the

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

derivative transporters. These two assumptions were applied to the determination of the cellular directional mode in the three cell populations: AIP, BIP, and DIP cell populations (Figure 3A).

To visualize urate handling across the regions, we clarified which cell populations were present in the regions. The BIP and DIP cell populations were highly dominant in the S1 region and gradually decreased along the anatomical region (Figure 3B). This suggests that the three cell populations are not only in the PT region but also in the DL region, and that their proportion decreases as anatomical localization progressed downward from the S1 to DL regions. In all PT segments, the proportion of the BIP cell population was approximately 50%, whereas the AIP cell population was the smallest among the cell populations (Figure 3B). These analyses indicate that the cells expressing AI transporters (the AIP and DIP cell populations) also express the BI transporters, but the cells expressing BI transporters (the BIP and DIP cell populations) do not necessarily express AI transporters.

To elucidate the expression profiles of the efflux transporters across the cell populations, we analyzed the frequencies and expression levels of BE and AE transporters (Figure 3C, Supplementary Figure 3B, C). In all PT and DL clusters, the BIP and DIP cell populations exhibited a high frequency and expression level of *SLC2A9*, the only known BE transporter, while the AIP cell population showed a low frequency and expression level of *SLC2A9* (Figure 3C). These results suggest that SLC2A9 is expressed not only in the renal tubular cells related to cellular urate reabsorption (the DIP and AIP cell populations), but also in the renal tubular cells involved in cellular urate secretion (the BIP cell population); SLC2A9 plays the role of lowering the secretion efficiency. To clarify the regions affected by SLC2A9 to decrease secretion efficiency, we showed *SLC2A9* expression in the BIP cell population with AE transporters. The expression had peaks on both ends, the S1 and DL regions (Figure 3D), suggesting that SLC2A9 lowers the secretion efficiency in the BIP cell population, especially at the beginning and end of urate handling regions (Figure 3E).

4 Contribution of urate transportome to urate handling

We considered the presence of regulatory factors that promoted the expression patterns of the transporters. We hypothesized that one of the key factors was PDZ proteins, which are scaffold proteins that interact with various transmembrane proteins, including transporters via PDZ motifs at the C-termini. PDZK1, a scaffold protein, forms the functional urate transport unit (*urate transportome*) by clustering urate transporters at the apical membrane (Anzai *et al.*, 2004; Srivastava *et al.*, 2019). We first investigated the correlation between PDZK1 expression patterns and the cell populations. The expression of *PDZK1* in each cell population revealed that the DIP cell population had the highest

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

297 frequency and expression level of *PDZK1;* PDZK1-positive cells accounted for approximately 60% of
298 the DIP cell population (Figure 4A), suggesting that the DIP cell population expresses PDZK1 more
299 frequently and strongly than the other cell populations.

300 In contrast to the BIP and AIP cell populations, the DIP cell population potentially directs both
301 secretion and reabsorption and changes the net cellular transport mode by the transporters expressed.
302 To clarify whether the PDZK1-mediated urate *transportome* favors the expression patterns of the
303 transporters, we showed the expression of *PDZK1* over the number of apical transporters. The more
304 diverse the apical urate transporters were, the higher the expression levels and frequencies of *PDZK1*
305 were (Figure 4B). *PDZK1*-positive cells had higher or equal frequencies of all apical transporters
306 compared to *PDZK1*-negative cells, and the AE transporters (*SLC17A1* and *SLC17A3*) were also more
307 strongly expressed in *PDZK1*-positive cells (Figure 4C). These results indicate the contribution of
308 PDZK1 to the formation of the cellular urate *transportome* at the apical membranes of the DIP cell
309 population.

310 To demonstrate that the positivity of PDZK1 is related to the net cellular transport mode in the
311 DIP cell population, we focused on the expression of the basolateral transporters in addition to the
312 apical transporters across the positivity of *PDZK1*. As shown in Figure 2E, the DIP cell population had
313 single or multiple kinds of BI transporters. *PDZK1*-positive cells more frequently had multiple kinds
314 of BI transporters than *PDZK1*-negative cells (Figure 4D). Also, *SLC2A9,* the AE transporter, was
315 expressed more frequently and strongly in *PDZK1*-positive cells (Supplementary Figure 4A). About
316 60% of the *PDZK1*-positive DIP cell population accounted for the positive expressions of both efflux
317 transporter types (the AE and BE transporters), while the *PDZK1*-negative DIP cell population more
318 frequently accounted for the negative expressions of the efflux transporter types, including the influx
319 transporter-only population (Figure 4E). These results suggest that PDZK1 has a positive relationship
320 with the basolateral transporters in addition to apical transporters and changes the net cellular transport
321 mode in the DIP cell population.

322 According to our cell population classification criteria, efflux transporters are not always expressed
323 in the DIP cell population. In fact, about 20% of the PDZK1-negative DIP cell population expressed
324 none of the efflux transporters (Figure 4E). It is interesting to consider the kind of cellular transport
325 modes the DIP cell population had without efflux transporters. Even if the cell population influxes
326 urate at an early stage, it would generate higher and higher intracellular urate concentration. We expect
327 that the change in intracellular and extracellular concentration gradient would then subsequently drive

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the role of the influx transporters to change the efflux physiologically. The fundamental principle of transporters is that the transport direction is determined by the substrate concentration gradient. The DIP cell population without efflux transporters could potentially switch the direction of urate transport from influx to efflux, for example, from a BI transporter to a BE transporter or from an AI transporter to an AE transporter (Figure 4F).

5 Design of a triple cell-unit model of urate handling, the "*CUTE model*"

The *Cellular Urate Transport Excretion* (*CUTE*) *model* proposed three representative cell populations: the BIP, DIP, and AIP cell populations (Figure 5A). The cellular transport modes were defined by the cell populations (either secretion or reabsorption); if the BIP (or AIP) cell population expressed AE (BE) transporters, the population worked for urate secretion (reabsorption) (Figure 5A). The DIP cell populations could work for urate secretion or reabsorption depending on the transporter types expressed. The DIP cell population with either AE (or BE) transporters could potentially play a role in urate secretion (or reabsorption), while the other DIP populations ("AE- BE-" or "AE+ BE+") could work bi-directionally (Supplementary Figure 5A). Our results suggest that not all renal tubular cells express the same set of urate transporters, and that the cells involved in the cellular reabsorption are distinct from those involved in the cellular secretion.

The clarification of the set of urate transporters at single-cell resolution provided some insights into transporter functions. First, the role of SLC2A9 could be significant to lower the secretion efficiency, which supports the clinical data of low urate excretion (Figure 5B: Upper panel). Second, the *CUTE model* proposed a property of transport reversibility in some transporters in physiological conditions (Figure 5B: Lower panel). These results were found only when the molecular function was taken into consideration on cell-units.

To clarify the cell proportion for urate handling across the regions, we quantified the counts for reabsorption cells, bi-directional cells, and secretion cells (Figure 5C, Supplementary Table 4). Proportions of the regions were shown (Supplementary Figure 5B), and the average proportions of the cell populations in the regions were as follows: secretion 63%, bi-direction 32%, and reabsorption 5%. The ratio of reabsorption cells per secretion cells had an upward trend along the renal tubule (S1 0.04, S2 0.09, S3 0.12, and DL 0.33); in particular, the ratio in the DL region was by far the highest among the regions. These results suggest that urate reabsorption is not as superior as urate secretion at the beginning of PT but becomes more significant at S3 and DL. Accordingly, secretion could be most active at the beginning of the PT and gradually declines along the PT segments.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

359

360 ## Discussion

361 The concept of having same sets of multiple transporter molecules in the same cells has been
362 proposed (Bobulescu & Moe, 2012; Hyndman *et al.*, 2016). Our study elucidated these sets of the
363 transporters in single cells and predicted the directions of the cellular urate transport. Four factors were
364 used to construct the cellular model. The first factor was the identification of renal anatomical regions
365 where known urate transporters were expressed, and the regions were expected to be the functional
366 regions for the reabsorption and secretion. The next factor was the identification of four transporter
367 types, which were defined based on the characteristics of the transporters. The third factor was the
368 classification of four cell populations based on the expression profiles of the transporters at single-cell
369 levels. The combination of the transporter types and the cell populations led to the prediction of cellular
370 direction modes of urate transport as the fourth factor. The *CUTE model* confirmed the expression of
371 the transporter molecules, except SLC22A13, ABCG2, and ABCC4, in human kidneys. In addition,
372 the *CUTE model* clarified the cellular inhomogeneity (Figure 5A) and complexity of cell population
373 distribution along the anatomical regions (Figure 5C). An advantage of having inhomogeneous cell
374 populations is probably the modulation of urate homeostasis.

375 The *CUTE model* proposes some concepts about the molecular function of transporters. One is a
376 reversal property of some transporters in human kidneys (Figure 5B: Lower panel). Bi-directional
377 functions of urate transporters *in vitro* have been previously reported (Enomoto *et al.*, 2002; Bakhiya
378 *et al.*, 2003; Anzai *et al.*, 2008; Jutabha *et al.*, 2010) while the urate transporters in physiological
379 conditions are proposed to restrict in uni-directional transport based on the transport affinity and
380 substrate availability at both the outside and inside of the cells (Bobulescu & Moe, 2012; Hyndman *et*
381 *al.*, 2016). The *CUTE model* suggests that the urate transporters function bi-directionally *in vivo* as
382 well as *in vitro*.

383 The reversibility is not applied to all urate transporters given in previous reports. For example,
384 SLC22A7 and SLC22A11 were reported to be unidirectional urate transporters (Hagos *et al.*, 2007;
385 Sato *et al.*, 2010). In addition, SLC22A12 and SLC2A9 tend to cooperate in urate reabsorption in the
386 co-expressing cells (Nakanishi *et al.*, 2013), although both transporters possibly divert in opposite
387 directions (Enomoto *et al.*, 2002; Anzai *et al.*, 2008). Cumulatively, these reports suggest that BI and
388 AE transporters, except SLC22A7, are capable of switching their transport directions, whereas AI and
389 BE transporters have restricted transport directions. The *CUTE model* relatively restricts the direction

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

modes yet accommodates some flexibility toward the reabsorption mode. Urate reabsorption is superior to urate secretion in the whole kidney. This model suggests a mechanism to keep the superiority of urate reabsorption.

Knowledge from human samples is essential to truly understanding human physiological systems. The *CUTE model* not only explains the molecular significance of urate transporters but also proposes clinical relevance. Currently, only two urate transporters related to the Mendelian disorders have been reported: SLC22A12 (OMIM 220150) which is responsible for renal hypouricemia type 1 (RHUC1) (Sperling, 2006), and SLC2A9 (OMIM612076) which is responsible for renal hypouricemia type 2 (RHUC2) (Kawamura *et al.*, 2011; Chiba *et al.*, 2014). Before the characterization of SLC2A9, most of the clinical studies focused on SLC22A12, and over 90% of renal hypouricemia was reported from *SLC22A12* mutations (Ichida *et al.*, 2004). Here, we refer to the clinical association of SLC2A9, especially in the elderly male population, because the datasets used in this study were from elderly males. The *CUTE model* advocates SLC2A9 over other transporters because of its single-player BE transporter, high expression profiles, and dual functions in reabsorption and secretion attenuation (Figure 5B: Upper panel). Future analysis of more datasets covering a wide range of ages will provide additional clues about other urate transporters.

The *CUTE model* provides some clues as to where novel urate transporter candidates might be located. Our model suggests the existence of novel transporter(s) involved in urate reabsorption. The most possible candidate is a novel BE transporter, which is likely expressed more frequently and strongly in the AIP and DIP cell populations than in the BIP and DIN cell populations due to the weak expression of SLC2A9 in current AIP and DIP cell populations within the S2 region (Table 2). The model itself can provide a ready platform for further analysis when a novel urate transporter is identified. The analytical model from this study is beneficial for both before and after the discovery of candidate transporters.

The series of analyses presented here is highly scalable, and the methodology itself can be a prototype for other metabolite analyses in all organs. It is well known that multiple transporters in the same cells coordinate to handle a substance across a tissue region, such as nephrons, intestinal epithelia, liver tissues, and blood-brain barrier. Our analytical strategy is useful for broad types of transporters and substrates, leading to the understanding of molecular transport mechanisms in actual physiological systems. Our study suggests the functional cooperation of SLC22A11 and SLC22A6 for urate reabsorption (Figure 5A). This functional cooperation is also the canonical transport pathway for other endogenous anion substrates or drugs (Hagos *et al.*, 2008, 2012; Willemin *et al.*, 2021). The

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methodology employed in this study is potentially a strong tool to elucidate the transport mechanism in physiological and pathological systems.

In interpreting the results, it is important to understand the limitations of the scRNA-seq data we used. We did not consider factors such as race/ethnicity, sex, and age, although they have been reported to impact serum uric acid levels (Barry *et al.*, 1992; Julius *et al.*, 2004). These three factors may cause different expression levels and/or frequencies of urate transporters, as seen in the case of SLC22A12 (Joseph *et al.*, 2015). Recently, SLC22A13 was found to contribute to urate reabsorption in an elderly Japanese male (Toyoda *et al.*, 2022), but the non-Asian scRNA-seq data we used in this study suggests that the low contribution of SLC22A13 is due to its low expression frequency (Figure 1A, Supplementary Figure 2B). In addition, as sequencing techniques and analyses are further developed, metadata analyses including these factors are expected. Further study is required to address these issues.

In conclusion, the *CUTE model* showed three representative cell populations with distinct transporter expression patterns and the differential distribution of the cell populations along the tubular regions. The construction of the model clarified that the molecule function of the transporters was variable and differed across the expression patterns of the transporters on single cell-units. The component of urate excretion was from 1) reabsorption cells (5% of the cell population) which certainly drive urate reabsorption; 2) secretion cells (63%) which are supposed to drive urate secretion, but whose process could be attenuated by the high affinity of SLC2A9; and 3) bi-directional cells (32%) which potentially function as both reabsorption and secretion, but tentatively drive the reabsorption over the secretion by the reversibility of some transporters. The inhomogeneity of the cell populations allowed us to predict distinct net cellular transport modes, suggesting that the cells for urate reabsorption are relatively different from the cells for urate secretion. The distribution indicates the most active region for urate transport is at proximal tubules and to some degree at the descending loop of Henle, which lead us to predict urate excretion dynamics in certain regions of the nephrons. Overall, the urate transport dynamics described by our model indicate the superiority of urate reabsorption over urate secretion, leading to high uric acid excretion. Our methodology reflects the basic concept that any process, whether secretion or reabsorption, must pass through at least two types of transporters: one at the apical membrane and the other at the basolateral membrane. We believe our model can be applied to the analysis of transporter systems in general, bridging the gap from molecular function to cellular function.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Anzai N, Ichida K, Jutabha P, Kimura T, Babu E, Chun JJ, Srivastava S, Kitamura K, Hisatome I, Endou H & Sakurai H (2008). Plasma urate level is directly regulated by a voltage-driven urate efflux transporter URATv1 (SLC2A9) in humans. *J Biol Chem* **283,** 26834–26838.

Anzai N, Miyazaki H, Noshiro R, Khamdang S, Chairoungdua A, Shin HJ, Enomoto A, Sakamoto S, Hirata T, Tomita K, Kanai Y & Endou H (2004). The multivalent PDZ domain-containing protein PDZK1 regulates transport activity of renal urate-anion exchanger URAT1 via its C terminus. *J Biol Chem* **279,** 45942–45950.

Ba J, Brown D & Friedman PA (2003). Calcium-sensing receptor regulation of PTH-inhibitable proximal tubule phosphate transport. *Am J Physiol - Ren Physiol*; DOI: 10.1152/ajprenal.00249.2003.

Bakhiya N, Bahn A, Burckhardt G & Wolff NA (2003). Human organic anion transporter 3 (hOAT3) can operate as an exchanger and mediate secretory urate flux. *Cell Physiol Biochem* **13,** 249–256.

Barry CL, Royle GA & Lake Y (1992). Racial Variation in Serum Uric Acid Concentration in Pregnancy: a Comparison between European, New Zealand Maori and Polynesian Women. *Aust New Zeal J Obstet Gynaecol* **32,** 17–19.

