論文の内容の要旨

論文題目　深層学習を用いた HLA imputation 法の開発と Parkinson 病と 1 型糖尿病の原因遺伝子変異解明への応用

氏名　内藤　龍彦

ゲノムワイド関連解析（genome-wide association study, GWAS）は，一塩基変異（single nucleotide variant，SNV）などのマーカー変異を用いてゲノム全体の領域と疾患の有無との相関を網羅的に検定することにより，疾患の発症リスクと関連する感受性領域を検出する遺伝統計学的手法である．GWAS で検出される領域の中でも，6 番染色体に位置し HLA 遺伝子を始めとした免疫機能に関わる多数の遺伝子がコードされている主要組織適合性複合体（major histocompatibility complex, MHC）領域は，自己免疫性疾患をはじめとした多様な疾患の発症に関わる重要な領域である．HLA 分子は大きくクラス I, II に分類され，それぞれ内在性抗原，外来性抗原を CD8 陽性，CD4 陽性 T 細胞に提示して免疫応答を引き起こす．GWAS で検出できるのは大まかな感受性領域のみであるため，疾患の発症と直接的に関連するバリアントを同定するためには，感受性領域を高密度にジェノタイピングしたデータを用いて追加解析を行わなければならない（fine-mapping）．MHC 領域を fine-mapping するためには，GWAS で対象とした個人について HLA 遺伝子配列を網羅的に決定する必要がある．一方，HLA 遺伝子の配列決定は PCR に基づいた手法や次世代シークエンサーを用いた手法により行うが，それらは費用や手間の観点から GWAS で対象とするような大規模な集団に対して適用するのは現実的ではない．このような中で開発された遺伝統計学的手法が HLA imputation 法である．HLA imputation 法とは，予め構築した SNV ジェノタイプと HLA アレルからなる参照パネルを用いることで，SNV ジェノタイプのみから HLA アレルを統計学的に推定する手法である．HLA imputation 法により MHC 領域の fine-mapping が可能となり，様々な疾患で MHC 領域におけるリスク関連バリアントが同定されるようになった．

一方，このような研究を行っていく中で，報告されるリスク関連 HLA バリアントが人種集団間で異なるということがしばしば見られるようになった．例えば 1 型糖尿病（type I diabetes, T1D）においては，欧米人集団では HLA-DRβ1 pos. 13，71，HLA-DQβ1 pos. 57 の 3 つのアミノ酸位置が主要なリスク関連 HLA バリアントとして報告されたが，日本人集団ではそのような関連は認めない．これは，MHC 領域が他の遺伝的領域に比して長大な連鎖不平衡を有し単一の集団における fine-mapping が困難であることや，集団により異なったアレル頻度分布を呈することに起因すると考えられる．異なった人種集団間に共通して普遍的に発症リスクに関わるバリアントを同定するには，異人種集団間の統合解析（trans-ethnic fine-mapping）が必要と考えられた．一方，従来の HLA imputation 法では希少アレルでの予測精度が著しく低下してしまうため，人種集団の統合解析を行う際に信頼性が低下してしまうという難点があった．これは，HLA アレルの頻度分布が集団間で大きく異なるため，ある集団で頻度が低く imputation 精度も

低いアレルをフィルタリングすることで，他の集団においては頻度の高いアレルも解析対象から外れてしまうということが起こりうるために生じる問題である．このように HLA imputation 法の精度の改善が望まれる状況の中で，本研究では深層学習に着目した．深層学習は多層ニューラルネットワークを用いた機械学習法の一種であり，複雑な特徴量の組み合わせを直接学習して予測することができる．深層学習モデルの一種である畳み込みニューラルネットワークが画像認識の分野で他の機械学習法を大きく上回る精度を達成したことを皮切りに注目を集めるようになり，翻訳や音声認識など多様な分野で応用されるようになった．本研究では，深層学習を応用した新規の HLA imputation 法（DEEP*HLA）を開発し，網羅的な性能評価を行うとともに，実際の GWAS ジェノタイプデータへ適用することで MHC 領域における trans-ethnic fine-mapping を行った．

本研究で開発した DEEP*HLA は，SNV ジェノタイプを入力データとし，複数の HLA 遺伝子のアレルの予測値を同時に出力する，マルチタスク畳み込みニューラルネットワークを利用したモデルである．従って，SNV ジェノタイプと HLA アレルの組み合わせである参照パネルを予め学習することで，SNV ジェノタイプのみから HLA アレルを予測することができる．精度評価は，日本人集団用の参照パネル（1118 例），欧米人集団用の参照パネル（5122 例）を用いて交差検証法により行った．また，日本人集団用の参照パネルを用いて独立サンプル（918 例）の imputation を行った場合の精度検証も行った．比較のために，従来の HLA imputation 法として SNP2HLA と HIBAG の精度評価も行った．結果，いずれの検証においても DEEP*HLA は既存手法と比較して，特に頻度が低い希少アレルにおいて精度の上昇を認めた（頻度 1%未満のアレルの陽性適中率において，各データセットでそれぞれ DEEP*HLA では 0.799，0.877，0.863 で，SNP2HLA は 0.624，0.686，0.640，HIBAG では 0.505，0.736，0.753 であった．さらに日本人集団用参照パネルと欧米人集団用参照パネルを試験的に統合した混合人種集団パネルを用いて 1000 ゲノムプロジェクトの多人種集団データに適用した場合も，多くの人種集団において DEEP*HLA が特に希少アレルの imputation において精度が上回っていた．計算負荷の評価も行ったところ，総実行時間の点ではサンプル数が増えるに従い既存手法と比べて有利であり，またメモリ負荷の点からもバイオバンクレベルのサンプル数に対しても実行可能であった．次に，DEEP*HLA が入力の SNV のどの領域を予測の判断根拠として用いているかを評価するために，深層学習モデルの入力感度マップの可視化手法である SmoothGrad を適用した．結果，DEEP*HLA は単純な連鎖不平衡に依らない様々な SNV を予測の判断根拠とし，HLA アレルと SNV との複雑な相関関係を学習して予測を行っている可能性が考えられた．

