博士論文

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

内　藤　龍　彦

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

所属：脳神経医学専攻　神経内科学

指導教員：戸田　達史

申請者：内藤　龍彦

目　次

要　旨

MHC 領域は，多様な疾患の発症に関わる重要な領域である．HLA imputation 法の開発により MHC 領域での fine-mapping が可能となったが，従来の HLA imputation 法では希少アレルに対する予測精度が低下してしまい，trans-ethnic fine-mapping の信頼性を下げる要因になっていた．本研究では深層学習を応用した新規の HLA imputation 法，DEEP*HLA を開発した．包括的な性能評価の結果，DEEP*HLA は従来法と比較し希少アレルに対する予測精度が改善していた．次に，DEEP*HLA をバイオバンク・ジャパンと UK バイオバンクの GWAS データに適用し，Parkinson 病と 1 型糖尿病の MHC 領域の trans-ethnic fine-mapping を行った．結果，両疾患で 2 つの異なった集団に共通して発症リスクに関わる HLA バリアントを同定した．本研究は，大規模 GWAS データ解析において深層学習が有用であることの一例を示す．

序　文

1. 主要組織適合遺伝子複合体 (major histocompatibility complex，MHC) 領域

1. 1. ゲノムワイド関連解析と MHC 領域

多因子疾患とは，遺伝的因子と環境的因子により発症する疾患であり，ゲノムワイド関連解析 (genome-wide association analysis，GWAS) は多因子疾患の感受性遺伝子領域を同定することにおいて有用な遺伝統計学的解析手法である．GWAS とは，形質と関連するゲノム座位を特定するために，ゲノム全体の領域のマーカー変異について形質との関連を網羅的に検定することである．GWAS により様々な疾患において多数の感受性遺伝子領域が同定されてきており，例えば Parkinson 病の GWAS においては，これまで約 40 箇所の感受性領域が指摘されている [1]．GWAS で検出される遺伝子領域の中でもヒト白血球抗原 (human leukocyte antigen，HLA) 遺伝子をはじめとした免疫機能に関わる多数の遺伝子をコードする MHC 領域は GWAS で最も多く感受性が検出される重要な領域である [2]．MHC 領域の機能に一致して，主に自己免疫疾患や感染症に対する発症リスクとの関連が指摘されてきたが，それ以外にも代謝性疾患や精神神経疾患など様々な多因子疾患の発症リスクとの関連も見出されてきており [3,4]，これらの疾患の病態において自己免疫学的な機序の関与が窺われる．MHC 領域に存在する遺伝子の中でも，HLA 遺伝子は MHC 領域の遺伝的リスクの大部分を説明すると考えられている [5,6]．

1. 2. MHC 領域と HLA 遺伝子

MHC 領域は，遺伝子の 6 番染色体上の位置 6p21.3 に位置し約 5 Mb に及ぶ長大な領域である．MHC 領域には，多数のヒト白血球抗原 (human leukocyte antigen，HLA) 遺伝子が含まれている．HLA 遺伝子によりコードされる HLA 分子は，T 細胞受容体を介して抗原を T 細胞に提示し，免疫反応を惹起する役割を持つ．MHC 領域は，クラス I，II，III 領域の 3 つの領域に大別される [7]．クラス I 領域には，古典的 HLA クラス I (*HLA-A*，*-B*，*-C*) ・非古典的 HLA クラス I (*HLA-E*，*-F*，*-G* など) が存在する．クラス II 領域には，古典的 HLA クラス II (*HLA-DR*，*-DP*，*-DQ*) ・非古典的 HLA クラス II (*HLA-DM*，*-DP*) が存在する．その他の領域がクラス III 領域となる．

HLA クラス I 分子は，有核細胞の表面に発現し，一般的に内在性抗原を CD8 陽性 T 細胞に提示する機能を持つ．HLA クラス I 分子は，多型性が高く明確な抗原提示能を持つ古典的 HLA クラス I 分子と，多型性が低く様々な機能を持つ非古典的 HLA クラス I 分子に分類される．HLA クラス I 分子の構造は，α1，α2，α3 の 3 つのドメインからなる重鎖と，1 つの免疫グロブリン様ドメインを構成する β2 ミクログロブリンからなり，α1，α2 からなるペプチド収容溝に抗原ペプチドが結合する．

HLA クラス II 分子は，マクロファージや樹状細胞などの抗原提示細胞の表面に発現し，外来性抗原を CD4 陽性 T 細胞に提示する機能を持つ．HLA クラス II 分子の構造は，α1，α2 の 2 つのドメインからなる α 鎖と，β1，β2 の 2 つのドメインからなる β 鎖からなり，α1，β1 からなるペプチド収容溝に抗原ペプチドが結合する．HLA クラス I の各分子が，1 つの HLA 遺伝子によってコードされるのに対し，HLA クラ

ス II の場合は，2 つの遺伝子の産物からヘテロ二量体が形成される．つまり，例えば，*HLA-DQA1* 遺伝子と *HLA-DQB1* 遺伝子が，それぞれ DQ 分子の α 鎖と β 鎖をコードする．

1. 3. HLA アレルの命名法

HLA アレルの命名は，HLA-A*01:01:01 などのように，HLA 遺伝子座名の後にアスタリスクを挟み，セミコロンで区切られた数字によって表現される．セミコロンで挟まれた数字の個数による 2，4，6，8 桁レベルの階層的な命名により，異なった解像度での配列情報が表現される．2 桁レベルは血清型，4 桁レベルはコーディング領域のアミノ酸配列の差異，6 桁レベルはコーディング領域の非同義置換も含めた塩基配列の差異，8 桁レベルは非コーディング領域も含めた塩基配列の差異を表す．HLA 分子においてアミノ酸配列が機能的に重要であるため，アミノ酸アレルもしくは，その組み合わせである 4 桁レベルの HLA アレルが疾患の発症リスクと最も関連することが多い．

2. MHC 領域と疾患リスク

前述のように，MHC 領域は様々な疾患の発症リスクと関わるが，ここでは本研究で解析対象となる 1 型糖尿病 (type 1 diabetes，T1D) と Parkinson 病 (Parkinson's disease，PD) について述べる．

2. 1. T1D と HLA

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疾患である [8]．T1D は，多因子疾患であるが，遺伝的要因が強く，特に MHC 領域のみで表現型分散の約 30%を占め，特に *HLA-DRB1*, *-DQA1*，*-DQB1* の領域に強いリスクが報告されている [5]．特定の HLA アレルが発症リスクに繋がる機序としては，自己抗原であるインスリンや抗グルタミン酸脱炭酸酵素 (glutamic acid decarboxylase，GAD) への HLA 分子の結合親和性の変化の他，遺伝子発現量の変化 [9]，DM 分子のペプチド編集機能への影響 [10]，DQ 分子の不安定性 [11] など様々な機序が報告されており，複数の要因が複雑に絡み合って病態に寄与していると考えられる．

欧米人集団では HLA-DQβ1 non-Asp57 と発症リスクとの間に強い相関があることが報告されている [7]．Hu らの欧米人集団の大規模コホートを対象とした研究でも，HLA-DRβ1 pos.13，71，HLA-DQβ1 pos.57 の 3 つのアミノ酸位置が *HLA-DRB1*，*-DQA1*，*-DQB1* 領域のリスクの大部分を説明し，その中でも特に HLA-DQβ1 pos.57 が最も強い関連を認めたと報告されている [7]．一方，中国人集団では，他の HLA ア

レルにより強い関連が報告されている．また，日本人集団ではむしろ，HLA-DQβ1 Asp57 はリスクと相関しており，欧米人集団とは逆の効果を示している[13].

2.2. PD と HLA

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の一つである[14]．有病率は 10 万人に 100-300 人であり，60 歳代に発症のピークがある．高齢になるほど罹患率が増す．約 5-10%は単一遺伝性であるが，残りの孤発性 PD にもしばしば遺伝的素因が見られる．病理学的には，黒質線条体のドパミン作動性神経の脱落と α-シヌクレインを含む凝集体 (Lewy 小体) の蓄積を特徴とする．様々な臨床的報告や基礎研究から，神経炎症も PD の病態に関与すると考えられている[15]．特に Sulzer らの報告では，PD 患者末梢血単核細胞を α-シヌクレイン由来の特定のエピトープで刺激した際に免疫細胞が活性化し，特にその反応性と HLA 遺伝子型には相関が見られており，PD の病態における HLA 分子の重要性が窺われる[16]．MHC 領域の遺伝的リスクとしては，HLA-DRB1*04 や[3]，HLA-C*03:04，HLA-DRB1*04:04[17] などに発症リスクとの関連が報告されているが，いずれもサンプルサイズが数千人にとどまっている．MHC 領域の PD の発症リスクにおける効果量に比してサンプルサイズが小さく，大規模集団を用いた統一的な見解に欠けていると考えられる．また，大半の報告が欧米人を対象とした研究であり，非欧米人の報告はさらに少ない[18,19].

3. HLA imputation 法と MHC 領域における fine-mapping

3. 1. HLA imputation 法

GWAS で同定できるのは，形質に関連がある大まかな遺伝子領域までであり，原因遺伝子変異の同定 (fine-mapping) には追加の解析が必要である．GWAS で MHC 領域に感受性が検出された場合には，HLA 遺伝子配列を網羅的に決定し，その中で最も発症に関連する HLA アレルを求めるのが一般的な方法である (図 1)．HLA 遺伝子型のタイピング法には，配列特異的オリゴヌクレオチド (sequence-specific oligonucleotide，SSO) ハイブリダイゼーション，サンガーシークエンシング，次世代シークエンシング (next-generation sequencing，NGS) など様々な方法があるが，手間や費用の観点から，GWAS で対象となる大規模なコホートへの適用は容易ではない[4,20]．従って，一塩基変異 (single nucleotide variant，SNV) と HLA 遺伝子型がタイピングされた個人ジェノタイプデータからなる参照パネルを予め構築し，それを用いて HLA 遺伝子型が未観測のサンプルにおいて SNV レベルの情報から HLA 遺伝子型を間接的に推定して解析するのが一般的である[4,21–23]．このように，SNV と HLA 遺伝子型の連鎖不平衡 (linkage disequilibrium，LD) の関係性を用いて，SNV の情報から統計学的に HLA 遺伝子型を推定することを HLA imputation 法という (図 1)．MHC 領域の LD 構造やアレル頻度は集団特異的であるため，HLA imputation を行う際は，参照パネルとターゲットの集団を揃えて行うのが一般的である．HLA imputation 法の開発により，それ以前は困難であった MHC 領域の fine-mapping が可能となり，多様な疾患のリスク HLA アレル解明に寄与した[22,23]．

図 1: MHC 領域の fine-mapping の概要

GWAS で MHC 領域が感受性領域として検出された場合，HLA 遺伝子配列型を網羅的に決定した上で，リスクと最も関連するアレルを探索する．

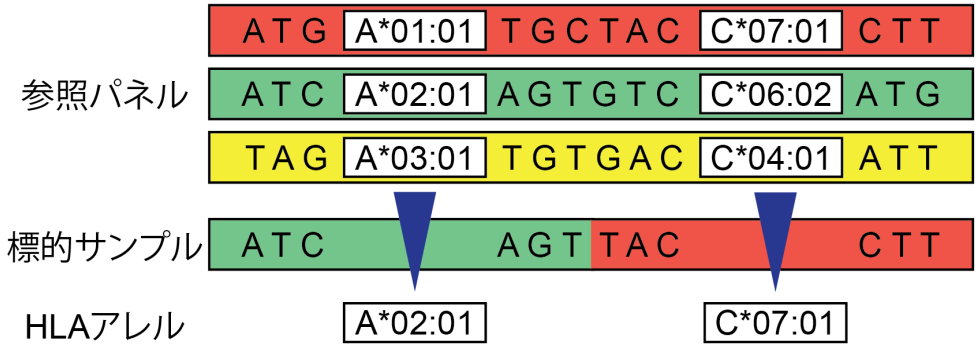

図 2: HLA imputation 法の概要

SNV のジェノタイプと HLA アレルの両方の情報が収載された HLA imputation 用の参照パネルを予め構築することにより，SNV ジェノタイプと HLA アレルの相関関係を用いて SNV ジェノタイプの情報のみ標的サンプルの HLA アレルを推測することができる．

HLA imputation 法は，タグ SNV を用いた簡易的な推論から始まり [24,25]，その後様々な手法が開発された．Leslie らは，Li & Stephens ハプロタイプモデルに基づいた確率的アプローチによる HLA imputation 法を初めて報告した [26]．Li & Stephens ハプロタイプモデルは，個人の遺伝子配列はその個人が属する集団の他の個人の遺伝

子配列の組換えと少数の突然変異で表現されるとする遺伝統計学モデルである [20]. HLA*IMP は，Leslie らの手法を元に，ヨーロッパの集団の HLA タイピングデータと SNV データを用いてソフトウェアとして確立したものである [21]．その後開発されたソフトウェアプログラム HLA*IMP:02 は，集団間の不均一性に対応するため，複数の集団からの SNV データを使用してハプロタイプグラフを元に imputation を行う [29]．現在使用可能な HLA*IMP シリーズは，各研究者が参照パネルを用いて独自にモデルを構築する機能は，実装されていない.

本研究では，他の標準的なソフトウェアである，SNP2HLA と HLA Genotype Imputation with Attribute Bagging (HIBAG) を後の解析で用いる．SNP2HLA は，各 HLA アレルやアミノ酸アレルをバイナリアレルと見なすことで，ハプロタイプグラフを用いた SNV genotype imputation ソフトウェアである Beagle を内部的に適用して，HLA アレルとアミノ酸アレルの imputation を同時に行う [30]．HIBAG もよく用いられるソフトウェアの一つであり，EM アルゴリズムに基づいたハプロタイプの推定を行う分類器を複数構築してアンサンブル学習を行うことにより imputation を行う [31]. SNP2HLA では，imputation を行うたびに参照データを必要とするが，HIBAG は一旦分類器を構築すれば参照データを使用せずに imputation を行うことができるというメリットがある.

3.2. 既存の HLA imputation 法と MHC 領域の fine-mapping の問題点

MHC 領域には，集団に特異的な長大なハプロタイプや複雑な LD 構造が存在するため，HLA アレルの分布や頻度は集団間で大きく異なる [6,32]．その結果，fine-mapping 研究を進めていく中で，疾患と関連する HLA アレルが集団間でばらつく，ということがしばしばみられる．代表的な例として前述の T1D においては，欧米人集団では発症リスクと最も関連があるアレルが，日本人集団では逆の相関を示している [13]．疾患の発症に中心的に作用するアレルであれば，集団に依らず普遍的に発症リスクに関連する可能性が高いと考えるのが自然である [33]．従って，集団間で統合的に fine-mapping (trans-ethnic fine-mapping) を行い集団間でリスクが共有されたバリアントを解明することは重要であると考えられる [34]．Trans-ethnic fine-mapping を行う一つの方法は，異なる集団の MHC 領域の複雑さを捉えうるような大規模な HLA imputation 用 multi-ethnic 参照パネルを構築し，それを用いて imputation を行うことである [35]．もう一つの方法は，各集団用の参照パネルで imputation された結果のデータを統合して解析を行うことである．後者の方法は簡便なようであるが，集団間でアレル頻度が大きく異なる MHC 領域のバリアントを頑健に評価するためには，頻度の低いアレルに対しても十分に精度の高い HLA imputation 法が必要である．これは，集団間で頻度が大きく異なる場合，片方の集団で頻度が低く imputation 精度が悪くても，他方の集団では頻度が高い場合があり解析対象から除きづらいためである (一般の関連解析では，頻度が低いアレルは効果量が小さいため解析対象から除いても大きな問題にならない) ．一方，前述のような従来の HLA imputation 法

は，全体としての imputation 精度はいずれも 90%以上程度はあるが [36]，後に示すように，集団中で頻度の低いアレルに対しては精度が著しく低下する傾向がある．MHC 領域特有の複雑な LD 構造を読み取って imputation を行うには，単純な確率論的推論にとどまらない，より高度なパターン認識アルゴリズムが必要である可能性が考えられた．

4. 深層学習

4.1. ニューラルネットワークと深層学習

ニューラルネットワークとは，脳の神経回路を模した数理モデルである．多量の訓練データを学習することで分類・予測や特徴抽出を行うコンピューターアルゴリズムを機械学習法と呼ぶが，深層学習は，入力・複数の中間・出力層からなる階層的なネットワーク構造 (多層ニューラルネットワーク) により，入力データから複雑な特徴を直接学習して予測を行う機械学習法の一種である．深層学習は，特に画像の分類問題において既存手法に比べて極めて高い精度を誇ったことを皮切りに注目されるようになり，様々な分野で応用研究が進んでいる．主に画像認識で用いられる畳み込みニューラルネットワーク (convolutional neural networks，CNN) や，音声認識や文字列処理で用いられる回帰型ニューラルネットワーク (recurrent neural network, RNN) が有名であるが，本研究では，CNN を用いるため，以下では CNN について述べる．

4.2. 畳み込みニューラルネットワーク

CNN は，局所受容野を持つ畳み込み層と，位置普遍性の役割を持つプーリング層という 2 つの層を交互に重ねた構造を持つ順伝播型のネットワークである [37]．この 2 つの層により，画像などの入力データから有益な特徴量を抽出し，その後全結合層により特徴量に重み付けを行い，最後に予測値を出力する．目的となるタスクがデータの所属するカテゴリを予測する問題 (クラス分類問題) であった場合，クラスの数と同数のノードを出力層に配置しその各ノードの出力値が各クラスに分類される確率として予測を行う．クラス分類問題において各クラスに所属する予測確率は 0〜1 の範囲を取りその総和は 1 である必要があるが，それは活性化関数であるソフトマックス関数を介することでなされる．

各ノードに存在する重みと呼ばれるパラメータを，訓練データを用いて最適化を行うことが学習である．訓練データの入力データをモデルに投入した際の出力と正解との差を誤差として計算し，それを元に逆誤差伝播法により重み更新を行う．クラス分類問題においては，誤差はクロスエントロピーを用いることが多い．モデルに訓練データを繰り返し投入して重み更新を行い，十分な予測能が得られたところで学習完了となる．

4.3. 遺伝学分野における深層学習の応用

遺伝学分野における深層学習の応用研究の代表的なものとしては，塩基配列パタンの学習による遺伝子変異の機能的影響の予測や [38–41]，bulk/single-cell RNA-Seq な

どの高次元データの非線形次元削減など [42,43] が挙げられ，いずれも既存の統計・機械学習手法を上回る精度を達成している．一方で，大規模 SNV ジェノタイプデータに対する応用研究の成功例はまだ限られている．SNV ジェノタイプの imputation 法の先行研究としては，雑音除去自己符号化器を用いた手法 [44]，RNN を用いた手法 [45] が挙げられるが，いずれも精度においては既存手法を上回ってはいない．

5. 本研究の目的

上述のように，単一の集団における fine-mapping では，発症に普遍的に関連する HLA バリアントを同定するのが困難であった．一方，信頼性の高い trans-ethnic fine-mapping の結果を得るためには，できるだけ精度の高い HLA imputation 法が必須である． 本研究では，imputation 精度の改善を図るため，深層学習を応用した新規の HLA imputation 法 (DEEP*HLA と命名) を開発し，包括的な性能評価を行った．さらに，DEEP*HLA をバイオバンクの GWAS ジェノタイプデータに適用し，PD，T1D の 2 つの疾患について，MHC 領域における trans-ethnic fine-mapping を行い，各々において発症リスクに関連する HLA バリアントの解明を図った．

方　法

1. 概要

DEEP*HLA は，HLA imputation 用の参照パネルを学習することで，SNV ジェノタイプデータを入力として HLA 遺伝子の遺伝子型の推定値を出力する CNN である．特に，DEEP*HLA は，マルチタスク学習により予め定めたグループ内に属する複数の HLA 遺伝子のアレルを同時に学習・出力することができる．頑健なベンチマークのために，日本人集団用，欧米人集団用の 2 つの異なる HLA imputation 用参照パネルを用いて交差検証法により精度評価を行った．また日本人集団用参照パネルについては，他の日本人データセットに適用した場合の精度評価も行った．さらに，第Ⅲ相 1000 ゲノムプロジェクト (1000 Genomes Project Phase III，1KGv3) のデータを用いて，多民族集団を対象とした HLA imputation の精度を検証した．

本研究の後半では，バイオバンク・ジャパン (BioBank Japan，BBJ) の日本人コホートと UK バイオバンク (UK Biobank，UKB) の英国人コホートについて，各集団の参照パネルを学習した DEEP*HLA モデルを用いて，HLA imputation を行い，T1D, PD の両疾患について，それぞれ MHC 領域の trans-ethnic fine-mapping を行った．T1D においては，BBJ と UKB のコホートの HLA imputation 結果を用いて，trans-ethnic fine-mapping を行った．PD においては，UKB コホートの HLA imputation 結果と，複数の研究の GWAS サマリ統計量を統合してメタアナリシスを行うことで，trans-ethnic fine-mapping を行った．

なお，本研究は，大阪大学研究倫理審査委員会の承認を得て実施された (承認番号：734-13) .

