2023 年 8 月 25 日

疾患関連遺伝子群における類似した特徴の抽出と検出に成功

～3層オミックスデータにデータ駆動型解析を適用～

学校法人　中央大学
岩手医科大学 いわて東北メディカル・メガバンク機構

概　　要

東北メディカル・メガバンク（TMM）計画[注1)]岩手医科大学 いわて東北メディカル・メガバンク機構（IMM）では、地域住民コホート参加者・約 100 名の血液細胞からゲノム・エピゲノム・トランスクリプトームの3層オミックスデータを網羅的に収集し、共同研究のリファレンスとして、あるいは分譲データとして利用促進することで、疾患関連研究に役立ててきた。

しかし、このコホート参加者の3層オミックスデータは症例対照研究で行われるように何れかの疾患患者と非患者である対照群の2群で収集しているわけではないため、3層オミックスデータのみでは、どのような異常が潜在しているかを特定することが難しかった。

本研究において中央大学 教授の田口 善弘とIMMの研究チームは、田口が開発したデータ駆動型[注2)]の解析方法により、患者群と非患者群に分類されていない対象者群の3層オミックスデータから類似した特徴を抽出することに成功した。さらに、これらのパターンに同期して変動している遺伝子群を特定したところ、さまざまな疾患関連遺伝子を抽出できることを明らかにした。この研究成果は、今後、追跡調査のデータを利用し、抽出した遺伝子群について発症者のデータとの比較を行うことによって、疾患発症を予測することにもつながるものと期待される。

本研究成果は、米国の科学雑誌「PLOS ONE」（日本時間の 2023 年 8 月 10 日付）に掲載されました。

＊＊

【研究者】

田口　善弘　中央大学理工学部　教授　（物理学科）
小巻　翔平　岩手医科大学医歯薬総合研究所　講師　（生体情報解析部門）
須藤　洋一　岩手医科大学 いわて東北メディカル・メガバンク機構　特命准教授（生体情報解析部門）
大桃　秀樹　岩手医科大学医歯薬総合研究所　特任准教授　（生体情報解析部門）
山﨑　弥生　岩手医科大学 いわて東北メディカル・メガバンク機構　特命助教（生体情報解析部門）
清水　厚志　岩手医科大学医歯薬総合研究所　教授　（生体情報解析部門）

【発表（雑誌）】

タイトル：Integrated analysis of human DNA methylation, gene expression, and genomic variation in iMETHYL database using kernel tensor decomposition-based unsupervised feature extraction
雑誌名：PLOS ONE
URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0289029

【研究内容】

1. 背景

複数種類のプロファイルをまとめて解析するいわゆるマルチオミックス[注3)]解析は、1つ1つのプロファイルの変数の数が著しく異なるため容易ではない。例えばよく一緒に解析されることが多い遺伝子発現プロファイルとゲノムの DNA メチル化[注4)]プロファイルの場合、遺伝子は数万個しかないが、DNA メチル化サイトは数千万カ所存在するなど、個数が3桁も異なっており、遺伝子との組み合わせの数は膨大である。そのため、なんらかの付加的な情報によって解析対象数を制限する、例えば、プロモーター[注5)]と呼ばれる遺伝子の発現量に影響を与えることが解っている領域の情報を事前知識として与えるなどしなければ統合解析が難しい状況であった。しかし、解析するゲノム領域を制限してしまうと、他の機能（例えば、エンハンサー[注6)]と呼ばれる、ゲノムをループ状に結合させることで遺伝子の発現量に影響を与える部分）への DNA メチル化の影響は除外されてしまい、さらには、未知の影響を検出することが不可能であった。

これに対してデータ駆動型の方法は事前知識を用いずともデータのみから想定し得ない関係性を抽出することができるため、相互関係にある遺伝子と DNA メチル化領域を特定できる可能性があった。しかし、膨大な数である DNA メチル化領域と遺伝子の組み合わせすべてを既存のデータ駆動型のアプローチでは調べることができず、手が届かなかった。

