Human leukocyte antigen (HLA) genes are associated with a variety of diseases, yet the direct typing of HLA alleles is both time-consuming and costly. Consequently, various imputation methods leveraging sequential single nucleotide polymorphisms (SNPs) data have been proposed, employing either statistical or deep learning models, such as the convolutional neural network (CNN)-based model, DEEP*HLA. However, these methods exhibit limited imputation efficiency for infrequent alleles and necessitate a large size of reference dataset. In this context, we have developed a Transformer-based model to HLA allele imputation, named “HLA Reliable IMputation by Transformer (HLARIMNT)” designed to exploit the sequential nature of SNPs data. We evaluated HLARIMNT’s performance using two distinct reference panels; Pan-Asian reference panel \((n = 530)\) and Type 1 Diabetes genetics Consortium (T1DGC) reference panel \((n = 5225)\), alongside a combined panel \((n = 1060)\). HLARIMNT demonstrated superior accuracy to DEEP*HLA across several indices, particularly for infrequent alleles. Furthermore, we explored the impact of varying training data sizes on imputation accuracy, finding that HLARIMNT consistently outperformed across all data size. These findings suggest that Transformer-based models can efficiently impute not only HLA types but potentially other gene types from sequential SNPs data.
precise imputations than existing models. However, there was still room for improvement in DEEP*HLA, especially in accuracy for infrequent alleles, and its dependency on large size reference panels for effective imputation.

In this study, we introduce a Transformer-based model, named "HLA Reliable IMpuatioN by Transformer (HLARIMNT)", designed to impute eight classical HLA alleles. HLARIMNT leverages the Transformer’s capability to exploit the sequential nature of SNPs. Our findings indicate that HLARIMNT consistently achieves higher imputation accuracy compared to CNN-based DEEP*HLA, showcasing the potential of Transformer-based models in enhancing the precision of HLA allele imputation.

### MATERIALS AND METHODS

#### Datasets

We used Pan-Asian reference panel [29, 30] and Type 1 Diabetes Genetics Consortium (T1DGC) reference panel [31]. The Pan-Asian panel was genotyped with the Illumina HumanCoreExome BeadChip and T1DGC panel was genotyped using Illumina ImmunoChip, and both contain 4-digit resolution typing data of eight classical HLA genes based on SSO method; HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1, and HLA-DPB1. Both were publicly distributed with Beagle format in a phased condition. Pan-Asian reference panel is downloadable with SNP2HLA software (https://software.broadinstitute.org/mpg/snp2hla/). T1DGC panel is available from the NIDDK Central Repository after the registration process (https://repository.niddk.nih.gov/studies/t1dgc-special/). Pan-Asian reference panel contains 530 unrelated individuals, i.e., 1060 haplotypes, of Asian ancestry. T1DGC reference panel contains 5225 unrelated individuals, i.e., 10450 haplotypes, of European ancestry. The mixed panel was generated by combining all 530 individuals from Pan-Asian panel and randomly selected 530 individuals from T1DGC panel, resulting in a total of 1060 individuals.

Number of all and infrequent (<0.01) alleles are shown in Table 1. Since all data used in this study have already been published and are publicly available, the ethics committee of our institute have waived an ethical review.

#### Model architectures

We adopted a basic structure of the model from DEEP*HLA and modified the CNN part of into Transformer-based architecture (the blue dotted line in Fig. 1, i.e., Embedding Layer and Transformer Layer).

HLA genes are divided into groups according to LD structure and physical distance as in DEEP*HLA: (1)HLA-A, (2)HLA-C, HLA-B, (3)HLA-DRB1, HLA-DQA1, HLA-DQB1 and (4)HLA-DPA1, HLA-DPB1, and genes in each group are imputed simultaneously. Therefore, the hyperparameters of the models are the same except for the number of outputs, but the weights trained are different for each group. For each group, the model takes the input of each haplotype SNP genotype from pre-phased data, which are expressed by two-dimensional vectors. The SNPs are expressed by 01 or 10 based on whether each base is consistent with a reference or alternative one. The range of SNPs used for training in each group is the same as DEEP*HLA; 500kbps each.