2. HLA imputation 用参照パネル

HLA imputation 法の精度評価，バイオバンクへの HLA imputation の適用に際して，日本人集団，欧米人集団の 2 つの HLA imputation 用参照パネルを使用した.

2. 1. 日本人集団の参照パネル

非血縁関係の日本人 1120 人からなり，NGS によって決定した 33 種の HLA 遺伝子 (9 種類が古典的 HLA 遺伝子 (class I: *HLA-A*，*HLA-B*，*HLA-C*，class II: *HLA-DRA*，*HLA-DRB1*，*HLA-DQA1*，*HLA-DQB1*，*HLA-DPA1*，*HLA-DPB1*)，24 種類が非古典的 HLA 遺伝子 (*HLA-E*，*HLA-F*，*HLA-G*，*HLA-H*，*HLA-J*，*HLA-K*，*HLA-L*，*HLA-V*，*HLA-DRB2*，*HLA-DRB3*，*HLA-DRB4*，*HLA-DRB5*，*HLA-DRB6*，*HLA-DRB7*，*HLA-DRB8*，*HLA-DRB9*，*HLA-DOA*，*HLA-DOB*，*HLA-DMA*，*HLA-DMB*，*MICA*，*MICB*，*TAP1*，*TAP2*)) の最大 6 桁レベルの HLA タイピングデータと，HumanCoreExome BeadChip (v1.1; Illumina) を用いてジェノタイピングされた MHC 領域の高密度 SNV データを含んでいる [4]．フェージングの段階で，HLA アレルのストランドが異なる解像度間で一意に定まらない 2 人のデータは除外した.

日本人集団用の参照パネルを用いた HLA imputation の精度評価にあたって，交差検証のみでなく，日本人集団の独立サンプルに適用したときの精度も評価した．非

血縁関係の日本人 908 人からなり，SSO 法によって決定した 7 種の古典的 HLA 遺伝子 (*HLA-A*，*HLA-B*，*HLA-C*，*HLA-DRB1*，*HLA-DQA1*，*HLA-DQB1*，*HLA-DPA1*) の 4 桁レベルの HLA タイピングデータと，4 つの SNV ジェノタイピングアレイ (Illumina HumanOmniExpress BeadChip，Illumina HumanExome BeadChip，Illumina Immunochip，Illumina HumanHap550v3 Genotyping BeadChip) を用いてジェノタイピングした高密度 SNV データから構成される．

2. 2. 欧米人集団用の参照パネル

欧米人種用参照パネルとして，The Type 1 Diabetes Genetics Consortium (T1DGC) 参照パネルを使用した．T1DGC パネルは，非血縁関係の欧米人集団 5225 人からなり，SSO 法によって決定した 8 種の古典的 HLA 遺伝子 (*HLA-A*，*HLA-B*，*HLA-C*，*HLA-DRB1*，*HLA-DQA1*，*HLA-DQB1*，*HLA-DPA1*，*HLA-DPB1*) の 4 桁レベルの HLA タイピングデータと，Illumina Immunochip を用いてジェノタイピングされた 5868 個の SNV から構成される [30]．フェージングの段階で，HLA アレルのストランドが異なる解像度間で一意に定まらない 103 人のデータは除外した．

2. 3. 混合集団パネルと 1000 Genomes Project データ

日本人集団参照パネルと欧米人集団参照パネルを統合した混合集団パネルを試験的に作成し，1000 Genomes Project Phase III (1KGv3) の多様な集団における imputation 精度を評価した．参照パネルを統合する際は，集団間で SNV のアレル頻度が異な

ることを考慮して，パリンドロミック SNV (アレルが，A-T もしくは C-G の組み合わせになっているもの) は除外した.

1KGv3 のコホートは，5 つの異なる集団 (AFR，AMR，EAS，EUR，SAS) からなる 2554 人で構成されている．各 HLA 遺伝子 (HLA クラス I 遺伝子: *HLA-A*，*HLA-B*，*HLA-C*，HLA クラス II 遺伝子: *HLA-DRB1*，*HLA-DQB1*) について，NGS に基づく手法で求められた 4 桁アレルのタイピングデータを正解データとして使用した [46].

3. GWAS データ

3. 1. BBJ コホート

BBJ (https://biobankjp.org/english/index.html) は，2003 年から 2007 年までに登録された日本人約 20 万人からなる多施設病院ベースのレジストリであり，血清，臨床情報など多彩な情報を含んでいる [47,48]．BBJ プロジェクトに登録された T1D 診断歴のある 831 例の GWAS データと，自己免疫疾患の診断歴のない対照群 61,556 例の GWAS データを用いた [4].

3. 2. UKB コホート

UKB (https://www.ukbiobank.ac.uk/) は，2006 年から 2010 年までに英国で登録された 40～69 歳の約 50 万人の健康関連情報から構成されている [49]．T1D 患者は，病院記録でインスリン依存性糖尿病と診断された個人のうち，病院記録でインスリン非依存性糖尿病，自己申告診断で 2 型糖尿病と診断された個人を除外した．PD 患者

は，病院記録で PD と診断された個人から選んだ．対照群は，病院記録にも自己申告診断にも自己免疫疾患の記録がない個人を選定した．PD の対照群は，そこからさらに自己申告診断で PD の記録がある個人を除外した．性染色体異数性，自己申告の性と遺伝的性の不一致，高品質マーカーのコール率における外れ値に該当する個人は除外した．

3. 3. PD の GWAS サマリ統計量

PD においては，GWAS サマリ統計量を使用した解析も行った．欧米人集団のデータとして，23andMe が保有する GWAS サマリ統計量を使用した．これには，2014 年の Nalls らの研究 (ジェノタイピングプラットフォーム毎に，PD 患者 866 例と対照 32,538 例，PD 患者 3,261 例と対照 29,499 例) [50]，Chang らの研究 (PD 患者 6,476 例と対照 302,042 例) [51]，およびその後の研究 (PD 患者 2,448 例と対照 571,411 例) [1] が含まれている．東アジア人集団のデータとして，日本人集団の PD GWAS 要約統計量 (PD 患者 988 例と対照 2521 例) [52] と東アジア人集団の GWAS メタアナリシスの要約統計量 (PD 患者 6724 例と対照 24,851 例) [53] を使用した．

4. DEEP*HLA の構造

DEEP*HLA は，imputation を行う対象の HLA 遺伝子をいくつかのグループに分け，同じグループ内の HLA 遺伝子の HLA ジェノタイプについて同時に imputation を行うマルチタスク学習を利用している．グループ内の共有部分は，2 つの畳み込

み層と全結合層からなり，グループ内の各 HLA 遺伝子のジェノタイプの推定値を出力するための全結合層からなる (図 3)．グループ分けは，LD 構造と物理的距離に基づいて行った: (1) {*HLA-F*，*HLA-V*，*HLA-G*，*HLA-H*，*HLA-K*，*HLA-A*，*HLA-J*，*HLA-L*，*HLA-E*}，(2) {*HLA-C*，*HLA-B*，*MICA*，*MICB*}，(3) {*HLA-DRA*，*HLA-DRB9*，*HLA-DRB5*，*HLA-DRB4*，*HLA-DRB3*．*HLA-DRB8*，*HLA-DRB7*，*HLA-DRB6*，*HLA-DRB2*，*HLA-DRB1*，*HLA-DQA1*，*HLA-DOB*，*HLA-DQB1*}，(4) {*TAP2*，*TAP1*，*HLA-DMB*，*HLA-DMA*，*HLA-DOA*，*HLA-DPA1*，*HLA-DPB1*}．

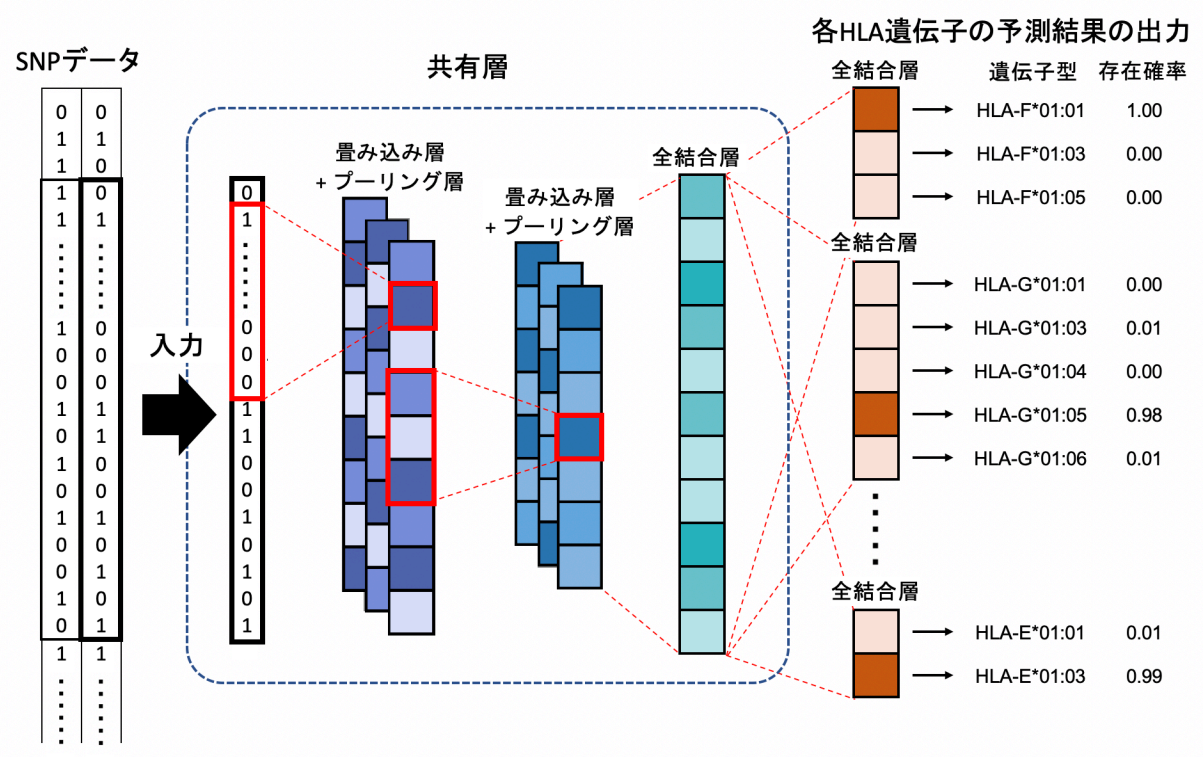

図 3: DEEP*HLA のアーキテクチャ

DEEP*HLA は，フェージングされた SNV ジェノタイプデータを入力として受け取り，HLA 遺伝子の各アレルの予測確率 (dosage) を出力する深層学習モデルである．

DEEP*HLA は，ハプロタイプ毎の SNV のジェノタイプを入力として受け取り，各 HLA 遺伝子についての個々のアレルのジェノタイプの存在確率 (dosage) を出力す

る．各グループについての入力ウィンドウ内の各 SNV は，reference allele または alternative allele と一致するかどうかに基づいて，0/1 にエンコードされる．入力ウィンドウのサイズは，グループ内の遺伝子の範囲の両側 ±500 Kb の領域に固定した．これは，参照パネルと対象集団とに共通して存在する SNV のタイピング密度に依るが，約 500~800 個程度の SNV に相当する．最大プーリング層と全結合層を有する２つの畳み込み層が，共有部分として入力層に続く．共有部分の最後にある全結合層は，各 HLA 遺伝子のアレル数と対応するノード数を有する個々の全結合層が続く．最終出力の前にソフトマックス関数を適用することで，1 つのハプロタイプについてその総和が 1 になるように，各アレルに対して 0.0～1.0 の範囲の dosage を出力するようにした．出力層を除くすべての層の出力には，活性化関数として Rectified Linear Unit 関数を用いた．畳み込み層と全結合層にはドロップアウト (dropout rate: 0.5) を適用し [54]，畳み込み層にはバッチ正規化を追加した [55].

訓練を終了する学習エポック数を決定するために，データセットの 5%をサブバリデーションデータとして残し，サブバリデーションデータにおいて 8 回連続で精度の改善がなかった場合に訓練終了とした．10 分割交差検証法を行うときは，公正な評価を行うために，サブバリデーションは訓練データから選定した [56]．グループ内の各 HLA 遺伝子の損失関数の更新にあたっては，全 HLA 遺伝子についてパレート最適解を見つけるため，multiple-gradient descent algorithm – upper bound を使用して，損失関数の重み付けを行い，それに対して最適化を行った [57]．2 桁，4 桁，6 桁アレルの順番に学習を行っていったが，階層性を利用するために下位の桁のアレル

の学習を行うときは，上位の桁のアレルの学習済みモデルのパラメータを初期値として開始した (転移学習) [58]．参照パネルの HLA アレルの中には，4 桁または 6 桁のアレルが決定されないものもあったが，他の HLA 遺伝子と同等の階層レベルを維持するために，それらの上位アレルを代わりに設定した．フィルタの数や畳み込み層のカーネルサイズ，全結合層のサイズなどのハイパーパラメータは，交差検証法の前に無作為に選んだ 90%のデータに対して Optuna を用いて最適化した [59]．結果，1 層目畳み込み層はフィルター数 128，フィルターサイズ 64，2 層目畳み込み層はフィルター数 64，フィルターサイズ 64，全結合層のサイズは 256 とした．深層学習アーキテクチャの構築にあたっては，Python のニューラルネットワークライブラリである Pytorch 1.4.1 (http://pytorch.org/) を用いて実装した．

5. HLA imputation ソフトウェアの設定

SNP2HLA (v1.0.3; http://software.broadinstitute.org/mpg/SNP2HLA/) は，始めに独自のアルゴリズムに基づいて参照パネルと目的のサンプルの間でストランドを揃えるが，交差検証法においては，参照データと検証先のデータの間でストランドは揃っているはずであるため，このステップは除外した． SNP2HLA の他の設定はデフォルトのままとした． HIBAG (1.22.0.) は，日本人集団参照パネルを用いるときは，分類器の数は，良好な精度を達成するのに十分とされている 25 個に設定した [60]．欧米人集団参照パネルにおいては，学習時間が非常に長いため，交差検証法の最初の検証セットで精度が分類器を 25 個に設定した場合と 2 個に設定した場合で精度がほと

んど変わらないことを確認して 2 個に設定して評価した．入力 SNV の両側の隣接領域は 500 kb に設定した．

6. 精度評価の方法と指標

様々な側面から imputation 精度を評価するために，感度，陽性適中率，r^2，最良推測遺伝子型の一致率を下記の通り定義して用いた．

SNP2HLA の原著論文では，遺伝子座ごとの精度は，各アレルの dosage の合計を真の総数で割ったものとして定義された [61]．この定義は，クロス集計表における感度と一致しているため，後で定義する陽性適中率と対比するために，ここでは感度 (sensitivity，Se) と定義した．

$$Se(L) = \frac{\sum_{i=1}^{n}\left(D_i\left(A1_{i,\ L}\right) + D_i\left(A2_{i,\ L}\right)\right)}{2n}$$

ここで，n はサンプル数を表し，D_i は個体 i のアレルの dosage を表し，アレル $A1_{i,L}$ および $A2_{i,L}$ は遺伝子座 L での個体 i の真の HLA アレルを表す．Imputation されたアレルの dosage は，真のアレル $A1_{i,L}$ および $A2_{i,L}$ との整合性を最適化するように配置された場合を仮定している．個々の HLA アレルの imputation 精度を評価するために，遺伝子座毎の感度 $Se\,(L)$ を分解して，各アレルの感度 $Se\,(A)$ として定義した．

$$Se(A) = \frac{\sum_{j=1}^{m} D_j(A)}{m}$$

上記の感度では，偽陽性の影響を評価できないため，陽性適中率 (positive predictive value，PPV) もクロス集計表に基づいて，次のように定義した．

$$PPV(A) = \frac{\sum_{j=1}^{m} D_j(A)}{\sum_{j=1}^{m} D_j(A) + \sum_{k=1}^{2n-m} D_k(A)}$$

ここで，m は，サンプル全体におけるアレル A の真の本数を示し，D_i は，アレル A を持つ個々のハプロタイプ j におけるアレル A の dosage を表す．D_k は，A 以外のアレルを持つハプロタイプ k における A の dosage を表す．

また，他によく用いられる imputation 精度評価指標として，各アレルの dosage と真の本数について Pearson の積率相関係数 r^2 を計算した [30]．

さらに，最良推測遺伝子型 (best-guess genotype) の精度を評価するために，各アレルの最良推測遺伝子型と真の遺伝子型の一致率 (concordance rate，CR) を次のように計算した．

$$CR(L) = \frac{\sum_{i=1}^{n}\left(B_i\left(A1_{i,\ L}\right) + B_i\left(A2_{i,\ L}\right)\right)}{2n}$$

ここで，B_i は，個人 i のアレルの最良推測遺伝子型を表す．定義上，dosage が最良推測遺伝子型に置換されている点以外は，感度と同じである．従って，感度と同様に分解して，各アレルの最良推測遺伝子型の一致率 $CR\,(A)$ も評価した．最良推測遺伝子型の陽性適中率は冗長であるため評価しなかった．

各遺伝子座または特定の範囲のアレル頻度における精度値は，個々のアレルの精度をアレル頻度に基づいて加重平均を計算して求めた． r^2 についてはバイアスを減らすため，Fisher の Z スコア変換を個々の値に適用してから加重平均を求め，その後にそれらを逆変換した [62]．

7. 計算負荷の測定

日本の参照パネルを使用して，BBJ のデータの一部 (*n* = 1,000，2,000，5,000，10,000，20,000，50,000，および 100,000 サンプル) を imputation した場合の計算負荷を測定した．各手法の計算不可の測定には，48 個の Central Processing Unit (CPU) コア (Intel®Xeon®E5-2687Wv4@ 3.00 GHz) と 256 GB の Random Access Memory (RAM) を搭載した CentOS 7.2.1511 サーバを用いた．さらに，20 個の CPU コア (Intel®Core™i9-9900X @ 3.50 GHz)，GPU NVIDIA®GeForce®RTX2080Ti，128 GB の RAM を搭載した Ubuntu 16.04.6 LTS を使用して，GPU を使用した場合の DEEP*HLA の学習時間も測定した． DEEP*HLA のモデルを使用して imputation を行うには，入力ジェノタイプデータをハプロタイプフェージングし，さらにモデルの学習をする必要がある．また DEEP*HLA の実行時間測定にあたっては，imputation のプロセスのみでなく，GWAS データのフェージング時間 (Eagle ソフトウェアで実施)，参照パネルの学習時間も合わせて総実行時間とした．同様に，HIBAG もモデルの学習を必要とするため，モデル学習時間，imputation 実行時間を合わせて総実行時間とした．SNP2HLA は使用可能メモリの最大値を 100 GB に設定して実行した．処理時間と最大メモリ使用量は，コマンドラインインターフェイスから実行した場合について，GNU Time ソフトウェアを使用して測定した．