Bobulescu IA & Moe OW (2012). Renal Transport of Uric Acid: Evolving Concepts and Uncertainties. *Adv Chronic Kidney Dis* **19,** 358–371. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3619397/ [Accessed July 31, 2020].

Butler A, Hoffman P, Smibert P, Papalexi E & Satija R (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36,** 411–420.

Chiba T et al. (2015). NPT1/SLC17A1 is a renal urate exporter in humans and its common gain-of-function variant decreases the risk of renal underexcretion gout. *Arthritis Rheumatol (Hoboken, NJ)* **67,** 281–287.

Chiba T, Matsuo H, Nagamori S, Nakayama A, Kawamura Y, Shimizu S, Sakiyama M, Hosoyamada M, Kawai S, Okada R, Hamajima N, Kanai Y & Shinomiya N (2014). Identification of a hypouricemia patient with SLC2A9 R380W, a pathogenic mutation for renal hypouricemia type 2. *Nucleosides Nucleotides Nucleic Acids* **33,** 261–265.

Diamond HS & Paolino JS (1973). Evidence for a postsecretory reabsorptive site for uric acid in man. *J Clin Invest* **52,** 1491–1499.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A & Huber W (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440.

Durinck S, Spellman PT, Birney E & Huber W (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat Protoc* **4,** 1184–1191.

Enomoto A, Kimura H, Chairoungdua A, Shigeta Y, Jutabha P, Cha SH, Hosoyamada M, Takeda M, Sekine T, Igarashi T, Matsuo H, Kikuchi Y, Oda T, Ichida K, Hosoya T, Shimokata K, Niwa T, Kanai Y & Endou H (2002). Molecular identification of a renal urate-anion exchanger that regulates blood urate levels. *Nature* **417,** 447–452.

Hafemeister C & Satija R (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20,** 1–15.

Hagos Y, Krick W, Braulke T, Mühlhausen C, Burckhardt G & Burckhardt BC (2008). Organic anion transporters OAT1 and OAT4 mediate the high affinity transport of glutarate derivatives accumulating in patients with glutaric acidurias. *Pflugers Arch Eur J Physiol* **457,** 223–231.

Hagos Y, Schley G, Schdel J, Krick W, Burckhardt G, Willam C & Burckhardt BC (2012). α-Ketoglutarate-related inhibitors of HIF prolyl hydroxylases are substrates of renal organic anion transporters 1 (OAT1) and 4 (OAT4). *Pflugers Arch* **464,** 367–374.

Hagos Y, Stein D, Ugele B, Burckhardt G & Bahn A (2007). Human renal organic anion transporter 4 operates as an asymmetric urate transporter. *J Am Soc Nephrol* **18,** 430–439.

Hunter JD (2007). Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9,** 99–104.

Hyndman D, Liu S & Miner JN (2016). Urate Handling in the Human Body. *Curr Rheumatol Rep*; DOI: 10.1007/s11926-016-0587-7. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4841844/ [Accessed July 31, 2020].

Ichida K, Hosoyamada M, Hisatome I, Enomoto A, Hikita M, Endou H & Hosoya T (2004). Clinical and Molecular Analysis of Patients with Renal Hypouricemia in Japan-Influence of URAT1 Gene on Urinary Urate Excretion. *J Am Soc Nephrol* **15,** 164–173.

Ichida K, Hosoyamada M, Kimura H, Takeda M, Utsunomiya Y, Hosoya T & Endou H (2003). Urate transport via human PAH transporter hOAT1 and its gene structure. *Kidney Int* **63,** 143–155.

Iharada M, Miyaji T, Fujimoto T, Hiasa M, Anzai N, Omote H & Moriyama Y (2010). Type 1 sodium-dependent phosphate transporter (SLC17A1 protein) is a Cl--dependent urate exporter. *J Biol Chem* **285,** 26107–26113.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Joseph S, Nicolson TJ, Hammons G, Word B, Green-Knox B & Lyn-Cook B (2015). Expression of drug transporters in human kidney: Impact of sex, age, and ethnicity. *Biol Sex Differ* **6,** 1–15.

Julius S, Alderman MH, Beevers G, Dahlöf B, Devereux RB, Douglas JG, Edelman JM, Harris KE, Kjeldsen SE, Nesbitt S, Randall OS & Wright JT (2004). Cardiovascular risk reduction in hypertensive black patients with left ventricular hypertrophy: The life study. *J Am Coll Cardiol* **43,** 1047–1055.

Jutabha P et al. (2010). Human sodium phosphate transporter 4 (hNPT4/SLC17A3) as a common renal secretory pathway for drugs and urate. *J Biol Chem* **285,** 35123–35132.

Kawamura Y, Matsuo H, Chiba T, Nagamori S, Nakayama A, Inoue H, Utsumi Y, Oda T, Nishiyama J, Kanai Y & Shinomiya N (2011). Pathogenic GLUT9 mutations causing renal hypouricemia type 2 (RHUC2). *Nucleosides Nucleotides Nucleic Acids* **30,** 1105–1111.

Koepsell H (2013). The SLC22 family with transporters of organic cations, anions and zwitterions. *Mol Aspects Med* **34,** 413–435.

Kratzer JT, Lanaspa MA, Murphy MN, Cicerchi C, Graves CL, Tipton PA, Ortlund EA, Johnson RJ & Gaucher EA (2014). Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc Natl Acad Sci U S A* **111,** 3763–3768.

Levinson DJ & Sorensen LB (1980). Renal handling of uric acid in normal and gouty subjects: Evidence for a 4-component system. *Ann Rheum Dis* **39,** 173–179.

Liao J, Yu Z, Chen Y, Bao M, Zou C, Zhang H, Liu D, Li T, Zhang Q, Li J, Cheng J & Mo Z (2020). Single-cell RNA sequencing of human kidney. *Sci Data* **7,** 1–9.

Mather A & Pollock C (2011). Glucose handling by the kidney. *Kidney Int Suppl*; DOI: 10.1038/ki.2010.509.

Matsubayashi M et al. (2021). 27-Hydroxycholesterol regulates human SLC22A12 gene expression through estrogen receptor action. *FASEB J* **35,** e21262.

Matsuo H, Chiba T, Nagamori S, Nakayama A, Domoto H, Phetdee K, Wiriyasermkul P, Kikuchi Y, Oda T, Nishiyama J, Nakamura T, Morimoto Y, Kamakura K, Sakurai Y, Nonoyama S, Kanai Y & Shinomiya N (2008). Mutations in glucose transporter 9 gene SLC2A9 cause renal hypouricemia. *Am J Hum Genet* **83,** 744–751.

Mueckler M & Thorens B (2013). The SLC2 (GLUT) family of membrane transporters. *Mol Aspects Med* **34,** 121–138.

Nakanishi T, Ohya K, Shimada S, Anzai N & Tamai I (2013). Functional cooperation of URAT1 (SLC22A12) and URATv1 (SLC2A9) in renal reabsorption of urate. *Nephrol Dial Transplant* **28,** 603–611.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

548 Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, Laurenti E, Wilson NK, Kent DG
549 & Göttgens B (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor
550 cell differentiation. *Blood* **128,** e20–e31.
551 Nigam SK & Bhatnagar V (2018). The systems biology of uric acid transporters: the role of remote
552 sensing and signaling. *Curr Opin Nephrol Hypertens* **27,** 305–313.
553 Reimer RJ (2013). SLC17: A functionally diverse family of organic anion transporters. *Mol Aspects*
554 *Med* **34,** 350–359.
555 Sato M, Mamada H, Anzai N, Shirasaka Y, Nakanishi T & Tamai I (2010). Renal secretion of uric
556 acid by organic anion transporter 2 (OAT2/SLC22A7) in human. *Biol Pharm Bull* **33,** 498–503.
557 Sender R & Milo R (2021). The distribution of cellular turnover in the human body. *Nat Med 2021*
558 *271* **27,** 45–48.
559 Sorensen LB (1965). Role of the intestinal tract in the elimination of uric acid. *Arthritis Rheum* **8,**
560 694–706.
561 Sperling O (2006). Hereditary renal hypouricemia. *Mol Genet Metab* **89,** 14–18.
562 Srivastava S, Nakagawa K, He X, Kimura T, Fukutomi T, Miyauchi S, Sakurai H & Anzai N (2019).
563 Identification of the multivalent PDZ protein PDZK1 as a binding partner of sodium–coupled
564 monocarboxylate transporter SMCT1 (SLC5A8) and SMCT2 (SLC5A12). *J Physiol Sci* **69,**
565 399–408.
566 Steele TH & Rieselbach RE (1967). The renal mechanism for urate homeostasis in normal man. *Am J*
567 *Med* **43,** 868–875.
568 Stegle O, Teichmann SA & Marioni JC (2015). Computational and analytical challenges in single-
569 cell transcriptomics. *Nat Rev Genet* **16,** 133–145. Available at:
570 https://www.nature.com/articles/nrg3833 [Accessed August 15, 2020].
571 Stuart T, Butler A, Hoffman P, Stoeckius M, Smibert P, Satija R, Hafemeister C, Papalexi E, Mauck
572 Iii WM & Hao Y (2019). Comprehensive Integration of Single-Cell Data Resource
573 Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888-1902.e21.
574 Toyoda Y, Kawamura Y, Nakayama A, Morimoto K, Shimizu S, Tanahashi Y, Tamura T, Kondo T,
575 Kato Y, Ichida K, Suzuki H, Shinomiya N, Kobayashi Y, Takada T & Matsuo H (2022).
576 OAT10/SLC22A13 Acts as a Renal Urate Re-Absorber: Clinico-Genetic and Functional
577 Analyses With Pharmacological Impacts. *Front Pharmacol* **13,** 577.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

578 Vallejos CA, Risso D, Scialdone A, Dudoit S & Marioni JC (2017). Normalizing single-cell RNA
579 sequencing data: Challenges and opportunities. *Nat Methods* **14,** 565–571. Available at:
580 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5549838/ [Accessed August 15, 2020].
581 Wickham H (2010). A Layered grammar of graphics. *J Comput Graph Stat* **19,** 3–28.
582 Willemin ME, Van Der Made TK, Pijpers I, Dillen L, Kunze A, Jonkers S, Steemans K, Tuytelaars
583 A, Jacobs F, Monshouwer M, Scotcher D, Rostami-Hodjegan A, Galetin A & Snoeys J (2021).
584 Clinical Investigation on Endogenous Biomarkers to Predict Strong OAT-Mediated Drug-Drug
585 Interactions. *Clin Pharmacokinet* **60,** 1187–1199.
586 Wilson PC, Wu H, Kirita Y, Uchimura K, Ledru N, Rennke HG, Welling PA, Waikar SS &
587 Humphreys BD (2019). The single-cell transcriptomic landscape of early human diabetic
588 nephropathy. *Proc Natl Acad Sci U S A* **116,** 19619–19625.
589 Wu H, Uchimura K, Donnelly EL, Kirita Y, Morris SA & Humphreys BD (2018). Comparative
590 Analysis and Refinement of Human PSC-Derived Kidney Organoid Differentiation with Single-
591 Cell Transcriptomics. *Cell Stem Cell* **23,** 869-881.e8.
592

593 Competing interests

594 EM received joint collaborative funding from Fujiyakuhin, Co. Ltd. EM is a CEO of molmir, Inc.
595

596 Author Contributions

597 **Y.M.S.**: Conceptualization, Formal analysis, Investigation, Software, Writing (original draft), Writing
598 (review and editing). **P.W.**: Conceptualization, Funding acquisition, Writing (original draft), Writing
599 (review and editing). **M.Ma.**: Conceptualization, Formal analysis, Writing (review and editing).
600 **M.Mi.**: Investigation, Software, Writing (review and editing). **N.S.**: Investigation, Software, Writing
601 (review and editing). **Y.S.**: Writing (review and editing). **M.T.**: Writing (review and editing). **K.K.**:
602 Funding acquisition, Writing (review and editing). **K.S.:** Writing (review and editing). **M.E.**: Writing
603 (review and editing). **K.T.**: Writing (review and editing). **H.K.**: Funding acquisition, Writing (review
604 and editing). **S.N.**: Conceptualization, Formal analysis, Funding acquisition, Project administration,
605 Supervision, Writing (original draft), Writing (review and editing). **E.M.**: Conceptualization, Formal
606 analysis, Funding acquisition, Project administration, Supervision, Writing (original draft), Writing
607 (review and editing). All authors contributed to the final manuscript.
608

609 Funding

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

610 This work was supported by grants from JSPS KAKENHI [JP19K07373 and JP22K06150 to P.W.,
611 JP19K23952 to K.K., JP21H03365 to S.N., and JP20H03199 to E.M], AMED [JP23wm0425004 to
612 E.M.], Mochida Memorial Foundation grant for Medical and Pharmaceutical research to P.W., Takeda
613 Science Foundation to E.M., Nara Medical University Grant-in-Aid for Collaborative Research
614 Projects to K.S. and H.K., and Gout and uric acid foundation of Japan to S.N.

615

616 Acknowledgments

617 The authors would like to thank Sotaro Kikuchi for the useful insights gained in group discussions.
618 The authors also thank Keren-Happuch E Fan Fen for her critical reading of the manuscript.

619

620 Keywords

621 transporter, single cell RNA-sequencing, uric acid, human kidney, gout, solute carriers, purine,
622 apicobasal polarity

