東北大学
TOHOKU UNIVERSITY

Press Release

2020 年 10 月 1 日

東北大学東北メディカル・メガバンク機構
日本医療研究開発機構

個人情報保護を尊重した高精度なゲノム配列推定法を実現！

深層学習を応用したこれまでにない手法

【発表のポイント】

- 多人数の個人の全ゲノム情報（参照パネル*1）を用いることなく、深層学習*2 技術によって全ゲノム配列を推定する画期的な手法を開発
- 個人情報保護上の懸念を払拭
- 多くの研究機関で高精度な遺伝子型インピュテーション*3 が可能になり、遺伝医学の進展に大きく貢献できると期待

【概要】

全ゲノム配列を、数万～数百万程度の限られた遺伝情報を取得する SNP*4 アレイ解析のデータから推定する遺伝子型インピュテーションは、遺伝医学の研究等で頻繁に用いられる研究手法です。この手法は、多人数の個人の全ゲノム情報（参照パネル）を用いる必要があり、しかも参照パネルは個人情報保護の観点で研究機関の間での共有が困難です。

東北大学東北メディカル・メガバンク機構は、その難点を克服し、たくさんの個人情報の集合体である参照パネルを使わない新たな遺伝子型インピュテーション手法 RNN-IMP（Recurrent Neural Network - IMPutation）法を開発し、従来法と遜色ない高い精度が実現できることを示しました。

今回開発された RNN-IMP 法は、深層学習技術を利用することで従来の手法で用いられていた参照パネルの代わりに個人識別が困難な数値パラメータ情報を用いる形で、遺伝子型インピュテーションをする手法です。本開発により、多くの研究機関においてもこれまでの数理モデルでは困難であった高精度な遺伝子型インピュテーションが可能となることが期待されます。

この成果は米国東部時間 2020 年 10 月 1 日に英国科学雑誌「PLOS Computational Biology」のオンライン版で公開されます。

【詳細】

＜背景＞

ゲノムワイド関連解析*5 や遺伝子変異による疾患発症リスク予測では、ゲノム情報取得のための計測技術としてSNPアレイが一般的に用いられます。SNPアレイでは、各個人の遺伝子変異情報の計測を安価に行うことができる一方、取得できる遺伝子変異情報は予め設計されたものに限られます。このため、SNP アレイを用いた解析では、計測された遺伝子変異情報から未観測の遺伝子変異情報を推定する遺伝子型インピュテーションにより擬似全ゲノム配列の推定を行うことが一般的です。

従来の遺伝子型インピュテーション手法では、Li and Stephens モデル*6 と呼ばれる遺伝学の理論モデルに基づき、参照パネルを構成する多検体の全ゲノム配列を用いて未観測な遺伝子変異情報の推定が行われています(図 1)。