8. 学習済み DEEP*HLA モデルの感度マップの生成

学習済み DEEP*HLA モデルに SmoothGrad[63] を適用することで，DEEP*HLA の予測において重要な SNV を推定した．SmoothGrad は，深層学習モデルの感度マップを生成する方法の一種であり，入力データにノイズを加えた場合に，出力にどれくらい変化が生じるかによって予測の判断根拠となる入力領域を，感度として評価するというものである．ガウジアンノイズを加えた入力を複数作成しサンプリングの上，平均を取ることで頑健な感度マップを作成する．本研究では，入力 SNV に対してガウジアンノイズを追加したサンプルを 200 個生成し，学習済み DEEP*HLA モデルに入力し，SNV の各位置の感度値を，ノイズ入りサンプルから得られた出力値と実際の正解ラベルとの差分によって生じる勾配の絶対値を平均することによって得た．ある HLA アレルの感度マップを作成するとき，ターゲット HLA アレルを持つ全てのハプロタイプに対して感度マップを生成し，それらの平均を求めた．

9. Monte Carlo ドロップアウト法による imputation の不確かさの推定

DEEP*HLA の予測の不確かさを推定するために，Monte Carlo ドロップアウト法を用いた [64]．ドロップアウトとは，深層学習モデルの学習時に各層の重みを一定確率でランダムに 0 として順伝播させることによって過学習を防ぐ手法である [54]．上述のように，DEEP*HLA モデルでは畳み込み層，全結合層の各層に適用している．通常予測時には，ドロップアウトをオフとするが，Monte Carlo ドロップアウト法では，予測を行う際もドロップアウトをオンとしたまま，複数回サンプリングを行う．

異なるサンプリングでは異なるユニットがドロップアウトされるため，モデルのパラメータをベルヌーイ分布の確率変数に基づいたベイズサンプリングとみなすことができる．以下のように，サンプリングの変動を，エントロピーを用いて定量化することで，最良推定遺伝子型の予測の不確かさを求めた．

$$H = -\left(\frac{t}{T}\log\frac{t}{T} + \frac{T-t}{T}\log\frac{T-t}{T}\right)$$

ここで，T は変分サンプリングの回数，t は出力された遺伝子型が最良推定遺伝子型と一致した回数である．今回の実装では $T = 200$ とした．

10. GWAS ジェノタイプデータに対する trans-ethnic fine-mapping

本研究では，HLA バリアントを「MHC 領域の SNV，2 桁および 4 桁の HLA アレル，それぞれの残基に対応する HLA アミノ酸アレル」と定義し関連解析の対象とした．BBJ，UKB の GWAS ジェノタイプデータに，それぞれ日本人，欧米人集団参照パネルを用いて学習した DEEP*HLA モデルを適用し，2 桁・4 桁の HLA アレルの dosage を求めた．アミノ酸アレルの dosage は，imputation された 4 桁 HLA アレルの dosage に対して，4 桁 HLA アレルとアミノ酸配列の対応表を用いて行列計算により算出した．10 分割交差検証法での r^2 が 0.7 以下であったアレルは，解析対象から除外した．MHC 領域の SNV は，Eagle (version 2.3) でプレフェージングした後，minimac3 (version 2.0.1) を用いて imputation した．それらについて，Minor allele frequency (MAF) ≧ 0.5%，r^2 ≧ 0.7 を満たさないものは除外した．

Trans-ethnic fine-mapping を行うにあたり，それぞれのコホートの imputation 結果を統合した．SNV については，集団間の MAF の相違を考慮して，ストランドを正確に揃えるために，すべてのパリンドロミック SNV を除去した．HLA アレル，アミノ酸アレルについては，片方の集団でしかみられないものは他方の集団では存在しないものとしてコードした．

疾患の感受性の対数オッズ比に対して，アレルの dosage による相加的効果を仮定して，ロジスティック回帰モデルを用いて HLA バリアントと疾患リスクとの関連を評価した．共変量として，個人の年齢と性別を加えた．さらに潜在的な集団的構造化を補正するために，各コホートの GWAS ジェノタイプデータ (MHC 領域は含まない) の第 1〜10 主成分も回帰モデルの共変量に含めた．また，UKB については，ascertainment centre と genotyping chip も共変量に含めた．trans-ethnic 解析にあたり，共変量として集団を示すカテゴリカル変数を追加した．また，他方の集団の主成分項は 0 とした．

HLA バリアントと遺伝子間の独立したリスクを評価するために，関連するバリアントの遺伝子型を共変量として追加したフォワードステップワイズ条件付け回帰分析を行った．HLA 遺伝子全体で条件付けする場合，4 桁アレル全てを共変量として含めた [4,34]．ただし，先行研究で，T1D のリスクは，*HLA-DRB1*，*-DQA1*，*-DQB1* の領域のバリアントの組み合わせと強く関連していることが報告されおり [5]，T1D についてこの領域を解析する際には，HLA 遺伝子ごとではなく，個々の HLA バリアントごとに条件付けを行った．特定の HLA アミノ酸位置で条件付けする場合は，

その位置の全てのマルチアアレルを共変量に含めた．ゲノムワイド有意水準 ($P = 5.0 \times 10^{-8}$) に基づいて，疾患のリスクと関連する HLA バリアントを評価した．

上記のステップワイズ回帰分析によって同定されたリスク関連 MHC 領域バリアントについて，多変量回帰モデルを用いてリスクを評価した．アミノ酸多型をモデルに含める際には，各アミノ酸位置の最も頻度の高いアレルを参照アレルとして除外した．各集団において，リスク関連 MHC 領域変異によって説明される表現型の分散は，各集団の有病率と多変量回帰モデルから得られた効果量を用いて，責任閾値モデルに基づいて推定した．

11. GWAS サマリ統計量に対する trans-ethnic fine-mapping

PD の GWAS サマリ統計量に対する MHC fine-mapping は，DISH ソフトウェアを使用して Z スコアの多次元正規分布への近似により，HLA アレルの Z スコアを直接 imputation した[65]．正則化項λは，偽陰性を防ぐため，またメタアナリシスのノイズ除去の性質を考慮して，比較的小さい値 ($\lambda = 0.05$) に設定した．本研究では，各集団に対応する集団の参照パネルを用いた．東アジア人集団の GWAS メタアナリシスのサマリ統計量に対しては，Pan-Asian 参照パネル ($n = 530$) [61,66] を用いた．Indel およびマルチアレル SNV は除外し，GWAS 要約統計量と参照パネルの間で SNV の strand は，SNP2HLA と同じ基準で揃えた[30]．Imputation 後さらに，MAF ≥ 0.01 および $r^2 \geq 0.7$ でフィルタリングを行った．

サマリ統計量の条件付け解析は，各集団の参照パネルを元に GCTA COJO ソフトウェア (デフォルト設定) を使用して行なった[67]．各アレルの関連における効果量と標準誤差は，Zhu らの方法で Z スコアから導出した[68]．

各集団の参照パネル間で共有されたバリアントのリスク関連性について，メタアナリシスを行った．その際，集団間の SNV の MAF の不均一性を考慮して，パリンドロミック SNV は除外した上で，strand を揃えた．効果量と標準誤差の分布の研究間の不均一性を考慮して，サンプルサイズベースの Z スコアメタアナリシス法を行った[69]．条件付けメタアナリシスは，各データの条件付け解析をそれぞれ行い，それらをメタアナリシスする形で行った．条件付け解析の各ステップでは，関連するバリアントを共変量として追加で含め，メタアナリシスで有意水準を満たすバリアントがなくなるまで行った．厳密な基準として，ゲノムワイド有意水準 ($P = 5.0 \times 10^{-8}$) を用いた．一方，これまでの検討から PD の発症リスクに対する HLA バリアントの効果量は比較的小さいと想定されるため，効果量が小さいが意味のある関連を検出するために，本研究での解析対象の HLA バリアントの総数の Bonferroni 補正に基づく MHC 領域有意水準 $P = 3.3 \times 10^{-6}$ (= 0.05 / 15,000) も使用して評価した．

12. *HLA-DRB1* アレルの α-シヌクレインエピトープに対する結合親和性予測

先行研究で指摘されている α-シヌクレインペプチドのアミノ酸位置 Y39 から始まるエピトープ (KTKEGVLYVGSKTKE)[16] に対する *HLA-DRB1* アレルの結合親和性を，NetMHCIIpan 4.0 (BA オプション，他デフォルト設定) を使用して *in silico* に予測し

て評価した [70]．アレルグループ間の結合親和性 (nM) の違いは，Mann-Whitney U 検定を使用して評価した．いずれかの 参照パネルに収載されており，NetMHIIpan4.0 で対応している全ての 4 桁 *HLA-DRB1* アレルを対象とした．

結　果

1. 精度評価

1.1. 日本人集団における精度

日本人集団の参照パネルを用いて 10 分割交差検証法で精度評価を行った[4]．4 桁 HLA アレル全体では，DEEP*HLA は感度: 0.987，陽性適中率: 0.986，r^2: 0.984，一致率: 0.988 であった．全体の精度の違いは，SNP2HLA，HIBAG と比べて大差はなかったが，DEEP*HLA は，希少アレルに対する精度においては，他手法を上回っていた (頻度 1%未満のアレルにおいて，DEEP*HLA では感度: 0.690，陽性適中率: 0.799，r^2: 0.911，一致率: 0.691 に対して，SNP2HLA，HIBAG では順に，感度: 0.628, 0.635; 陽性適中率: 0.624，0.505; r^2 = 0.862，0.792; 一致率: 0.621，0.675; 図 4a) .

さらに，日本人集団の参照パネルを用いて独立した日本人サンプルに適用した場合の精度も評価した[23]．DEEP* HLA は，感度: 0.973，陽性適中率: 0.972，r^2: 0.986, 一致率: 0.973 で，他の手法を上回っていた (図 4a) ．この場合においても，DEEP*HLA は希少アレルに対する精度において他手法を上回っていた (頻度 1%未満のアレルにおいて，感度: 0.799; 陽性適中率: 0.877; r^2 = 0.981; 一致率: 0.691 に対して，SNP2HLA，HIBAG では順に，感度: 0.673，0.699; 陽性適中率: 0.686，0.736; r^2: 0.891，0.961; 一致率: 0.636，0.719; 図 4a) .

1.2. 欧米人集団における精度

欧米人参照パネルを用いて 10 分割交差検証法で精度評価を行った [61]. DEEP*HLA は，4 桁 HLA アレル全体では，DEEP*HLA は感度: 0.979，陽性適中率: 0.976，r^2: 0.981，一致率: 0.979 で，他の手法を上回っていた (図 4b) ．DEEP*HLA は，特に陽性適中率と r^2 において，希少アレルに対してより有利であった (頻度 1%未満の 4 桁 HLA アレルにおいて，DEEP*HLA では感度: 0.830; 陽性適中率: 0.863; r^2: 0.908; 一致率: 0.832，SNP2HLA と HIBAG で順に感度: 0.793，0.745; 陽性適中率: 0.640，0.753; r^2: 0.745，0.886; 一致率: 0.804，0.769; 図 4b) .

1.3. HLA 遺伝子座間の精度の比較

図 5 に示すように，HLA 遺伝子座間の精度の違いを評価した．既存手法では，*HLA-B* や *HLA-DRB1* の精度が他の遺伝子座よりも低いのに対し，DEEP*HLA ではその精度は比較的保たれていた．その結果，DEEP*HLA は他手法に比べて HLA 遺伝子座間の精度の平均値が最も高くばらつきは最も小さい傾向にあった．日本人の独立サンプルの希少アレルについてのみ，感度と一致率の分散が SNP2HLA よりも高かったが，SNP2HLA のほぼ全ての HLA 遺伝子の精度指標は DEEP*HLA よりも低くかった.

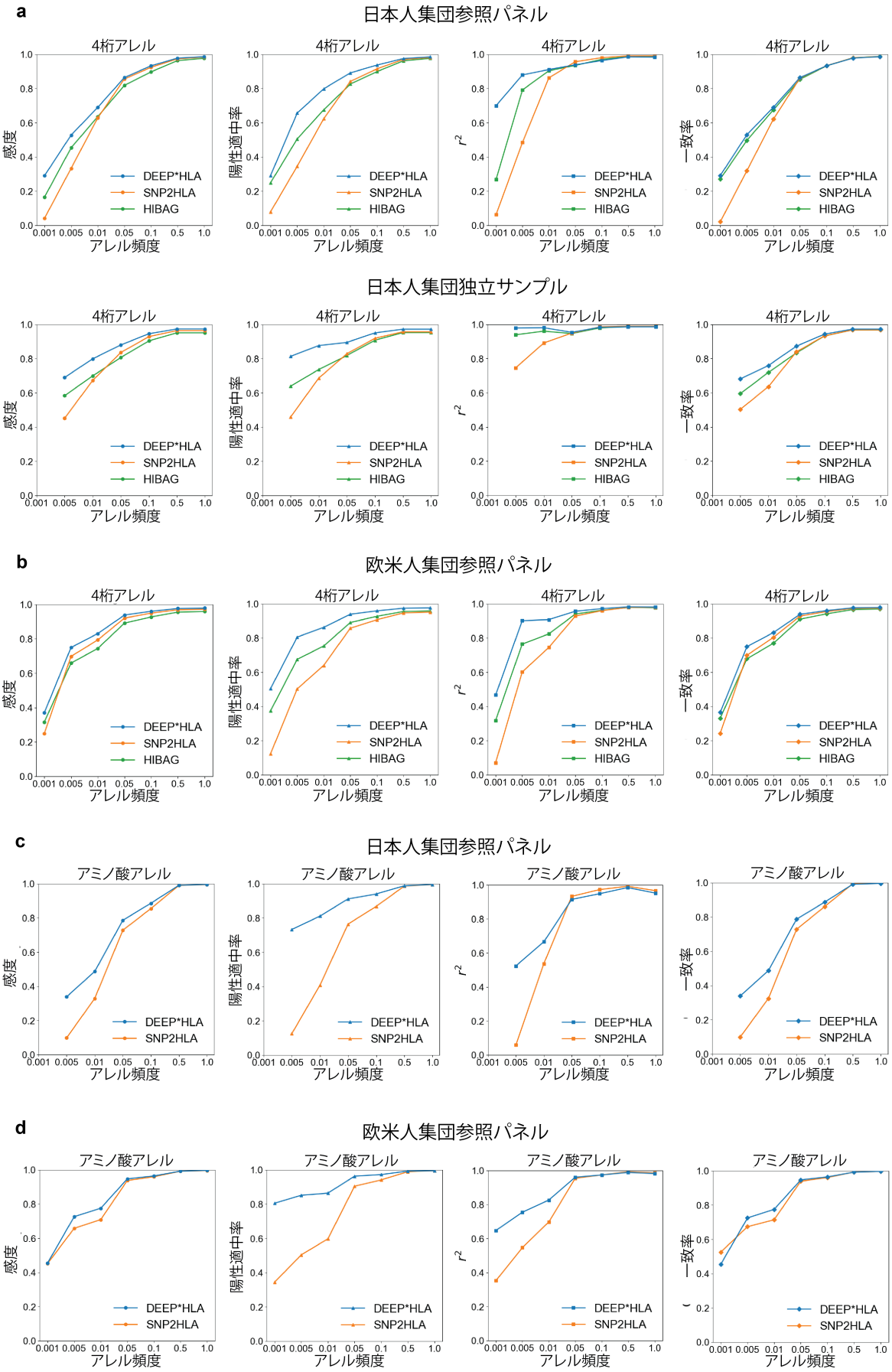
a
日本人集団参照パネル
4桁アレル
感度
陽性適中率
r^2
一致率
アレル頻度
0.001 0.005 0.01 0.05 0.1 0.5 1.0
DEEP*HLA
SNP2HLA
HIBAG
日本人集団独立サンプル
b
欧米人集団参照パネル
c
アミノ酸アレル
d

図 4: 各 HLA imputation 法の精度評価

日本人集団参照パネル (a, c) と欧米人集団参照パネル (b, d) で評価した 4 桁 HLA アレル (a, b) とアミノ酸アレル (c,d) の感度，陽性適中率，r^2，一致率を示した．各指標について，横軸にアレル頻度，縦軸に各頻度未満のアレルの精度値の平均値を示した．DEEP*HLA は特に頻度の低いアレルに対して有利であった．

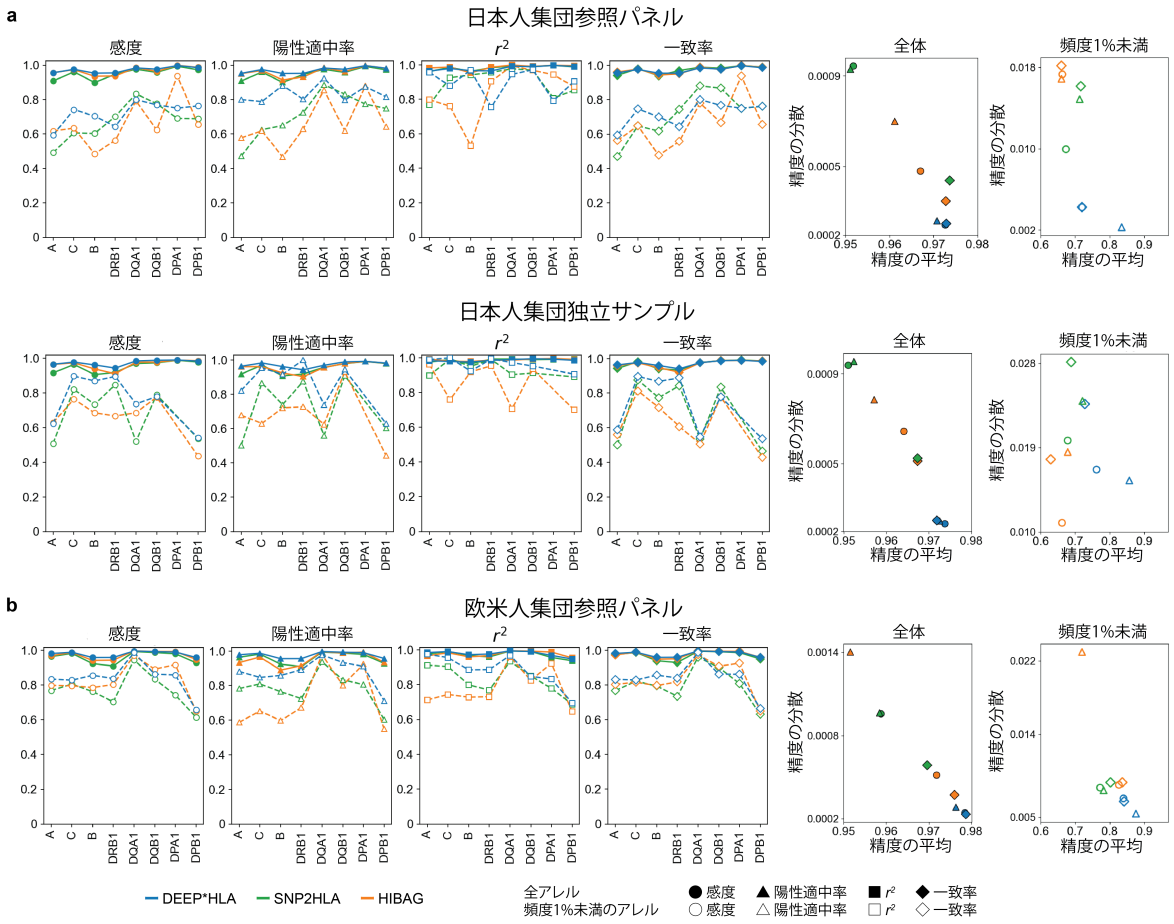

図 5: HLA 遺伝子座間の imputation 精度の比較

各パネルは，日本人集団参照パネル (a，上)，日本人集団独立サンプル (a，下)，交差検証法による欧米人集団参照パネル (b) における 8 つの古典的 HLA 遺伝子の精度を示す．実線と破線は，それぞれ全てのアレルと頻度 1%未満のアレルの平均の精度を示す．右の 2 つの散布図は，各 HLA imputation 法の各精度指標の HLA 遺伝子間で平均と分散の関係を表した．R^2 は相加的統計量ではないので示さなかった．