2. 研究内容と成果

今回は文献1で開発した方法をいわて東北メディカル・メガバンク機構（IMM）が 100 名の地域住民から網羅的に収集した3層オミックスデータ（遺伝子発現プロファイル、DNA メチル化プロファイル、一塩基多型（SNP）[注7)]プロファイル）に適用して、疾患関連遺伝子との関係を把握できるかを確認した。この方法はカーネルテンソル分解を用いた教師無し学習による変数選択法（以下、「テンソル分解法」）と呼ばれるデータ駆動型で、全員が健常者である今回のような対象者に特段のラベルが付与されていない場合でも適用可能である。

また、一プロファイルあたりカーネルサイズ（具体的には対象者数の二乗）程度のメモリーがあれば実行可能で、SNP や DNA メチル化サイトの様に変数が数十万から数千万になるような巨大なプロファイルであってもデータ駆動型解析で対象者群の固有パターンの特定と、そのパターンに相同的なプロファイルをもつ変数（遺伝子発現量、DNA メチル化プロファイル、SNP プロファイル）を特定できる方法でもある（図）。

IMM はプロファイル取得後、現在にいたるまで対象者の継続的な観察を行っている。抽出された遺伝子、DNA メチル化サイト、SNP の疾患依存性が将来の対象者の発症傾向とどの程度一致するかなどを観察することは、テンソル分解法を3層オミックスデータに適用することが未病を含めた対象者の健康状態把握にどの程度有効であるかを検証する、稀な機会となることが期待される。

テンソル分解法を $CD4^+$ T 細胞、単球、好中球の3種類の細胞の22本の常染色体ごとの3層オミックスデータに適用した。その結果、2種類の対象者パターンが同定され、22本の常染色体ごとに得られたこれらの2種類の対象者パターンは常染色体間で非常に強く相関し（保存され）ていた。常染色体に存在する遺伝子は互いに全く異なるので、同じようなプロファイルを得られるのは偶然とは考えにくい。この様な場合、もっとも疑われるバイアスはバッチ効果（対象者ごとに検体の採取量が違っているのを検出してしまっている等）であるが、2種類のパターンは互いに直交しているため、単なるバッチ効果ではこのような2種類の対象者パターンの出現を説明できないこと、サンプル調整方法も取得方法も異なる3種類

のオミックスデータに同じバッチ効果が乗っている可能性は極めて小さいことなどから、単なるバッチ効果であるとは考えにくい。

この2種類の対象者パターンはテンソル分解で得られる第2、第3特異値ベクトルであり、前者は3つの細胞種すべてで、後者は単球以外の2種の細胞で検出された。次にこれらの対象者パターンと相同なプロファイルをもつ遺伝子とDNAメチル化領域を選択したところ、多くの転写因子[注8)]の共通標的になっており、かつ、これらはエンリッチメント解析により、さまざまな疾患との関連がある転写因子であることが明らかとなった。また、同定されたSNPはこれらの転写因子のDNA結合部位と統計的に有意に重なっていることも分かった。

以上のことから、テンソル分解法の3層オミックスデータの統合解析への応用が有効であることが示されたと思われる。

参考文献:

1） Taguchi, Yh., Turki, T. Novel feature selection method via kernel tensor decomposition for improved multi-omics data analysis. BMC Med Genomics 15, 37 (2022). https://doi.org/10.1186/s12920-022-01181-4

3. 今後の展開

IMM では本研究成果に利用したデータ以外に臍帯血データなども蓄積しており、今後も3層オミックスデータの蓄積を行っていく予定である。将来的には3層オミックスデータのテンソル分解法による解析が未病検出も含めた国民的な健康管理の手段として確立させることを目指していく。

