A global atlas of genetic associations of 220 deep phenotypes

Saori Sakaue[1-5, 46, *], Masahiro Kanai[1, 5-9, 46], Yosuke Tanigawa[10], Juha Karjalainen[5-7,9], Mitja Kurki[5-7,9], Seizo Koshiba[11,12], Akira Narita[11], Takahiro Konuma[1], Kenichi Yamamoto[1,13], Masato Akiyama[2,14], Kazuyoshi Ishigaki[2-5], Akari Suzuki[15], Ken Suzuki[1], Wataru Obara[16], Ken Yamaji[17], Kazuhisa Takahashi[18], Satoshi Asai[19,20], Yasuo Takahashi[21], Takao Suzuki[22], Nobuaki Shinozaki[22], Hiroki Yamaguchi[23], Shiro Minami[24], Shigeo Murayama[25], Kozo Yoshimori[26], Satoshi Nagayama[27], Daisuke Obata[28], Masahiko Higashiyama[29], Akihide Masumoto[30], Yukihiro Koretsune[31], FinnGen, Kaoru Ito[32], Chikashi Terao[2], Toshimasa Yamauchi[33], Issei Komuro[34], Takashi Kadowaki[33], Gen Tamiya[11,12,35,36], Masayuki Yamamoto[11,12,35], Yusuke Nakamura[37,38], Michiaki Kubo[39], Yoshinori Murakami[40], Kazuhiko Yamamoto[15], Yoichiro Kamatani[2,41], Aarno Palotie[5,9,42], Manuel A. Rivas[10], Mark J. Daly[5-7,9], Koichi Matsuda[43, *], Yukinori Okada[1,2,41,44,45, *]

1. Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan.
2. Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
3. Center for Data Sciences, Harvard Medical School, Boston, MA, USA
4. Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
5. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA
6. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
7. Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA
8. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
9. Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
10. Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, CA, USA
11. Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan
12. The Advanced Research Center for Innovations in Next-Generation Medicine (INGEM), Sendai, Japan

13. Department of Pediatrics, Osaka University Graduate School of Medicine, Suita, Japan.
14. Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan
15. Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
16. Department of Urology, Iwate Medical University, Iwate, Japan
17. Department of Internal Medicine and Rheumatology, Juntendo University Graduate School of Medicine, Tokyo, Japan
18. Department of Respiratory Medicine, Juntendo University Graduate School of Medicine, Tokyo, Japan
19. Division of Pharmacology, Department of Biomedical Science, Nihon University School of Medicine, Tokyo, Japan
20. Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan
21. Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan
22. Tokushukai Group, Tokyo, Japan
23. Department of Hematology, Nippon Medical School, Tokyo, Japan
24. Department of Bioregulation, Nippon Medical School, Kawasaki, Japan
25. Tokyo Metropolitan Geriatric Hospital and Institute of Gerontology, Tokyo, Japan
26. Fukujuji Hospital, Japan Anti-Tuberculosis Association, Tokyo, Japan
27. The Cancer Institute Hospital of the Japanese Foundation for Cancer Research, Tokyo, Japan
28. Center for Clinical Research and Advanced Medicine, Shiga University of Medical Science, Otsu, Japan
29. Department of General Thoracic Surgery, Osaka International Cancer Institute, Osaka, Japan
30. Aso Iizuka Hospital, Fukuoka, Japan
31. National Hospital Organization Osaka National Hospital, Osaka, Japan
32. Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
33. Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
34. Department of Cardiovascular Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

35. Graduate School of Medicine, Tohoku University, Sendai, Japan
36. Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan
37. Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan
38. Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan
39. RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
40. Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan
41. Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan
42. Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Analytic and Translational Genetics Unit, Department of Medicine, and the Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
43. Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, the University of Tokyo, Tokyo, Japan
44. Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan
45. Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan
46. These authors contributed equally: S Sakaue and M Kanai.

* Corresponding authors

Saori Sakaue, M.D., Ph.D.	Koichi Matsuda, M.D., Ph.D.	Yukinori Okada, M.D., Ph.D.
ssakaue@bwh.harvard.edu	kmatsuda@edu.k.u-tokyo.ac.jp	yokada@sg.med.osaka-u.ac.jp
Center for Data Sciences, Harvard Medical School	Department of Computational Biology and Medical Sciences, Graduate school of Frontier Sciences, The University of Tokyo	Department of Statistical Genetics, Osaka University Graduate School of Medicine

Abstract

The current genome-wide association studies (GWASs) do not yet capture sufficient diversity in terms of populations and scope of phenotypes. To address an essential need to expand an atlas of genetic associations in non-European populations, we conducted 220 deep-phenotype GWASs (disease endpoints, biomarkers, and medication usage) in BioBank Japan (n = 179,000), by incorporating past medical history and text-mining results of electronic medical records. Meta-analyses with the harmonized phenotypes in the UK Biobank and FinnGen (n_{total} = 628,000) identified over 4,000 novel loci, which substantially deepened the resolution of the genomic map of human traits, benefited from East Asian endemic diseases and East Asian specific variants. This atlas elucidated the globally shared landscape of pleiotropy as represented by the MHC locus, where we conducted fine-mapping by HLA imputation. Finally, to intensify the value of deep-phenotype GWASs, we performed statistical decomposition of matrices of phenome-wide summary statistics, and identified the latent genetic components, which pinpointed the responsible variants and shared biological mechanisms underlying current disease classifications across populations. The decomposed components enabled genetically informed subtyping of similar diseases (e.g., allergic diseases). Our study suggests a potential avenue for hypothesis-free re-investigation of human disease classifications through genetics.

Main

Introduction

Medical diagnosis has been shaped through the description of organ dysfunctions and extraction of shared key symptoms, which categorizes a group of individuals into a specific disease to provide an optimal treatment. The earliest physicians in ancient Egypt empirically made disease diagnoses based on clinical symptoms, palpitation, and auscultation (~2600 BC)[1]. Since then, continuous efforts by physicians have sophisticated the disease classifications through empirical categorization. An increased understanding of organ functions and the availability of diagnostic tests including biomarkers and imaging techniques have further contributed to the current disease classifications, such as ICD10[2] and phecode[3].

In the past decades, genome-wide association studies (GWASs) have provided new insights into the biological basis underlying disease diagnoses. The genetic underpinnings enable us to re-interrogate the validity of historically- defined disease classifications. To this end, a comprehensive catalog of disease genetics is warranted. However, current genetic studies still lack the comprehensiveness in three ways; (i) population, in that the vast majority of GWASs have been predominated by European populations[4], (ii) scope of phenotypes, which have been limited to target diseases of a sampling cohort, and (iii) a systematic method to interpret a plethora of summary results for understanding disease pathogenesis and epidemiology. We thus need to promote equity in genetic studies by sharing the results of genetic studies of deep phenotypes from diverse populations.

To expand the atlas of genetic associations, here we conducted 220 deep-phenotype GWASs in BioBank Japan project (BBJ), including 108 novel phenotypes in East Asian

populations. We then conducted GWASs for corresponding harmonized phenotypes in UK Biobank (UKB) and FinnGen, and finally performed trans-ethnic meta-analyses (n_{total} = 628,000). The association results elucidated trans-ethnically shared landscape of the pleiotropy and genetic correlations across diseases. Furthermore, we applied DeGAs[5] to perform truncated singular-value decomposition (TSVD) on the matrix of GWAS summary statistics of 159 diseases each in Japanese and European ancestries, and derived latent components shared across the diseases. We interpreted the derived components by (i) functional annotation of the genetic variants explaining the component, (ii) identification of important cell types in which the genes contributing to the component are specifically regulated, and (iii) projection of GWASs of biomarkers or metabolomes into the component space. The latent components recapitulated the hierarchy of current disease classifications, while different diseases sometimes converged on the same component which implicated the shared biological pathway and relevant tissues. We also classified a group of similar diseases (e.g., allergic diseases) into subgroups based on these components. Analogous to the conventional hierarchical classification of diseases based on the shared symptoms, an atlas of genetic studies resolved the shared latent structure behind human diseases, which elucidated the genetic variants, genes, organs, and biological functions underlying human diseases.

Results

GWAS of 220 traits in BBJ and trans-ethnic meta-analysis

Overview of this study is presented in **Extended Data Figure 1**. BBJ is a nationwide biobank in Japan, and recruited participants based on the diagnosis of at least one of 47 target diseases (**Supplementary Note**)[6]. Along with the target disease status, deep phenotype data, such as past medical history (PMH), drug prescription records (~ 7 million), text data retrieved from electronic medical records (EMR), and biomarkers, have been collected. Beyond the collection of case samples based on the pre-determined target diseases, the PMH and EMR have provided broader insights into disease genetics, as shown in recently launched biobanks such as UKB[7] and BioVU[8]. In this context, we curated the PMH, performed text-mining of the EMR, and merged them with 47 target disease status.[9] We created individual-level phenotype on 159 disease endpoints (38 target diseases with median 1.25 times increase in case samples and 121 novel disease endpoints) and 23 categories of medication usage. We then systematically mapped the disease endpoints into phecode and ICD10, to enable harmonized GWASs in UKB and FinnGen. We also analyzed a quantitative phenotype of 38 biomarkers in BBJ, of which individual phenotype data are available in UKB[10]. Using genotypes imputed with the 1000 Genome Project phase 3 data (n = 2,504) and population-specific whole-genome sequencing data (n = 1,037) as a reference panel[11], we conducted the GWASs of 159 binary disease endpoints, 38 biomarkers, and 23 medication usages in ~179,000 individuals in BBJ (**Figure 1a–c**, **Supplementary Table 1** and **2** for phenotype summary). To maximize the statistical power, we used a linear mixed model implemented in SAIGE[12] (for binary traits) and BOLT[13] (for quantitative traits). By using linkage disequilibrium (LD)-score regression[14],