1.5. 混合集団パネルを用いた imputation の多様な集団における精度

日本人集団参照パネルと欧米人集団参照パネルを統合することで試験的に混合集団参照パネルを構築し，1KGv3 のデータを用いて多様な集団に対する imputation 精度を評価した．4 桁 HLA アレルにおいては，JPT では，感度: 0.965，陽性適中率: 0.940，r^2: 0.996，一致率: 0.964，EUR では，感度: 0.964，陽性適中率: 0.918，r^2: 0.983，一致率: 0.963 であり，他手法を上回っていた (図 6)．また，頻度 1%未満の 4 桁 HLA アレルにおいても，JPT では，感度: 0.965，陽性適中率: 0.940，r^2: 0.996，一致率: 0.964，EUR では，感度: 0.964，陽性適中率: 0.918，r^2: 0.983，一致率: 0.963 であり，他手法を上回っていた (図 6)．興味深いことに，参照パネルに含まれていない集団においても，多くの場合，DEEP*HLA が最も精度が高く特に希少アレルで有利であった．今回構築した混合集団参照パネルは，異なるタイピング法を用いた参照パネルを統合した試験的なものであるが，DEEP*HLA が多様な集団の参照パネルにおいても優位であることの可能性を示すと考える．

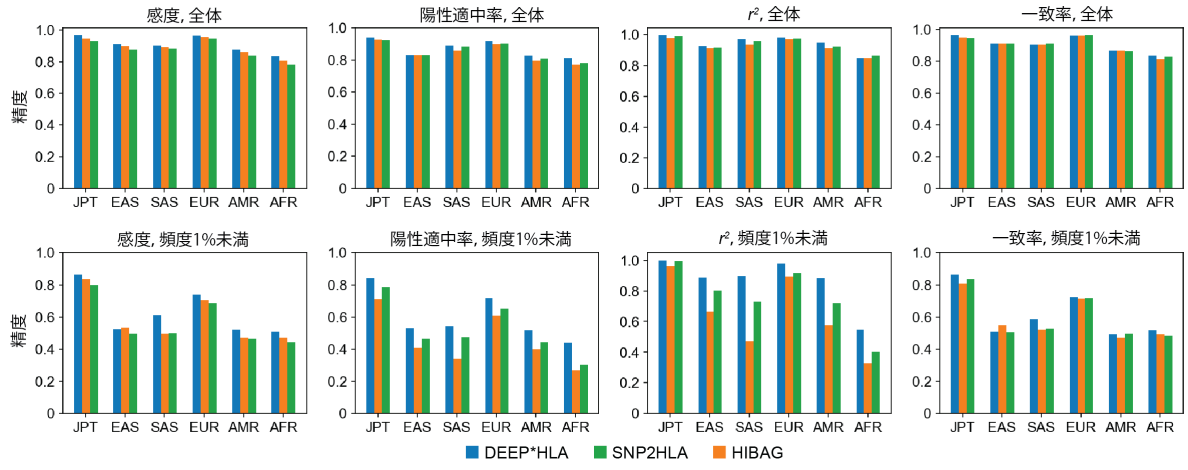

図 6: 混合集団参照パネルによる 1KGv3 における imputation 精度

各パネルは，上段が全ての 4 桁 HLA アレル，下段が頻度 1%未満の 4 桁 HLA アレルの精度の平均を示す． 各棒は，各集団を示すが，EAS は，JPT を除いたもの示した．JPT，Japanese in Tokyo; EAS，East Asian; SAS，South Asian; EUR，European; AMR，Ad Mixed American; AFR，African.

1.6. 総括

以上をまとめると，DEEP*HLA による全体的な imputation 精度の改善は比較的小さかったが，特に imputed dosage の精度指標において，頻度の低いアレルでの精度の改善は明らかであった． SNP2HLA では陽性適中率が大きく低下する傾向にあったが，これはおそらく SNP2HLA は各アレルをバイナリアレルとして個別に imputation するため，HLA 遺伝子の各アレルの dosage の総和が，理想値 (= 2.0) を超える可能性があるためであると考えられる．関連解析においては，最良推定遺伝子型ではなく dosage を用いることを考えると [3]，dosage の指標における精度の上昇は有意義である．また，HLA 遺伝子座間の imputation 精度の不均一さは fine-mapping を行うときのフィルタリングの偏りにつながるため，DEEP*HLA で遺伝子座間の精度差が小さかったことも fine-mapping における利点であると考えられる．

2. DEEP*HLA の入力感度マップ

次に，SmoothGrad を使用して，DEEP*HLA の予測において入力 SNV のどの領域が重要かを，感度マップとして評価した．DEEP*HLA モデルは，強力な LD を持つ周囲の SNV のみでなく，距離の離れた SNV のノイズにも反応した (図 7) ．強く反

応した SNV は，必ずしも LD の強さに依存せず，入力領域全体に広がっていた．この結果の解釈として，ある HLA アレルを予測することは他のアレルではないことを予測することということであるため，他の HLA アレルのいずれかとの LD 関係にある SNV が反応した可能性が考えられた．他の解釈としては，DEEP*HLA は，HLA アレルと SNV の間の単純な LD 関係ではなく，複数の SNV の複雑な組み合わせを学習して認識している可能性も考えられた．

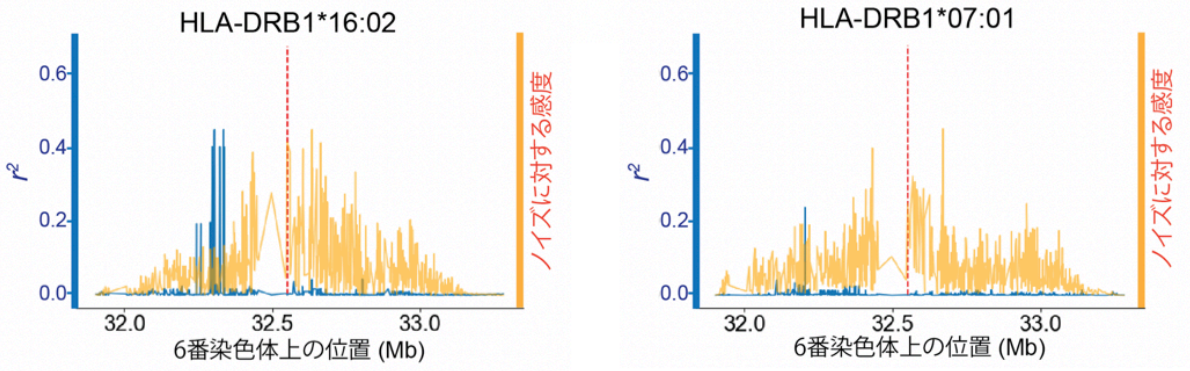

図 7: DEEP*HLA の感度マップ
DEEP*HLA の入力ノイズに対する感度マップ (オレンジ色) と HLA アレル (中央の赤い破線) に対する LD 定数 r^2 (青線) の比較．感度は正規化して示した．いずれの例においても，DEEP*HLA は LD に関係なく広範囲に渡ってノイズに反応した．

3. Imputation の信頼性の試験的な定量

深層学習モデルに共通する課題の一つは，予測の信頼性を定量化するということである [71]．DEEP*HLA による imputation の不確かさを，Bayesian 深層学習の一種である Monte Carlo ドロップアウト法を用いて試験的に評価した [64,72]．Monte Carlo ドロップアウトは，深層学習モデルにおいて，ドロップアウトをオンにした状態でサ

ンプリングを複数回行うことにより，その予測の変動をエントロピーとして計算することで，予測の不確かさを定量化する．この不確かさの指標は，HLA 遺伝子の各バイナリアレルではなく，個人の各 HLA 遺伝子のジェノタイプの予測に対応する．そこで，この不確かさが，対象となる HLA 遺伝子の最良推測遺伝子型が正解と一致しているかどうかをどの程度推測できるかを評価した．比較のために，各 HLA 遺伝子において予測したアレルの dosage が高いほど正解と一致している可能性が高いと判定する，dosage に基づいた判別の精度も評価した．Monte Carlo ドロップアウトによるエントロピーに基づいた不確かさでは，4 桁 HLA アレルの Receiver Operatorating Characteristic 曲線下面積は日本人集団参照パネルでは 0.851，欧米人集団参照パネルでは 0.883 となり，dosage に基づいた判別 (日本人集団参照パネルでは 0.722，欧米人集団参照パネルでは 0.754) よりも優れた判別能を示した．深層学習モデルの予測の不確かさの推定は，現在も発展中の分野であるが [72]，本研究の結果は，Bayesian 深層学習が，深層学習による imputation 法の信頼性指標の確立に適用できる可能性を示していると考えられる．

4. 計算負荷の評価

n = 1,000，2,000，5,000，10,000，20,000，50,000，100,000 人の GWAS ジェノタイプサブセットを使用して，各 HLA imputation 手法の計算負荷を評価した．図 8 左に示すように，DEEP*HLA は，サンプル数が大きくなるにつれて，総処理時間において有利であることがわかった (図 8a)．さらに，GPU を使用して学習を行った場合，

DEEP*HLA の学習時間は 153 分から 36 分に短縮した．最大メモリ使用量に関しては，すべての方法で，サンプル数にほぼ比例して増加した (図 8 右)． HIBAG は，すべてのサンプル数で最もメモリ効率において優れていた． SNP2HLA は，サンプル数が 20,000 を超える場合，100 GB のメモリ内では実行できなかったが，DEEP*HLA は 100,000 人のサンプル数でも実行可能であった．本実験では，比較のため実行時間をシングルスレッドで計測しているが，実用上は並列計算・実行でさらなる実行時間の短縮が見込まれるためより多くのサンプルに対応可能である．

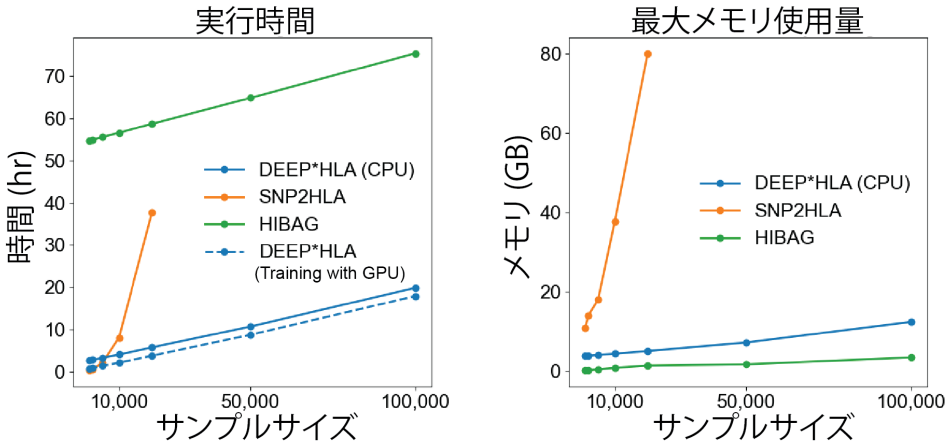

図 8: 各 HLA imputation 法の計算負荷の比較

日本人集団参照パネルを用いて BBJ のサンプルに対して HLA imputation を行った場合の実行時間 (左) と最大メモリ使用量 (右) を示した．実行時間の青の破線は，DEEP*HLA のモデル学習で GPU を使用した場合を示した．DEEP*HLA は，実行時間においてサンプルサイズが大きくなるにつれて有利であった．全ての手法で最大メモリ使用量はサンプルサイズにほぼ比例して増加していた．SNP2HLA はサンプルサイズが 20,000 以上の場合，100 GB 以内のメモリ使用量では実行できなかった．

5. T1D の MHC 領域における fine-mapping

5.1. Trans-ethnic fine-mapping の結果

DEEP*HLA で希少アレルの imputation 精度が上昇した点を活かして，T1D の trans-ethnic fine-mapping に応用し，T1D のリスク関連アレルの不均一性の問題に取り組んだ．日本人集団参照パネルと欧米人集団参照パネルで学習した DEEP*HLA モデルを，BBJ コホート (T1D 患者 831 人と対照 61,556 人) と UKB コホート (T1D 患者 732 人と対照 353,727 人) から GWAS ジェノタイプデータに適用して HLA imputation を行い，集団間のデータを統合し，trans-ethnic fine-mapping を実行した (計 T1D 患者 1,563 人と対照 415,283 人) .

HLA バリアントと T1D の関連解析により，HLA-DRβ1 pos. 71 で最も強い関連を認めた ($P_{omnibus} = 7.5 \times 10^{-120}$; 図 9a，10) ．HLA-DRβ1 pos. 71 は欧米人におけるリスク関連アミノ酸多型として以前より報告されているものの一つであった[5]．HLA-DRβ1 pos. 71 で条件付けすると，次に HLA-DQβ1 pos. 185 で最も強い独立した関連を認めた ($P_{omnibus} = 3.1 \times 10^{-69}$; 図 10) ．同様に，*HLA-DRB1*，*-DQA1*，*-DQB1* 領域内でフォワードスッテプワイズ条件付け解析を行い，HLA-DQβ1 Tyr30 ($P_{binary} = 6.7 \times 10^{-20}$; 図 10) ，HLA-DRβ1 pos. 74 ($P_{omnibus} = 1.2 \times 10^{-11}$; 図 10) ，HLA-DQβ1 Arg70 ($P_{omnibus} = 3.3 \times 10^{-9}$; 図 10) で有意な独立した関連を検出した．

上記の結果は，欧米人集団の大規模な T1D コホートを対象とした先行研究で HLA-DQβ1 pos. 57，HLA-DRβ1 pos. 13，HLA-DRβ1 pos. 71 の 3 つのアミノ酸位置が上位のリスク関連バリアントとして報告されていたのとは異なっていた．これら 3

つのリスク関連バリアントの T1D のリスクに対するオッズ比は，BBJ コホートと UKB コホートでは正の相関を示さなかった (Pearson'r = -0.59，P = 0.058)．一方，今回 trans-ethnic fine-mapping により，コホート間で正の相関を持つバリアントセットを同定することができた (Pearson'r = 0.76，$P = 6.8 \times 10^{-3}$)．

さらに，*HLA-DRB1*，*-DQA1*，*-DQB1* で条件付けすると，HLA-A pos. 62 で有意な独立した関連を認めた ($P_{omnibus} = 5.9 \times 10^{-13}$; 図 9b)． HLA-A pos. 62 で条件付けした後には，HLA-A 領域内には他の独立した関連性は認めなかった． 次に，*HLA-DRB1*，*-DQA1*，*-DQB1*，*-A* で条件付けした場合，HLA-B*54:01 とそれに特異的なアミノ酸アレル (HLA-B Gly45 と Val52) で有意な独立した関連を認めた ($P_{binary} = 1.3 \times 10^{-9}$; 図 9c)．*HLA-DRB1*，*-DQA1*，*-DQB1*，*-A*，*-B* で条件付けした場合，ゲノムワイド有意水準を満たすバリアントは存在しなかった ($P > 5.0 \times 10^{-8}$; 図 9d)．

本研究で検出されたリスク関連バリアントにおいて，アミノ酸位置の大部分は HLA 分子の構造においてペプチド結合溝に位置していた (図 11)．図 12，表 1 に各リスク関連バリアントの T1D に対するオッズ比を示す．同定されたリスク関連バリアントを用いて多変量回帰モデルを構築し，日本人と欧米人の T1D 有病率をそれぞれ 0.014%[73] と 0.4%[74] と仮定した場合，T1D の表現型分散の 10.3%と 27.6%を説明した． T1D リスクに関するオッズ比は，集団間で正の相関を示した (Pearson'r = 0.71，$P = 4.4 \times 10^{-3}$)．

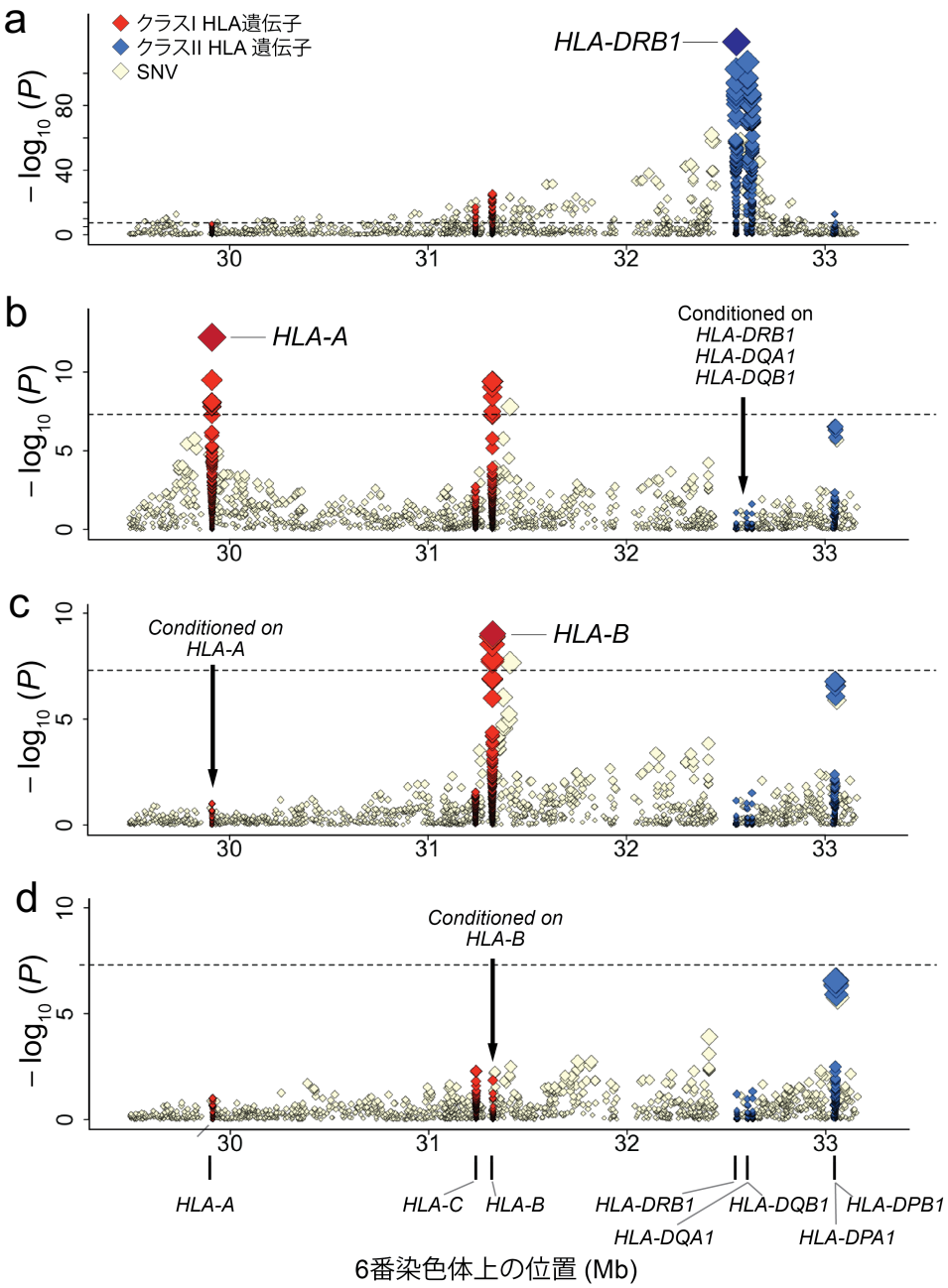

図 9: T1D の MHC 領域の trans-ethnic fine-mapping の結果

各パネルは，trans-ethnic fine-mapping におけるステップワイズ条件付き回帰分析: (a) 条件付けなしの結果，(b) *HLA-DRB1*，*-DQA1*，*-DQB1* で条件付けした結果，(c) *HLA-A*，*-DRB1*，*-DQA1*，*-DQB1* で条件付けした結果，(d) *HLA-A*，*-B*，*-DRB1*，*-DQA1*，*-DQB1* で条件付けした結果を示す．各点は，検定した HLA バリアントの-log10 (P 値) を表す．黒の破線は，ゲノムワイド有意水準 ($P = 5.0 \times 10^{-8}$) を表す．

