182. 深層学習による、遺伝子発現変化原因変異の網羅的同定

王　青波

大阪大学 大学院医学系研究科 遺伝統計学教室

Key words：eQTL，fine-mapping，マルチタスク深層学習，Transformer，遺伝子発現制御

緒　言

ゲノムワイド関連解析（GWAS）の発展により、ヒトゲノムの非コード領域に存在する多くの変異と種々の疾患の関連が示唆されているが、どの変異が因果的に寄与する変異であるかの予測は困難を極める。本研究では、従来手法では解釈が難しいヒト非コード変異を、近傍遺伝子の細胞種特異的発現制御という文脈で包括的に理解し、その応用として人種間での遺伝子発現機構の違いや形質への影響を解き明かすことを目的とした。マルチタスク深層学習の利用、ロス関数の改良、そして新規特徴量の利用により、遺伝子発現制御変異の予測において従来手法を上回る精度を達成した。また、遺伝的背景の異なる集団においても遺伝子発現制御変異の予測モデルは高い精度で適用可能であることを示した。

方　法

1. 学習データ

欧米人集団の代表的な eQTL（expression quantitative loci：遺伝子発現に関連するゲノム領域）データベースである GTEx に関して、我々が 2 種類の統計的 fine-mapping 手法（SuSiE, FINEMAP）を適用し、因果的変異である確率（Posterior Inclusion Probability：PIP）を割り当てたデータ[1]を学習対象とした（SuSiE, FINEMAP のうち小さい方の PIP を割り当てている）。各遺伝子-変異-組織ペアに関して、PIP＜0.0001 であるものをランダムにサンプルしたものをネガティブデータとしている。ポジティブデータとしては、PIP＞0.9 で定義したが、学習器の訓練に関しては、方法 3 に示すように先行研究と異なり、やや確度の低いものも含み PIP＞0.1 を閾値として学習データに含んだ。

2. マルチタスク深層学習

Keras を用いマルチタスク深層学習器を実装、訓練した。モデルアーキテクチャは図 1 に示す通りで、50 エポックの学習を行い、学習器を構築した。性能評価に関しては、ランダムな 90%を訓練データ、10%をテストデータとした。

3. ロス関数の改良

学習器の訓練においては、ポジティブラベルを PIP＞0.1 と緩く定義し、不正解時のロスを PIP に応じた値とすることで、PIP の大きさに応じたペナルティの大きさを表現し、ラベル間の不均衡を調整する重みづけも行った。訓練自体は二値分類問題であり、具体的には以下のように定義された（n_{neg}はネガティブデータのサンプル数、$\hat{y}_i$はモデルの出力）。

ラベリング:

$y_i = 0$ if $PIP_i < 0.0001$

$y_i = 1$ if $PIP_i > 0.1$

重み付け:

$$w_i = \frac{\sum_{i \mid y\_i = 1} PIP_i}{n_{neg}} \text{ if } y_i = 0$$

$$w_i = PIP_i \text{ if } y_i = 1$$

ロス関数:

$$loss = -\sum_i w_i \{y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i})\}$$

