nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0336-0

Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population

Jun Hirata[1,2], Kazuyoshi Hosomichi[3], Saori Sakaue[1,4,5], Masahiro Kanai[1,4,6], Hirofumi Nakaoka[7], Kazuyoshi Ishigaki[4], Ken Suzuki[1,4,8], Masato Akiyama[4,9], Toshihiro Kishikawa[1,10], Kotaro Ogawa[1,11], Tatsuo Masuda[1,12], Kenichi Yamamoto[1,13], Makoto Hirata[14], Koichi Matsuda[15], Yukihide Momozawa[16], Ituro Inoue[7], Michiaki Kubo[17], Yoichiro Kamatani[4,18] and Yukinori Okada[1,4,19]*

To perform detailed fine-mapping of the major-histocompatibility-complex region, we conducted next-generation sequencing (NGS)-based typing of the 33 human leukocyte antigen (HLA) genes in 1,120 individuals of Japanese ancestry, providing a high-resolution allele catalog and linkage-disequilibrium structure of both classical and nonclassical HLA genes. Together with population-specific deep-whole-genome-sequencing data ($n = 1{,}276$), we conducted NGS-based HLA, single-nucleotide-variant and indel imputation of large-scale genome-wide-association-study data from 166,190 Japanese individuals. A phenome-wide association study assessing 106 clinical phenotypes identified abundant, significant genotype–phenotype associations across 52 phenotypes. Fine-mapping highlighted multiple association patterns conferring independent risks from classical HLA genes. Region-wide heritability estimates and genetic-correlation network analysis elucidated the polygenic architecture shared across the phenotypes.

Genetic variants of the major histocompatibilty complex (MHC) region at 6p21.3 confer the largest number of associations that explain substantial phenotypic variations of a wide range of complex human diseases and quantitative traits[1]. The MHC region is one of the most polymorphic sites in the human genome and is characterized by population-specific complex linkage disequilibrium (LD) structure and long-range haplotypes[2–5]. Among the >200 genes densely contained in the MHC region[6,7], human leukocyte antigen (HLA) genes are considered to explain most of the genetic risk of MHC. Fine-mapping efforts to identity causal variants within the MHC region reported many HLA alleles and amino acid polymorphisms associated with complex human traits[8]. In particular, development of the HLA imputation method and construction of population-specific reference panels have successfully accelerated the identification of causal variants that should be useful for personalized medicine[9–12].

However, several points have yet to be implemented in genetic and phenotypic studies of MHC. The first point is the use of NGS for fine-mapping MHC risk. Compared with traditional HLA typing methods, such as sequence-specific oligonucleotide hybridization (SSO) and sequencing-based typing, HLA typing by NGS could provide higher resolution of alleles for a wider spectrum of HLA and HLA-related genes beyond a limited number of classical HLA genes[13–16]. Population-specific whole-genome sequencing (WGS) data contribute to imputing functional rare variants with high accuracy[17]. Given that variants of the nonclassical HLA genes are responsible for disease risk, as well as those of the classical HLA genes, and that functional variants of non-HLA genes within the MHC region affect clinical phenotypes[18,19], MHC risk analyses using the NGS-based reference panel are warranted to achieve more accurate fine-mapping of the causal variants.

The second point is the application of the HLA imputation method to large-scale genome-wide association study (GWAS) data that represent all the participants of population-level cohorts. Many nation-wide biobanks have recently been launched to capture the genetic and phenotypic variation of these populations. To date, large-scale GWAS data from >100,000 samples have been publicly released from several biobanks (for example, >500,000 from UK

[1]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. [2]Pharmaceutical Discovery Research Laboratories, Teijin Pharma Limited, Hino, Japan. [3]Department of Bioinformatics and Genomics, Graduate School of Advanced Preventive Medical Sciences, Kanazawa University, Ishikawa, Japan. [4]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [5]Department of Allergy and Rheumatology, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan. [6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [7]Division of Human Genetics, National Institute of Genetics, Shizuoka, Japan. [8]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [9]Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Fukuoka, Japan. [10]Department of Otorhinolaryngology, Head and Neck Surgery, Osaka University Graduate School of Medicine, Osaka, Japan. [11]Department of Neurology, Osaka University Graduate School of Medicine, Osaka, Japan. [12]Department of Obstetrics and Gynecology, Osaka University Graduate School of Medicine, Osaka, Japan. [13]Department of Pediatrics, Osaka University Graduate School of Medicine, Osaka, Japan. [14]Laboratory of Genome Technology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [15]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [16]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [17]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [18]Kyoto-McGill International Collaborative School in Genomic Medicine, Sakyo-ku, Kyoto, Japan. [19]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan. *e-mail: yokada@sg.med.osaka-u.ac.jp

Biobank[17,20] and >170,000 from BioBank Japan Project (BBJ)[21,22]. Although HLA imputation of such big genotype data needs further tuning in the analytic pipeline, achievement of this task should enhance the knowledge of the genetic landscape of MHC in these populations.

The third point is a phenome-wide assessment of risk variants in the MHC region. Cross-phenotype analysis has identified shared genetic correlations among human traits, which are represented as pleiotropic associations of the variants and cross-phenotype network that are linked to disease biology[23–26]. Phenome-wide association studies (PheWASs) that use electronic medical records or medical information collected throughout a cohort have successfully identified clinically useful genotype–phenotype correlations[27,28]. MHC is one of the most pleiotropic sites in the genome[1], and thus application of the PheWAS approach should elucidate the phenotypic landscape of the MHC variants as well[29].

Here we report a comprehensive analysis that characterizes the genetic and phenotypic landscape of MHC in the Japanese population. We newly constructed an HLA imputation reference panel of Japanese individuals ($n = 1{,}120$) through high-resolution NGS typing of both classical and nonclassical HLA genes ($n = 33$). Together with accurate imputation of single-nucleotide variants (SNVs) and indels in a broad allele-frequency spectrum by using the population-specific deep-WGS reference data ($n = 1{,}276$)[30], HLA imputation of the 166,190 Japanese individuals from the BBJ genotype data was conducted to apply a PheWAS of 106 complex human diseases and quantitative traits extracted from clinical records.

Results

NGS typing of HLA genes in the Japanese population. For the 1,120 unrelated Japanese individuals, we conducted high-resolution typing of 33 HLA-related genes with up to six-digit-level allele information (study design in Supplementary Fig. 1). We adopted target-capture technique and sequencing with relatively longer read lengths (350 base pairs (bp) and 250 bp for paired-end, an average depth of 260.1×)[31,32]. By conducting validation with the traditional SSO method for some individuals ($n = 182$), we observed higher accuracy in classical HLA allele typing than that in previous NGS-based reports (<0.56% potentially inaccurate typing). NGS-based HLA typing was able to update allele information that was incorrectly assigned by traditional typing methods (for example, HLA-DRB1*14:01 by SSO was corrected as HLA-DRB1*14:54 by NGS[33]; details in Supplementary Table 1).

Among the 33 sequenced HLA genes, 9 are classical HLA genes (3 for class I and 6 for class II), and 24 are nonclassical HLA genes (Supplementary Table 2; HLA gene classification criteria in Methods). Whereas alleles of classical HLA genes were highly polymorphic (on average, there were 9.7, 20.1 and 21.6 alleles per gene for two-digit, four-digit and six-digit-level allele information, respectively), those of nonclassical HLA genes showed lower variations (1.4, 3.1 and 4.0 alleles per gene, respectively; Fig. 1a and Supplementary Tables 2 and 3). Of these, *HLA-B*, *HLA-DRB1* and *MICA* had the largest numbers of alleles for class I and II classical HLA genes and nonclassical HLA genes, respectively ($n = 39$, 33 and 15 in four-digit-level allele information). Because there was inconsistent definition of the registered sequences for one of the nonclassical HLA genes of *TAP2*, it was difficult to consistently define the four-digit (and also six-digit) alleles of *TAP2* (details in Supplementary Table 4). Although elucidation of six-digit allele distribution is one of the topics that was finally achieved by introduction of NGS, we found that increments of HLA allele variations from four to six digits (+1.4 and +0.9 for classical and nonclassical HLA alleles, respectively) were limited as compared with those from two to four digits (+10.4 and +1.7 alleles, respectively).

High-dimensional compression elucidates HLA-variant patterns. Systematic visualization of LD patterns among HLA genes contributes to the understanding of population-specific LD structure of genetic variants within MHC[4]. Thus, we introduced an entropy-based LD-measurement index (ε) to assess distributions of the four-digit HLA alleles and to quantify pairwise LD between the HLA genes. Within MHC, there exist four major LD blocks of the HLA genes ($\varepsilon > 0.15$): *HLA-G*, *HLA-H*, *HLA-K* and *HLA-A* for block 1; *HLA-C*, *HLA-B*, *MICA* and *MICB* for block 2; *HLA-DRA*, *HLA-DRB* family genes, *HLA-DQA1*, *HLA-DQB1* and *HLA-DOB* for block 3; and *HLA-DPA1* and *HLA-DPB1* for block 4 (Fig. 1b), thus demonstrating that classical and nonclassical HLA genes together constitute the LD patterns within MHC.

One challenge in HLA-polymorphism characterization in personalized regenerative medicine or organ transplantation is an optimized classification of the haplotypes based on HLA typing data[34]. Classifying haplotypes according to simple combinations of multiple HLA alleles and genes is likely to subdivide samples into clusters that are too segmented. Thus, we introduced a machine-learning-based clustering approach. We adopted *t*-distributed stochastic neighbor embedding (tSNE), a machine-learning method for high-dimensionality compression and visualization[35,36], to the HLA typing data. We then performed unsupervised clustering of the haplotypes by using tSNE components (tSNE_1 and tSNE_2) and the DBSCAN algorithm[37].

For classical HLA alleles, 3, 10 and 11 clusters were constructed for two-digit, four-digit and six-digit alleles, respectively (frequency > 0.01; Fig. 2a). Although haplotypes of higher- and lower-digit alleles were clustered separately, clusters of the higher-digit alleles were subsets of those of the lower-digit alleles, corresponding to the original definition of HLA allele nomenclature (Fig. 2b)[5]. The clusters of the six-digit classical HLA alleles had lower increments in variations than those of the four-digit classical HLA alleles (+1 cluster), whereas variations substantially increased from two-digit to four-digit alleles (+7 clusters). Given that the highly polymorphic nature of the HLA alleles is derived from balancing selection such as heterozygosity advantage, four-digit alleles (that is, amino acid polymorphisms) of the HLA genes might be main targets of the selection pressure rather than two-digit or six-digit alleles.