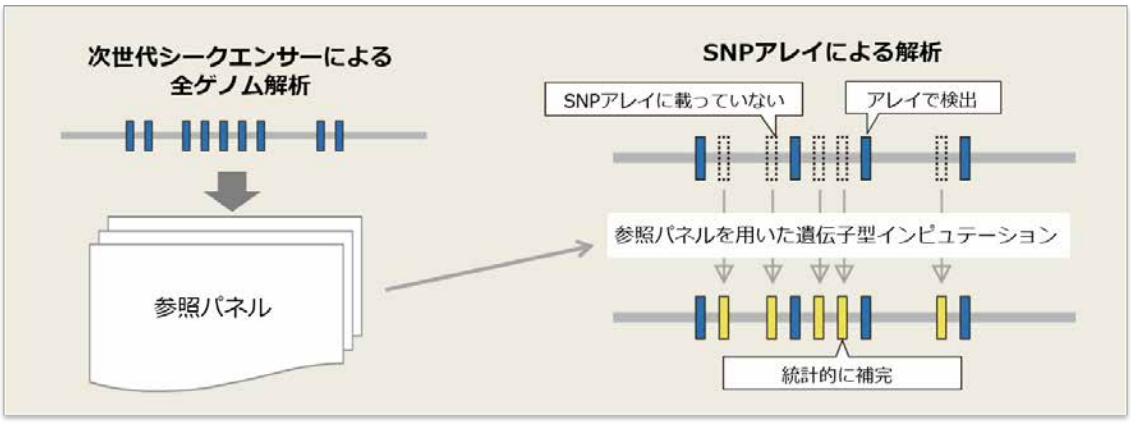

図 1. 左図の全ゲノム解析に対して右図の SNP アレイ解析では解析できる変異情報が限られますが、参照パネルを用いた遺伝子型インピュテーションにより統計的に情報を補完することができます。

また、対象となる民族集団と参照パネルとして用いる民族集団の遺伝的背景が近く、参照パネルに含まれる検体数が多数であるほど推定精度が高いことが知られています。例えば、日本人集団を対象とした場合、日本人集団から構成される参照パネルを用いることが望ましく、ヨーロッパ集団等の他集団から構成される参照パネルを用いた場合、推定精度は限られたものとなります。しかしながら、個人情報保護の観点から、各研究機関で作成された参照パネルを異なる研究機関で共有することや、一般に公開することは困難です。また、1000 人ゲノムプロジェクト*7 では 2,504 検体から構成される参照パネルが公共データとして公開されていますが、その中で、東アジア集団は 504 検体にとどまることから、日本人をはじめとする東アジア集団を対象とした遺伝子型インピュテーションの精度は限られたものとなっていました。

＜今回の成果＞

この度、参照パネルの情報を個人識別が困難な形で保持し、遺伝子型インピュテ

ーションを行う RNN-IMP 法を開発しました。RNN-IMP 法では、参照パネルの情報を数理モデルのパラメータとして学習することで、遺伝子型インピュテーションを行います。このため、参照パネルの情報は個人識別が困難な数値パラメータ情報として保持され、公共データとして共有することが可能です(図 2)。

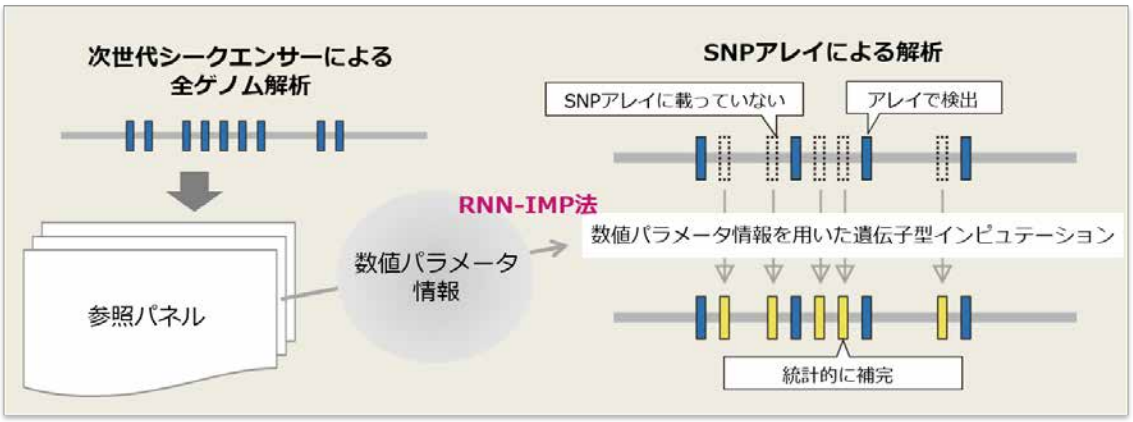

図 2. RNN-IMP 法は、参照パネルではなく数値パラメータ情報を用いて遺伝子型インピュテーションを行います。

これまで、サポートベクターマシン*8 等の従来のパターン認識で用いられてきた手法を数理モデルとして参照パネルを学習する遺伝子型インピュテーション手法は提案されていますが、Li and Stephens モデルに基づく手法と比べて遺伝子型インピュテーションの推定精度が大きく低下する問題がありました。RNN-IMP 法では、近年、自然言語処理、音声認識、動画像解析で用いられている人工知能技術の一つであるリカレントニューラルネットワーク*9 を数理モデルとして用いており、さらに最新の人工知能研究の知見を導入することで推定精度の向上が図られています。

