A multi-task convolutional deep learning method for HLA allelic imputation and its application to trans-ethnic MHC fine-mapping of type 1 diabetes.

Tatsuhiko Naito[1, 2], Ken Suzuki[1], Jun Hirata[1,3], Yoichiro Kamatani[4], Koichi Matsuda[5], Tatsushi Toda[2], Yukinori Okada[1,6,7*].

1) Department of Statistical Genetics, Osaka University Graduate School of Medicine, 565-0871, Suita, Japan.
2) Department of Neurology, Graduate School of Medicine, The University of Tokyo, 113-8655, Tokyo, Japan.
3) Pharmaceutical Discovery Research Laboratories, Teijin Pharma Limited, 191-8512, Hino, Japan
4) Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 108-8639, Tokyo, Japan
5) Laboratory of Clinical Genome Sequencing, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 108-8639, Tokyo, Japan.
6) Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, 565-0871, Suita, Japan.
7) Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, 565-0871, Suita, Japan.

* Corresponding author:

Yukinori Okada, MD, PhD

Address: Department of Statistical Genetics, Osaka University Graduate School of Medicine,

2-2 Yamadaoka, Suita, Osaka 565-0871, Japan.

Tel: +81-6-6879-3971

E-mail: yokada@sg.med.osaka-u.ac.jp

Abstract

Conventional HLA imputation methods drop their performance for infrequent alleles, which reduces reliability of trans-ethnic MHC fine-mapping due to inter-ethnic heterogeneity in allele frequency spectra. We developed DEEP*HLA, a deep learning method for imputing HLA genotypes. Through validation using the Japanese and European HLA reference panels (*n* = 1,118 and 5,112), DEEP*HLA achieved the highest accuracies in both datasets (0.987 and 0.976) especially for low-frequency and rare alleles. DEEP*HLA was less dependent of distance-dependent linkage disequilibrium decay of the target alleles and might capture the complicated region-wide information. We applied DEEP*HLA to type 1 diabetes GWAS data of BioBank Japan (*n* = 62,387) and UK Biobank (*n* = 356,855), and successfully disentangled independently associated class I and II HLA variants with shared risk between diverse populations (the top signal at HLA-DRβ1 amino acid position 71; $P = 6.2 \times 10^{-119}$). Our study illustrates a value of deep learning in genotype imputation and trans-ethnic MHC fine-mapping.

Introduction

Genetic variants of the major histocompatibility complex (MHC) region at 6p21.3 contribute to the genetic of a wide range of human complex traits.[1] Among the genes densely contained in the MHC region, human leukocyte antigen (HLA) genes are considered to explain most of the genetic risk of MHC.[1] Strategies for direct typing of HLA alleles, including sequence specific oligonucleotide (SSO) hybridization, Sanger sequencing, and next-generation sequencing, do not easily scale for large cohorts since they are labor-intensive, time-consuming, expensive, and limited in terms of allele resolution and HLA gene coverage.[2,3] As a result, in many cases, the genotypes of HLA allele are indirectly imputed from single nucleotide variant (SNV)-level data using population-specific HLA reference panels.[3–6]

The MHC region harbors unusually complex sequence variations and haplotypes that are specific to individual ancestral populations; thus, the distribution and frequencies of the HLA alleles are highly variable across different ethnic groups.[1,7] This causes heterogeneity in reported HLA risk alleles of human complex diseases across diverse populations.[8] For example, in type I diabetes (T1D), the strong association between non-Asp57 in HLA-DQβ1 and T1D risk has been found in Europeans[9,10] but not in the Japanese population, where the T1D susceptible HLA-DQβ1 alleles carry Asp57.[11] Although elucidation of risk alleles beyond ethnicities would contribute to further understanding of genetic architecture of the MHC region associated with pathologies of complex diseases, few trans-ethnic MHC fine-mappings have been reported yet.[12] One of the ways of conducting trans-ethnic fine-mapping in the comprehensive MHC region is to newly construct a large HLA reference panel which captures the complexities of the MHC region across different populations.[13] The other is to integrate data of different populations which are imputed with a reference panel specific for each population. Although the latter way

seems straightforward, we need an HLA imputation method accurate enough for infrequent alleles to robustly evaluate HLA variants which are highly heterogenous in allele frequency across ethnicities.

Various methods for HLA allelic imputation have been developed. SNP2HLA is one of the standard software, which uses the imputation software package Beagle to impute both HLA alleles and the amino acid polymorphisms for those classical alleles.[14] HLA Genotype Imputation with Attribute Bagging (HIBAG)[15] is also promising software, which employs multiple expectation-maximization-based classifiers to estimate the likelihood of HLA alleles. While SNP2HLA explicitly uses reference haplotype data, of which public accessibility is often limited, HIBAG does not require them once the trained models are generated. Both methods have achieved high imputation accuracy;[16] however, are less accurate for rare alleles as shown later. Given the complex linkage disequilibrium (LD) structures specific for the MHC region, a more sophisticated pattern recognition algorithm beyond simple stochastic inference seems to be necessary to overcome this situation.

After boasting of its extremely high accuracy in image recognition, deep learning has been attracting attention in various fields, and a lot of successful applications in the field of genomics have been reported.[17] It can learn a representation of input data and discover relevant features of high complexity through deep neural networks. Its typical application for genomic problems is the prediction of the effects of non-coding and coding variants, where the models encodes the inputs of flanking nucleotide sequence data.[18–21] Another example is non-liner unsupervised learning of high-dimensional quantitative data of transcriptome.[22,23] However, successful representation learnings for SNV-data in the field of population genetics has been limited.[24] Here, we developed DEEP*HLA, a multi-task convolutional deep learning

92 method to accurately impute genotypes of HLA genes from SNV-level data. Through
93 application to the two HLA reference panels of different populations, DEEP*HLA achieved
94 higher imputation accuracy both in sensitivity and specificity than conventional methods.
95 Notably, it was more advantageous especially in imputing low frequent or rare alleles. As also a
96 value of our method, it was by far the fastest in total processing time, which indicates its
97 applicability to biobank-scale data. We applied the trained models of DEEP*HLA to the
98 large-scale T1D GWAS data of BioBank Japan (BBJ) and UK Biobank (UKBB), and conducted
99 trans-ethnic HLA association analysis.
100
101

Results

An overview of our study

An overview of our study is presented in **Fig. 1**. Our method, DEEP*HLA, is convolutional neural networks which learn an HLA referenced panel, and impute genotypes of HLA genes from pre-phased SNV data. Its framework uses a multi-task learning which can learn and impute alleles of several HLA genes which belong to the same group simultaneously (see Method). Multi-task learning is presumed to have two advantages in this situation. First, the genotypes of some flanking HLA genes, which often have strong LD for each other, are correlated; and the shared features of individual tasks would be informative. Second, it helps reduce the processing time by grouping tasks especially in our latest reference panel, which consists of more than thirty HLA genes. For robust benchmarking, we targeted the two different HLA imputation reference panels: (i) our Japanese reference panel (n = 1,118);[3] (ii) the Type 1 Diabetes Genetics Consortium (T1DGC) reference panel (n = 5,112),[25] respectively. We evaluated its performance in comparison with other HLA imputation methods by 10-fold cross-validation and an independent HLA dataset (n = 908).[6] In the latter part, we performed MHC fine-mappings of Japanese cohort from BBJ and British cohort from UKBB by applying the trained models specific for individual populations. We integrated the imputed GWAS genotypes and performed trans-ethnic HLA association analysis.

DEEP*HLA achieved high imputation accuracy especially in low-frequency or rare alleles

First, we applied DEEP*HLA to the Japanese panel, which is a high-resolution allele catalog of the 33 classical and non-classical HLA genes in 1,118 individuals of Japanese ancestry.[3] We compared imputation accuracy of DEEP*HLA in sensitivity and specificity (see Method) with SNP2HLA and HIBAG in 10-fold cross-validation. DEEP*HLA achieved sensitivity and specificity of 0.987 in 4-digit allelic resolution, which were superior to SNP2HLA (sensitivity of 0.985 and specificity of 0.984) and HIBAG (sensitivity and specificity of 0.979; **Supplementary Table 1**). Remarkably, DEEP*HLA was best through all ranges of allele frequencies; and was more advantageous as alleles were low frequent or rare (**Fig. 2a** and **Supplementary Table 1**). In addition to the cross-validation, to investigate whether DEEP*HLA could impute well when applied to independent samples, we applied the model trained with our Japanese reference panel to a dataset of 908 Japanese individuals (1,816 haplotypes) with 4-digit resolution alleles of 8 classical HLA genes and SNP genotype data.[6] Similarly, DEEP*HLA performed better than the other methods; and was more advantageous as alleles were low frequent or rare (**Fig. 2a** and **Supplementary Table2**).

