2025 年 9 月 30 日

報道関係者各位

国立大学法人筑波大学
国立研究開発法人理化学研究所

ヒト転写因子データの未測定範囲を体系化し研究戦略を提示

ヒトの遺伝子調節を担うタンパク質である転写因子のゲノム結合データを精査し、未測定データが多数存在することを明らかにしました。これにより、データの偏りが研究や疾患理解に及ぼす影響と、今後の測定対象選択の戦略を提案しました。また、本成果を公開データベースとして整備しました。

ヒトのゲノムには約 1,600 種類の転写因子があり、400 種類以上の組織・細胞型で遺伝子の働きを調節しています。その役割を理解するために重要な手法が、クロマチン免疫沈降シーケンス（ChIP-seq）です。しかしながら、抗体や細胞数などの制約からすべての組み合わせを網羅的に測定することは困難で、多くの重要な組織・細胞型が未測定のまま残されています。

本研究では、公開されているヒトの ChIP-seq データを大規模に解析し、十分に発現しているにもかかわらず測定されていない「未測定転写因子－組織・細胞型ペア」を体系的に定義しました。その結果、血球細胞では測定が比較的進んでいる一方、膵臓や筋肉、胎盤などでは 80%以上が未測定であり、大きな研究の空白が存在することが明らかになりました。

また、他の実験データと統合解析したところ、未測定の転写因子であっても多くの遺伝子の発現に影響を及ぼしうることが示されました。これは、現在得られているデータだけでは遺伝子調節ネットワークの全体像を十分に把握できないことを意味します。さらにシミュレーション解析からは、測定順序を工夫し、初期段階から多様な転写因子を幅広く測定する戦略が、疾患関連変異の解釈を効率的に改善できることが示されました。

この研究は、未測定データがゲノム機能解析の応用研究や疾患理解に与える影響を初めて包括的に示したものであり、今後の研究資源の戦略的な活用に重要な指針を提供します。

研究代表者
筑波大学医学医療系
尾崎 遼　客員准教授（兼 理化学研究所 生命機能科学研究センター　チームディレクター）
筑波大学附属病院
田原 沙絵子　初期研修医

研究の背景

転写因子[注1)]は DNA に結合して遺伝子の働きを制御するタンパク質であり、病気の発症や進行、さらには治療薬の標的探索にも関わる極めて重要な因子です。ヒトゲノムには約 1,600 種類の転写因子が存在し、400 種類以上の組織・細胞型でそれぞれ異なる働きをします。これらの働きを網羅的に明らかにするための手法として、クロマチン免疫沈降シーケンス（ChIP-seq）[注2)]が用いられています。ChIP-seq 測定データは公共データベースに蓄積され、遺伝子制御ネットワークの再構築、疾患関連ゲノム領域の解釈、SNP（一塩基多型）[注3)]の機能影響評価、さらには AI を用いた創薬など、幅広い応用に活用されています。

しかしながら、これまで世界中の研究者によって蓄積されてきた ChIP-seq 測定データの集合も、生物学的に重要な転写因子と組織・細胞型のすべての組合せは網羅できていません。これは、一度の実験で得られるのは単一の「転写因子－組織・細胞型ペア」に限られる一方で、可能な組合せは膨大であり、また、ChIP-seq 測定において転写因子ごとの特異的抗体と数百万規模の細胞試料が必要とされるためです。その結果、十分に発現しているにもかかわらず測定が行われていない「未測定転写因子－組織・細胞型ペア（未測定データ）」[注4)]が多数存在しています。未測定データの存在は、データベースの網羅性を制限し、データ解析や機械学習の精度に影響を与え、研究活動における新発見を妨げる可能性があります。

これまでのゲノム科学において、ヒトゲノムプロジェクトや ENCODE（The Encyclopedia of DNA Elements）といった国際的なヒトゲノム解読の取り組みは、包括的データの公開によって研究を加速させてきました。従って、ChIP-seq データについても同様に、未測定データの範囲を明確に定義し、戦略的に補完することは、データ駆動型研究の信頼性を高め、疾患理解や創薬研究に大きな波及効果をもたらすと期待されます。