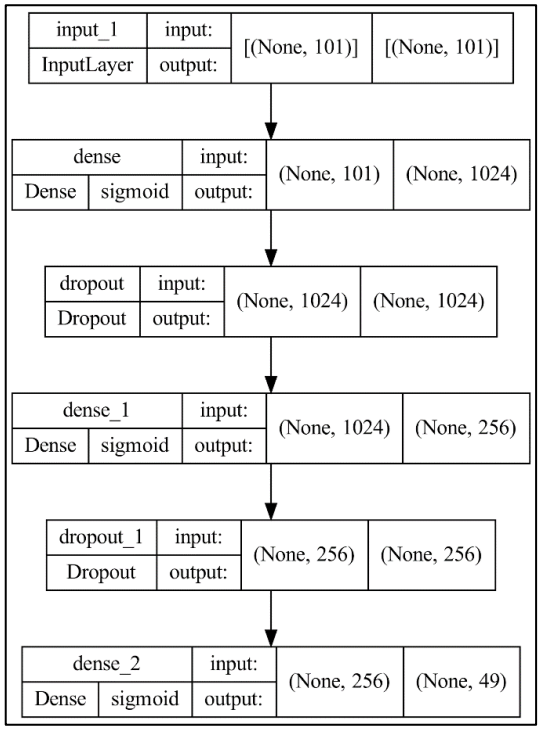

図 1. モデルアーキテクチャ

今回予測に用いたモデルの構成を示した。

4. 新規特徴量の付加

Enformer［2］における主成分の前半 100 次元、及び GTEx が公開する変異-遺伝子間の距離［3］を特徴量として用いた。その他、遺伝子発現量平均や、遺伝子 constraint、Activity By Contact（ABC）値等を特徴量として付加することを試みたが、予測能の向上は見られなかったため、上記のように単純な特徴量の構成となった（データ未公開）。

5. 日本人集団 eQTL データの構築と利用

我々の先行研究［4］にて、COVID-19 陽性である日本人集団 465 人の genotype 及び末梢血 RNA-seq データを用い、eQTL call 及び統計的 fine-mapping を行った（Japan COVID-19 Task Force：JCTF)。Fine-mapping 手法は［1］と同様であった。ポジティブ、ネガティブデータに関しては従来と同様に PIP＜0.0001、PIP＞0.9 で定義した。Functionally-informed fine-mapping に関しては、計算された PIP を post-hoc にスコアで重み付けして補正する形で実行した。

6. 性能評価

方法 1 に述べたとおり、GTEx においてネガティブをランダムに 100,000 変異遺伝子取得し、ポジティブと合わせたデータセットにおいて 10 のスプリットに分ける操作を行った。9 つのスプリットから学習し、残りのスプリットにおける予測を行うタスクを 10 回繰り返し、性能評価を行った（JCTF においても同様)。Area Under Receiver-Operating Characteristic（AUROC）及び Area Under Precesion-Recall Curve（AUPRC）値を評価値とした。

結　果

1. 遺伝子発現制御能予測スコア（Expression Modifier Score v2）の構築

GTEx においてサンプル数の比較的少ない脳関連組織において、マルチタスク深層学習と、各組織で別々に学習を行なった際の性能評価の比較を図 2a に示す。単一組織における学習と比べて、脳関連組織をまとめたマルチタスク深層学習、さらにはその他の全組織を含んだ学習によって精度が向上することが示された。

次に、訓練データにおけるポジティブサンプルの閾値を PIP＝0.9 から順々に下げた際の予測能の推移を図 2b に示す。閾値を下げること、そしてロス関数の改変により予測能の向上が達成できることが示された。

また、用いる特徴量による予測能の違いに関して評価した結果を図 2c に示す（ヴァイオリンは 49 組織での分布のカーネル密度推定に基づく)。Enformer 由来の特徴量が、AUPRC の評価においては転写開始点距離 (TSS) 以上の予測能を持つことが示された。以上により個々の改変が予測能の向上に寄与することを確かめた。

2. Expression Modifier Score v2 性能評価

結果 1 の要素を組み込み、方法に示したモデルアーキテクチャで学習した予測器の予測能を、GTEx 内のデータで評価し、既存の有効な手法（EMS v1）と比較した結果を図 2d に示す。ほとんどの組織（48／49）において AUPRC における評価では上回っていることが確認された。AUROC による評価では、EMSv1 に及ばない組織も多いが、幾つかの組織においては v1 を大きく上回る AUROC が達成されていることが確認出来る。AUPRC を主な評価軸と捉えると、構築したスコア（Expression Modifier Score v2＝EMSv2）が既存スコアに比べ有効であることが示唆される。