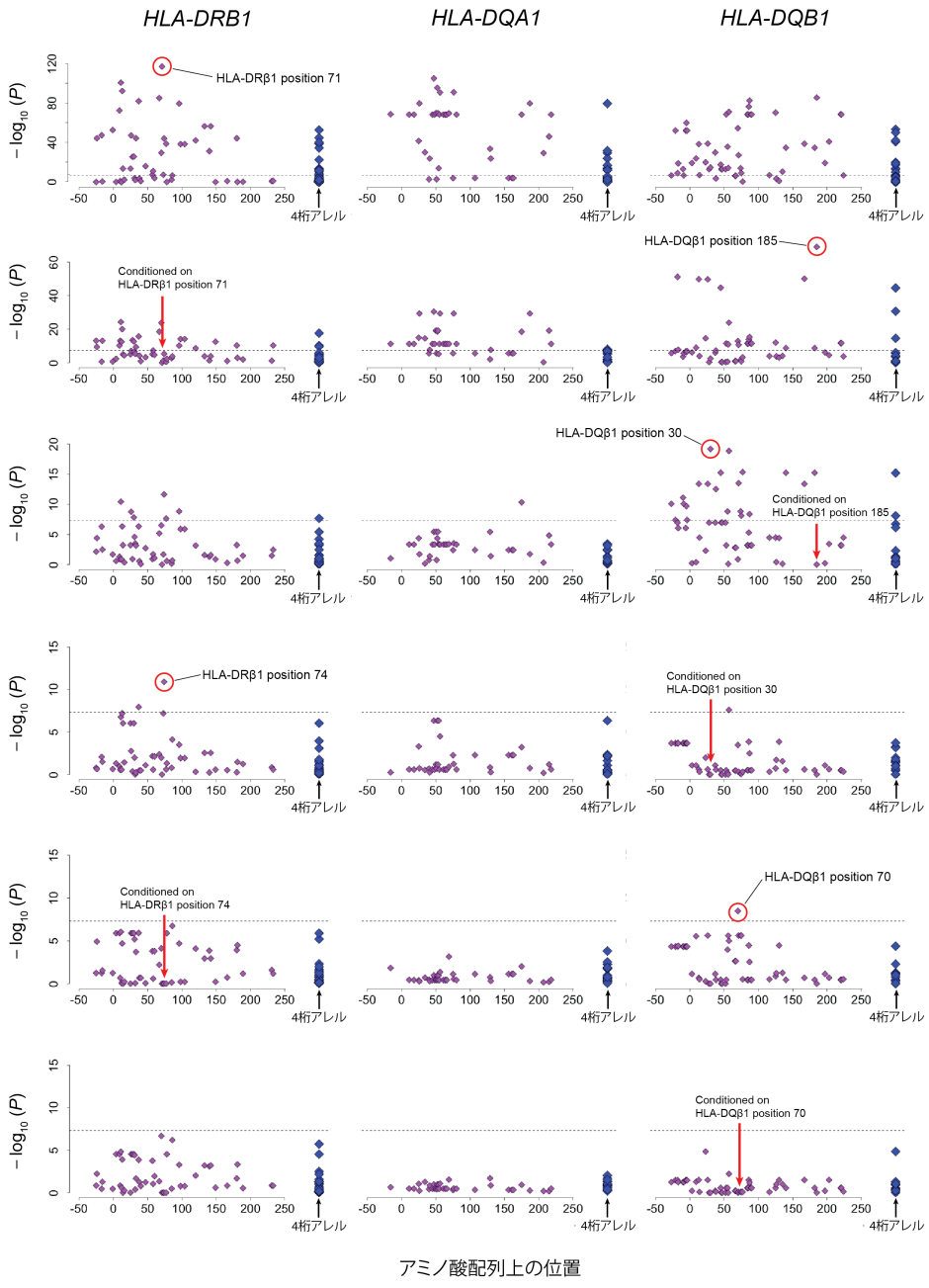

図 10: T1D の *HLA-DRB1*，*-DQA1*，*-DQB1* 領域の trans-ethnic fine-mapping の結果

各点は，HLA アミノ酸アレル (紫) と 4 桁 HLA アレル (青) の-log10 (P) 値を示す．アミノ酸アレルについては，各位置での最小の P 値を示した．ステップワイズ条件付け回帰分析の各ステップにおける最小 P 値を示すアレルを赤丸で示している．破線の水平線はゲノムワイド有意水準 ($P = 5.0 \times 10^{-8}$) を表す．

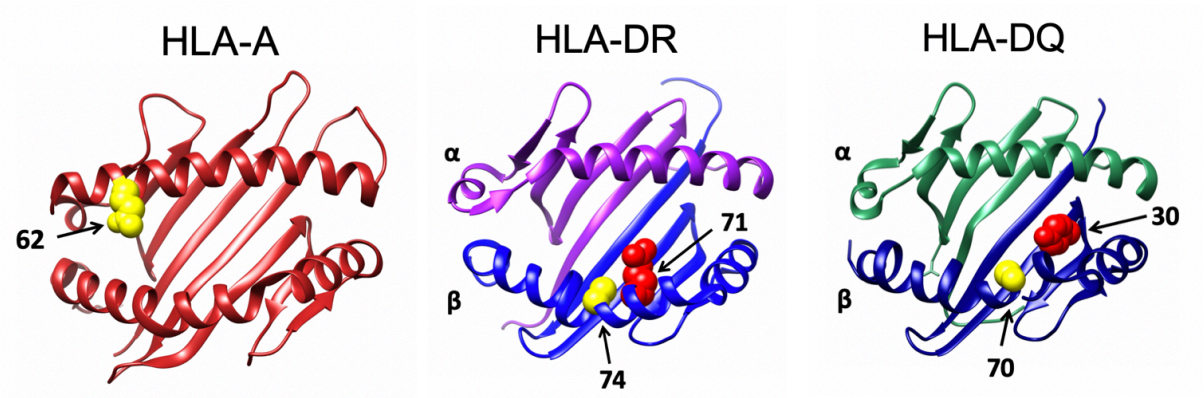

図 11: Trans-ethnic fine-mapping で同定された T1D のリスク関連アミノ酸アレルの位置

T1D のリスク関連アミノ酸アレルの位置を黄色または赤色の矢印で示した．HLA-A，HLA-DR，HLA-DQ のタンパク質構造は，Protein Data Bank のエントリ 1X7Q，3PDO，1UVQ に基づいており，UCSF Chimera version 1.14 を使用して表示した．

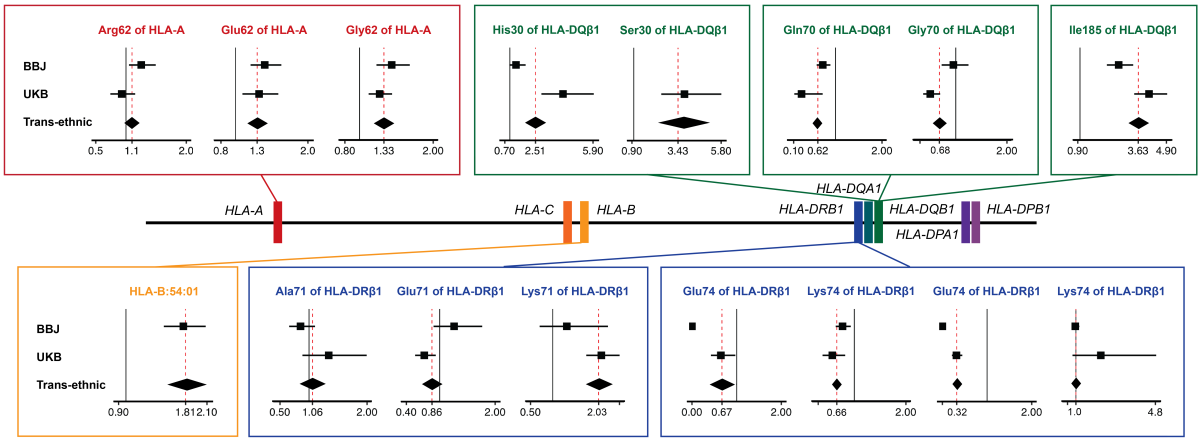

図 12: Trans-ethnic fine-mapping で同定された T1D の発症リスクに関連する HLA バリアント

T1D の各リスク関連 HLA バリアントのフォレストプロットを HLA 遺伝子の染色体上の位置に沿って示した．各ボックスは，同じ位置のアミノ酸アレルもしくは古典的 HLA アレルを表す．各フォレストプロットでは，BBJ と UKB の各コホートにおけるロジスティック回帰モデル，および trans-ethnic 集団におけるロジスティック回帰モデルに基づくオッズ比と 95%信頼区間を示した．赤色の破線は，trans-ethnic 集団におけるオッズ比，黒の実線は オッズ比 = 1 を示す．

表 1: Trans-ethnic fine-mapping で検出された T1D のリスク関連 HLA バリアント

	アレル頻度 (BBJ)		アレル頻度 (UKB)					
	T1D 患者	対照	T1D 患者	対照	オッズ比 (95%信頼区間)		P 値†	
	831 人	61,556 人	732 人	353,727 人	BBJ	UKB	BBJ	UKB
HLA-DRβ1 pos.71								
Ala	0.10	0.18	0.043	0.15	0.85 (0.66-1.10)	1.34 (0.89-1.99)	0.23	0.16
Arg	0.82	0.73	0.33	0.45	(参照)			
Glu	0.073	0.074	0.083	0.12	1.26 (0.89-1.77)	0.72 (0.56-0.93)	0.019	0.0013
Lys	0.0096	0.011	0.54	0.28	1.31 (0.71-2.24)	2.11 (1.77-2.53)	0.035	1.9×10^{-16}
HLA-DQβ1 pos.185								
Iso	0.39	0.57	0.68	0.83	2.74 (2.21-3.40)	4.12 (3.49-4.99)	3.5×10^{-20}	7.0×10^{-55}
Thr	0.61	0.43	0.32	0.17	(参照)			
HLA-DQβ1 pos.30								
His	0.16	0.19	0.18	0.23	1.36 (0.97-1.93)	4.16 (2.86-5.96)	0.0078	3.0×10^{-14}
Ser	0.0042	0.0038	0.34	0.25	inf	3.82 (2.53-5.87)	0.079	3.8×10^{-10}
Tyr	0.83	0.80	0.48	0.52	(参照)			
HLA-DRβ1 pos.74								
Ala	0.56	0.59	0.59	0.65	(参照)			
Arg	0.0018	0.00088	0.28	0.15	0 (0-0.045)	0.64 (0.42-0.96)	0.08	0.0036
Glu	0.32	0.27	0.021	0.036	0.77 (0.64-0.93)	0.57 (0.38-0.82)	0.00065	0.0004
Gln	0.0024	0.0030	0.079	0.15	0 (0-0.0029)	0.31 (0.21-0.44)	0.079	4.5×10^{-10}
Leu	0.12	0.14	0.023	0.023	0.97 (0.81-1.16)	2.20 (0.85-4.84)	0.074	0.0077
HLA-DQβ1 pos.70								
Arg	0.60	0.62	0.79	0.63	(参照)			
Glu	0.26	0.17	0.020	0.020	0.73 (0.59-0.9)	0.27 (0.11-0.71)	0.0002	0.0052
Gly	0.14	0.20	0.19	0.35	0.95 (0.72-1.25)	0.50 (0.36-0.70)	0.073	3.1×10^{-5}
HLA-A pos.62								
Arg	0.19	0.20	0.064	0.086	1.25 (1.05-1.49)	0.93 (0.74-1.16)	0.0012	0.53
Glu	0.39	0.37	0.094	0.093	1.40 (1.21-1.63)	1.33 (1.10-1.60)	9.2×10^{-6}	0.0025
Gln	0.15	0.19	0.46	0.49	(参照)			
Gly	0.26	0.24	0.33	0.29	1.44 (1.23-1.68)	1.27 (1.12-1.44)	6.6×10^{-6}	1.6×10^{-4}
Leu	0	0	0.055	0.044	-	2.01 (1.57-2.55)	1.5×10^{-12}	1.9×10^{-8}
HLA-B*54:01	0.14	0.073	0	0	1.78 (1.51-2.08)	-	-	-

BBJ，BioBank Japan; UKB，UK Biobank; HLA，human leucocyte antigen.
†記載の全ての HLA バリアントを含めた多変量回帰モデルに基づく.

5.2. 個々の集団の fine-mapping の結果

Trans-ethnic fine-mapping の利点を評価するために，各コホートに対して個別に fine-mapping を行い，それらの結果を trans-ethnic fine-mapping の結果と比較した．いずれのコホートでも，*HLA-DRB1* および *HLA-DQB1* において，最も強い関連を認めた (図 13，14)．BBJ では HLA-DQβ1 pos.185 ($P = 8.3 \times 10^{-47}$)，UKB では HLA-DRβ1 pos.71 ($P = 4.1 \times 10^{-107}$) で，いずれも trans-ethnic fine-mapping で同定されたリスク関連バリアントと一致していた．一方，その後の条件付き解析で検出されたリスク関連バリアントセットは同一ではなかった．個々の集団の fine-mapping においては，ステップワイズ条件付き解析において最相関バリアントと強い LD 関係にあり同程度の関連性を示す候補バリアントが複数存在するため，明確な fine-mapping は困難であった (図 13，14)．一方，trans-ethnic fine-mapping では，LD 構造の集団間の相違を利用して交絡を軽減し，より明確な関連シグナルとしてバリアントを特定することができたといえる．*HLA-DRB1*，*-DQA1*，*-DQB1* で条件付けした場合，BBJ コホートでは *HLA-B* (HLA-B*54:01 が最相関; $P = 4.1 \times 10^{-10}$)，UKB コホートでは *HLA-A* (HLA-A pos. 62 が最相関; $P = 1.4 \times 10^{-8}$) に，それぞれ有意な独立した関連を認めた (図 13，14)．これらは，trans-ethnic fine-mapping で同定されたものと一致していた．この結果は，単一集団を対象とした fine-mapping に比べて，trans-ethnic fine-mapping でより多くのリスク関連遺伝子座を同定できることを示しており，特に HLA-A pos. 62 の T1D の発症リスクが日本人集団 (つまり東アジア人集団) においても共有されていることを検出できたのは特記すべき点である．以上のように，より明確なシグ

ナルを検出できうる点，より多くのリスク関連遺伝子座位を検出できうる点は，trans-ethnic fine-mapping の利点であると考えられる．

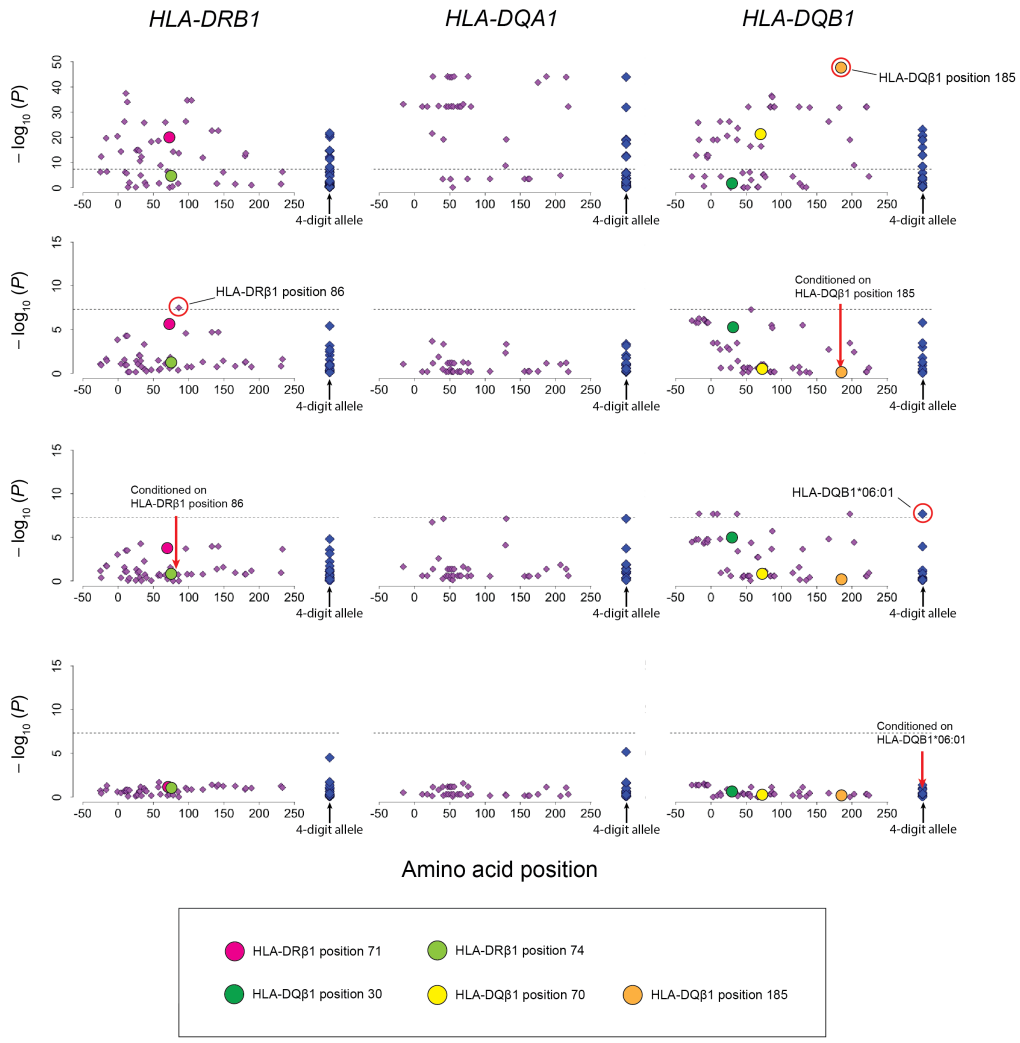

図 13: BBJ コホートにおける T1D の *HLA-DRB1*, *-DQA1*, *-DQB1* 領域の fine-mapping の結果

各点は，HLA アミノ酸アレル (紫) と 4 桁 HLA アレル (青) の-$\log_{10}$ (P) 値を示す．アミノ酸アレルについては，各位置での最小の P 値を示した．ステップワイズ条件付け回帰分析の各ステップにおける最小 P 値を示すアレルを赤丸で囲った．また，下に示すように trans-ethnic fine-mapping で検出されたリスク関連バリアントをラベル付けした．破線の水平線はゲノムワイド有意水準 ($P = 5.0 \times 10^{-8}$) を表す．