Next, we applied DEEP*HLA to the Type 1 Diabetes Genetics Consortium (T1DGC) reference panel of 5,122 unrelated individuals of European ancestries.[25] It consists of 2- and 4-digit alleles of the 8 classical HLA gene. DEEP*HLA achieved sensitivity and specificity of 0.976 in 4-digit resolution, which were superior to SNP2HLA (sensitivity of 0.972 and specificity of 0.935) and HIBAG (sensitivity and specificity of 0.959), was more advantageous as the alleles were low frequent or rare (**Fig.2b** and **Supplementary Table 3**). There were significant declines in the specificity of SNP2HLA especially for imputing infrequent alleles, because the

sum of the allele dosages of each HLA gene of an individual can exceed the expected value (i.e. = 2.0) since it imputes each allele separately as a binary allele.

DEEP*HLA can define HLA amino acid polymorphisms without ambiguity

DEEP*HLA separately imputes classical alleles of each HLA gene, as a multi-label classification in the field of machine learning. Thus, it has an advantage that the sum of imputed allele dosages of each HLA gene is definitely set as an ideal value of 1.0 per a haplotype. This feature enables us to define a dosage of amino acid polymorphisms from the imputed 4-digit allele dosages without ambiguity. Then, we compared this method of imputing amino acid polymorphisms with SNP2HLA, which imputes them as binary alleles. Although DEEP*HLA was equivalent with SNP2HLA in both accuracy metrics in imputing amino acid polymorphisms in total (0.997 vs 0.997 in the Japanese panel; 0.996 vs 0.996 in T1DGC panel; **Supplementary Table 4, 5**), it achieved more accurate imputation for low-frequency and rare alleles (**Fig. 2c, d**). As well as in imputing classical HLA alleles, the performance improvement was remarkable in specificity evaluated in T1DGC data.

High performance of DEEP*HLA in computational costs

We benchmarked the computational costs of DEEP*HLA against SNP2HLA and HIBAG using subset of GWAS dataset from BBJ containing *n* = 1,000, 2,000, 5,000, 10,000, 20,000, 50,000, and 100,000 samples (2,000 SNPs consistent with the reference panel). Unlike SNP2HLA, DEEP*HLA and HIBAG require pre-phased GWAS data and the models trained with reference data. Thus, we compared the total processing time including pre-phasing of GWAS data, training the models, and imputation of DEEP*HLA and HIBAG, with the running time of

SNP2HLA. We used a state-of-art GPU, GeForce RTX 2080 Ti in training DEEP*HLA. As shown in **Fig. 2e**, DEEP*HLA imputation was by far the fastest in total processing time as the sample size increased. When comparing the pure imputation times, it was faster than HIBAG (**Supplementary Table 6**). As for memory cost, all methods exhibited maximum memory usage scaling roughly linearly with sample size (**Fig. 2e** and **Supplementary Table 6**), and HIBAG was the most memory-efficient through all the sample sizes. While SNP2HLA did not work within 100 GB memory of our machine for the sample size of more than 20,000, DEEP*HLA was able to impute even the biobank-scale sample size that reached 100,000.

Characteristics of the alleles where DEEP*HLA was advantageous to impute

We focused on the characteristics of the HLA alleles of which accuracy was improved by our method in comparison with SNP2HLA, which was second to our method in total accuracy metrics. SNP2HLA runs Beagle intrinsically, which performs imputation based on hidden Markov model of a localized haplotype-cluster. We hypothesized that this kind of methods works better for imputing alleles of which LDs with the surrounding SNVs are stronger in close positions and get weaker as more distant from the target HLA allele (we termed this feature as distant-dependent LD decay). Conversely, it could be limited at imputing alleles which have sparse LD structures throughout the MHC region. To verify this hypothesis, we defined the area under curve (AUC) representing distant-dependent LD decay. The AUC values become higher when LDs with the surrounding SNVs get stronger as they get closer to the target HLA allele (**Fig. 3b**). We evaluated how much two accuracies of DEEP*HLA and SNP2HLA are affected by the AUC values and allele frequency with a multivariate linear regression analysis. As expected, both sensitivity and specificity were positively correlated with AUC in SNP2HLA. On

the other hand, the specificity in DEEP*HLA were less dependent on AUC, and there was no significant correlation with the specificity in cross-validation on the Japanese panel (P = 0.069; **Fig. 3a** and **Supplementary Table 7**).

Next, to investigate our assumption that DEEP*HLA performs better imputation by recognizing distant SNVs as well as close SNVs of strong LD, we applied SmoothGrad, a method for generating a sensitivity map of a deep learning model.[26] It is a simple approach based on the idea of adding noise to the input data and taking the average of the resulting sensitivity maps for each sampled data. As displayed in its application to example HLA alleles, a trained DEEP*HLA model reacted to the noises of not only the surrounding SNVs with strong LD, but also the distant SNVs (**Fig. 3c**). Interestingly, the strongly reacted SNVs were not always those of even moderate LD, but also spread across the entire the input region. While the validity of SmoothGrad for a deep learning model of genomic data has under investigation, one probable explanation is that predicting an allele by our method conversely means predicting absence of the other alleles of the target HLA gene; thus, any SNV positions in LD with any of the other HLA alleles could be informative. Another explanation is that DEEP*HLA might recognize complicated combinations of multiple distinct SNVs within the region, rather than the simple HLA allele-SNV LD correlations.

Empirical evaluation of imputation uncertainty in deep learning models

A common issue of deep learning models is how to quantify the reliability of their predictions; and one potential solution is uncertainty inferred from the idea of Bayesian deep learning.[27] Then, we experimentally evaluated the uncertainty of imputation by DEEP*HLA using Monte Carlo (MC) dropout, which could be applied following general implementation of neural

networks with dropout units.[28,29] In MC dropout, uncertainty was presented as entropy of sampling variation with keeping dropout turned on. This uncertainty index corresponds not to each binary allele of a gene, but to the prediction of genotype of a gene of an individual. Thus, we evaluated whether it could guess the correctness of best-guess genotypes of the target HLA genes. We compared it with a dosage-based discrimination, in which we assume that a best-guess imputation of higher genotype dosage (probability) is more likely to be correct. The entropy-based uncertainty identified incorrectly imputed genotypes in areas under the curve of the receiver operating characteristic (ROC-AUC) of 0.851 in the Japanese panel, and of 0.883 in T1DGC reference panel in 4-digit alleles, which were superior to dosage-based discrimination (ROC-AUC = 0.722 in the Japanese panel and = 0.754 in T1DGC panel; **Supplementary Fig. 1**). Whereas the estimation of prediction uncertainty of a deep learning model is still developing;[29] our results might illustrate its potential applicability to establishment of a reliability score for genotype imputation by deep neural networks.