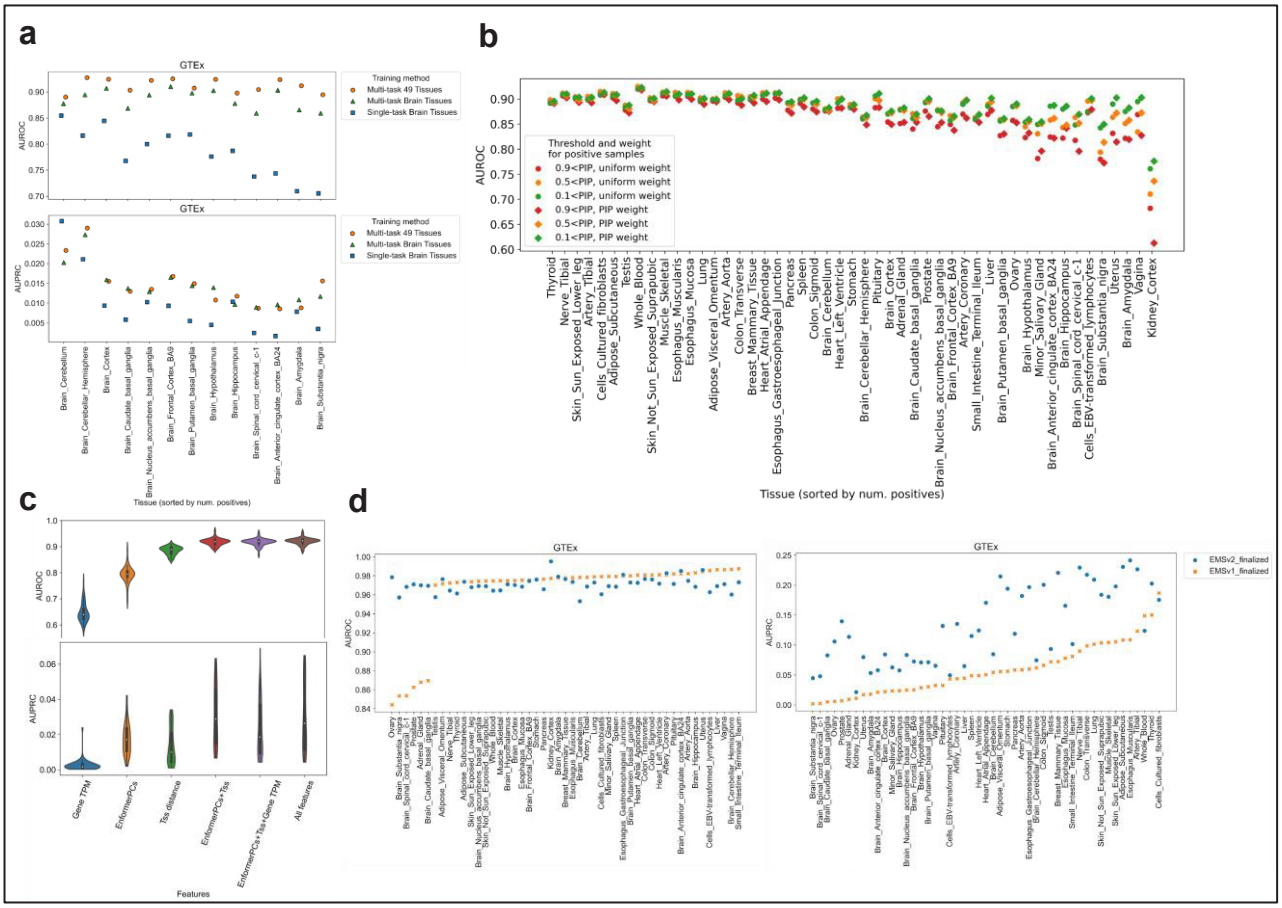

図 2.　Expression Modifier Score（EMS）v2 の構築と性能評価

a）マルチタスク深層学習による性能向上。

b）閾値と下げ、ロス関数を改変することによる性能向上。

c）新規特徴量を利用することによる性能向上。

2）　既存スコア（EMSv1）との性能比較。

3.　Expression Modifier Score v2 の日本人集団データへの適用

JCTF における因果的 eQTL（PIP＞0.9）の予測性能評価を図 3a、b に示す。類似スコア（TSS 距離及び Sei [5]）と比べて、日本人集団データにおいても EMSv2 が高い予測能を示すことが確認された。また、EMSv1 はその計算コストから、GTEx に存在しない変異には適用が出来ないため、評価対象外であり、そうした適用範囲の面でも有効性が示唆される。

次に、予測スコアを事前情報として functionally-informed fine-mapping （実際には post-hoc な重み付け）をした結果を図 3c に示す。EMSv2 を利用することで、fine-mapping 結果が変わりうることが示された。

