118. 非コードゲノム領域でのがんドライバー変異の探索

鬼丸　洸

*名古屋大学 大学院医学系研究科 附属神経疾患・腫瘍分子医学研究センター

Key words：癌ゲノム，畳み込みニューラルネットワーク，非コード領域，体細胞変異，エピジェノム

緒　言

がんはゲノムの変異を要因とする病気であり、個々人のゲノム配列を理解することにより、より正確な診断、治療方針の決定を行うことが出来る。本研究では、現在理解の進んでいない非コード領域のゲノム配列変異とがんの関係について着目する。非コード領域のゲノム配列変異のがん化への寄与に関しては、現在議論を呼んでいる。近年、変異、転座または増幅した転写制御配列によるがん遺伝子の制御異常が、がん化に貢献をしていることが明らかとなった［1］。一方で、2,500 以上のがん組織のゲノム配列解析から、ドライバー変異（がん化に直接的に関わる変異）のうち非コード領域に属するものは 13%のみであることが報告されている［2］。しかしながら、genome-wide association study（GWAS）では、90％以上の表現型に関連する single nucleotide polymorphisms（SNPs）が非コード領域に存在し、非コード領域が生物学的機能において著しい役割を担っていることがわかっている。このため、がん化においても、非コード領域のゲノム配列変異が大きな役割を担っている可能性は非常に高く、その是非を明確にする必要がある。

現在までの研究で、非コード領域のドライバー変異があまり多く見つかっていない原因として、非コード領域の変異が転写制御に及ぼす影響を評価する技術は非常に限られていることがあげられる。本研究では、深層学習を用いた非コード領域のドライバー変異の同定方法の確立を目指した。深層学習は、AI 技術のコアをなす手法で、ゲノム配列を解釈する上でも有用であることを、我々の最近の研究で示している［3］。我々は、独自の畳み込みニューラルネットワークにより、ゲノム配列における転写制御機能を予測する手法を開発し、これにより、エピジェノミックシグナルの有無を予測したり、それぞれの塩基の転写因子結合への寄与度を評価したりするなど、様々なことが可能となった。本研究計画では、こうした非コード領域の機能予測アルゴリズムを、公開されているがんゲノム配列に応用することによって、非コード領域のがん化への貢献度を検証した。1,072 の乳がんゲノムの small nucleotide variants（SNVs）を解析したところ、合計約 1,300 箇所において、CCCTC 結合因子（CTCF）結合が変動している可能性があることが予測された。

方　法

1. SNV データの処理

TCGA における 1,072 の乳がんゲノムの small nucleotide variants（SNVs：塩基置換および小規模のゲノムの欠損、挿入を含む）データをダウンロードし、すべてのデータを一つの vcf file にマージした。このデータを、ヒトゲノム配列（GRCh38）を基に vcftools 0.1.16 によって塩基配列に変換した。SNVs を塩基配列に置換した後、ゲノム配列を 1,000 塩基毎に区切り、A、G、C、T、N の塩基をそれぞれ（1000）、（0100）、（0010）、（0001）、（0000）の one-hot vector に変換することで、4×1,000 のテンソルデータに変換し、畳み込みニューラルネットワークのインプット配列とした（図 1a）。また、比較対象として、レファレンスゲノム配列自体も同様の処理を行った。今回はレファレンスゲノム配列にある R や W といったイレギュラーな塩基配列については N と同等の扱いをしたが、今後、これらの配列をどのように扱っていくかについては検討の余地がある。

*現在の所属：株式会社 ちとせ研究所

2. 深層学習モデル

深層学習モデルとして、畳み込みニューラルネットワーク（CNN）を採用した（図 1b）。特に我々が独自開発した forward- and reverse-sequence scan（FRSS）レイヤーにより、畳み込みの第一段階で、ゲノム配列の forward strand と reverse complementary strand を同時に読み込むデザインを採用した（図 1c）。