1000 人ゲノムプロジェクトで公開されているゲノム配列データを用いた検証の結果、Li and Stephens モデルに基づく代表的手法と比較してサポートベクターマシンを用いた手法では遺伝子型インピュテーションの推定精度が大きく下がっている一方、RNN-IMP 法では同等の推定精度で遺伝子型インピュテーションが実現できていることが分かりました(図 3)。なお、本検証では、遺伝子型インピュテーションの評価指標として、真の遺伝子変異情報と推定された遺伝子変異情報の相関係数の二乗値 R^2 を評価に用いました。また、参照パネルの一部について個人識別が困難な形でしか利用できない条件のもとでは、RNN-IMP 法が既存の手法よりも高い推定精度で遺伝子型インピュテーションが可能であることを確認しています(図 4)。

個人情報保護の観点からこれまで共有が困難であった参照パネルについて、RNN-IMP 法により数値パラメータ情報として共有して利用することが可能となるため、独自の参照パネルを持たない研究機関においても適切な参照パネル情報を用いた形で高精度な遺伝子型インピュテーションが可能となることが期待されます。

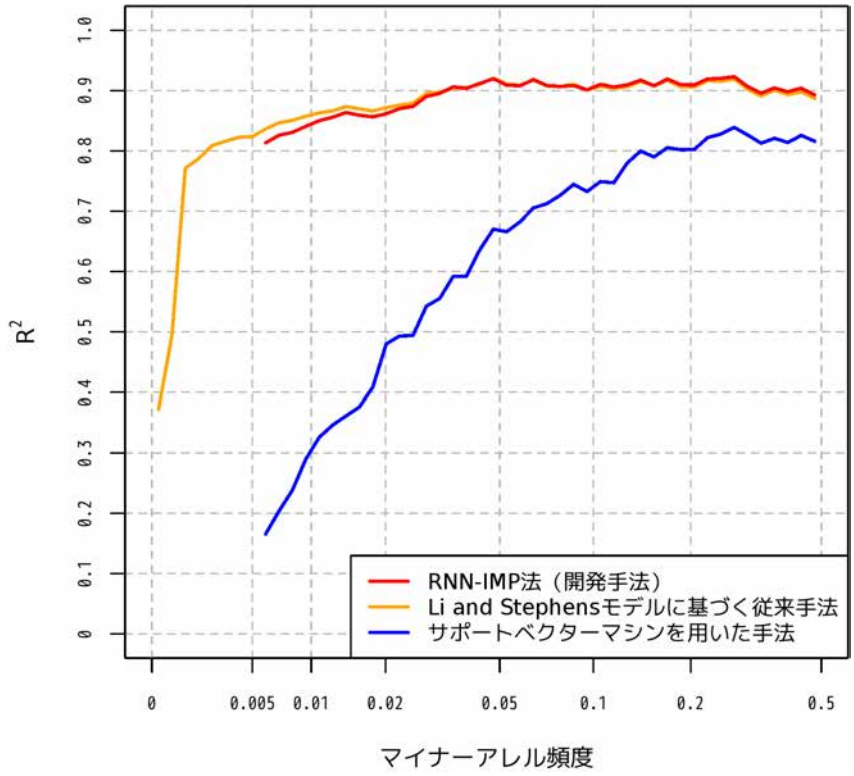

図 3. 1000 人ゲノムプロジェクトで公開されている参照パネルにより、RNN-IMP 法、Li and Stephens モデルに基づく従来手法、サポートベクターマシンを用いた手法と比較を行いました。どの手法もマイナーアレル頻度の低い遺伝子変異では遺伝子型インピュテーションの推定精度が下がる傾向にありますが、RNN-IMP 法と Li and Stephens モデルに基づく従来手法では同等の精度である一方、サポートベクターマシンを用いた手法では精度が大きく下がっていることが確認されました。