623 Figures and Tables

624 Figures

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

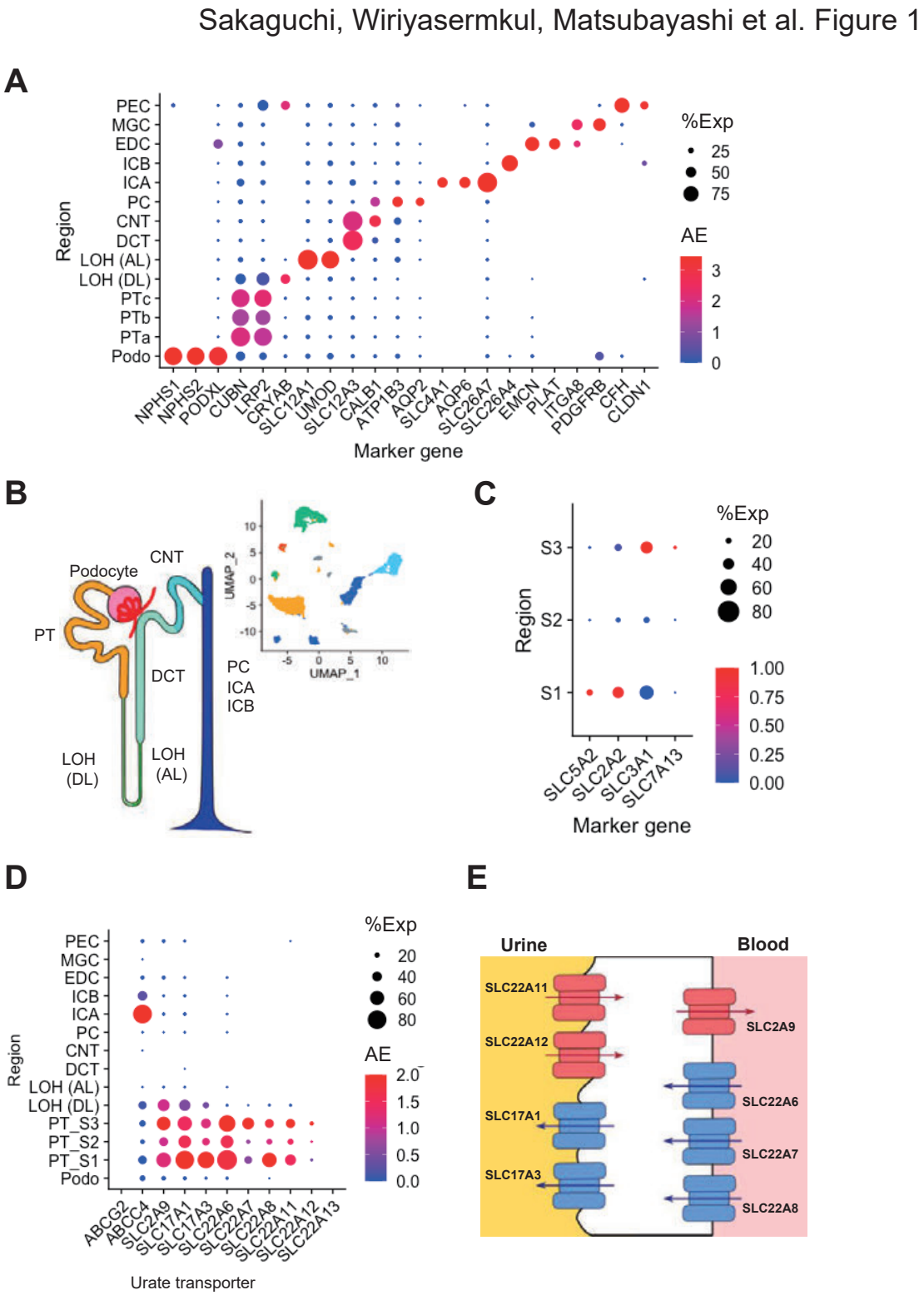
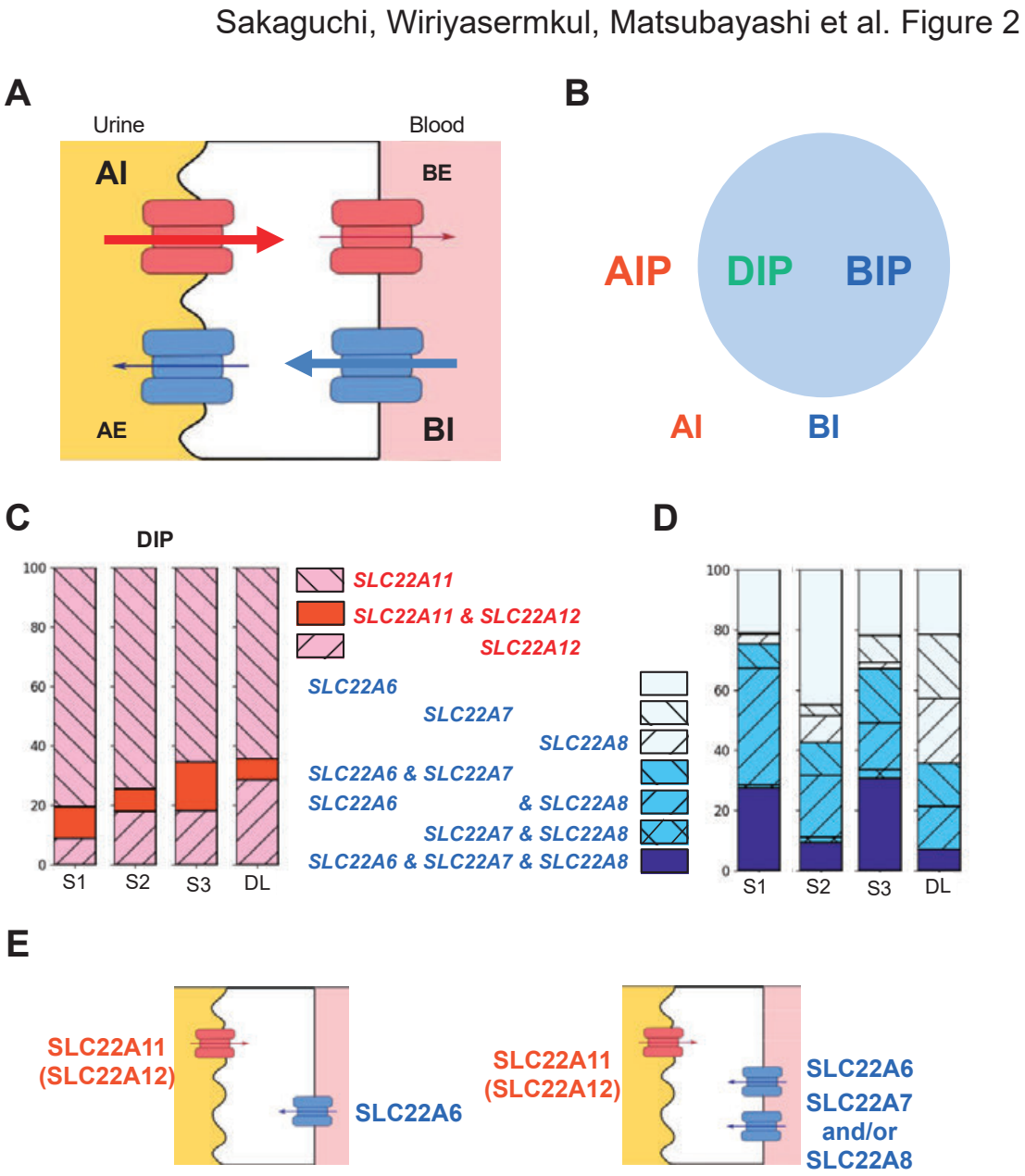


Figure 1. Identification of regions and transporters for urate handling in human kidney (A) Annotation of the 14 clusters by the marker genes. Podo, podocyte; PTa-c, three clusters of proximal

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

628 tubule; LOH (DL), the loop of Henle (descending loop); LOH (AL), the loop of Henle (ascending
629 loop); DCT, distal convoluted tubule; CNT, connecting tubule; PC, principal cell; ICA, intercalated
630 cell type A; ICB, intercalated cell type B; EDC, endothelial cell; MGC, mesangial cell; and PEC,
631 parietal epithelial cell. (B) A schematic diagram of human nephron anatomy and unsupervised
632 clustering of healthy human adult renal cells. (C) The dot-plot of PT marker genes: S1 markers and S3
633 markers. S1-S3, S1-S3 segments of proximal tubules. Dot sizes indicate the gene frequencies (%Exp)
634 and dot colors indicate average expression levels (AE) in the cluster. (D) The dot-plot of the gene
635 expression of urate transporters in the renal regions. (E) A schematic diagram of urate transporters
636 expressed in the PT and DL clusters. The apical membrane is to the left of the cells and the basolateral
637 membrane is to the right. Red symbols represent transporters which are reconstituted to urate
638 reabsorption. Blue symbols represent transporters which are reconstituted to urate secretion. Arrows
639 indicate directions of urate flow.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



640

641 **Figure 2. Classification of transporter types and cell populations for urate handling.** (A) A
642 schematic diagram of four types of urate transporters classified based on their subcellular localization
643 and physiological roles (apical/basolateral and influx/efflux). The eight transporters were classified
644 into one of the following four transporter types: (1) SLC22A11 and SLC22A12 as AI transporters; (2)
645 SLC17A1 and SLC17A3 as AE transporters; (3) SLC22A6, SLC22A7, and SLC22A8 as BI
646 transporters; or (4) SLC2A9 as BE transporter. The cooperation of an Apical Influx (AI) transporter
647 and a Basolateral Efflux (BE) transporter leads to urate reabsorption (red). The cooperation of a

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Basolateral Influx (BI) transporter and an Apical Efflux (AE) transporter leads to urate secretion (blue). (B) Venn diagram indicating cell populations, which were classified based on the positivity of influx transporters. Cell clusters in the PT and DL were classified into one of the following cell populations: AIP cell population, BIP cell population, DIP cell population, or DIN cell population. (C-D) Bar plots indicating the percentage (y-axis) of AI transporters (C) and BI transporters (D) in the DIP cell population across the regions (x-axis) in the DIP cell population. (E) Expression pattern schema of influx transporters. The apical membrane is to the left of the cells and the basolateral membrane is to the right of the cells. Red symbols represent the AI transporters, which are only single kind expressed in single cells. Blue symbols represent the BI transporters, which are single or multiple kind(s) expressed in single cells. Arrows indicate directions of urate flow.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

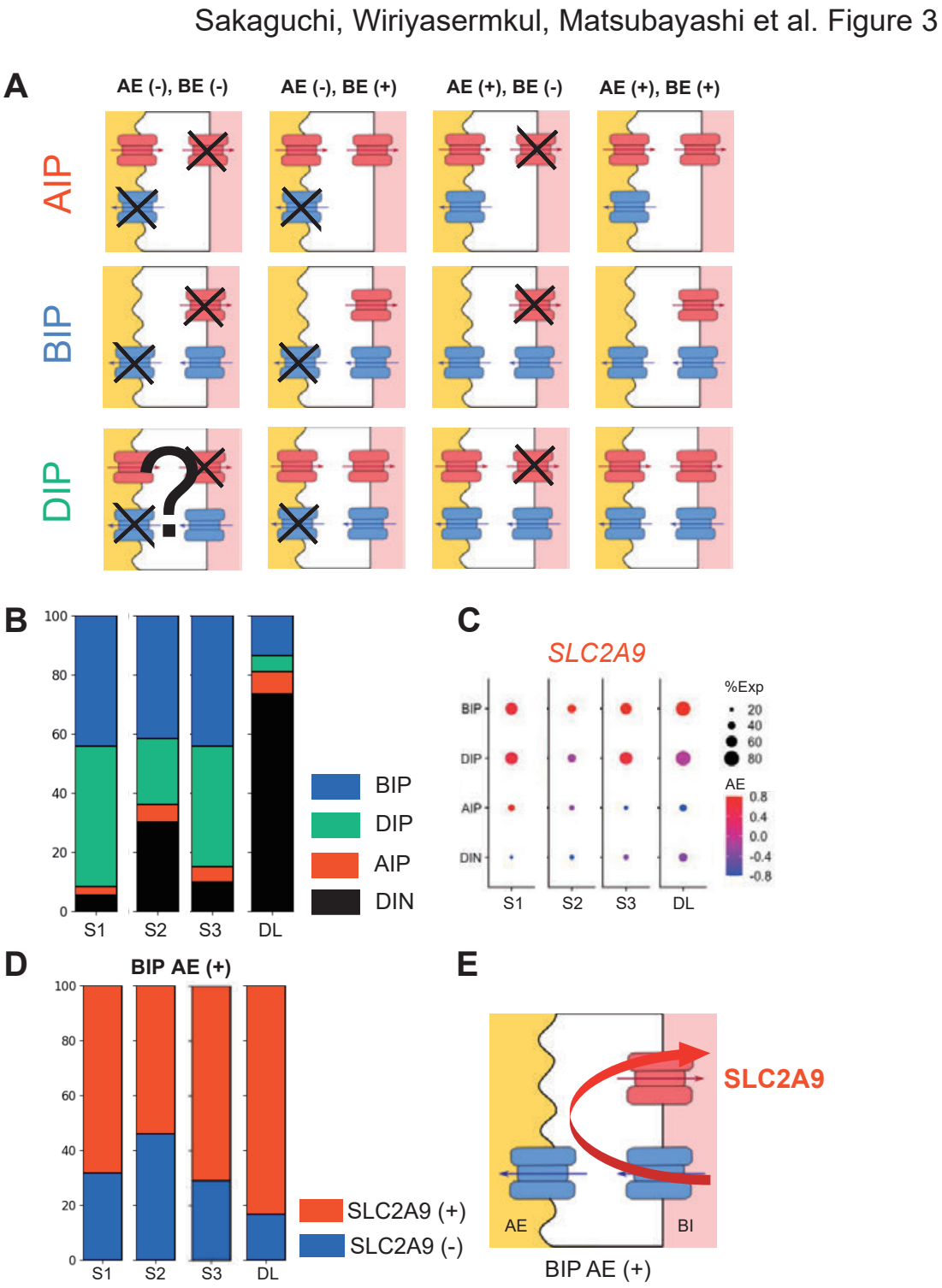


Figure 3. Prediction of cellular urate transport direction from the expression of the transporters

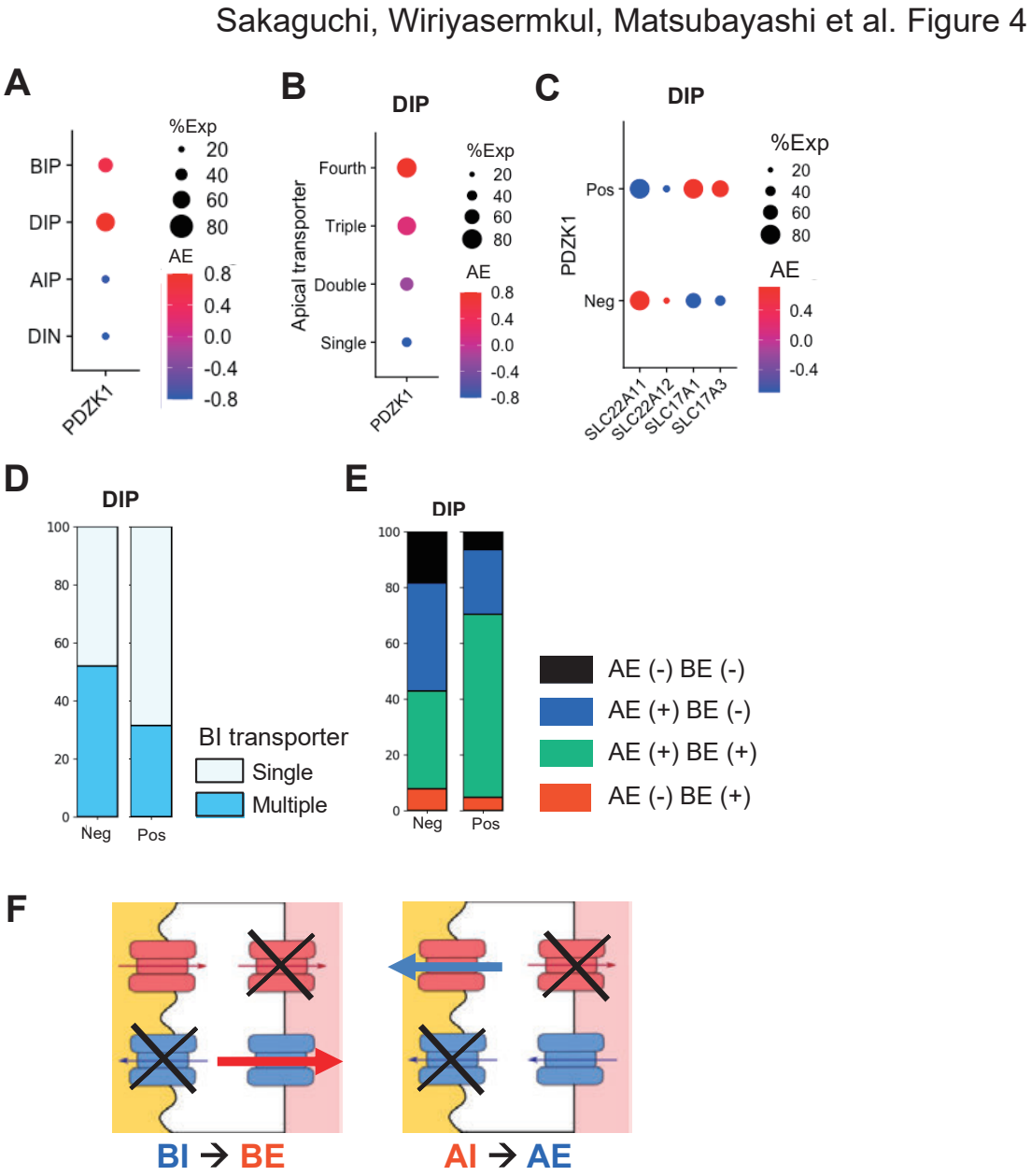
(A) Models indicating all potential expression patterns of urate transporters in each of the following cell populations: AIP cell population, BIP cell population, and DIP cell population. Colors indicate

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

types of transporters: red, AI or BE transporters; blue, BI or AE transporters. Arrows show urate transport directions. (B) Proportion of the cell populations in the regions. The bar plot shows the percentage of cells (y-axis) in each cell population across the PT and LOH (DL) clusters (x-axis). (C) The dot-plot indicates the frequency and expression levels of SLC2A9 (x-axis) across the cell populations (y-axis) along the three PT segments (S1-S3) and DL. Dot sizes refer to the frequency of a molecule expressed in the cell population (%Exp), while dot colors indicate expression levels (AE). (D) Positivity proportion of SLC2A9 in the BIP cell population with the AE transporters. The bar plot shows the percentage of cells (y-axis) in each cell population across the PT and LOH (DL) clusters (x-axis). (E) A schematic diagram of SLC2A9 cell function in the BIP cell population. The apical membrane is to the left of the cells and the basolateral membrane is the right of the cells. Red symbols represent the AI transporters. Blue symbols represent the BI and AE transporters. Arrows indicate directions of urate flow.