Embedding Layer (Fig. 1b) consolidates the neighboring SNPs together to divide them into 50 chunks, and adds a classification (cls) token at the head of SNPs to learn the features. The cls token has the same shape as a single chunk, and its elements are all zeros. Then a common linear layer projects SNPs to the size of (51, 512).

Transformer Layer (Fig. 1c) applies positional encoding and the encoder portion of Transformer to the data, after which the feature vector of cls token is extracted.

Classification Layer (Fig. 1d), which is prepared for each HLA gene, is a combination of a linear layer and a softmax layer. Output dimension of the layer is the same as the number of alleles the HLA gene has. Output values are imputation dosages for the alleles, each of which takes a value from 0.0 to 1.0 that should be 1.0 when summed up.

To search hyperparameters, we utilized Optuna (https://github.com/optuna). The searched parameters and the values are as follows; Number of Transformer heads: 64, Number of layer of Transformer encoder: 2, Dimension of feedforward in Transformer encoder: 64, Batch size: 64, Dimension of embedding in Embedding Layer: 512, Learning rate: 0.0005, and Number of epochs for early stopping: 50.

#### Training

In the training, as with DEEP*HLA, we adopted hierarchical fine-tuning, in which the parameters for classifying 2-digit alleles were transferred to the model for 4-digit alleles. This allowed the model to take advantage of the hierarchical nature of HLA alleles. We only used SNP data for training and evaluation, and removed other information from the reference panel. We used Cross Entropy Loss as a loss function and Adam to optimize the loss. The Cross Entropy Loss is expressed by the following equation, whether it indicates correct output that should be 0 or 1, \( y_i \) indicates the output of softmax layer, and \( n \) indicates the number of classes.

\[
\text{Loss} = - \sum_{i=1}^{n} t_i \log y_i
\]

The data was first split into test data and other data, then 90% of the other data was used as training data and 10% as validation data. The validation data was used for early stopping and model updates during the training.

First, a variant ‘count’, which is an indicator used for early stopping, is initialized to 0. At the beginning of the epoch, the training data is batched and trained using back propagation. At the end of one epoch, the model’s correctness rate (the percentage of data that the model could impute correctly) in

| HLA     | Pan Asian         | T1DGC         | Mixed panel |
|---------|-------------------|---------------|-------------|
|         | All alleles | Infrequent alleles | All alleles | Infrequent alleles | All alleles | Infrequent alleles |
| A       | 26        | 7              | 50          | 15          | 39        | 22              |
| C       | 24        | 9              | 33          | 15          | 32        | 14              |
| B       | 48        | 20             | 97          | 74          | 72        | 44              |
| DRB1    | 32        | 11             | 51          | 15          | 48        | 28              |
| DQA1    | 8         | 0              | 8           | 1           | 8         | 0               |
| DQB1    | 15        | 1              | 18          | 5           | 18        | 3               |
| DPA1    | 5         | 1              | 7           | 4           | 7         | 3               |
| DPB1    | 21        | 6              | 34          | 20          | 27        | 16              |

Table 1. Number of all and infrequent (<0.01) alleles
the validation data is calculated. If this value is greater than the value of the 'best model', the weights are overwritten and saved as the 'best model'. If not, add 1 to the 'count', and when the value of the 'count' reaches specified number, the training is terminated. Also, at regular intervals, the learning rate is decreased. This process is repeated until the number of epochs exceed specified number, at which point training is terminated. The 'best model' stored at the end of training is used to examine the accuracy for the test data. For each gene, the average imputation accuracy was obtained by weighted average based on allele frequencies.

All experiments were conducted on a computer equipped CPU: AMD EPYC 7252 8-Core; memory: 125GB; and GPU: NVIDIA A5000 x1.

Evaluation
In the experiments, we used 4 indices for the evaluations; $r^2$, PPV, sensitivity, and probability, values of which were calculated for each allele.