we confirmed that the confounding biases were controlled in the GWASs (**Supplementary Table 3**). In this expanded scope of GWASs in the Japanese population, we identified 396 genome-wide significant loci across 159 disease endpoints, 1891 across 38 biomarkers, of which 92 and 156 loci were novel, respectively ($P < 1.0 \times 10^{-8}$; see **Methods**, **Supplementary Table 4**). We conducted the initial medication-usage GWASs in East Asian populations, and detected 171 genome-wide significant loci across 23 traits (see **Methods**). These signals underscore the value of (i) conducting GWASs in non-Europeans and (ii) expanding scope of phenotypes by incorporating biobank resources such as PMH and EMR. For example, we detected an East Asian-specific variant, rs140780894, at the MHC locus in pulmonary tuberculosis (PTB; Odds Ratio [OR] = 1.2, $P = 2.9 \times 10^{-23}$, Minor Allele Frequency[MAF]$_{EAS}$ = 0.24; **Extended Data Figure 2**), which was not present in European population (Minor Allele Count [MAC]$_{EUR}$ = 0)[15]. PTB is a serious global health burden and relatively endemic in Japan[16] (annual incidence per 100,000 was 14 in Japan whereas 8 in the United Kingdom and 3 in the United States in 2018 [World Health Organization, Global Tuberculosis Report]). Because PTB, an infectious disease, can be treatable and remittable, we substantially increased the number of cases by combining the participants with PMH of PTB to the patients with active PTB at the time of recruitment (from 549[9] to 7,800 case individuals). Similar to this example, we identified a novel signal at rs190894416 at 7p14.2 (OR = 16, $P = 6.9 \times 10^{-9}$; **Extended Data Figure 3**) in dysentery, which is bloody diarrhea caused by infection with *Shigella bacillus* that was once endemic in Japan when poor hygiene had been common[17]. We also identified novel signals in common diseases that have not been target diseases but were included in the PMH record, such as rs715 at 3'UTR

of *CPS1* in cholelithiasis (**Extended Data Figure 4**; OR = 0.87, $P = 9.6\times10^{-13}$) and rs2976397 at the *PSCA* locus in gastric ulcer, gastric cancer , and gastric polyp (**Extended Data Figure 5**; OR = 0.86, $P = 6.1\times10^{-24}$). We detected pleiotropic functionally impactful variants, such as a deleterious missense variant, rs28362459 (p.Leu20Arg), in *FUT3* associated with gallbladder polyp (OR = 1.46, $P = 5.1\times10^{-11}$) and cholelithiasis (OR = 1.11, $P = 7.3\times10^{-9}$; **Extended Data Figure 6**), and a splice donor variant, rs56043070 (c.89+1G>A), causing loss of function of *GCSAML* associated with urticaria (OR = 1.24, $P = 6.9\times10^{-12}$; **Extended Data Figure 7**), which was previously reported to be associated with platelet and reticulocyte counts[18]. Medication-usage GWASs also provided interesting signals as an alternative perspective for understanding disease genetics[19]. For example, individuals taking HMG CoA reductase inhibitors (C10AA in Anatomical Therapeutic Chemical Classification [ATC]) were likely to harbor variations at *HMGCR* (lead variant at rs4704210, OR = 1.11, $P = 2.0\times10^{-27}$). Prescription of salicylic acids and derivatives (N02BA in ATC) were significantly associated with a rare East Asian missense variant in *PCSK9*, rs151193009 (p.Arg93Cys; OR = 0.75, $P = 7.1\times10^{-11}$, MAF_{EAS}=0.0089, MAF_{EUR}=0.000; **Extended Data Figure 8**), which might indicate a strong protective effect against the thromboembolic diseases in general.

To confirm that the signals identified in BBJ were replicable, we conducted GWASs of corresponding phenotypes (i.e., disease endpoints and biomarkers) in UKB and FinnGen, and collected summary statistics of medication usage GWAS recently conducted in UKB[19] (**Supplementary Table 5**). We then compared the effect sizes of the genome-wide significant variants in BBJ with those in a European dataset across binary and quantitative

traits (see **Methods**). The loci identified in our GWASs were successfully replicated in the same effect direction (1,830 out of 1,929 [94.9%], $P < 10^{-325}$ in sign test) and with high effect-size correlation (**Extended Data Figure 9**).

Motivated by the high replicability, we performed trans-ethnic meta-analyses of these 220 harmonized phenotypes across three biobanks (see **Methods**). We identified 1,362 disease-associated, 10,572 biomarker-associated, and 841 medication-associated loci in total, of which 356, 3,576, and 236 were novel, respectively (**Figure 1d**, **Supplementary Table 6**). All these summary results of GWASs are openly shared without any restrictions. Together, we successfully expanded the genomic map of human complex traits in terms of populations and scope of phenotypes through conducting deep-phenotype GWASs across trans-ethnic nationwide biobanks.

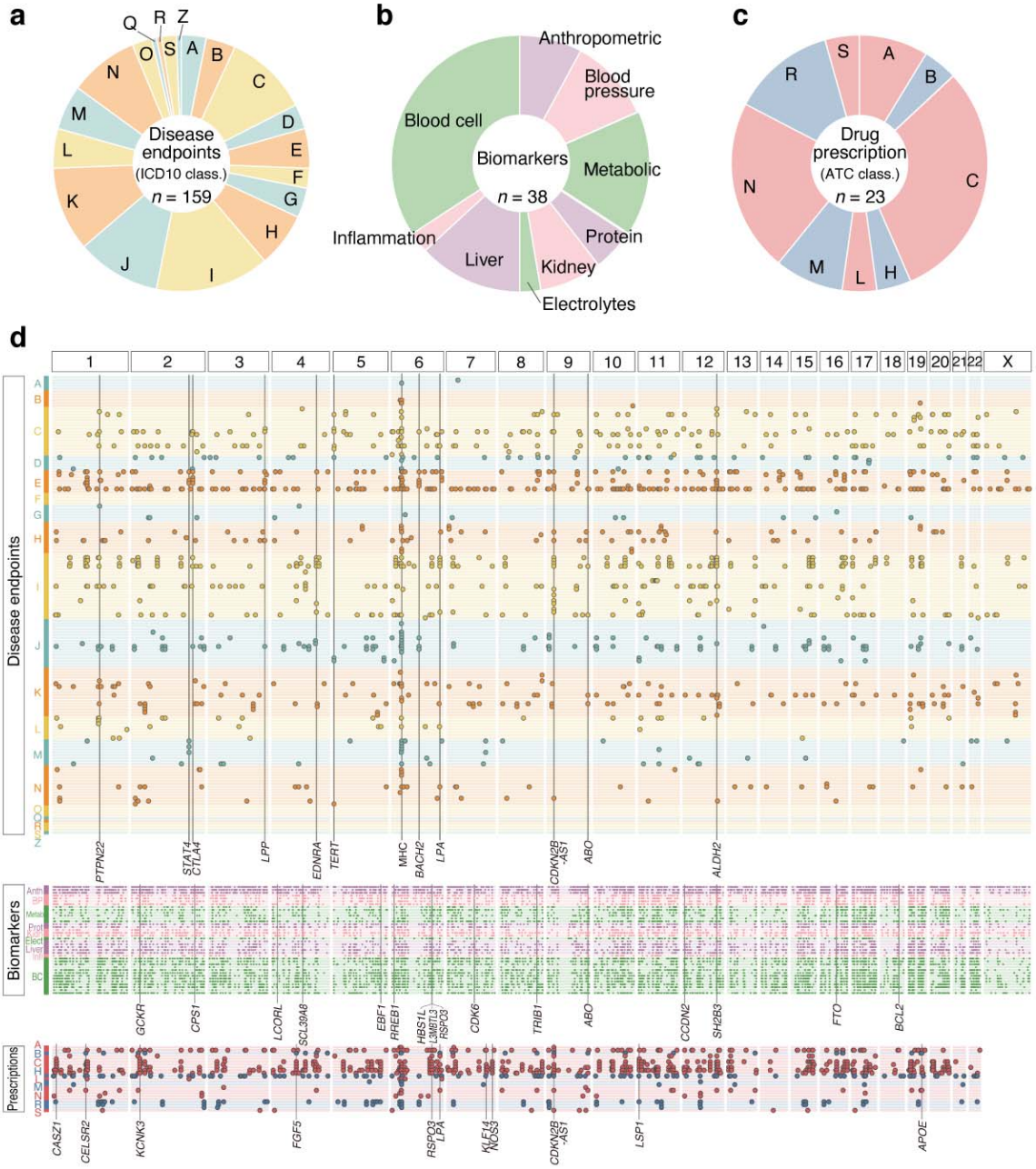

Figure 1. Overview of the identified loci in the trans-ethnic meta-analyses of 220 deep phenotype GWASs.

(**a**-**c**) The pie charts describe the phenotypes analyzed in this study. The disease endpoints (**a**; n_{trait} = 159) were categorized based on the ICD10 classifications (A to Z; **Supplementary Table 1a**), the biomarkers (**b**; n_{trait} = 38; **Supplementary Table 1b**) were classified into nine categories, and medication usage was categorized based on the ATC system (A to S; **Supplementary Table 1c**). (**d**) The genome-wide significant loci identified in the

trans-ethnic meta-analyses and pleiotropic loci ($P < 1.0\times10^{-8}$). The traits (rows) are sorted as shown in the pie chart, and each dot represents significant loci in each trait. Pleiotropic loci are annotated by lines with a locus symbol.

The regional landscape of pleiotropy.