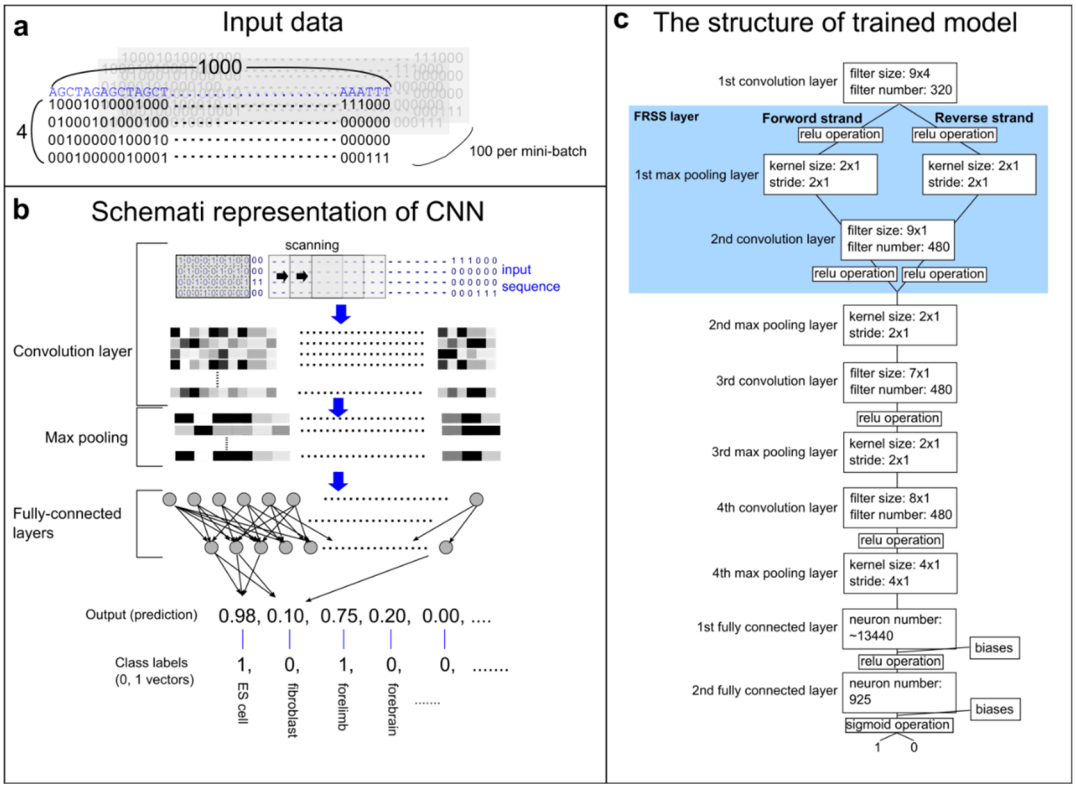

図 1. ゲノム配列を入力とする畳み込みニューラルネットワークの概要

a) 入力データの模式図。1,000 塩基の AGCTN に対して、それぞれ（1000）、（0100）、（0010）、（0001）、（0000）のベクターが割り当てられている。

b) 畳み込みニューラルネットワークによるゲノム配列の読み込みと予測の図。上から、畳み込みカーネルで配列情報を読み込み、Max pooling、Fully-connected layer などを通り、最後に、エピジェノミックシグナルの因子やシグナルのある組織を 0 から 1 のスコアで予測する。

c) 今回の研究で使用した CNN モデルの全容。青でハイライトされた場所が FRSS レイヤーで、ゲノム配列の相補鎖を右側の畳み込みレイヤーで並行して読んでいる。

3. 深層学習モデルの訓練と予測

深層学習モデルの訓練データとして ENCODE プロジェクトから、マウスの 54 個の正常組織や細胞株における CTCF の ChIP-seq データをダウンロードし、MACS2 で peak call を行い、教師データとした。また、訓練の際のレファレンスゲノムは、マウスゲノム（mm10）を用いた。ヒトを用いなかった理由としては、マウスのデータを使用したほうが精度の高い訓練が出来るためである。CTCF を選んだのも同様の理由である。また、今回、モデルをヒトゲノムに適用するため、過学習の影響を最低限に抑える意図もある。

予測では、方法 1 で用意したデータを用いて、レファレンスゲノム配列と乳がんゲノム配列に対して、CTCF の結合予測を行った。NVIDIA の GPU を用いて訓練、および予測を行った。

4. 出力データの後処理

モデルによって予測された、乳がんゲノムにおける CTCF 結合領域のゲノム座標は、DNA 配列欠損や挿入よって、ずれてしまっているので、直接比較することが出来ない。そこで本研究では、乳がんゲノム配列をレファレンスゲノム配列にアラインメントすることで、座標補正を行った。アラインメントには GenAlign（v1.0.22）を用いた。アラインメント結果を chain ファイルに変換し、liftOver で乳がんゲノムにおける CTCF 結合領域のゲノム座標を補正した。その上で 1,000 塩基区切りとなっているゲノム領域のうち、レファレンスと乳がんゲノム間で最もオーバーラップしているもの同士を対応するゲノム区間であるとし、モデルによる予測結果を比較した。