$r^2(A)$ represents the square of Pearson's product moment correlation coefficient between imputed and typed dosages and is expressed as follows;

$$r^2(A) = \frac{\sum_{i=1}^{2n} x_i(A) y_i(A) - (\sum_{i=1}^{2n} x_i(A) (\sum_{i=1}^{2n} y_i(A))}{\left(\sum_{i=1}^{2n} x_i^2(A) - \frac{\sum_{i=1}^{2n} x_i^2(A)}{2n}\right)\left(\sum_{i=1}^{2n} y_i^2(A) - \frac{\sum_{i=1}^{2n} y_i^2(A)}{2n}\right)}$$

where $n$ is the number of individuals (i.e., $2n$ is the number of haplotypes), $A$ is the type of allele (e.g., HLA-A*01:23), $x_i(A)$ is the imputed dosage of allele A for haplotype $i$, which is obtained from the output of the softmax layer, and $y_i(A)$ is the typed dosage of allele A for haplotype $i$, taking values 1 if haplotype $i$ has allele $A$ and otherwise 0.

Sensitivity($A$) represents the percentage of haplotypes that were correctly imputed to have allele $A$ in all the haplotypes that have allele $A$, and PPV($A$) represents the percentage of haplotypes with allele $A$ in all the haplotypes predicted to have allele $A$ (Naito 2021).

Probability($A$) represents the imputed dosage of allele $A$ for each haplotype with allele $A$, and is expressed as follows;

$$\text{probability}(A) = \frac{\sum_{i=1}^{m} x_i(A)}{m}$$

where $m$ represents the number of haplotypes with allele $A$.

Relevant codes are available at https://github.com/seitalab/HLARIMNT.

RESULTS
We first conducted a comparative analysis of imputation accuracy between CNN-based DEEP*HLA and Transformer-based HLARIMNT against three datasets: the Pan-Asian reference panel, the T1DGC reference panel, and the mixed panel derived from both.
Fig. 2  Average accuracy for 4-digit alleles in each of the HLA genes.  a Pan Asian reference panel, b T1DGC reference panel, and c Mixed reference panel. HLARIMNT outperformed DEEP*HLA on most genes and indices for all data datasets, though there were some exceptions depending on the combination of indices and genes.
After the training of the models, imputation was executed against the test dataset for 4-digit alleles for eight HLA genes. Accuracy for each allele was assessed using four indices; $r^2$, PPV, sensitivity, and probability. This training-test session was repeated 5 times individually, then calculated the simple additive average of each index to compare the two methods.

Imputation accuracy for each HLA gene is shown in Fig. 2. Generally, HLARIMNT demonstrated superior performance for nearly all genes across the three reference panels, with certain exceptions based on the specific gene-reference panel-index combinations.

Table 2 is the summary of 5 experiments. HLARIMNT demonstrated superior accuracy across all indices for all 4-digit alleles across the reference panels, albeit DEEP*HLA exhibited equal Probability in the T1DGC reference panel and slightly higher Probability in the Mixed reference panel. Imputation accuracy for infrequent 4-digit alleles (with frequencies less than 0.01) was also tested using T1DGC panel (Table 3). The size of test data of the other panels were too small to test infrequent alleles. Table 3 highlights HLARIMNT's enhanced performance for these alleles compared to its performance across all alleles.

The influence of allele frequency on imputation accuracy was further examined using the T1DGC panel (Fig. 3). This analysis displayed the accuracy for 4-digit alleles with frequencies at or above the threshold indicated on the vertical axis, plotted against the horizontal axis (noting that <0.005 represents a frequency of less than 0.005). Regardless of the allele frequencies, HLARIMNT achieved comparable or superior accuracy to DEEP*HLA.