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

675



676

677 **Figure 4. Expression profiles of urate transporters and PDZK1 in DIP cell population.** (A)
678 Expression levels and frequencies of P DZK1 in the cell populations. (B) Expression levels and
679 frequencies of PDZK1 in the DIP cell population are rearranged by the numbers of apical membrane
680 transporters. (C) The dot-plot indicates the frequency and expression levels of the apical transporters
681 (x-axis) across the positivity of PDZK1 (y-axis). Dot sizes refer to the frequency of a molecule
682 expressed in the cell population (%Exp), while dot colors indicate the expression levels (AE). (D) Cell

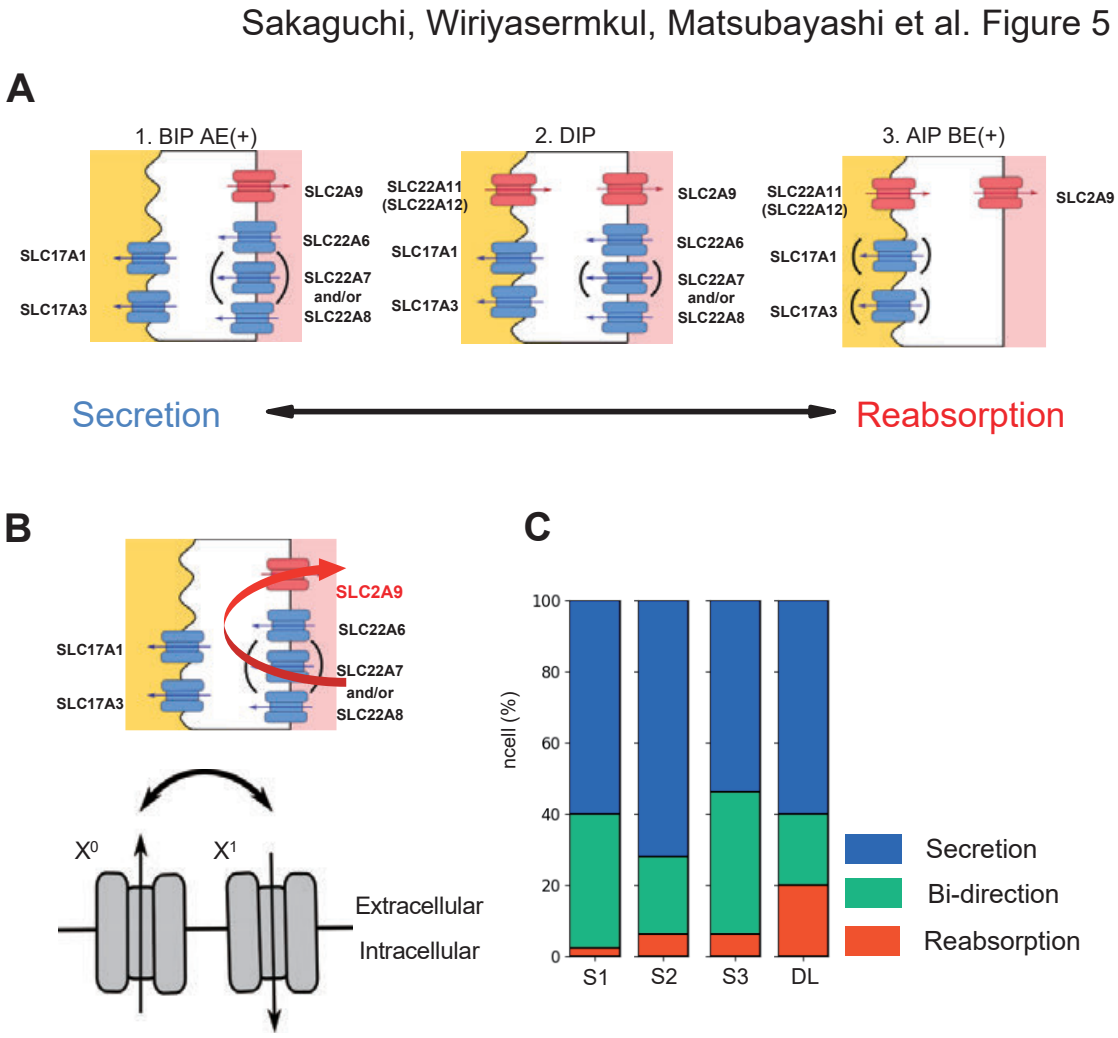
bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

683 proportion of the BI transporter kind numbers. The bar plot shows the percentage of cells (y-axis)
684 across the positivity of PDZK1 (x-axis). (E) Cell proportion by the efflux transporters (AE and BE
685 transporters). The bar plot shows the percentage of cells (y-axis) across the positivity of PDZK1 (x-
686 axis). (F) A schematic diagram of transporter reversibility from influx transporter to efflux transporter
687 in the DIP cell population. the apical membrane is to the left of the cells and the basolateral membrane
688 is the right of the cells. Red symbols represent the AI and BE transporters. Blue symbols represent the
689 BI and AE transporters. Arrows indicate directions of urate flow.

690
691

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

692



693

694 **Figure 5. Major cell populations explaining renal urate transport dynamics.** (A) Representative
695 cell populations that contributed to urate reabsorption and secretion. Colors indicate types of
696 transporters: red, AI or BE transporters; blue, BI or AE transporters. (B) A schematic diagram
697 illustrating the feasibility of the superiority of urate reabsorption over urate secretion. Upper panel:
698 Secretion attenuation of BIP cell population by SLC2A9 at all tubular regions. Lower panel:
699 Reversibility of the transporter function between urate influx and efflux in human kidney. (C) The bar
700 plot showing cell proportion by cellular urate handling in three PT segments and LOH (DL) regions.

701

702

703

bioRxiv preprint doi: https://doi.org/10.1101/2023.06.29.545826; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

Table 1. Transport properties of urate transporters expressed in renal tubular regions of human males in this analysis.

SLC name	Protein name	Function	Type	Transport system*	Substrates*	Av. (PT_S1)	Av. (PT_S2)	Av. (PT_S3)	Av. (LOH (DL))	Km for urate (μM)	Ref
SLC22A11	OAT4	Reabsorption	AI	Exchanger/Facilitate	Organic anions (e.g. NSAIDs, PG, urate)	2.99	3.97	4.52	1.94	3779.6	Hagos et al., 2007
SLC22A12	URAT1	Reabsorption	AI	Exchanger	Urate	0.36	0.92	1.42	0.70	371	Enomoto et al., 2002
SLC2A9	GLUT9	Reabsorption	BE	Facilitate	Urate (glucose, fructose)	4.68	4.53	7.35	5.47	365	Anzai et al., 2008
SLC22A6	OAT1	Secretion	BI	Exchanger	Organic anions (e.g. PAH, PG, urate)	10.25	12.93	12.50	1.50	943	Ichida et al., 2003
SLC22A7	OAT2	Secretion	BI	Exchanger/Facilitate	Organic anions (e.g. PAH, PG, urate)	1.29	1.91	4.74	0.79	1168	Sato et al., 2010
SLC22A8	OAT3	Secretion	BI	Exchanger	Organic anions (e.g. PAH, PG, urate)	3.75	2.96	3.13	1.25	2888	Bakhiya et al., 2003
SLC17A1	NPT1	Secretion	AE	Facilitate (Δψ-driven / Cl-dependent)	Organic anions (e.g. PAH, urate)	10.60	8.54	7.66	5.89	1100	Iharada et al., 2010
SLC17A3	NPT4	Secretion	AE	Facilitate (Δψ-driven)	Organic anions (e.g. PAH, urate)	6.02	3.82	3.56	2.18	Unknown	Jutabha et al., 2010

Av. : Average expression of three functional cell groups, ITR cell group, DP cell group, and ITS cell group, in each cluster.

* : A reference to the SLC table (http://slc.bioparadigms.org/).

Column "Ref" shows the reference lists to refer the K_m values for urate of each urate transporters.

5. Appendix

5-2 シンポジウム資料

5-2-1 キックオフシンポジウム …… 255
5-2-2 中間成果発表会 …… 259
5-2-3 最終成果報告会 …… 281

5-2-1　キックオフシンポジウム

後援：公益財団法人中谷医工計測技術振興財団
技術開発研究助成【長期大型研究】

キックオフシンポジウム

「生体5次元情報」を解読する医工計測技術を創出する「知・もの・人」づくり計画

公立大学法人
奈良県立医科大学

2019年
5月16日 木 13:00–17:10

奈良県立医科大学【大講堂】 奈良県橿原市四条町840
レセプション 17:20-18:20 大学附属病院2階【レストランロイヤル】

奈良県立医科大学 理事長・学長
細井 裕司

(公財)中谷医工計測技術振興財団 事務局長
齋田 馨

奈良県立医科大学 V-iCliniX講座 教授 / ATR 佐藤匠徳特別研究所 所長
佐藤 匠徳 「Virtual Human InformatiX Project: ヒト生体5次元情報ネットワークの計測・解読から拓ける世界」

奈良県立医科大学 腎臓内科学 教授
鶴屋 和彦 「腎臓からみた心腎連関：CKD-MBD」

奈良県立医科大学 循環器内科学 教授
斎藤 能彦 「心腎連関の分子機序」

けいはんなリサーチコンプレックス事業 事務局長
(公財)関西文化学術研究都市推進機構 超快適スマート社会推進室 部長
森田 芳文

MBT研究所 研究教授
梅田 智広

東京医科歯科大学 大学院医歯学総合研究科 疾患多様性遺伝学分野 教授
田中 敏博 「ゲノム医学・ゲノム医療」

(株)国際電気通信基礎技術研究所 佐藤匠徳特別研究所 主任研究員
浦山 恭次 「The Body-wide Transcriptome Landscape of Disease Models」

奈良県立医科大学 口腔外科学 教授
桐田 忠昭 「唾液腺の臨床医学」

奈良県立医科大学 発生・再生医学 教授
栗本 一基 「単一細胞遺伝子発現解析の展望」

奈良県立医科大学 副学長
車谷 典男

問合せ先：tns-sec@atr.jp / 0774-95-2330
責任者：奈良県立医科大学 V-iCliniX講座 教授(寄附講座) 佐藤 匠徳
主催：奈良県立医科大学・けいはんなリサーチコンプレックス事業・MBT研究所

開会	13:00〜13:10　開会の挨拶　奈良県立医科大学　理事長・学長　細井 裕司 13:10〜13:20　中谷財団紹介 (公財)中谷医工計測技術振興財団　事務局長　寶田 馨
研究概要紹介	座長:細井 裕司（奈良県立医科大学　理事長・学長 ） Talk45分、Q&A5分 13:20〜14:10 奈良県立医科大学 V-iCliniX講座 教授 ATR 佐藤匠徳特別研究所 所長　佐藤 匠徳
休憩	14:10〜14:15
Session1	座長:嶋 緑倫（奈良県立医科大学小児科 教授） Talk 15分、Q&A 5分 14:15〜14:35　奈良県立医科大学　腎臓内科学　教授　鶴屋 和彦 14:35〜14:55　奈良県立医科大学　循環器内科学　教授　斎藤 能彦
休憩	14:55〜15:00
けいはんなRC MBT研究所	座長:佐藤 匠徳（奈良県立医科大学 V-iCliniX講座 教授 ATR 佐藤匠徳特別研究所 所長） Talk15分 15:00〜15:15　けいはんなRC事業紹介 (公財)関西文化学術研究都市推進機構 超快適スマート社会推進室　部長　森田 芳文 15:15〜15:30　MBT紹介　MBT研究所　研究教授　梅田 智広
Session2	座長:佐藤 匠徳（奈良県立医科大学 V-iCliniX講座 教授 ATR 佐藤匠徳特別研究所 所長） Talk25分、Q&A5分 15:30〜16:00　東京医科歯科大学 大学院医歯学総合研究科 疾患多様性遺伝学分野　教授　田中 敏博 Talk 10分、Q&A5分 16:00〜16:15　ATR 佐藤匠徳特別研究所　主任研究員　浦山 恭次
休憩	16:15〜16:20
Session3	座長:吉栖 正典（奈良県立医科大学 薬理学講座 教授） Talk 15分、Q&A 5分 16:20〜16:40　奈良県立医科大学　口腔外科学　教授　桐田 忠昭 16:40〜17:00　奈良県立医科大学　発生・再生医学　教授　栗本 一基
閉会	17:00〜17:10 閉会の挨拶　奈良県立医科大学　副学長　車谷 典男
レセプション	17:20〜18:20　場所:大学附属病院2階【レストランロイヤル】

5-2-2　中間成果発表会

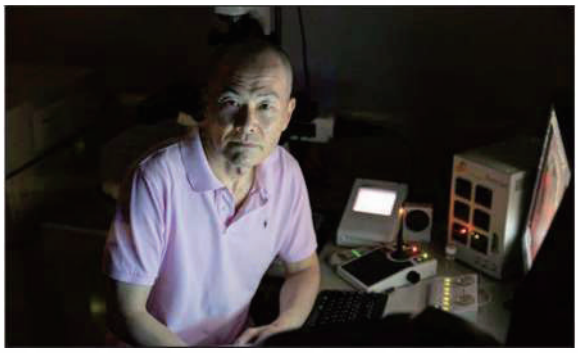


PROGRAM & ABSTRACT

The 2nd International Symposium on Human InformatiX

Sensing, Modeling and Controlling Omics in Biomedicine

October 24, 2022



Welcome to the 2nd International Symposium on Human InformatiX

The Darwinism, arguably the only theory in Biology, was proposed in the mid-19th century. The genetic code was cracked in the mid-20th century. The human genome was decoded at the beginning of the 21st century.

In the 21st century, Biology and Medicine is rapidly transforming themselves at an unprecedented pace – mapping all 37 trillion cells, all signaling pathways, all genes and proteins in human to generate the google map of the human body. These endeavors are all to ultimately crack the biological principles of the human and gain insights into "WHAT IS LIFE".

Furthermore, scientific community and system is transforming.
"Old boys club" "Imperial big labs" are all falling down, getting replaced by virtual transparent lab consortium.

So, in this age, it is becoming absolutely essential to integrate all scientific disciplines such as mathematics, physics, chemistry, engineering, informatics and biology to decode the fundamental principles of life and to understand us, The Human, and THIS IS becoming realistic.

To this end, we organize this meeting, Human InformatiX.

This is the 2nd of the series – the first one was held in February, 2020 in Kyoto, Japan, After the 2-years hiatus due to the COVID-19 pandemic, we are now organizing the 2nd of the series.

The 1st meeting focused on mathematical and systems biology areas and this 2nd will focus on "sensing, modeling and controlling" aspects of the functions and behaviors of the molecules, cells and organs inside the human body. To this end we are bring in engineering scientists to join biology, medical, informatics and mathematics scientists as our speakers.

Hence, as the keynote speaker, we are pleased to welcome Professor Cynthia Reinhart-King. She is the President of the Biomedical Engineering Society, USA. She is also Cornelius Vanderbilt Professor of Engineering, Professor of Biomedical Engineering and Director of Graduate Studies in Biomedical Engineering at Vanderbilt University in USA.

The meeting is organized and funded by one of the major non-profit agencies funding biomedical engineering: Nakatani Foundation (https://www.nakatani-foundation.jp/en/).

We welcome participation from broad and diverse areas of scientific and medical disciplines and invite you to join us in virtual space or Osaka, Japan. We look forward to seeing you in October, 2022!