研究内容と成果

本研究では、公開されている大規模なヒト ChIP-seq データベースを対象に網羅的な解析を行い、未測定データの体系的な定義と定量化を試みました。解析には、ヒトで最も多くの ChIP-seq 測定データを収載している データベース ChIP-Atlas を中心に用い、2023 年 10 月時点で 27,865 件の実験データ（1,810 種類の転写因子、1,126 種類の細胞型を対象）を精査しました。その際、別のデータベースから取得した遺伝子発現データにおいて十分に発現している転写因子を「潜在的に測定可能」とみなし、その上で ChIP-seq 測定が未実施である場合を「未測定」と定義しました。

解析の結果、実験数は年々増加しているものの、ChIP-seq 測定は一部の転写因子や血球細胞系に集中しており、測定対象の偏りが確認されました。実際に、血球細胞では数百種類の転写因子が測定されていた一方、膵臓や筋肉、胎盤といった組織では 80%以上が未測定のままであることが分かりました。さらに解析を時系列で追跡すると、この偏りは時間とともに緩和されるのではなく、むしろ保持・強化されるという傾向（The rich gets richer）が見られました。つまり、過去に多くの研究が行われた転写因子や組織ほど新たな実験が追加されやすく、逆に未測定データは長期間取り残され続けることが明らかになりました。また、転写因子ごとの出版論文数の偏りも、ChIP-seq 測定の対象選択に大きな影響を与えており、研究資源が限られた範囲に集中する傾向を強める要因となっていることが示されました。

また、転写因子の機能的重要性を評価するため、他の公開実験データとの統合解析を行いました。具体的には、転写因子の機能を遺伝子ノックアウト[注5)]等で阻害した後の遺伝子発現変動数（DEGs）を指標とし、さらに出版論文数を加えて解析しました。その結果、未測定の転写因子であっても多数の遺伝子に影響を及ぼす場合があり、研究の注目度によらず機能的に重要な因子が存在することが確認されました。加えて、十分に発現しており、多くの遺伝子発現に影響し、組織特異的マーカーとして知られる転写因子

を「隠れた宝石（hidden gems）」と定義しました。この基準に基づいて、ATM（乳腺）、SOX9（肝臓）、PTEN（肺）といった転写因子について、疾患理解や創薬に直結するにもかかわらず未測定の重要ペアが抽出されました。

データの偏りが遺伝子制御機構や疾患関連領域の探索といった応用解析に与える影響を評価するため、新たに「Reg-TF カバー率」（発現遺伝子の転写開始点付近に ChIP-seq ピークが存在する割合）と「GWAS-SNP カバー率」（疾患関連一塩基多型に ChIP-seq ピークが重なる割合）を定義し、組織・細胞型ごとに算出しました。その結果、測定数が豊富な血球細胞系では両カバー率が高く、疾患関連変異の解析に有用である一方、膵臓や神経系では低値にとどまり、重要な組織・細胞型に対する情報不足が浮き彫りとなりました。

さらに、シミュレーション解析により、測定順序を工夫することでデータの偏りを早期に緩和できることを示しました。特に、測定の初期段階から多様な転写因子を幅広く測定する戦略は、疾患関連 SNP のカバー率を効率的に高め、AI による予測精度の向上にもつながることが確認されました（参考図）。

最後に、未測定転写因子－組織・細胞型ペアの包括的リストを作成し、誰でもアクセス可能な公開データベース（https://moccs-db.shinyapps.io/Unmeasured_shiny_v1/）を整備しました。これにより、研究者が不足データを容易に把握し、限られた研究資源を戦略的に配分できると期待されます。

今後の展開

これまでのメタサイエンス研究[注6)]は、論文や特許情報などの文献解析が主流でした。しかし近年、オミクス測定技術[注7)]やオープンサイエンスの進展によって、大規模な測定データが蓄積しつつあります。これらのデータは個別化医療や AI 創薬の基盤となると同時に機械学習・生成 AI の学習資源にもなることから、データの蓄積状況や偏り・不足を体系的に調べることは今後の生命科学においてますます重要になります。