However, haplotype clusters of nonclassical HLA alleles had different patterns than those of classical HLA alleles (Fig. 2a), and parsimonious correspondences of the clusters between classical and nonclassical HLA alleles seemed to be difficult to define (Fig. 2b). This result suggests that nonclassical HLA genes have independent genetic landscapes in their variations compared with those of classical HLA genes, and that risk assessments of nonclassical HLA-gene variants should additionally contribute to fine-mapping efforts to identify causal functional variants in the MHC region.

NGS-based HLA and SNV imputation of Japanese GWAS data. Motivated by the newly identified genetic architecture of both classical and nonclassical HLA genes, we constructed a new HLA imputation reference panel of the Japanese population ($n = 1{,}120$). Whereas previous studies have focused primarily on the core MHC region for risk fine-mapping (around 29–33 Mb on chromosome 6, NCBI Build 37), we extended the target region into the MHC and its flanking region (24–36 Mb), which we define as the 'entire MHC' herein. Together with genotyping of the SNPs in the entire MHC region, we incorporated sequenced variants of the HLA genes and constructed the reference panel by using SNP2HLA[9]. The imputation accuracy of the constructed HLA imputation reference panel was empirically evaluated by a cross-validation approach[12]. Whereas previous studies have reported limited accuracy of NGS-based HLA typing[14,38], the newly constructed reference panel achieved high imputation accuracy (96.4 and 99.1% for the four-digit classical and nonclassical HLA alleles, respectively; Supplementary Table 3). This concordance

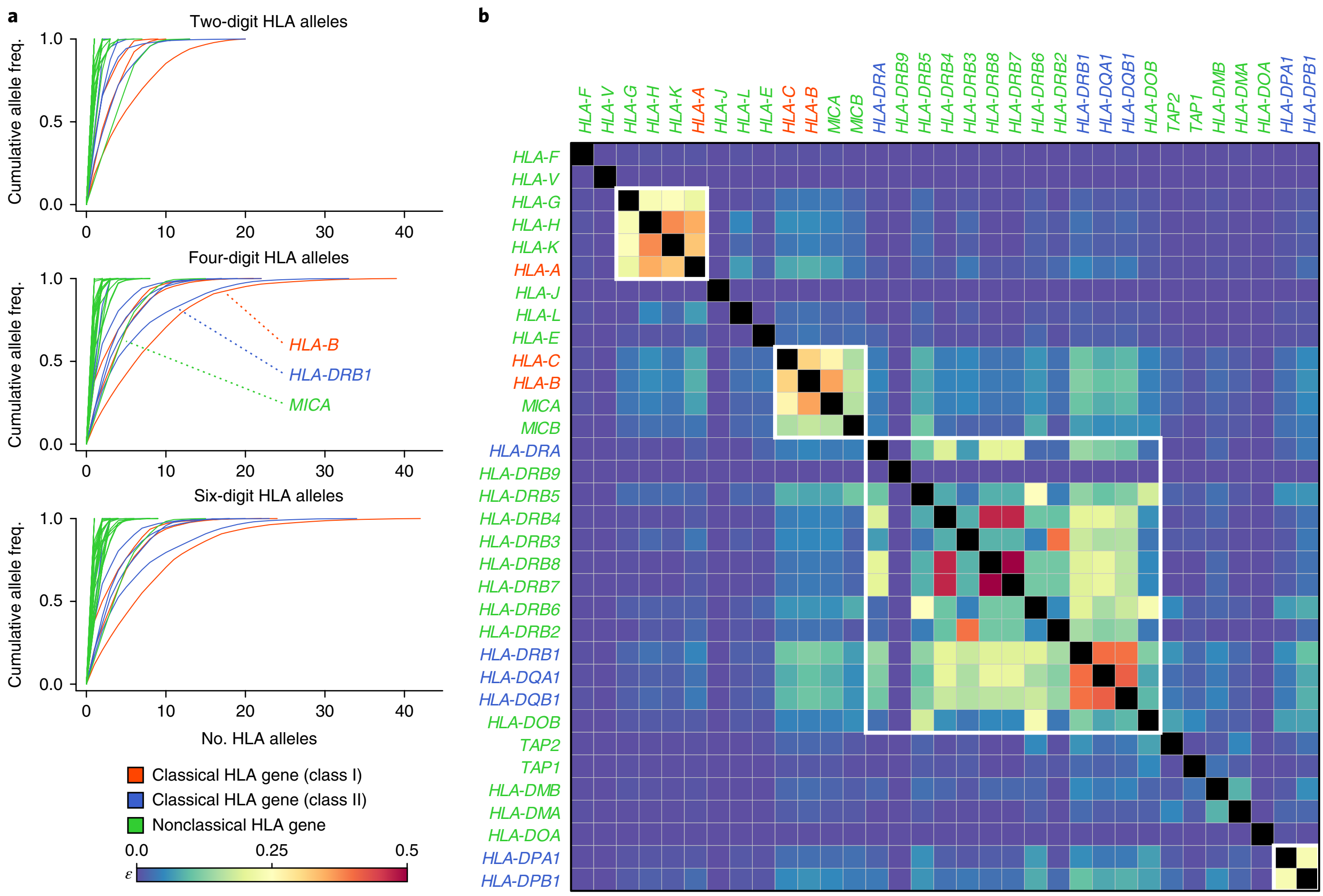

Fig. 1 | High-resolution allele-frequency spectra and linkage disequilibrium of HLA genes. a, Cumulative frequency (freq.) spectra of two-digit, four-digit and six-digit HLA alleles obtained by using NGS-based typing. Genes with the largest numbers of alleles are labeled separately for classical HLA genes (class I and class II) and nonclassical HLA genes. **b**, Pairwise evaluation of LD measurement, ε, among the HLA genes. ε uses normalized entropy of the haplotype frequency, and a higher ε value represents stronger LD. LD blocks ($\varepsilon > 0.15$) are highlighted with white boundaries.

was even better than that of the previously constructed SSO-method-based reference panel of Japanese individuals (95.9% for the four-digit classical HLA alleles, $n = 908$ for independent samples)[4].

Using the constructed reference panel, we densely imputed the HLA variants of the GWAS genotype data of the Japanese population constructed by BBJ ($n = 166,190$)[21,22]. To apply HLA imputation to such large-scale GWAS data, we updated the protocol to incorporate multiple software for genotype phasing and imputation (SNP2HLA, Eagle and minimac3; details in Methods). Furthermore, to complement SNP-microarray-based incomplete coverage of the variants, we densely imputed SNV and indels within the entire MHC region by using the deep-WGS data of the Japanese population as a reference ($n = 1,276$, average depth = 24.6×)[30]. After application of strict postimputation variant filtering (minor allele frequency (MAF) ≥ 0.5% and imputation score $Rsq \geq 0.7$), we obtained genotype dosages of 108 two-digit, 184 four-digit and 200 six-digit alleles and 2,273 amino acid polymorphisms of classical and nonclassical HLA genes, as well as 62,030 SNV and 4,203 indels in the entire MHC region (68,998 variants in total).

PheWAS identifies pleiotropy of MHC with human phenotypes. Using the NGS-based HLA, SNV and indel imputation data of the BBJ GWAS, we conducted PheWAS to comprehensively elucidate the genetic and phenotypic landscapes of the entire MHC. We incorporated data on 106 phenotypes collected from medical records of nationwide hospitals belonging to BBJ (Supplementary Table 5). Of these, 46 were complex diseases classified into four categories (immune related, metabolic and cardiovascular, cancers and other diseases)[21,22], and 60 were quantitative traits classified into ten categories (anthropometric, metabolic, protein, kidney related, electrolyte, liver related, other biochemical, hematological, blood pressure and echocardiographic)[25,26].

In the PheWAS, we evaluated associations of the entire MHC region with all of the 106 phenotypes. Approximately half of the phenotypes ($n = 52$; 16 diseases and 36 quantitative traits) indicated the association signals that satisfied the genome-wide-significance threshold ($P < 5.0 \times 10^{-8}$; ref. [39]; Table 1 and Supplementary Fig. 2), thus demonstrating substantial pleiotropic roles of MHC in a wide range of human phenotypes. Furthermore, stepwise conditional analysis identified multiple independent association signals in as many as 20 phenotypes (Supplementary Table 6). On average, 2.0 independent signals per phenotype were observed, with the largest number of seven signals observed for adult height and alkaline phosphatase. This result suggests that the genetic risk in MHC may reflect polygenic combinations of multiple functional and biological origins. Applying a multivariate regression model fitting nonadditive effects of the HLA alleles, we found significant nonadditive effects of HLA-DPB1*05:01 and HLA-DPB1*02:02 alleles on the risk of Graves' disease ($P < 3.7 \times 10^{-16}$; Supplementary Figure 3). Despite limited increments in allele variations from four-digit to six-digit alleles, several six-digit HLA alleles indicated more significant associations than those observed for the ancestral four-digit