結　果

1. 体細胞変異による CTCF 結合領域の獲得と欠損

訓練した CNN モデルで予測した結果をもとに、予測スコアがレファレンスあるいは乳がんゲノムの少なくともどちらかで 0.5 以上であり、かつ、体細胞変異に起因したスコアの変動が 2 倍以上であるゲノム領域をカウントしたところ、650 箇所の CTCF 結合領域の獲得と 682 箇所の欠損を予測することが出来た（図 2a）。図 2b、c は CTCF 結合領域の欠損と獲得の例である。CNN モデルは複数のマウスの組織、細胞株の CTCF 結合シグナルを基に訓練を受けており、下図は、特にマウスの線維芽細胞であれば CTCF 結合シグナルがあると予測するものである。このため、解釈には十分な注意を払わなければならない。線維芽細胞以外にも、脳や ES 細胞での CTCF 結合を予測しているデータもあるので、どういった組織・細胞株のデータを基に訓練すべきかは、今後の課題となる。

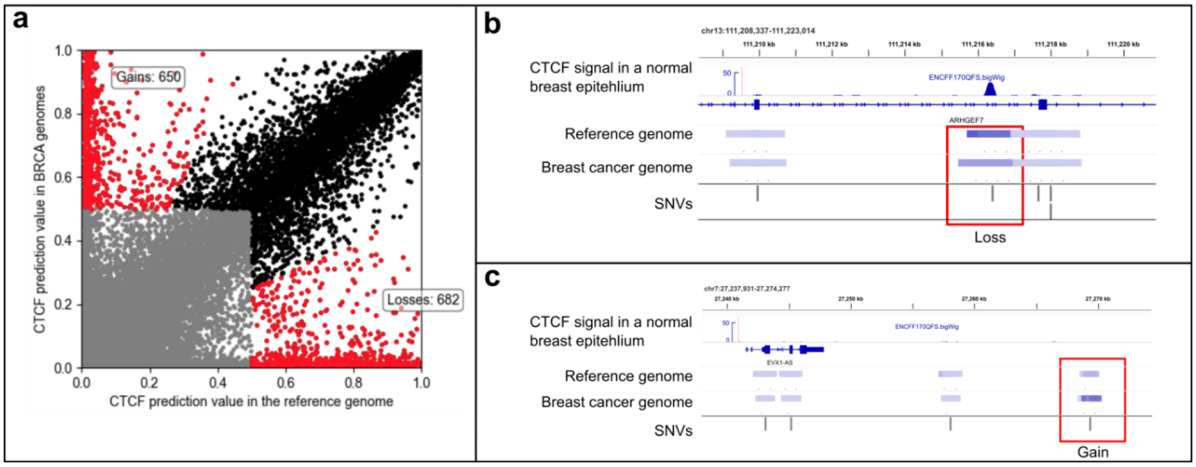

図 2. 乳がんゲノムにおける CTCF 結合領域の潜在的獲得と欠損

a) レファレンスゲノムと乳がんゲノムにおける CTCF 結合予測スコアの比較。1 つの点は 1 つの 1,000 塩基に区切られたゲノム領域に相当する。2 つのゲノム間でともにスコアが 0.5 以下のゲノム領域は灰色、0.5 以上だがスコアが変動しないものを黒、スコアが 2 倍以上変動したものを赤で示している。

b、c) それぞれ CTCF 結合領域の欠損と獲得の例。“CTCF signal in normal breast epithelium”は、ENCODE プロジェクトのヒトの正常な乳房上皮における CTCF の ChIP-seq データである。“Reference genome”はレファレンスゲノムにおける CTCF 結合予測、“Breast cancer genome”は乳がんゲノムにおける CTCF 結合領域の予測である。青色が濃いほど、結合する可能性が高いことを示している。SNVs は乳がんゲノムにおける体細胞変異した塩基を示している。赤いボックスで囲われた部分が予測スコアが変動したゲノム領域を示している。

また、CTCF 結合の欠損、獲得があったと予測された近傍の遺伝子について、gene ontology（GO）解析を行ったところ、細胞のコンタクトや細胞接着に関する GO のエンリッチメントが見られた（図 3）。

