

機械学習 と 次世代シーケンス技術 の活用により 日本人集団の HLA を解明

キーワード: 機械学習、次世代シーケンス技術、白血球血液型、HLA 遺伝子、個別化医療

【研究成果のポイント】

- ◆ 日本人集団 1,120 名を対象に、次世代シーケンス技術^{※1}を駆使して白血球の血液型を決定する HLA 遺伝子^{※2}を解読
- ◆ 機械学習^{※3}により、日本人集団の白血球の血液型が 11 パターンの組み合わせで構成されることを解明
- ◆ バイオバンク・ジャパン^{※4}が構築した日本人集団 17 万人を対象としたフェノムワイド関連解析(PheWAS)^{※5}で、白血球の血液型の個人差が、病気や量的形質を含む 50 以上の表現型^{※6}に関わることを示した

❖ 概要

大阪大学 大学院医学系研究科遺伝統計学の岡田随象 教授(IFReC 免疫統計学兼任)らの研究グループは、次世代シーケンス技術と機械学習を用いて、日本人集団における白血球の血液型が 11 パターンで構成されており、その個人差が、病気や量的形質を含む 50 以上の表現型に関わっていることを明らかにしました。

ヒトの血液に含まれる白血球には血液型が存在し、ヒトゲノム上の HLA 遺伝子の配列の個人差で決定されます。白血球の血液型は移植医療や個別化医療^{※7}に際して重要ですが、HLA 遺伝子構造が複雑で解読に専門技術が必要なことや高額な実験費用により、HLA 遺伝子配列の詳細な個人差の解明は遅れていました。

岡田教授らの研究グループは、最先端のゲノム配列解読技術である次世代シーケンス技術を駆使して、日本人集団 1,120 名を対象に 33 の HLA 遺伝子におけるゲノム配列を決定することに成功しました。得られた HLA 遺伝子ゲノム配列情報に対して機械学習手法である tSNE^{※8}を適用した結果、日本人集団の白血球の血液型を 11 パターンの組み合わせに分類可能なことが明らかになりました。これは、複雑なヒトゲノム情報の解釈を、機械学習手法を用いて実現した先進的な成功例と評価することができます。

さらに研究グループは、日本人集団 17 万人のゲノムデータを対象に、白血球の血液型をコンピュータ上で高精度に推定することに成功しました。推定された血液型パターンに基づき、多彩な表現型との関連を調べるフェノムワイド関連解析を実施しました。その結果、50 以上の表現型において、白血球の血液型が発症に関与していることが明らかになりました。

本研究成果により、日本人集団における白血球の血液型の全容が解明されました。機械学習による白血球の血液型の分類に成功したことは、生命科学研究における機械学習の画期的な応用例と考えられます。さらに、白血球の血液型を用いた個別化医療の実現に貢献するものと期待されます。