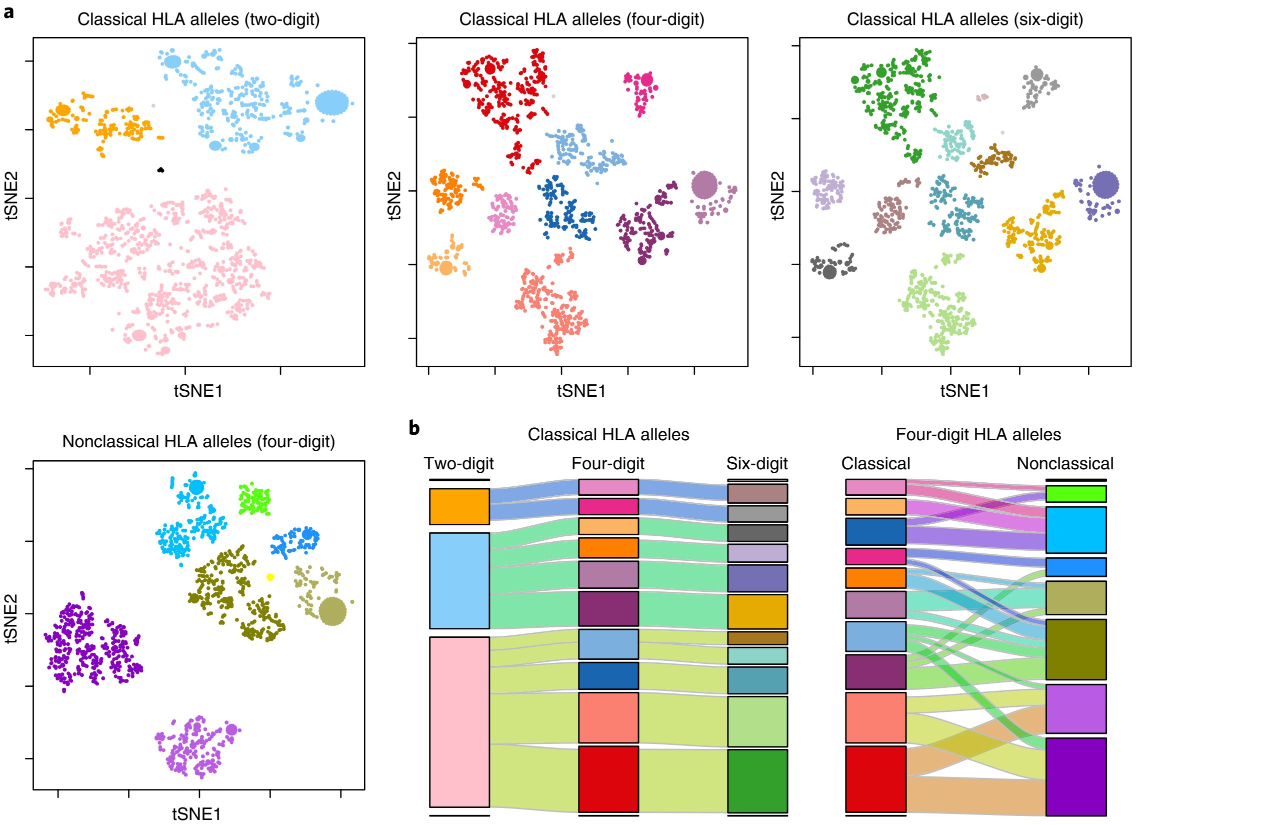

Fig. 2 | Machine-learning-based clustering of haplotypes by using HLA allele information. a, Unsupervised clustering results by machine learning (ML) using NGS-based HLA-typing data as inputs. Haplotypes are plotted on the basis of the two components of tSNE and clustered according to the DBSCAN algorithm. Clustering was separately conducted for each digit of classical or nonclassical HLA genes. **b**, Connections between machine-learning-based clusters of haplotypes. Each rectangle corresponds to the clusters identified in **a**. Rectangle height reflects the number of haplotypes included in each cluster.

alleles (for example, odds ratio = 1.32 and $P = 4.0 \times 10^{-28}$ at HLA-DRB4*01:03:02 but odds ratio = 1.13 and $P = 8.4 \times 10^{-11}$ at HLA-DRB4*01:03 with asthma).

PheWAS-based classifications of MHC-association patterns. Although our PheWAS approaches identified abundant association signals, their association patterns could be classified according to the types of the responsible genes (Fig. 3). (i) Associations of classical HLA genes were most evident (28 of the 52 top association signals and 52 of the 97 independent association signals). We observed that a series of quantitative traits, including hematological and blood pressure traits, were enriched in associations with the class I classical HLA gene variant (for example, $P = 6.7 \times 10^{-24}$ at HLA-C Tyr116 with basophil count and $P = 5.0 \times 10^{-40}$ at HLA-B amino acid position 116 with eosinophil count). As for the class II classical HLA genes, associations with diseases such as immune-related diseases and cancers were more evident than those with quantitative traits (for example, $P = 3.3 \times 10^{-43}$ at HLA-DQβ1 amino acid position 57 with chronic hepatitis B and $P = 1.1 \times 10^{-16}$ at rs9273367 in LD with HLA-DQβ1 Ile185 ($r^2 = 0.81$) with type 1 diabetes). (ii) Nonclassical HLA gene variants showed significant associations as well (for example, $P = 4.0 \times 10^{-28}$ at HLA-DRB4*01:03:02 with asthma and $P = 6.7 \times 10^{-10}$ at rs2844726 in LD with HLA-E amino acid position 107 ($r^2 = 0.76$) with red-blood-cell count). (iii) Associations of non-HLA gene variants were observed within each class of the MHC region (for example, $P = 9.5 \times 10^{-14}$ at rs2233965 at *C6orf15* with type 2 diabetes in the class I region, $P = 2.2 \times 10^{-20}$ at rs3830041 at *NOTCH4* with aspartate aminotransferase in the class III region, and $P = 2.6 \times 10^{-13}$ at rs3864302 at *C6orf10* with atopic dermatitis in the class II region). Such top association signals observed at the non-HLA gene variant within MHC still remained significant when conditioned on nearby HLA gene variants with the strongest association, thus confirming their independent phenotypic effects from the HLA genes. (iv) Non-HLA genes in the extended MHC region showed associations (for example, $P = 5.7 \times 10^{-18}$ at rs1799945 at *HFE* with mean corpuscular hemoglobin and $P = 4.4 \times 10^{-29}$ at rs2762353 at *SLC17A1* with uric acid). (v) Furthermore, non-HLA genes in the region flanking the MHC also showed associations (for example, $P = 1.9 \times 10^{-12}$ at rs73743323 at *IP6K3* with phosphorus and $P = 5.4 \times 10^{-77}$ at rs139458943 at *GPLD1* with alkaline phosphatase). In pattern 5, we observed the contribution of rare SNVs (MAF < 0.01) on several traits (for example, *GPLD1* for alkaline phosphatase and *GRM4* and *HMGA1* for adult height and estimated glomerular filtration rate). Our NGS-based HLA, SNV and indel imputation enabled us to detect such independent association signals from classical HLA genes. (vi) Population-specific long-range haplotypes characterize the LD structure of MHC[4]. Here, we show that a long-range haplotype that spans the entire MHC region specific to the Japanese population[2,4] had pleiotropic effects on multiple phenotypes ($P = 3.6 \times 10^{-23}$ with estimated glomerular filtration rate and $P = 7.3 \times 10^{-17}$ with triglyceride). (vii) Analogously to the identification of multiple independent association signals for a single phenotype, several traits confer combinations of multiple association patterns (for example, associations with adult height in the class II classical HLA variant ($P = 6.7 \times 10^{-17}$ at HLA-DRβ 1 74Ala, pattern 2), the extended MHC region ($P = 2.0 \times 10^{-39}$ at rs9379833 at *HIST1H2BE*, pattern 5) and the region flanking the MHC ($P = 5.7 \times 10^{-72}$ at rs4713762 at *HMGA1*, pattern 4)).

Table 1 | Significant association signals in the entire MHC region identified by PheWAS

Trait	No. samples (cases, controls)	No. independent signals in MHC	Top associated variant[a]	Position (hg19)	Allele 1/2	Gene	Allele 1 frequency (cases, controls)	Effect size (allele 1)	P[b]
Immune-related diseases									
Asthma	(7,207, 62,407)	1	HLA-DRB4*01:03:02	32,502,549	–	*HLA-DRB4,* *HLA-DRB1*	(0.173, 0.140)	1.32 (1.25–1.38)	4.0×10^{-28}
Atopic dermatitis	(2,358, 62,407)	1	rs3864302	32,278,792	T/C	*C6orf10*	(0.266, 0.316)	0.79 (0.74–0.84)	2.6×10^{-13}
Chronic hepatitis B	(1,238, 62,407)	4	HLA-DQβ1 position 57	32,631,702	–	*HLA-DQB1*	–	–	3.3×10^{-43}
Chronic hepatitis C	(5,333, 62,407)	2	HLA-B position 156	31,323,307	–	*HLA-B*	–	–	4.2×10^{-12}
Graves' disease	(1,938, 62,407)	4	rs72500561	33,039,958	G/A	*HLA-DPA1,* *HLA-DPB1*	(0.587, 0.464)	1.63 (1.53–1.74)	7.0×10^{-50}
Pollinosis	(5,139, 58,556)	1	rs9272544	32,606,878	A/G	*HLA-DQA1*	(0.417, 0.439)	0.89 (0.85–0.93)	1.8×10^{-8}
Rheumatoid arthritis	(2,346, 62,407)	2	HLA-DQA1*03:03	32,605,398	–	*HLA-DQA1,* *HLA-DRB1*	(0.322, 0.157)	2.65 (2.49–2.83)	2.0×10^{-173}
Type 1 diabetes mellitus	(106, 62,407)	1	rs9273367	32,626,438	T/A	*HLA-DQB1*	(0.707, 0.423)	3.29 (2.44–4.43)	1.1×10^{-16}
Metabolic and cardiovascular diseases									
Hyperlipidemia	(43,939, 62,407)	1	HLA-DQβ1 position −21	32,631,702	–	*HLA-DQB1*	–	–	3.5×10^{-18}
Myocardial infarction	(11,868, 62,407)	1	HLA-B position 80	31,323,307	–	*HLA-B*	–	–	3.4×10^{-14}
Stable angina	(14,461, 62,407)	1	rs1362104	30,101,656	C/T	Long-range haplotype	(0.386, 0.369)	1.08 (1.05–1.11)	1.5×10^{-8}
Type 2 diabetes mellitus	(36,698, 62,407)	1	rs2233965	31,080,899	G/T	*C6orf15*	(0.328, 0.346)	0.93 (0.91–0.95)	9.5×10^{-14}
Cancers									
Lung cancer	(3,615, 62,407)	1	HLA-DQβ1 position 67	32,631,702	–	*HLA-DQB1*	–	–	6.1×10^{-9}
Liver cancer	(1,587, 62,407)	1	rs9271377	32,587,165	G/T	*HLA-DQA1*	(0.148, 0.186)	0.75 (0.68–0.83)	9.4×10^{-9}
Other diseases									
Liver cirrhosis	(1,824, 62,407)	1	rs3129943	32,338,695	G/A	*C6orf10*	(0.415, 0.368)	1.22 (1.14–1.31)	7.2×10^{-9}
Nephrotic syndrome	(871, 62,407)	1	HLA-DQA1*05:05:01	32,605,398	–	*HLA-DQA1*	(0.069, 0.041)	1.81 (1.49–2.19)	2.1×10^{-8}
Anthropometric QTL									
Adult height	151,336	7	rs4713762	34,231,661	A/G	*HMGA1*	0.131	0.096 (0.0053)	5.7×10^{-72}
Body mass index	150,369	3	HLA-DQβ1 position 185	32,631,702	–	*HLA-DQB1*	–	–	2.8×10^{-13}
Metabolic QTL									
Total cholesterol	123,854	1	HLA-DQβ1 position 30	32,631,702	–	*HLA-DQB1*	–	–	1.3×10^{-9}
HDL cholesterol	68,016	1	rs4947340	32,435,338	C/T	*HLA-DRA*	0.451	-0.034 (0.0054)	7.5×10^{-10}
Triglyceride	101,870	1	rs9469053	31,755,776	G/A	Long-range haplotype	0.075	0.069 (0.0081)	7.3×10^{-17}
Blood sugar	89,917	1	rs28360985	30,993,244	T/C	*MUC22*	0.211	−0.039 (0.0057)	1.8×10^{-11}
Hemoglobin A1c	41,121	1	rs2844542	31,347,274	C/G	*MICA*	0.348	0.040 (0.0073)	2.9×10^{-8}
Protein QTL									
Total protein	109,640	2	rs13197513	32,990,121	C/T	*HLA-DPB1*	0.051	−0.074 (0.0097)	1.5×10^{-14}
Albumin	98,739	1	rs77849299	31,456,345	C/G	*MICB*	0.224	−0.043 (0.0054)	2.2×10^{-15}
Nonalbumin protein	95,151	2	rs28752797	31,291,172	C/T	*HLA-C*	0.193	0.061 (0.0058)	6.8×10^{-26}
Albumin/globulin ratio	95,238	4	6:31468859	31,469,859	T/TC	*MICB*	0.189	−0.070 (0.0059)	7.4×10^{-33}
Kidney-related QTL									
Serum creatinine	137,322	2	rs28360975	30,978,834	T/G	Long-range haplotype	0.081	0.066 (0.0067)	9.6×10^{-22}
Estimated glomerular filtration rate	138,827	3	rs28360975	30,978,834	T/G	Long-range haplotype	0.081	−0.068 (0.0069)	3.6×10^{-23}
Uric acid	105,190	3	rs2762353	25,794,431	T/C	*SLC17A1*	0.160	−0.066 (0.0058)	4.4×10^{-29}
Electrolyte QTL									
Potassium	128,510	1	rs3129943	32,338,695	G/A	*C6orf10*	0.367	0.022 (0.0041)	4.7×10^{-8}
Phosphorus	41,346	1	rs73743323	33,705,355	T/C	*IP6K3*	0.032	−0.138 (0.0195)	1.9×10^{-12}
Liver-related QTL									
Total bilirubin	106,555	1	HLA-DQA1*03:03:01	32,605,398	–	*HLA-DQA1*	0.162	0.039 (0.0059)	2.5×10^{-11}
Aspartate aminotransferase	129,615	1	rs3830041	32,191,339	A/G	*NOTCH4*	0.185	0.047 (0.0049)	2.2×10^{-20}
Alanine aminotransferase	129,662	1	rs206769	32,961,104	T/C	*HLA-DMA*	0.214	0.026 (0.0047)	4.7×10^{-8}