Trans-ethnic MHC fine-mapping of T1D

We applied the DEEP*HLA models trained with the Japanese panel and T1DGC panel to HLA imputation of T1D GWAS data of BBJ (831 cases and 61,556 controls) and UKBB (732 cases and 356,123 controls), respectively. T1D is a highly heritable autoimmune disease that results from T cell–mediated destruction of insulin-producing pancreatic β cells.[30] We separately imputed GWAS data of the cohorts and then combined them to perform trans-ethnic MHC fine-mapping (1,563 cases and 417,679 controls). Association analysis of the imputed HLA variants with T1D found the most significant association at the HLA-DRβ1 amino acid position 71 ($P_{omnibus} = P = 6.2 \times 10^{-119}$; **Fig. 4a and Supplementary Table 8**), one of the T1D risk amino

236 acid polymorphisms in the European population.[10] In T1D, the largest HLA gene associations
237 were reported in the *HLA-DRB1*, *-DQA1*, and *-DQB1*;[10,31] thus, we further investigated
238 independently associated variants within these HLA genes. When conditioning on HLA-DRβ1
239 amino acid position 71, we observed the most significant independent association in HLA-DQβ1
240 amino acid position 185 ($P_{omnibus}$ = 8.9 × 10^{-69}). Through stepwise forward conditional analysis
241 in the class II HLA region, we found significant independent associations in on Tyr30 in
242 HLA-DQβ1 (P_{binary} = 9.6 × 10^{-20}), HLA-DRβ1 amino acid position 74 ($P_{omnibus}$ = 1.4 × 10^{-11}), and
243 Arg70 in HLA-DQβ1 ($P_{omnibus}$ = 4.5 × 10^{-9}; **Supplementary Fig.2** and **Supplementary Table 9**).
244 The association of HLA-DRβ1 amino acid position 74 has been previously reported in
245 Europeans.[32]

246 These results were different from a previous study of large T1D cohort of European
247 ancestries, which reported three amino acid polymorphisms at HLA-DQβ1 position 57,
248 HLA-DRβ1 position 13, and HLA-DRβ1 position 71 were top-associated amino acid
249 polymorphisms in the *HLA-DRB1*, *-DQA1*, and *-DQB1* region. We then constructed multivariate
250 regression models for individual population that incorporated our T1D risk-associated HLA
251 amino acid polymorphisms and classical alleles of *HLA-DRB1* and *HLA-DQB1*, and compared
252 the effects of these variants. Whereas the odds ratios of the risk-associated variants reported
253 previously did not show any positive correlation between different populations (Pearson's r =
254 -0.59, P = 0.058; **Supplementary Fig.3** and **Supplementary Table 10**), those observed in our
255 analyses presented significant positive correlation (Pearson's r = 0.76, P = 6.8 × 10^{-3};
256 **Supplementary Fig.3**).

257 We further investigated whether T1D risk was associated with other HLA genes
258 independently of *HLA-DRB1*, *-DQA1*, and *-DQB1*. When conditioning on *HLA-DRB1*, *-DQA1*,

and -*DQB1*, we identified a significant independent association at HLA-A amino acid position 62 ($P_{omnibus}$ = 5.4 × 10^{-13}; **Fig. 4b** and **Supplementary Table 8**). After conditioning on HLA-A amino acid position 62, we did not observe any additional independent association in HLA-A alleles. When we conditioned on *HLA-DRB1*, *-DQA1*, *-DQB1*, and *-A*, we identified a significant independent association at HLA-B*54:01 (P_{binary} = 1.3 × 10^{-9}; **Fig. 4c** and **Supplementary Table 8**), and its unique amino acid alleles (Gly45 and Val52 at HLA-B). HLA-B*54:01 has traditionally been suggested as a risk allele in Japanese by a candidate HLA gene approach.[11] Its independent association through the MHC region-wide fine-mapping was first proven here. When conditioning on *HLA-DRB1*, *-DQA1*, *-DQB1*, *-A*, and *-B*, no variants in the MHC region satisfied the genome-wide significance threshold (P > 5.0 × 10^{-8}; **Fig. 4d** and **Supplementary Table 8**). Multivariate regression analysis of the identified risk variants explained 10.3% and 27.6% of the phenotypic variance in T1D under assumption of disease prevalence of 0.014%[33] and 0.4%[34] for Japanese and British cohorts, respectively. Their odds ratios on T1D risk were also correlated between different populations (Pearson's r = 0.71, P = 4.4 × 10^{-3}; **Table 1**).

Discussion

In this study, we demonstrated that DEEP*HLA, a multi-task convolutional deep learning method for HLA imputation, outperformed conventional HLA imputation methods both in sensitivity and specificity. DEEP*HLA was more advantageous when the target HLA variants, including classical alleles and amino acid polymorphisms, were low frequent or rare. Our study demonstrated that a conventional method dropped its performance for the alleles which did not exhibit distant-dependent LD decay features with the target HLA allele. DEEP*HLA was not restricted to this point, and comprehensively captures the relationships among distinct multiple variants regardless of LD.

To date, technical application of deep neural networks to population genetics data has been limited. In a previous attempt for genotype imputation, a sparse convolutional denoising autoencoder was only compared with reference-free methods.[24] There might be two possible reasons for the success of our DEEP*HLA. First unlike genotype imputation by denoising autoencoders, which assumed various positions of missing genotypes in a reference panel to impute, the prediction targets were fixed to the HLA allele genotypes as a classification problem. Second, convolutional neural networks, which leverage a convolutional kernel that is capable of learning various local patterns, might be suited for learning the complicated LD structures of the MHC region.

We filtered alleles of poor imputation quality based on the results of cross-validation in the current application; however, an indicator of reliability could be further utilized. We demonstrated that the uncertainty of prediction inferred from a Bayesian deep learning method had potential capability of distinguishing incorrectly-imputed alleles in per-gene of individuals.

Our future work should establish a method to quantify per-allele uncertainty of imputation which could be practically used as a filtering threshold for subsequent analyses.

Taking advantage of the significant improvement of imputation accuracy for rare alleles, we conducted trans-ethnic MHC fine-mapping in T1D. Our study successfully disentangled a set of independently associated amino acid polymorphisms and HLA alleles. This approach could be performed as well using the conventional HLA imputation methods. However, the results obtained by our method should be more reliable since there were several risk-associated alleles which were rare only in one population. As a result, the catalogue of the T1D risk-associated variants by our trans-ethnic approach were different from those of the previous study in Europeans.[10] We admit the possibility that the smaller sample size in our study and different definition of the phenotypes (between studies, and between cohorts in our study) might also contribute to this disparity. Especially, we note potential distinctiveness of Japanese T1D phenotypes.[35] Considering that our observed variants shared the effects on the T1D risk between different populations, however, we might gain a novel insight into the issue of inter-ethnic heterogeneity of T1D risk allele in the MHC region.

In terms of trans-ethnic analysis, we targeted the two major populations of Europeans and east Asians. As a next step, multi-ethnic MHC fine-mapping integrating further diverse ancestry should be warranted for robust prioritization of risk-associated HLA variants.[13] Given their high learning capacity of deep neural networks, our method should be helpful not only when integrating the imputation results of multiple references, but also when using a more comprehensive multi-ethnic reference. We expect that highly accurate imputation realized by learning of complex LDs in the MHC region using neural networks will enable us to further

318 elucidate the involvement of common genetic features in the MHC region that affect complex
319 traits beyond ethnicity.
320
321 **Acknowledgements**
322 We would like to thank all the participants involvement in this study. We thank the members of
323 Biobank Japan and RIKEN Center for Integrative Medical Sciences for their supports on this
324 study.
325
326 **Conflicts of interests**
327 The authors declare no conflicts of interests.
328
329 **Data availability**
330 The Japanese HLA data have been deposited at the National Bioscience Database Center
331 (NBDC) Human Database (research ID: hum0114). Independent HLA genotype data of
332 Japanese population is available in the Japanese Genotype-phenotype archive (JGA;
333 accession ID: JGAS00000000018). T1DGC HLA reference panel can be download at a NIDDK
334 central repository with a request (https://repository.niddk.nih.gov/studies/t1dgc-special/).
335 GWAS data of the BBJ are available at the NBDC Human Database (research ID: hum0014).
336 UKBB GWAS data is available upon request (https://www.ukbiobank.ac.uk/).
337
338 **Code availability**
339 Python scripts for training a model and performing imputation with our method are in
340 DEEP*HLA GitHub repository (https://github.com/tatsuhikonaito/DEEP-HLA).

Methods

The architecture of DEEP*HLA

DEEP*HLA is a multitask convolutional neural network with a shared part of two convolutional layers and a fully-connected layer, and individual fully-connected layers which output allelic dosages of individual HLA genes to impute simultaneously HLA genes of the same group (**Supplementary Fig.4**). The grouping was based on the LD structure[3] and physical distance in the current application: (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*}, and (4) {*TAP2*, *TAP1*, *HLA-DMB*, *HLA-DMA*, *HLA-DOA*, *HLA-DPA1*, *HLA-DPB1*}. Genes which were not typed or had only one allele in individual reference panels were excluded from the group.