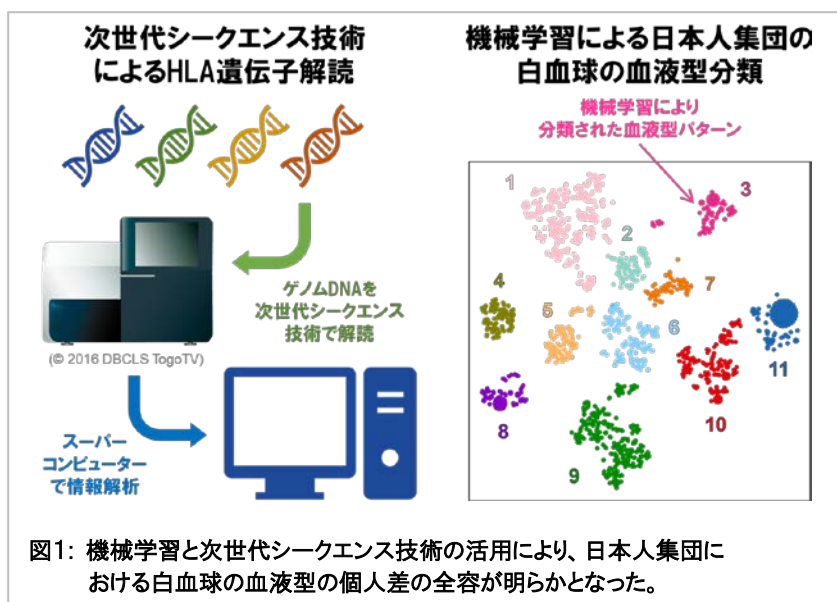


図1: 機械学習と次世代シーケンス技術の活用により、日本人集団における白血球の血液型の個人差の全容が明らかとなった。

❖ 研究の背景

ヒトの血液には、赤血球、白血球、血小板といった血球細胞が含まれています。赤血球の血液型(A、B、O、AB 型)がヒトゲノム配列上の特定の遺伝子(=ABO 遺伝子)のゲノム配列の個人差で決まると同様に、白血球にも血液型が存在し、HLA 遺伝子のゲノム配列の個人差で決定されることが知られています。白血球の血液型の個人差は免疫アレルギー疾患をはじめ多彩な表現型の発症に強いリスクを有することが知られており、個人のゲノム情報を医療に活用する個別化医療の先駆けとして社会実装が期待されています。また、骨髄移植や iPS 細胞を用いた再生医療などの臓器移植を伴う医療行為においては、臓器の提供者(ドナー)と移植を受ける人(レシピエント)との間で白血球の血液型を合致させる必要があることが知られています。白血球の血液型の構成は人種間で大きく異なることから、日本人集団における白血球の血液型の構成の全容解明が強く望まれていました。

しかし、多数の種類 of HLA 遺伝子が存在し、各遺伝子が数十種類以上の配列パターンを有するため、白血球の血液型の組み合わせが膨大な数になってしまうことから、その全体像は明らかになっていませんでした。さらに、HLA 遺伝子配列の構造が複雑で、解読に多額の実験費用と専門的なデータ解析技術が必要となることも、全容解明を遅らせる原因となっていました。

岡田教授らの研究グループではこれまでに、7 個の主要な HLA 遺伝子(=古典的 HLA 遺伝子)を対象に、日本人集団における白血球の血液型の構成を決定し、日本人集団に特異的な白血球の血液型が存在することを報告していました(Okada Y et al. *Nat Genet* 2015)。今回、HLA 遺伝子解析やゲノム情報解析を進めてきた研究者との共同研究により、HLA 遺伝子の構成や病気との関連を詳細に検討しました。

❖ 本研究の成果

岡田教授らは、国立遺伝学研究所(井ノ上逸朗 教授)および金沢大学医薬保健研究域医学系(細道一善 准教授)との共同研究で、最先端のゲノム配列解読技術である次世代シーケンシング技術を駆使することで、**日本人集団 1,120 名を対象に 33 の HLA 遺伝子における 720 種類のゲノム配列を詳細に決定**することに成功しました(図1左)。これまでに解明が進んでいなかった非古典的 HLA 遺伝子も解析対象に含まれているなど、日本人集団における HLA 遺伝子配列のデータベースとして、最大級の情報を含んだ成果になります。

これらのゲノム配列の膨大な組み合わせを効率的に分類する目的で、得られた **HLA 遺伝子ゲノム配列情報に対して機械学習手法の一つである t-SNE を適用**しました。その結果、**日本人集団の白血球の血液型を 11 パターンの組み合わせに分類可能**なことが明らかになりました(図1右)。これは、複雑なヒトゲノム情報の解釈を、機械学習手法を用いて実現した先進的な成功例と評価することができます。