Thomas N. Sato, Ph.D.
The Chair of the organizing committee

Director, The Thomas N, Sato BioMEC-X Laboratories,
Advanced Telecommunications Research Institute International (ATR)
Professor, Virtual Human InformatiX Clinic, Contribution Course, Nara Medical University
Director, Grant Program for Biomedical Engineering Research (Nakatani Foundation)

Contents

Speakers List	03
Program	04
Keynote	05
Speakers	06
Blitz talk	17

Speakers List

Keynote Speaker:

• Cynthia Reinhart-King, Ph.D.

Speakers:

• Thomas N. Sato, Ph.D.

•Kyoji Urayama, Ph.D.

• Satoshi Kozawa, Ph.D.

• Kengo Tejima, Ph.D.

• Mari Nakanishi, Ph.D.

• Eiichiro Mori, M.D./Ph.D.

• Atsushi Maruyama, Ph.D.

• Takashi Tokuda, Ph.D.

• Rei Narikawa, Ph.D.

• Yoshihiro Maruyama, D.Phil.

Blitz Talk by Medical Students:

• Hirona Yokoyama

• Hotaka Doi

• Miyu Terashima

PROGRAM

9:00-9:10 Opening Remarks	**Nobuhiro Matsumori** Secretary General, Nakatani Foundation
9:10-10:00 Keynote	**Cynthia Reinhart-King, Ph.D.** Professor, Biomedical Engineering, Vanderbilt University, USA President, The Biomedical Engineering Society, USA
10:00-10:15	**Thomas N. Sato, Ph.D.** Director, The Thomas N. Sato BioMEC-X Laboratories, ATR Professor, V-iCliniX, Nara Medical University
10:15-10:25	**Kyoji Urayama, Ph.D.** Senior Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
10:25-10:35	**Satoshi Kozawa, Ph.D.** Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
10:35-10:45	**Kengo Tejima, Ph.D.** Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
10:45-10:55	**Mari Nakanishi, Ph.D.[1] and Kengo Tejima, Ph.D.** [1]Assistant Professor, V-iCliniX, Nara Medical University [1]Collaborating Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
10:55-11:05	**Break**
11:05-11:15	**Eiichiro Mori, M.D./Ph.D.** Associate Professor, Nara Medical University Collaborating Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
11:15-11:25	**Atsushi Maruyama, Ph.D.** Professor, Tokyo Institute of Technology Visiting Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
11:25-11:35	**Takashi Tokuda, Ph.D.** Professor, Tokyo Institute of Technology Visiting Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
11:35-11:45	**Rei Narikawa, Ph.D.** Associate Professor, Tokyo Metropolitan University Collaborating Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
11:45-12:00	**Yoshihiro Maruyama, D.Phil.** Senior Lecturer, Australian National University, Australia Visiting Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR
12:05-12:20	**Hirona Yokoyama** Medical Student, Nara Medical University Student Intern, The Thomas N. Sato BioMEC-X Laboratories, ATR Student Intern, Karydo TherapeutiX, Inc. **Hotaka Doi** Medical Student, Nara Medical University Student Intern, The Thomas N. Sato BioMEC-X Laboratories, ATR Student Intern, Karydo TherapeutiX, Inc. **Miyu Terashima** Medical Student, Medical School, Okayama University Student Intern, The Thomas N. Sato BioMEC-X Laboratories, ATR Student Intern, Karydo TherapeutiX, Inc.
12:20-12:25 Closing Remarks	**Thomas N. Sato, Ph.D.**



The Role of Cellular-Level Forces in Human Health

Cynthia Reinhart-King, Ph.D.

Professor, Biomedical Engineering, Vanderbilt University, USA
President, The Biomedical Engineering Society, USA

【Personal Introduction】
Cynthia Reinhart-King is the Cornelius Vanderbilt Professor and Senior Associate Dean for Research in Engineering at Vanderbilt University. She obtained undergraduate degrees in Chemical Engineering and Biology at MIT and her PhD at the University of Pennsylvania. Her lab's interests are in cell mechanics in cancer and atherosclerosis. She has won numerous awards for research, teaching, and service. She is President-elect of BMES.

【Biography】

2008	Assistant Professor, Cornell University
2013	Associate Professor with Tenure, Cornell University
2017	Cornelius Vanderbilt Professor, Vanderbilt University
2022	co-Director, Center on Mechanobiology, Vanderbilt University
2022	Senior Associate Dean for Research, Vanderbilt School of Engineering

【Abstract】
The organization of molecules into cells, cells into tissues, and tissues into organs requires coordination between both chemical and mechanical signals in the body. Perturbations to those signals can cause large, long range disruptions that can result in numerous diseases, including cancer and atherosclerosis. Cells can exert forces against both adjacent cells and their surrounding environment, and those forces contribute to both the health of the cells and the organization of the tissue. In this talk, I will discuss my lab's work studying the role of cellular forces in dictating and maintaining the structure and function of tissues. I will discuss the importance of cellular forces in maintaining blood vessel structure and integrity and the role of forces in cancer metastasis. Given the importance of force in biology and health, the field of mechanomedicine is emerging, where mechanical cues, forces, and transducers are therapeutically targeted to prevent or treat disease. I will discuss approaches toward targeting mechanical cues within cells and tissues, and the connection of mechanobiology with other rapidly growing research areas including metabolomics and transcriptomics.

MEMO



Introduction to Virtual Human InformatiX Clinics (V-iCliniX) Project

Thomas N. Sato, Ph.D.

Director, The Thomas N. Sato BioMEC-X Laboratories, ATR
Professor, V-iCliniX, Nara Medical University

【Personal Introduction】
Tom is an established scientist who has been listed among the top 1% most cited authors for journals in Molecular Biology and Genetics. He made a seminal-discovery of vascular endothelial-specific receptors, Tie1 and Tie2. In a series of collaborations with Regeneron scientists, he uncovered physiological functions of angiopoietins, ligands for Tie2. Most recently, he is studying inter-organ communications at the systems and organismal levels and their roles in human diseases. He is also an accomplished triathlete at the Ironman distance.

【Biography】
EDUCATION:

1985	B.S. (Biology, Cellular Biochemistry Program) University of Tsukuba, Tsukuba, Japan
1987	Cold Spring Harbor Summer Course, "Learning and Memory", Cold Spring Harbor Lab, NY
1988	Ph.D. (Biology, Neurobiology Program) Georgetown University, Washington, DC

PROFESSIONAL EXPERIENCE:

1989-1990	Postdoctoral Fellow with M.C. Wilson The Scripps Research Institute La Jolla, CA 92037
1991-1995	Assistant Member (Assistant Professor Level) Roche Institute of Molecular Biology Nutley, NJ 07110
1995-1997	Assistant Professor of Medicine Harvard Medical School Cardiovascular Division Beth Israel Deaconess Medical Center Boston, MA 02215
1997-2002	Associate Professor of Medicine and Molecular Biology (Tenured in 1999) Integrative Biology Graduate Program UT Southwestern Medical Center at Dallas Dallas, TX 75390-8573
2003-2004	Professor of Medicine and Molecular Biology Integrative Biology Graduate Program UT Southwestern Medical Center at Dallas Dallas, TX 75390-8573
2005-2009	Professor of Cell and Developmental Biology Weill Cornell Medicine New York, NY 10021
2006-2009	Joseph C. Hinsey Professor of Cell and Developmental Biology Weill Cornell Medicine New York, NY 10021
2007-2008	Member, Nanobiotechnology Center Cornell University Ithaca, NY 14853
2008-2019	Adjunct Professor Department of Biomedical Engineering Cornell University Ithaca, NY 14853
2009-2014	Professor Graduate School of Biological Sciences Nara Institute of Science and Technology (NAIST) Ikoma, Nara, Japan
2009-Present	Affiliate Professor Centenary Institute Sydney, Australia
2013-2014	Visiting Scientist Advanced Telecommunications Research Institute International (ATR) Kyoto, Japan
2013-2020	Research Director ERATO SATO Live Bio-Forecasting Project Japan Science Technology Agency (JST) Kyoto, Japan
2014-Present	Director The Thomas N. Sato BioMEC-X Laboratories Advanced Telecommunications Research Institute International (ATR) Kyoto, Japan
2015-Present	Scientific Founder and Chair, Board of Directors Karydo TherapeutiX, Inc. Tokyo, Japan
2019-Present	Professor V-iCliniX (Virtual Human InformatiX Clinics) Nara Medical University Nara, Japan

【Abstract】
The V-iCliniX is a major interdisciplinary scientific project challenging one of the hardest biomedical goals: The precision human digital-twins system

The project is funded by Nakatani Foundation, the sponsor of this symposium, and several other Japanese government agencies such ATLA, JST, AMED, JSPS, etc.

The precision human digital-twins system allows for monitoring and preventing/treating nano to micro to macro level dysfunctions inside the individual human bodies at anywhere and anytime.

This system consists of three modules and their network:
1) Sensor nano-machines that monitors molecular, cellular and organ dysfunctions inside the human body at real-time.
2) Regulatory nano-machines that provides therapeutics to prevent/treat the dysfunctions identified by the above sensor nano-machines inside the body.
3) The real-time multi-dimensional computational model of the individuals across all scales (i.e., molecular, cellular, organ, organismal levels) at the meta-space.
4) The network of the above three modules.

The project team consists of scientists who are the experts across diverse scientific disciplines across biology, medicine, informatics, mathematics, chemistry, engineering, electronics, etc.

As the introduction to the presentations by each of the scientists in the project, I will briefly summarize each of their works and also describe how we integrate such diverse scientific works into the one-team effort all towards the realization of “the precision human digital-twins system”.

MEMO



Salivary proline-rich low complexity proteins and phosphorus homeostasis

Kyoji Urayama, Ph.D.

The Thomas N. Sato BioMEC-X Laboratories,
Advanced Telecommunications Research Institute International (ATR)

【Present Affiliation】
1. Senior Researcher,
The Thomas N. Sato BioMEC-X Laboratories, ATR, Japan
2. Scientist, Karydo TherapeutiX, Inc., Japan

【Personal Introduction】
My research backgrounds are pharmacology, cardiology, molecular and cellular biology, and phenotype analysis of genetically mutated mice. Currently I am interested in and engaged in the studies combining experimental and bioinformatics approaches to decipher multi-organ crosstalk mechanisms in health and diseases.

【Biography】
2005-2008 Ph.D. Louis Pasteur University Strasbourg 1, Strasbourg, France
2009-2011 Post-doctoral fellow, University of Strasbourg, Strasbourg, France
2011-2014 Research Assistant Professor, NAIST, Nara, Japan
2014-Present Senior Researcher, ATR, Kyoto, Japan
2019-Present Scientist, Karydo TherapeutiX, Inc., Kyoto, Japan

【Abstract】
The salivary glands (SGs) secrete saliva and preserve oral hygiene and facilitate food digestion. In addition to these canonical functions, their important roles in the body-wide health maintenance has recently been implicated. However, this latter non-canonical function remains enigmatic.

In the course of conducting the body-wide transcriptome analysis of the mouse, we identified a family of proteins that are specifically expressed in the SGs across 24 different organs analyzed. They are all consists of low-complexity (LC) sequences of 25 – 40% prolines and repetitive sequences rich (70 – 80%) in proline, glutamine and glycine. Furthermore, these salivary proline-rich low complexity proteins (sPRPs) reside on the same genomic locus in a tandem sequence, and this structural and genomic organization is highly conserved among all vertebrate species, implicating their essential function in health.

To investigate the function of sPRPs, we generated the mouse models that are deficient in each or combinations of sPRPs genes by targeted gene-editing. The body-wide phenotype analyses found their critical role in maintaining phosphorus homeostasis, a physiological mechanism that controls functions of various organs such bone and kidney, and also aging.

In my presentation, I will show the result of the mouse phenotype analyses and discuss their roles in health with a particular emphasis on the SGs-regulated remote-organ functions.

MEMO



Latent disease similarities and therapeutic repurposing possibilities uncovered by multi-modal generative topic modeling of human diseases

Satoshi Kozawa, Ph.D.

Researcher,
The Thomas N. Sato BioMEC-X Laboratories, ATR

【Present Affiliation】
1. The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International
2. Karydo TherapeutiX, Inc.

【Personal Introduction】
I studied statistical machine learning, mainly Bayesian statistics, and applied it to biological problems in graduate school. I am currently working on biological/medical problems using various information science approaches.

【Biography】

2009-2014	M.S. and Ph.D., NARA Institute of Science and Technology, Japan
2014-Present	Reseacher, Advanced Telecommunications Research Institute International, Japan
2017-Present	Scientist, Karydo TherapeutiX, Inc., Japan

【Abstract】
Human diseases are characterized by multiple features such as their pathophysiological, molecular, and genetic changes. The rapid expansion of such multi-modal disease-omics space provides an opportunity to re-classify diverse human diseases and uncover their latent molecular similarities, which could be exploited to repurpose a therapeutic-target for one disease to another.

Herein, we probed this underexplored space by soft-clustering 6,955 human diseases by multi-modal generative topic-modeling. Focusing on chronic kidney disease and myocardial infarction, two most life-threatening diseases, unveiled are their previously underrecognized molecular similarities to neoplasia and mental/neurological-disorders, and 69 repurposable therapeutic-targets for these diseases.

In this presentation, I will explain birief summary of this topic-modeling method and discuss the results.

MEMO



Towards the identification of molecular/cellular/organ check-points of organismal life and death

Kengo Tejima, Ph.D.

The Thomas N. Sato BioMEC-X Laboratories (TNSL),
Advanced Telecommunications Research Institute International (ATR)

【Present Affiliation】
1. The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International, Japan
2. Karydo TherapeutiX, Inc., Japan

【Personal Introduction】
I studied molecular biology, especially mycological biofilms, in graduate school. The biofilm study got me fascinated by the complex but sophisticated cellular systems. During the course of this graduate study, I have developed further interest in biological networks such as inter-organ/cellular gene networks. Hence, I am currently engaged in the study to decode these complex networks, using both molecular biological and computational methods.

【Biography】

2014-2018	M.S. and Ph.D. in Science, School of Life Science and Technology, Tokyo Tech, JPN
2014-2018	Education Academy of Computational Life Sciences, Tokyo Tech, JPN
2016	Visiting Student, National Heart and Lung Institute, Imperial College London, UK
2018-Present	Researcher, TNSL, Advanced Telecommunications Research Institute International, JPN
2019-Present	Scientist, Karydo TherapeutiX, Inc., JPN

【Abstract】
All organisms die and the understanding of the death mechanism facilitates our deeper understanding of life, ultimately addressing “What is life?”, an ultimatum of all living system raised by Erwin SchrÖdinger in 1944. Due to such far-reaching importance in biology, molecular mechanisms of cell death have been studied for decades, providing the comprehensive molecular pictures of the cell death mechanisms such as apoptosis, necrosis, ferroptosis, etc. However, the death mechanism at the organismal level of multicellular organisms remains elusive.

What is the sequence of events at the molecular, cellular, and organ levels leading to the ultimate death of organisms? Is there a critical core of events that are decisive for the irreversibility of the death process? More specifically, is there an organ, cells, and/or molecules of which dysfunction propagate across whole body to determine the death of the organism? More practically, could we delay, prevent, reverse the death process of the whole organism or even organs? Furthermore, a philosophical question could be that “what is the definition of death, and yet life?”.

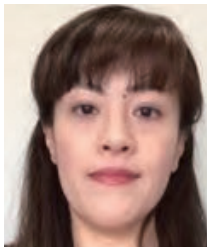
In this presentation, I present our approach towards these unfathomably complex questions.

Our central hypothesis is that there are multiple molecular, cellular, and organ levels of check-points of which dysfunction results in the sequential irreversibility of the organismal death.

To identify such “death/life check-points”, we first aim to obtain the comprehensive gene expression landscapes across the whole body over the course of organismal death. In this study, we analysed two organismal death models of zebrafish: 1) the death by the perturbation of the circulation, 2) the death by starvation. The comprehensive gene expression is analysed by the whole-body RNA-seq and single cell RNA-seq methods.

In my presentation, I will present the first draft of organismal death atlas and several informatics analyses of the atlas and discuss how we test our central hypothesis and identify the putative “death/life check-points”.