For each group, SNPs within its window are encoded to one-hot vectors based on whether each genotype is consistent with a reference or alternative allele. The window sizes on each side were set to 500 kb in the current investigation. Two convolutional layers with max-pooling layers and a fully-connected layer follow the input layer as a shared part. The fully-connected layer in the end of shared part is followed by each fully-connected layer which has nodes consistent with the number of alleles of each HLA gene. To return a dosage of imputation, which ranges from 0.0 to 1.0 for a haplotype, softmax activation was added before the last output. Dropout was used on the convolutional and fully-connected layers,[36] and batch normalization was added to the convolutional layers.[37]

During training, 5% of data set were spared for validation to determine the point for early-stopping training (i.e. we used 85% of data were used for training in 10-fold

cross-validation). Categorical cross entropy loss function of each HLA gene was minimized using the Adam optimizing algorithm.[38] As a multi-task learning to find a Pareto optimal solution of all tasks, we used the multiple-gradient descent algorithm – upper bound (MGDA-UB), where the loss function of each task is scaled based on its optimization algorithms.[39] To taking advantage of the hierarchical nature of HLA alleles (i.e. 2-digit, 4-digit, and 6-digit), we implemented hierarchical fine-tuning, in which the parameters of model of upper hierarchical structures were transferred to those of the lower one.[40] We transferred the parameters of shared networks of 2-digit alleles to 4-digit alleles, and of 4-digit alleles to 6-digit alleles during training successively. Although some HLA alleles in our reference panel were not determined in 4-digit or 6-digit resolution, we set their upper resolution instead to keep equivalent hierarchical levels with other HLA genes. Hyperparameters, including the number of filters and kernel sizes of convolutional layers, fully-connected layer size, were tuned with Optuna.[41] The hyperparameters of the Japanese model were determined using an randomly sampled set before cross-validation, and the same values were used for hyper-parameters of the European model. Our deep learning architectures were implemented using Pytorch 1.4.1 (see URLs), a Python neural network library.

Empirical evaluation of HLA imputation accuracy

We defined two metrics to evaluate the imputation accuracy of the gene-level dosage in various aspects. First, the accuracy was calculated by summing across all individuals the dosage of each true allele in the individual, and divided by the total number of observation, as proposed in the paper of SNP2HLA.[25] We defined this as sensitivity *Se* because it counts positives that are correctly identified as such.

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

where *n* denotes the number of individuals, D_i represents the imputed dosage of an allele in individual *i*, and alleles $A1_{i,L}$ and $A2_{i,L}$ represent the true HLA alleles for individual *i* at locus *L*. In contrast, we defined specificity *Sp* as

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

where alleles $\overline{A1_{i,L}}$ and $\overline{A2_{i,L}}$ represent the HLA alleles which are incorrectly imputed dosage for individual *i* at locus *L*. Due to the nature of formula, total sensitivity and specificity of each HLA gene should be the same value for DEEP*HLA and HIBAG, in which the sum of dosage in each HLA gene of each individual is constant.

We extended these metrics for each gene to evaluate imputation performance of each allele *A*.

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

$$Sp(A) = 1 - \frac{\sum_{k=1}^{2n-m} D_k(A)}{m}$$

where *m* denotes the number of true observations of allele *A* in total sample, and D_i represents imputed dosage of allele *A* in individual haplotype *j* which has allele *A*. D_k represents imputed dosage of allele *A* in individual haplotype *k* of which true allele is not *A* (note, *Sp*(*A*) can be a negative value). Although these metrics are different from their general definitions, they are adjusted for bias due to allele frequency by dividing by true number of alleles.

When averaging the accuracy metrics, we weighted them by allele frequency.

Estimation of HLA imputation uncertainty of DEEP*HLA using MC dropout method

In order to provide uncertainty of prediction, we adopted the entropy of sampling variation of MC dropout method.[28] In MC dropout, dropout are kept during prediction to perform multiple model calls. Different units are dropped across different model calls; thus, it can be considered as Bayesian sampling with treating the parameters of a CNN model as random variables of Bernoulli distribution. The uncertainty of a best-guess genotype inferred from the entropy of sampling variation is determined as

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

where T is the number of variational samplings and t is the number of times in which obtained genotype was same as the best-guess genotype. We set T = 200 in the current investigation.

AUC metric representing distant-dependent LD decay

To evaluate whether the LD between an HLA allele and its surrounding SNVs gets weaker as the SNVs are distant to it, we calculated the area under the curve (AUC) of the cumulative curve of r^2 from the HLA allele (AUC for distance-dependent LD decay). When the LD of flanking SNVs of an HLA allele has such a characteristic, r^2 measure of LD tends to decline from the HLA allele. In other words, the bilateral cumulative curve of r^2 from the HLA allele should be more likely to be convex upward; then the AUC tends to be higher. We determined the AUC by normalizing the maximum values of r^2 sum and window sizes to 1. We evaluated its association with accuracies of each imputation method by linear regression model adjusted with an allele frequency and the maximum value of r^2. We set window size as the range of its input for evaluating the association with DEEP*HLA, and 1,000 for SNP2HLA.

Regional sensitivity maps of DEEP*HLA

We applied SmoothGrad approach to estimate which SNVs were important for DEEP*HLA to impute genotypes of each HLA gene.[26] For each haplotype, we generated 200 samples which were added Gaussian noise to encoded SNV data and input them to a trained model, and obtained the sensitivity values for individual SNV positions by averaging the absolute values of gradients caused by the difference from the true label. When we obtained the sensitivity of an allele, we averaged the maps of all haplotypes which truly has the allele.

HLA imputation software and parameter settings

We tested the latest version of each software available in Jun 2020 to compare with our method. SNP2HLA (v1.0.3) first arranges the strand in its own algorithm; however, we removed this step data during cross-validation, in which the strands must be the same between training and test data. Other settings of SNP2HLA were set to the default values. HIBAG (1.22.0.) receives phased genotypes data as input; and we used phased data generated using Beagle as well as our method. The number of classifiers were set to 25, which is sufficient to provide good performance,[42] in testing with the Japanese. For T1DGC panel, training time was extremely long with 25 classifiers; thus, we set 2 of classifiers after we confirmed that the imputation accuracy was almost unchanged in the first set of cross-validation. Flanking regions on each side was set to 500 kb.

Computational costs measurement

We measured the computational costs of imputation of subset of BioBank Japan (BBJ) Project data set (n = 1,000, 2,000, 5,000, 10,000, 20,000, 50,000, and 100,000 samples) by our Japanese reference panel (2,000 SNVs were consistent). All our runtime analyses except

449 model training of DEEP*HLA were performed on a dedicated server running CentOS 7.2.1511,
450 with 48 CPU cores (Intel ® Xeon ® E5-2687W v4 @ 3.00 GHz) and 256 GB of RAM without
451 GPU. The model training of DEEP*HLA was conducted on Ubuntu 16.04.6 LTS with 20 CPU
452 cores (Intel ® Core ™ i9-9900X @ 3.50 GHz), 2 GPUs (NVIDIA ® GeForce ® RTX 2080 Ti),
453 and 128 GB of RAM. DEEP*HLA and HIBAG require pre-phased GWAS data and the models
454 trained with reference data; thus, we measured the process not only of imputation, but also of
455 pre-phasing of GWAS data (conducted by Eagle) and training the models with a reference
456 panel. In SNP2HLA, the maximum of available memory was set to 100 GB. The processing
457 time and maximum memory usage was measured using GNU Time software when running
458 from a command line interface.
459
460 **HLA imputation reference data**
461 (i) Our Japanese reference panel and a validation dataset
462 Our Japanese reference panel contains NGS-based 6-digit resolution HLA typing data of 33
463 classical and non-classical HLA genes, of which 9 were classical HLA genes (*HLA-A*, *HLA-B*,
464 and *HLA-C* for class I; *HLA-DRA*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and
465 *HLA-DPB1* for class II) and 24 were nonclassical HLA genes (*HLA-E*, *HLA-F*, *HLA-G*, *HLA-H*,
466 *HLA-J*, *HLA-K*, *HLA-L*, *HLA-V*, *HLA-DRB2*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DRB6*,
467 *HLA-DRB7*, *HLA-DRB8*, *HLA-DRB9*, *HLA-DOA*, *HLA-DOB*, *HLA-DMA*, *HLA-DMB*, *MICA*, *MICB*,
468 *TAP1*, and *TAP2*), along with high-density SNP data of the MHC region by genotyping with the
469 Illumina HumanCoreExome BeadChip (v1.1; Illumina) of 1,120 unrelated individuals of
470 Japanese ancestry.[3] Among them, we excluded 2 individuals' data in which sides of some HLA
471 alleles were inconsistent among different resolutions after pre-phasing.