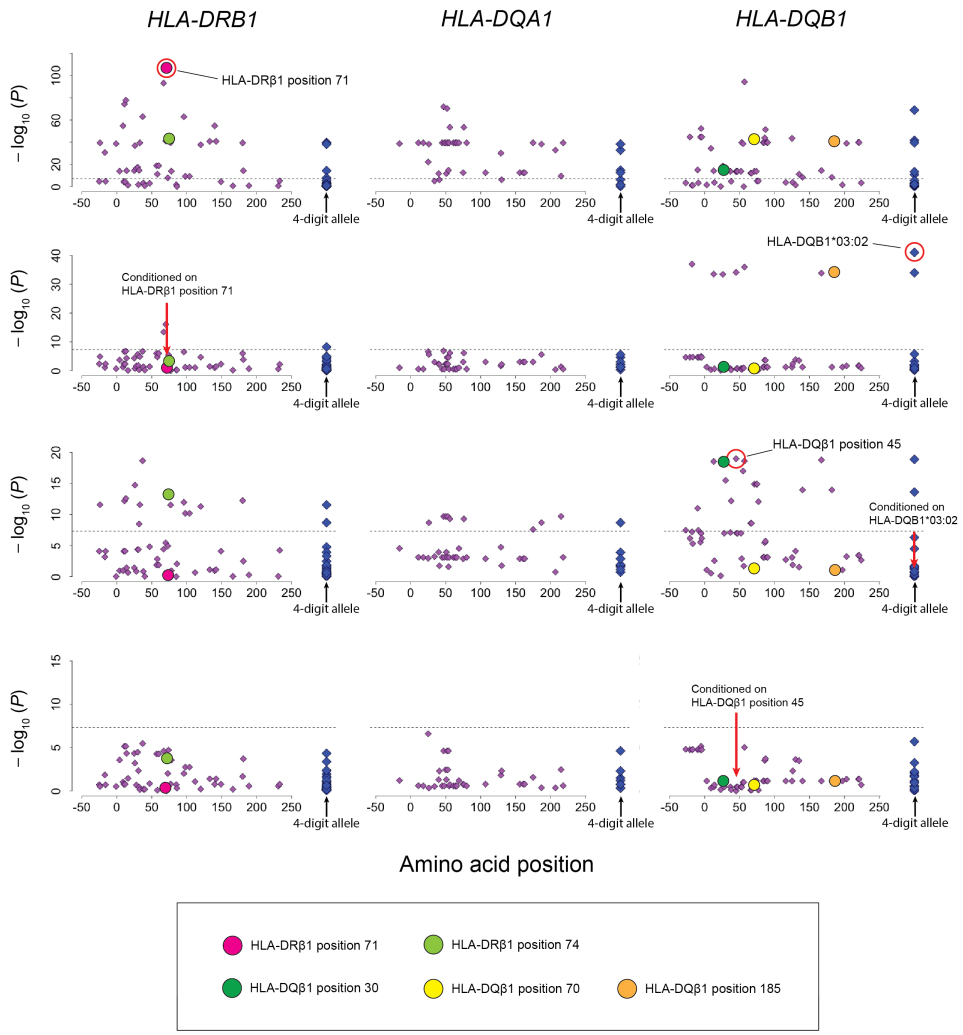

図 14: UKB コホートにおける T1D の *HLA-DRB1*, *-DQA1*, *-DQB1* 領域の fine-mapping の結果

各点は，HLA アミノ酸アレル (紫) と 4 桁 HLA アレル (青) の-$\log_{10}$ (P) 値を示す．アミノ酸アレルについては，各位置での最小の P 値を示した．ステップワイズ条件付け回帰分析の各ステップにおける最小 P 値を示すアレルを赤丸で囲った．また，下に示すように trans-ethnic fine-mapping で検出されたリスク関連バリアントをラベル付けした．破線の水平線はゲノムワイド有意水準 ($P = 5.0 \times 10^{-8}$) を表す．

6. PD の MHC 領域における fine-mapping

6.1. Trans-ethnic fine-mapping の結果

PD と MHC 領域との関連については，十分なサンプル数を対象とした fine-mapping による統一的な見解に欠けており，本研究で大規模集団を対象として trans-ethnic fine-mapping を行った．結果，HLA-DRβ1 His13 で最も強い関連を認めた ($P = 6.0 \times 10^{-15}$; 図 15a)．また，同等の関連を，HLA-DRB1*04 およびそれに対応するアミノ酸アレルである HLA-DRβ1 の Asn/His 33 でも認めた ($P = 6.1 \times 10^{-15}$)． HLA-DRβ1 His13 は HLA-DRB1*04 ($r^2 = 0.9995$) と強い LD 関係にあるため，今回の結果は HLA-DRB1*04 の保護効果を報告した先行研究の結果と合致していた [3,17]． HLA-DRβ1 pos.13 で条件付けしたところ，*HLA-DRB1* 内のバリアントの関連が著しく弱まり ($P > 0.01$)，HLA-DRβ1 pos.13 が *HLA-DRB1* 領域が持つ PD のリスクの大部分を説明することが示唆された.

HLA-DRβ1 pos.13 で条件付けしたところ，HLA-B Ala69 で最も強い独立した関連を認めた ($P = 1.0 \times 10^{-7}$; 図 15b)． HLA-DRβ1 pos.13 と HLA-B pos. 69 で条件付けした後は，有意な関連を示すバリアントを認めなかった ($P > 3.3 \times 10^{-6}$; 図 15c)．本研究で検出された 2 つのアミノ酸位置はいずれも HLA 分子の構造においてペプチド結合溝に位置していた (図 16)．図 17，表 2 に trans-ethnic fine-mapping で得られた全てのリスク関連バリアントの PD のリスクに対するオッズ比を示す．また表 3 に各集団におけるリスク関連バリアントのアレル頻度を示した (参照パネルより判断).

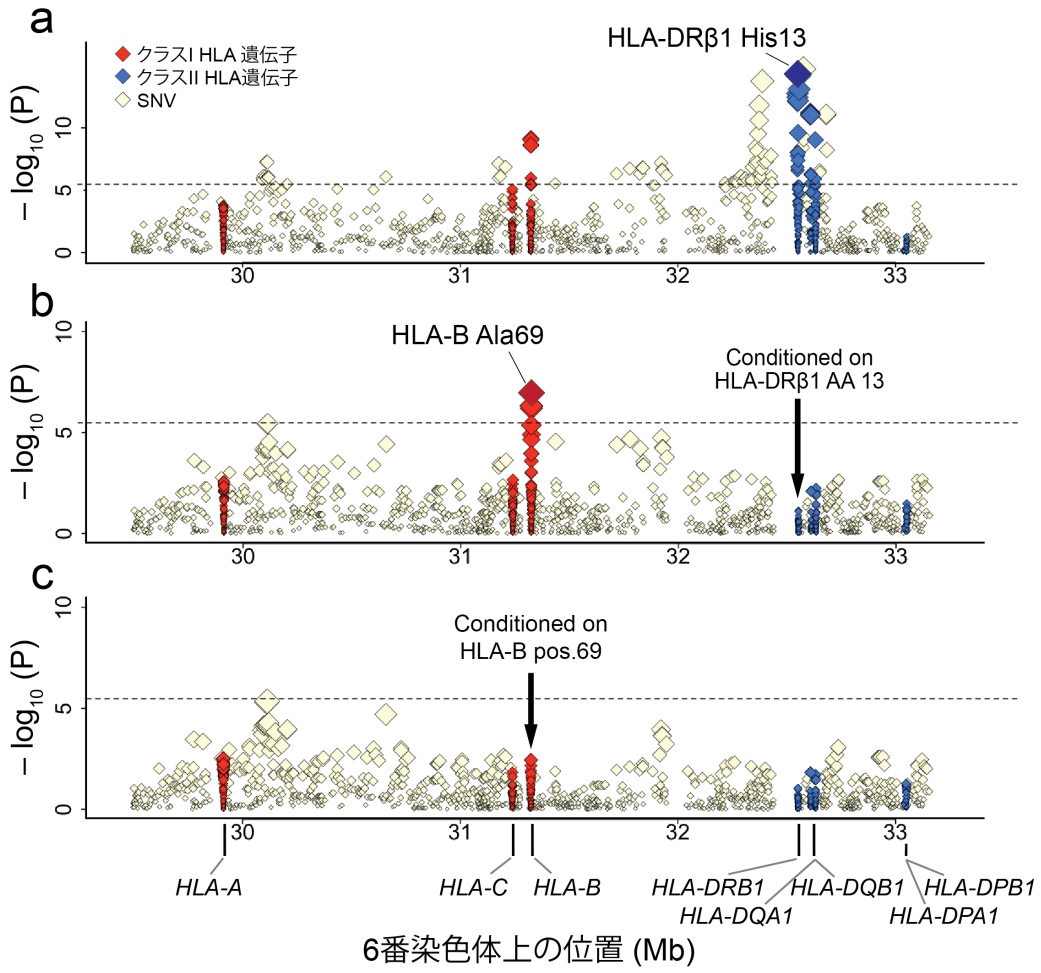

図 15: PD の MHC 領域の trans-ethnic fine-mapping の結果

各パネルは，trans-ethnic fine-mapping におけるステップワイズ条件付き回帰分析: (a) 非条件付き結果，(b) HLA-DRβ1 pos.13 で条件付けした結果，(c) HLA-DRβ1 pos.13 と HLA-B pos.69 で条件付けした結果を表す．各点は，検定した HLA バリアントの-log10 (P 値) を表す．破線の水平線は，$P = 3.3 \times 10^{-6}$ は MHC 領域有意水準を表す．

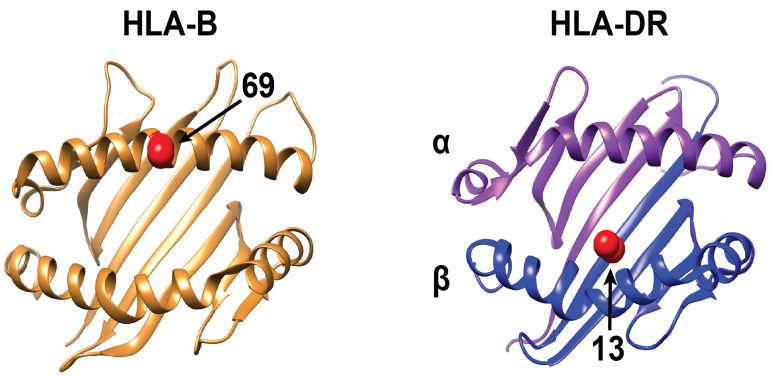

図 16: *HLA-DRB1* アレルの α-シヌクレインエピトープへの結合親和性の *in silico* 予測と PD のリスク関連 HLA アミノ酸アレルの位置

PD のリスク関連アミノ酸アレルの位置を赤色の矢印で示した．HLA-B，HLA-DR のタンパク質構造は，Protein Data Bank のエントリ 2BVP，3PDO に基づいており，UCSF Chimera version 1.14 を使用して表示した．

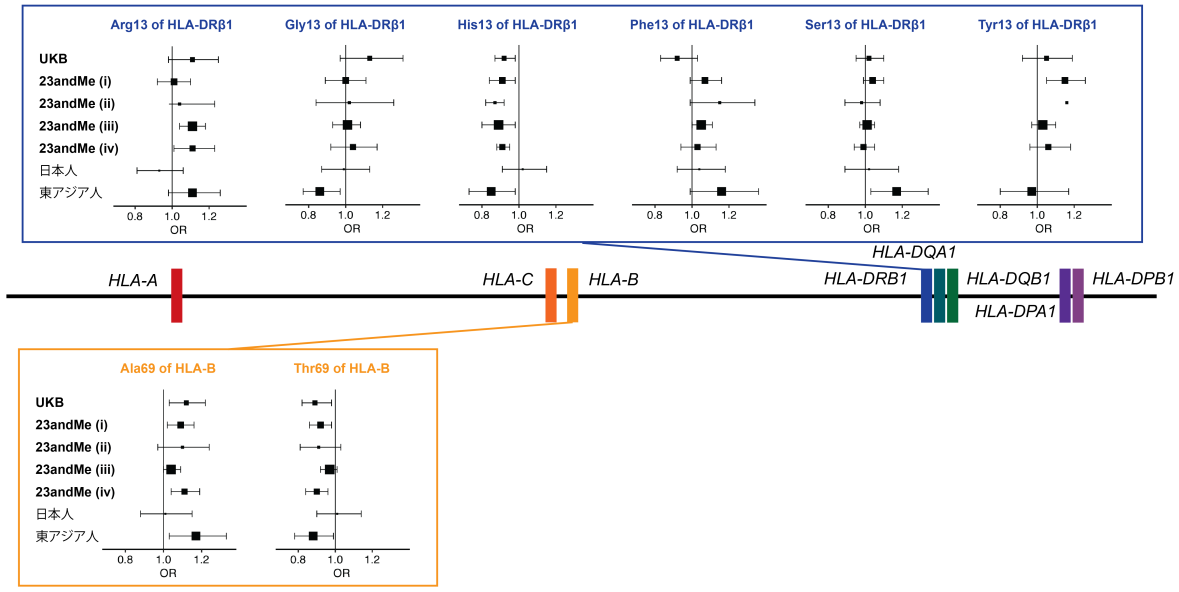

図 17: Trans-ethnic fine-mapping で同定された PD の発症リスクに関連する HLA バリアント

PD の各リスク関連 HLA バリアントのフォレストプロットを HLA 遺伝子の染色体上の位置に沿って示した．各ボックスは，同じ位置のアミノ酸アレルを表す．各フォレストプロットでは，各コホートにおけるオッズ比と 95%信頼区間 (Z スコアから近似的に算出した効果量と標準誤差による) を示した．

表 2: Trans-ethnic fine-mapping で検出された PD のリスク関連 HLA バリアント

	欧米人集団										東アジア人集団				メタ解析	
	UKB		Nalls et al. 2014		Nalls et al. 2014		Chang et al. 2017		Nalls et al. 2019		Satake et al. 2009		Foo et al. 2017			
	PD 患者 1,599 人	対照 352,325 人	PD 患者 3,261 人	対照 29,499 人	PD 患者 866 人	対照 32,538 人	PD 患者 6,476 人	対照 302,042 人	PD 患者 2,448 人	対照 571,411 人	PD 患者 988 人	対照 2,521 人	PD 患者 779 人	対照 13,227 人		
	オッズ比 (95%信頼区間) [a]	P 値	オッズ比 (95%信頼区間) [a]	P 値	オッズ比 (95%信頼区間) [a]	P 値	オッズ比 (95%信頼区間) [a]	P 値	オッズ比 (95%信頼区間) [a]	P 値	オッズ比 (95%信頼区間) [a]	P 値	オッズ比 (95%信頼区間) [a]	P 値	効果 [b]	P 値
PD の発症リスクとの関連																
HLA-DRβ1 pos.13																
Arg	1.11 (0.98-1.25)	0.10	1.01 (0.92-1.10)	0.82	1.04 (0.88-1.23)	0.66	1.11 (1.04-1.18)	7.5×10^{-4}	1.11 (1.01-1.23)	0.029	0.93 (0.81-1.06)	0.28	1.11 (0.98-1.26)	0.10	リスク	9.1×10^{-5}
Gly	1.13 (0.97-1.31)	0.11	1.00 (0.89-1.11)	0.95	1.02 (0.84-1.26)	0.82	1.01 (0.93-1.08)	0.89	1.04 (0.92-1.17)	0.56	0.99 (0.87-1.13)	0.86	0.86 (0.77-0.97)	0.013	-	0.58
His	0.91 (0.84-0.98)	0.012	0.87 (0.83-0.92)	2.9×10^{-6}	0.89 (0.80-0.99)	0.025	0.91 (0.88-0.95)	5.5×10^{-6}	0.92 (0.87-0.98)	0.0098	1.02 (0.91-1.15)	0.75	0.81 (0.70-0.93)	0.0041	保護的	6.0×10^{-15}
Phe	0.92 (0.83-1.03)	0.16	1.07 (0.99-1.16)	0.11	1.15 (0.99-1.34)	0.063	1.05 (1.00-1.11)	0.063	1.03 (0.94-1.13)	0.50	1.04 (0.92-1.18)	0.52	1.16 (0.99-1.36)	0.073	リスク	0.0036
Ser	1.02 (0.95-1.10)	0.52	1.04 (0.99-1.10)	0.11	0.98 (0.89-1.08)	0.67	1.01 (0.97-1.05)	0.64	0.99 (0.94-1.05)	0.81	1.02 (0.89-1.18)	0.74	1.17 (1.03-1.34)	0.017	リスク	0.030
Tyr	1.05 (0.92-1.19)	0.45	1.15 (1.05-1.26)	0.0040	1.16 (0.97-1.38)	0.10	1.03 (0.97-1.10)	0.37	1.06 (0.96-1.18)	0.26	0.63 (0.23-1.73)	0.37	0.97 (0.80-1.17)	0.74	リスク	0.027
PD の発症リスクとの関連 (HLA-DRβ1 pos.13 で条件付け)																
HLA-B pos.69																
Ala	1.12 (1.03-1.22)	0.012	1.09 (1.02-1.16)	0.0076	1.10 (0.97-1.24)	0.12	1.04 (1.00-1.09)	0.072	1.11 (1.04-1.19)	0.0031	1.01 (0.88-1.15)	0.91	1.17 (1.03-1.33)	0.018	リスク	1.0×10^{-7}
Thr	0.89 (0.82-0.98)	1.2×10^{-2}	0.92 (0.86-0.98)	0.0089	0.91 (0.81-1.03)	0.12	0.97 (0.92-1.01)	0.11	0.90 (0.84-0.96)	0.0022	1.01 (0.90-1.14)	0.81	0.88 (0.78-0.99)	0.032	保護的	4.8×10^{-7}

UKB，UK Biobank; PD，Parkinson 病.

[a] オッズ比，95%信頼区間は Z スコアから導出した.

[b] 有意な関連を示すアレルのみ，効果量の正負に基づいて「リスク」もしくは「保護的」と記載した.

表 3: Trans-ethnic fine-mapping で検出された PD のリスク関連 HLA バリアントの集団におけるアレル頻度

	アレル頻度 [a]		
	欧米人集団	日本人集団	東アジア人集団
HLA-DRβ1 pos.13			
Arg	0.091	0.18	0.21
Gly	0.058	0.19	0.27
His	0.29	0.26	0.14
Phe	0.11	0.21	0.11
Ser	0.37	0.16	0.19
Tyr	0.081	0.0027	0.081
HLA-B pos.69			
Ala	0.20	0.20	0.19
Thr	0.20	0.27	0.25

[a] 各集団の参照パネルを元に記載した.