MEMO



Salivary proline-rich low complexity proteins in human disease

Mari Nakanishi, Ph.D.[1] and Kengo Tejima, Ph.D.[2]

[1] V-iCliniX, Nara Medical University, [2] The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International (ATR)

【Present Affiliation】
1. Assistant Professor, V-iCliniX, Nara Medical University, Japan
2. Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR, Japan, and Scientist, Karydo TherapeutiX, Inc., Japan

【Personal Introduction】
I received my M.S. in Science. After the graduation, I conducted researches on disease association of HLA genes in the Department of Legal Medicine. During this study, I was interested in the diversity of the HLA genes. Recently, I have been studing the relationship of disease polymorphism and molecular structure.

▶ **Kengo Tejima** See his other presentation.

【Biography】

-2018	Research Student, Department of Legal Medicine, Nara Medical University, Japan
2018	Ph.D. Nara Medical University, Japan
2018-2020	Post-doctoral fellow, Department of Future Basic Medicine, Nara Medical University, Japan
2020-Present	Assistant Professor, V-iCliniX, Nara Medical University, Japan

▶ **Kengo Tejima** See his other presentation.

【Abstract】
Recently, we discovered a unique family of salivary glands-specific proline-rich proteins (sPRPs) of low complexity sequences that appears to control the body-wide homeostasis on the basis of mouse knockout studies (see the presentation by Kyoji Urayama). To investigate the relevance of this finding to human diseases, we conducted clinical studies.

We recruited approximately 700 patients of renal and/or cardiovascular diseases and 1,000 healthy subjects. We collected both saliva and serum from the subjects and analyzed the concentration of one of the sPRPs, hPRB1.

In our presentation, we will describe various clinical data that we collected and machine-learing methods that we applied to these data to evaluate the possible usefulness of hPRB1 as a biomarker for human diseases.

MEMO



Recognition of cross-β structure

Eiichiro Mori, M.D., Ph.D.

Associate Professor, Nara Medical University
Collaborating Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR

【Present Affiliation】
Department of Future Basic Medicine and V-iCliniX Laboratory, Nara Medical University, Japan

【Personal Introduction】
Biological phase separation is considered as a major driving force of membrane-less organelles in cells. Intrinsically disordered proteins with low-complexity sequence (LC domains) form labile cross-β structures to drive phase separation, and these polymeric conformations are transient but specific. We are trying to understand how cross-β polymers of LC domains are recognized and regulated.

【Biography】

2003-2011	Nara Medical University, Japan (2003-2009 Medical Student; 2009-2011 Resident)
2011-2017	Postdoctoral Fellow, University of Texas Southwestern Medical Center, USA
2017-2018	Assistant Professor, Nara Medical University, Japan
2019-Present	Associate Professor, Nara Medical University, Japan
2022-Present	CEO, molmir, Inc., Japan

【Abstract】
Proteins are folded based on primary amino acid sequences following Anfinsen's dogma, but proteins with low-complexity in amino acid sequence (LC domains) are intrinsically disorederd. Although LC domains (LCDs) are recognized as biologically important regions, usually found in DNA/RNA binding proteins, their functional understanding from structural aspects has been challenging. LCDs are major drivers of membrane-less organelles such as nucleoli and RNA granules, and this phenomenon is called phase separation in the context of physical chemistry. Genetic analysis of neurological diosorders offers series of disease-related amino acid substitutions compromize physiologically optimized phase separation, leading to protein aggregations. Protein aggregates in neurological disoreders are fibrils enriched in β-sheets, called cross-β polymers. Studies of solid-state nuclear magnetic resonance (NMR) and cyro-electron microscopy combined with protein chemistry revealed that cross-β structures of LCDs are fundamental structural unit of phase separation. However, how cells recognize these LCD cross-β structures remains unknown. Here we show that Zn-finger domains (ZnF) recognize cross-β structures of LCDs. ZnF with two cysteins and two histidines (C2H2-type) are mostly found in transcription factors and recognize DNA. Solution NMR analysis reveals that C2H2-ZnF binds LCDs with different regions from DNA binding site. Moreover, non-polymeric mutant LCDs are failed to interact with C2H2-ZnF, suggesting that ZnF recognize cross-β polymeric form but not monomeric form of LCDs. These results offer a model that DNA binding ZnF recognize cross-β structures of LCDs to mediate weak and transient but specific interaction and transduce signals in phase-separated compartments.

MEMO



Design of artificial chaperones for nanobiotechnology

Atsushi Maruyama, Ph.D.

Professor, Tokyo Institute of Technology
Visiting Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR

【Present Affiliation】
School of Life Science and Technology, Tokyo Institute of Technology

【Personal Introduction】
Research interests: Nanobioscience and technology, Biofunctional polymers, Artificial chaperones, Stimuli-responsive polymers
Recent publications: Smart protein refolding system based on UCST-type ureido polymers, *Biomacromolecules*, 2022, doi: 10.1021/acs.biomac.2c00694, Cationic copolymer-augmented DNA hybridization chain reaction, *ACS Appl. Mater. Interfaces*, 2022, doi: 10.1021/acsami.2c11548, Label-free detection of HPV mRNA with artificial chaperone-enhanced MNAzyme (ACEzyme)-based electrochemical sensor, *Biosens. Bioelectron.*, 2022, doi: 10.1016/j.bios.2022.114352, Light-regulated liquid-liquid phase separation for spatiotemporal protein recruitment and cell aggregation, *ACS Appl. Mater. Interfaces*, 2021, doi: 10.1021/acsami.0c22314

【Biography】

1986	Assistant Professor, Sophia University
1990	Assistant Professor, Tokyo Institute of Technology
2000	Associate Professor, Tokyo Institute of Technology
2004	Professor, Kyushu University
2013-Present	Professor, Tokyo Institute of Technology

【Abstract】
Biological macromolecules must fold into native structures to gain functional activities. In living cells, proteins called molecular chaperones mediate productive folding by preventing undesired interactions and aggregation and by facilitating refolding of misfolded macromolecules into their bioactive forms. Inspired by natural molecular chaperones, artificial chaperones that mimic some features of their biological counterparts have been designed. We have designed artificial chaperones employing synthetic polymers and evaluated their promising applications in enhancing macromolecular assembly of proteins, polypeptides, and nucleic acids.

Cationic copolymers as nucleic acid chaperones
We synthesized a polymer consisting of a polycationic backbone and a high density of hydrophilic graft chains, poly (l-lysine)-graft-dextran (PLL-g-Dex). The cationic copolymer forms soluble interpolyelectrolyte complexes with negatively charged oligonucleotides without changing the conformation of the DNA. The copolymer shields the electrostatic repulsion between negatively charged DNA strands to facilitate DNA hybridization and enhance thermal stability of DNA structures. The copolymer also accelerates the DNA strand displacement reaction, in which a complementary DNA strand binds to a partially or fully hybridized DNA complex displacing a pre-hybridized strand. The ability to catalyze strand displacement is a feature of natural nucleic acid chaperones such as NCp7. The chaperoning activity of the copolymer was applied to improve DNA-nanodevices such as DNA nanomachines, DNA computers, and DNA enzymes.

Cationic copolymers as peptide chaperones
Peptides composed of short linear chains of amino acids and functions as highly selective and effective signaling molecules. Increasing number of peptide-based drugs had been approved. In general, peptides have short plasma half-lives, poor physical and chemical stability, and tend to aggregate, limiting their use as therapeutics. These limitations have been mitigated by several strategies including conjugation, rational peptide drug design, and chemical modifications. Artificial chaperones have the potential to improve efficacies of peptide-based therapeutics. We designed cationic copolymer, PAA-g-Dex, to chaperone peptides having opposite ionic charge. The E5 peptide is an amphiphilic, anionic peptide derived from hemagglutinin peptide of influenza virus. The protonation of glutamic acid residues under acidic conditions converts E5 from an inactive random coil to an active α-helix structure that exhibits membrane-disruptive activity. The protonation, however, increases hydrophobicity and causes aggregation and precipitation, and insufficient solubility limits the bioactivity of E5 and its application. The addition of PAA-g-Dex, having a cationic poly(allylamine) backbone and dextran graft chains, increased solubility, promotes coil-to-helix transformation, and significantly enhances membrane disruptive activity of E5 peptide under acidic and neutral pH.
Recently, we prepared smart protein refolding system based on temperature-responsive polymers. Upper critical solution temperature (UCST)-type ureido polymers undergo liquid–liquid phase separation (LLPS) upon cooling rather than the liquid–solid phase transition of the typical lower critical solution temperature (LCST)-type polymers. Hydrophobized ureido polymers assisted refolding of proteins. The refolded protein was readily separated from the polymers by cooling.

MEMO



Milli-machines / Implantable Devices in Biomedicine

Takashi Tokuda, Ph.D.

Professor, Tokyo Institute of Technology
Visiting Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR

【Present Affiliation】
Tokyo Institute of Technology (Professor)
The Thomas N.Sato BioMEC-X Lab. ATR (Visiting Researcher)

【Personal Introduction】
Takashi Tokuda received his B.E. M.E., and Ph.D degrees from Kyoto University, Japan, in 1993, 1995, and 1998. He was an Assistant Professor during 1999–2008, as well as an Associate Professor during 2008–2019 in Nara Institute of Science and Technology. Since 2019, he has been a Professor at Tokyo Institute of Technology. His research interests include CMOS image sensors, implantable devices, wireless technology, and IoT.

【Biography】

1993	Faculty of Electronics, Kyoto University
1995	Master Program, Graduate School of Electronics, Kyoto University
1998	Ph. D. Program, Graduate School of Electronics, Kyoto University
1998	Postdoctral Research Fellow, Japan Society of Promotion of Science.
1999-2008	Assistant Professor, Nara Institute of Science and Technology
2008-2019	Associate Professor, Nara Institute of Science and Technology
2019-	Professor, Tokyo Institute of Technology

【Abstract】
Ultra-small, implantable electronic systems have been attracting interest and expectation for new kinds of medical diagnosys / therapeutic technology. As the most intelligent multifunctional electronic platform, CMOS LSI technolgy plays the core role in the implantable device technologies. We have been devloping various kinds of implantable devices such as implantable image sensor, implantable optical and electric neural stimulator etc. In collaboration with Thomas N. Sato BioMEC-X Lab., we are developing an implantable electronic device called as Milli-machines, that can handle light and electricity, to interact with various kinds of functional molecular-based systems called as Nano-machins.
The Milli-machines are expected to have a cylindlical shape that can be implanted into living bodies with limited invasion. It is battery-less or equipped with ultra-small battery. A micro-controller chip, light source LED and photodiode (PD) are expected to implement onto the Milli-machines. As an interaction hub in the digital-twin system, the safety and operational reliability is most important. Adopting a industrial, well-validated microcontoller is expected to fulfil the requirements.
We desgined and fabricated an non-implantable functional varification device. Currently, the miniaturized fluorescence measrement function is under evalation. The current device configuration is compatible with UV – Visible excitation and Visible detection. After the basic proof of concept was completed, it is expected to shrink to implantable device format. Electric interfacing will be also implemented for both electrochemical measurements. It can be also applied for heater-based drug / Nano-machines delivery. The peripheral building blocks such as interface electrodes or heaters will be develped.

MEMO



Development of bio-imaging and optogenetic tools based on the cyanobacterial photoreceptors

Rei Narikawa, Ph.D.

Associate Professor, Tokyo Metropolitan University
Collaborating Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR

【Present Affiliation】
Department of Biological Sciences, Graduate School of Science, Tokyo Metropolitan University

【Personal Introduction】
Based on the photobiology, I would like to understand the light-responsive strategy of the photosynthetic microorganisms, especially cyanobacteria, in the basic science, and link to the applied science. I love scientific discussion with beers.

【Biography】

1997	The University of Tokyo, College of Arts and Sciences
2001-2006	The University of Tokyo, Graduate School of Arts and Sciences
2007	Assistant Professor, The University of Tokyo, Graduate School of Arts and Sciences
2014	Lecturer, Shizuoka University, Graduate School of Science and Technology
2021-Present	Associate Professor,Tokyo Metropolitan University, Graduate School of Science

【Abstract】
Photosynthetic organisms utilize light as an energy source, and so developed highly organized photoresponsive systems to utilize light as information. In the case of the oxygen-evolving prokaryotes, cyanobacteria, linear tetrapyrrole-binding cyanobacteriochrome photoreceptors play central roles in regulating various light acclimation processes such as phototaxis and acclimation of photosynthetic apparatuses. We have focused on the chromophore-binding GAF domains of the cyanobacteriochrome photoreceptors, and identified various photoreceptors widely covering UV and visible regions (Fushimi & Narikawa (2019) *Curr. Opin. Struct. Biol.*, 57, 39-46). In the initial phase of our study, we could determine crystal structures of the typical cyanobacteriochrome GAF domains (Narikawa et al. (2013) *Proc. Natl. Acad. Sci. U.S.A.*, 110, 918-923). Further, we recently succeeded in modifying binding chromophore species and color tuning mechanisms to create photoreceptors to sense various light qualities. In this lecture, after briefly summarizing the natural diversity of the cyanobacteriochromes, I will introduce three engineering works. One focused on isomerization of the chromophore from initially incorporated one and several color tuning mechanisms. In this work, we have succeeded in creating seven artificial photoreceptors based on the single cyanobacteriochrome GAF domain (Fushimi et al. (2020) *Proc. Natl. Acad. Sci. U.S.A.*, 117, 15573-15580). One of the engineered molecules is applicable to photoregulation of cAMP synthesis. The second one focused on mammalian intrinsic chromophore, biliverdin. In this work, we succeeded in the rational conversion of the cyanobacteriochrome GAF domains to accept the biliverdin chromophore and in determining the crystal structure of the biliverdin-binding cyanobacteriochrome GAF domain, enabling the application to the mammalian in vivo imaging (Fushimi et al. (2019) *Proc. Natl. Acad. Sci. U.S.A.*, 116, 8301-8309). We are further trying to perform site-directed random mutagenesis to obtain molecules with better biliverdin incorporation and better expression yield than the background molecule. I will introduce the serendipitous discovery of an interesting mutation in this ongoing study (Suzuki et al. Submitted). As a final topic, I will also introduce challenges to design circular permutated cyanobacteriochromes, which would open the door for various bio-imaging and optogenetic tool development.

MEMO



Category Theory for Neural Networks

Yoshihiro Maruyama, D.Phil.

Senior Lecturer, Australian National University, Australia
Visiting Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR

【Present Affiliation】
Senior Lecturer, Australian National University, Australia
Visiting Researcher, The Thomas N. Sato BioMEC-X Laboratories, ATR

【Personal Introduction】
I am a Senior Lecturer at the School of Computing of the Australian National University. Prior to this, I held Hakubi Assistant Professorship at Kyoto University. Prior to that, I studied for a PhD at the University of Oxford. I work on formal logic, category theory (esp. categorical duality) and foundations of artificial intelligence and cognitive science.

【Biography】

2011-2017	DPhil in Computer Science, University of Oxford, UK
2015-2019	Assistant Professor, The Hakubi Centre for Advanced Research, Kyoto University, Japan
2019-2021	Lecturer, School of Computing, Australian National University, Australia
2022-Present	Senior Lecturer, School of Computing, Australian National University, Australia

【Abstract】
Category theory is a new mathematical language that emerged as an alternative to set theory in foundations of mathematics and has recently been actively applied in various fields of science. Traditional applications of category theory in computer science include those in semantics of programming languages, and more recently, it has been applied in artificial intelligence and machine learning as well. Applications of category theory in the life sciences, however, have been limited so far; put another way, they are frontiers of applied category theory. In this talk we shall explore applications of category theory in symbolic logical representations of neural networks (which exemplifies the category-theoretical integration of symbolic and statistical AI).

MEMO



Prediction of therapeutic indications and side effects for new drug targets in silico: Edit-distance based pathway-classifier

Hirona Yokoyama

Medical Student, Nara Medical University
Student Intern, The Thomas N. Sato BioMEC-X Laboratories, ATR
Student Intern, Karydo TherapeutiX, Inc.