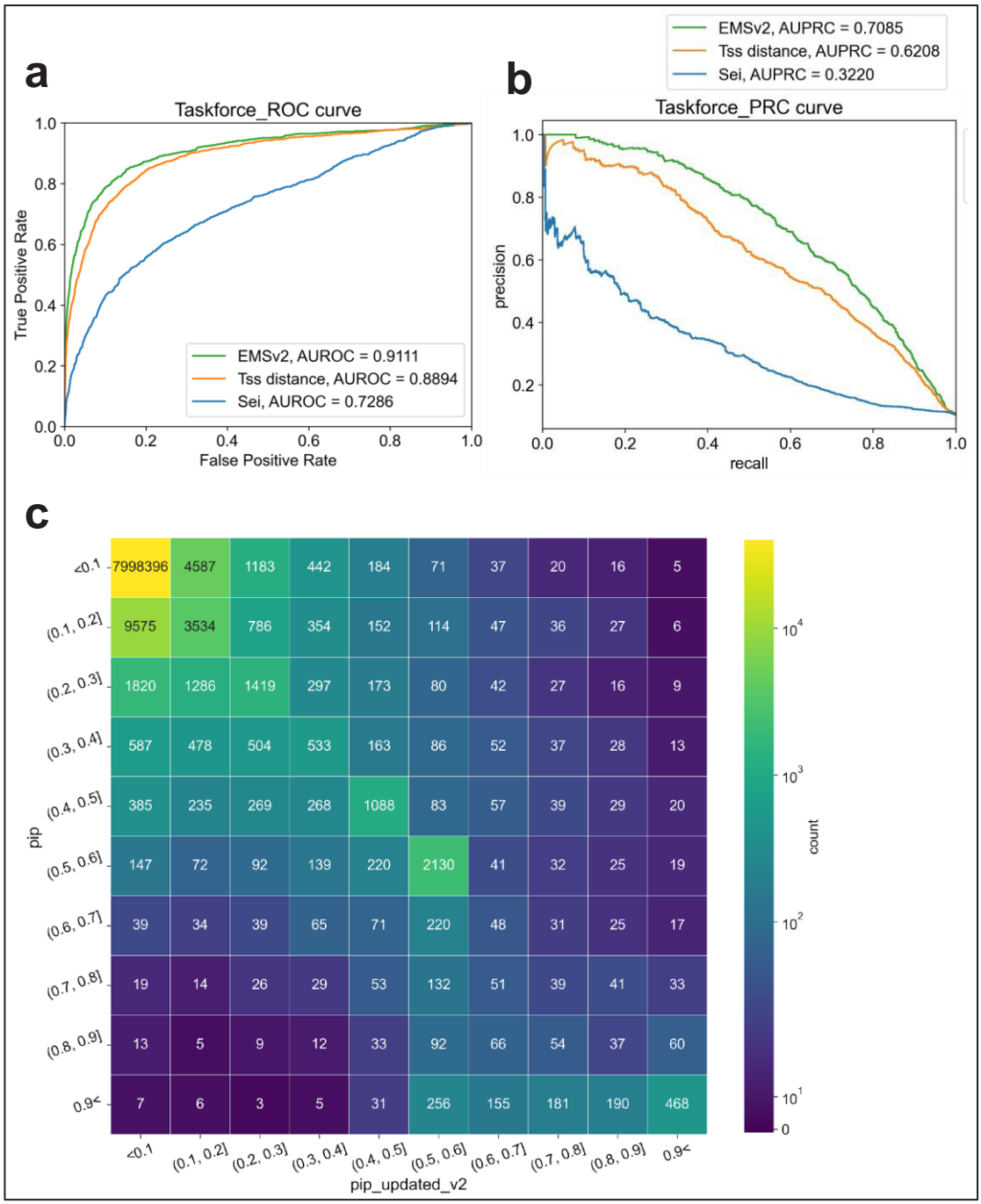

図 3. EMSv2 の日本人集団 eQTL データへの適用

a）日本人集団における因果的 eQTL（PIP＞0.9）予測能の AUROC による評価。

b）日本人集団における因果的 eQTL（PIP＞0.9）予測能の AUPRC による評価。

c）EMSv2 を用い PIP を重み付けした際の元の PIP との比較。

考　察

本研究では、ヒトゲノムの非コード領域に存在する変異が近傍遺伝子発現に制御を及ぼすか否かの評価というチャレンジングなタスクに対して、マルチタスク深層学習の適用、ロス関数の連続性の付加とそれに伴う教師データの増強、そして Transformer と主成分分析を利用した新規特徴量の利用のそれぞれが、従来手法と比べた精