Continued

Table 1 | Significant association signals in the entire MHC region identified by PheWAS (Continued)

Trait	No. samples (cases, controls)	No. independent signals in MHC	Top associated variant[a]	Position (hg19)	Allele 1/2	Gene	Allele 1 frequency (cases, controls)	Effect size (allele 1)	P[b]
Alkaline phosphatase	101,464	7	rs139458943	24,497,823	A/G	*GPLD1*	0.065	−0.168 (0.0090)	5.4×10^{-77}
Other biochemical QTL									
Creatine kinase	102,511	1	HLA-DQβ1 position −17	32,631,702	–	*HLA-DQB1, HLA-DRB1*	–	–	1.1×10^{-24}
Lactate dehydrogenase	122,047	1	HLA-DQβ1 position −4	32,631,702	–	*HLA-DQB1*	–	–	6.7×10^{-29}
Hematological QTL									
White-blood-cell count	104,453	4	rs2524084	31,241,639	A/G	*HLA-C*	0.425	0.052 (0.0043)	1.1×10^{-32}
Neutrophil count	60,350	2	rs2853946	31,247,203	T/A	*HLA-C*	0.275	0.059 (0.0064)	5.6×10^{-20}
Eosinophil count	60,350	3	HLA-B position 116	31,323,307	–	*HLA-B, HLA-C*	–	–	5.0×10^{-40}
Basophil count	60,350	1	HLA-C Tyr116	31,238,217	–	*HLA-C*	0.428	−0.059 (0.0058)	6.7×10^{-24}
Monocyte count	60,350	2	rs2524084	31,241,639	A/G	*HLA-C*	0.425	0.066 (0.0058)	1.0×10^{-29}
Lymphocyte count	60,350	2	rs4959105	32,583,146	T/C	*HLA-DRB1*	0.456	−0.053 (0.0057)	1.7×10^{-20}
Red-blood-cell count	105,252	1	rs2844726	30,444,357	T/C	*HLA-E*	0.322	0.029 (0.0046)	6.7×10^{-10}
Hemoglobin	105,146	1	rs2302398	31,088,232	A/G	*CDSN*	0.228	0.030 (0.0051)	1.1×10^{-8}
MCV	104,487	1	rs9264579	31,235,746	A/G	*HLA-C*	0.346	0.031 (0.0045)	1.5×10^{-11}
Mean corpuscular hemoglobin	104,308	2	rs1799945	26,091,179	G/C	*HFE*	0.029	0.112 (0.0127)	5.7×10^{-18}
MCHC	104,912	1	rs1799945	26,091,179	G/C	*HFE*	0.029	0.093 (0.0127)	5.1×10^{-13}
Platelet count	104,696	5	rs5745568	33,548,394	A/C	*BAK1*	0.230	0.070 (0.0051)	3.3×10^{-41}
Blood-pressure QTL									
Systolic blood pressure	132,148	1	rs2523557	31,331,257	G/A	*HLA-B*	0.143	0.032 (0.0054)	3.6×10^{-9}
Mean arterial pressure	132,033	1	rs4947311	31,326,166	C/T	*HLA-B*	0.158	0.029 (0.0053)	2.6×10^{-8}

Significantly associated variants identified by the PheWAS are indicated. [a]When the omnibus *P* value of the HLA amino acid position indicated the most significant associations, no amino acid residue was indicated. [b]Two-tailed *P* values calculated with logistic or linear regression that satisfied the genome-wide-significance threshold ($P < 5.0 \times 10^{-8}$) are indicated without adjustment.

Cluster visualization of the observed association patterns could help illustrate the overall genetic and phenotypic landscape within the entire MHC region (Fig. 4). Significant MHC associations with 11 traits were newly identified by our study (that is, pollinosis, hyperlipidemia, myocardial infarction, stable angina, type 2 diabetes, liver cancer, liver cirrhosis, nephrotic syndrome, total protein, potassium and creatine kinase). In addition, we newly identified trait-associated signals on previously unreported HLA variants or other MHC variants in 37 phenotypes (Supplementary Table 7).

Our NGS-based MHC fine-mapping efforts were able to refine responsible risk variants that had not been identified earlier (Supplementary Table 6). For example, previous studies on hepatitis B in Japanese individuals have suggested that *HLA-DRB1*, *HLA-DQB1* and *HLA-DPB1* allele haplotypes can explain the risk embedded within the MHC class II region[40]. However, our study shows that the amino acid polymorphisms of HLA-DQβ1 (position 57), HLA-DPα1 (position 111) and HLA-DQα1 (position 160) independently explained the risk. Although a contribution of the HLA-C allele was originally suggested for monocyte count[41], our study additionally identifies risk at *MICB* (rs2395040), which would support the roles of monocytes in disease pathophysiology[42].

Genetic correlation within MHC highlights phenotype networks. Another approach to infer genetic and phenotypic overlap is to estimate genetic correlation[24–26]. Contrary to the PheWAS approach that assesses point-by-point connections between single variants and phenotypes, genetic correlation could account for shared polygenic architecture across the phenotypes. To that end, we estimated region-wide polygenic heritability of phenotypes that was explained by variants within the entire MHC (Fig. 5a and Supplementary Table 5). As reported previously[4,18,40,43], immune-related diseases such as type 1 diabetes, rheumatoid arthritis, Graves' disease, chronic hepatitis B and asthma showed the highest region-wide heritability (9.8, 9.5, 4.6, 3.5 and 1.1%, respectively). Although single-variant associations were not significant, possibly because of the small sample size of the cases ($n = 547$), uterine cervical cancer showed relatively high heritability among the phenotypes (1.6%). When the proportions of the heritability explained by classical HLA gene variants and other MHC variants not in LD with them ($r^2 < 0.1$) were quantified, immune-related diseases showed the largest proportions of heritability derived from classical HLA gene variants (on average 0.69), whereas metabolic and cardiovascular diseases showed the smallest proportions (on average 0.32).

Finally, we estimated genetic correlations of the entire MHC region across the phenotypes and visualized cross-phenotype networks reflecting shared polygenic architecture and embedded biological information. As suggested by single-phenotype heritability analysis, the genetic-correlation network of classical HLA gene variants and that of other MHC variants showed different patterns of connections (Fig. 5b). In the former, several tight connections among the phenotypes belonging to the same categories (for example, immune related, metabolic and cardiovascular, hematological and protein) together configure the entire network. In the latter, the entire network was divided into subnetworks constituted separately by diseases and quantitative traits. As an example of a specific trait, rheumatoid arthritis showed positive correlations with asthma, type 1 and 2 diabetes mellitus, and total bilirubin but a negative correlation with body mass index in classical HLA gene variants, whereas it showed negative correlations with hyperlipidemia, stable angina, myocardial infarction, lactate dehydrogenase and eosinophil count in other MHC variants. These results indicate that polygenic architecture of the entire MHC region confers pleiotropic diversity according to the