Because human traits are highly polygenic and the observed variations within the human genome are finite in number, pleiotropy, where a single variant affects multiple traits, is pervasive[20]. While pleiotropy has been intensively studied in European populations by compiling previous GWASs[20,21], the landscape of pleiotropy in non-European populations has remained elusive. By leveraging this opportunity for comparing the genetics of deep phenotypes across populations, we sought to investigate the landscape of regional pleiotropy in both Japanese and European populations. We defined the degree of pleiotropy as the number of significant associations per variant ($P < 1.0\times10^{-8}$)[21]. In the Japanese, rs11066015 harbored the largest number of genome-wide significant associations (45 traits; **Figure 2a**), which was in tight LD with a missense variant at the *ALDH2* locus, rs671. Following this, rs117326768 at the MHC locus (23 traits) and rs1260326 at the *GCKR* locus (18 traits) were most pleiotropic. In Europeans, rs3132941 at the MHC locus harbored the largest number of genome-wide significant associations (46 traits; **Figure 2b**), followed by rs4766578 at the *ATXN2*/*SH2B3* locus (38 traits) and rs4665972 at the *GCKR* locus (28 traits). Notably, the *ALDH2* locus (pleiotropic in Japanese) and the MHC locus (pleiotropic in Japanese and Europeans) are known to be under recent positive selection[22,23]. To systematically assess whether pleiotropic regions in the genome were likely to be under selection pressure in each of the populations, we investigated the enrichment of the signatures of recent positive selection quantified by the metric singleton density score (SDS)[22] values within the pleiotropic loci (see **Methods**). Intriguingly, when compared with those under the null hypothesis, we observed significantly higher values of SDS χ^2 values within the pleiotropic loci, and this fold change increased as the number of associations

increased (i.e., more pleiotropic) in both Japanese and Europeans (**Figure 2c** and **2d**). To summarize, the trans-ethnic atlas of genetic associations elucidated the broadly shared landscape of pleiotropy, which implied a potential connection to natural selection signatures affecting human populations.

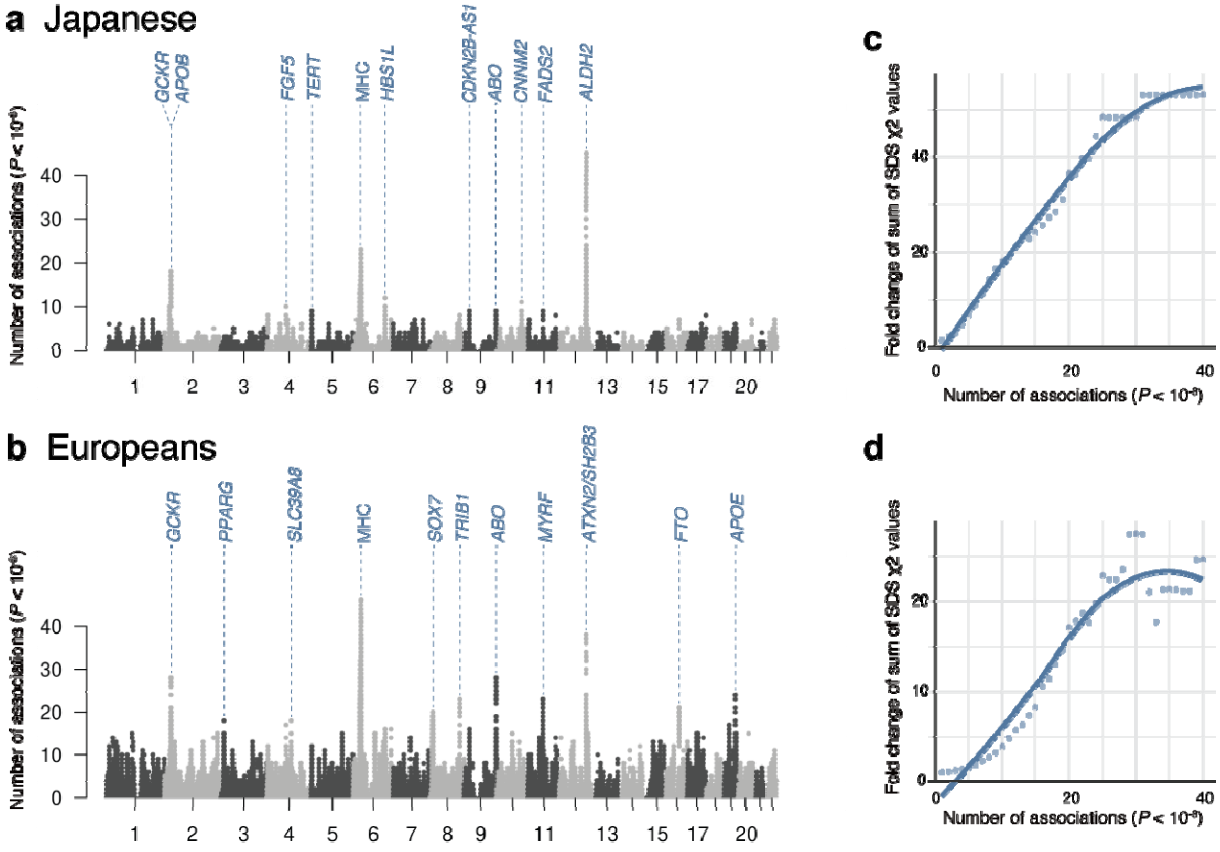

Figure 2. Number of significant associations per variant.

(**a**, **b**) The Manhattan-like plots show the number of significant associations ($P < 1\times10^{-8}$) at each tested genetic variant for all traits (n_{trait} = 220) in Japanese (**a**) and in European GWASs (**b**). Loci with a large number of associations were annotated based on the closest genes of each variant. (**c**, **d**) The plots indicate the fold change of the sum of SDS χ^2 within variants with a larger number of significant associations than a given number on the x-axis compared with those under the null hypothesis in Japanese (**c**) and in Europeans (**d**). We also illustrated a regression line based on local polynomial regression fitting.

Pleiotropic associations in HLA and ABO locus.

Given the strikingly high number of associations in both populations, we next sought to fine-map the pleiotropic signals within the MHC locus. To this end, we imputed the classical HLA alleles in BBJ and UKB, and performed association tests for 159 disease endpoints and 38 biomarkers (**Figure 3a** and **3b**). After the fine-mapping and conditional analyses (see **Methods**), we identified 94 and 153 independent association signals in BBJ and UKB, respectively (the regional threshold of significance was set to $P < 1.0\times10^{-6}$; **Supplementary Table 7**). Overall, *HLA-B* in class I and *HLA-DRB1* in class II harbored the largest number of associations in both BBJ and UKB. For example, we successfully fine-mapped the strong signal associated with PTB to HLA-DRβ1 Ser57 (OR = 1.20, $P = 7.1\times10^{-19}$) in BBJ. This is the third line of evidence showing the robust association of HLA with tuberculosis identified to date[24,25], and we initially fine-mapped the signal to *HLA-DRB1*. Interestingly, HLA-DRβ1 at position 57 also showed pleiotropic associations with other autoimmune and thyroid-related diseases, such as Grave's disease (GD), hyperthyroidism, Hashimoto's disease, hypothyroidism, Sjogren's disease, chronic hepatitis B, and atopic dermatitis in BBJ. Of note, the effect direction of the association of HLA-DRβ1 Ser57 was the same between hyperthyroid status (OR = 1.29, $P = 2.6\times10^{-14}$ in GD and OR = 1.37, $P = 1.4\times10^{-8}$ in hyperthyroidism) and hypothyroid status (OR = 1.50, $P = 9.0\times10^{-8}$ in Hashimoto's disease and OR = 1.31, $P = 1.5\times10^{-7}$ in hypothyroidism), despite the opposite direction of thyroid hormone abnormality. This association of HLA-DRβ1 was also observed in Sjogren's syndrome (OR = 2.04, $P = 7.9 \times 10^{-12}$), which might underlie the epidemiological

comorbidities of these diseases[26]. Other novel associations in BBJ included HLA-DRβ1 Asn197 with sarcoidosis (OR = 2.07, $P = 3.7\times10^{-8}$), and four independent signals with chronic sinusitis (i.e., *HLA-DRA*, *HLA-B*, *HLA-A*, and *HLA-DQA1*).

Another representative pleiotropic locus in the human genome is the *ABO* locus. We performed ABO blood-type PheWAS in BBJ and UKB (**Figure 3c** and **3d**). We estimated the ABO blood type from three variants (rs8176747, rs8176746, and rs8176719 at 9q34.2)[27], and associated them with the risk of diseases and quantitative traits for each blood group. A variety of phenotypes, including common diseases such as myocardial infarction as well as biomarkers such as blood cell traits and lipids, were strongly associated with the blood types in both biobanks (**Supplementary Table 8**). We also replicated an increased risk of gastric cancer in blood-type A as well as an increased risk of gastric ulcer in blood-type O in BBJ[28].