6.2. 個々の集団の fine-mapping の結果

PD においても，欧米人集団，東アジア人集団で個々に fine-mapping を行い，結果を評価した．

欧米人集団のメタアナリシス (PD 患者 14,650 例と対照 1,288,625 例) では，HLA-DRβ1 His13 ($P = 2.3 \times 10^{-14}$，図 18a) と，その強い LD 関係にある SNV (rs3104413，$P = 1.3 \times 10^{-16}$，$r^2 = 0.97$) に最も強い関連が見られた．さらに，trans-ethnic fine-mapping の際と同様に，HLA-DRβ1 pos.13 と HLA-B pos.69 について条件付けしたところ，MHC クラス III 領域内においても，MHC 領域有意水準を満たして独立にリスクに関連するバリアントを検出した (rs535586，*EHMT2*; $P = 2.5 \times 10^{-7}$; 図 18b)．この SNV は，東アジア人集団の参照パネルで収載されていなかかったため，trans-ethnic fine-mapping では関連性を評価できていなかった．

東アジア人集団のメタアナリシス (PD 患者 7712 例，対照 27,372 例) では，HLA クラス II 遺伝子群よりも，HLA クラス I 遺伝子群である HLA-A*33:03 ($P = 2.9 \times 10^{-5}$) や HLA-B Ser67 ($P = 3.2 \times 10^{-5}$) が上位のシグナルを呈した (図 18c)．HLA-B Ser67 は HLA-B Ala69 と中等度に LD 関係にあるので (Pan-Asian 参照パネルで $r^2 = 0.30$)，HLA-B Ala69 のシグナルを反映している可能性があった．一方，HLA-A*33:03 は，欧米人集団のメタアナリシスでは PD リスクとの関連性は認めなかった ($P = 0.92$)．これは，HLA-A*33:03 の，欧米人集団におけるアレル頻度 (= 0.019) が東アジア人集団のアレル頻度 (= 0.10) よりも著しく低いことが一因と考えられ，PD のリスクに関与する HLA 遺伝子構造の集団間の相違を示している可能性がある．ただし，これ

らの関連は，MHC 領域有意水準 ($P < 3.3 \times 10^{-6}$; 図 18c) を満たしておらず，さらなる検証が必要と考えられる．

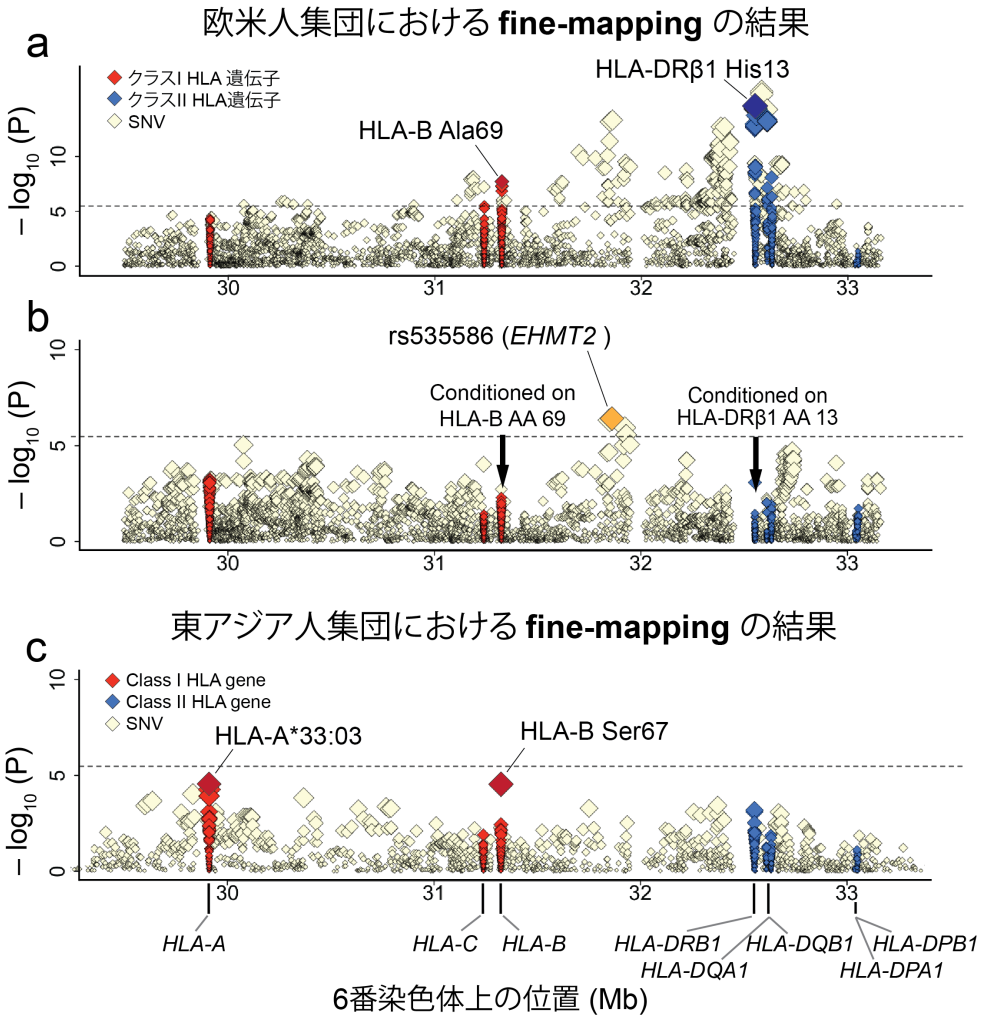

図 18: 個々の集団における，PD の MHC 領域における fine-mapping の結果

各パネルは，個々の集団における fine-mapping におけるステップワイズ条件付き回帰分析: (a) 欧米人集団，非条件付き解析の結果， (b) 欧米人集団，HLA-DRβ1 pos.13 と HLA-B pos.69 で条件付けした結果， (c) 東アジア人集団，非条件付き解析結果を表す．各点は，検定した HLA バリアントの-log10 (P 値) を表す．破線の水平線は，$P = 3.3 \times 10^{-6}$ は MHC 領域有意水準を表す．

6. 3. HLA 分子と α-シヌクレインの結合親和性の *in silico* 予測

Sulzer らの報告では，α-シヌクレイン由来のエピトープ，特に 2 つのアミノ酸位置 Y39 と S129 から始まる 2 つのペプチドが，PD 患者の T 細胞応答を誘導し，さらにその反応性が HLA クラス II アレルによって異なることが示唆されている [16]．例えば，HLA-DRB1*15:01 は， Y39 エピトープへの結合親和性が他の *HLA-DRB1* アレルのよりも強く，Y39 エピトープに応答した PD 患者は HLA-DRB1*15:01 を持っている頻度が高かった．なお HLA-DRB1*15:01 は，本研究でも PD リスクとの示唆的な関連を示していた ($P = 4.1 \times 10^{-6}$) ．その研究で評価された *HLA-DRB1* アレルの数は限られていたため，今回 *in silico* ツールである NetMHC II pan-4.0 を使用して，*HLA-DRB1* アレルの Y39 エピトープへの結合親和性を包括的に評価した．なお S129 エピトープについては，Sulzer らの報告ではリン酸化された状態でのみ T 細胞応答を誘導したとのことであり，リン酸化ペプチドは NetMHC II pan-4.0 で未対応であるため評価しなかった [16].

in silico 予測では，以前の *in vitro* アッセイの結果と一致して，HLA-DRB1*15:01 が最も強い結合親和性 (97 nM) を示した．興味深いことに，HLA-DRβ1 pos.13 の保護アレルである His13 を持つ *HLA-DRB1* アレルは，His13 を持たないアレルよりも有意に弱い結合親和性を示した ($P = 9.6 \times 10^{-4}$; 図 19) ．また逆に，HLA-DRβ1 pos.13 のリスクアレルである Arg13 を持つ *HLA-DRB1* アレルは，13His を持たないアレルよりも有意に弱い結合親和性を示した ($P = 1.0 \times 10^{-3}$; 図 19) ．これらの結果は，pos. 13 のアミノ酸アレルが α-シヌクレイン由来のエピトープの抗原提示能とそれに続く

免疫応答を変化させることによって，PD リスクと関連していることを示唆する可能性がある．

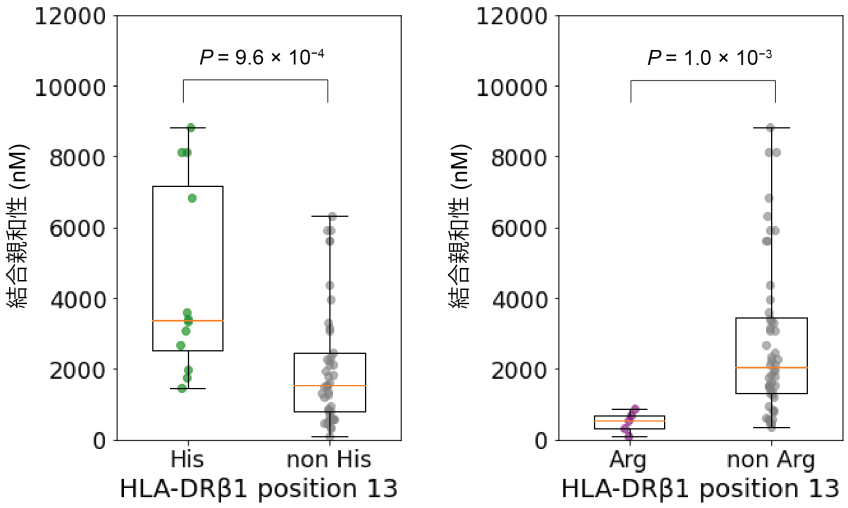

図 19: *HLA-DRB1* アレルの α-シヌクレインエピトープへの結合親和性の *in silico* 予測

α-シヌクレインの Y39 エピトープに対する HLA-DRB1 アレルの結合親和性の *in silico* 予測結果を，HLA-DRβ1 pos.13 に His (左) および Arg (右) を有するアレル間で比較したものをボックスプロットで示した．HLA-DRβ1 His13 を有するアレルと HLA-DRB1*04 アレルはほぼ同一である．また，HLA-DRB1*15:01 は HLA-DRβ1 Arg13 を有する．

考　察

本研究では，マルチタスク畳み込み深層ニューラルネットワークを応用した新規の HLA imputation 法，DEEP*HLA を開発し，包括的な性能評価の上，疾患データへの適用を行った． DEEP*HLA は，HLA アレル，アミノ酸アレルのいずれにおいても，特に希少アレルで従来の手法よりも高い精度を達成した．希少アレルの imputation 精度が改善したことを利用して，T1D と PD の発症リスクにおいて，集団間のデータの統合による trans-ethnic MHC fine-mapping を実施した．このような trans-ethnic fine-mapping のアプローチは，従来の HLA imputation 法を用いても可能ではあるが，特に T1D のリスク関連アレルにおいて，HLA-DRβ1 Lys71 や HLA-DQB1 Ser30 や HLA-DRβ1 Arg71・Glu74 など，片方の集団でのみ稀であったものがいくつか含まれており，DEEP*HLA を用いたことでより信頼性の高い fine-mapping 結果が得られた可能性がある．

大規模 GWAS ジェノタイプデータに対する深層学習の応用研究は，現時点でまだ多くなされていない．SNV のジェノタイプ imputation の以前の研究では，雑音除去自己符号化器や [44]，RNN を用いた手法 [45] で挙げられるが，いずれも制度において既存の imputation 法を上回ってはいなかった．今回，DEEP*HLA が既存の HLA imputation 法よりも高い精度を達成できた要因は 2 点考えられる．第一に，DEEP*HLA は HLA アレルを判別する分類問題として，予測対象が HLA アレのみに固定されていたことが挙げられる．第二に，様々な局所的特徴を学習できる畳み込

みカーネルを持つ CNN が，MHC 領域特有の複雑な LD 構造の学習に適していた可能性が挙げられる．また本研究では，MHC 領域を標的としたが，本手法の枠組みは他の複雑領域にも適用可能である．例えば，Killer cell immunoglobulin-like receptor (KIR) 領域は，19 番染色体長腕に位置し多数の KIR 遺伝子を含む領域であるが，MHC 領域と類似の強固で長大な LD をもち [75]，KIR アレルの予測は DEEP*HLA と類似のフレームワークがそのまま適用可能であると考えられる．KIR は正常細胞の古典的 HLA クラス I 分子を認識し NK 細胞による破壊除去を防ぐなど免疫においてはたらく分子である．

本研究では DEEP*HLA の入力領域は 1000 kb 程度 (グループ内の遺伝子自体の長さにより変動)，1 層目の畳み込み層のフィルターサイズは 128 (約 200-300kb 相当) としたが，MHC 領域は 5 mb 以上まで広がっており長大なハプロタイプ構造の存在も考えると，入力領域を拡大し，またそれに伴い畳み込みの受容野を大きくすることで予測精度が改善する可能性は考えられる．一般に畳み込みフィルターのサイズを大きくし受容野を拡大すると計算量が膨大となる可能性があるが，計算量を抑えたまま受容野を拡大する方法として，dilated covolution フィルターの使用が挙げられる [76]．Dilated convolution とは，隙間の空いた畳み込みフィルターを用いることで，パラメータ数と計算負荷を抑えたまま受容野を拡大する方法である．ただし，入力ピクセル間の繋がりが滑らかである画像処理においては，隙間の空いたフィルターでも適切に特徴抽出できると考えられるが，本研究で対象とするような離散的な SNV 入力データで上手く特徴抽出できるかはわからない．入力領域を広げるという

観点では，近年用いられるようになった attention 機構を用いたモデルの適用も有用な手段であると期待する．Attention 機構とは，ニューラルネットワークモデルにおいて入力の重要な箇所を注意喚起する仕組みである [77]．Encoder-decoder モデルに self-attention と multi-head を組み込み発展させた Transformer モデルは自然言語処理において高い精度を達成しており [77]，さらにゲノムの塩基配列を入力とした変異の機能予測にも応用されている [78]．長大な入力領域から，遠方であっても各 HLA 遺伝子の予測に重要な箇所を学習させるのに役立つ可能性が考えられる．

本研究で DEEP*HLA の imputation 結果を fine-mapping に用いるときには，交差検証法における結果を用いてアレルのフィルタリングを行ったが，フィルタリングのための別の信頼性の指標があった方がより実用的である．今回，ベイズ深層学習法に基づいた予測の不確かさの指標が，誤って imputation されたものを HLA 遺伝子レベルで識別する能力がありうることを示した．アレルレベルのフィルタリングとして実際に使用できる不確かさの定量化を実装することが，次の課題の一つである．

T1D の発症リスクに関連する MHC 領域の遺伝的特徴に関しては，欧米人では DR3-DQA1*05-DQB1*02 および DR4-DQA1*03-DQB1*03:02 ハプロタイプが [79,80]，日本人では DR9 -DQA1*03-DQB1*03:03 および DR4-DQA1*03-DQB1*04:01 ハプロタイプがリスクハプロタイプとして報告されている [81]．Hu らによる欧米人の大規模コホートを対象とした先行研究では，HLA-DRβ1 pos.13，HLA-DRβ1 pos.71，HLA-DQβ1 pos.57 の 3 つのアミノ酸アレルが，*HLA-DRB1*，*-DQA1*，*-DQB1* 領域のリスクの大部分を説明し，HLA-DQβ1 pos.57 が Asp ではないことが最もリスクに相関すると報告

されている[5]．一方，日本人における上記のリスクハプロタイプはHLA-DQβ1 Asp57を有しており，欧米人集団と日本人集団でリスクアレルの不一致が見られる[81]．本研究の trans-ethnic fine-mapping では，HLA-DRβ1 pos.71 と pos.74，HLA-DQβ1 pos.30 と pos.70 と pos.185 の 5 つの独立したリスク関連アミノ酸アレルを検出した．これらの 5 つのアミノ酸位置のうちの 4 つは，HLA 分子においてペプチド結合溝に位置しており，抗原提示能力へ機能的関与を示唆している (図 12)．HLA-DRβ1 pos.71 の関連は，欧米人集団においては，Hu らの先行研究と同様であったが，日本人集団においては同様の効果は認めなかった．他に Hu らの先行研究で報告されていたリスクアレルである HLA-DRβ1 pos.13 と HLA-DQβ1 pos.57 は本研究では検出されなかったが，HLA-DRβ1 pos.13 の代表アレルである HLA-DRβ1 His13 は欧米人集団では HLA-DQB1 Ile/Tyr185 と LD 関係にあり (それぞれ r^2 = 0.54, 0.35)，HLA-DQβ1 pos.57 の代表アレルである HLA-DQβ1 Asp57 は欧米人・日本人両集団で HLA-DQβ1 Tyr30 と LD 関係にあり (それぞれ r^2 = 0.20, 0.34)，各集団で LD 関係にあり，集団間でリスクがより共有されたアレルに修正された可能性がある．HLA-DRβ1 pos.74 の関連は，漢民族および特定の欧米人集団で報告されていたが[82,83]，Hu らの研究では，HLA-DRβ1 pos.74 と関連するアレルである HLA-DRB1*04:03 が稀であるため関連性は評価されなかった．一方，本研究では，trans-ethnic fine-mapping を行うことで，両集団で同様の効果量を持つことを示すことができた． HLA-DQβ1 pos. 185 はペプチド結合溝には位置していないが，Ile/Thr の変異は，隣接する残基と相互作用することで，HLA-DQ-DM 分子のアンカリングを変化させ，他の自己免疫疾患に対する

感受性に寄与することが示唆されている [84,85]．今回検出されたリスクアレルである HLA-DQ β1 Ile185 は，日本人集団，欧米人集団のリスクハプロタイプを構成する HLA-DQA1*03 と強い LD 関係にあり，先行研究とも矛盾しない．HLA クラス I 遺伝子に関しては，HLA-A pos.62 の独立した関連性は，Hu らの欧米人集団の先行研究と合致しており [5]，本研究では日本人集団でも同様の関連があることを示すことができた． HLA-B*54:01 は，候補遺伝子探索に基づいた研究で日本人集団におけるリスクアレルとして以前から指摘されていたが [13]，本研究では，MHC 領域全体の fine-mapping を介した独立した関連性を初めて示すことができた．

以上のように，本研究で得られた T1D のリスク関連バリアントのセットは，Hu らの欧米人集団の先行研究とは異なるものとなり [5]，本研究で検出されたバリアントは異なる集団間で全体的に共有されたリスクを有するものとなった．ただし以下の点には留意する必要がある．一点は，Hu らの研究と比較して本研究はサンプルサイズが比較的小さく，また表現型の定義が異なる (研究間および研究のコホート間) ことが，この差異の一因となっている可能性があることである [86]．特に日本人の T1D では，遅発性 T1D や fulminant diabetes mellitus が欧米に比して多いことなど病型の分布が特異性である点は以前から指摘されており留意しなければならない [86]．この点に関しては，表現型や抗体の違いに着目したサブアナリシスを行うことで，病型に応じたより詳細な fine-mapping の結果を得られる可能性がある．

本研究では，PD の発症リスクにおいても trans-ethnic MHC fine-mapping を行った．その結果， HLA-DRβ1 His13 (≒ HLA-DRB1*04) の保護効果が，最も有意なリスク関

連バリアントであることを特定した．HLA-DRβ1 pos.13 は，HLA-DR 分子のペプチド結合溝の底部に位置し，結合したペプチドと直接相互作用すると考えられる (図 16)．以前から指摘されていることだが，関節リウマチの発症リスクは HLA-DRβ1 His13 と強く相関することが知られており [3,87–89]，PD と関節リウマチの疫学的逆相関と合致している [89]．Hollenbach らの近年の報告では，*HLA-DRB1* の「共有エピトープ」と PD との関連が示されていたが [89]，そのような関連は本研究では見られなかった．さらに本研究では，PD の発症リスクと HLA クラス II 遺伝子の関連として，HLA-B Ala69 の関連の可能性を初めて示した． HLA-B Ala69 も，HLA-B 分子のペプチド結合溝を構成している (図 2b)．欧米人を対象とした先行研究の中には，HLA-B*07:02 を含むハプロタイプと PD の発症リスクとの相関を報告したものもあったが [17]，HLA-B*07:02 は HLA-B Ala69 と中程度の LD を示しており (r^2 = 0.35)，HLA-B Ala69 の持つリスク効果を反映していた可能性がある．

PD の病態における HLA 分子の役割は完全には解明されていないが，近年の研究では，α-シヌクレイン由来の Y39 エピトープとリン酸化 S129 エピトープが HLA アレルに応じて T 細胞介在性の免疫反応を誘発する可能性があることが示唆されている [16]．また，同著者はこの HLA が介在する免疫応答が PD の発症に寄与する可能性も示している [90]．本研究で行った *HLA-DRB1* アレルと Y39 エピトープの *in silico* 結合親和性予測では，保護効果を持つ His13 (HLA-DRB1*04 サブタイプ) を有するアレルは，他のアレルに比べて結合親和性が弱く，また逆にリスク効果を持つ Arg13 を有するアレルは結合親和性が強いという傾向があった．このことは，HLA-DRβ1

His13 もしくは HLA-DRB1*04 が，α-シヌクレインエピトープへの結合親和性の低下を通じて PD の発症に対して保護効果を示すという仮説を支持していると考えられる．なお，自己抗原ペプチドであまりに強く結合する T 細胞受容体を持つ T 細胞は，胸腺における負の選択で除去されうるので [91]，強い結合親和性を持つアレル・ペプチドの組み合わせが必ずしも生体でもそのように働くというわけではないため，本研究におけるモデルは単純化したモデルである．また，リン酸化 S129 エピトープは，以前の研究で HLA-DQB1*04:02 および HLA-DQB1*05:01 と高い結合親和性が示されていたが [16]，本研究では，*in silico* ツールがリン酸化ペプチドに未対応であったため評価しなかった．リン酸化 S129 ペプチドは Lewy 小体に高密度で存在し，毒性を引き起こすことによって PD の病態に寄与しうることが知られており重要である [92,93]．従って，その免疫応答の違いに焦点を当てた，様々な HLA アレルを対象とした包括的な解析は，PD の免疫学的病態機序のさらなる解明に役立つ可能性があると考えられる．また，本研究では HLA-B 分子の特定のアミノ酸位置と PD の発症リスクとの関連も示唆されており，α-シヌクレインペプチドに対する HLA-B 分子の抗原提示の実験的検証も有用である可能性がある．

α-シヌクレインが HLA 分子へ抗原提示され免疫応答の起点となるためには，マイクログリアの介在が考えられる．マイクログリアは中枢神経に存在する組織マクロファージであり活性化することで，通常のマクロファージと同様に作用する．従って，α-シヌクレインを含む神経細胞が死滅し，活性化したマイクログリアにより貪食・分解され，HLA クラス II 分子を介して CD4 陽性 T 細胞に抗原提示されるこ