また近年、実験自動化技術や高効率な実験プロトコルの導入によって、より効率的なデータ収集が可能になってきており、本研究成果は、どのような測定を優先して実験すべきかをデータ解析や AI によって合理的かつ戦略的に判断する際に有用であり、国際共同プロジェクトや他分野のデータ駆動型研究にも応用可能な視座を与えるものと期待されます。

参考図

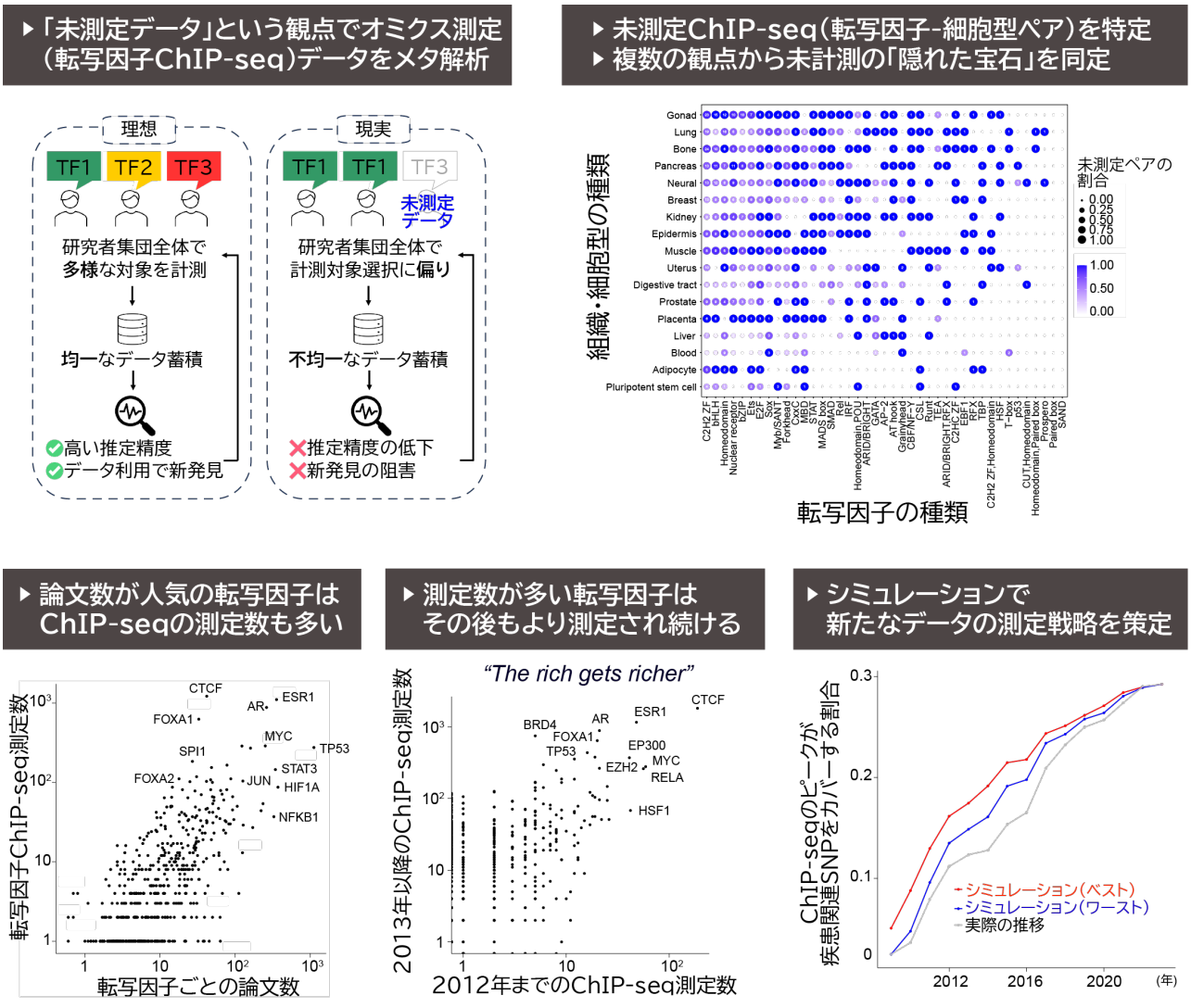

図　本研究で行った未測定転写因子－組織・細胞型ペアの抽出と戦略的活用の概要

ヒト転写因子 ChIP-seq データベースと遺伝子発現データを統合し（上左図）、特定の組織・細胞型で十分に発現しているにもかかわらず ChIP-seq が未測定である「未測定転写因子－組織・細胞型ペア（未測定データ）」を体系的にリスト化した（上右図）。これらのリストから論文数が多い「人気」の転写因子が ChIP-seq 測定数も多いこと（下左図）、測定数の多い転写因子はその後も測定され続けること（下中図）を明らかにした。さらに、シミュレーション解析を行い、測定の順序を工夫して多様な転写因子を早期に調べる戦略が、疾患関連変異のカバー率を向上させることを示した（下右図）。

用語解説

注１）　転写因子（Transcription factorF）

DNA に結合して、遺伝子の働きを調節するタンパク質。

注２）　クロマチン免疫沈降シーケンシング（ChIP-seq）

DNA とタンパク質の相互作用を解析するクロマチン免疫沈降法（ChIP）と、大量の DNA の塩基配列を迅速に解析する次世代シーケンシングを組み合わせ、ゲノム全体で転写因子が結合している場所を特定する方法。

注３）　一塩基多型（Single Nucleotide Polymorphism, SNP）

ゲノム DNA の配列において、ある一塩基（A、T、G、C のいずれか）が集団内で一定の頻度（通常 1%

以上）で置き換わっている遺伝的多様性。ヒトゲノムでは数百万か所に存在し、多くは機能に影響しないが、一部は病気のなりやすさや薬の効き方などの個人差に関わる。

注４） 未測定転写因子－組織・細胞型ペア（未測定データ）

特定の転写因子と組織・細胞型の組み合わせのうち、まだ ChIP-seq 測定が行われていないもの。

注５） 遺伝子ノックアウト（gene knockout）

ある遺伝子の働きを完全に失わせる実験手法。特定の遺伝子を欠損させた細胞や動物をつくり、その遺伝子が生命活動や疾患にどのように関与しているかを調べることができる。