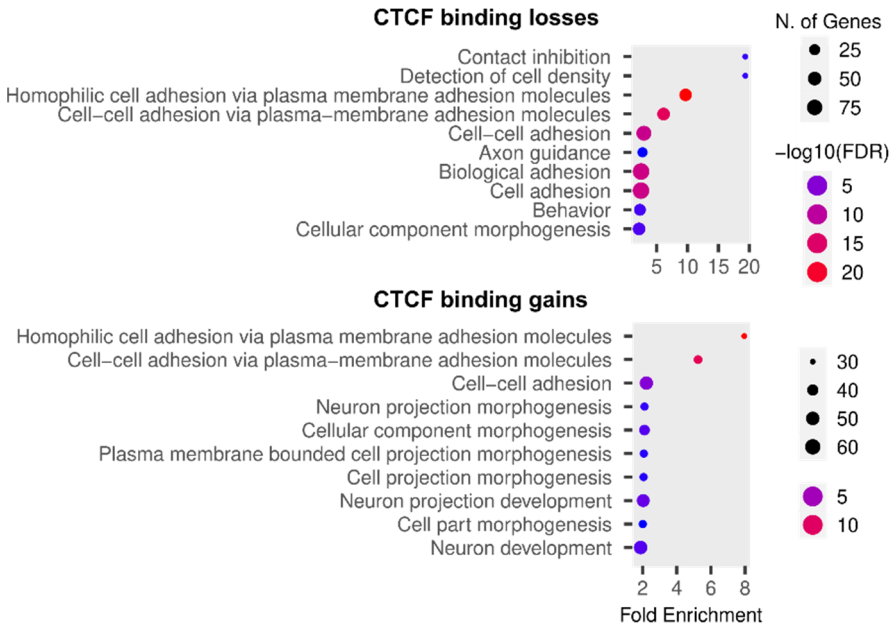

図 3. 乳がんゲノムにおける CTCF 結合領域で変動が予測された遺伝子の GO 解析
CTCF 結合の変動が予測された近傍遺伝子セットにたいして、Biological process の GO 解析を行った結果。上が CTCF 欠損、下が CTCF 獲得遺伝子セット。横軸が Fold enrichment、色が－log10（FDR）、ドットの大きさが該当する遺伝子の数。

考　察

今回、畳み込みニューラルネットワークを応用することで、TCGA の 1,072 の乳がんゲノムから得られた SNVs を基に、およそ 1,300 箇所において、CTCF 結合領域が獲得または欠損することを予測出来た。このことは、深層学習を用いて、癌ゲノムデータに新たな解釈を加えることが出来ることの概念実証となる。しかしながら、1,072 の症例に対して、1,300 箇所という数字は、およそ 2 万から 10 万箇所、CTCF 結合領域が正常ゲノムに存在するのを考えると、比較的少ないようにも思われる。これらの CTCF 結合領域の変動が実際に起こっているのか、起こっている場合に転写制御への影響があるのかといった問いは、今後検討していく予定である。

また、今回の研究で、SNVs が入ったゲノムでは座標がずれるので、エピジェノミック情報を比較解析することに、テクニカルな問題があることがわかった。今回はアラインメント情報をもとに座標の補正を行ったが、この方法は、アライメントを行うプログラムの正確性に依存したり、リピート配列の多いところでは、適用できなかったりするため、方法の改善が必要であるあることがわかった。SNVs の情報が入った vcf ファイルから直接座標補正を行うという方法を今後検討していく予定である。

また、今回は乳がんゲノムと CTCF のみに着目したが、より多くの癌種、エピジェノミック因子を用いて、今後解析を拡大していくことで、癌ゲノムにおける非コード領域の変異の全容が明らかになることが期待される。

共同研究者・謝辞

本研究は、名古屋大学大学院医学系研究科分子腫瘍学研究室の鈴木洋教授の指導の基で行われたので深く感謝を申し上げます。

文　献

1) Tang F, Yang Z, Tan Y, Li Y. Super-enhancer function and its application in cancer targeted therapy. NPJ Precis Oncol. 2020 Feb 12;4:2. DOI: 10.1038/s41698-020-0108-z. PMID: 32128448

2) ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82-93. DOI: 10.1038/s41586-020-1969-6. PMID: 32025007

3) Onimaru K, Nishimura O, Kuraku S. Predicting gene regulatory regions with a convolutional neural network for processing double-strand genome sequence information. PloS One 2020 Jul 23;15(7):e0235748. DOI: 10.1371/journal.pone.0235748. PMID: 32701977