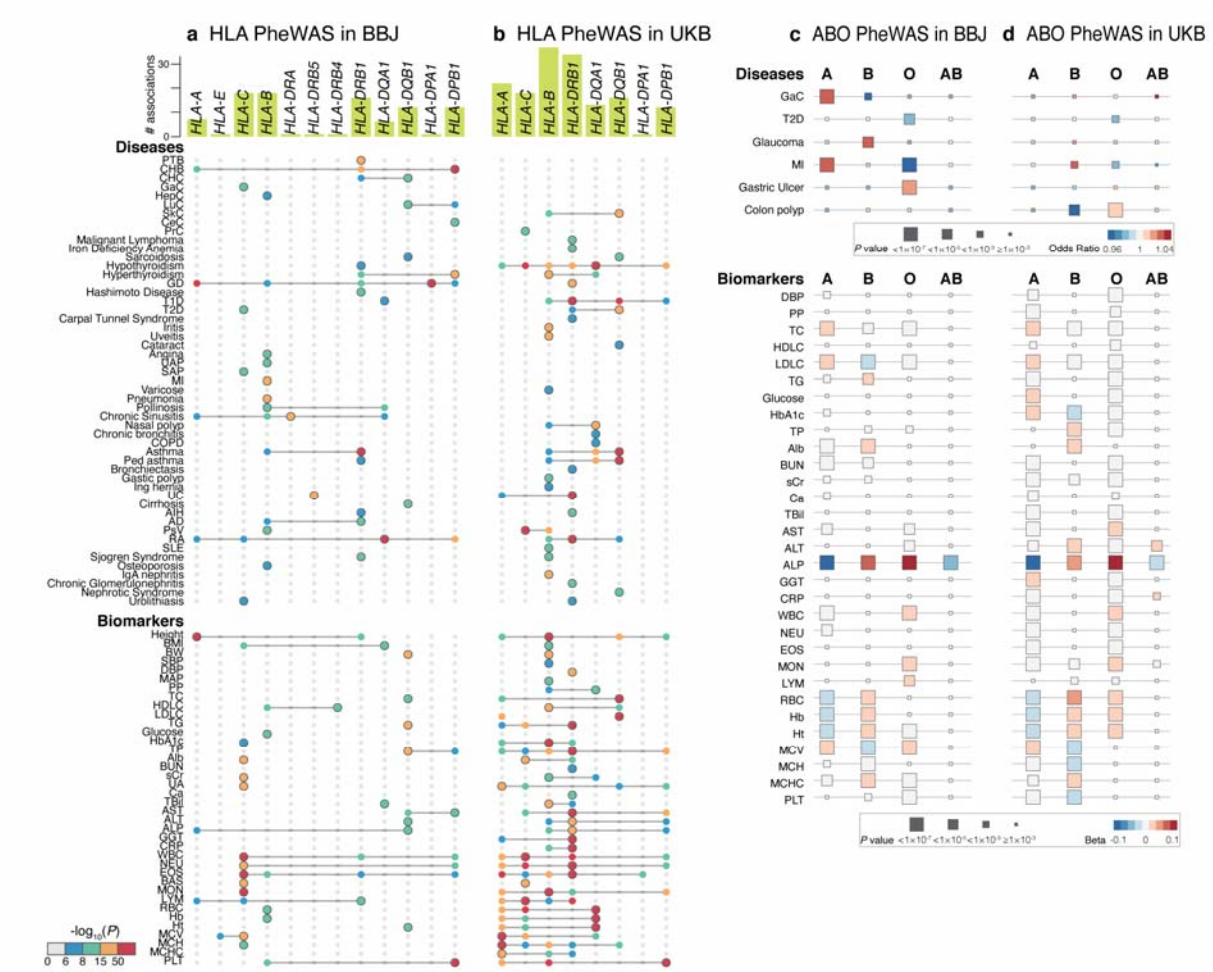

Figure 3. HLA and ABO association PheWAS.

(**a**,**b**) Significantly associated HLA genes identified by HLA PheWAS in BBJ (**a**) or in UKB (**b**) are plotted. In addition to the top association signals of the phenotypes, independent associations identified by conditional analysis are also plotted, and the primary association signal is indicated by the plots with a gray border. The color of each plot indicates two-tailed *P* values calculated with logistic regression (for binary traits) or linear regression (for quantitative traits) as designated in the color bar at the bottom. The bars in green at the top indicate the number of significant associations per gene in each of the populations. The detailed allelic or amino acid position as well as statistics in the association are provided in **Supplementary Table 7**.

(**c**,**d**) Significant associations identified by ABO blood-type PheWAS in BBJ (**c**) or in UKB (**d**) are shown as boxes and colored based on the odds ratio. The size of each box indicates two-tailed *P* values calculated with logistic regression (for binary traits) or linear regression (for quantitative traits).

Genetic correlation elucidates the shared phenotypic domains across populations.

The interplay between polygenicity and pleiotropy suggests widespread genetic correlations among complex human traits[29]. Genetic relationships among human diseases have contributed to the refinement of disease classifications[30] and elucidation of the biology underlying the epidemiological comorbidity[29]. To obtain deeper insights into the interconnections among human traits and compare them across populations, we computed pairwise genetic correlations (r_g) across 106 traits (in Japanese) and 148 traits (in Europeans) with Z-score for $h^2_{SNP} > 2$, using bivariate LD score regression (see **Methods**). We then defined the correlated trait domains by greedily searching for the phenotype blocks with pairwise $r_g > 0.7$ within 70% of r_g values in the block on the hierarchically clustered matrix of pairwise r_g values (**Extended Data Figure 10**). We detected domains of tightly correlated phenotypes, such as (i) cardiovascular- acting medications, (ii) coronary artery disease, (iii) type 2 diabetes- related phenotypes, (iv) allergy- related phenotypes, and (v) blood-cell phenotypes in BBJ (**Extended Data Figure 10a**). These domains implicated the shared genetic backgrounds on the similar diseases and their treatments (e.g., (ii) diseases of the circulatory system in ICD10 and for coronary artery disease and their treatments) and diagnostic biomarkers (e.g., (iii) glucose and HbA1c in type 2 diabetes). Intriguingly, the corresponding trait domains were mostly identified in UKB as well (**Extended Data Figure 10b**). Thus, we confirmed that the current clinical boundaries for a spectrum of human diseases broadly reflect the shared genetic etiology across populations, despite differences in ethnicity and despite potential differences in diagnostic and prescription practices.

Deconvolution of a matrix of summary statistics of 159 diseases provides novel insights into disease pathogenesis.

A major challenge in genetic correlation is that the r_g is a scholar value between two traits, which summarizes the averaged correlation over the whole genome into just one metric[31]. This approach is not straightforward in specifying a set of genetic variants driving the observed correlation, which should pinpoint biological pathways and dysfunctional organs explaining the shared pathogenesis. To address this, gathering of the genetic association statistics of hundreds of different phenotypes can dissect genotype-phenotype association patterns without a prior hypothesis, and identify latent structures underlying a spectrum of complex human traits. In particular, matrix decomposition on the summary statistics is a promising approach[5,32,33], which derives orthogonal components that explain association variance across multiple traits while accounting for linear genetic architectures in general. This decomposition can address two challenges in current genetic correlation studies. First, it informs us of genetic variants that explain the shared structure across multiple diseases, thereby enabling functional interpretation of the component. Second, it can highlight sub-significant associations and less powered studies, which are important in understanding the contribution of common variants in rare disease genetics with a small number of case samples[32] or in genetic studies in underrepresented populations where smaller statistical power is inevitable.

Therefore, we applied DeGAs[5] on a matrix of our disease GWAS summary statistics in Japanese and the meta-analyzed statistics in Europeans ($n_{disease}$ = 159; **Figure 4a** and **4b**). To interpret the derived latent components, we annotated the genetic variants explaining each component (i) through GREAT genomic region ontology enrichment analysis[34], (ii)

through identification of relevant cell types implicated from tissue specific regulatory DNA (ENCODE3[35]) and expression (GTEx[36]) profiles, and (iii) by projecting biomarker GWASs and metabolome GWASs into the component space ($n_{biomarker}$=38, $n_{metabolite_EAS}$=206, $n_{metabolite_EUR}$=248; **Figure 4a**). We applied TSVD on the sparse Z score matrix of 22,980 variants, 159 phenotypes each in 2 populations (Japanese and Europeans), and derived 40 components that together explained 36.7% of the variance in the input summary statistics matrix (**Extended Data Figure 11, 12**).

Globally, hierarchically similar diseases as defined by the conventional ICD10 classification were explained by the same components, based on DeGAs trait squared cosine scores that quantifies component loadings[5] (**Figure 4c**, **d**). This would be considered as a hypothesis-free support of the historically defined disease classification. For example, component 1 explained the genetic association patterns of diabetes (E10 and E11 in ICD10) and component 2 explained those of cardiac and vascular diseases (I00-I83), in both populations. Functional annotation enrichment of the genetic variants explaining these components by GREAT showed that component 1 (diabetes component) was associated with abnormal pancreas size (binomial $P_{enrichment}$ =7.7×10^{-19}) as a human phenotype, whereas component 2 (cardiovascular disease component) was associated with xanthelasma (i.e., cholesterol accumulation on the eyelids; binomial $P_{enrichment}$ =3.0×10^{-10}). Further, the genes comprising component 1 were enriched in genes specifically expressed in the pancreas ($P_{enrichment}$ =5.5×10^{-4}), and those comprising component 2 were enriched in genes specifically expressed in the aorta ($P_{enrichment}$ =1.9×10^{-3}; **Extended Data Figure 13**). By projecting the biomarker and metabolite GWASs into this component space, we

observed that component 1 represented the genetics of glucose and HbA1c, and component 2 represented the genetics of blood pressure and lipids, all of which underscored the biological relevance. Thus, this deconvolution-projection analysis elucidated the latent genetic structure behind human diseases, which highlighted the underlying biological functions, relevant tissues, and associated human phenotypes.

The latent components shared across diseases explained the common biology behind etiologically similar diseases. For example, we identified that component 10 explained the genetics of cholelithiasis (gall stone), cholecystitis (inflammation of gallbladder), and gall bladder polyp (**Figure 4e**). The projection of European metabolite GWASs into the component space identified that component 10 represented the metabolite GWAS in the bilirubin metabolism pathway. Component 10 was composed of variants involved in intestinal cholesterol absorption in the mouse phenotype (binomial $P_{\text{enrichment}} = 3.8\times10^{-10}$). This is biologically relevant, since increased absorption of intestinal cholesterol is a major cause of cholelithiasis, which also causes cholecystitis[37]. This projection analysis was also applicable to the Japanese metabolites GWASs, which showed the connection between the component 1 (diabetes component) and arginine and glucose levels, and between the component 10 (gallbladder disease component) and glycine, which conjugates with bile acids[38].