注６） メタサイエンス研究（Meta-scientific research）

研究の方法、データの公開・共有、論文の引用や再現性など、科学活動の進め方や科学知識の蓄積の仕組みそのものを客観的に調べることで、研究の効率化や透明性の向上、バイアスの低減などを目指す学問分野。

注７） オミクス測定技術（Omics technologies）

ゲノム（genomics）、トランスクリプトーム（transcriptomics）、プロテオーム（proteomics）、メタボローム（metabolomics）など、生体内の分子を網羅的かつ高精度に測定する技術の総称。次世代シーケンシングや質量分析計などの解析技術を用いて、大規模な分子情報を取得し、生命現象の包括的理解や疾患研究、創薬に活用される。本研究で用いた ChIP-seq も、オミクス測定技術の一つ。

研究資金

本研究は、科研費（JP22K17992）および JST 未来社会創造事業による研究プロジェクト「ロボティックバイオロジーによる生命科学の加速」（JPMJMI20G7）による研究プロジェクトの一環として実施されました。

掲載論文

【題　名】 *Unmeasured* human transcription factor ChIP-seq data shape functional genomics and demand strategic prioritization
（未測定のヒト転写因子 ChIP-seq データが機能ゲノム学に与える影響と戦略的優先順位付けの必要性）
【著者名】 Saeko Tahara and Haruka Ozaki
【掲載誌】 *Briefings in Functional Genomics*
【掲載日】 2025 年 9 月 30 日
【DOI】 10.1093/bfgp/elaf016

問合わせ先

【研究に関すること】
尾崎　遼（おざき　はるか）
筑波大学 医学医療系 客員准教授
URL: https://sites.google.com/view/ozakilab-jp
https://www.bdr.riken.jp/ja/research/labs/ozaki-h/index.html

【取材・報道に関すること】
筑波大学広報局
TEL: 029-853-2040
E-mail: kohositu@un.tsukuba.ac.jp

理化学研究所広報部 報道担当
TEL: 050-3495-0247
E-mail: ex-press@ml.riken.jp