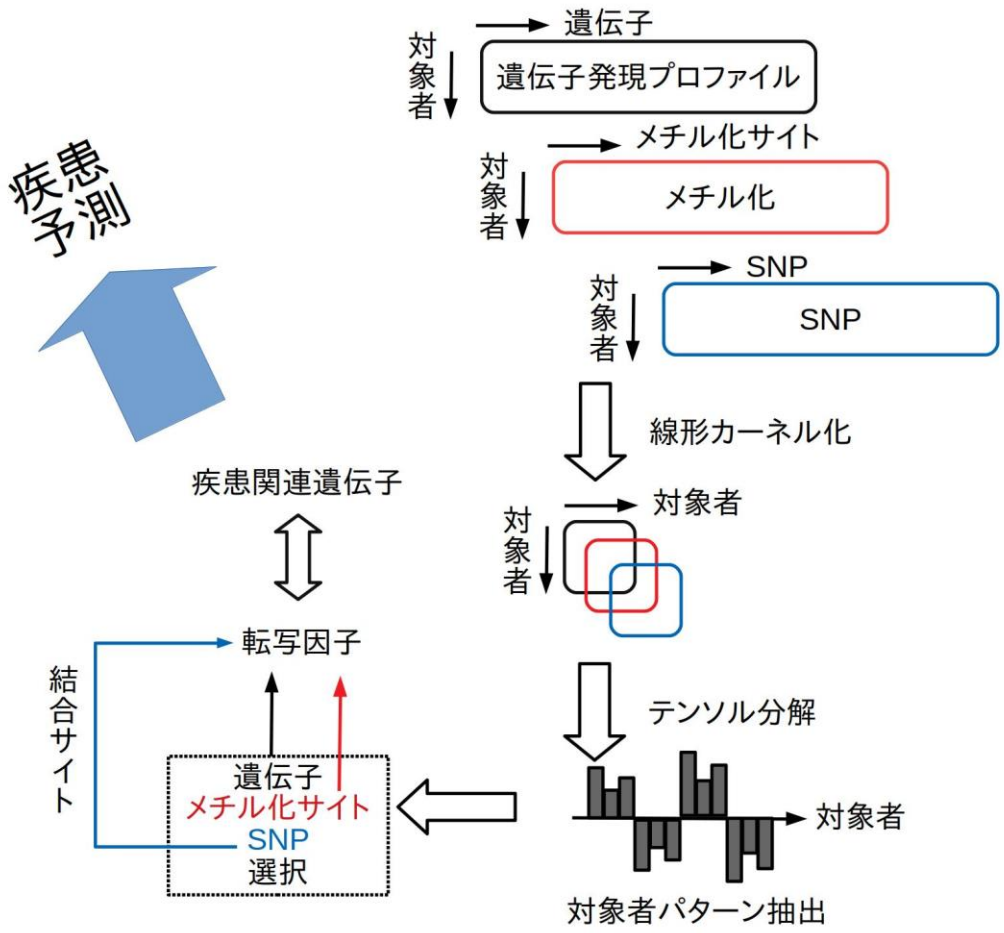

図　遺伝子発現プロファイル、DNA メチル化プロファイル、SNP プロファイルをそれぞれ線形カーネル化[注9)]で対象者数の２乗の行列に変換した後、テンソル分解を行い対象者パターンを抽出し、それらと相同な変動パターンを持つ、遺伝子、メチル化サイト、SNP を選択することで外部情報なしのデータ駆動型の方法で選択できる。遺伝子、メチル化サイトは多くの転写因子の標的となっており、それらの転写因子の DNA への結合部位は検出された SNP と統計的に有意に重なっていたことから３層オミックスデータの統合解析は成功したとみなされる。また転写因子は疾患とも関係しており、３層オミックスデータ＋テンソル分解法は将来の疾患発症を予測可能な方法として期待される解析法である。