度向上に寄与することが示された。また、それらの要素を全て組み込んだ学習器を構築し、その予測能の向上を評価した。そして、欧米ベースのゲノムデータで学習した学習器を、新規に日本人集団のゲノムデータから同定された制御変異の予測に適用することで、集団の遺伝的背景を超えて、機能情報ベースの制御活性予測が可能になることが示された。

以上の成果はヒトゲノム変異解釈において大きな意義を有するが、同時に課題も多く提示された。(1) マルチタスク深層学習を利用することで、サンプル数の少ない組織での予測能の大きな向上と、全体としての予測能の向上を達成したが、個別の組織で評価すると、血液等、従来よりも予測能が落ちてしまう組織が存在した。これらは主に元々サンプル数が比較的多く、組織にマッチした特徴量も多いといった共通点が挙げられる。即ち、マルチタスク深層学習により、簡単な問いへの正解率を「犠牲」にして、困難な問いへの正解率を上げているという傾向が見られた。(2) 日本人集団においてスコアを事前情報として使った fine-mapping においては、新規に同定された制御変異の数は限られた。即ち、欧米集団で学習したスコアは、多少集団特異的なシグナルを学習している可能性が示唆される。また、(3) 得られたスコアは制御変異の有無を予測するものであり、最終的に疾患や種々の形質に影響を及ぼす変異であるかどうかの判定は行われない。より強固かつ実用的なスコアとするために、アルゴリズムや特徴量の引き続きの改良に加え、予測が失敗する変異に共通する性質の考察や、変異間相互作用の検討 [6]、遺伝子の機能情報を取り入れた疾患関連スコアへの転移学習等が必要と考えられる。引き続き取り組んでいく。

共同研究者・謝辞

DeepMind 社の Ziga Avsec 氏、Calico 社の David Kelley 氏に多くのアドバイスを頂いたことを感謝する。大阪大学大学院医学系研究科遺伝統計学研究室の岡田随象教授及び研究室メンバーの指導と議論に感謝する。また、本研究結果は主に、大阪大学医学部医学科学生の高橋勇伍氏に産生して頂いたため、最大の感謝を表する。

文　献

1) Wang QS, Kelley DR, Ulirsch J, Kanai M, Sadhuka S, Cui R, Albors C, Cheng N, Okada Y, Aguet F, Ardlie KG, MacArthur DG, Finucane HK. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. Nat Commun. Nature Publishing Group; 2021 Jun 7;12(1):3394. PMID: 34099641 DOI: 10.1038/s41467-021-23134-8

2) Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021 Oct;18(10):1196–1203. PMID: 34608324 DOI: 10.1038/s41592-021-01252-x

3) THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. American Association for the Advancement of Science; 2020 Sep 11;369(6509):1318–1330. PMID: 32913098 DOI: 10.1126/science.aaz1776

4) Wang QS, Edahiro R, Namkoong H, Hasegawa T, Shirai Y, et al., The whole blood transcriptional regulation landscape in 465 COVID-19 infected samples from Japan COVID-19 Task Force. Nat Commun. Nature Publishing Group; 2022 Aug 22;13(1):4830. PMID: 35995775 DOI: 10.1038/s41467-022-32276-2

5) Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. Nature Publishing Group; 2022 Jul;54(7):940–949. PMID: 35817977 DOI: 10.1038/s41588-022-01102-2

6) Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, Hill AJ, O'Donnell-Luria AH, Karczewski KJ, MacArthur DG. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun. Nature Publishing Group; 2020 May 27;11(1):2539. PMID: 32461613 DOI: 10.1038/s41467-019-12438-5