次に DEEP*HLA を用いて T1D の MHC 領域における trans-ethnic fine-mapping を行い，先述の異人種集団間のリスク関連 HLA バリアントの不均一性の問題に取り組んだ．T1D はインスリン産生膵臓 β 細胞の T 細胞介在性破壊によりインスリン分泌不全をきたし高血糖を来す自己免疫性疾患である．BioBank Japan，UK Biobank の両コホートの SNV ジェノタイプデータにそれぞれ日本人集団用参照パネル，欧米人集団用参照パネルで学習した DEEP*HLA を適用して HLA imputation を行った．両集団の imputation 結果を統合の上，trans-ethnic fine-mapping を行った（計

PD 患者 1563 例と対照 415,283 例）．結果，HLA クラスⅡ遺伝子の HLA-DRβ1 pos. 71 に最も強い関連を認めた（$P = 7.5 \times 10^{-120}$）．さらに，同領域内では，HLA-DRβ1 pos. 71 に加えて，HLA-DRβ1 pos. 74，HLA-DQβ1 pos. 30，70，185 にも独立した関連を認めた．また，HLA クラスⅠ遺伝子として，HLA-A pos. 62，HLA-B*54:01 にも独立した関連を認めた．それらは人種集団間で共通したリスク効果を持っていた．HLA-DRβ1 pos. 74 や HLA-A pos. 62 などの一部のバリアントは一方の人種集団で感受性が報告されていたが，他方の集団でも同様のリスク効果を持つことを示すことが出来た．検出されたリスク関連アミノ酸位置の多くは，HLA 分子の構造においてペプチド結合溝に位置しており，抗原提示能への機能的影響を介して T1D の発症に関わると考えられた．

さらに本研究では，Parkinson 病（Parkinson's disease, PD）についても MHC 領域における trans-ethnic fine-mapping を行った．PD は，パーキンソニズムと表現される運動症状と非運動症状を呈する代表的な神経変性疾患である．PD の発症リスクにおいても MHC 領域との関連が以前より指摘されていたが，報告されるリスク関連 HLA アレルにはばらつきがあり統一的な見解に欠けていた．これは，MHC 領域が占める PD の遺伝的リスクに対して，先行研究ではサンプル数が少なく十分な検出力を得られていないことが一因と考えられた．また，先行研究の多くが欧米人を対象としており他の人種集団におけるエビデンスに乏しかった．本研究では，UK Biobank の SNV ジェノタイプデータに対して DEEP*HLA を適用し fine-mapping を行い，さらに他の複数の GWAS 研究の要約統計量と合わせてメタアナリシスを行い trans-ethnic fine-mapping を行った（計 PD 患者 22,362 例と対照 1,315,997 例）．結果，HLA クラスⅡ遺伝子の HLA-DRβ1 pos. 13（$P = 6.0 \times 10^{-15}$）に最も強い関連を認め，それ以外に HLA-B pos. 69（$P = 1.0 \times 10^{-7}$）にも独立した関連を認めた．HLA 分子が PD の発症にどのように関わるかにおいては先行研究では，HLA 分子が PD の脳組織内蓄積物質である α-シヌクレインの断片をエピトープとして認識し，抗原提示・免疫応答を引き起こす可能性が報告されていた．従って，本研究では *in silico* の HLA-ペプチド結合親和性予測ツールである NetMHCpan II を用いて，*HLA-DRB1* アレルと α-シヌクレイン・エピトープの結合親和性を網羅的に評価した．結果，HLA-DRβ1 pos.13 においてヒスチジン（PD の発症に保護的な効果をもつアレル）を有する *HLA-DRB1* アレルはその他のアレルに比べて弱い結合親和性を持ち（$P = 9.6 \times 10^{-4}$），逆にアルギニン（PD の発症にリスク効果をもつアレル）を有する *HLA-DRB1* アレルは，他のアレルに対して強い結合親和性を呈した（$P = 1.0 \times 10^{-3}$）．このことから，HLA-DRβ1 pos.13 は，α-シヌクレインエピトープへの結合親和性の変化を介して PD の発症に関わっている可能性が考えられた．

本研究では深層学習を応用した HLA imputation 法，DEEP*HLA を開発した．特に頻度の低いアレルの予測精度の改善に成功し，その長所を活かして MHC 領域における trans-ethnic fine-mapping に実用した．Trans-ethnic fine-mapping の利点は，各集団の LD による交絡を調整することにより，普遍的にリスクに関連するバリアントを検出できうることである．結果，PD と T1D において，異なる集団間で共通して発症リスクとなる HLA バリアントのセットを同定し

た．本研究では，いずれの疾患においても 2 つの異なる人種集団を対象として fine-mapping を行ったが，リスク関連バリアントのより頑健な検出のための次の段階としては，より多くの異人種集団を対象とした multi-ethnic fine-mapping が有効であると考えられる．本研究結果が PD と T1D の病態における免疫学的機序のさらなる解明に寄与するとともに，DEEP*HLA が様々な疾患のリスク関連 HLA アレルの同定に役立つことを期待している．