本研究は、科学研究費補助金（文部科学省）、新学術領域研究(研究領域提案型)、19H05270、20H04848、科学研究費基金（文部科学省）、基盤研究(C)、20K12067 の支援を受けて実施された。

【お問い合わせ先】

＜研究に関すること＞
田口善弘（タグチ ヨシヒロ）
中央大学理工学部　教授（物理学科）
TEL：　03-3817-1791
E-mail: tag@granular.com

＜広報に関すること＞
中央大学　研究支援室
TEL　03-3817-7423 または 1675，FAX　03-3817-1677
E-mail: kkouhou-grp@g.chuo-u.ac.jp

【用 語 解 説】

注1）　東北メディカル・メガバンク（TMM）計画

東北メディカル・メガバンク（TMM）計画は、東日本大震災からの復興事業として2011（平成23）年度から始められ、被災地の健康復興と、個別化予防・医療の実現を目指している。TMM 計画は、東北大学東北メディカル・メガバンク機構（ToMMo）といわて東北メディカル・メガバンク機構（IMM）を実施機関として、東日本大震災被災地の医療の創造的復興および被災者の健康増進に役立てるために、合計 15 万人規模の地域住民コホート調査および三世代コホート調査を 2013（平成 25）年より実施し、収集した試料・情報をもとにバイオバンクを整備している。なお、TMM 計画は、2015（平成 27）年度より、日本医療研究開発機構（AMED）が本計画の研究支援担当機関の役割を果たしている。

IMM ウェブサイト:http://iwate-megabank.org/

注2）　データ駆動型

通常の機械学習やデータサイエンスでは教師あり学習や強化学習を用いてある目的に従ってモデルをトレーニング（学習）させるが、データ駆動型の研究では最初になんらかの目的を置くのではなく、まず初めにデータを解析し、その結果を見てデータの中で何が起きているかを観察することで結果を出していく。このような研究方法をデータ駆動型の研究という。

注3）　マルチオミックス

遺伝子発現プロファイル以外の DNA メチル化プロファイルやSNPプロファイルをまとめてよぶ呼称。それ以外にも microRNA の発現プロファイル、ヒストン修飾など多岐に渡るものがマルチオミックスを構成しうる。

注4）　DNA メチル化

ヒトの DNA を構成する4分子（アデニン・チミン・グアニン・シトシン）のうち、シトシンの多くはメチル化される。これが起きた場合、転写制御に関わるタンパク質の結合を阻害することで遺伝子の発現に影響を与える（一般には抑止）。

注5）　プロモーター

ゲノムを構成する DNA のうち遺伝子としての機能領域はごく一部分である。どの遺伝子を適切な時期、適切な組織で発現させるかを決めるためには遺伝子に隣接した領域が重要であり、その領域が「プロモーター」と呼ばれている。この部分がメチル化されたり、新規突然変異が生じると発現量が変化する。

注6） エンハンサー、サイレンサー

プロモーターと異なり、遺伝子から遠く離れている領域でも DNA のループ構造形成を通じて遺伝子の発現に影響を与える領域が存在する。このうち、発現を上昇させるものはエンハンサー、抑制させるものはサイレンサーと呼ばれている。この部分に対する DNA メチル化や新生突然変異も遺伝子の発現に影響を与える。

注7） 一塩基多型（SNP）

個人によって、DNA の4種類の分子の1つが他の分子に置き換わっている場合がある箇所。生体内のタンパク質は DNA 配列が示す順番で 20 種類のアミノ酸を連ねて作成されるため、SNP の存在する DNA 上の位置によっては、タンパク質の構造が変わり、その機能に大きな影響を生じることがある。

注8） 転写因子

遺伝子の発現をつかさどるタンパク質群の総称。転写因子がプロモーター／エンハンサー領域に結合することで遺伝子の発現が制御される。

注9） カーネル化（カーネルトリック）

非常に高次元の空間を少数のサンプルの次元で解析できる数学的なトリックで、機械学習では日常的に使われている。