表現型カテゴリ	表現型名	表現型カテゴリ	表現型名	表現型カテゴリ	表現型名
アレルギー疾患	アトピー性皮膚炎 喘息 花粉症	血液検査値	赤血球数 ヘモグロビン濃度 平均赤血球容積 平均赤血球ヘモグロビン値 平均赤血球ヘモグロビン濃度	生化学検査値	アルブミン 非アルブミン蛋白 アルブミン/グロブリン比 血清クレアチニン 推定糸球体濾過量 尿酸 カリウム 無機リン 総ビリルビン アスパラギン酸アミノ基転移酵素 アラニンアミノ基転移酵素 アルカリフォスファターゼ クレアチンキナーゼ 乳酸脱水素酵素
自己免疫疾患	関節リウマチ バセドウ病 1型糖尿病		白血球数 好中球数 好酸球数 好塩基球数 単球数 リンパ球数 血小板数	生理検査結果	収縮期血圧 平均血圧
感染症	B型肝炎 C型肝炎	生化学検査値	総コレステロール HDLコレステロール 中性脂肪 血糖 ヘモグロビンA1c 総蛋白		
心血管障害	心筋梗塞 安定狭心症				
生活習慣病	2型糖尿病 高脂血症				
悪性腫瘍	肺癌 肝臓癌				
臓器疾患	肝硬変 ネフローゼ症候群				
身体測定値	身長 肥満				

図2:日本人集団 17 万人のフェノムワイド関連解析により、白血球の血液型の発症への関与が明らかとなった 52 の表現型。

さらに研究グループは、理化学研究所生命医科学研究センター(鎌谷洋一郎 チームリーダー)との共同研究で、1,120 名で構築した白血球の血液型のパターン情報を学習データとして用いることで、[バイオバンク・ジャパンが構築した日本人集団 17 万人における大規模ゲノムワイド関連解析\(GWAS\)^{※9}のゲノムデータを対象に、白血球の血液型をスーパーコンピューター上で高精度に推定することに成功しました。](#)得られた 17 万人の白血球の血液型パターンに基づき、病気(免疫疾患・生活習慣病・悪性腫瘍、etc.)や量的形質(身長・肥満、血液検査値、生理検査結果、etc.)を含む 100 を超える多彩な表現型との関連を網羅的に調べる、フェノムワイド関連解析を実施しました(図2)。本解析は、アジア人集団で実施されたフェノムワイド関連解析として、サンプル数および表現型数において過去最大規模の解析となります。その結果、これまで報告されていた免疫アレルギー疾患だけでなく、半数を超える [52 の表現型において、白血球の血液型が発症に関与していることが明らかになりました。](#)これは、今まで想定されていたより広範囲の表現型の発症に、白血球の血液型の個人差が密接に関わっていることを示した結果と考えられます。

❖ 本研究成果が社会に与える影響(本研究成果の意義)

本研究成果により、[日本人集団における白血球の血液型の全容が明らかとなり、移植医療への応用が期待されます。](#)機械学習による白血球の血液型の分類に成功したことから、[複雑なヒトゲノム情報に対する機械学習の応用研究におけるマイルストーンとなること](#)が期待されます。さらに、[白血球の血液型を用いて個人の表現型を予測し適切な医療を施す個別化医療の実現に貢献するもの](#)と期待されます。

❖ 用語説明

※1 次世代シーケンス技術

生物のゲノムを構成する DNA 配列を高速に解読する技術。従来のゲノム解読手法であるサンガー法と比較して桁違いのスループットを誇り、幅広い生命科学研究における重要なツールとなっている。

※2 HLA 遺伝子(human leukocyte antigen gene)

ヒトの血球細胞の一種である白血球の表面に発現する分子で、白血球の血液型を規定する。生体内における自己と非自己の認識や外来性の病原菌に対する免疫反応を司り、多彩な表現型の個人差を規定している。主要な古典的 HLA 遺伝子(classical HLA gene)においては生物学的な役割の研究や検査方法の開発が進んでいるが、その他の非古典的 HLA 遺伝子(non-classical HLA gene)については説明が遅れている。