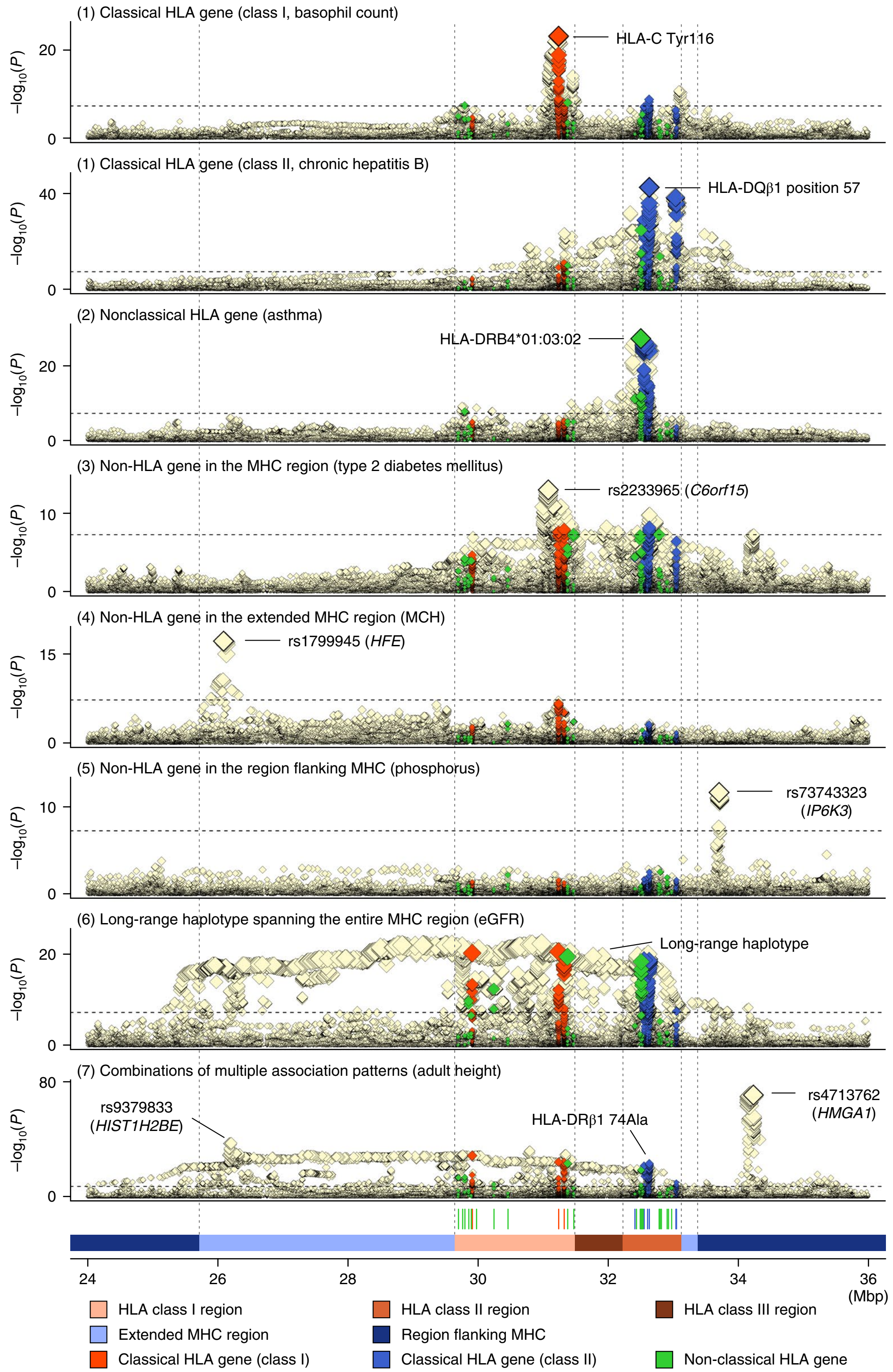

Fig. 3 | Genotype–phenotype association patterns identified by PheWAS with NGS-based HLA, SNV and indel imputation. Regional association plots of the entire MHC region in the PheWAS on the large-scale GWAS of the BBJ. Horizontal bar represents significance threshold. NGS-based HLA, SNV and indel imputation enabled classification of the association patterns of genetic risk factors within MHC (from top to bottom): (1) classical HLA gene (class I and II), (2) nonclassical HLA gene, (3) non-HLA gene in the MHC region, (4) non-HLA gene in the extended MHC region, (5) non-HLA gene in the region flanking MHC, (6) long-range haplotype spanning the entire MHC region and (7) combinations of the multiple association patterns. Two-tailed *P* values calculated with logistic or linear regression are indicated without adjustment ($n = 166{,}190$ independent Japanese individuals). Dotted horizontal lines indicate genome-wide-significance threshold of $P = 5.0 \times 10^{-8}$. MCH, mean corpuscular hemoglobin; eGFR, estimated glomerular filtration rate.

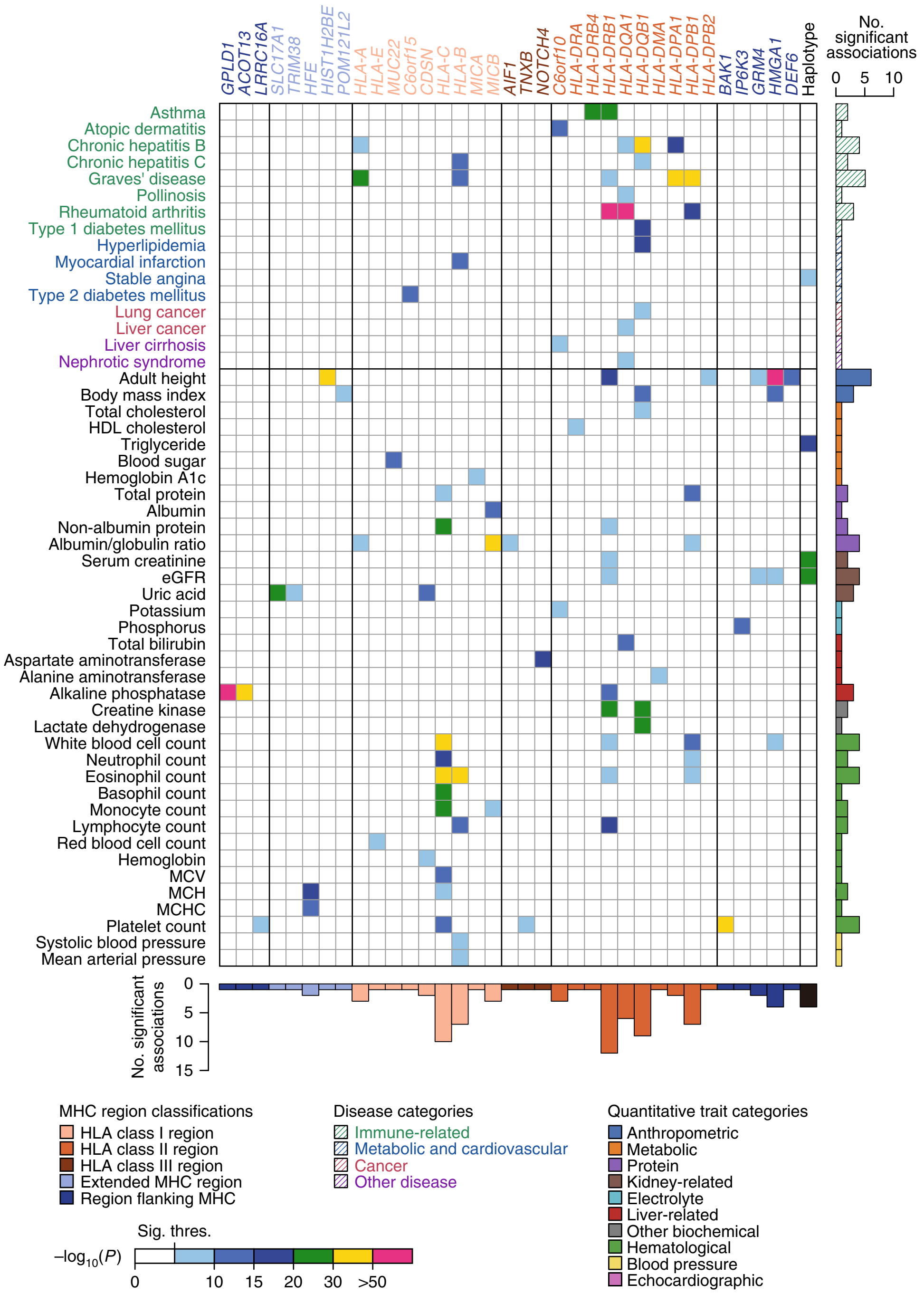

Fig. 4 | Matrix plot of gene and phenotype associations in the entire MHC region. Significantly associated gene and phenotype pairs identified by PheWAS are plotted in the matrix. In addition to the top association signals of the phenotypes, independent associations identified by conditional analysis are indicated. The bars at the right and bottom show the number of association signals per phenotype and gene, respectively. Two-tailed *P* values calculated with logistic or linear regression are indicated without adjustment ($n = 166{,}190$ independent Japanese individuals). MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; sig. thres., significance threshold.

phenotypes, phenotype categories and functional categories of the responsible genes.

Discussion

Through NGS-based typing of high-resolution HLA gene polymorphisms and implementation of the imputation reference panel in the Japanese population, our PheWAS approach using large-scale GWAS data successfully fine-mapped the genetic risk embedded in the entire MHC region and excavated the cross-phenotype genetic-correlation network.

Our study highlights several new findings. First, we constructed a catalog of NGS-based high-resolution frequency spectra of both classical and nonclassical HLA alleles. Our resources should contribute to the understanding of the biological and clinical roles of nonclassical HLA genes, a challenging area of MHC yet to be investigated[13]. Our next steps will include (i) direct construction of a highly accurate HLA imputation reference panel from WGS data without target sequencing of HLA and (ii) application of long-read sequencing technology to copy number variants and other complex genomic regions such as killer cell immunoglobulin-like receptor[44].