Some components could be further utilized to boost understanding of the underpowered GWASs with the use of well-powered GWAS, and for identifying the contributor of shared genetics between different diseases. For example, we complemented underpowered varicose GWAS in BBJ (n_{case} = 474, genome-wide significant loci = 0) with higher-powered GWAS in Europeans (n_{case} = 22,037, genome-wide significant loci = 54), since both GWASs

were mostly represented by component 11, which was explained by variants related to abnormal vascular development (binomial $P_{enrichment} = 4.2\times10^{-7}$; **Figure 4f**). Another example is component 27, which was shared with rheumatoid arthritis and systemic lupus erythematosus, two distinct but representative autoimmune diseases. Component 27 was explained by the variants associated with interleukin secretion and plasma cell number (binomial $P_{enrichment} = 6.1\times10^{-10}$ and 9.3×10^{-10}, respectively), and significantly enriched in the DNase I hypersensitive site (DHS) signature of lymphoid tissue ($P_{enrichment} =1.3\times10^{-4}$; **Figure 4g**). This might suggest the convergent etiology of the two autoimmune diseases, which could not be elucidated by the genetic correlation alone.

Finally, we aimed at hypothesis-free categorization of diseases based on these components. Historically, hypersensitivity reactions have been classified into four types (e.g., types I to IV)[39], but the clear sub-categorization of allergic diseases based on this pathogenesis and whether the categorization can be achieved solely by genetics were unknown. In our TSVD results, the allergic diseases (mostly J and L in ICD10) were represented by the four components 3, 16, 26, and 34. By combining these components as axis-1 (e.g., components 3 and 16) and axis-2 (e.g., components 26 and 34), and comparing the cumulative variance explained by these axes, we defined *axis-1 dominant allergic diseases* (e.g., asthma and allergic rhinitis) and *axis-2 dominant allergic diseases* (metal allergy, contact dermatitis, and atopic dermatitis; **Figure 4h**). Intriguingly, the *axis-1 dominant diseases* corresponded etiologically well to type I allergy (i.e., immediate hypersensitivity). The variants explaining *axis-1* were biologically related to IgE secretion and Th_2 cells (binomial $P_{enrichment} = 9.9\times10^{-46}$ and 2.9×10^{-44}, respectively). Furthermore,

GWAS of eosinophil count was projected onto *axis-1*, which recapitulated the biology of type I allergy[40]. In contrast, the *axis-2 dominant diseases* corresponded to type IV allergy (i.e., cell-mediated delayed hypersensitivity). The variants explaining *axis-2* were associated with IL-13 and interferon secretion (binomial $P_{\text{enrichment}} = 1.6\times10^{-10}$ and 5.2×10^{-9}, respectively), and GWAS of C-reactive protein was projected onto *axis-2*, which was distinct from *axis-1*[41]. To summarize, our deconvolution approach (i) recapitulated the existing disease classifications, (ii) clarified the underlying biological mechanisms and relevant tissues shared among a spectrum of related diseases, and (iii) showed potential application for genetics-driven categorization of human diseases.

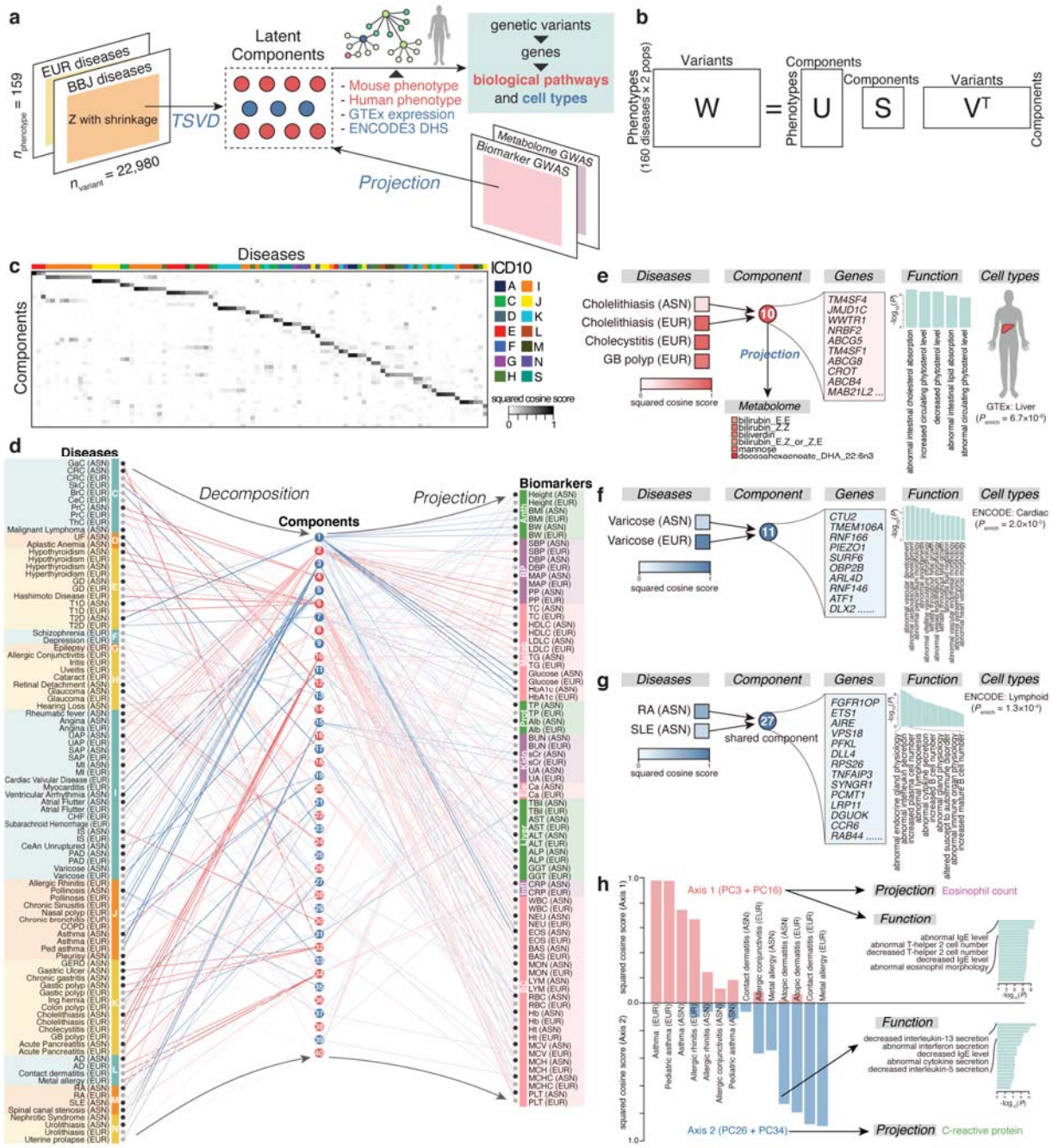

Figure 4. The deconvolution analysis of a matrix of summary statistics of 159 diseases across populations.

(**a**) An illustrative overview of deconvolution-projection analysis. Using DeGAs framework, a matrix of summary statistics from two populations (EUR: European and BBJ: Biobank Japan) was decomposed into latent components, which were interpreted by annotation of a

set of genetic variants driving each component and in the context of other GWASs through projection. (**b**) A schematic representation of TSVD applied to decompose a summary statistic matrix **W** to derive latent components. **U**, **S**, and **V** represent resulting matrices of singular values (**S**) and singular vectors (**U** and **V**). (**c**) A heatmap representation of DeGAs squared cosine scores of diseases (columns) to components (rows). The components are shown from 1 (top) to 40 (bottom), and diseases are sorted based on the contribution of each component to the disease measured by the squared cosine score (from component 1 to 40). Full results with disease and component labels are in **Extended Data Figure 14**. (**d**) Results of TSVD of disease genetics matrix and the projection of biomarker genetics. Diseases (left) and biomarkers (right) are colored based on the ICD10 classification and functional categorization, respectively. The derived components (middle; from 1 to 40) are colored alternately in blue or red. The squared cosine score of each disease to each component and each biomarker to each component is shown as red and blue lines. The width of the lines indicates the degree of contribution. The diseases with squared cosine score > 0.3 in at least one component are displayed. Anth; anthropometry, BP; blood pressure, Metab; metabolic, Prot; protein, Kidn; kidney-related, Ele; Electrolytes, Liver; liver-related, Infl; Inflammatory, BC; blood cell. (**e**-**h**) Examples of disease-component correspondence and the biological interpretation of the components by projection and enrichment analysis using GREAT. A representative component explaining a group of diseases based on the contribution score, along with responsible genes, functional enrichment results GREAT, relevant tissues, and relevant biobarkers/metabolites is shown. GB; gallbladder. RA; rheumatoid arthritis. SLE; systemic lupus erythematosus.

Discussion

Here, we performed 220 GWASs of human complex traits by incorporating the PMH and EMR data in BBJ, substantially expanding the atlas of genotype-phenotype associations in non-Europeans. We then systematically compared their genetic basis with GWASs of corresponding phenotypes in Europeans. We confirmed the global replication of loci identified in BBJ, and discovered 4,170 novel loci through trans-ethnic meta-analyses, highlighting the value of conducting GWASs in diverse populations. The results are openly shared through web resources, which will be a platform to accelerate further research such as functional follow-up studies and drug discovery[42]. Of note, leveraging these well-powered GWASs, we observed that the genes associated with endocrine/metabolic, circulatory, and respiratory diseases (E, I, and J by ICD10) were systematically enriched in targets of approved medications treating those diseases (**Extended Data Figure 15**). This should motivate us to use this expanded resource for genetics-driven novel drug discovery and drug repositioning.