とで免疫応答の起点となると考えられる．これを支持する先行研究としては，Zhang らは，ラット中脳の神経・グリア細胞培養を用いた実験で細胞外の凝集α-シヌクレインがマイクログリアを活性化しドーパミン作動性ニューロンの神経変性を促進し，さらにマイクログリアの活性には，α-シヌクレインの貪食を介していることを示した[94]．Zhang らの研究では HLA 分子の介在については検討されていないが，間接的に支持する知見としては PD 患者の剖検脳における HLA-DR 分子を発現したマイクログリアが黒質線条体において増加していること[95]などが報告されている．

また欧米人集団においては，MHC クラス III 領域にも独立して PD のリスクと関連するシグナルが存在することが示唆された．最も関連が強いバリアントである rs535586 は，*EHMT2* に位置していた．Sugeno らは，核内の α-シヌクレインが EHMT2 タンパクを介して H3K9 を活性化し，神経細胞接着分子やシナプス関連タンパク質に影響を与え，PD におけるシナプス機能障害につながる可能性を報告しており，関連性が疑われる[96]．ただし，rs535586 は東アジア人集団の参照パネルでは収載されていないため，trans-ethnic fine-mapping では関連を評価できなかった．MHC クラス III 領域は，ヒトの様々な複雑形質において HLA 遺伝子とは独立したリスクを持つことが知られており[97]，この領域を fine-mapping することにより PD の発症における遺伝的背景のさらなる解明につながる可能性がある．

本研究では，PD の trans-ethnic fine-mapping において，バイオバンクデータの疾患分類を元にした GWAS もしくは多様な先行 GWAS 研究に基づく fine-mapping のメ

タアナリシスを行っており，PD の様々な病型を同じ表現型として扱っている．一方，PD では発症年齢の違いや病型に多様性があることが知られており，そのような表現型の違いに着目して解析を行うことで，各病型に特異的な遺伝的背景に迫れる可能性がある．例えば，Blauwendraat らは，PD 患者の発症年齢について GWAS を行うことで，PD のリスク関連遺伝子として確立されているもののなかでも発症年齢に影響を及ぼすものとそうでないものがあることを報告した [98]．また，Alfradique らは振戦優位型の PD と姿勢反射障害優位型の PD に分類して GWAS を行うことで，PD の運動型の相違に影響を及ぼしうる遺伝子をいくつか報告している [99]．HLA においても病型に応じた解析を行うことで，各病型に特異的な HLA バリアントを解明できる可能性がある．ただし，病型の細分化によるサンプルサイズ・検出力の低下とのトレードオフにはなる点には留意すべきである．

PD の trans-ethnic fine-mapping においては，今回サマリ統計量による fine-mapping, メタアナリシスを行ったため，その点に関して本研究の解析上の limitation がいくつか挙げられる．効果量に基づくメタアナリシスではなく，サンプルサイズに基づく Z スコアのメタアナリシスを行ったため，多様な集団におけるリスクの効果量の不均一性をモデリングすることはできなかった．また，条件付き回帰分析では，COJO は数千人以上を対象とした参照パネルを用いることが推奨されており，東アジア人集団参照パネルのサンプルサイズは，COJO の信頼性に維持するのに十分ではない可能性があった．またステップワイズ条件付き解析においては，遺伝子座で条件付けして別の遺伝子座の独立した関連を探索する場合は，その遺伝子座を説明

する全アレルで条件付けする方がより頑健な解析法ではあるが，COJO で多くのバリアントを含めて条件付けした場合の信頼性が保証されていないため，最も関連が強いバリアントのみで条件付けした．これらの問題は，PD の GWAS データは UKB のデータを除いて個々のジェノタイプデータが入手できなかったことに起因しており，安全なデータ共有スキームの開発の必要性であると考えられる [100].

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 [33]．本研究では，日本人もしくは東アジア人集団と欧米人集団という 2 つの集団を対象とした．従って，リスク関連 HLA バリアントのより頑健な検出のための次の段階として，より多くの異なった集団を対象とした multi-ethnic fine-mapping が有効である [35]．また，本研究における trans-ethnic fine-mapping の目的は，片方の集団にのみ強い効果を持つバリアントを発見することではなく，集団間で共通したリスクを持つバリアントを検出することであったため，集団間の効果の異質性は明示的にはモデリングしなかった．より多くの集団を用いた解析においては，単一の集団によるバイアスが減少するため，不均一性のモデリングがより効果的に検出力を上昇させることができると考えられるため，その観点からも multi-ethnic fine-mapping が有効であると考えられる．深層学習の高い学習能力を考えると，DEEP*HLA は本研究のように個々の集団に対して imputation を行い統合する場合だけでなく，multi-ethnic 参照パネルを用いて多様な集団に対して一括で imputation を行う場合 も有効であろうと考

えられる．本手法を通じて，様々な多因子疾患において，集団を超えて発症リスクに寄与する MHC 領域の遺伝的特徴がさらに解明されると期待している．

結　語

深層学習を応用した HLA imputation 法，DEEP*HLA を開発した．集団中で頻度の低いアレルの imputation 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．結果，Parkinson 病と 1 型糖尿病において，異なる集団間で共通して発症リスクとなるバリアントセットを同定した．本研究結果が両疾患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 バリアントの同定に役立つことを期待している．

略 語 集

AUC — Area Under the Curve (曲線下面積)

BBJ — BioBank Japan (バイオバンク・ジャパン)

CNN — Convolutional Neural Network (畳み込みニューラルネットワーク)

CPU — Central Processing Unit (中央処理装置)

GB — Giga Byte (ギガバイト)

GPU — Graphics Processing Unit (画像処理装置)

GWAS — Genome-Wide Association Study (ゲノムワイド関連解析)

HLA — Human Leukocyte Antigen (ヒト白血球抗原)

KIR — Killer cell Immunoglobulin-like Receptor (キラー細胞免疫グロブリン様受容体)

LD — Linkage Disequibrilium (連鎖不平衡)

MHC — Major Histocompatibility Complex (主要組織適合遺伝子複合体)

NGS — Next-Generation Sequencing (次世代シークエンシング)

PD — Parkinson's Disease (Parkinson 病)

RAM — Random Access Memory (ランダムアクセスメモリ)

RNN — Recurrent Neural Network (回帰型ニューラルネットワーク)

SNV — Single Nucleotide Variant (一塩基変異)

SSO — Sequence-Specific Oligonucleotide (配列特異的オリゴヌクレオチド)

T1D — Type 1 Diabetes (1 型糖尿病)

T1DGC — The Type 1 Diabetes Genetics Consortium (1 型糖尿病遺伝学コンソーシアム)

UKB — UK Biobank (UK バイオバンク)

1KGv3 — Phase III 1000 Genomes Project (第Ⅲ相 1000 ゲノムプロジェクト)

引 用 文 献

1. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

2. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

3. Ahmed, I. *et al.* Association between Parkinson's disease and the HLA-DRB1 locus. *Mov. Disord.* **27**, 1104–1110 (2012).

4. Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).

5. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905 (2015).

6. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).

7. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).

8. Atkinson, M. A., Eisenbarth, G. S. & Michels, A. W. Type 1 diabetes. *Lancet* **383**, 69–82 (2014).

9. Britten, A. C., Mijovic, C. H., Barnett, A. H. & Kelly, M. A. Differential expression of HLA-DQ alleles in peripheral blood mononuclear cells: alleles associated with susceptibility to and protection from autoimmune type 1 diabetes. *Int. J.*

Immunogenet. **36**, 47–57 (2009).

10. Zhou, Z. & Jensen, P. E. Structural Characteristics of HLA-DQ that May Impact DM Editing and Susceptibility to Type-1 Diabetes. *Front. Immunol.* **4**, (2013).

11. Miyadera, H., Ohashi, J., Lernmark, Å., Kitamura, T. & Tokunaga, K. Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA. *J. Clin. Invest.* **125**, 275–91 (2015).

12. Todd JA, Bell JI & McDevitt HO. HLA-DQbeta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* **329**, 599–604 (1987).

13. Kawabata, Y. *et al.* Differential association of HLA with three subtypes of type 1 diabetes: Fulminant, slowly progressive and acute-onset. *Diabetologia* **52**, 2513–2521 (2009).

14. Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *Lancet* (2021) doi:10.1016/S0140-6736(21)00218-X.

15. Tan, E. K. *et al.* Parkinson disease and the immune system — associations, mechanisms and therapeutics. *Nat. Rev. Neurol.* **16**, 303–318 (2020).

16. Sulzer, D. *et al.* T cells from patients with Parkinson's disease recognize α-synuclein peptides. *Nature* **546**, 656–661 (2017).

17. Wissemann, W. T. *et al.* Association of parkinson disease with structural and regulatory variants in the hla region. *Am. J. Hum. Genet.* **93**, 984–993 (2013).

18. Zhao, Y. *et al.* Association of HLA locus variant in parkinson's disease. *Clin. Genet.* **84**, 501–504 (2013).

19. Chang, K.-H., Wu, Y.-R., Chen, Y.-C., Fung, H.-C. & Chen, C.-M. Association of genetic variants within HLA-DR region with Parkinson's disease in Taiwan. *Neurobiol. Aging* **87**, 140.e13-140.e18 (2020).

20. Erlich, H. HLA DNA typing: Past, present, and future. *Tissue Antigens* **80**, 1–11 (2012).

21. Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science (80-.).* **330**, 1551–1557 (2010).

22. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).

23. Okada, Y. *et al.* Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nat. Genet.* **47**, 798–802 (2015).

24. De Bakker, P. I. W. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).

25. Monsuur, A. J. *et al.* Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoS One* **3**, 1–6 (2008).

26. Leslie, S., Donnelly, P. & McVean, G. A Statistical Method for Predicting Classical HLA Alleles from SNP Data. *Am. J. Hum. Genet.* **82**, 48–56 (2008).

27. Li, Na (Department of Biostatistics, University of Washington, Seattle, W. 98195) & Stephens, Matthew (Department of Statistics, University of Washington, Seattle, W. 98195). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**, 2213–2233 (2003).

28. Dilthey, A. T., Moutsianas, L., Leslie, S. & McVean, G. HLA*IMP-an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics* **27**, 968–972 (2011).

29. Dilthey, A. *et al.* Multi-Population Classical HLA Type Imputation. *PLoS Comput.*

Biol. **9**, e1002877 (2013).

30. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One* **8**, e64683 (2013).

31. Levin, A. M. *et al.* Performance of HLA allele prediction methods in African Americans for class II genes HLA-DRB1, -DQB1, and -DPB1. *BMC Genet.* **15**, 1–11 (2014).

32. Gourraud, P.-A. *et al.* HLA Diversity in the 1000 Genomes Dataset. *PLoS One* **9**, e97282 (2014).

33. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 1–14 (2014).

34. Okada, Y. *et al.* Contribution of a Non-classical HLA Gene, HLA-DOA, to the Risk of Rheumatoid Arthritis. *Am. J. Hum. Genet.* **99**, 366–374 (2016).

35. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. *medRxiv Prepr.* (2020) doi:https://doi.org/10.1101/2020.07.16.20155606.

36. Karnes, J. H. *et al.* Comparison of HLA allelic imputation programs. *PLoS One* **12**, 1–12 (2017).

37. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *NIPS Proc.* 1097–1105 (2012).

38. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**, 831–838 (2015).

39. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk [supplementary]. *Nat. Genet.* **50**, 1171–1179 (2018).

40. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).

41. Naito, T. Predicting the impact of single nucleotide variants on splicing via sequence-based deep neural networks and genomic features. *Hum. Mutat.* **40**, 1261–1269 (2019).

42. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

43. Dwivedi, S. K., Tjärnberg, A., Tegnér, J. & Gustafsson, M. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nat. Commun.* **11**, (2020).

44. Chen, J. & Shi, X. Sparse convolutional denoising autoencoders for genotype imputation. *Genes (Basel)*. **10**, 1–16 (2019).

45. Kojima, K. *et al.* A genotype imputation method for de-identified haplotype reference information by using recurrent neural network. *PLOS Comput. Biol.* **16**, e1008207 (2020).

46. Abi-Rached, L. *et al.* Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLoS One* **13**, e0206512 (2018).

47. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

48. Hirata, M. *et al.* Cross-sectional analysis of BioBank Japan clinical data: A large cohort of 200,000 patients with 47 common diseases. *J. Epidemiol.* **27**, S9–S21 (2017).

49. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, 1–10

(2015).

50. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).

51. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).

52. Satake, W. *et al.* Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.* **41**, 1303–1307 (2009).

53. Foo, J. N. *et al.* Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk between Asians and Europeans: A Genome-Wide Association Study. *JAMA Neurol.* (2020) doi:10.1001/jamaneurol.2020.0428.

54. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

55. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. ICML* 448–456 (2015).

56. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *Int. Conf. Learn. Represent.* (2015).

57. Sener, O. & Koltun, V. Multi-task learning as multi-objective optimization. *Adv. Neural Inf. Process. Syst.* **2018-Decem**, 527–538 (2018).

58. Shimura, K., Li, J. & Fukumoto, F. HFT-CNN: Learning Hierarchical Category Structure for Multi-label Short Text Categorization. 811–816 (2019) doi:10.18653/v1/d18-1093.

59. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proc. ACM SIGKDD Int. Conf. Knowl.*

Discov. Data Min. 2623–2631 (2019) doi:10.1145/3292500.3330701.

60. Zheng, X. *et al.* HIBAG - HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).

61. Han, B. *et al.* Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. *Am. J. Hum. Genet.* **94**, 522–532 (2014).

62. Silver, N. C. & Dunlap, W. P. Averaging correlation coefficients: Should Fisher's z transformation be used? *J. Appl. Psychol.* **72**, 146–148 (1987).

63. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. (2017).

64. Gal, Y. & Ghahramani, Z. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. 1–17 (2015).

65. Lim, J., Bae, S.-C. & Kim, K. Understanding HLA associations from SNP summary association statistics. *Sci. Rep.* **9**, 1337 (2019).

66. Pillai, N. E. *et al.* Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. *Hum. Mol. Genet.* **23**, 4443–4451 (2014).

67. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).

68. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

69. de Bakker, P. I. W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, 122–128 (2008).

70. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and

NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).

71. Kendall, A. & Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017-Decem**, 5575–5585 (2017).

72. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd Int. Conf. Mach. Learn. ICML 2016* **3**, 1651–1660 (2016).

73. Onda, Y. *et al.* Incidence and prevalence of childhood-onset Type 1 diabetes in Japan: the T1D study. *Diabet. Med.* **34**, 909–915 (2017).

74. Sivertsen, B., Petrie, K. J., Wilhelmsen-Langeland, A. & Hysing, M. Mental health in adolescents with Type 1 diabetes: Results from a large population-based study. *BMC Endocr. Disord.* **14**, 1–8 (2014).

75. Parham, P. MHC class I molecules and kirs in human history, health and survival. *Nat. Rev. Immunol.* **5**, 201–214 (2005).

76. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018).

77. Vaswani, A. *et al.* Attention Is All You Need. *IEEE Ind. Appl. Mag.* **8**, 8–15 (2017).

78. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

79. Thomson, G. *et al.* Relative predispositional effects of HLA class II DRB1-DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. *Tissue Antigens* **70**,

110–127 (2007).

80. Erlich, H. *et al.* HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk analysis of the type 1 diabetes genetics consortium families. *Diabetes* **57**, 1084–1092 (2008).

81. Miyadera, H. & Tokunaga, K. Associations of human leukocyte antigens with autoimmune diseases: Challenges in identifying the mechanism. *J. Hum. Genet.* **60**, 697–702 (2015).

82. Cucca, F. A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins. *Hum. Mol. Genet.* **10**, 2025–2037 (2001).

83. Zhu, M. *et al.* Identification of novel T1D risk loci and their association with age and islet function at diagnosis in autoantibody-positive T1D individuals: Based on a two-stage genome-wide association study. *Diabetes Care* **42**, 1414–1421 (2019).

84. Wang, H. yu *et al.* Risk HLA class II alleles and amino acid residues in myeloperoxidase–ANCA-associated vasculitis. *Kidney Int.* **96**, 1010–1019 (2019).

85. Kachooei-mohaghegh-yaghoobi, L., Rezaei-rad, F. & Zamani, M. The impact of the HLA DQB1 gene and amino acids on the development of narcolepsy. *Int. J. Neurosci.* **0**, 1–8 (2020).

86. Kawasaki, E. & Eguchi, K. Is type 1 diabetes in the Japanese population the same as among Caucasians? *Ann. N. Y. Acad. Sci.* **1037**, 96–103 (2004).

87. Okada, Y. *et al.* Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. *Hum. Mol. Genet.* **23**, 6916–6926 (2014).

88. Sung, Y. *et al.* Reduced Risk of Parkinson Disease in Patients With Rheumatoid

Arthritis. *Mayo Clin. Proc.* **91**, 1346–1353 (2016).

89. Hollenbach, J. A. *et al.* A specific amino acid motif of HLA-DRB1 mediates risk and interacts with smoking history in Parkinson's disease. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7419–7424 (2019).

90. Lindestam Arlehamn, C. S. *et al.* α-Synuclein-specific T cell reactivity is associated with preclinical and early Parkinson's disease. *Nat. Commun.* **11**, (2020).

91. Takaba, H. & Takayanagi, H. The Mechanisms of T Cell Selection in the Thymus. *Trends Immunol.* **38**, 805–816 (2017).

92. Fujiwara, H. *et al.* A-Synuclein Is Phosphorylated in Synucleinopathy Lesions. *Nat. Cell Biol.* **4**, 160–164 (2002).

93. Ma, M. R., Hu, Z. W., Zhao, Y. F., Chen, Y. X. & Li, Y. M. Phosphorylation induces distinct alpha-synuclein strain formation. *Sci. Rep.* **6**, 1–11 (2016).

94. Zhang, W. *et al.* Aggregated α-synuclein activates microglia: a process leading to disease progression in Parkinson's disease. *FASEB J.* **19**, 533–542 (2005).

95. McGeer, P. L., Itagaki, S., Boyes, B. E. & McGeer, E. G. Reactive microglia are positive for HLA-DR in the substantia nigra of Parkinson's and Alzheimer's disease brains. *Neurology* **38**, 1285–1285 (1988).

96. Sugeno, N. *et al.* α-Synuclein enhances histone H3 lysine-9 dimethylation and H3K9me2-dependent transcriptional responses. *Sci. Rep.* **6**, 1–11 (2016).

97. Kamitaki, N. *et al.* Complement genes contribute sex-biased vulnerability in diverse disorders. *Nature* **582**, 577–581 (2020).

98. Blauwendraat, C. *et al.* Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and α-synuclein mechanisms. *Mov. Disord.* **34**, 866–875 (2019).

99. Alfradique-Dunham, I. *et al.* Genome-Wide Association Study Meta-Analysis for Parkinson Disease Motor Subtypes. *Neurol. Genet.* **7**, e557 (2021).

100. Bonomi, L., Huang, Y. & Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* **52**, 646–654 (2020).

謝　辞

本研究の遂行にあたっては，多くの方々のご指導とご協力を頂きました．本論文は東京大学大学院医学系研究科神経内科学教室・戸田達史教授，大阪大学大学院医学系研究科遺伝統計学・岡田随象教授のご指導，ご鞭撻のもとに纏められたものです．謹んで心からの感謝の意を申し上げます.

論文審査にあたっては，東京大学大学院医学系研究科人類遺伝学・藤本明洋教授，同神経病理学・岩坪威教授，同衛生学・石川俊平教授，同神経生化学・藤井哉講師，東京大学医学部附属病院脳神経外科・宮脇哲講師の貴重なご助言を賜りました．深く感謝申し上げます.

また，患者検体および対照検体の収集においては，バイオバンク・ジャパンに多大なご協力を頂きました．特に，東京大学大学院新領域創成科学研究科複雑形質ゲノム解析分野・鎌谷洋一郎教授，同クリニカルシークエンス分野・松田浩一教授に多大なご協力を頂きました．ここにお礼申し上げます.

最後になりますが，本研究の趣旨に快く賛同して頂き，検体協力にご協力頂いた全ての方々に，深く感謝，お礼を申し上げたいと思います.