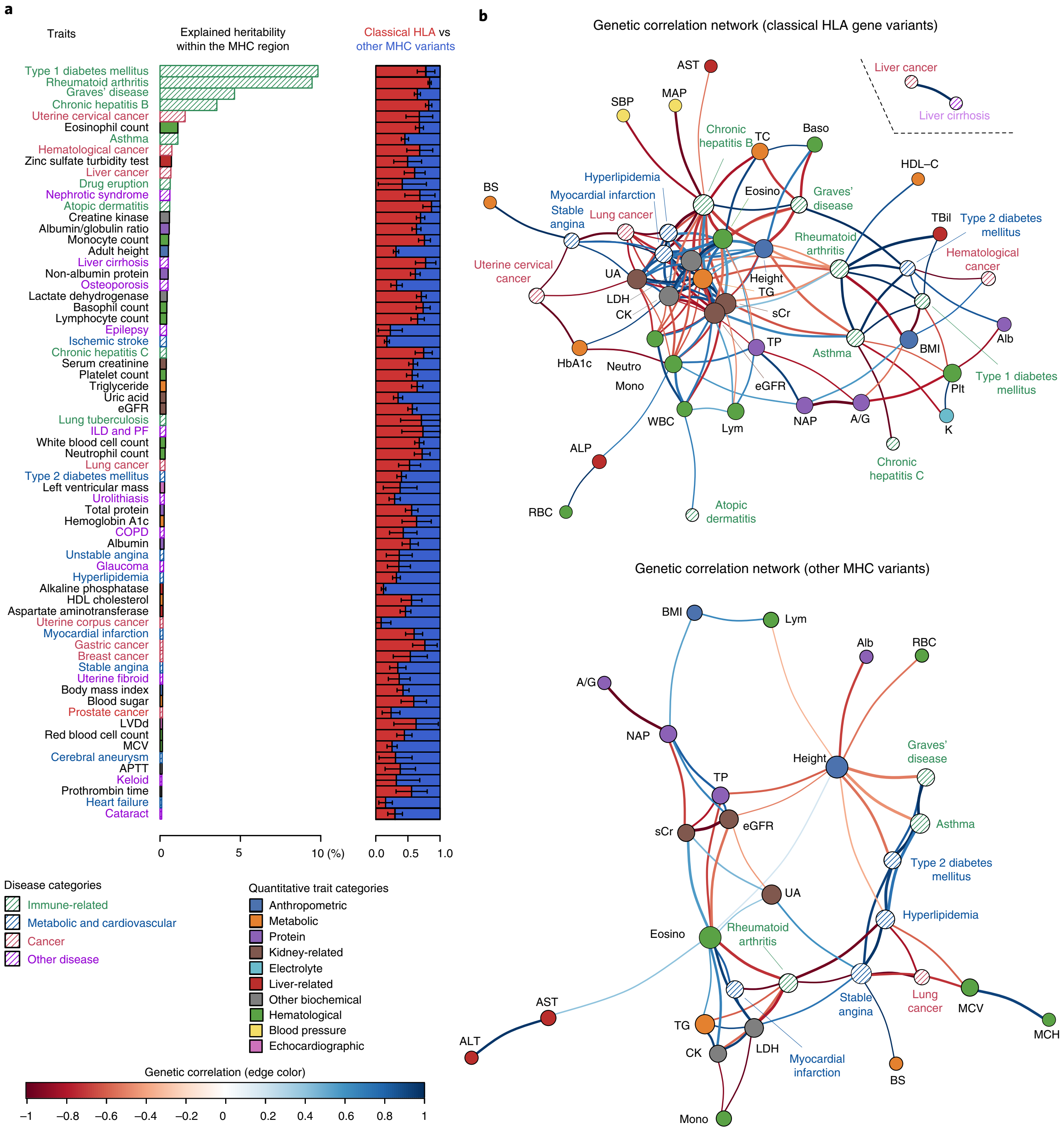

Fig. 5 | Region-wide heritability and genetic-correlation networks across phenotypes. a, Heritability estimates of phenotypes, based on variants within the entire MHC region. Phenotypes with >0.1% of the explained heritability are indicated at left. Right, heritability proportions and empirically estimated standard errors (s.e.) between classical HLA gene variants and other MHC variants ($n = 166{,}190$ independent Japanese individuals). **b**, Region-wide genetic-correlation networks across phenotypes depicted from classical HLA gene variants (top) and other MHC variants (bottom). Genetically correlated phenotypes are clustered close together as circles, and each edge represents a significant genetic correlation. Positive and negative genetic correlations are indicated by color according to the legend, and thicker edges correspond to more significant correlations. Abbreviations are provided in Supplementary Table 5.

The strategy in this study to separately impute HLA and SNV by using different reference panels could disrupt LD among the variants, and imputation of all the variants of interest by using a single panel is warranted. Second, application of a high-dimensional compression technique to the HLA data, such as tSNE originally applied for epigenetic data[45,46], effectively configured unbiased clustering of the haplotypes. The result is notable because machine-learning-based unsupervised clustering successfully recaptured the original definition of HLA-allele nomenclature and identify the independent genetic landscapes of classical and nonclassical HLA genes without prior biological or genetic knowledge. This finding indicates that trans-omics sharing of analytical methods between genomics and epigenomics fields may yield innovative findings[47]. Third, NGS-based HLA, SNV and indel imputation followed by the PheWAS approach successfully demonstrated a wide range of genotype–phenotype correlations in complex human traits. Approximately half of the phenotypes examined in our PheWAS showed significant associations; this proportion was larger than we expected on the basis of similar previous approaches[27,29]. Our study indicates the value of PheWAS focusing on large-scale genotype data on sites with pleiotropic features. Further accumulation of genotype and clinical data is warranted to achieve larger study scales. Fourth, dense fine-mapping efforts highlighted several patterns of association signals within the entire MHC region. In particular, we confirmed independent phenotype risk from classical HLA genes, namely nonclassical HLA genes and non-HLA genes within the core MHC, extended MHC and flanking regions. Finally, MHC-region-wide heritability and genetic-correlation estimates depicted cross-phenotype networks in a manner complementing those obtained from single-variant and multiple-phenotype associations such as PheWAS. As an intermediate approach between single-variant analysis and genome-wide polygenic assessments, region-wide or locus-based approaches may be promising as well[48].

In conclusion, our study comprehensively elucidated the genetic and phenotypic landscapes of MHC in the Japanese population.

URLs. The BioBank Japan Project (BBJ), https://biobankjp.org/english/index.html; Japan Biological Informatics Consortium (JBIC), http://www.jbic.or.jp/english/; Omixon Target software, https://www.omixon.com/; BWA, http://bio-bwa.sourceforge.net/; GATK, https://software.broadinstitute.org/gatk/; IPD-IMGT/HLA database, https://www.ebi.ac.uk/ipd/imgt/hla/; OptiType, https://github.com/FRED-2/OptiType/; POLYSOLVER, https://software.broadinstitute.org/cancer/cga/polysolver/; HLA-HD, https://www.genome.med.kyoto-u.ac.jp/HLA-HD/; Kourami, https://github.com/Kingsford-Group/kourami/; eLD, http://www.sg.med.osaka-u.ac.jp/tools.html; Rtsne R package, https://cran.r-project.org/web/packages/Rtsne/index.html; DBSCAN R package, https://cran.r-project.org/web/packages/dbscan/index.html; Alluvial R package, https://github.com/mbojan/alluvial/; SNP2HLA, http://software.broadinstitute.org/mpg/snp2hla/; Eagle, https://data.broadinstitute.org/alkesgroup/Eagle/; Minimac3, https://genome.sph.umich.edu/wiki/Minimac3#Download/; R statistical software, https://cran.r-project.org/; GCTA, http://cnsgenomics.com/software/gcta/; Igraph R package, https://cran.r-project.org/web/packages/igraph/index.html.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0336-0.

Received: 27 February 2018; Accepted: 13 December 2018;
Published online: 28 January 2019

References

1. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
2. Okada, Y. et al. HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology* **141**, 864–871 (2011).
3. 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).
4. 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).
5. Robinson, J., Soormally, A. R., Hayhurst, J. D. & Marsh, S. G. The IPD-IMGT/HLA Database: new developments in reporting HLA variation. *Hum. Immunol.* **77**, 233–237 (2016).
6. The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923 (1999).
7. Horton, R. et al. Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).
8. Nejentsev, S. et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* **450**, 887–892 (2007).
9. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
10. 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).
11. Okada, Y. et al. Fine mapping major histocompatibility complex associations in psoriasis and its clinical subtypes. *Am. J. Hum. Genet.* **95**, 162–172 (2014).
12. Hirata, J. et al. Variants at HLA-A, HLA-C, and HLA-DQB1 confer risk of psoriasis vulgaris in Japanese. *J. Invest. Dermatol.* **138**, 542–548 (2018).
13. Hosomichi, K., Shiina, T., Tajima, A. & Inoue, I. The impact of next-generation sequencing technologies on HLA research. *J. Hum. Genet.* **60**, 665–673 (2015).
14. Zhou, F. et al. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat. Genet.* **48**, 740–746 (2016).
15. Robinson, J. et al. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.* **13**, e1006862 (2017).
16. Schofl, G. et al. 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. *BMC Genomics* **18**, 161 (2017).
17. Walter, K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
18. Okada, Y. et al. Contribution of a nonclassical HLA gene, HLA-DOA, to the risk of rheumatoid arthritis. *Am. J. Hum. Genet.* **99**, 366–374 (2016).
19. Okushi, Y. et al. Circulating and renal expression of HLA-G prevented chronic renal allograft dysfunction in Japanese recipients. *Clin. Exp. Nephrol.* **21**, 932–940 (2017).
20. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
21. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
22. 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).
23. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
24. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
25. Akiyama, M. et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).
26. 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).
27. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
28. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016).
29. Karnes, J. H. et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci. Transl. Med.* **9**, eaai8708 (2017).
30. Okada, Y. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
31. Hosomichi, K. et al. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. *BMC Genomics* **14**, 355 (2013).
32. Hosomichi, K., Mitsunaga, S., Nagasaki, H. & Inoue, I. A bead-based normalization for uniform sequencing depth (BeNUS) protocol for multi-samples sequencing exemplified by HLA-B. *BMC Genomics* **15**, 645 (2014).
33. Yang, K. L. et al. New allele name of some HLA-DRB1*1401: HLA-DRB1*1454. *Int. J. Immunogenet.* **36**, 119–120 (2009).
34. Morizane, A. et al. MHC matching improves engraftment of iPSC-derived neurons in non-human primates. *Nat. Commun.* **8**, 385 (2017).
35. van der Maaten, L. & Hilton, G. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
36. van der Maaten, L. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
37. Ester, M., Kriegel, H. P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *KDD '96 Proc. Second Int. Conf. Knowl. Discov. Data Min.*, 226–231 (AAAI Press, Palo Alto, CA, USA, 1996).

38. Bauer, D. C., Zadoorian, A., Wilson, L. O. W. & Thorne, N. P. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Brief. Bioinform.* **19**, 179–187 (2018).
39. Kanai, M., Tanaka, T. & Okada, Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J. Hum. Genet.* **61**, 861–866 (2016).
40. Nishida, N. et al. Understanding of HLA-conferred susceptibility to chronic hepatitis B infection requires HLA genotyping-based association analysis. *Sci. Rep.* **6**, 24767 (2016).
41. Okada, Y. et al. Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. *PLoS Genet.* **7**, e1002067 (2011).
42. Ikeshita, S., Miyatake, Y., Otsuka, N. & Kasahara, M. MICA/B expression in macrophage foam cells infiltrating atherosclerotic plaques. *Exp. Mol. Pathol.* **97**, 171–175 (2014).
43. 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).
44. Roe, D. et al. Revealing complete complex KIR haplotypes phased by long-read sequencing technology. *Genes Immun.* **18**, 127–134 (2017).
45. See, P. et al. Mapping the human DC lineage through the integration of high-dimensional techniques. *Science* **356**, eaag3009 (2017).
46. Takeuchi, Y. et al. Clinical response to PD-1 blockade correlates with a sub-fraction of peripheral central memory CD4+ T cells in patients with malignant melanoma. *Int. Immunol.* **30**, 13–22 (2018).
47. Platzer, A. Visualization of SNPs with t-SNE. *PLoS One* **8**, e56883 (2013).
48. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).