【Methods】
For the prediction of therapeutic indications and side effects for new drug targets, machine learning models were developed and trained by using the edit-distance based similarity metric between biological pathways, and their known therapeutic indications or side-effects. This classifier uses the edit-distance, specifically Levenshtein distance, to measure the similarity metric between the pathways. In this method, the similarities of all possible pathways of the target candidates to each clinically-approved therapeutic pathway for each disease were computed. The binary classifier of therapeutic indications and side effects was developed using LightGBM.

【Results】
Herein, I examined the prediction performance of the method with the 176 serious adverse outcomes and the 1,168 therapeutic indications by hold-out cross-validation. The result shows F1 scores>0.5 for 124 out of the 176 side effects and for 616 out of the 1168 therapeutic indications, indicating the effectiveness and the usefulness of the method.

【Abstract】
Drug development is essential for medical treatment, but is a severely labor intensive and cost-ineffective process. To overcome this challenge, I developed a biological pathway based-classifier algorithm that enables to predict potential therapeutic targets and side-effects elicited by any molecules when they are therapeutically targeted. In this presentation, I report its algorithm and show the effectiveness of the method as determined by the hold-out cross-validation.

MEMO



Single cell transcriptome analysis

Hotaka Doi

Medical Student, Nara Medical University
Student Intern, The Thomas N. Sato BioMEC-X Laboratories, ATR
Student Intern, Karydo TherapeutiX, Inc.

【Abstract】
Single-cell transcriptome analysis based on next-generation sequencing platforms was first published in 2009. Since then, this technology has revolutionized transcriptome research by yielding vast amounts of molecular information of biology at the single cell resolution. This information enables the identification of rare cell populations, gene regulatory networks, and cellular heterogeneity. Furthermore, they provide unprecedented insights into the pathogenesis of diseases and their mechanisms. A project called “Human Cell Atlas”, which aims to map all 37 trillion cells in the human body, is also underway. Single-cell RNA sequencing (scRNA-seq) is of great importance in the study of biology. I will describe my efforts to work with scRNA-seq data.

MEMO



Analysis of cell type composition in gastric cancer biopsies by scRNA-sequencing

Miyu Terashima

Medical Student, Year 3, Medical School, Okayama University
Student Intern, The Thomas N. Sato BioMEC-X Laboratories, ATR
Student Intern, Karydo TherapeutiX, Inc.

【Abstract】
Single-cell RNA sequencing (scRNA-seq) is a powerful tool that can be used to discover features such as population heterogeneity, monoallelic gene expressions, and splicing patterns of various tissues. Since the first scRNA-seq study was published in 2009, it has fueled much discovery and innovation in medicine over recent years to identify molecular features associated with the development of complex diseases including cancer.

The aim of this analysis was to establish a single-cell atlas using scRNA-seq for cells from gastric antral mucosa biopsies of patients in varying progressions of gastric premalignant lesions and early gastric cancer using a publicly available database.

All scRNA-seq computational analysis was conducted in R (Version 4.1.2) and were performed with the methods implemented in Seurat, a graph-based clustering software suite. Gene expression data was obtained from GEO (Zhang et al 2019; Accession number GSE134520), where all low-quality cells were excluded through filtering. After the removal of unwanted cells, the data was normalized and scaled in order to remove uninteresting sources of variation. Finally, dimensionality reduction, unsupervised cell clustering, and annotation of clusters were conducted to identify cell types present in the gastric biopsies.

A total of 22,910 genetic features were identified across the database of 32,331 high quality cells. By applying the Seurat suite to these datasets, 38 main cell clusters were identified based on their expression patterns of known cell markers. These results show that the gastric cancer biopsies consist mainly of two cell-types, epithelial and non-epithelial cells. Epithelial cells include pit mucous cells marked as TFF1 and MUC5AC, antral basal gland mucous cells marked as TFF2 and MUC6, enteroendocrine cells marked by CHGA and CHGB, proliferative cells marked by MKI67 and BRIC5, cancer cells marked by CEACAM5 and CEACAM6, and chief cells marked by PGA3 and PGA4. Non-epithelial cells mainly consist of B cells as marked by CD79A, T cells marked by CD2 and CD3D, macrophages as marked by CSF1R and CD68, fibroblasts as marked by DCN and PDPN, smooth muscle cells as marked by ACTA2, endothelial cells marked as VWF and ENG, and mast cells marked by TPSAB1.

Early detection of gastric cancer is important in clinical practice, and knowledge of representative cell types in gastric microenvironments can provide us with detailed information of insight into molecular markers that could be used for early diagnosis. However, there are certain limitations that need to be considered when conducting scRNA-seq. In the process of selecting relevant significant components and annotating cell clusters, there are steps that require subjective parameters set by the user. Samples obtained were limited and may not accurately reflect the development of cells in GC, and thus, more samples from various longitudinal databases should also be considered in future computational analyses.

MEMO

5-2-3　最終成果報告会

奈良県立医科大学
V-iCliniX講座

最終成果報告会

日時
2024年2月16日(金)
10：00–16：40

場所
奈良県立医科大学
厳橿会館 3階
（ハイブリッド開催）

※QRコードからのご登録で、
どなたでもご参加いただけます。

問合せ
奈良県立医科大学　V-iCliniX講座
中西　nmari@naramed-u.ac.jp

SPEAKERS

第一部 10：00-12：20

中谷財団 代表理事 家次 恒
V-iCliniX講座 佐藤 匠徳

（座長：口腔外科 桐田 忠昭）
ATR/V-iCliniX　浦山 恭次
ATR/V-iCliniX　小沢 哲
ATR/V-iCliniX　手島 健吾

（座長：腎臓内科 鶴屋 和彦）
腎臓内科　江里口 雅裕

（座長：V-iCliniX 佐藤 匠徳）
東京工業大学　徳田 崇
東京都立大学　成川 礼
物質･材料研究機構　土井 康太郎

第二部 13：20-16：40

V-iCliniX　森 英一朗

（座長：V-iCliniX 森 英一朗）
学長　細井 裕司
産業技術総合研究所　冨田 峻介
徳島大学　齋尾 智英

（座長：脳神経内科 杉江 和馬）
脳神経内科　井口 直彦

（座長：解剖学第二 和中 明生）
解剖学第二　田中 達英

（座長：血栓止血分子病態学 嶋 緑倫）
血液内科　松本 雅則

2024.1.10　差替版

第一部 10:00-12:20

10:00
開会挨拶
（公財）中谷医工計測技術振興財団：代表理事　家次 恒

V-iCliniX講座挨拶
奈良県立医科大学 V-iCliniX 佐藤 匠徳

（座長：口腔外科 桐田 忠昭）

10:10～10:20
ATR / V-iCliniX　浦山 恭次
「Salivary proteins of low complexity sequences for distant organ-homeostasis」

10:20～10:30
ATR / V-iCliniX　小沢 哲
「Physics-informed neural network (PINNs) のバイオインフォマティクスへの応用」

10:30～10:40
ATR / V-iCliniX　手島 健吾
「On the death check-points of the multi-organ system」

（座長：腎臓内科 鶴屋 和彦）

10:40～10:50
奈良県立医科大学 腎臓内科 江里口 雅裕
「腎臓病における唾液腺-骨連関」

（座長：V-iCliniX 佐藤 匠徳）

10:50～11:20
東京工業大学 徳田 崇
「多角的生体センシング・メディケーションのための生体埋込デバイス」

11:20～11:50
東京都立大学 成川 礼
「哺乳類生体内で利用可能なリン酸バイオセンサーの開発」

11:50～12:20
物質・材料研究機構 土井 康太郎
「電気化学的手法を用いたリン酸イオンセンサの開発」

第二部 13:20-16:40

13:20～13:30 and Beyond 概要説明
奈良県立医科大学 V-iCliniX 森 英一朗
「分子の動きをミルことによる創薬」

（座長：V-iCliniX 森 英一朗）

13:30～13:40
奈良県立医科大学 学長 細井 裕司
「軟骨伝導聴覚とMBT（Medicine-Based Town）」

13:40～14:10
産業技術総合研究所 冨田 峻介
「光パターンを基にバイオサンプルの複雑性を解読する分析技術「Chemical tongue」」

14:10～14:40
徳島大学 齋尾 智英
「生命を駆動する動的相互作用の機序解明」

14:40～14:55 休憩

（座長：脳神経内科 杉江 和馬）

14:55～15:25
奈良県立医科大学 脳神経内科 井口 直彦
「ジンクフィンガードメインはLCドメインポリマーを認識する」

（座長：解剖学第二 和中 明生）

15:25～15:55
奈良県立医科大学 解剖学第二 田中 達英
「神経免疫連関による痛覚制御の新しいメカニズム」

（座長：血栓止血分子病態学 嶋 緑倫）

15:55～16:25
奈良県立医科大学 血液内科 松本 雅則
「地方公立単科大学での抗体製剤の開発ー血栓症研究から出血症治療薬」

16:25～16:40 閉会挨拶
奈良県立医科大学 V-iCliniX 佐藤 匠徳

2024.1.10 差替版

奈良県立医科大学
V-iCliniX講座

最終成果報告会

日時

2024年2月16日(金)
10：00–16：40

場所

奈良県立医科大学　厳橿会館 3階
（ハイブリッド開催）

アブストラクト

問合せ
奈良県立医科大学　V-iCliniX講座
中西　nmari@naramed-u.ac.jp

第一部

10:00～10:10

開会挨拶
公益財団法人　中谷医工計測技術振興財団
代表理事　家次　恒

V-iCliniX講座挨拶
奈良県立医科大学　V-iCliniX講座
教授　佐藤 匠徳

第一部

座長：奈良県立医科大学　口腔外科　教授　桐田 忠昭

10:10～10:20

株式会社国際電気通信基礎技術研究所（ATR）　佐藤匠徳特別研究所　主任研究員
Karydo TherapeutiX株式会社
奈良県立医科大学　V-iCliniX講座　招聘助教
浦山 恭次

演題：

Salivary proteins of low complexity sequences for distant organ-homeostasis

Kyoji Urayama[1,2,3,4], Kengo Tejima[1,2,3,4], Shunki Takagi[1, 2], Mari Nakanishi[4], Yosuke Nakagawa[5], Masahiro Eriguchi[6], Makoto Watanabe[7], Takaaki Kosugi[6], Tadaaki Kirita[5], Kazuhiko Tsuruya[6], Yoshihiko Saito[7], Makoto Kuro-o[8], Thomas N. Sato[1,2,3,4]

1 Karydo TherapeutiX, Inc., Kyoto, Japan; 2 The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan; 3 ERATO Sato-Live Bio-Forecasting Project, Kyoto, Japan; 4 V-iCliniX Laboratory, 5 Department of Oral and Maxillofacial Surgery, 6 Department of Nephrology, 7 Department of Cardiovascular Medicine, Nara Medical University (NMU), Nara, Japan; 8 Division of Anti-aging Medicine and Center for Molecular Medicine, Jichi Medical University, Tochigi, Japan;

Abstract

Distant organs influence each other via inter-organ cross-talks. However, the salivary-glands function locally by secreting saliva for oral-hygiene/lubrication and food-digestion. Herein, we challenge this dogma. We find that the salivary-glands secrete a family of evolutionarily-conserved proteins of proline-rich low-complexity sequences (LCS), S-PRPs (Salivary Proline-Rich-Proteins), into saliva and blood. Using mouse models, we demonstrate that blood S-PRPs form unusually stable high-molecular-weight-complex and titrate the bone phosphate-levels in a diet-dependent manner. This S-PRPs function is also bolstered by a negative-feedback loop established with FGF23. Moreover, comprehensive multi-organ transcriptome analyses indicate their additional functions beyond mineral-homeostasis. In human, one of S-PRPs, PRB1, is associated with chronic kidney disease, a disease linked to mineral-disorders. Hence, our findings unveil a previously unappreciated “endocrine-like” function of the salivary-glands and provide a mechanistic insight into their roles in the whole-body scale well-being. Furthermore, our results illustrate a physiological function of “secreted”, rather than typical intra-cellular, LCS proteins.

第一部

座長：奈良県立医科大学　口腔外科　教授　桐田 忠昭

10:20～10:30

株式会社国際電気通信基礎技術研究所（ATR）　佐藤匠徳特別研究所　研究員
Karydo TherapeutiX株式会社
奈良県立医科大学　V-iCliniX講座　招聘助教
小沢 哲

演題：

Physics-informed neural network (PINNs) のバイオインフォマティクスへの応用

生物学的・医学的な現象は微分方程式の系で記述されることが多い。微分方程式の系は一般的には解析解が存在せず、解の数値情報が欲しい場合は何らかの方法で近似解のシミュレーションを行う必要がある。ルンゲクッタ法等の既存のシミュレーション手法でも十分に高精度な近似解を導くことができるが、解くべき問題の式・変数の数が多い場合や計算精度を高くしようとする場合、多大な計算コストが必要となってしまう。

Physics-informed neural network (PINN) は深層学習のフレームワークを利用し、微分方程式の系の近似解を推定する手法である。通常の深層学習のフレームワークは与えられたデータの入出力関係をうまく再現できるようにニューラルネットワークを学習させるが、PINNはデータを原理的には必要とせず、与えられた式をうまく再現できるようにニューラルネットワークを学習させる。PINNはニューラルネットワークの学習こそ計算コストがかかるものの、一度学習が完了してしまえば高速にシミュレーション結果を推定することが可能となる。

第一部

座長：奈良県立医科大学　口腔外科　教授　桐田 忠昭

10:30～10:40

株式会社国際電気通信基礎技術研究所（ATR）　佐藤匠徳特別研究所　研究員
Karydo TherapeutiX株式会社
奈良県立医科大学　V-iCliniX講座　招聘助教
手島 健吾

演題：

On the death check-points of the multi-organ system

Kengo Tejima, Kyoji Urayama, Thomas N. Sato

Karydo TherapeutiX, Inc., Kyoto, Japan; ERATO Sato Live Bio-Forecasting Project, Kyoto, Japan; The Thomas N. Sato BioMEC-X Laboratories, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan; V-iCliniX Laboratory, Nara Medical University, Nara, Japan

Abstract

Slowing and/or reversing dying-process extend life. However, how vertebrates including human die is a question of unfathomable complexity. Furthermore, organismal-scale revitalization from death remains unrealized. Zebrafish, despite being a vertebrate, possesses extraordinary regenerative capacities. It also survives initial organogenesis without the cardiovascular system. Hence, we hypothesized that zebrafish could recover from prolonged no-heartbeat back to heart-beating/swimming. Here, we show that zebrafish indeed has this ability and developed a reversible death-model. Using this model, we compared the body-wide single-cell transcriptome-landscapes between the reversible and irreversible death-states, and identified a unique class of cells, “zombie”-cells, that emerge and increase in their number during their transition. Zombie-cells are also found in another death-model induced by starvation, suggesting their generality and death-specificity, despite their distinct cellular origins in two death-models as inferred by bioinformatics analyses. In addition to the cellular dying-signatures, we found the changes of transcriptome-landscapes during the dying-process, and a reverse-genetics approach identified esm1 vascular permeability-regulator and slc16a9a transporter genes, of which mutations slow the dying-process, suggesting their targeting to extend life. Our results illustrate the utilities of these zebrafish models to study dying-process, and also the “death-atlas” herein provides a comprehensive resource to gain further insights into death-mechanism.

