日本医療研究開発機構　　ゲノム創薬基盤推進研究事業
事後評価報告書

公開

I 基本情報

研究開発課題名：（日本語）大規模集団ゲノムデータを利用した遺伝子発現制御文法の機械学習による、VUS 病原性の網羅的評価と実験検証

（英　語）Machine learning-based annotation of gene expression regulation grammar, applied to comprehensive evaluation of VUS pathogenicity and experimental validation

研究開発実施期間：令和　4 年 4 月 22 日～令和　7 年 3 月 31 日

研究開発代表者　氏名：(日本語) 岡田随象

（英　語）Yukinori OKADA

研究開発代表者　所属機関・部署・役職：

（日本語）国立研究開発法人理化学研究所・生命医科学研究センター・チームリーダー

（英　語）Team Leader, RIKEN Center for Integrative Medical Sciences

II 研究開発の概要

研究開発の成果およびその意義等

本研究は大規模計算/実験手法による、variant of uncertain significance (VUS)の遺伝子発現制御を伴う病原性を統一的に評価できる基盤作りを目的とした。遺伝性疾患診断において全ゲノムシーケンス(WGS)から同定される機能未知の変異(VUS)のうち、アミノ酸配列やスプライシングの変化を生じないものに関する統一的な評価は未だ存在せず、WGS の結果の解釈は難を要する。本研究では、VUS のうち近傍遺伝子発現の制御を通じて病原性を有する変異が一定数存在することに注目し、(1)集団ゲノム/エピゲノムと深層学習を利用した遺伝子発現制御文法の包括的な理解 (2)大規模並列レポーターアッセイ(MPRA)を利用した複数細胞種での実験的検証を通じて、遺伝子発現制御に関する一般的な予測モデルの構築、遺伝性疾患診断への適用までを行い、VUS から真に病原性のある変異を同定するパイプラインの確立を目指すものとして始まった。本研究により、遺伝子発現制御を介して病原性を有する変異の新規多数同定が期待された。また、得られたスコアは公開し、遺伝性疾患診断に広く利用されることを目指すものとして行った。

主な成果を以下に記す:

(1) eQTL ファインマッピング手法及び予測モデルの向上と適用

GTEx に存在する eQTL の統計的 fine-mapping 及び、配列やエピゲノム情報を入力とする学習器を訓練する操作を繰り返し、制御活性を持つ確率を定量する学習器を構築が行われた。学習器による予測スコアが、遺伝子制御活性の予測において従来手法の多くを上回ることが検証され、予測スコアを利用することで複雑疾患の原因

となりうるゲノム変異を従来手法と比べて効率よく捜索出来ることが示された。これらの内容は論文として集積され現在投稿中（In Review)である。

(2) 新規同定された制御活性変異の MPRA による検証

(1)で予測された万単位の制御活性を持つ変異に関して、大規模並列レポーターアッセイ MPRA を用い変異によるエンハンサ活性の変化を大規模並列的に定量が行われた。また、二年目以降には、同様手法を拡張して、遺伝子発現制御機能を持つ変異の範囲として、血液疾患との相関が知られる変異や、身長との関連が知られる変異、そして天然には存在しない複数変異の組み合わせ等を考慮しそれらに対する MPRA とその結果の評価が行われた。

また、活用する細胞種として、従来より広く利用される HepG2, K562 等の細胞種に加えて、Chondrocyte を利用することで、従来手法では取得出来ない細胞種特異的な結果を取得することを試みた。

MPRA により制御活性の定量により、遺伝統計的に予測される fine-mapped 変異が実際に細胞種特異的な制御活性を示す確率が有意に高いことが実験的に示された。この結果を含む MPRA データの解析結果は、論文の一部としてまとめられ出版された（Wang et al Nature Genetics 2024)。

(3) (1～2)で得られたスコアの可視化と二次利用のための整備

本研究で得られたスコアや MPRA 結果を可視化するためのプラットフォームとして、Japan Omics Browser (JOB; https://japan-omics.jp/)を整備した。JOB においては、機械学習等に関する専門知識を有さなくても簡便なクリックにより研究者が興味のある遺伝子や変異に関してスコアが取得可能なよう設計されており、公開から 1 年程度で 1000 近くのアクティブユーザを獲得する等、広く二次利用がされている。また、JOB の構築や使用プロトコルをまとめた論文が研究期間終了直後(2025/05)出版された（Takahashi et al BMC Genomics 2025)。

The research aimed to construct a basis for pathogenicity evaluation of variants of uncertain significance (VUS), utilizing large-scale computational and experimental approaches.

Among many VUS that are identified via whole-genome-sequencing (WGS) of disease patients in clinical genomics, non-coding VUS are relatively hard to evaluate. Assuming that a number of VUS could present their pathogenicity through regulation of nearby gene expression, this research aimed to provide a general model for gene expression regulation with application to clinical genomics setting and to provide a pipeline for pathogenic regulatory variants from VUS.

We set two key methods in our research: (1) fundamental understanding of gene regulatory grammars through deep learning of genome and epigenome data, and (2) massively parallel reporter assay in multiple cell lines enabling large-scale experimental validations. In addition, we planned to make the results public, for wider use in the genetics community.

Main results:

(1) Improved fine-mapping resolution of eQTLs
We trained on statistically fine-mapped eQTLs in GTEx while allowing wide range of epigenetic features. We showed that the output score has higher prediction accuracy compared to alternative tools in predicting regulatory variants, and showed that we can utilize the score for prioritization of complex-trait causal variants. The publication summarizing the result is under review.

(2) Massively parallel reports assay
We surveyed hundreds of thousands of predicted regulatory variants identified in (1). We also included blood-trait associated variants as well as height-associated variants in the list, and expanded our cell types of investigation to include chondrocyte. In addition, we included variants pairs that are un-observed in human population.
Our analysis showed that statistically fine-mapped eQTLs are enriched for MPRA-validated variants. These results were included as part of the Nature Genetics paper (Wang et al 2024).

(3) Browser for public use of our results
We implemented the Japan Omics Browser (JOB: [JOB](https://japan-omics.jp/)) for visualization of our machine-learning and MPRA results. JOB is a user-friendly platform not requiring any machine-learning knowledge, and has accumulated near 1,000 active users in a year after release. The manuscript summarizing the protocols and details of JOB is published after the research period (2025/05; Takahashi et al BMC Genomics 2025)