The landscape of regional pleiotropy was globally shared across populations, and pleiotropic regions tended to have been under recent positive selection. Further elucidation of pleiotropy in other populations is warranted to replicate our results. To highlight the utility of deep phenotype GWASs, we finally decomposed the multi-ethnic genotype–phenotype association patterns by TSVD. The latent components derived from TSVD pinpointed the convergent biological mechanisms and relevant cell types across diseases, which can be utilized for re-evaluation of existing disease classifications. The incorporation of biomarker and metabolome GWAS summary statistics enabled further interpretation of the latent components. Our approach suggested a potential avenue for restructuring of the medical

diagnoses through dissecting the shared genetic basis across a spectrum of diseases, as analogous to the current disease diagnostics historically shaped through empirical categorization of shared key symptoms across a spectrum of organ dysfunctions.

In conclusion, our study substantially expanded the atlas of genetic associations, supported the historically-defined categories of human diseases, and should accelerate the discovery of the biological basis contributing to complex human diseases.

Acknowledgments

We sincerely thank all the participants of BioBank Japan, UK Biobank, and FinnGen. 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). The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and nine industry partners (AbbVie, AstraZeneca, Biogen, Celgene, Genentech, GSK, MSD, Pfizer and Sanofi). Following biobanks are acknowledged for collecting the FinnGen project samples: Auria Biobank (https://www.auria.fi/biopankki/), THL Biobank (https://thl.fi/fi/web/thl-biopank), Helsinki Biobank (https://www.terveyskyla.fi/helsinginbiopankki/), Northern Finland Biobank Borealis (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki), Finnish Clinical Biobank Tampere (https://www.tays.fi/biopankki), Biobank of Eastern Finland (https://ita-suomenbiopankki.fi), Central Finland Biobank (https://www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (https://www.bloodservice.fi/Research%20Projects/biobanking),

Terveystalo Biobank Finland (https://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). S.S. was in part supported by The Mochida Memorial Foundation for Medical and Pharmaceutical Research．M.Kanai was supported by a Nakajima Foundation Fellowship and the Masason Foundation. Y.Tanigawa is in part supported by a Funai Overseas Scholarship from the Funai Foundation for Information Technology and the Stanford University School of Medicine. M.A.R. is in part supported by National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under award R01HG010140 (M.A.R.), and a National Institute of Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Y.O. was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (19H01021, 20K21834), and AMED (JP20km0405211, JP20ek0109413, JP20ek0410075, JP20gm4010006, and JP20km0405217), Takeda Science Foundation, and Bioinformatics Initiative of Osaka University Graduate School of Medicine, Osaka University.

Author Contributions

S.S., M. Kanai, and Y.O. conceived the study. S.S., M. Kanai, Y. Tanigawa., M.A.R., and Y.O. wrote the manuscript. S.S., M. Kanai, J.K., M. Kurki, T.Konuma, Kenichi Yamamoto, M.A., K.Ishigaki, Kazuhiko Yamamoto, Y. Kamatani, A.P., M.J.D., and Y.O. conducted GWAS data studies. S.S., Y. Tanigawa., and M.A.R. conducted statistical decomposition analysis. S.S., S.T., A.N., G.T., and Y.O. conducted metabolome analysis. A.S., K.S., W.O.,

Ken Yamaji, K.T., S.A., Y.Takahashi, T.S., N.S., H.Y., S.Minami, S.Murayama, Kozo Yoshimori, S.N., D.O., M.H., A.M., Y.Koretsune, K.Ito, C.T., T.Y., I.K., T.Kadowaki, M.Y., Y.N., M.Kubo, Y.M., Kazuhiko Yamamoto, and K.M. collected and managed samples and data. A.P. and M.J.D. coordinated collaboration with FinnGen.

Competing Financial Interests

M.A.R. is on the SAB of 54Gene and Computational Advisory Board for Goldfinch Bio and has advised BioMarin, Third Rock Ventures, MazeTx and Related Sciences. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The genotype data of BBJ used in this study are available from the Japanese Genotype-phenotype Archive (JGA; http://trace.ddbj.nig.ac.jp/jga/index_e.html) with accession code JGAD00000000123 and JGAS00000000114. The UKB analysis was conducted via application number 47821. This study used the FinnGen release 3 data. Summary statistics of BBJ GWAS and trans-ethnic meta-analysis will be publicly available without any restrictions.

Code availability

We used publicly available software for the analyses. The used software is listed and described in the **Method** section of our manuscript.

Methods

Genome-wide association study of 220 traits in BBJ

We conducted 220 deep phenotype GWASs in BBJ. BBJ is a prospective biobank that collaboratively collected DNA and serum samples from 12 medical institutions in Japan and recruited approximately 200,000 participants, mainly of Japanese ancestry (**Supplementary Note**). All study participants had been diagnosed with one or more of 47 target diseases by physicians at the cooperating hospitals. We previously conducted GWASs of 42 out of the 47 target diseases[9]. In this study, we newly curated the PMH records included in the clinical data, and performed text-mining to retrieve disease records from the free-format EMR as well. For disease phenotyping, we merged this information with the target disease status, and defined the case status for 159 diseases with a case count > 50 (**Supplementary Table 2**). As controls, we used samples in the cohort without a given diagnosis or related diagnoses, which was systematically defined by using the phecode framework[3] (**Supplementary Table 1**). For medication-usage phenotyping, we again retrieved information by text-mining of 7,018,972 medication records. Then, we categorized each medication trade name by using the ATC, World Health Organization. For biomarker phenotyping, we used the same processing and quality control method as previously described (**Supplementary Table 2** for phenotype summary)[10,43]. In brief, we excluded measurements outside three times of interquartile range (IQR) of upper/lower quartile. For individuals taking anti-hypertensive medications, we added 15 mmHg to systolic blood pressure (SBP) and 10 mmHg to diastolic blood pressure (DBP). For individuals taking a statin, we applied the following correction to the lipid measurements: i) Total cholesterol was divided by 0.8; ii) measured LDL-cholesterol (LDLC) was adjusted as LDLC / 0.7; iii) derived

LDLC from the Friedewald was re-derived as (Total cholesterol / 0.8) - HDLC - (Triglyceride/ 5).

We genotyped participants with the Illumina HumanOmniExpressExome BeadChip or a combination of the Illumina HumanOmniExpress and HumanExome BeadChips. Quality control of participants and genotypes was performed as described elsewhere[11]. In this project, we analyzed 178,726 participants of Japanese ancestry as determined by the principal component analysis (PCA)-based sample selection criteria. The genotype data were further imputed with 1000 Genomes Project Phase 3 version 5 genotype (n = 2,504) and Japanese whole-genome sequencing data (n = 1,037) using Minimac3 software. After this imputation, we excluded variants with an imputation quality of Rsq < 0.7.

We conducted GWASs for binary traits (i.e., disease endpoints and medication usage) by using a generalized linear mixed model implemented in SAIGE (version 0.37), which had substantial advantages in terms of (i) maximizing the sample size by including genetically related participants, and (ii) controlling for case–control imbalance[12] , which was the case in many of the disease endpoints in this study. We included adjustments for age, age^2, sex, age×sex, age^2×sex, and top 20 principal components for as covariates used in step 1. For sex-specific diseases, we alternatively adjusted for age, age^2, and the top 20 principal components as covariates used in step 1, and we used only controls of the sex to which the disease is specific. For the X chromosome, we conducted GWASs separately for males and females, and merged their results by inverse-variance fixed-effects meta-analysis[44]. We conducted GWASs for quantitative traits (i.e., biomarkers) by using a linear mixed model implemented in BOLT-LMM (version 2.3.4). We included the same covariates as used in the binary traits above.

All the participants provided written informed consent approved from ethics committees of the Institute of Medical Sciences, the University of Tokyo and RIKEN Center for Integrative Medical Sciences.

Harmonized genome-wide association study of 220 traits in UKB and FinnGen

We conducted the GWASs harmonized with BBJ in UKB and in FinnGen. The UK Biobank project is a population-based prospective cohort that recruited approximately 500,000 people across the United Kingdom (**Supplementary Note**). We defined case and control status of 159 disease endpoints, which were originally retrieved from the clinical information in UKB and mapped to BBJ phenotypes via phecode (**Supplementary Table 1**). We also analyzed 38 biomarker values provided by the UKB. The genotyping was performed using either the Applied Biosystems UK BiLEVE Axiom Array or the Applied Biosystems UK Biobank Axiom Array. The genotypes were further imputed using a combination of the Haplotype Reference Consortium, UK10K, and 1000 Genomes Phase 3 reference panels by IMPUTE4 software[7]. In this study, we analyzed 361,194 individuals of white British genetic ancestry as determined by the PCA-based sample selection criteria (see **URLs**). We excluded the variants with (i) INFO score ≤ 0.8, (ii) MAF ≤ 0.0001 (except for missense and protein-truncating variants annotated by VEP[45], which were excluded if MAF ≤ 1×10^{-6}), and (iii) $P_{HWE} \leq 1 \times 10^{-10}$. We conducted GWASs for 159 disease endpoints by using SAIGE with the same covariates used in the BBJ GWAS. For biomarker GWASs, we used publicly available summary statistics of UKB biomarker GWAS when available (see **URLs**), and otherwise performed linear regression using PLINK software with the same covariates, excluding the genetically related individuals (the 1st, 2nd, or 3rd degree)[7]. For medication

usage GWASs, we used publicly available summary statistics of medication usage in UKB[19], which was organized by the ATC and thus could be harmonized with BBJ GWASs.