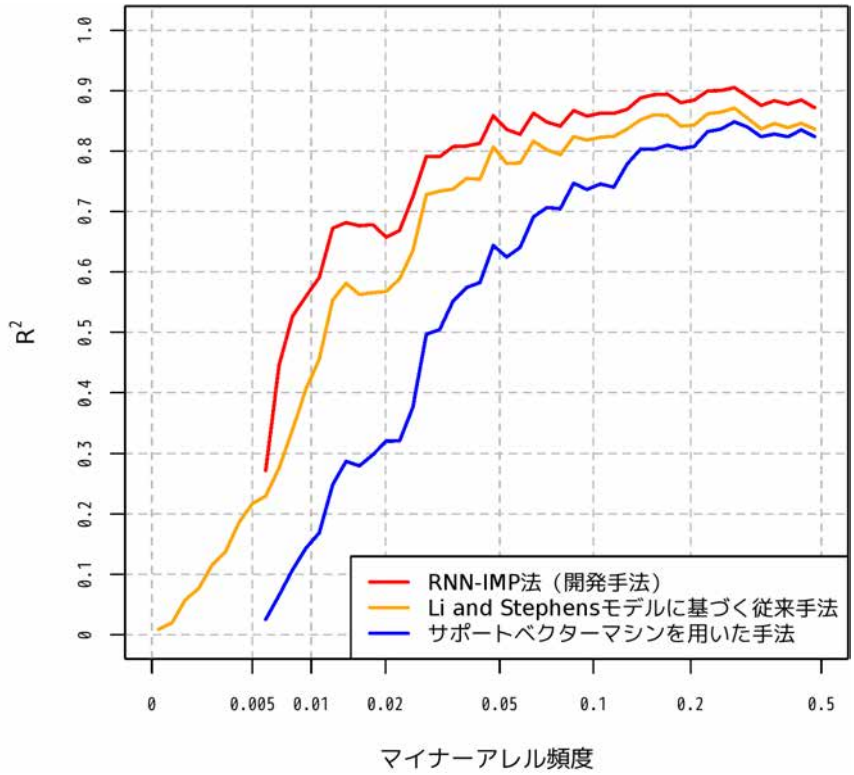

図 4. 1000 人ゲノムプロジェクトで公開されている参照パネルについて、東アジア人集団の参照パネル情報が個人情報保護の観点から個人識別が困難な形でしか利用できない条件のもと、東アジア人集団を対象とした遺伝子型インピュテーションの推定精度を検証しました。検証の結果、Li and Stephens モデルに基づく従来手法では参照パネルを全て利用できないことから遺伝子型インピュテーションの推定精度が低下しており、個人識別が困難な形で全ての参照パネル情報の利用が可能な RNN-IMP 法がより高い精度での遺伝子型インピュテーションが可能であることを確認しました。

＜今後の展望＞

本成果により、個人情報保護上の懸念を払拭して、多くの研究機関で高精度な遺伝子型インピュテーションが可能になり、遺伝医学の進展に大きく貢献できると期待されます。

なお、本研究は、日本医療研究開発機構（AMED）が研究支援する「東北メディカル・メガバンク計画」の一環として行われ、さらにゲノム医療実現推進プラットフォーム事業　研究開発課題「AMED が行うゲノム医療研究支援サービスを支える研究開発基盤の整備」 の支援も受けています。

【参考】

＜東北メディカル・メガバンク計画について＞

東北メディカル・メガバンク計画は、東日本大震災からの復興事業として平成 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ます。東北大学東北メディカル・メガバンク機構と岩手医科大学いわて東北メディカル・メガバンク機構を実施機関として、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ます。

東北メディカル・メガバンク計画は、平成 27 年度より、AMED が本計画の研究支援担当機関の役割を果たしています。

【用語説明】

*1 参照パネル：数千以上の大規模検体についての全ゲノム配列から構成される遺伝子変異情報。遺伝子型インピュテーションを行う上で必要となる。参照パネルを構成する各検体の全ゲノム配列情報は、次世代シークエンサーにより計測されることが一般的である。なお、ToMMoより公開されている全ゲノムリファレンスパネルは、数千人分の変異情報を頻度情報として公開しているものであり、ここでいう「参照パネル」とは異なる。

*2 深層学習：ニューラルネットワークと呼ばれる脳内のニューロンを模した数理モデルを多層で構成することで高い精度でのパターン認識を可能とする人工知能技術。2012 年に開催された画像認識の精度を競うコンテストにおいて従来の手法と比べ大幅な精度向上が可能であることが認知され、研究が急速に活発化し、現在の人工知能研究における主流の技術となっている。

*3 遺伝子型インピュテーション：観測済みの遺伝子変異情報から未観測の遺伝子変異情報を推定する手法。

*4 SNP アレイ：全ゲノム配列上に存在する一塩基多型（Single Nucleotide Polymorphism: SNP）を主とした遺伝子変異情報を計測する手法。計測できる遺伝子変異情報は、予め設計されたものに限られるが、ハイスループットシークエンサーよりも安価に計測が可能であることから、大規模検体を対象とした遺伝子変異情報の取得に用いられる。

*5 ゲノムワイド関連解析：身長、血圧、疾患の有無等の表現型情報について一般的に数千検体以上の大規模検体の全ゲノム配列の遺伝子変異情報を網羅的に解析することで、関連する遺伝子変異の同定を行う解析手法。

*6 Li and Stephens モデル：ある個人の全ゲノム配列はその他の個人の全ゲノム配列の組換えと少数の突然変異で表現が可能であるとする遺伝学の理論モデル。既存の遺伝子型インピュテーション手法はこの理論に基づき、参照パネル内の全ゲノム配列の組換えで対象検体の全ゲノム配列を推定し、遺伝子変異情報の推

定を行っている。

*7 1000 人ゲノムプロジェクト：複数の民族集団から構成される 1,000 を越える検体を対象とした、全ゲノム配列情報の計測と遺伝子変異情報の網羅的な解析を行う国際研究プロジェクト。2008 年 1 月より開始された。現在、2,504 検体について各検体の全ゲノム配列と遺伝子変異情報等が公開されている。

*8 サポートベクターマシン：人工知能研究の一分野である機械学習における代表的なパターン認識手法の一つ。

*9 リカレントニューラルネットワーク：前述のニューラルネットワークの中で音声データ、動画データ、文字列データ等の系列データに対応した数理モデルで、音声信号解析、動画像解析、自然言語処理において用いられる。

【論文題目】

タイトル: A genotype imputation method for de-identified haplotype reference information by using recurrent neural network

日本語タイトル：「個人識別性を排したハプロタイプ参照パネル情報を対象とするリカレントニューラルネットワークを用いた遺伝子型インピュテーション法」

著者：小島 要、田高 周、勝岡史城、田宮 元、山本雅之、木下賢吾
掲載誌：PLOS Computational Biology
掲載日：2020 年 10 月 1 日
DOI: 10.1371/journal.pcbi.1008207

【お問い合わせ先】
（研究に関すること）
東北大学東北メディカル・メガバンク機構
ゲノムプラットフォーム連携センター
センター長　木下 賢吾（きのした けんご）
電話番号：022-274-5952
E メール：kengo@ecei.tohoku.ac.jp

（報道担当）
東北大学東北メディカル・メガバンク機構
長神 風二（ながみ ふうじ）
電話番号：022-717-7908
ファクス：022-717-7923
E メール：pr@megabank.tohoku.ac.jp

（AMED 事業に関すること）
日本医療研究開発機構（AMED）
ゲノム・データ基盤事業部 ゲノム医療基盤研究開発課
電話番号：03-6870-2228
E メール：tohoku-mm＠amed.go.jp