※3 機械学習(machine learning)

コンピューターに高次元データを学習させることで、データの内部に潜む特徴的なパターンを見つけ出すデータ解析手法の総称。人工知能(artificial intelligence)研究におけるデータ解析手法としても広く活用されている。

※4 バイオバンク・ジャパン

日本人集団 27 万人を対象とした生体試料バイオバンクで、東京大学医科学研究所内に設置されている。ゲノム DNA や血清サンプルを臨床情報と共に収集し、研究者へのデータ提供や分譲を行っている。

※5 フェノムワイド関連解析(phenome-wide association study; PheWAS)

特定の遺伝子変異に着目し、多数の表現型との関連を網羅的に検討する解析手法。

※6 表現型

生物の外見や特徴として表現された形態的・生理的性質。代表的なヒトの表現型として、病気や身体的特徴(身長・肥満)、血液検査結果、生理検査結果などが含まれる。

※7 個別化医療

画一的な標準医療でなく、ヒトゲノム情報の違いなど患者さん一人一人の個性を考慮して施す次世代の医療。

※8 tSNE(t-distributed stochastic neighbor embedding)

機械学習手法の一つ。高次元データの次元を効率的に圧縮することで、低次元のデータ(例:2次元の画像データ)に変換する方法。

※9 ゲノムワイド関連解析(genome-wide association study; GWAS)

ヒトゲノム配列上の数百万カ所の遺伝子変異と特定の表現型との関係を網羅的に検討する解析手法。数千人～百万人を対象に大規模に実施され、多数の表現型に対する遺伝子変異が同定されている。

【研究者のコメント】 <岡田随象 教授>

HLA 遺伝子の配列の個人差の解明は、移植医療の推進や個別化医療の実現など、日本人集団におけるゲノム情報の社会実装に不可欠と考えられています。本研究は、これまで HLA 遺伝子を巡る疾患ゲノム研究を主導してきた私たちのチームと、長年にわたって日本人集団の HLA 遺伝子解読技術開発に取り組んできた国立遺伝学研究所および金沢大学、日本を代表するゲノムコホートであるバイオバンク・ジャパンのゲノム情報解析を担ってきた理化学研究所が、共同研究を行うことにより達成することができました。すべての共同研究者や研究支援機構、サンプルを提供してくださった方々に感謝を申し上げます。

❖ 掲載論文

【掲載誌】 *Nature Genetics* (オンライン) 2019 年 1 月 29 日(日本時間)に掲載。

【タイトル】 “Genetic and phenotypic landscape of the MHC region in the Japanese population.”

【著者名】 Jun Hirata, Kazuyoshi Hosomichi, Saori Sakaue, Masahiro Kanai, Hirofumi Nakaoka, Kazuyoshi Ishigaki, Ken Suzuki, Masato Akiyama, Toshihiro Kishikawa, Kotaro Ogawa, Tatsuo Masuda, Kenichi Yamamoto, Makoto Hirata, Koichi Matsuda, Yukihide Momozawa, Ituro Inoue, Michiaki Kubo, Yoichiro Kamatani, Yukinori Okada*.

(* 責任著者)

❖ 特記事項

本研究は、日本医療研究開発機構(AMED)が支援するゲノム研究バイオバンク事業「オーダーメイド医療の実現プログラム(疾患関連遺伝子等の探索を効率化するための遺伝子多型情報の高度化)」の一環として行われ、文部科学省が推進する新学術領域研究「ゲノム科学の総合的推進に向けた大規模ゲノム情報生産・高度情報解析支援(ゲノム支援)」および「がんの複雑性のシステムの理解を目指した新次元の統合的研究(システム癌新次元)」、大阪大学先導的学際研究機構 生命医科学融合フロンティア研究部門、大阪大学大学院医学系研究科 バイオインフォマティクスイニシアティブの協力を得て行われました。