FinnGen is a public–private partnership project combining genotype data from Finnish biobanks and digital health record data from Finnish health registries (**Supplementary Notes**). For GWASs, we used the summary statistics of FinnGen release 3 data (see **URLs**). The disease endpoints were mapped to BBJ phenotypes by using ICD10 code, and we defined 129 out of 159 endpoints in BBJ. We did not conduct biomarker and medication-related GWASs because the availability of these phenotypes was limited.

Meta-analysis, definition of significant loci, and annotation of the lead variants with genome-wide significance

First, we performed intra-European meta-analysis when summary statistics of both UKB and FinnGen were available, and then performed trans-ethnic meta-analysis across three or two cohorts in 159 disease endpoints, 38 biomarker values, and 23 medication usage GWASs. We conducted these meta-analyses by using the inverse-variance method and estimated heterogeneity with Cochran's Q test with metal software[44]. The summary statistics of primary GWASs in BBJ and trans-ethnic meta-analysis GWASs are openly shared without any restrictions.

We adopted the genome-wide significance threshold of $< 1.0\times10^{-8}$, as previously used in similar projects in BBJ and UKB[9,21]. We defined independent genome-wide significant loci on the basis of genomic positions within ±500 kb from the lead variant. We considered a trait-associated locus as novel when the locus within ±1 Mb from the lead variant did not include any variants that were previously reported to be significantly associated with the

same disease. We basically searched for previous reports of known loci in the GWAS catalog[18], but also referred to PubMed or preprints when the corresponding trait was not included in GWAS catalog or when the large-scale GWASs were released in the preprint server as of July 2020 (**Supplementary Table 9**).

We annotated the lead variants using ANNOVAR software, such as rsIDs in dbSNP database (see **URLs**), the genomic region and closest genes, and functional consequences. We also supplemented this with the gnomAD database[15], and also looked for the allele frequencies in global populations as an independent resource.

Replication of significant associations in BBJ

For 2,287 lead variants in the genome-wide significant loci of 159 disease endpoints and 38 biomarkers in BBJ, we compared the effect sizes and directions with European-only meta-analysis when available and with UKB-based summary statistics otherwise. Of them, 1,929 variants could be compared with the corresponding European GWASs. Thus, we performed the Pearson's correlation test for these variants' beta in the association test in BBJ and in European GWAS. We also performed the correlation tests with variants with $P_{EUR} < 0.05$ and to those with $P_{EUR} < 1.0\times10^{-8}$.

Evaluation of regional pleiotropy

We assessed the regional pleiotropy based on each tested genetic variant separately for BBJ GWASs and for European GWASs (i.e., intra- European meta-analysis when FinnGen GWAS was available and UKB summary statistics otherwise). We quantified the degree of pleiotropy per genetic variant by aggregating and counting the number of genome-wide

significant associations across 220 traits. We then annotated loci from the largest number of associations ($n_{associations}$ > 9 in BBJ and > 18 in Europeans) in **Figure 2a**, **b**.

Next, we assessed the recent natural selection signature within the pleiotropic loci separately for Japanese and for Europeans. To do this, we first defined the pleiotropic loci by identifying genetic variants that harbored a larger number of significant associations than a given threshold. We varied this threshold from 1 to 40. Then, at each threshold, we calculated the sum of SDS χ^2 values within the pleiotropic loci, and compared this with the χ^2 distribution under the null hypothesis with a degree of freedom equal to the number of variants in the loci. We thus estimated the SDS enrichment within the pleiotropic loci defined by a given threshold as fold change and *P* value. The SDS values were obtained from the web resource indicated in the original article on Europeans (see **URLs**) and provided by the authors on Japanese[23]. The raw SDS values were normalized according to the derived allele frequency as described previously.

Fine-mapping of HLA and ABO loci

We performed the fine-mapping of MHC associations in BBJ and UKB by HLA imputation[46]. In BBJ, we imputed classical HLA alleles and corresponding amino acid sequences using the reference panel recently constructed from 1,120 individuals of Japanese ancestry by the combination of SNP2HLA software, Eagle, and minimac3 , as described previously[47]. We applied post-imputation quality control to keep the imputed variants with minor allele frequency (MAF) ≥ 0.5% and Rsq > 0.7. For each marker dosage that indicated the presence or absence of an investigated HLA allele or an amino acid sequence, we performed an association test with the disease endpoints and biomarkers. We assumed

additive effects of the allele dosages on phenotypes in the regression models. We included the same covariates as in the GWAS. In UKB, we imputed classical HLA alleles and corresponding amino acid sequences using the T1DGC reference panel of European ancestry (n = 5,225)[48]. We applied the same post-imputation quality control and performed the association tests as in BBJ.

Heritability and genetic correlation estimation

We performed LD score regression (see **URLs**) for GWASs of BBJ and Europeans to estimate SNP-based heritability, potential bias, and pairwise genetic correlations. Variants in the MHC region (chromosome 6:25–34 Mb) were excluded. We also excluded variants with $\chi^2 > 80$, as recommended previously[49]. For heritability estimation, we used the baselineLD model (version 2.2), which included 97 annotations that correct for bias in heritability estimates[50]. We note that we did not report liability-scale heritability, since population prevalence of 159 diseases in each country was not always available, and the main objective of this analysis was an assessment of bias in GWAS, rather than the accurate estimation of heritability. We calculated the heritability Z-score to assess the reliability of heritability estimation, and reported the LDSC results with Z-score for h^2_{SNP} is > 2 (**Supplementary Table 3**). For calculating pairwise genetic correlation, we again restricted the target GWASs to those whose Z-score for h^2_{SNP} is > 2, as recommended previously[49]. In total, we calculated genetic correlation for 106 GWASs in BBJ and 148 in European GWASs, which resulted in 5,565 and 10,878 trait pairs, respectively.

To illustrate trait-by-trait genetic correlation, we hierarchically clustered the r_g values with hclust and colored them as a heatmap (**Extended Data Figure 10**). To adopt reliable

genetic correlations, we restricted the r_g values that had $P_{cor} < 0.05$. Otherwise, the r_g values were replaced with 0. We then defined the tightly clustered trait domains by greedily searching for the phenotype blocks with pairwise $r_g > 0.7$ within 70% of r_g values in the block from the top left of the clustered correlation matrix. We manually annotated each trait domain by extracting the characteristics of traits constituting the domain (**Extended Data Figure 10**).

Deconvolution of a matrix of summary statistics by TSVD

We performed the TSVD on the matrix of genotype-phenotype association Z scores as described previously as DeGAs framework[5]. In this study, we first focused on 159 disease endpoint GWASs in BBJ and European GWAS (i.e., 318 in total) to derive latent components through TSVD. On constructing a Z-score matrix, we conducted variant-level QC. We removed variants located in the MHC region (chromosome 6: 25–34 Mb), and replaced unreliable Z-score estimates with zero when one of the following conditions were satisfied:

- *P* value of marginal association ≥ 0.001
- Standard error of beta value ≥ 0.2

Considering that rows and columns with all zeros do not contribute to matrix decomposition, we excluded variants that had all zero Z-scores across 159 traits in either in BBJ or Europeans. We then performed LD pruning using PLINK software[51] ("--indep-pairwise 50 5 0.1") with an LD reference of 5,000 randomly selected individuals of white British UKB participants to select LD-independent variant sets, which resulted in a total of 22,980 variants. Thus, we made a Z-score matrix (= **W**) with a size of 318 (*N*: 159 diseases × 2

populations) × 22,980 (*M*: variants). With a predetermined number of *K*, TSVD decomposed **W** into a product of three matrices: **U**, **S**, and $\mathbf{V}^{\mathbf{T}}$: $\mathbf{W} = \mathbf{USV}^{\mathbf{T}}$. $\mathbf{U} = (u_{i,k})_{i,k}$ is an orthonormal matrix of size $N \times K$ whose columns are phenotype singular vectors, **S** is a diagonal matrix of size $K \times K$ whose elements are singular values, and $\mathbf{V} = (v_{j,k})_{j,k}$ is an orthonormal matrix of size $M \times K$ whose columns are variant singular vectors. Here we set *K* as 40, which together explained 36.7% of the total variance of the original matrix. This value was determined by experimenting with different values from 20 to 100 and selecting the informative and sufficient threshold. We used the TruncatedSVD module in the sklearn.decomposition library of python for performing TSVD.

To interpret and visualize the results of TSVD, we calculated the squared cosine scores. The phenotype squared cosine score, $cos_i^{2^{phe}}(k)$, is a metric to quantify the relative importance of the *k*th latent component for a given phenotype *i*, and is defined as follows;

$$cos_i^{2^{phe}}(k) = \frac{\left(f_{i,k}^{p}\right)^2}{\sum_{k'}\left(f_{i,k'}^{p}\right)^2}$$

where

$$\mathbf{F}_p = \mathbf{US} = \left(f_{i,k}^{p}\right)_{i,k}.$$

Annotation of the components by using GREAT and identification of relevant cell types

We calculated the variant contribution score, which is a metric to quantify the contribution of a given variant *j* to a given component *k* as follows;

$$contr_k^{var}(j) = \left(v_{i,k}\right)^2$$

For each component, we can thus rank the variants based on their contribution to the component and calculate the cumulative contribution score. We defined a set of *contributing variants* to a given component to include top-ranked variants that had high contribution scores until the cumulative contribution score to the component exceeded 0.5. For these variant sets contributing to the latent components, we performed the GREAT (version 4.0.4) binomial genomic region enrichment analysis[34] based on the size of the regulatory domain of genes and quantified the significance of enrichment in terms of binomial fold enrichment and binomial *P* value to biologically interpret these components. We used the human phenotype and mouse genome informatics phenotype ontology, which contains manually curated knowledge about the hierarchical structure of phenotypes and genotype-phenotype mapping of human and mouse, respectively. The enriched annotation with a false discovery rate (FDR) < 0.05 is considered significant and displayed in the figures.