第一部

座長：奈良県立医科大学　腎臓内科　教授　鶴屋 和彦

10:40～10:50

奈良県立医科大学　腎臓内科　講師
江里口 雅裕

演題：
腎臓病における唾液腺-骨連関

１奈良県立医科大学腎臓内科学，2国際電気通信基礎技術研究所，3奈良県立医科大学V-iCliniX，4奈良県立医科大学未来基礎医学
江里口 雅裕[1]，玉城 裕之[1]，孤杉 公啓[1]，浦山 恭次[2,3]，中西 真理[3,4]，森 英一郎[3,4]，佐藤 匠徳[2,3]，鶴屋 和彦[1]

高リン負荷をかけた腎臓病モデルマウスにおける全臓器のbulk RNA sequencingにより、破骨細胞にある破骨細胞関連受容体（OSCAR）と、唾液腺から特異的に分泌されるカルシウムとリンの担体タンパク質である唾液中Proline-rich proteins（S-PRPs）が、リン代謝に関与する新たな因子として同定されました。OSCARは破骨細胞の分化に必須のNFATc1の活性化やアポトーシスの阻害などが報告されています。しかし、OSCARと慢性腎臓病に伴う骨ミネラル代謝異常（CKD-MBD）との関係を調べた研究はありません。

我々は、佐藤先生よりOSCAR KOマウスをいただき、慢性腎臓病マウスモデルにおいて、高リン負荷におけるリン代謝に対するOSCARの役割を検討いたしました。

OSCARは骨より分泌されるリン利尿ホルモンであるFGF-23を介して腎臓からの尿中リン排泄量を調節していました。また、OSCARは耳下腺におけるS-PRPsおよびリン輸送体のmRNA発現レベルを調節していました。OSCARは、腎臓、唾液腺、骨との相互作用を通じてリン恒常性を調節していることがわかりました。

OSCARのリン代謝に関する知見は現在研究中ですが、中間報告させていただきます。

第一部

座長：奈良県立医科大学 V-iCliniX講座 教授　佐藤 匠徳

10:50～11:20

東京工業大学　科学技術創成研究院　未来産業技術研究所　教授
徳田 崇

演題：
多角的生体センシング・メディケーションのための生体埋込デバイス

各種のバイオセンシングや薬剤供給などのハブとして機能することを目指す、生体埋め込み型デバイスについてアーキテクチャを提案し、基礎実証を行った。
侵襲性低減の観点からは、生体埋め込みデバイスのサイズはできるだけ小さいことが求められるが、本研究においては概念実証デバイスとして直径10mm程度、長さ50mmを目標サイズとした。
アーキテクチャとしてはBLE通信機能を備えたマイクロコントローラをコアとし、これに増幅回路・スイッチ等を備えた生体内エレクトロニクスステーションの位置づけとなるデバイスを研究開発した。
試作したデバイスでは、生体計測・薬剤放出等に利用できるマルチチャンネル電流・電圧源機能および電流・電圧計測機能を登載しており、充電式小型バッテリで駆動可能である。また電流源機能によりLEDを駆動することが可能であり、蛍光計測のための励起光やイメージングのための生体内光源、さらには光遺伝学を援用した神経刺激を実現することができる。
また温度応答型の薬剤保持機構を組み合わせることで、薬剤放出システムを構築することにも対応する。本研究においては当該システムを試作し、基礎的な出力・計測機能を評価した上で、リン酸濃度計測によるin vitro機能実証を行った。

第一部

座長：奈良県立医科大学 V-iCliniX講座 教授　佐藤 匠徳

11:20～11:50

東京都立大学　大学院理学研究科生命科学専攻　准教授
成川 礼

演題：
哺乳類生体内で利用可能なリン酸バイオセンサーの開発

生命はさまざまな環境変化に応じて、自身の細胞や体を作り変えることで、環境に応答する高度な機構を備えている。このような機構において、環境シグナルを感知するタンパク質が重要な役割を担う。これらの受容体タンパク質は、バイオセンサーや制御マシンを構築する上で、開発土台の分子として用いることができる。我々の研究では、バクテリア由来のリン酸の受容体タンパク質と光受容体タンパク質を用いて、リン酸センサーや光制御マシンの構築を進めている。リン酸センサーの構築においては、ヒトの細胞内濃度をモニター可能な分子を構築するためのエンジニアリングを進めており、光制御マシンの構築においては、多様な光質に応答する分子の構築と多様なアウトプットを制御する系の構築を進めている。どちらの開発においても、キメラ分子の作出と変異の導入という2つの実験系を組み合わせて分子を改変している。本講演では、主にリン酸バイオセンサーの開発について発表しつつ、光制御マシンの開発についても簡単に紹介したい。

第一部

座長：奈良県立医科大学 V-iCliniX講座 教授　佐藤 匠徳

11:50～12:20

国立研究開発機構物質・材料研究機構　構造材料研究センター　主任研究員
土井 康太郎

演題：
電気化学的手法を用いたリン酸イオンセンサの開発

土井康太郎*1、廣本祥子*1、堤祐介*2
*1 国立研究開発法人物質・材料研究機構 構造材料研究センター 耐食材料グループ
*2 国立研究開発法人物質・材料研究機構 構造材料研究センター 腐食研究グループ

老化や疾病などの身体的変化により生体内のリン酸イオン濃度が上昇することが近年明らかになってきた。すなわち、生体内のリン酸イオン濃度の変化をいち早く得ることができれば、より早い治療が期待できるということである。そこで我々は、金属ジルコニウム（Zr）を電極としたリン酸イオンセンサの開発を行った。リン酸イオンセンサの原理を示す。金属基板を溶液に浸漬させると、一定時間後に金属表面と生成する化合物は平衡状態に達する。平衡状態にある金属表面はある浸漬電位を示すが、この浸漬電位はネルンストの式より溶液中の物質のイオン濃度によって決定される。本研究では、金属基板にZrを用いてリン酸イオンを含む溶液中でZr表面に生成するリン酸ジルコニウムとZrの平衡状態の変化、すなわち浸漬電位の変化から溶液中のリン酸イオン濃度変化を測定するセンサの開発を目的とした。Zrは生体適合性・耐食性に優れ、表面に電気化学反応を阻害するリン酸カルシウムを形成しないため、リン酸イオンのセンシングに有効である。表面を安定させるため、予めリン酸イオンを含む溶液中で前処理を施したZrとリン酸イオンの反応は可逆性を示し、1-10 mMのリン酸イオン濃度の範囲でよい電位応答を示した。また、塩化物イオン変化やpH変化、温度変化などの浸漬電位を変化させ得る因子に関しても、生体内での変化範囲ではリン酸イオンセンサに及ぼす影響は小さいことが分かった。

第二部

13:20～13:30

奈良県立医科大学　V-iCliniX講座　准教授
森 英一朗

演題：
分子の働きをミルことによる創薬

分子の可視化は、分子の機能理解には欠かせない。これまでのタンパク質の構造解析は、X線結晶構造解析やクライオ電子顕微鏡などの技術により進められてきた。2020年代に入りAIによる構造予測の精度が急激に上昇し、予測構造をベースとした計算機上での創薬開発が、現実のものとなってきた。一方で、タンパク質の中には安定な構造を有する領域とそうでない領域が存在しており、これらの領域はその特性から天然変性領域や天然変性タンパク質（IDR/IDP：intrinsically disordered region/protein）と呼ばれる。IDRの存在は2000年頃に注目されるようになり、2010年代には動的な細胞内構造体の駆動原理としての生物学的相分離における重要な因子として研究が進められてきた。実際、IDRが構造を有する分子領域と同等かそれ以上に重要な生理機能を有していることが分かってきている。特定の構造を持たないIDRを標的とした創薬開発にはIDRの可視化が必要であるが、現状ではIDRを可視化する技術には課題が多く残されている。本講演では、IDRを含めた分子動態評価の最近の動向について、他の演者による講演を含め議論したい。

第二部

座長：奈良県立医科大学 V-iCliniX講座 准教授　森 英一朗

13:30～13:40

奈良県立医科大学　学長
細井 裕司

演題：
軟骨伝導聴覚とMBT（Medicine-Based Town）

500年以上前から、音が内耳に達する経路として気導経路と骨（伝）導経路の2つの経路が知られていた。2004年に私が音情報を含む振動を耳軟骨に与えると音が内耳に良好に伝達されることを発見し、第3の聴覚「軟骨伝導」という概念と用語を提唱した。この医学的発見は、補聴器などの医療機器に応用できるだけでなく、巨大なマーケットを持つ音響・通信機器に革命的な変化をもたらす。つまり、新しい音伝導メカニズムが新しいスマートホン、スマートグラス、腕時計型・指輪型電話やコンピュータ端末、ペン型電話、メガネ型電話・補聴器、通信用インカム、介護ロボット、水中音楽鑑賞・通信装置、サウンドジュエリー（聞こえる宝石）を生み出すことになり、すでにいくつかが上市されている。

現在、MRIやCTのように工学が医学をサポート、発展させるME（Medical Engineering）は広く知られている。軟骨伝導の発見が産業の発展につながることから、医学が工学や産業を発展させることをMBE(Medicine-Based Engineering)と名付けた。それをまちづくりに生かすことをMBT（Medicine-Based Town）と呼び、奈良医大を中心に、実現のためのMBTコンソーシアムを組織している。

第二部

座長：奈良県立医科大学 V-iCliniX講座 准教授　森 英一朗

13:40～14:10

国立研究開発法人産業技術総合研究所　健康医工学研究部門 ナノバイオデバイス研究グループ
主任研究員
冨田 峻介

演題：

光パターンを基にバイオサンプルの複雑性を解読する分析技術「Chemical tongue」

これまでに我々は、ヒトの味覚の仕組みを模倣したバイオ分析技術 'Chemical tongue' の開発を進めてきた。この技術は、合成プローブ群と機械学習を組み合わせることで、バイオサンプルの微細な違いを捉えることを可能にする。典型的なChemical tongueは、多様な構造の蛍光ポリマーの溶液を配置したアレイを用いる。サンプルをアレイに加えると、各ポリマーは様々な強さでサンプル中の成分と相互作用し、その結果、サンプルの性質を反映した '光パターン' が出力される。機械学習によってこれらのパターンを解析することで、サンプルの正確な識別が可能になる。

Chemical tongueは、未知の組成を持つサンプルでも、得られる光パターンに基づいて比較分析できるという、味覚と類似した特長を有する。これまでに、治療用抗体、血清、培養細胞、腸内細菌叢などの複雑なバイオサンプルの高精度な識別や分類、さらには変化をモニタリングすることに成功している。迅速かつ簡便、そして安価に複雑なバイオサンプルの正確な分析を実現するChemical tongueは、医療分野をはじめとする幅広い産業での応用が期待される。本講演では、Chemical tongueのための材料開発と、構築したChemical tongueによる多様なバイオサンプル評価に関する最近の成果を紹介したい。

第二部

座長：奈良県立医科大学 V-iCliniX講座 准教授　森 英一朗

14:10～14:40

徳島大学　先端酵素学研究所　教授
齋尾 智英

演題：
生命を駆動する動的相互作用の機序解明

生体分子の作用機序を分子レベルから明らかにする上で，タンパク質複合体の立体構造は決定的な情報を与え得る．しかし，溶液中でのタンパク質は柔軟であり，弱く動的な相互作用が機能発現に重要な場合も多いが，そうしたタンパク質の動的な姿については技術的制約により理解が進んでいない．一方で，生物学的相分離に関する理解の広がりと，神経変性疾患やがん悪性化などとの関連が明らかになるにつれ，生命科学研究や医学研究の分野におけるタンパク質の動的相互作用の重要性が認知されつつある．本研究では，生物学的相分離やタンパク質フォールディングにおける制御因子として機能するシャペロンに注目し，シャペロンによる動的基質認識を通した機能発現のメカニズム解明を目指している．特に，核磁気共鳴（NMR）法や分光法などを用いた溶液中でのタンパク質の構造解析，相互作用解析を用いた研究を推進している．本発表では，NMRを用いたタンパク質の動的構造解析手法について簡単に解説するとともに，NMRを用いた研究成果，特に筋萎縮性側索硬化症の関連因子の生物学的相分離とそれを制御するシャペロンに関する研究の成果を紹介する．

第二部

座長：奈良県立医科大学 脳神経内科　教授　杉江 和馬

14:55～15:25

奈良県立医科大学　脳神経内科　医員
井口 直彦

演題：
ジンクフィンガードメインはLCドメインポリマーを認識する

相分離性タンパク質は限られた数種類のアミノ酸で構成された低複雑性（LC）ドメインを持ち、このLCドメインが自己会合することでポリマーを形成する。LCドメインを持つhnRNPA2やFUS、TDP43などのRNA結合タンパク質は、相分離の破綻により病原性線維を形成し、筋萎縮性側索硬化症などの神経変性疾患につながることが示唆されている。これまでに複数の分子シャペロンがLCドメインポリマーの形成を制御することが分かっているが、LCドメインポリマーが認識されるメカニズムは不明な点が多い。そこで我々はDNA結合モチーフであるジンクフィンガードメイン（ZnF）がLCドメインと構造依存的に結合することを明らかにした。ZnFはLCドメインのヒドロゲルに結合し、LCドメインのポリマー形成を阻害した。さらに、ZnFはモノマー状態よりポリマー状態のLCドメインを強く認識した。これらの結果からZnFはLCドメインポリマーの形成における調節因子として機能していると考えられる。

第二部

座長：奈良県立医科大学 解剖学第二　教授　和中 明生

14:55～15:25

奈良県立医科大学　解剖学第二　講師
田中 達英

演題：
神経免疫連関による痛覚制御の新しいメカニズム

様々な感覚刺激は、感覚情報として符号化され一次求心性神経を介して脳へと伝達される。痛み研究はこれまで痛覚伝導路に着目した研究が主流であったが、近年、神経細胞と空間的に近接する非神経細胞、特に脊髄におけるミクログリアやアストロサイトが慢性疼痛の発症と維持に重要であることが数多く報告されている。一方で、末梢における非神経細胞が痛みの伝導にどのように関与しているのか、いまだ不明な点が多い。末梢組織の損傷や感染により発痛物質や炎症メディエーターが痛みに関与することは古くから知られているが、生理的な痛覚閾値や急性痛において末梢の非神経細胞が感覚受容にどのように関与するのか不明な点が多い。我々は、病理的疼痛のみならず生理的な機械刺激に対しても神経周囲に局在するマクロファージが痛覚の制御に重要な役割を持つことを明らかにした。さらに、マクロファージ-ニューロン間のシグナル伝達に重要な因子としてSNX25/Nrf2/NGFを同定した。本講演では皮膚真皮層および遠位結腸の粘膜固有層における一次求心性神経周囲のマクロファージに着目した痛覚センシング機構について我々の研究成果を報告する。

第二部

座長：奈良県立医科大学 血栓止血分子病態学　教授　嶋 緑倫

14:55～15:25

奈良県立医科大学　血液内科・輸血部　教授
松本 雅則

演題：
地方公立単科大学での抗体製剤の開発ー血栓症研究から出血症治療薬

地方の公立の単科大学というのは、研究を行う上で3重苦であるという話を聞いたことがある。奈良医大もこの範疇に入るが、私は周囲に恵まれて、自分たちで一から作った抗体製剤の開発を行い、AMED予算を獲得し、現在非臨床試験を実施している。

止血因子von Willebrand因子（VWF）を切断する酵素ADAMTS13の活性が著減することで、全身の血管に血栓ができる血栓性血小板減少性紫斑病（TTP）が発症する。TTP研究のためADAMTS13に対するモノクローナル抗体を学内で作成し、A10という抗体はADAMTS13活性を阻害することから2004年に特許申請した。

TTPは、ADAMTS13が効かないことによりVWFが切断されずに血栓症となる。逆にADAMTS13が効きすぎて出血症である後天性von Willebrand症候群(AVWS)となることが2010年頃から注目されるようになった。AVWSは大動脈弁狭窄症などに合併し症例数も多いが、有効な治療法が無いため、A10を使った治療法を想起した。

AMED研究費獲得後、ヒト血液還流実験、カニクイザルモデルなどを用いた実験を行い、A10の有効性を証明した。2021年にA10のヒト化に成功し、これを製剤化するためにGLPグレードでの大量培養の方法を確立し、非臨床試験を開始した。今後2，3年以内にFirst-in-human試験を実施したいと考えている。

第二部

16:25～16:40

閉会挨拶
奈良県立医科大学　V-iCliniX講座
教授　佐藤 匠徳

公益財団法人
中谷医工計測技術振興財団

第2回長期大型研究助成　奈良県立医科大学　最終成果報告書

発　行　令和6年9月
発行所　公益財団法人 中谷医工計測技術振興財団
〒141-0032　東京都品川区大崎一丁目2番2号　アートヴィレッジ大崎セントラルタワー8階
ＵＲＬ　https://www.nakatani-foundation.jp