To benchmark the imputation performance when the Japanese panel is applied to independent dataset, we used 908 individuals of Japanese ancestries with 4-digit resolution alleles of classical HLA genes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*), which was used as a HLA reference panel in our previous study.[6] It contains high-density SNP data genotyped with four SNP genotyping arrays (the Illumina HumanOmniExpress BeadChip, the Illumina HumanExome BeadChip, the Illumina Immunochip, and the Illumina HumanHap550v3 Genotyping BeadChip). This study was approved by the ethical committee of Osaka University Graduate School of Medicine.

(ii) The Type 1 Diabetes Genetics Consortium (T1DGC) reference panel.

T1DGC panel contains 5,868 SNPs (genotyped with Illumina Immunochip) and 4-digit resolution HLA typing data of classical HLA genes (*HLA-A*, *HLA-B*, and *HLA-C* for class I, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQB1*, and *HLA-DRB1* for class II) of 5,225 unrelated individuals of European ancestries.[14] Among them, we excluded 103 individuals' data in which sides of some HLA alleles were inconsistent among different resolutions after pre-phasing.

T1D GWAS data in the Japanese population

The BioBank Japan (BBJ) is a multi-institutional hospital-based registry that comprised DNA, serum, and clinical information of approximately 200,000 individuals of Japanese ancestry in 2003-2007.[43,44] We used GWAS data from 831 cases who had record of T1D diagnosis and 61,556 controls of Japanese genetic ancestry enrolled in BBJ Project. The controls were included in those enrolled in our previous study that investigated the association of the MHC region to comprehensive phenotypes, and the number of T1D cases was increased.[3] The

process of patient registration, the GWAS data, and the QC process have been described elsewhere.[43–45]

T1D GWAS data in the British population

The UK Biobank (UKBB) comprises health related information approximately 500,000 individuals aged between 40-69 who were recruited from across the United Kingdom in 2006-2010.[46] We used GWAS data from 732 T1D patients and 356,123 controls of British genetic ancestry enrolled in UKBB. We selected T1D patients as individuals who were diagnosed as insulin-dependent diabetes mellitus in hospital records, and neither as non-insulin-independent diabetes mellitus in hospital records nor as type 2 diabetes in self-reported diagnosis. The controls were selected as individuals who did not have record of any autoimmune diseases neither in hospital records nor in self-reported diagnosis. We included only individuals of British ancestry according to self-identification and criteria based on principal component (PC).[47] We excluded individuals of ambiguous sex (sex chromosome aneuploidy and inconsistency between self-reported and genetic sex), and outlier of heterozygosity or call rate of high quality markers.

Imputation of the HLA variants of GWAS data of T1D and control individuals

In this study, we defined the HLA variants as SNVs in the MHC region, classical 2-digit and 4-digit biallelic HLA alleles, biallelic HLA amino acid polymorphisms corresponding to the respective residues, and multi-allelic HLA amino acid polymorphisms for each amino acid position. We applied DEEP*HLA to the GWAS data to determine classical 2-digit and 4-digit biallelic HLA alleles. The dosages of biallelic HLA amino acid polymorphisms corresponding to

518 the respective residues and multiallelic HLA amino acid polymorphisms for each amino acid
519 position were determined from the imputed 4-digit classical allele dosages. We applied
520 post-imputation filtering as the biallelic alleles of which both the sensitivity and specificity in
521 10-fold cross-validation were higher than 0.7. The sensitivity and specificity of the current
522 definition could be overestimated if an allele frequency is above 0.5; thus, we calculated those
523 with allele reversed (i.e. flipping reference/alternative alleles) and filtered also by them. The
524 SNVs in the MHC region were imputed using minimac3 (version 2.0.1) after pre-phased with
525 Eagle (version 2.3). We applied stringent post-imputation QC filtering of the variants (minor
526 allele frequency ≥ 0.5% and imputation score Rsq ≥ 0.7). For trans-ethnic fine-mapping, we
527 integrated the results of imputation of individual cohorts by including the HLA genes, amino acid
528 position, and SNVs which were typed in both reference panels. Regarding the HLA alleles and
529 amino acid polymorphisms that existed in one population, they were regarded as absent on the
530 other population. Considering the disparity in allele frequency of SNVs among different
531 populations, we removed all palindromic SNVs to align the strands correctly without fail.
532
533 **Association testing of the HLA variants**
534 We assumed additive effects of the allele dosages on the log-odds scale for susceptibility of
535 T1D; and evaluated associations of the HLA variants with the risk of T1D using a logistic
536 regression model. To robustly account for potential population stratification, we included the top
537 ten PCs obtained from the GWAS genotype data of each cohort (not including the MHC region)
538 as covariates in the regression model. For trans-ethnic analysis, PC terms of each other
539 population were set to 0; and, besides, we added a categorical variable indicating a population
540 as a covariate. We also included sex of individuals as a covariate.

To evaluate independent risk among the HLA variants and genes, we conducted a forward-type stepwise conditional regression analysis that additionally included the binary HLA variant genotypes as covariates. When conditioned on HLA gene(s), we included all the 4-digit alleles as covariates to robustly condition the associations attributable to the HLA genes, as previously described.[3,12] When conditioning on the specific HLA amino acid position(s), we included the multi-allelic variants of the amino acid residues. We applied a forward stepwise conditional analysis for the HLA variants and then HLA genes, based on the genome-wide association significance threshold ($P = 5.0 \times 10^{-8}$).

We tested a multivariate full regression model by including the risk-associated HLA variants in *HLA-DRB1*, *HLA-DQB1*, *HLA-A*, and *HLA-B*, which were identified through the stepwise regression analysis. When we included amino acid polymorphisms in the model, we excluded the most frequent residue in the British cohort from each amino acid position as the reference allele. The phenotypic variance explained by the identified risk-associated HLA variants was estimated on the basis of a liability threshold model assuming the population-specific prevalence of T1D and using the effect sizes obtained from the multivariate regression model.

URLs

DEEP*HLA, https://github.com/tatsuhikonaito/DEEP-HLA

Pytorch, http://pytorch.org/

SNP2HLA, http://software.broadinstitute.org/mpg/snp2hla/

HIBAG, https://www.bioconductor.org/packages/release/bioc/html/HIBAG.html

Eagle, https://data.broadinstitute.org/alkesgroup/Eagle/

564 Minimac3, https://genome.sph.umich.edu/wiki/Minimac3

565 Biobank Japan, https://biobankjp.org/english/index.html

566 UK biobank, https://www.ukbiobank.ac.uk/

References

1. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
2. Erlich, H. HLA DNA typing: Past, present, and future. *Tissue Antigens* **80**, 1–11 (2012).
3. Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
4. International HIV Controllers Study *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
5. 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).
6. 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).
7. Gourraud, P. A. *et al.* HLA diversity in the 1000 genomes dataset. *PLoS One* **9**, (2014).
8. 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).
9. Todd JA, Bell JI & McDevitt HO. HLA-DQbeta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* **329**, 599–604 (1987).
10. 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).
11. 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).

12. 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).

13. Luo, Y. *et al.* A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. Preprint at https://www.medrxiv.org/content/10.1101/2020.07.16.20155606v1 (2020).

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

15. 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).

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

17. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).

18. 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).

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

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

21. 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).

22. 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).

23. 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).

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

25. 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).

26. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise. Preprint at https://arxiv.org/abs/1706.03825 (2017).

27. 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).

28. Gal, Y. & Ghahramani, Z. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. Preprint at https://arxiv.org/abs/1506.02158 (2015).

29. 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).

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

31. 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).

636 32. Cucca, F. A correlation between the relative predisposition of MHC class II alleles to type
637 1 diabetes and the structure of their proteins. *Hum. Mol. Genet.* **10**, 2025–2037 (2001).
638 33. Onda, Y. *et al.* Incidence and prevalence of childhood-onset Type 1 diabetes in Japan:
639 the T1D study. *Diabet. Med.* **34**, 909–915 (2017).
640 34. Sivertsen, B., Petrie, K. J., Wilhelmsen-Langeland, A. & Hysing, M. Mental health in
641 adolescents with Type 1 diabetes: Results from a large population-based study. *BMC*
642 *Endocr. Disord.* **14**, 1–8 (2014).
643 35. Kawasaki, E. & Eguchi, K. Is type 1 diabetes in the Japanese population the same as
644 among Caucasians? *Ann. N. Y. Acad. Sci.* **1037**, 96–103 (2004).
645 36. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout□:
646 A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**,
647 1929–1958 (2014).
648 37. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by
649 reducing internal covariate shift. *Proc. ICML* 448–456 (2015).
650 38. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *Int. Conf. Learn.*
651 *Represent.* (2015).
652 39. Sener, O. & Koltun, V. Multi-task learning as multi-objective optimization. *Adv. Neural Inf.*
653 *Process. Syst.* **2018-Decem**, 527–538 (2018).
654 40. Shimura, K., Li, J. & Fukumoto, F. HFT-CNN: Learning Hierarchical Category Structure
655 for Multi-label Short Text Categorization. 811–816 (2019) doi:10.18653/v1/d18-1093.
656 41. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation
657 Hyperparameter Optimization Framework. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov.*
658 *Data Min.* 2623–2631 (2019).

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

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

44. 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).

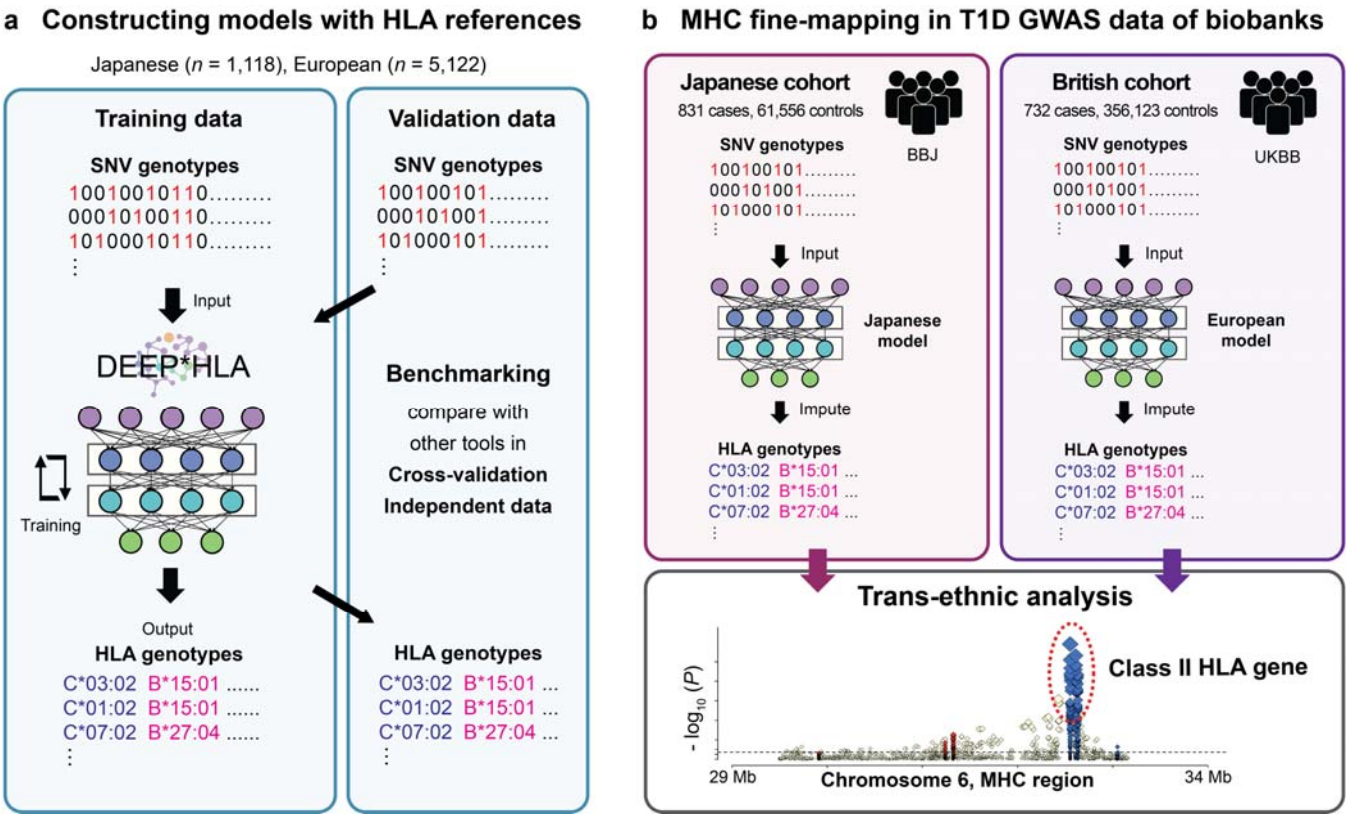
45. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

46. 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).

47. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

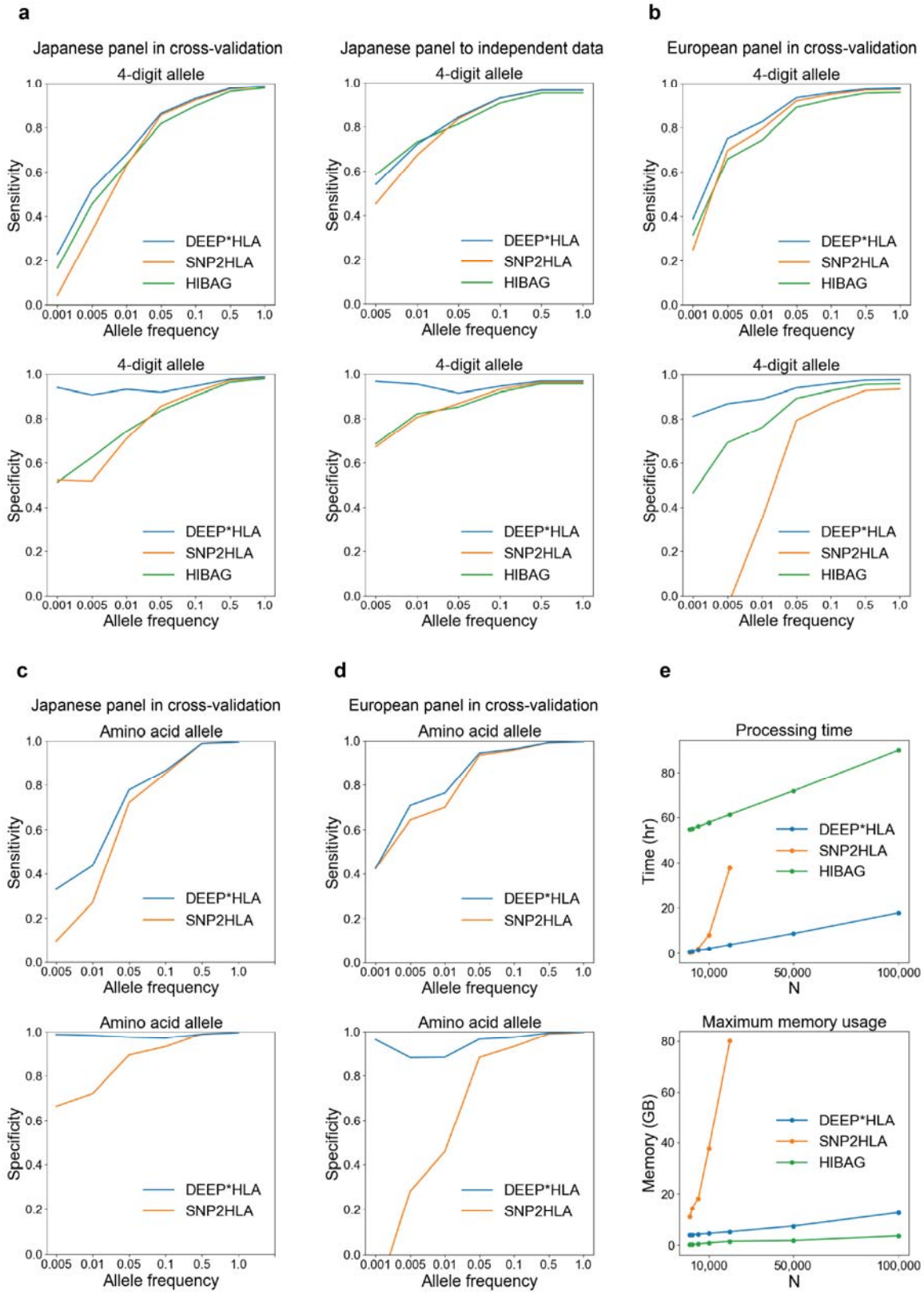
Figure Legends

Figure 1. An overview of the study



(**a**) Our method, DEEP*HLA, is a deep learning architecture that takes an input of genotypes of SNVs and outputs the genotype dosages of HLA genes. To train a model and benchmark its performance, we used Japanese and European HLA reference panels respectively, and evaluated its accuracies in cross-validation with compared to other tools. In the Japanese panel, we also evaluated its accuracy by applying the trained model to the independent Japanese HLA data. (**b**) We conducted trans-ethnic MHC fine-mapping in T1D GWAS data of BBJ and UKBB. We performed HLA imputation for the Japanese cohort from BBJ and the British cohort from UKBB using the models specific for individual populations, respectively. We integrated the individual results of imputed genotypes and performed trans-ethnic association analysis.

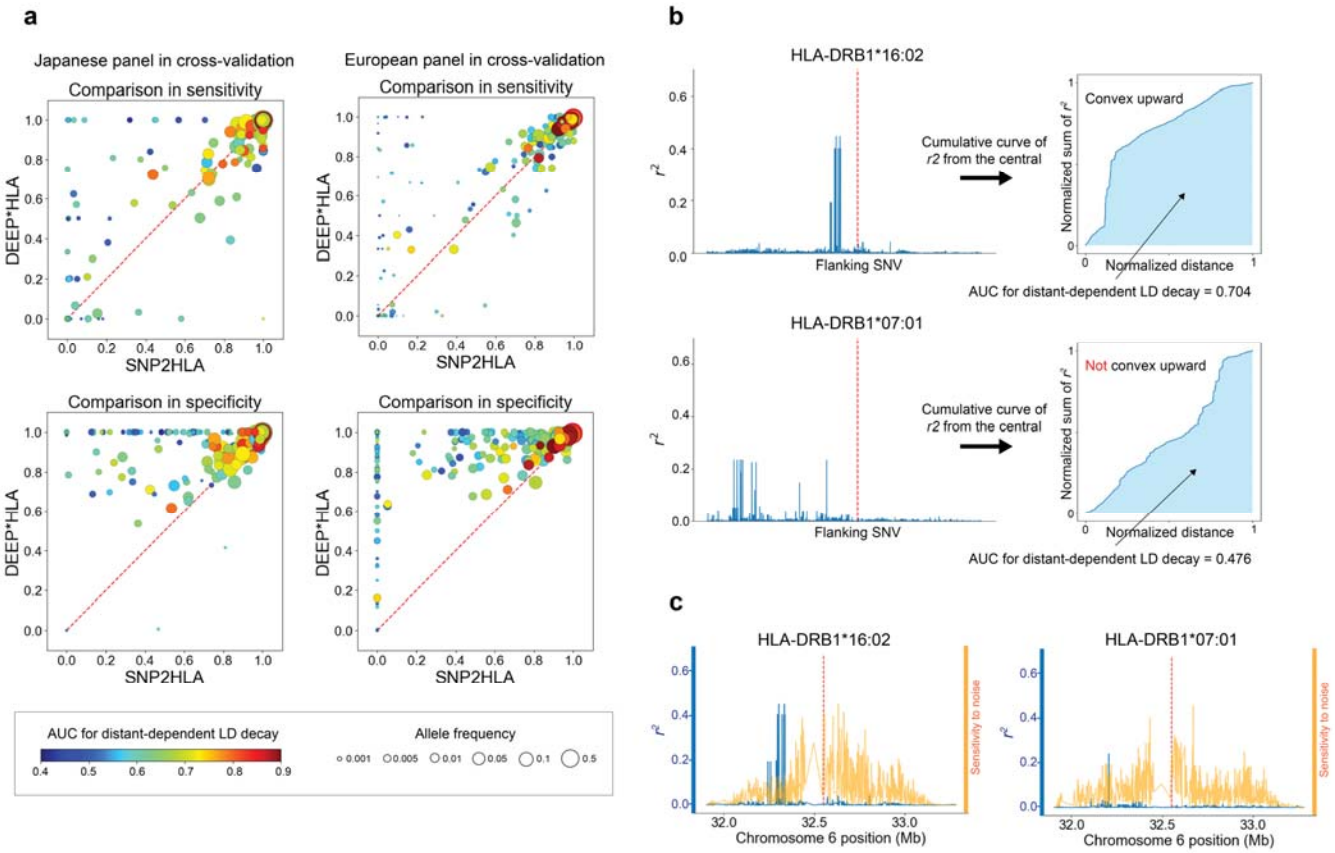
684 **Figure 2. Performance evaluations of HLA imputation methods**



685

(**a-d**) Sensitivity (upper) and specificity (lower) for the 4-digit alleles (**a, b**) and the amino acid polymorphisms (**c, d**) evaluated in our Japanese reference panel (**a, c**) and T1DGC reference panel (**b, d**). For each metrics, those for alleles of which frequency is less than a value on the horizontal axis are shown on the vertical axis. As a whole, DEEP*HLA outperformed other methods especially in specificity and imputing infrequent alleles. (**e**) Processing time (upper) and maximum memory usage (lower) evaluated on imputing the BBJ samples using the Japanese panel. DEEP*HLA imputed by far the fastest in total processing time as the sample size increased. All methods exhibited maximum memory usage scaling roughly linearly with sample size**.** SNP2HLA did not work within 100 GB in our machine for the sample size of more than 20,000.

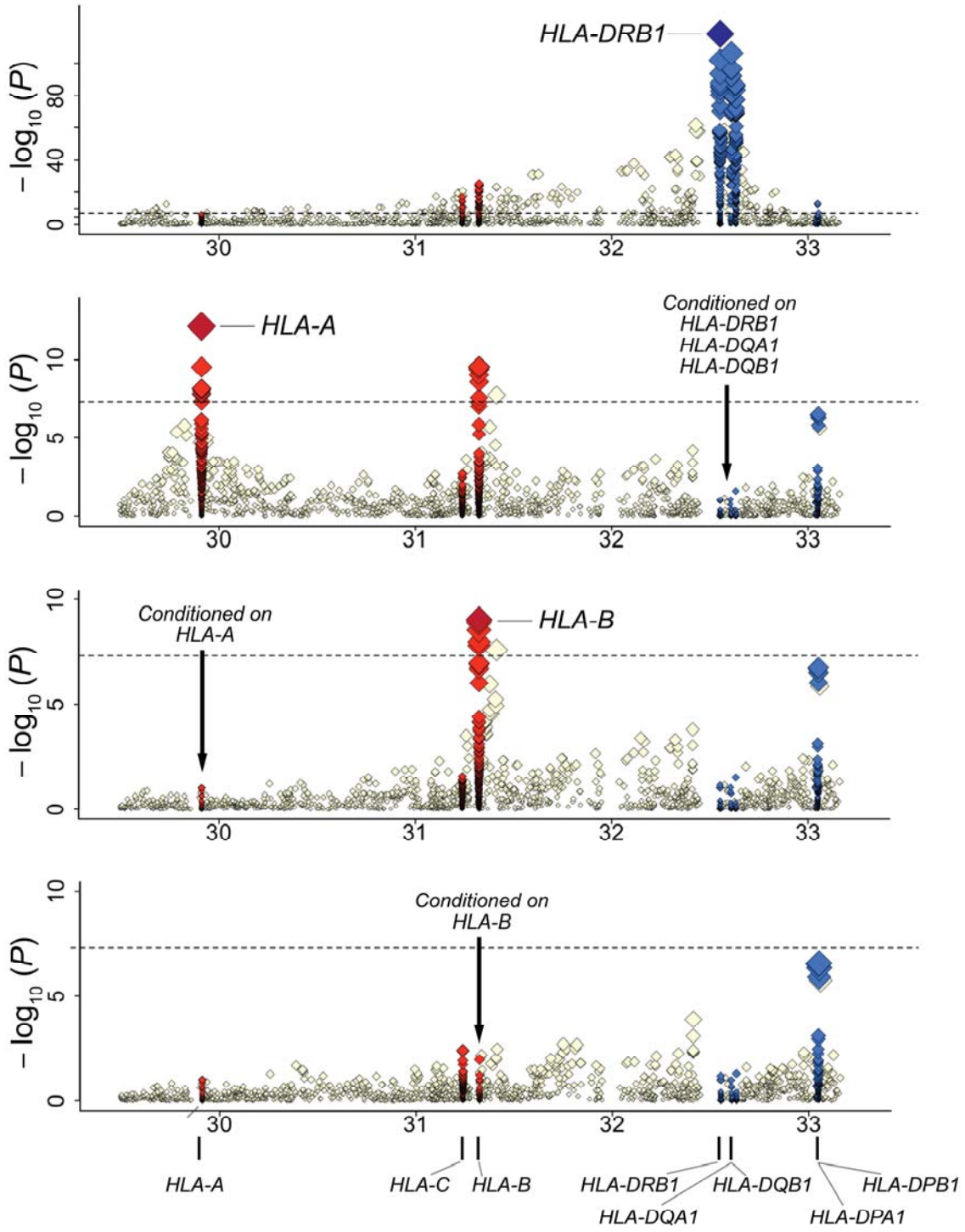
Figure 3. Comparison between DEEP*HLA and SNP2HLA displayed with allele frequencies and AUC for distance-dependent LD decay



(**a**) Comparisons of imputation accuracy between DEEP*HLA and SNP2HLA in sensitivity (upper) and specificity (lower) for 4-digit allele imputation for cross-validation on the Japanese panel (left) and T1DGC panels (right). Each dot corresponds to one allele, displayed with allele frequencies (size) and AUC for distance-dependent LD decay (color). Those of which specificities were less than 0 are shown with converted to 0 for visibility. Performance of SNP2HLA was limited when imputing the alleles with low frequency and low AUC, DEEP*HLA was relatively accurate even for the less frequent alleles regardless of the AUC. (**b**) Example illustrations of AUC for distance-dependent LD decay. The left figures illustrate r^2 of LD between an HLA allele (red dash line in the central) and flanking SNVs. HLA-DRB1*16:02 has strong LD

in close positions and weaker in the distance; and cumulative curve of r^2 of bilateral SNVs becomes convex upward; and the AUC becomes bigger. In contrast, HLA-DRB1*07:01 has moderate LD in distant or sparse positions; and the curve does not become convex upward; and the AUC becomes smaller. (**c**) Comparison between r^2 (blue line) and sensitivity maps of DEEP*HLA (orange line) for example alleles (red dash line in the central). The sensitivities are normalized for visibility. In both examples, DEEP*HLA reacted to noises across an extensive area regardless of LD.

715 **Figure 4. Trans-ethnic association plots of the HLA variants with T1D in the MHC region.**



716

717

718 Diamonds represent $-\log_{10}$ (*P* values) for the tested HLA variants, including SNPs, classical
719 alleles and amino acid polymorphisms of the HLA genes. The dashed black horizontal lines
720 represent the genome-wide significance threshold of $P = 5.0\times10^{-8}$. The physical positions of the
721 HLA genes on chromosome 6 are shown at the bottom. (**a–e**) Each panel shows the
722 association plot in the process of stepwise conditional regression analysis: nominal results. (**a**)
723 Results conditioned on *HLA-DRB1*, *HLA-DQA1*, and *HLA-DRB1*. (**b**) Results conditioned on
724 *HLA-DRB1*, *HLA-DQA1*, *HLA-DRB1*, and *HLA-A*. (**c**) Results conditioned on *HLA-DRB1*,
725 *HLA-DQA1*, *HLA-DRB1*, *HLA-A*, and *HLA-B.* (**d**) Our study identified independent contribution
726 of multiple HLA class I and class II genes to T1D risk in a trans-ethnic cohort, of which the
727 impacts of class II HLA genes was more evident. Detailed association results are available in
728 **Supplementary Table 4**.

729 **Tables 1. Associations of the HLA variants with T1D risk identified through the**

730 **trans-ethnic fine-mapping study.**

	Frequency (BBJ)		Frequency (UKBB)					
	Case	Control	Case	Control	OR (95% CI)		P†	
HLA variant	n = 831	n = 61,556	n = 732	n = 356,123	BBJ	UKBB	BBJ	UKBB
HLA-DRβ1 amino acid position 71								
Alanine	0.10	0.18	0.04	0.15	0.85 (0.66-1.10)	1.34 (0.89-1.99)	0.23	0.16
Arginine	0.82	0.73	0.33	0.45	(reference)			
Glutamic acid	0.073	0.074	0.083	0.12	1.26 (0.89-1.77)	0.72 (0.56-0.93)	0.019	0.0013
Lysine	0.0096	0.011	0.54	0.28	1.31 (0.71-2.24)	2.09 (1.75-2.50)	0.035	4.2×10^{-16}
HLA-DQβ1 amino acid position 185								
Isoleucine	0.39	0.57	0.68	0.83	2.74 (2.21-3.40)	4.12 (3.45-4.93)	3.5×10^{-20}	3.8×10^{-54}
Threonine	0.61	0.43	0.32	0.17	(reference)			
HLA-DQβ1 amino acid position 30								
Histidine	0.16	0.19	0.18	0.23	1.36 (0.97-1.93)	4.13 (2.86-5.95)	0.0078	3.2×10^{-14}
Serine	0.0042	0.0038	0.34	0.25	inf	3.78 (2.51-5.81)	0.079	5.3×10^{-10}
Tyrosine	0.83	0.80	0.48	0.52	(reference)			
HLA-DRβ1 amino acid position 74								
Alanine	0.56	0.59	0.59	0.65	(reference)			
Arginine	0.0018	0.00088	0.28	0.15	0 (0-0.045)	0.65 (0.42-0.97)	0.08	0.0039
Glutamic acid	0.32	0.27	0.021	0.036	0.77 (0.64-0.93)	0.57 (0.38-0.82)	0.00065	0.0004
Glutamine	0.0024	0.0030	0.0795	0.15	0 (0-0.0029)	0.31 (0.21-0.45)	0.079	5.3×10^{-10}
Leucine	0.12	0.14	0.023	0.023	0.97 (0.81-1.16)	2.19 (0.84-4.84)	0.074	0.0079
HLA-DQβ1 amino acid position 70								
Arginine	0.60	0.62	0.79	0.63	(reference)			
Glutamic acid	0.26	0.17	0.020	0.020	0.73 (0.59-0.9)	0.27 (0.11-0.72)	0.00020	0.00057
Glycine	0.14	0.20	0.19	0.35	0.95 (0.72-1.25)	0.50 (0.36-0.69)	0.073	2.9×10^{-5}
HLA-A amino acid position 62								
Arginine	0.19	0.20	0.06	0.09	1.25 (1.05-1.49)	0.93 (0.74-1.15)	0.0012	0.052
Glutamic acid	0.39	0.37	0.09	0.09	1.40 (1.21-1.63)	1.33 (1.10-1.59)	9.2×10^{-6}	0.0003
Glutamine	0.15	0.19	0.46	0.49	(reference)			
Glycine	0.26	0.24	0.33	0.29	1.44 (1.23-1.68)	1.27 (1.12-1.44)	6.6×10^{-6}	1.5×10^{-4}
Leucine	0	0	0.055	0.044	-	2.01 (1.57-2.55)	1.5×10^{-12}	1.8×10^{-8}
HLA-B*54:01	0.14	0.073	0	0	1.78 (1.51-2.08)	-	-	-

HLA, human leucocyte antigen; OR, odds ratio; 95% CI, 95% confidence interval.
†Obtained from the multivariate regression model that included all the variants listed here.

731