For a gene set associated with the contributing variants with a given component ($P<0.05$), we sought to identify relevant cell types by integrating two datasets: (i) ENCODE3 DHS regulatory patterns across human tissues from non-negative matrix factorization (NFM)[35] and (ii) specifically expressed genes defined from GTEx data[36]. In brief, a vocabulary (i.e., DHS patterns) for regulatory patterns was defined from the NFM of 3 million DHSs × 733 human biosamples encompassing 438 cell and tissue types. Then, for each regulatory vocabulary, GENCODE genes were assigned based on their overlying DHSs. The gene labeling result was downloaded from the journal website[35]. We also defined genes specifically expressed in 53 tissues from GTEx version 7 data, based on the top 5% of the *t*-statistics in each tissue as described elsewhere[52]. Then, for (i) each regulatory vocabulary

and (ii) each tissue, we performed Fisher's exact tests to investigate whether the genes associated with a given component are significantly enriched in the defined gene set.

Projection of biomarker and metabolite GWASs into the component space

To further help interpret the latent components derived from disease-based TSVD, we projected the Z-score matrix of biomarker GWASs and metabolite GWASs into the component space. Briefly, we constructed the Z-score matrices (**W'**) of 38 biomarkers of BBJ and European GWASs (i.e., 76 rows) and 248 known metabolites of independent previous GWASs in the European population[53] × 22,980 variants (**Supplementary Table 10**). Then, using the **V** from the disease-based TSVD, we calculated the phenotype contribution as follows;

$$\mathbf{F}_p^{projection} = \mathbf{W}'\mathbf{V} = \left(f_{i,k}^{projection}\right)_{i,k}$$

We note that for metabolite GWASs, since the GWASs were imputed with the HapMap reference panel, we imputed Z-scores of missing variants using ssimp software[54] (version 0.5.5 --ref 1KG/EUR --impute.maf 0.01), and otherwise we set the missing Z-scores to zero.

Projection of Metabolite GWASs in Japanese into the component space

To investigate whether the projection analysis is applicable to independent dataset, we conducted metabolite GWASs in Tohoku Medical Megabank Organization (ToMMo). ToMMo is a community-based biobank that combines medical and genome information from the participants in the Tohoku region of Japan[55]. Detailed cohort description is presented in **Supplementary Notes**. In this study, we analyzed a total of 206 metabolites[56] measured by proton nuclear magnetic resonance (NMR) or liquid chromatography (LC)–MS

(**Supplementary Table 11**). For sample QC, we excluded samples meeting any of the following criteria: (1) genotype call rate < 95%, (2) one individual from each pair of those in close genetic relation (PI_HAT calculated by PLINK[51] ≥ 0.1875) based on call rate, and (3) outliers from Japanese ancestry cluster based on the principal component analysis with samples of 1KGP phase 3 data. For phenotype QC, we excluded (1) the measurements in pregnant women, (2) those which took time from sampling to biobanking ≥ 2 days, and (3) phenotypic outlier defined as log-transformed measurements laying more than 4 SD from the mean for each metabolite. The participants were genotyped with a custom SNP array for the Japanese population (i.e., Japonica Array v2). For genotype QC, we excluded variants meeting any of the following criteria: (1) call rate < 98%, (2) *P* value for Hardy–Weinberg equilibrium < 1.0×10^{-6}, and (3) MAF < 0.01. The QCed genotype data were pre-phased by using SHAPEIT2 software (r837), and imputed by using IMPUTE4 software (r300.3) with a combined reference panel of 1KGP phase3 (*n* = 2,504) and population specific WGS data (i.e., 3.5KJPNv2; *n* = 3,552)[56]. After imputation, we excluded variants with imputation INFO < 0.7.

For GWASs, we obtained the residuals from a linear regression model of each of log-transformed metabolites adjusted for age, age^2, sex, time period from sampling to biobanking, and top 20 genotype PCs. The residuals were then transformed by rank-based inverse normalization. Association analysis of imputed genotype dosage with the normalized residual of each metabolite was performed using PLINK2 software. We constructed the Z-score matrices (**W'**) of the Japanese metabolites GWASs (i.e., 206 rows) × 22,980 variants, in which we applied the same QC to the Z-scores and set the missing Z-scores to zero again. We then performed the projection as described above.

Drug target enrichment analysis

To investigate whether disease-associated genes are systematically enriched in the targets of the approved drugs for the treatment of those diseases, the Genome for REPositioning drugs (GREP)[57] was used. A list of genes closest to the lead variants from GWAS, which was concatenated based on the alphabetical category of ICD10 (A to N), was used as an input gene set to test the enrichment for the target genes of approved drugs for diseases of a given ICD10 category.

References

1. Berger, D. A brief history of medical diagnosis and the birth of the clinical laboratory. Part 1--Ancient times through the 19th century. *MLO. Med. Lab. Obs.* **31**, (1999).
2. Organización Mundial de la Salud. International statistical classification of diseases and related health problems, 10th revision (ICD-10). *World Heal. Organ.* (2016).
3. 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).
4. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
5. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 1–14 (2019).
6. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
7. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
8. Ritchie, M. D. *et al.* Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *Am. J. Hum. Genet.* **86**, 560–572 (2010).
9. Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
10. 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).
11. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, 4393 (2019).
12. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
13. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
14. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

15. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
16. Hagiya, H. *et al.* Trends in incidence and mortality of tuberculosis in Japan: A population-based study, 1997-2016. *Epidemiology and Infection* vol. 147 (2019).
17. Kudoh, Y. & Sakai, S. Current Status of Bacterial Diarrheal Diseases in Japan. in *Bacterial Diarrheal Diseases* 83–93 (Springer Netherlands, 1985). doi:10.1007/978-94-009-4990-4_8.
18. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001-6 (2014).
19. Wu, Y. *et al.* Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10**, 1–10 (2019).
20. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
21. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
22. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science (80-.).* **354**, 760–764 (2016).
23. 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).
24. Qi, H. *et al.* Discovery of susceptibility loci associated with tuberculosis in Han Chinese. *Hum. Mol. Genet.* **26**, 4752–4763 (2017).
25. Sveinbjornsson, G. *et al.* HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat. Genet.* **48**, 318–322 (2016).
26. Baldini, C., Ferro, F., Mosca, M., Fallahi, P. & Antonelli, A. The association of Sjögren syndrome and autoimmune thyroid disorders. *Frontiers in Endocrinology* vol. 9 121 (2018).
27. Nakao, M. *et al.* ABO blood group alleles and the risk of pancreatic cancer in a Japanese population. *Cancer Sci.* **102**, 1076–1080 (2011).
28. Edgren, G. *et al.* Risk of gastric cancer and peptic ulcers in relation to ABO blood type: A cohort study. *Am. J. Epidemiol.* **172**, 1280–1285 (2010).
29. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
30. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science (80-.).* **360**, (2018).

31. 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).
32. Burren, O. S. & Wallace, C. Informed dimension reduction of clinically-related genome-wide association. *bioRxiv* (2020).
33. Chasman, D. I., Giulianini, F., Demler, O. V. & Udler, M. S. Pleiotropy-Based Decomposition of Genetic Risk Scores: Association and Interaction Analysis for Type 2 Diabetes and CAD. *Am. J. Hum. Genet.* **106**, 646–658 (2020).
34. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
35. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
36. GTEx Consortium, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
37. Portincasa, P. & Wang, D. Q. H. Intestinal absorption, hepatic synthesis, and biliary secretion of cholesterol: Where are we for cholesterol gallstone formation? *Hepatology* vol. 55 1313–1316 (2012).
38. Vessey, D. A. The biochemical basis for the conjugation of bile acids with either glycine or taurine. *Biochem. J.* **174**, 621–626 (1978).
39. Coombs, R. R. A. & Gell, P. G. . The Classification of Allergic Reactions Underlying Disease. in *Clinical Aspects of Immunology* 317–337 (1963).
40. Stone, K. D., Prussin, C. & Metcalfe, D. D. IgE, mast cells, basophils, and eosinophils. *J. Allergy Clin. Immunol.* **125**, S73 (2010).
41. Kobayashi, K., Kaneda, K. & Kasama, T. Immunopathogenesis of delayed-type hypersensitivity. *Microsc. Res. Tech.* **53**, 241–245 (2001).
42. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
43. Sakaue, S. *et al.* Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. *Nat. Med.* **26**, 542–548 (2020).
44. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
45. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

46. 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).
47. Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
48. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One* **8**, e64683 (2013).
49. Zheng, J. *et al.* LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
50. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
51. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
52. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
53. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
54. Rüeger, S., McDaid, A. & Kutalik, Z. Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genet.* **14**, e1007371 (2018).
55. Kuriyama, S. *et al.* The Tohoku Medical Megabank Project: Design and mission. *J. Epidemiol.* **26**, 493–511 (2016).
56. Tadaka, S. *et al.* JMorp: Japanese Multi Omics Reference Panel. *Nucleic Acids Res.* **46**, D551–D557 (2018).
57. Sakaue, S. & Okada, Y. GREP: Genome for REPositioning drugs. *Bioinformatics* **35**, 3821–3823 (2019).

URLs

- SDS values in UK10K provided by Pritchard's lab; http://web.stanford.edu/group/pritchardlab/UK10K-SDS-values.zip
- Summary statistics of biomarker GWASs in UKB by Neale's lab ; http://www.nealelab.is/uk-biobank/ukbround2announcement
- LDSC software; https://github.com/bulik/ldsc
- FinnGen release 3 data; https://www.finngen.fi/en/access_results
- dbSNP; https://www.ncbi.nlm.nih.gov/snp/
- World Health Organization, Global Tuberculosis Report; https://www.who.int/tb/publications/global_report/en/