Acknowledgements

We thank T. Aoi for kind support of the study. This research was supported by the Tailor-Made Medical Treatment program (the BioBank Japan Project) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Agency for Medical Research and Development (AMED), MEXT KAKENHI (221S0002), Bioinformatics Initiative of Osaka University Graduate School of Medicine, and Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University. Y.O. was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (15H05670, 15H05911 and 15K14429), AMED (18gm6010001h0003 and 18ek0410041h0002), Takeda Science Foundation, the Uehara Memorial Foundation, the Naito Foundation, Daiichi Sankyo Foundation of Life Science, Senri Life Science Foundation and Suzuken Memorial Foundation. K.H. was supported by JSPS KAKENHI grant no. P16H06502 'Neo-self'. J.H. is an employee of Teijin Pharma Limited. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Author contributions

Y.O. supervised the study. J.H., K.H., Y.K. and Y.O. wrote the manuscript. J.H., K.H., S.S., M. Kanai, K.I., K.S., M.A., T.K., K.O., T.M., K.Y., Y.K. and Y.O. conducted data analysis. M.H., K.M., M. Kubo and Y.K. provided data. Y.O., K.M. and M. Kubo collected samples. Y.O., K.H., H.N., Y.M. and I.I. conducted experiments.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0336-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to Y.O.

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

Methods

Cohort. To construct the NGS-based HLA typing data, we enrolled 1,120 unrelated individuals of Japanese ancestry. Genomic DNA was obtained from Epstein–Barr virus–transformed B-lymphoblast cell lines of unrelated Japanese individuals established by the Japan Biological Informatics Consortium (JBIC)[12]. In the PheWAS, 166,190 individuals were enrolled from BBJ, and participants were affected with any of the 45 target diseases defined by the project (Supplementary Table 5)[21,22]. As for the WGS-based SNV imputation reference panel, 1,276 independent individuals of BBJ were enrolled (patients with myocardial infarction, drug eruption, colorectal cancer, breast cancer, prostate cancer or gastric cancer)[30]. Individuals determined to be of non-Japanese origin either by self-reporting or by principal component analysis were excluded, as described[12,25,26,30]. All the BBJ individuals provided written informed consent, as approved by the ethical committees of RIKEN Yokohama Institute and the Institute of Medical Science, University of Tokyo. This study was approved by the ethical committee of Osaka University Graduate School of Medicine.

NGS-based HLA typing of Japanese individuals. We conducted high-resolution allele typing (two-digit, four-digit and six-digit alleles) of 33 HLA and HLA-related genes, of which 9 were classical HLA genes (*HLA-A*, *HLA-B* and *HLA-C* for class I; *HLA-DRA*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1* for class II) and 24 were nonclassical HLA genes (*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* and *TAP2*; Supplementary Table 2). Although current definitions of the HLA gene classifications are ambiguous (for example, classical HLA gene, nonclassical HLA gene, HLA-like gene or pseudo-HLA gene)[6,7], in this study, we defined the major classical HLA genes as classical HLA genes and other genes as nonclassical HLA genes. We also defined alleles of classical HLA genes as classical HLA alleles and those of nonclassical HLA genes as nonclassical HLA alleles for simplicity.

Entire HLA gene sequencing with the sequence-capture method was used for high-resolution HLA typing[13]. The sequence-capture method was based on hybridization between DNA of an adapter-ligated library (KAPA Hyper Prep Kit, Roche) and a biotinylated DNA probe (SeqCap EZ choice kit, Roche) custom designed on the basis of target sequences of 33 HLA genes (length of total target regions = 236,885 bp; Supplementary Table 8). Paired-end sequence reads (read 1, 350 bp; read 2, 250 bp) were obtained by using a MiSeq sequencer (Illumina). Typing of two-digit, four-digit and six-digit HLA alleles was conducted in Omixon Target software version 1.9.3 (Omixon) with IPD-IMGT/HLA Database release 3.21.0. Phase-defined HLA gene analysis was also used to resolve the phase ambiguity[31,32]. In parallel, to complement the HLA allele information that was specific to the Japanese population and not correctly implemented in Omixon Target software, we obtained SNV genotypes in PCR-amplified regions according to the variant-calling pipeline[31], and partially updated the HLA typing results on the basis of those obtained according to the sequencing-based typing method. The sequence reads were aligned to the reference human genome with the contig sequences of the MHC region (GRCh37 (human_g1k_v37.fasta), hap2_cox contig and hap5_mcf contig) using BWA (version 0.7.15). Variant calling was conducted with GATK HaplotypeCaller and UnifiedGenotyper (version 3.6). HLA allele sequences were obtained from the IPD-IMGT/HLA database[5].

We empirically confirmed the accuracy of HLA typing by evaluating concordance rates of the four-digit HLA alleles with those additionally genotyped with the SSO method (a WAKFlow HLA typing kit (Wakunaga) together with the Luminex Multi-Analyte Profiling system (xMAP, Luminex); $n = 182$ for *HLA-A*, *HLA-B*, *HLA-C* and *HLA-DRB1*, and $n = 144$ for *HLA-DQA1*, *HLA-DQB1* and *HLA-DPB1*). We observed a high concordance rate of 98.2% between typed alleles of NGS and SSO (2,278 of 2,320 alleles in total). We confirmed that most mismatched alleles (29 of 42) derived from wrong typing of SSO but not NGS as previously reported (for example, HLA-DRB1*14:01 by SSO was corrected to HLA-DRB1*14:54 by NGS[33]; details in Supplementary Table 1). This provides confidence in the accuracy of our NGS-based HLA typing protocol (≤0.56% of potentially inaccurate typing). Although we further attempted to verify these ambiguous mismatched alleles by using tools to estimate HLA alleles from WGS or whole-exome-sequencing data (OptiType (version 1.3.1)[49], Polysolver (version 4)[50], HLA-HD (version 1.2.0.1)[51] and Kourami (version 0.9.6)[52]), it was difficult to determine the correct alleles, owing to inconsistent outputs of the tools.

In addition, we assessed concordance rates of the SNV genotypes between microarray-based SNP genotyping data (described below) and those obtained by target sequencing used for NGS-based HLA typing. Among the 203 SNVs genotyped by both SNP microarray and NGS, the genotype concordance was as high as 0.997. Of these, 29 and 45 SNVs were included in the coding regions of classical and nonclassical HLA genes, with concordance rates of 0.994 and 0.998, respectively.

Assessment of LD structure on the basis of normalized entropy index. To evaluate LD structure among HLA genes, we introduced an LD-measurement index called ε, which uses differences in the normalized entropy of the haplotype-frequency distributions between LD and the null hypothesis of linkage equilibrium[53], by using eLD software (version 1.0)[54]. ε was originally developed to assess LD among multiple biallelic markers, and we previously showed that ε is also applicable to assess LD between two multiallelic markers such as the HLA alleles[4]. For each pair of HLA genes, we calculated ε to quantify LD between the HLA genes, by using the observed frequency of the four-digit HLA alleles. Because the estimation of ε can be biased when the haplotype frequency distribution is sparse, we combined the HLA alleles with frequency <0.01 into a single dummy allele. The value of ε ranges between 0 and 1, and a higher ε value represents stronger LD.

Machine-learning-based clustering by using HLA allele information. We performed unsupervised clustering of haplotypes with NGS-based HLA typing data by using tSNE, a machine-learning method for high-dimensionality compression and visualization[35,36]. tSNE is usually used to classify cells by using single-cell transcriptome or immunoprofiling data (cytometry by time of flight)[45,46], and in this study we applied tSNE to classify haplotypes to obtain unbiased classification patterns based on HLA allele information[47]. We conducted tSNE for phased haplotype data of HLA alleles separately for classical or nonclassical HLA genes and for each digit by using the Rtsne R package (version 0.13). On the basis of the two components obtained from the tSNE results (tSNE_1 and tSNE_2), we conducted unsupervised clustering by adopting the DBSCAN R package (version 1.1.1)[37]. We first determined the following parameters to optimize the average silhouette width score by using the four-digit classical HLA alleles: a perplexity value = 25, a minimum number of reachable points = 3, and a reachable epsilon neighborhood parameter = 8.62. We fixed the perplexity value and the minimum number of reachable points, and then determined the reachable epsilon neighborhood parameters for two-digit classical, six-digit classical and four-digit nonclassical HLA alleles separately to optimize the average silhouette width score (10.0, 8.96 and 8.94, respectively). Parsimonious connections of the clusters were constructed with the alluvial R package (version 0.1–2).

Construction of population-specific NGS-based HLA imputation reference panel. For individuals with NGS-based HLA typing data, we obtained high-density SNP data of the MHC region by genotyping with the Illumina HumanCoreExome BeadChip (v1.1; Illumina). We applied stringent quality control (QC) filters as previously described[12,55]. Briefly, we applied QC filters to the individuals (call rates >0.99, exclusion of outliers by principal component analysis, exclusion of closely related individuals) and then applied QC filters to the SNPs (call rates ≥0.99, MAF ≥0.01, Hardy–Weinberg-equilibrium P value ≥ 1.0×10^{-7}). We extracted the genotyped SNPs in the entire MHC region (24–36 Mb on chromosome 6, NCBI Build 37). In addition to the HLA alleles typed by NGS (two-digit, four-digit and six-digit), we incorporated HLA gene amino acid polymorphisms corresponing to the four-digit HLA alleles according to the IPD-IMGT/HLA database[5]. We encoded both HLA alleles and HLA amino acid polymorphisms, and constructed the NGS-based HLA imputation reference panel of the Japanese population together with SNP genotype data with SNP2HLA software (version 1.0.3; $n = 1{,}120$ for the 33 HLA genes)[9].

The imputation accuracy of the constructed HLA imputation reference panel was empirically evaluated by a cross-validation approach[12]. We randomly split the panel into two data sets ($n = 560$ for each data set). HLA alleles from one of the data sets were masked and then imputed by using another data set as an imputation reference. The concordance between imputed and genotyped HLA allele dosages was calculated separately for each HLA gene and each allele digit. To relatively compare the imputation accuracy among the different reference panels, we evaluated accuracy in the previously reported HLA imputation reference panel of independent Japanese individuals in the same way ($n = 908$)[4,12].