Subsequently, we explored the effect of training data size using the T1DGC reference panel (Fig. 4). The training dataset comprised various subsets of individuals, including sizes of 530, 1300, 2600, and 4180, randomly selected from the T1DGC reference panel. The test dataset remained constant and independent across all training data sizes. Figure 4b shows the average accuracy for all 4-digit alleles across different training data sizes. HLARIMNT consistently achieved equal or higher accuracy than DEEP*HLA across all sizes of training data. Moreover, the advantage of HLARIMNT over DEEP*HLA was more pronounced with smaller training data sizes across all indices. For infrequent alleles (with frequencies less than 0.01), HLARIMNT exhibited superior performance across all sizes of training data, as shown in Fig. 4c. Similar

Table 2. Average imputation accuracy for all alleles ($n = 5$)

| Dataset   | Indices | DEEP*HLA | HLARIMNT |
|-----------|---------|----------|----------|
|           | Mean    | S.D.     | Mean     | S.D.    |
| Pan-Asian | $r^2$   | 0.884    | 0.024    | 0.922   | 0.015   |
|           | PPV     | 0.894    | 0.023    | 0.917   | 0.020   |
|           | Sensitivity | 0.895  | 0.023    | 0.922   | 0.015   |
|           | Probability | 0.893  | 0.023    | 0.899   | 0.021   |
| T1DGC     | $r^2$   | 0.970    | 0.004    | 0.976   | 0.003   |
|           | PPV     | 0.974    | 0.003    | 0.976   | 0.003   |
|           | Sensitivity | 0.974  | 0.003    | 0.977   | 0.003   |
|           | Probability | 0.974  | 0.004    | 0.974   | 0.003   |
| Mixed     | $r^2$   | 0.919    | 0.004    | 0.935   | 0.003   |
|           | PPV     | 0.926    | 0.003    | 0.933   | 0.003   |
|           | Sensitivity | 0.929  | 0.003    | 0.934   | 0.003   |
|           | Probability | 0.928  | 0.004    | 0.918   | 0.003   |

The best result of each index is highlighted in bold

Table 3. Average imputation accuracy for infrequent (<0.01) alleles ($n = 5$)

| Dataset | Indices | DEEP*HLA | HLARIMNT |
|---------|---------|----------|----------|
|         | Mean    | S.D.     | Mean     | S.D.    |
| T1DGC   | $r^2$   | 0.860    | 0.030    | 0.880   | 0.022   |
|         | PPV     | 0.893    | 0.047    | 0.900   | 0.028   |
|         | Sensitivity | 0.846  | 0.043    | 0.860   | 0.036   |
|         | Probability | 0.846  | 0.040    | 0.853   | 0.037   |

The best result of each index is highlighted in bold

Fig. 3 Average values of the 4 indices calculated for each allele frequency. At almost all of the frequencies, the accuracy of HLARIMNT was equal to or better than DEEP*HLA. Furthermore, the advantage of HLARIMNT was more noticeable at lower frequencies.
Fig. 4  a Preparation of small size training data. b The average accuracy for 4-digit alleles using various sizes of training data and (c) shows that of alleles which have frequencies less than 0.01. The accuracy of HLARIMNT was generally higher than that of DEEP*HLA on any training data sizes.
to previous findings, HLARIMNT’s advantage was more significant with smaller training datasets. This trend suggests that HLARIMNT’s effectiveness is particularly notable when the available training data is limited, underscoring its efficiency in imputing HLA alleles, especially those that are infrequent.

**DISCUSSION**

A significant observation from our study is that HLARIMNT demonstrated superior performance in a mixed panel combining two ethnic groups, suggesting its capability to capture features of the SNPs without overfitting to a specific ethnic background. This attribute is particularly beneficial for HLA imputation, which typically requires large reference panels for accuracy. In fact, both methods in this paper were more accurate with a larger number of training data. On the other hand, it is known that the distribution and frequency of the HLA alleles are variable across different ethnic groups [32], which results in heterogeneity in HLA risk among populations and data volumes, highlighting its potential utility in processing sequencing data. However, to create large reference panels, it is necessary to analyze SNPs of many individuals at high density, which is very expensive. In this regard, if we can use the mixture of reference panels including various ethnic groups for training, it will be possible to perform an accurate HLA imputation without performing new SNP sequencing to make the reference panel larger. HLARIMNT has the potential to meet this requirement.

Our results further underscore HLARIMNT’s enhanced accuracy with smaller datasets, irrespective of allele frequencies. Although haplotypes vary by ethnicity—necessitating ethnically specific reference panels—the financial implications of creating new panels for each ethnicity are considerable. HLARIMNT’s efficacy in accurately capturing characteristics with less data offers a potential solution to this problem, positioning it as a viable deep learning method for HLA imputation. This study aligns with recent research highlighting the Transformer model’s utility in processing sequential data, suggesting its superiority over traditional methods like RNNs and CNNs in genomic applications. This finding invites further exploration of Transformer models within genomics.

However, several aspects warrant discussion, including the limited size of the Pan-Asian reference panel and the potential unreliability of accuracy for alleles with very low frequencies. A detailed analysis of these errors could illuminate the model’s allele discrimination process and guide improvements in imputation accuracy. Further validation of HLARIMNT’s utility could involve testing with data from different ethnicities or examining accuracy variations among SNP sequences. Such an approach would necessitate careful selection of closely related ethnic groups for training and testing, given the significant allele frequency differences across ethnicities.

The study demonstrates HLARIMNT’s superior accuracy across all reference panels, notably for infrequent alleles, and its consistent performance across various training data sizes. These findings suggest HLARIMNT’s capacity for efficient HLA imputation across diverse populations and data volumes, highlighting its potential as a practical tool in the field.

**REFERENCES**

1. Dendrou CA, Petersen J, Rossjohn J, Fuegger L. HLA variation and disease. Nat Rev Immunol. 2018;18:325–39. https://doi.org/10.1038/nri.2017.143
2. Fan WL, Shiao MS, Hui RC, Su SC, Wang CW, Chang YC, et al. HLA association with drug-induced adverse reactions. J Immunol Res. 2017;2017:3186328. https://doi.org/10.1155/2017/3186328.
3. Ko TM, Tsai CY, Chen SY, Chen KS, Yu KH, Chu CS, et al. Use of HLA-B*58:01 genotyping to prevent allaporin-induced severe cutaneous adverse reactions in Taiwan: National prospective cohort study. BMJ. 2015;351. https://doi.org/10.1136/bmj.h4848.
4. Hirata J, Hosomichi K, Sakase K, Kanai M, Nakaoka H, Ishigaki K, et al. Genetic and phenotypic landscape of the major histocompatibility complex region in the Japanese population. Nat Genet. 2019;51:470–80. https://doi.org/10.1038/s41588-018-0336-0.
5. Erlich H. HLA DNA typing: past, present, and future. Tissue Antigens. 2012;80:1–11. https://doi.org/10.1111/j.1399-0039.2012.01881.x
6. Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PW, Walker BD, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science. 2010;330:1551–7. https://doi.org/10.1126/science.1195371.
7. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet. 2012;44:291–6. https://doi.org/10.1038/ng.1076.
8. Okada Y, Momozawa Y, Ashikawa K, Kanai M, Matsuda K, Katamany T, et al. Construction of a population-specific HLA imputation reference panel and its application to Graves’ disease risk in Japanese. Nat Genet. 2015;47:798–802. https://doi.org/10.1038/ng.3310.
9. Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet. 2008;82:48–56. https://doi.org/10.1016/j.ajhg.2007.09.001.
10. Dillthey AT, Moutsianas L, Leslie S, McVean GHL. A*IMP: an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinformatics. 2011;27:968–72. https://doi.org/10.1093/bioinformatics/btr061.
11. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165:2213–33. https://doi.org/10.1093/genetics/165.4.2213. Erratum in: Genetics. 2004 Jun;167:2:1039.
12. Dillthey A, Leslie S, Moutsianas L, Shen J, Cox C, Nelson MR, et al. Multi-population classical HLA type imputation. PLoS Comput Biol. 2013;9:e1002877 https://doi.org/10.1371/journal.pcbi.1002877.
13. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, et al. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS ONE. 2013;8:e64683 https://doi.org/10.1371/journal.pone.0064683.
14. Zheng X, Shen J, Cox C, Wakefeld JC, Ehm MG, Nelson MR, et al. HIBAG - HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014;14:192–200. https://doi.org/10.1007/s10958-013-0734-0.
15. Cook S, Choi W, Lim H-J, Luo Y, Kim K, Jia X, et al. Accurate imputation of human leukocyte antigens with COOKHLA. Nat Commun. 2021;12 https://doi.org/10.1038/s41467-021-21541-5.
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need [Internet]. 2017. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/1706.03762.
17. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners [Internet]. 2020. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/2005.14165.
18. Adwardiana D, Luong M-T, So DR, Hall J, Fiedel N, Thoppiplan R, et al. Towards a human-like open-domain chatbot [Internet]. 2020. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/2001.09977.
19. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale [Internet]. 2020. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/2010.11929.
20. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.
21. Huang C-ZA, Vaswani A, Uszkoreit J, Shazeer N, Simon I, Hawthorne C, et al. Music transformer [Internet]. 2018. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/1809.04281
22. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, et al. Photorealistic text-to-image diffusion models with deep language understanding [Internet]. 2022. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/2111.07048.
23. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with CLIP latents [Internet]. 2022. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/2204.06125.
24. Egger I, Gxanzer C, Pepe A, Pomykala KL, Jonske F, Kurz M, et al. Medical deep learning—a systematic meta-review. Comput Methods Prog Biomed. 2022;221:106874. https://doi.org/10.1016/j.cmpb.2022.106874.
25. Wang P, Liao R, Moyer D, Berkowitz S, Honig S, Golland P Image classification with consistent supporting evidence [Internet]. 2021. [cited 2024 Jan 27]. Available from: http://arxiv.org/abs/2111.07048.
26. Chu Y, Zhang Y, Wang Q, Zhang L, Wang X, Wang Y, et al. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. Nat Mach Intell. 2022;4:300–11. https://doi.org/10.1038/s42256-022-00459-7

27. Kojima K, Tadaka S, Katsuoka F, Tamiya G, Yamamoto M, Kinoshita K. A genotype imputation method for de-identified haplotype reference information by using recurrent neural network. PLoS Comput Biol. 2020;16:e1008207 https://doi.org/10.1371/journal.pcbi.1008207.

28. Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. Nat Commun. 2021;12:1639 https://doi.org/10.1038/s41467-021-21975-x.

29. Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. Nat Commun. 2021;12:1639 https://doi.org/10.1038/s41467-021-21975-x.

29. Pillai NE, Okada Y, Saw WY, Ong RTH, Wang X, Tantoso E, et al. Predicting HLA alleles from high-resolution SNP data in three southeast Asian populations. Hum Mol Genet. 2014;23:4443–51. https://doi.org/10.1093/hmg/ddu149.

30. Okada Y, Kim K, Han B, Pillai NE, Ong RTH, Saw WY, et al. Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. Hum Mol Genet. 2014;23:6916–26. https://doi.org/10.1093/hmg/ddu387.

31. Rich SS, Concannon P, Erlich H, Julier C, Morahan G, Nerup J, et al. The type 1 diabetes genetics consortium. Ann NY Acad Sci. 2006;1079 https://doi.org/10.1196/annals.1375.001.

32. Gourraud PA, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA diversity in the 1000 genomes dataset. PLoS ONE. 2014;9. https://doi.org/10.1371/journal.pone.0097282.

33. Todd JA, Beir JI. HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature. 1987;329:599–604. https://doi.org/10.1038/329999a0

34. Hu X, Deutsch AJ, Lenz TL, Onengut-Gumuscu S, Han B, Chen WM, 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. 2015;47:898–905. https://doi.org/10.1038/ng.3353.

35. Kawabata Y, Ikegami H, Awata T, Imagawa A, Maruyama T, Kawasaki E, et al. Differential association of HLA with three subtypes of type 1 diabetes: Fulminant, slowly progressive and acute-onset. Diabetologia. 2009;52:2513–21. https://doi.org/10.1007/s00125-009-1539-9.

ACKNOWLEDGEMENTS

We would like to thank all the participants involved in this study. We thank the Mathematics-Based Creation of Science Program at the Kyoto University Graduate School of Science and RIKEN Interdisciplinary Theoretical and Mathematical Sciences Program for their support of this study.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Jun Seita.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024