HLA and SNV imputation of GWAS data of BBJ individuals. Using the constructed NGS-based HLA imputation reference panel, we imputed the HLA variants of the large-scale GWAS data of the BBJ individuals ($n = 166{,}190$). Detailed characteristics of the GWAS data and the QC process are described elsewhere[21,22]. Although we usually use SNP2HLA software for HLA imputation because of the high imputation accuracy and ability to impute HLA amino acid polymorphisms[9,56], SNP2HLA is currently not applicable to such large-scale GWAS data, owing to a very large requirement of memory resources. Therefore, we initially used SNP2HLA to align SNP-strand and position information between the GWAS data and the reference panel, and then imputed the HLA variants with standard genome-wide imputation software. Specifically, we phased the GWAS data with Eagle (version 2.3) and imputed the variants with minimac3 (version 2.0.1). In addition, we densely imputed SNV and indels within the entire MHC region by using the deep-WGS data of the Japanese population as a reference ($n = 1{,}276$, average depth = 24.6×, sequenced on the Illumina HiSeq2500 platform (Illumina))[30]. For the PheWAS, we applied stringent postimputation QC filtering of the variants (MAF ≥0.5% and imputation score Rsq ≥0.7).

PheWAS of HLA variants by using imputed BBJ GWAS data. PheWAS was conducted by using clinical information of the individuals included in the imputed BBJ GWAS data. Associations of the imputed variants in the MHC region with 106 phenotype datasets (46 diseases and 60 quantitative traits; Supplementary Table 5) were examined. The diseases comprise four major categories (immune related

($n=10$), metabolic and cardiovascular ($n=10$), cancers ($n=13$) and other diseases ($n=13$)). The quantitative traits comprised ten major categories (anthropometric ($n=2$), metabolic ($n=6$), protein ($n=4$), kidney related ($n=4$), electrolyte ($n=5$), liver related ($n=6$), other biochemical ($n=6$), hematological ($n=13$), blood pressure ($n=4$) and echocardiographic ($n=10$)). Definitions of the diseases and the process of patient registration have been described elsewhere[21,22]. For the controls in disease association studies, we constructed a shared control group by excluding individuals affected by diseases known to have associations in the MHC region. Detailed processes of outlier exclusion, adjustment with clinical status and normalization methods of the quantitative traits have been described elsewhere[25,26].

We evaluated associations of the HLA variants with the risk of the diseases, by using a logistic regression model, and with dosage effects on the normalized values of the quantitative traits, by using a linear regression model[18], with a glm() function implemented in R statistical software (version 3.2.3). We defined the HLA variants as biallelic SNVs in the entire MHC region (24–36 Mbp at chromosome 6, NCBI build 37), two-digit, four-digit and six-digit biallelic alleles of the HLA genes, biallelic HLA amino acid polymorphisms corresponding to the respective residues and multiallelic HLA amino acid polymorphisms for each amino acid position. We assumed additive effects of the allele dosages on phenotypes in the regression models. We included the top ten principal components obtained from the GWAS genotype data (not including the MHC region) as covariates in the regression models to correct potential population stratification. An omnibus P value for each HLA amino acid position was obtained by a log likelihood-ratio test for the likelihood between the null model and the fitted model, followed by a χ^2 distribution with $m-1$ degree(s) of freedom for an amino acid position with m residues. To evaluate the nonadditive effects of the HLA alleles, we conducted a multivariate regression analysis that additionally included nonadditive genotype dosages of the HLA alleles as previously described[18,57]. We adopted a genome-wide-significance threshold of $P<5.0\times10^{-8}$ in our study[39].

Assignments of the candidate responsible genes to the top-associated variants of the phenotypes in the nominal and conditional analyses were conducted in the following manner: (i) when the variant was in moderate LD with any of the HLA alleles or amino acid polymorphisms ($r^2\geq0.5$), or located in the coding region of the HLA gene, the HLA gene was assigned; (ii) when the variant was in LD with the coding variants of the non-HLA gene, the non-HLA gene was assigned; and (iii) when the variant was located in an intergenic region, the nearest gene was assigned. Considering the strong functional effects of the HLA gene polymorphisms on human phenotypes, our assignment protocol puts relatively higher weights on HLA genes than on non-HLA genes. We note that r^2 values (that is, correlation of haplotypes) between the imputed dosages were approximately estimated by calculating Pearson's correlation of genotype dosages (R^2).

Conditional-association analysis of HLA variants. To evaluate independent risk among variants (and genes), we conducted a forward-type stepwise conditional regression analysis for phenotypes that satisfied the genome-wide-significance threshold. In each conditional step, we additionally included the associated variants as covariates in the regression model and repeated the analysis until no variants satisfied the significance threshold. When the top-associated variant itself was the HLA gene polymorphism or the SNV and indel in strong LD with any of the HLA gene polymorphisms ($r^2\geq0.7$), we additionally included all the two-digit, four-digit and six-digit alleles and the amino acid polymorphisms of the corresponding HLA gene as covariates in the regression to robustly condition the associations attributable to the HLA gene, as previously described[4,18]. Otherwise, the top-associated SNV and indels were additionally included as the associated variants.

Heritability estimates of the variants within the MHC region. We estimated the heritability of the phenotypes in the PheWAS that was explained by the variants within the entire MHC region, as well as calculating pairwise genetic correlations among the phenotypes. We adopted a Haseman–Elston regression implemented in GCTA software (version 1.91.1beta)[58], because a genomic restricted maximum-likelihood method, a typical method for estimating SNP-based heritability[59], was difficult to apply to the large sample size of our study. The estimated heritability of the diseases was adjusted according to disease prevalence in the Japanese population (Supplementary Table 5)[59]. In addition to the heritability estimation using all the MHC variants, we repeated the analysis separately for classical HLA variants (using the polymorphisms of classical HLA genes) and other variants (using the MHC variants not in LD with any of the classical HLA variants ($r^2<0.1$)), and quantified their relative proportions. Standard errors (s.e.) of the proportions were estimated by simulating the distribution of the proportion values according to random sampling from the mean and s.e.m. of the heritability estimates (×100,000 iterations). Although there have been discussions on how to precisely estimate heritability within a genetic locus with strong LD[60], because our main focus was on relative comparison of heritability across traits rather than quantification of absolute heritability values, we adopted GCTA as a standard method, as previously applied[61].

Using the matrix of pairwise genetic correlations among the phenotypes, we constructed a network of phenotypes representing shared genetic backgrounds of MHC across the phenotypes. We assigned each phenotype to a node, and the nodes were connected by edges weighted according to the magnitude of the corresponding genetic correlation. To effectively extract biological information embedded in the network and to avoid dense visualization, we used only highly significant genetic correlations (top 10% of the significance in the phenotype pairs and $P<0.05$ after adjustment of Bonferroni's correction). Network visualization was conducted according to the Fruchterman–Reingold algorithm, with the igraph R package (version 1.1.2).

Statistical analyses. Two-tailed logistic and linear regression was applied by using a glm() function implemented in R statistical software (version 3.2.3).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

Software and codes used for this study are available from URLs or upon request to the authors.

Data availability

HLA data have been deposited at the National Bioscience Database Center (NBDC) Human Database (research ID: hum0114) as open data without any access restrictions. GWAS data and phenotype data of the BBJ individuals are available at the NBDC Human Database (research ID: hum0014).

References

49. Szolek, A. et al. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
50. Shukla, S. A. et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
51. Kawaguchi, S. et al. HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).
52. Lee, H. & Kingsford, C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome. Biol.* **19**, 16 (2018).
53. Nothnagel, M. & Rohde, K. The effect of single-nucleotide polymorphism marker selection on patterns of haplotype blocks and haplotype frequency estimates. *Am. J. Hum. Genet.* **77**, 988–998 (2005).
54. Okada, Y. eLD: entropy-based linkage disequilibrium index between multi-allelic sites. *Hum. Genome Var.* **5**, 29 (2018).
55. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
56. Karnes, J. H. et al. Comparison of HLA allelic imputation programs. *PLoS One* **12**, e0172444 (2017).
57. Lenz, T. L. et al. Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat. Genet.* **47**, 1085–1090 (2015).
58. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
59. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
60. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
61. Hinks, A. et al. Fine-mapping the MHC locus in juvenile idiopathic arthritis (JIA) reveals genetic heterogeneity corresponding to distinct adult inflammatory arthritic diseases. *Ann. Rheum. Dis.* **76**, 765–772 (2017).

nature research

Corresponding author(s): Yukinori Okada

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	We used publicly available software for the data analysis (Omixon Target software version 1.9.3, BWA version 0.7.15, GATK version 3.6, OptiType version 1.3.1, POLYSOLVER version 4, HLA-HD version 1.2.0.1, Kourami version 0.9.6, eLD version 1, Rtsne R package version 0.13, DBSCAN R package version 1.1.1, Alluvial R package version 0.1-2, SNP2HLA version version 1.0.3, Eagle version 2.3, Minimac3 version 2.0.1, Rtsne R package version 3.2.3, GCTA version 1.91.1beta, Igraph R package version 1.1.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

HLA data is deposited at the National Bioscience Database Center (NBDC) Human Database (Research ID: hum0114) as open data without any access restrictions. GWAS data of the BBJ subjects is available at the NBDC Human Database (Research ID: hum0014).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Regarding the samples used for HLA sequencing, we used publicly available DNA obtained from cell lines of unrelated Japanese individuals established by the Japan Biological Informatics Consortium (JBIC, n = 1120). Regarding the samples used in the GWAS and PheWAS analysis, the BioBank Japan project has recruited roughly 200,000 participants. Clinical informations extracted from medical records and DNA / serum samples were also collected. Most of them were also genotyped by genome-wide SNP genotyping array, and relatively a small fraction of them were whole-genome-sequenced. For both, we selected samples as many as possible if they have available genotype or phenotype data.

Data exclusions

We excluded the samples and variants based on the standard quality control procedure in GWAS. Detailed information on quality controls were sufficiently described in our manuscript. The exclusion criteria were pre-established in the field of GWAS.

Replication

We used all the available data in the study, and did not conduct the two-staged discovery and replication studies.

Randomization

We did not apply randomization. All the samples with available accessibility to genotype and phenotype data were included in the analysis.

Blinding

We did not apply blinding of the samples because no intervention was conducted in our study.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging