nature genetics

LETTERS

https://doi.org/10.1038/s41588-020-0640-3

Check for updates

Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases

Kazuyoshi Ishigaki[1,2,3,4], Masato Akiyama[1,5], Masahiro Kanai[1,4,6], Atsushi Takahashi[1,7], Eiryo Kawakami[8,9,10], Hiroki Sugishita[9], Saori Sakaue[1,11,12], Nana Matoba[1,13], Siew-Kee Low[1,14], Yukinori Okada[1,11,15,16], Chikashi Terao[17], Tiffany Amariuta[2,3,4,6,18], Steven Gazal[4,19], Yuta Kochi[20,21], Momoko Horikoshi[22], Ken Suzuki[1,11,22,23], Kaoru Ito[24], Satoshi Koyama[24], Kouichi Ozaki[25], Shumpei Niida[25], Yasushi Sakata[26], Yasuhiko Sakata[27], Takashi Kohno[28], Kouya Shiraishi[28], Yukihide Momozawa[29], Makoto Hirata[30], Koichi Matsuda[31], Masashi Ikeda[32], Nakao Iwata[32], Shiro Ikegawa[33], Ikuyo Kou[33], Toshihiro Tanaka[34,35], Hidewaki Nakagawa[36], Akari Suzuki[20], Tomomitsu Hirota[37], Mayumi Tamari[37], Kazuaki Chayama[38], Daiki Miki[38], Masaki Mori[39], Satoshi Nagayama[40], Yataro Daigo[41,42], Yoshio Miki[43], Toyomasa Katagiri[44], Osamu Ogawa[45], Wataru Obara[46], Hidemi Ito[47,48], Teruhiko Yoshida[49], Issei Imoto[50,51,52], Takashi Takahashi[53], Chizu Tanikawa[54], Takao Suzuki[55], Nobuaki Sinozaki[55], Shiro Minami[56], Hiroki Yamaguchi[57], Satoshi Asai[58,59], Yasuo Takahashi[59], Ken Yamaji[60], Kazuhisa Takahashi[61], Tomoaki Fujioka[46], Ryo Takata[46], Hideki Yanai[62], Akihide Masumoto[63], Yukihiro Koretsune[64], Hiromu Kutsumi[65], Masahiko Higashiyama[66], Shigeo Murayama[67], Naoko Minegishi[68], Kichiya Suzuki[68], Kozo Tanno[69], Atsushi Shimizu[69], Taiki Yamaji[70], Motoki Iwasaki[70], Norie Sawada[70], Hirokazu Uemura[71,72], Keitaro Tanaka[73], Mariko Naito[74,75], Makoto Sasaki[69], Kenji Wakai[74], Shoichiro Tsugane[76], Masayuki Yamamoto[68], Kazuhiko Yamamoto[20], Yoshinori Murakami[77], Yusuke Nakamura[78], Soumya Raychaudhuri[2,3,4,6,79,84] ✉, Johji Inazawa[80,81,84] ✉, Toshimasa Yamauchi[23,84] ✉, Takashi Kadowaki[23,84] ✉, Michiaki Kubo[82,84] ✉ and Yoichiro Kamatani[1,83,84] ✉

The overwhelming majority of participants in current genetic studies are of European ancestry. To elucidate disease biology in the East Asian population, we conducted a genome-wide association study (GWAS) with 212,453 Japanese individuals across 42 diseases. We detected 320 independent signals in 276 loci for 27 diseases, with 25 novel loci ($P < 9.58 \times 10^{-9}$). East Asian–specific missense variants were identified as candidate causal variants for three novel loci, and we successfully replicated two of them by analyzing independent Japanese cohorts; p.R220W of *ATG16L2* (associated with coronary artery disease) and p.V326A of *POT1* (associated with lung cancer). We further investigated enrichment of heritability within 2,868 annotations of genome-wide transcription factor occupancy, and identified 378 significant enrichments across nine diseases (false discovery rate < 0.05) (for example, *NKX3-1* for prostate cancer). This large-scale GWAS in a Japanese population provides insights into the etiology of complex diseases and highlights the importance of performing GWAS in non-European populations.

Currently, large-scale genetic studies are dominated by European-descent samples, and fail to capture the level of diversity that exists globally[1–5]. Due to differential genetic architectures, transferability of genetic findings between populations is generally limited. Therefore, this imbalance poses a limitation in our understanding of the genetic architecture of complex diseases in non-European populations. Moreover, this imbalance could result in unequal benefits of precision medicine, as polygenic risk scores based on large-scale genetic studies in European populations have high predictive power of clinical outcomes in European samples[6–10] but poor predictive power in non-European samples[1,11]. Therefore, increasing the ancestral diversity of participants is an essential direction of genetic studies for the equality of genetic findings.

In addition, diversifying the ancestry of participants is important for the discovery of novel disease etiology[12]. Even in large-scale European studies, causal variants might be missed if they have low frequencies or are monomorphic in European populations. Such examples include p.E508K of *HNF1A* (identified in Latino populations[13]) and p.R684* of *TBC1D4* (identified in a Greenlandic

A full list of affiliations appears at the end of the paper.

population[14]), both of which are associated with type 2 diabetes. Therefore, differences in allele frequencies across populations can be an advantage for discovering genetic signals that were not identified in European populations.

Here, we report a genome-wide association study (GWAS) of 42 common diseases in the BioBank Japan Project (BBJ)[15,16]—one of the largest non-European biobanks, consisting of around 200,000 individuals. We provide detailed discussion of the biology of these diseases using multiple genomic annotations. We also examine inter-sex differences in genetic signals. Moreover, by incorporating previous genetic findings, we discuss the extent to which genetic signals are shared across populations while also investigating East Asian–specific genetic signals. Our study provides multiple insights into the etiology of complex traits, and highlights the importance of conducing genetic studies in non-European populations.

Results

GWAS of 42 diseases. We conducted a GWAS of 42 diseases in a Japanese population comprising 179,660 patients who participated in BBJ and 32,793 population-based controls (Table 1 and Supplementary Table 1). The 42 diseases encompassed a wide range of disease categories: 13 neoplastic diseases, five cardiovascular diseases, four allergic diseases, three infectious diseases, two autoimmune diseases, one metabolic disease and 14 uncategorized diseases. By including patients with unrelated diagnoses into control samples, we maximized the power of our GWAS (Methods, Extended Data Fig. 1 and Supplementary Table 1). We employed a generalized linear mixed model (GLMM) in our association analysis using SAIGE[17]. After imputing our genotypes with reference data from the 1000 Genomes Project Phase 3 (1KG Phase3)[18], we tested 8,712,794 autosomal variants and 207,198 X chromosome variants for association with 42 diseases. For 35 diseases for which we had both male and female patients, we also conducted male- and female-specific GWASs.

To quantify the heritability and bias in our GWAS results, we analyzed them using linkage disequilibrium score regression (LDSC) analysis[19] (Supplementary Table 2). Consistent with a recent finding in the European population[20], heritability estimation was improved by incorporating the baselineLD model[21], which includes functional annotations, linkage disequilibrium-dependent architectures and minor allele frequency (MAF)-dependent architectures (Supplementary Fig. 1 and Supplementary Table 2). Although we observed high genomic inflation factors (λ_{GC}) for some diseases (for example, $\lambda_{GC}=1.3$ for type 2 diabetes; Supplementary Table 2), LDSC analysis indicated that the majority of the inflated chi-squared statistics originated from polygenic effects rather than confounding biases (for example, intercept = 1.01 for type 2 diabetes; Supplementary Table 2).

To confirm that our GWAS produced reasonable signals, we examined how many of the previously identified risk alleles were replicable in our GWAS results (Extended Data Fig. 2, Table 1 and Supplementary Table 3). By analyzing all diseases together, 1,219 out of 1,396 previously reported risk alleles were replicated with the same effect direction (sign test; $P=1.47\times10^{-191}$). In East Asian populations of 1KG Phase3, MAFs of non-replicated alleles were significantly lower than those of replicated alleles (Mann–Whitney *U*-test, $P=9.5\times10^{-9}$; Extended Data Fig. 3). Therefore, the replication failures might be due to insufficient statistical power. The high replicability of previous GWAS signals suggested that genetic etiologies are generally shared across populations.

Considering that more than 1.5 million variants in our study are rare variants (MAF < 1%) (Supplementary Fig. 2), applying the conventional genome-wide significance threshold ($P<5\times10^{-8}$), which assumes 1 million independent tests, might increase type I errors. Therefore, to empirically estimate the appropriate *P* value threshold, we conducted a GWAS using 1,000 random binary phenotypes and analyzed the distributions of minimum *P* values (P_{min}) for each phenotype. The 95th percentile of P_{min} was 2.87×10^{-8}, and we defined this *P* value as an empirical genome-wide significance threshold at a significance level of $\alpha=0.05$ (Extended Data Fig. 4). In addition, we considered the multiple testing burden of analyzing sex-specific GWASs; each variant was tested for sex-combined, male-specific and female-specific analyses. Therefore, we set the significance threshold for our GWAS at $P=2.87\times10^{-8}/3$ ($=9.58\times10^{-9}$) and considered $P=5\times10^{-8}$ as a threshold of suggestive associations.

We defined a locus as a genomic region within ±1 megabase (Mb) from the lead variant, and we considered a locus as novel when it did not include any previously reported variants (*P* in previous GWAS $<5.0\times10^{-8}$). In the sex-combined analysis, we detected significant associations for 27 diseases at 260 autosomal loci (outside of the *HLA* region) and nine loci on the X chromosome ($P<9.58\times10^{-9}$; Supplementary Tables 4 and 5). Associations at the *HLA* region have been investigated in detail in a separate article[22]. We further performed conditional analyses in these 269 loci to explore associations independent of the lead variants. We detected 44 additional independent signals for nine diseases ($P<9.58\times10^{-9}$; Supplementary Table 6). The largest number of independent signals in a single locus was seven, found in the *FAM84B/POU5F1B* locus associated with prostate cancer. In the sex-specific analysis (in which male and female cases were analyzed separately), we detected four additional loci for three diseases that were not identified in a sex-combined analysis ($P<9.58\times10^{-9}$; Supplementary Table 7). We tested heterogeneity between effect size estimates for males and females using Cochran's *Q* test. This analysis found that all of the four loci showed nominally significant differences in effect size estimates between sexes (*P* values of heterogeneity (P_{het}) < 0.003). As we will introduce below, three variants with novel suggestive associations ($P<5.0\times10^{-8}$) passed the significance threshold after meta-analyzing with independent replication studies ($P<9.58\times10^{-9}$). In total, we detected 320 independent significant signals in 276 loci for 27 diseases, of which 25 loci were novel ($P<9.58\times10^{-9}$; Fig. 1a and Tables 1 and 2). At three novel significant loci, the lead variants were rare variants with a large effect size (MAF < 0.01 and odds ratio > 2; Fig. 1b), and two of them were missense variants.

To understand the characteristics of novel and known disease-associated variants in our study, we examined their allele frequencies in East Asian and European populations of 1KG Phase3. An intra-population MAF comparison showed that novel variants have significantly lower allele frequencies than known variants in European populations but not in East Asian populations (Extended Data Fig. 5). Trans-ancestral MAF comparison showed that both novel and known variants have higher MAF in East Asian populations than in European populations (Fig. 1c,d). However, trans-ancestral MAF differences are more pronounced in novel variants (Fig. 1e). These observations suggested that the high allele frequencies of disease-associated variants in our cohorts increased the statistical power to detect their significance, especially for novel variants. This highlights the importance of performing GWASs in non-European populations.

We sought to refine the previously identified association signals in European GWASs. We counted the number of variants in linkage disequilibrium with the lead variants in our GWAS and those of previous European GWASs ($r^2>0.8$ in respective populations in 1KG Phase3) (Supplementary Table 4). The average number of variants in linkage disequilibrium with the lead variants was 25.9 in European GWASs and 29.3 in our GWAS. In contrast, the average number of variants in linkage disequilibrium with both lead variants was 12.9. Therefore, our study successfully limited the number of potential causative variants.

Since a disproportionate number of patients with type 2 diabetes and coronary artery disease (CAD) were included in the controls of

Table 1 | Overview of the findings in this GWAS

Disease category	Disease	Sample size		Number of loci				
		Cases	Controls	Previous GWAS		BBJ GWAS		Additional signal
				All	Replicated	All	Novel	
Allergic	Asthma	8,216	201,592	66	57	7	2	2
Allergic	Atopic dermatitis	2,385	209,651	21	17	7	0	0
Allergic	Drug eruption	430	209,651	0	0	0	0	0
Allergic	Pollinosis	5,746	206,707	28	24	0	0	0
Autoimmune	Graves' disease	2,176	210,277	8	8	9	3	0
Autoimmune	Rheumatoid arthritis	4,199	208,254	72	63	5	0	0
Cardiovascular	Cerebral aneurysm	2,820	192,383	8	7	4	2	0
Cardiovascular	Congestive heart failure	9,413	203,040	0	0	0	0	0
Cardiovascular	CAD	29,319	183,134	184	167	53	1	7
Cardiovascular	Ischemic stroke	17,671	192,383	12	9	3	0	0
Cardiovascular	Peripheral artery disease	3,593	208,860	13	10	1	0	0
Infectious	Chronic hepatitis B	1,394	211,059	1	1	0	0	0
Infectious	Chronic hepatitis C	5,794	206,659	2	2	1	0	0
Infectious	Pulmonary tuberculosis	549	211,904	4	4	0	0	0
Metabolic	Type 2 diabetes	40,250	170,615	234	220	89	7	20
Neoplastic	Biliary tract cancer	339	195,745	0	0	0	0	0
Neoplastic	Breast cancer	5,552	89,731	121	102	7	0	0
Neoplastic	Cervical cancer	605	89,731	4	4	0	0	0
Neoplastic	Colorectal cancer	7,062	195,745	73	68	11	0	1
Neoplastic	Endometrial cancer	999	89,731	12	7	0	0	0
Neoplastic	Esophageal cancer	1,300	195,745	14	10	2	0	0
Neoplastic	Gastric cancer	6,563	195,745	10	9	4	1	1
Neoplastic	Hematological malignancy	1,236	211,217	45	32	0	0	0
Neoplastic	Hepatocellular carcinoma	1,866	195,745	2	0	1	1	0
Neoplastic	Lung cancer	4,050	208,403	18	15	6	1	1
Neoplastic	Ovarian cancer	720	89,731	4	3	0	0	0
Neoplastic	Pancreatic cancer	442	195,745	20	17	0	0	0
Neoplastic	Prostate cancer	5,408	103,939	107	97	20	0	9
Other	Arrhythmia	17,861	194,592	114	105	16	1	0
Other	Cataract	24,622	187,831	0	0	1	1	0
Other	COPD	3,315	201,592	70	54	5	1	2
Other	Cirrhosis	2,184	210,269	3	2	2	0	0
Other	Endometriosis	734	102,372	11	11	0	0	0
Other	Epilepsy	2,143	210,310	4	1	0	0	0
Other	Glaucoma	5,761	206,692	55	43	5	0	0
Other	Interstitial lung disease	806	211,647	10	7	1	1	0
Other	Keloid	812	211,641	3	3	4	1	1
Other	Nephrotic syndrome	957	211,496	0	0	0	0	0
Other	Osteoporosis	7,788	204,665	2	1	1	1	0
Other	Periodontal disease	3,219	209,234	2	0	0	0	0
Other	Urolithiasis	6,638	205,815	23	23	7	1	0
Other	Uterine fibroids	5,954	95,010	16	16	4	0	0

We considered a previous GWAS signal to be replicated when the signal in the previous studies had the same effect direction in this study. We utilized a GLMM in our GWAS, and set the genome-wide significance threshold at $P < 9.58 \times 10^{-9}$ for our study. We also included the variants that passed this significance threshold after meta-analyzing with the replication study. Detailed information is also provided in Supplementary Tables 3–7. Additional signal denotes the number of independent significant signals identified by conditioning analyses. COPD, chronic obstructive pulmonary disease.

GWASs for other diseases, our study design might create spurious associations mirroring the effects of risk alleles of type 2 diabetes and CAD. However, this possibility was ruled out by the following observations: (1) excluding all patients from control samples did not affect effect size estimates (Extended Data Fig. 1); (2) risk loci detected in our GWAS for other diseases were not enriched within

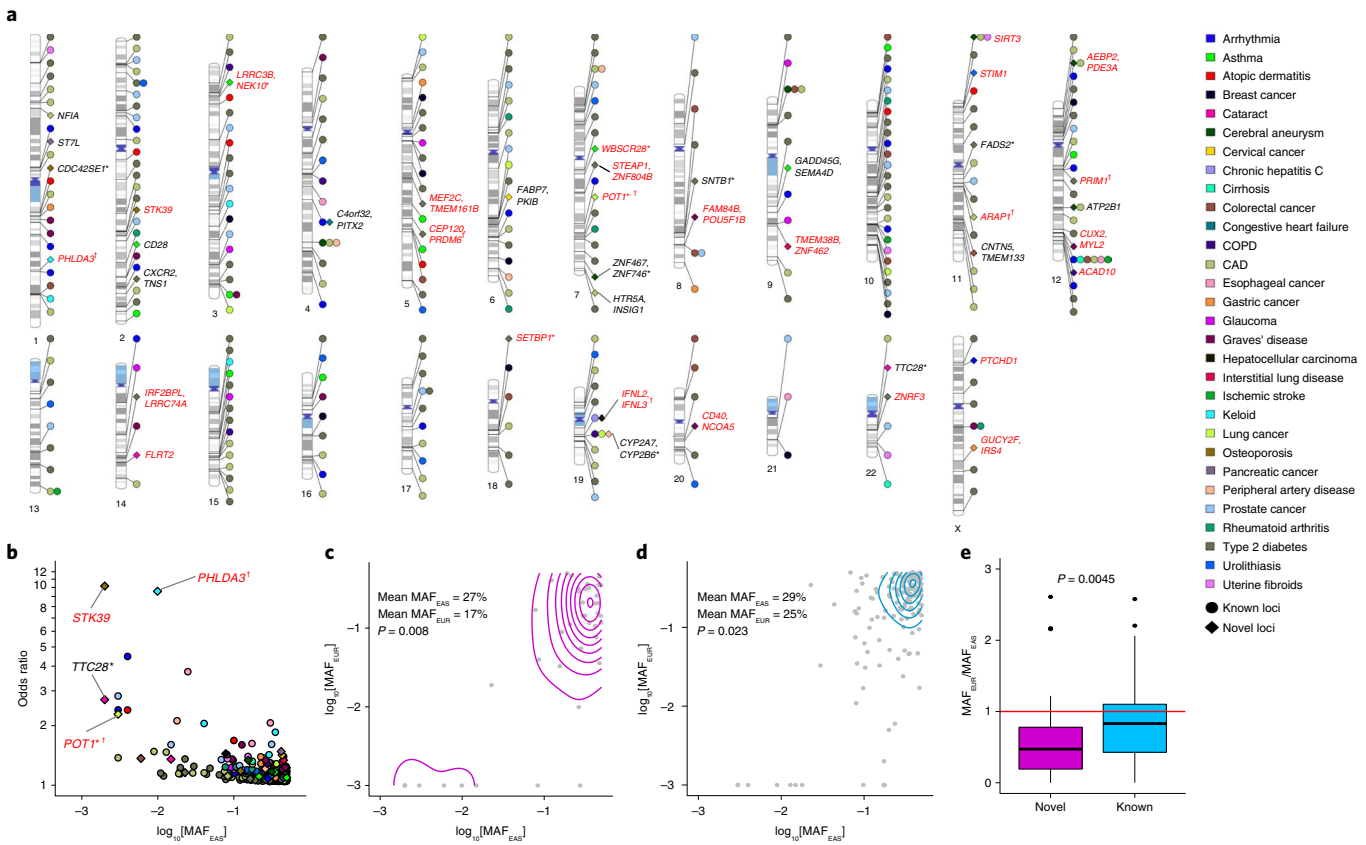


Fig. 1 | Disease-associated loci detected in this GWAS. a, Phenogram[52] of 331 suggestive loci detected in this GWAS ($P<5.0\times10^{-8}$). Pleiotropic associations were plotted at the same position (Methods). **b**, Allele frequencies and the odds ratios of the lead variants at 331 suggestive loci detected in this GWAS ($P<5.0\times10^{-8}$). The odds ratio of the risk allele was used. In **a** and **b**, novel loci (diamond symbols) are annotated with the closest gene names (only genes with an odds ratio of >2 are highlighted in **b**). Genes with significant associations have red labels ($P<9.58\times10^{-9}$). The sample size of the GWAS is provided in Table 1. We utilized a GLMM in our GWAS. Asterisks indicate loci detected in sex-specific GWASs. Dagger symbols show the lead variants that were linked to missense variants (see text for criteria). **c**–**e**, Trans-ancestral MAF comparisons of disease-associated variants at novel (**c** and **e**; $n=41$) and known loci (**d** and **e**; $n=153$) with suggestive significance ($P<5\times10^{-8}$). For known loci, we restricted this analysis to loci where the closest reported variants were discovered by GWAS in European populations. *P* values are provided (two-sided Mann–Whitney *U*-test). When MAF < 0.001, MAF was adjusted to 0.001 to fit the log scale. MAF_{EAS}, MAF in East Asian population (1KG Phase3). MAF_{EUR}, MAF in European population (1KG Phase3). In **e**, the center line in each box indicates the median, and the box limits indicate the upper and lower quartiles. Whiskers indicate interquartile range (IQR) × 1.5. COPD, chronic obstructive pulmonary disease.

type 2 diabetes or CAD known loci (Supplementary Fig. 3); and (3) effect directions of the known protective alleles of type 2 diabetes or CAD were not significantly biased to positive values in our GWAS for other diseases (Supplementary Fig. 4). Thus, we confirmed that our study results were not biased by having many patient samples in control groups.

Biological interpretation of disease-associated variants. Next, we investigated the potential impact of the disease-associated variants on protein functions (Supplementary Table 8). We linked the GWAS association and the missense variant when the lead variant and the missense variant were in linkage disequilibrium ($r^2>0.6$ in East Asians of 1KG Phase3) and the missense variant was included in the 95% credible set (Methods). Using these criteria, seven novel significant signals ($P<9.58\times10^{-9}$) were linked to missense variants. Although four missense variants were not the lead variant, conditioning on these missense variants canceled the signal of the lead variant (Fig. 2a and Supplementary Fig. 5). Importantly, three missense variants were monomorphic in Europeans and Africans (1KG Phase3); p.R220W of *ATG16L2* (rs11235604) associated with CAD; p.V326A of *POT1* (rs75932146) associated with lung cancer; and p.E62G of *PHLDA3* (rs192314256) associated with keloid (Fig. 2, Extended Data Fig. 6 and Table 3). Considering the relevance of these findings, we additionally included two independent cohorts in a Japanese population (2,855 CAD cases and 15,211 controls; and 2,440 lung cancer cases and 467 controls). This replication study successfully confirmed the associations at *ATG16L2* and *POT1* loci, and fixed-effects meta-analysis improved the statistical significance; the suggestive association at the *POT1* locus passed the significance threshold ($P<9.58\times10^{-9}$) (Supplementary Tables 9 and 10). Here, we discuss each of the three East Asian–specific missense variants in detail. First, *ATG16L2* is an autophagy-related gene highly expressed in immune cells, and previous studies reported that p.R220W of *ATG16L2* is also associated with immune-related traits (that is, serum levels of non-albumin protein in a Japanese population[23] and Crohn's disease in a Chinese population[24]). A previous GWAS for CAD in European populations did not detect significant associations at the *ATG16L2* locus[25] (Fig. 2a), suggesting that p.R220W of *ATG16L2*, which is absent in Europeans, may be the causal variant. Therefore, dysregulated autophagy in immune cells might play an important role in CAD. Second, *POT1* is a member of the telombin family and this protein binds to telomeres, regulating telomere length. Missense variants of *POT1* have been described as being responsible for several familial cancers[26–28]. In addition, our

Table 2 | Summary data of the lead variants in the 25 novel loci detected in this GWAS

									Allele frequency			
Disease	**Variant**	**REF**	**ALT**	**Gene**	**OR**	**L95**	**U95**	***P***	**EAS**	**EUR**	**AFR**	**Distance (Mbp)**
Loci detected in sex-combined analysis												
Arrhythmia	rs73205368	T	C	*PTCHD1*	1.08	1.06	1.10	4.25×10^{-15}	0.281	0.055	0.047	NA
CAD	rs11235571 (rs11235604)	G (C)	A (T)	*ARAP1* (*ATG16L2*)	0.90 (0.91)	0.87 (0.88)	0.93 (0.94)	2.64×10^{-9} (1.73×10^{-8})	0.083 (0.100)	0 (0)	0 (0)	2.9
Cataract	rs75812946	G	A	*FLRT2*	1.35	1.22	1.50	3.41×10^{-9}	0.015	0	0	NA
Cerebral aneurysm	rs12226402	G	A	*SIRT3*	1.34	1.23	1.45	1.57×10^{-12}	0.155	0.033	0.099	68.9
Cerebral aneurysm	rs78535549	C	T	*AEBP2*, *PDE3A*	0.85	0.81	0.90	7.97×10^{-9}	0.528	0.036	0.055	12.2
COPD	rs11066008	A	G	*ACAD10*	1.29	1.21	1.37	4.34×10^{-17}	0.275	0	0.001	3.8
Gastric cancer	rs1205528	T	C	*GUCY2F*, *IRS4*	0.92	0.89	0.94	2.80×10^{-10}	0.354	0.884	0.654	NA
Graves' disease	rs10673095	T	TTC	*FAM84B*, *POU5F1B*	0.81	0.76	0.87	2.11×10^{-9}	0.476	0.362	0.772	5.9
Graves' disease	rs11065783	A	G	*CUX2*, *MYL2*	1.34	1.24	1.44	7.23×10^{-14}	0.264	0.010	0	NA
Graves' disease	rs1569723	C	A	*CD40*, *NCOA5*	1.20	1.13	1.28	4.06×10^{-9}	0.565	0.743	0.976	NA
Hepatocellular carcinoma	rs8107030	A	G	*IFNL2*, *IFNL3*	1.44	1.28	1.62	7.96×10^{-10}	0.078	0.170	0.027	NA
Interstitial lung disease	rs6477542	C	T	*TMEM38B*, *ZNF462*	1.34	1.21	1.48	6.90×10^{-9}	0.451	0.207	0.123	NA
Keloid	rs192314256	T	C	*PHLDA3*	9.56	5.91	15.45	3.28×10^{-20}	0.010	0	0	20.8
Osteoporosis	rs578031265	C	T	*STK39*	10.16	4.74	21.74	2.38×10^{-9}	0.002	0.001	0	31.8
Type 2 diabetes	rs7721099	T	C	*MEF2C*, *TMEM161B*	1.05	1.04	1.07	1.41×10^{-9}	0.512	0.143	0.255	1.4
Type 2 diabetes	rs200525873	GT	G	*CEP120*, *PRDM6*	0.91	0.88	0.94	4.90×10^{-9}	0.086	0.040	0.037	11.2
Type 2 diabetes	rs39218	T	C	*STEAP1*, *ZNF804B*	1.06	1.04	1.08	1.28×10^{-9}	0.191	0.503	0.311	12.6
Type 2 diabetes	rs5762925	A	C	*ZNRF3*	1.05	1.03	1.07	3.93×10^{-9}	0.462	0.353	0.262	1.0
Type 2 diabetes*	rs2277339	T	G	*PRIM1*	1.05	1.04	1.07	2.67×10^{-10}	0.206	0.111	0.199	9.0
Type 2 diabetes*	rs17105012	C	A	*IRF2BPL*, *LRRC74A*	1.04	1.03	1.06	8.84×10^{-9}	0.297	0.143	0.034	2.6
Urolithiasis	rs12290747	T	C	*STIM1*	0.89	0.85	0.92	3.24×10^{-9}	0.317	0.313	0.017	107.3
Loci detected in sex-specific analysis												
Asthma	rs13227841	T	C	*WBSCR28*	0.86	0.81	0.90	2.04×10^{-9}	0.650	0.677	0.334	32.4
Asthma	rs9836823	A	G	*LRRC3B*, *NEK10*	0.86	0.82	0.91	5.19×10^{-9}	0.337	0.362	0.116	6.2
Lung cancer*	rs75932146	A	G	*POT1*	2.42	1.87	3.13	1.69×10^{-11}	0.003	0	0	NA
Type 2 diabetes	rs202209118	T	TCC	*SETBP1*	1.16	1.10	1.22	7.78×10^{-9}	0.023	0.019	0.002	6.1

Detailed information on these variants is provided in Supplementary Tables 4, 5 and 7. For variants detected in the sex-specific GWAS, statistics of sex with significant associations are provided. For a lead variant of CAD (rs11235571), we have also provided data of a missense variant (rs11235604) in linkage disequilibrium with the lead variant ($r^2=0.68$ in East Asian populations of 1KG Phase3; Table 3). The sample size is provided in Table 1. We utilized a GLMM in our GWAS, and set a genome-wide significance threshold at $P<9.58\times10^{-9}$. Disease names are marked by an asterisk when the variants passed the significance threshold after meta-analyzing with replication studies (Supplementary Tables 10 and 13), and statistics of meta-analysis are provided for such variants. The distance between the lead variant in this study and the closest reported variant in the previous GWAS is also provided. When there were no reported associations on the same chromosome, the distance information is set to NA. Allele frequencies of 1KG Phase3 are provided. AFR, African populations; ALT, alternative allele; COPD, chronic obstructive pulmonary disease; EAS, East Asian populations; EUR, European populations; L95, lower 95% confidence interval; Mbp, megabase pairs; OR, odds ratio relative to the alternative allele; REF, reference allele; U95, upper 95% confidence interval.

study showed that p.V326A of *POT1* is also positively associated with the risk of five other neoplastic diseases ($P<0.05$; Extended Data Fig. 7). These findings suggest that this variant might increase the risk of neoplastic diseases in general. p.V326A of *POT1* is more strongly associated with lung cancer in females than males (odds ratio for females: 2.29; odds ratio for males: 1.26; $P_{het}=7.7\times10^{-4}$) (Fig. 2b and Supplementary Table 7). We sought to figure out whether the sex-dependent effect can be explained by other factors, and conducted an association test stratified by histological and smoking status (Supplementary Table 10). However, we could not reach a definitive conclusion due to limited statistical power, and hence further large-scale studies will be required to answer this question. Together with a known association at the *TERT* locus (Supplementary Table 4), we provide additional evidence that telomere dysregulation is pathogenic for lung cancer. Third, p.E62G of *PHLDA3* is predicted to have a deleterious effect to its protein function (SIFT score (ref. [29])=0; CADD score (ref. [30])=33), and we detected a large effect size for keloid (odds ratio=9.56; 95% CI=5.91–15.45). We confirmed that genotyping of rs192314256 (p.E62G of *PHLDA3*) was not biased by batches of genotyping experiments or geographic areas (Supplementary Fig. 6). PHLDA3 is known to be a suppressor of AKT[31], and an upregulated AKT signaling pathway is related to increased collagen production from dermal fibroblasts[32]. Therefore, damaged PHLDA3 may activate the AKT pathway, promoting the development of keloid. Together, our study successfully identified novel potential causal genes that it would be difficult to discover by GWAS in European populations due to restrictive European allele frequencies.

Table 3 | Population-specific missense variants linked to disease-associated variants

							BBJ GWAS						Replication analysis						Meta-analysis				Allele frequency		
Disease	Variant	Gene	Amino acid change	REF	ALT	Case	Control	OR	L95	U95	*P*	Case	Control	OR	L95	U95	*P*	OR	L95	U95	P	P_{het}	EAS	EUR	AFR
Loci detected in sex-combined analysis																									
CAD	rs11235604	*ATG16L2*	p.R220W	C	T	29,319	183,134	0.91	0.88	0.94	1.73×10^{-8}	2,855	15,211	0.83	0.74	0.94	3.33×10^{-3}	0.91	0.88	0.93	5.69×10^{-10}	0.16	0.100	0	0
Keloid	rs192314256	*PHLDA3*	p.E62G	T	C	812	211,641	9.56	5.91	15.45	3.28×10^{-20}	-	-	-	-	-	-	-	-	-	-	-	0.010	0	0
Loci detected in sex-specific analysis																									
Lung cancer	rs75932146	*POT1*	p.V326A	A	G	1,340	101,766	2.29	1.71	3.05	2.21×10^{-8}	2,440	467	2.99	1.71	5.24	1.26×10^{-4}	2.42	1.87	3.13	1.69×10^{-11}	0.40	0.003	0	0

We conducted a meta-analysis with a fixed-effects model using independent Japanese cohorts (Supplementary Tables 9 and 10) and tested heterogeneity using Cochran's *Q* test (P_{het}). We utilized a GLMM in the BBJ GWAS and the replication study of CAD. We utilized a GLM for the replication analysis of lung cancer. We set a genome-wide significance threshold at $P<9.58\times10^{-9}$. In addition to the statistics, the sample size and allele frequencies of 1KG Phase3 are provided. Detailed information about missense variants is provided in Supplementary Table 8. AFR, African populations; ALT, alternative allele; EAS, East Asian populations; EUR, European populations; L95, lower 95% confidence interval; OR, odds ratio relative to the alternative allele; REF, reference allele; U95, upper 95% confidence interval.

We also investigated the potential impacts of the disease-associated variants on the messenger RNA levels using the Genotype-Tissue Expression (GTEx) database of expression quantitative trait loci (eQTL)[33]. Since the eQTL data are generated in European populations, we could not apply formal colocalization tests[34,35], which assume the same linkage disequilibrium structures between GWASs and eQTL studies. Therefore, we linked the GWAS association and the eQTL variant when the GWAS lead variant and the eQTL variant were in linkage disequilibrium ($r^2>0.6$ both in East Asian and European populations of 1KG Phase3) and the eQTL variant was included in the 95% credible set. We found that seven novel significant signals ($P<9.58\times10^{-9}$) and five novel suggestive signals ($P<5\times10^{-8}$) could be explained by at least one eQTL variant (Supplementary Table 11). Among them, the eQTL signals for *ATP2B1*, which were linked to a novel, suggestive variant of cerebral aneurysm (rs11105352; $P=1.22\times10^{-8}$), were highly specific to arterial tissues (Fig. 3). Since the loss of *ATP2B1* in vascular smooth muscle cells induced blood pressure elevation in mice[36], decreased expression of *ATP2B1* in arteries might induce hypertension, which leads to increased risk of cerebral aneurysm.

Replication with European GWAS results. Replication analysis in the same population is a critical part of genetic studies. Although we included two independent replication studies for CAD and lung cancer in a Japanese population, we were not able to prepare replication cohorts in a Japanese population for other diseases. Therefore, we conducted replication studies using previous European GWAS results. We utilized publicly available GWAS summary statistics of European populations for ten diseases (asthma, atrial fibrillation, breast cancer, CAD, congestive heart failure, glaucoma, ischemic stroke, prostate cancer, rheumatoid arthritis and type 2 diabetes; see Methods for selection of diseases) and tested for consistency in direction of effect. For these ten diseases, our GWAS detected suggestive associations at 218 known and 19 novel loci ($P<5\times10^{-8}$). Among them, statistics of European GWASs were available at 149 known and 15 novel loci. We first conducted replication analysis at the known loci. We restricted this analysis to 112 known loci with significant associations also in European GWASs ($P<5\times10^{-8}$) to exclude loci for which the European GWASs had insufficient power. Effect directions were consistent between BBJ and European GWASs at 109 out of 112 loci, but opposite at three loci (Extended Data Fig. 8 and Supplementary Table 12). These three replication failures were probably due to differences in linkage disequilibrium structure between populations (Extended Data Fig. 8). We then conducted replication analysis at the novel loci. Among 15 novel variants, 12 were replicated with the same effect direction (Supplementary Table 13). Meta-analysis using a fixed-effect model increased the level of significance in six of them, and two suggestive novel variants passed the significance threshold ($P<9.58\times10^{-9}$) (rs2277339 and rs17105012, associated with type 2 diabetes; Table 2 and Supplementary Table 13). Among the three variants that failed replication, rs13227841 is a missense variant originally identified as a potential causal variant at this locus (p.W78R of *WBSCR28*; Supplementary Table 8), which suggests that variants in linkage disequilibrium with rs13227841, not rs13227841 itself, may be responsible for the observed associations. The other replication failures might be due to different linkage disequilibrium structures or the absence of the causal variants in European populations. Further efforts to conduct a replication analysis in a Japanese population will be required to confirm the associations that we failed to replicate in these European studies.

Genetic correlation between male- and female-specific GWASs. To understand the differences in the genetic risks between males and females, we assessed genetic correlations using LDSC[37] between the results of the sex-specific GWASs for the 20 diseases

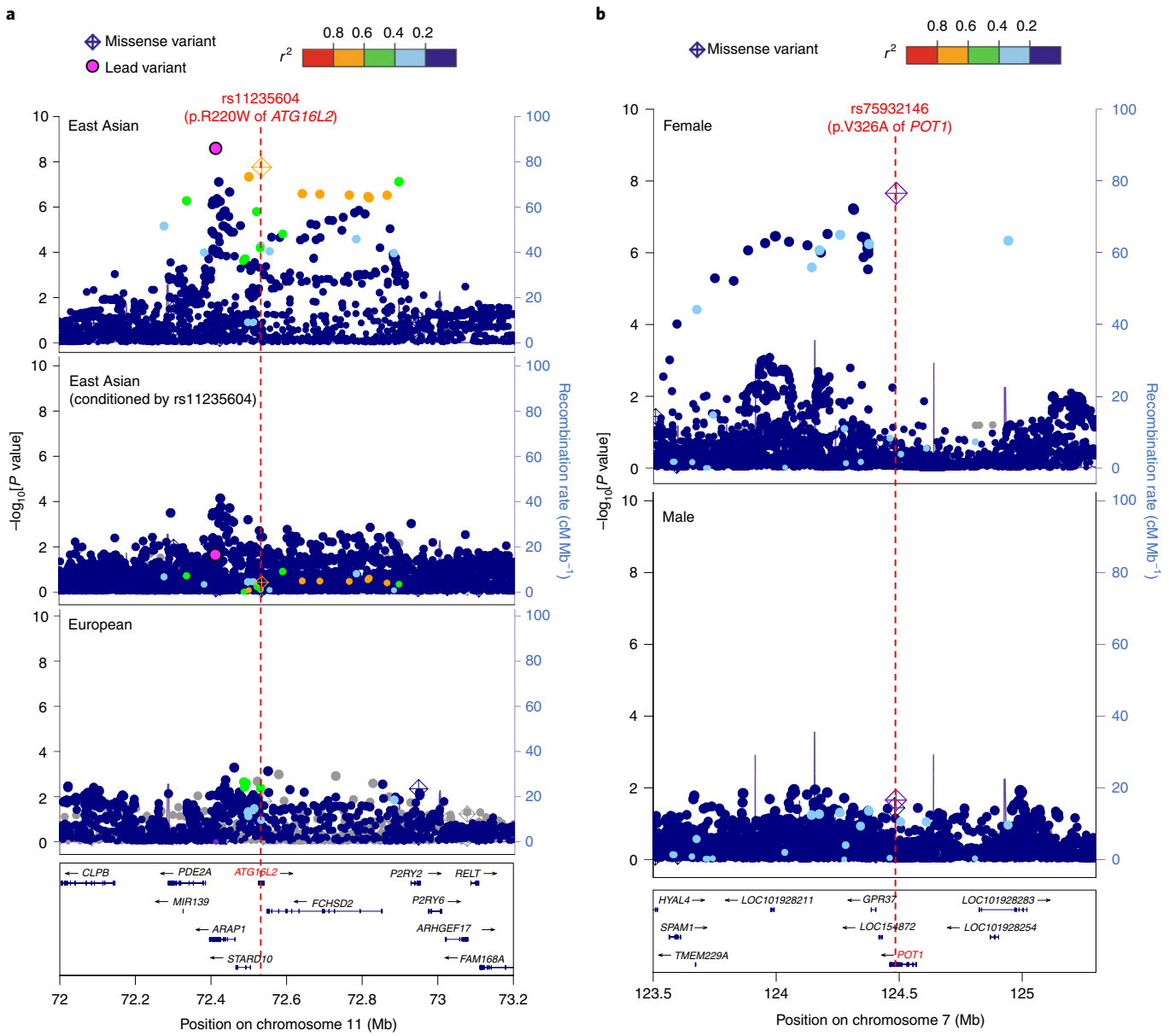


Fig. 2 | Novel associations that can be explained by East Asian–specific missense variants. a,**b**, Regional association plots for CAD (29,319 cases versus 183,134 controls) (**a**) and lung cancer (2,710 male cases versus 106,637 male controls; 1,340 female cases versus 101,766 female controls) (**b**). In **a**, *P* values were determined by conditional analysis and those from the European GWAS[25] were plotted separately. In **b**, *P* values from the female- and male-specific GWASs were plotted separately. We utilized a GLMM in our GWAS.

(see Methods for selection of diseases). Although most correlations were close to 1, the correlation of asthma was significantly smaller than 1 (genetic correlation = 0.63 (s.e. = 0.12); $P = 2.2 \times 10^{-3} < 0.05/20$; Extended Data Fig. 9). This finding suggested that genetic risks of asthma might be different between males and females. To explore the biological mechanism underlying this finding, we estimated the enrichment of the heritability of male or female asthma in the 220 cell type-specific regulatory regions using stratified linkage disequilibrium score regression (S-LDSC)[38]. We found significant enrichments for either male or female asthma in three annotations; T_H0, T_H1 and colonic mucosa ($P < 0.05/220$; Extended Data Fig. 9). Among them, the colonic mucosa annotation showed significant heterogeneity in the enrichment of heritability ($P_{het} = 0.006 < 0.05/3$). Recent studies suggested that host–microbiome interactions at intestinal mucosa (gut–lung axis) have important roles in the development of asthma[39,40], and our study suggested that the gut–lung axis might have sex-dependent roles in asthma. Considering their marginal significance, a replication study will be required to confirm these findings.

Transcription factors (TFs) underlying the etiology of diseases. To acquire more insights into disease biology, we estimated the heritability enrichments in the binding sites of a variety of TFs using S-LDSC. We included TF binding sites defined by 2,868 publicly available chromatin immunoprecipitation sequencing datasets for 410 unique TFs (Supplementary Table 14). To ensure the data was mutually comparable, we began our analysis from the raw sequencing data and defined TF binding sites using a uniform protocol (Methods). Using linkage disequilibrium scores of all of the TF binding sites, we grouped them into 15 clusters (the cluster name was defined by the most dominant TF), and performed uniform

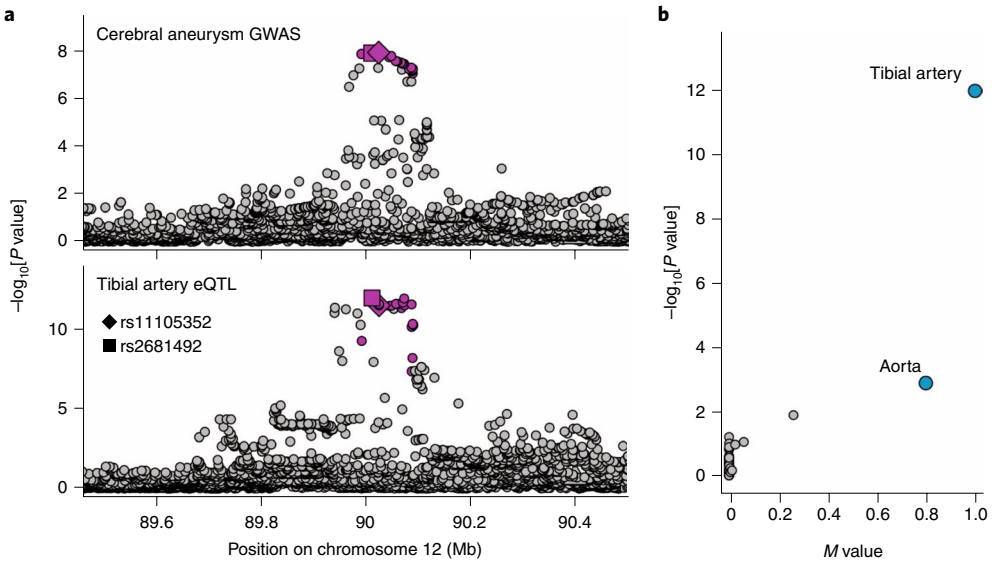


Fig. 3 | A novel suggestive association of cerebral aneurysm can be explained by artery-specific eQTL signals for *ATP2B1*. a, Regional association plots of the cerebral aneurysm GWAS (2,820 cases versus 192,383) at the *ATP2B1* locus (top) and eQTL signals for *ATP2B1* in the tibial artery (bottom). The lead variant of GWAS (rs11105352) is indicated by a diamond symbol, and the lead variant of eQTL (rs2681492) is indicated by a square symbol. Variants in linkage disequilibrium with rs11105352 are highlighted in purple ($r^2 > 0.6$ both in the East Asian and European populations of 1KG Phase3). We utilized a GLMM in our GWAS. **b**, Tissue specificity of eQTL signals for *ATP2B1* at rs2681492 (the lead variant of eQTL in the tibial artery (square symbol in **a**)). *P* values in the eQTL analysis and *M* values (the posterior probability that an eQTL effect existed in each tissue tested in the cross-tissue meta-analysis) in all tissues in the GTEx project[33] are provided. Each dot indicates a separate tissue. All statistics of eQTL analysis were derived from release version 7 of the GTEx project[33].

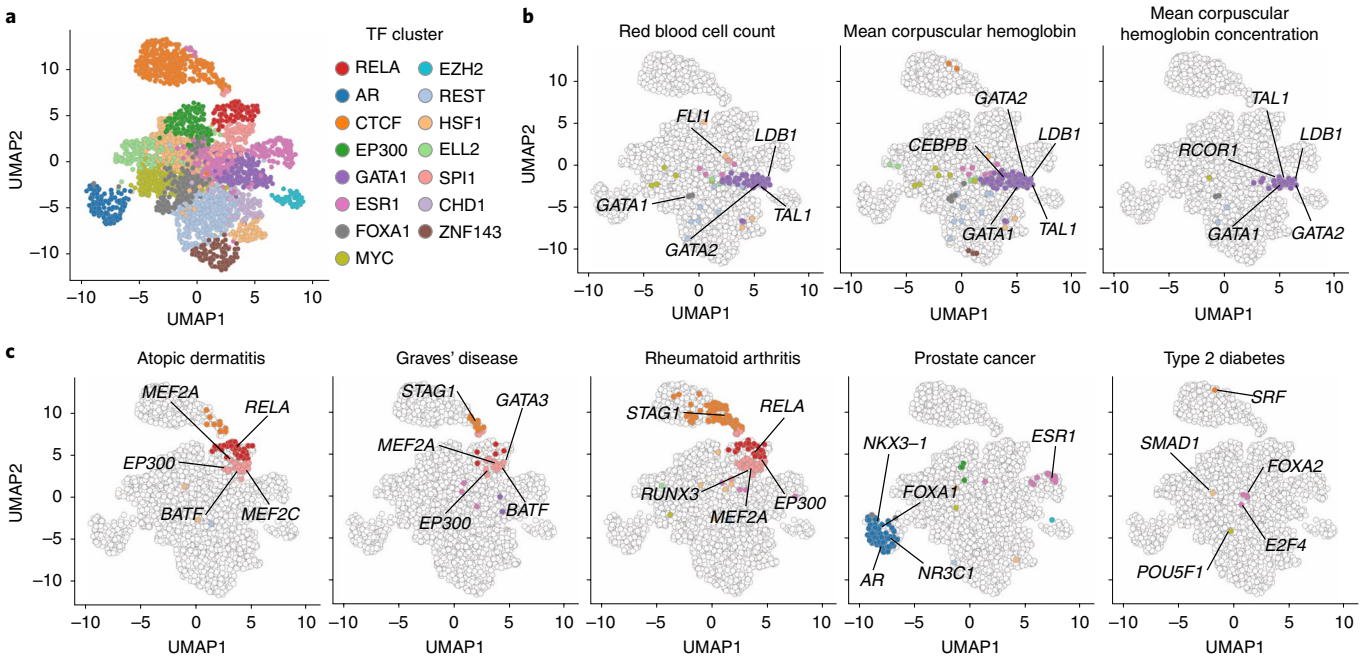


Fig. 4 | TFs whose binding sites were enriched for heritability of diseases. a, All of the 2,868 sets of TF binding sites grouped into 15 clusters were plotted in the UMAP space. **b**,**c**, The results of S-LDSC were plotted on the UMAP space for red blood cell-related traits (**b**) and diseases in this GWAS that had more than five significant TF binding site tracks (**b**; the results of the other diseases are provided in Extended Data Fig. 10). The significant results (FDR < 0.05) are highlighted using the cluster-specific colors shown in the legend in **a**. The names of the top five most significant TFs are also shown in each plot.

manifold approximation and projection (UMAP)[41] to project all TF binding sites into a two-dimensional space (Methods, Fig. 4a and Supplementary Fig. 7). To scale the performance of this analysis, we first analyzed previously reported GWASs for red blood cell-related traits[23] where the critical role of *GATA1* was supported by multiple pieces of evidence[42–46], and we successfully recapitulated this biology (Fig. 4b). We then applied this analysis to our 24 GWAS results (see Methods for selection of diseases) and detected 378 significant enrichments for nine diseases (false discovery rate (FDR) < 0.05) (Fig. 4c, Extended Data Fig. 10 and Supplementary Table 15). Biologically plausible TFs were highlighted by this analysis: *RELA*, a subunit of nuclear factor-κB, for atopic dermatitis, rheumatoid

arthritis and Graves' disease; sex hormone receptors (*AR* and *ESR1*) for prostate cancer; and *FOXA2*, which regulates insulin secretion in pancreatic β cells[47], for type 2 diabetes (Fig. 4c). This analysis also suggested that *NKX3-1* (a prostate-specific homeobox gene) has an important role in the biology of prostate cancer (Fig. 4c). In addition to this polygenic analysis, the importance of *NKX3-1* was also suggested by the regional analysis integrating eQTL databases; the risk allele of prostate cancer at the *NKX3-1* locus (rs4872174-C) was suggested to decrease the expression of *NKX3-1* (Supplementary Table 11). Consistently, loss of *NKX3-1* expression in human prostate cancers was reported to be correlated with tumor progression[48]. Together, our results confirmed and expanded our understanding of complex traits in the context of TF activity.

Discussion

Our study showed the advantages of conducting genetic studies in non-European populations. Typically, linkage disequilibrium acts as a major hurdle limiting the identification of causal variants in GWAS. However, jointly analyzing GWAS results from populations with different linkage disequilibrium structures can narrow down causal variants[12]. Indeed, when we consider variants in linkage disequilibrium with a lead variant as candidate causal variants ($r^2 > 0.8$), our study successfully reduced the number of candidate causal variants at 68 loci that were originally discovered in previous European GWASs (Supplementary Table 4). In addition, some novel variants in our study have been missed in larger GWASs in European populations due to restrictive European allele frequencies. Therefore, diversifying the ancestry of participants is important not only for the equality of genetic findings but also for the discovery of novel disease etiology.

Although previous studies already reported important roles of TFs in the etiology of complex traits[49–51], our TF enrichment analysis has two distinguishing features from previous studies. One feature is the comprehensiveness; we included 2,868 TF annotations, which is more than those used in most previous studies. The second feature is the method of the enrichment test; we utilized S-LDSC, whereas most previous studies utilized naive enrichment tests using genome-wide significant variants. S-LDSC evaluates enrichment of GWAS signals irrespective of significance, and it is robust to the biases coming from the overlapping annotations. Therefore, by incorporating a comprehensive catalog of TF annotations with a sophisticated method to test heritability enrichment, we provided evidence of TF importance in complex diseases from a polygenic angle.

The critical limitation of this study is insufficient replication analyses to validate novel signals. Among 25 novel loci ($P < 9.58 \times 10^{-9}$), we were able to prepare East Asian replication datasets for only two of them; p.R220W of *ATG16L2*, associated with CAD, and p.V326A of *POT1*, associated with lung cancer. To supplement this insufficiency, we utilized European GWAS results when data were available. We tested the replicability of eight novel signals ($P < 9.58 \times 10^{-9}$) and observed evidence of heterogeneity in effect size estimates for three of them ($P_{het} < 0.05$; Supplementary Table 13). This may be the case for several reasons: the locus might possess different linkage disequilibrium structures between populations and the variant might tag the causal variant only in East Asian populations (as illustrated in Extended Data Fig. 8); effect sizes might be truly different between populations; or they might be false positives. Therefore, until further replication studies in East Asian populations are conducted, we need to be cautious about the validity of these putatively novel variants since we were not able to provide evidence of replicability.

In summary, we conducted a large-scale GWAS of 42 diseases in a non-European population and provided rich public resources for genetic studies. Our study provides multiple insights into the etiology of complex traits by integrating annotations of missense variants, eQTL variants and TF binding site tracks. Currently, genetic studies are overwhelmed by European-descent samples, making the clinical translation of genetic findings far more beneficial to European individuals than other populations[1]. Our study contributes to broadening the population diversity in genetic studies and should potentially mitigate the problems originating from this imbalance.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0640-3.

Received: 21 February 2020; Accepted: 1 May 2020;
Published online: 08 June 2020

References

1. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
2. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
3. Morales, J. et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
4. Diversity matters. *Nat. Rev. Genet.* **20**, 495 (2019).
5. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
6. Maas, P. et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the united states. *JAMA Oncol.* **2**, 1295–1302 (2016).
7. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
8. Kullo, I. J. et al. Incorporating a genetic risk score into coronary heart disease risk estimates clinical perspective. *Circulation* **133**, 1181–1188 (2016).
9. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
10. Natarajan, P. et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101 (2017).
11. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
12. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
13. Estrada, K. et al. Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population the SIGMA Type 2 Diabetes Consortium. *J. Am. Med. Assoc.* **311**, 2305–2314 (2014).
14. Moltke, I. et al. A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
15. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
16. 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).
17. 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).
18. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
19. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
20. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).
21. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
22. Hirata, J. et al. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
23. 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).

24. Ma, T., Wu, S., Yan, W., Xie, R. & Zhou, C. A functional variant of *ATG16L2* is associated with Crohn's disease in the Chinese population. *Color. Dis.* **18**, O420–O426 (2016).
25. Van der Harst, P. & Verweij, N. The identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* **122**, 433–443 (2018).
26. Calvete, O. et al. The wide spectrum of *POT1* gene variants correlates with multiple cancer types. *Eur. J. Hum. Genet.* **25**, 1278–1281 (2017).
27. Bainbridge, M. N. et al. Germline mutations in shelterin complex genes are associated with familial glioma. *J. Natl Cancer Inst.* **107**, 384 (2015).
28. Robles-Espinoza, C. D. et al. *POT1* loss-of-function variants predispose to familial melanoma. *Nat. Genet.* **46**, 478–481 (2014).
29. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863–874 (2001).
30. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
31. Kawase, T. et al. PH domain-only protein PHLDA3 is a p53-regulated repressor of Akt. *Cell* **136**, 535–550 (2009).
32. Bujor, A. M. et al. Akt blockade downregulates collagen and upregulates MMP1 in human dermal fibroblasts. *J. Invest. Dermatol.* **128**, 1906–1914 (2008).
33. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
34. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
35. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
36. Kobayashi, Y. et al. Mice lacking hypertension candidate gene *ATP2B1* in vascular smooth muscle cells show significant blood pressure elevation. *Hypertension* **59**, 854–860 (2012).
37. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
38. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
39. Frati, F. et al. The role of the microbiome in asthma: the gut–lung axis. *Int. J. Mol. Sci.* **20**, E123 (2018).
40. Stokholm, J. et al. Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).
41. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
42. Matsuda, M., Sakamoto, N. & Fukumaki, Y. Delta-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the delta-globin gene promoter. *Blood* **80**, 1347–1351 (1992).
43. De Gobbi, M. et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **312**, 1215–1217 (2006).
44. Pevny, L. et al. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* **349**, 257–260 (1991).
45. Elhanati, Y., Marcou, Q., Mora, T. & Walczak, A. M. RepgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics* **32**, 1943–1951 (2016).
46. Welch, J. J. et al. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**, 3136–3147 (2004).
47. Lantz, K. A. et al. *Foxa2* regulates multiple pathways of insulin secretion. *J. Clin. Invest.* **114**, 512–520 (2004).
48. Bowen, C. et al. Loss of NKX3.1 expression in human prostate cancers correlates with tumor progression. *Cancer Res.* **60**, 6111–6115 (2000).
49. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016).
50. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
51. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
52. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information across chromosomes with PhenoGram. *BioData Min.* **6**, 18 (2013).

[1]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [2]Center for Data Sciences, Harvard Medical School, Boston, MA, USA. [3]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan. [6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [7]Department of Genomic Medicine, Research Institute, National Cerebral and Cardiovascular Center, Osaka, Japan. [8]Medical Sciences Innovation Hub Program (MIH), RIKEN, Yokohama, Japan. [9]Laboratory for Developmental Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [10]Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba, Japan. [11]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Osaka, Japan. [12]Department of Allergy and Rheumatology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [13]Department of Genetics and UNC Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [14]Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan. [15]Laboratory of Statistical Immunology, WPI Immunology Frontier Research Center, Osaka University, Osaka, Japan. [16]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Osaka, Japan. [17]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [18]Graduate School of Arts and Sciences, Harvard University, Cambridge, MA, USA. [19]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [20]Laboratory for Autoimmune Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [21]Department of Genomic Function and Diversity, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. [22]Laboratory for Genomics of Diabetes and Metabolism, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [23]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. [24]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [25]Medical Genome Center, National Center for Geriatrics and Gerontology, Obu, Japan. [26]Department of Cardiovascular Medicine, Osaka University Graduate School of Medicine, Osaka, Japan. [27]Department of Cardiovascular Medicine, Tohoku University Graduate School of Medicine, Tohoku, Japan. [28]Division of Genome Biology, National Cancer Center Research Institute, Tokyo, Japan. [29]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [30]Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [31]Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [32]Department of Psychiatry, Fujita Health University School of Medicine, Aichi, Japan. [33]Laboratory for Bone and Joint Diseases, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan. [34]Laboratory for Cardiovascular Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [35]Department of Human Genetics and Disease Diversity, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan. [36]Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan. [37]Laboratory for Respiratory and Allergic Diseases, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [38]Department of Gastroenterology and Metabolism, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan. [39]Department of Surgery and Sciences, Graduate School of Medicine, Kyushu University, Fukuoka, Japan. [40]Department of Gastroenterological Surgery, The Cancer Institute Hospital of the Japanese Foundation for Cancer Research, Tokyo, Japan. [41]Department of Medical Oncology and Cancer Center, and Center for Advanced Medicine against Cancer, Shiga University of Medical Science, Shiga, Japan. [42]Center for Antibody and Vaccine Therapy, Research Hospital, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [43]Department of Genetic Diagnosis, The Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan. [44]Division of Genome Medicine, Institute for Genome Research, Tokushima University, Tokushima, Japan. [45]Department of Urology, Kyoto University Graduate School of Medicine, Kyoto, Japan. [46]Department of Urology, Iwate Medical University School of Medicine, Iwate, Japan. [47]Division of Cancer Information and Control, Aichi Cancer Center Research Institute, Nagoya, Japan.

[48]Division of Descriptive Cancer Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, Japan. [49]Division of Genetics, National Cancer Center Research Institute, Tokyo, Japan. [50]Division of Molecular Genetics, Aichi Cancer Center Research Institute, Nagoya, Japan. [51]Risk Assessment Center, Aichi Caner Center Hospital, Nagoya, Japan. [52]Division of Cancer Genetics, Nagoya University Graduate School of Medicine, Nagoya, Japan. [53]Aichi Cancer Center, Nagoya, Japan. [54]Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [55]Tokushukai Group, Tokyo, Japan. [56]Department of Bioregulation, Nippon Medical School, Kawasaki, Japan. [57]Department of Hematology, Nippon Medical School, Tokyo, Japan. [58]Division of Pharmacology, Department of Biomedical Science, Nihon University School of Medicine, Tokyo, Japan. [59]Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan. [60]Department of Internal Medicine and Rheumatology, Juntendo University Graduate School of Medicine, Tokyo, Japan. [61]Department of Respiratory Medicine, Juntendo University Graduate School of Medicine, Tokyo, Japan. [62]Fukujuji Hospital, Japan Anti-Tuberculosis Association, Tokyo, Japan. [63]Aso Iizuka Hospital, Fukuoka, Japan. [64]National Hospital Organization Osaka National Hospital, Osaka, Japan. [65]Center for Clinical Research and Advanced Medicine, Shiga University of Medical Science, Shiga, Japan. [66]Department of General Thoracic Surgery, Osaka International Cancer Institute, Osaka, Japan. [67]Department of Neurology and Neuropathology (the Brain Bank for Aging Research), Tokyo Metropolitan Geriatric Hospital and Institute of Gerontology, Tokyo, Japan. [68]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan. [69]Iwate Tohoku Medical Megabank Organization, Iwate Medical University, Iwate, Japan. [70]Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan. [71]Department of Preventive Medicine, Institute of Biomedical Sciences, Tokushima University Graduate School, Tokushima, Japan. [72]College of Nursing Art and Science, University of Hyogo, Akashi, Japan. [73]Department of Preventive Medicine, Saga University Faculty of Medicine, Saga, Japan. [74]Department of Preventive Medicine, Nagoya University Graduate School of Medicine, Nagoya, Japan. [75]Department of Oral Epidemiology, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan. [76]Center for Public Health Sciences, National Cancer Center, Tokyo, Japan. [77]Division of Molecular Pathology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [78]Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [79]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. [80]Department of Molecular Cytogenetics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan. [81]Bioresource Research Center, Tokyo Medical and Dental University, Tokyo, Japan. [82]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [83]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [84]These authors jointly supervised this work: Soumya Raychaudhuri, Johji Inazawa, Toshimasa Yamauchi, Takashi Kadowaki, Michiaki Kubo, Yoichiro Kamatani. ✉e-mail: soumya@broadinstitute.org; johinaz.cgen@mri.tmd.ac.jp; tyamau-tky@umin.net; kadowaki-3im@h.u-tokyo.ac.jp; michiaki.kubo@riken.jp; yoichiro.kamatani@riken.jp

Methods

Subjects. All case samples in this GWAS were collected in the BBJ (https://biobankjp.org/english/index.html)[15,16], which is a biobank that collaboratively collects DNA and serum samples from 12 medical institutions in Japan and recruited approximately 200,000 patients with a diagnosis of at least one of 47 diseases. Among them, cases with dyslipidemia were not analyzed in this study because it was already reported as a quantitative trait in our previous study[23]. Amyotrophic lateral sclerosis and febrile seizure were also not analyzed due to limited sample sizes. Cases with myocardial infarction, stable angina and unstable angina were re-classified into a single disease category (CAD). Thus, we analyzed 42 diseases in this study. As controls, we used samples from population-based prospective cohorts: the Tohoku Medical Megabank Organization (Tohoku University), Iwate Tohoku Medical Megabank Organization (Iwate Medical University)[53], Japan Public Health Center-based Prospective Study and Japan Multi-Institutional Collaborative Cohort Study. In addition, we also included samples in BBJ without related diagnoses in the control group (Extended Data Fig. 1 and Supplementary Table 1). The sample sizes and demographic data are provided in Supplementary Table 1. For all participating studies, we obtained informed consent from all participants by following the protocols approved by their institutional ethical committees. We obtained approval from the ethics committees of the RIKEN Center for Integrative Medical Sciences and the Institute of Medical Sciences at The University of Tokyo. We have complied with all of the relevant ethical regulations.

Genotyping. We genotyped samples with the Illumina HumanOmniExpressExome BeadChip or a combination of the Illumina HumanOmniExpress and HumanExome BeadChips. For quality control of samples, we excluded those with: (1) a sample call rate of <0.98; and (2) outliers from East Asian clusters identified by principal component analysis (PCA) using the genotyped samples and the three major reference populations (Africans, Europeans and East Asians) in the International HapMap Project[54]. For quality control of genotypes, we excluded variants meeting any of the following criteria: (1) call rate < 99%; (2) *P* value for Hardy–Weinberg equilibrium (HWE) $< 1.0 \times 10^{-6}$; and (3) fewer than five heterozygotes. Using 939 samples whose genotypes were also analyzed by whole-genome sequencing (WGS), we added additional quality control based on the concordance rate between genotyping array and WGS. Variants with a concordance rate < 99.5% or a non-reference discordance rate ≥ 0.5% were excluded. We note that the allele frequency of rs671 (the East Asian–specific functional missense variant at *ALDH2*) substantially varies among the domestic regions within Japan due to strong selection pressure[55] and that genotypes of rs671 did not follow HWE. We thus did not apply the HWE quality control for rs671. We had confirmed the 100% concordance of rs671 genotypes between the single-nucleotide polymorphism microarray data used in this study and our internal WGS data ($n = 2{,}798$; see details in the discussion in ref. [56]).

Imputation. We utilized all samples in 1KG Phase3 (version 5; www.1000genomes.org/)[18] as a reference for imputation. We first prephased the genotypes with SHAPEIT2 (version 2.778; https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html) and then imputed dosages with minimac3 (version 2.0.1; https://genome.sph.umich.edu/wiki/Minimac). After imputation, we excluded variants with an imputation quality of Rsq < 0.7. For the X chromosome, we performed prephasing and imputation separately for males and females, and we excluded variants with an imputation quality of Rsq < 0.7 in either of them.

Genome-wide association analysis. We conducted a GWAS by employing a GLMM using SAIGE (version 0.29.4.2; https://github.com/weizhouUMICH/SAIGE)[17]. This strategy enabled us to maintain related samples in our GWAS, and the sample sizes were increased by 6% on average compared with removing related samples. Briefly, there are two steps in SAIGE. In step 1, we fit a null logistic mixed model using genotype data, and we added covariates in this step (see below). In step 2, we performed the single-variant association tests using imputed variant dosages. We applied the leave-one-chromosome-out approach. For the X chromosome, we conducted GWASs separately for males and females, and merged their results by inverse-variance fixed-effects meta-analysis. We used only female control samples for the GWAS of female-specific diseases (that is, breast cancer, cervical cancer, endometrial cancer, ovarian cancer, endometriosis and uterine fibroids). Similarly, we used only male control samples for the GWAS of prostate cancer. We incorporated age and the top five principal components as covariates. We also used sex as a covariate for the GWAS of diseases, which included both male and female samples. We also conducted male-specific and female-specific GWASs using the same pipeline as described above, and estimated heterogeneity in the effect size estimates using Cochran's *Q* test. In each GWAS, we excluded variants with a minor allele count of <10 based on the recommendation from thee developers of SAIGE. We created regional association plots using LocusZoom (version 1.2; http://locuszoom.sph.umich.edu/locuszoom/)[57]. We performed stepwise conditional analysis within ±1 Mb from the lead variant and repeated the association test by additionally incorporating the dosages of the identified variants as covariates in SAIGE step 1 until we did not detect any significant associations.

For each disease, we defined a significantly associated locus as a genomic region within ±1 Mb from the lead variant. When a locus did not include any variants that were previously reported to be significantly associated with the same disease ($P < 5.0 \times 10^{-8}$), we defined it as a novel locus. Since we tested each variant for disease association three times (sex-combined, female-specific and male-specific analysis), we considered the multiple testing burden on the empirical significance threshold ($P = 2.87 \times 10^{-8}$; see next paragraph), and we set the genome-wide significance threshold for our study at $P = 2.87 \times 10^{-8}/3$ ($= 9.58 \times 10^{-9}$).

Estimation of the empirical significance threshold by permutation test. Using the identical statistical method and imputed genotype data as were used in the main analysis, we conducted a GWAS using 1,000 simulated phenotypes. We utilized downsampled individuals ($n = 10{,}000$) because a permutation test using all samples (~200,000) was not computationally tractable. We simulated binary phenotypes with 1,920 cases and 8,080 controls (that is, the same case–control ratio as in the type 2 diabetes GWAS in our study). For each of the 1,000 simulated phenotypes, P_{min} was recorded, and the distributions of 1,000 P_{min} were analyzed. This analysis showed that the 95th percentile of P_{min} was 2.87×10^{-8} (Extended Data Fig. 4). We defined this value as an empirical genome-wide significance threshold at a significance level of $\alpha = 0.05$. The 95% confidence interval was estimated by 1,000 bootstraps using the R package boot (version 1.3-20).

To test the potential effect of downsampling on the P_{min} distributions, we compared the P_{min} distributions using all samples ($n = 198{,}137$) with those using 10,000 samples. To increase the computational efficiency, we restricted this analysis to imputed genotype data in chromosome 22. For this analysis, we utilized Plink2 (https://www.cog-genomics.org/plink/2.0/)[58] because SAIGE requires whole genotype data to estimate relatedness even when we restrict the analysis to chromosome 22. This analysis confirmed that downsampling does not have substantial impact on the P_{min} distributions (Extended Data Fig. 4).

Estimation of heritability. We estimated heritability and confounding bias in our GWAS results with LDSC (version 1.0.0; https://github.com/bulik/ldsc/)[19] using the baselineLD model (version 2.1; https://data.broadinstitute.org/alkesgroup/LDSCORE/)[21], which includes 86 annotations, including ten MAF- and six linkage disequilibrium-related annotations that correct for bias in heritability estimates[20]. Heritability and confounding bias were calculated using 481 East Asian samples in 1KG Phase3. For the analysis using LDSC, we excluded variants in the *HLA* region (chromosome 6: 26–34 Mb). We also calculated the heritability *z* score to assess the reliability of heritability estimation.

Absolute quantification of heritability estimation using GWAS results using GLMM can be biased because the effective sample size could be different from the true sample size (relative quantification is not biased; hence, GWAS results using GLMM can be applied for genetic correlation analysis and S-LDSC safely). Therefore, to confirm the robustness of heritability estimation in our analysis, we also performed a GWAS using a generalized linear regression model (GLM). As a simple GLM does not account for the bias caused by genetic relationships, we further excluded related samples (pi-hat by >0.187), and we analyzed genotype data with Plink2 using the same covariates as described above. Heritability estimates based on GWAS using two different methods (SAIGE versus PLINK) were comparable (Supplementary Table 2).

Replication of the previously reported variants by this GWAS. We included data in the GWAS Catalog (https://www.ebi.ac.uk/gwas/) that satisfied the following criteria: (1) *P* in previous GWAS $< 5 \times 10^{-8}$; (2) risk allele information was reported; (3) outside of major histocompatibility complex region (chromosome 6: 23–37 Mb); and (4) variants were analyzed in this study. When multiple variants were reported within the 1-Mb window, we included one variant for each disease. We considered a previous GWAS signal as replicated when the signal in the previous GWAS had the same effect direction in our GWAS.

Replication of the findings in this GWAS by independent cohorts in a Japanese population. We included an independent Japanese cohort of CAD and controls who enrolled in the Osaka Acute Coronary Insufficiency Study (OACIS)[59] and the National Center for Geriatrics and Gerontology (NCGG) Biobank[60]. OACIS is a study that examined patients with myocardial infarction at 25 collaborating hospitals in Osaka, Japan from April 1998 to April 2006. The NCGG Biobank is one of the facilities belonging to the National Center Biobank Network (https://ncbiobank.org/en/home.php). It has been running since 2012. The participants were recruited from NCGG hospital, which is located in Obu city, and the other nearby medical institutes. We also included 1,392 control DNA sequences from the Health Science Research Resources Bank in Osaka, Japan. Samples in NCGG were genotyped by Infinium Asian Screening Array-24 version 1.0 (Illumina), and samples in OACIS were genotyped using the same platform as was used for the BBJ samples. We extracted bi-allelic, shared variants genotyped in these studies. We excluded variants with: (1) a Hardy–Weinberg disequilibrium ($P < 1 \times 10^{-6}$); and (2) a low call rate (<99%). We excluded samples using the following criteria: samples with low call rate (<99%); PCA outliers; heterozygosity outliers; and sex-discordant samples. After QC, 2,855 CAD cases, 15,211 controls, and 111,041

single-nucleotide polymorphisms remained. After prephasing with Eagle (version 2.3), we performed imputation by minimac4 (version 1.0.0) using the 1KG Phase3 reference panel. An association test was conducted using SAIGE (version 0.36.3), including age, sex and the top five principal components as covariates. We tested the influence of bias using LDSC. The intercept was 1.008 (s.e. = 0.014), and λ_{GC} was 1.053, suggesting there was no substantial bias in the association results.

We also included a Japanese cohort with 2,440 female cases of lung cancer and 467 female controls enrolled in the study of the National Cancer Center Hospital. All cases were adenocarcinoma. Genotyping of rs75932146 was conducted by invader assay. An association test was conducted by logistic regression. Meta-analysis was conducted using a fixed-effect model via inverse-variance weighting. Heterogeneity of effect size estimates was tested using Cochran's *Q* test.

Replication of the findings in this GWAS by the previous European GWASs. We searched for European GWASs whose summary statistics were publicly available and whose disease affection statuses were based on physician diagnosis (excluding GWASs based on self-reported phenotypes). The latter criterion was added because all cases in the BBJ were diagnosed by a physician, and we wanted to prepare European GWASs of comparable phenotypes. We were able to prepare European GWAS summary statistics for ten diseases. Summary statistics for eight diseases were downloaded from the GWAS Catalog (https://www.ebi.ac.uk/gwas/) and their names and PMIDs were as follows; atrial fibrillation (30061737), breast cancer (29059683), CAD (29212778), glaucoma (29891935), ischemic stroke (29531354), prostate cancer (29892016), rheumatoid arthritis (24390342) and type 2 diabetes (30054458). Summary statistics of two diseases were downloaded from UK Biobank GWAS summary statistics at Neale Lab (http://www.nealelab.is/uk-biobank) and their names and phenotype code were as follows: asthma (22127); and congestive heart failure (I50). A meta-analysis was conducted using the fixed-effects model via inverse-variance weighting, and heterogeneity in effect size estimates was tested using Cochran's *Q* test.

Pleiotropy. We utilized the following variants detected in GWASs for each disease: (1) lead variants in the significantly associated loci; (2) independent signals detected by conditional analysis; and (3) lead variants detected in sex-specific GWASs. We defined pleiotropic association when these variants were in linkage disequilibrium ($r^2 > 0.6$). We calculated r^2 using East Asian samples in 1KG Phase3 (ref. [18]) by PLINK[58].

Functional annotation of associated variants. We calculated r^2 using East Asian samples (r^2_{EAS}) and European samples (r^2_{EUR}) in 1KG Phase3 (ref. [18]) by PLINK[58]. We also identified 95% credible sets using the R package corrcoverage (version 1.2.1). We linked the GWAS association and the missense variant when the lead variant and the missense variant were in linkage disequilibrium ($r^2_{EAS} > 0.6$) and the missense variant was included in the 95% credible set. For the annotation of non-synonymous variants, we used ANNOVAR (http://annovar.openbioinformatics.org/en/latest/)[61]. GRCh37 (hg19) coordinates were used in this study.

We also annotated GWAS variants with eQTL detected in the European population (release version 7 of the GTEx project)[33] under the following conditions: (1) the lead variants of the eQTL study were in linkage disequilibrium ($r^2_{EAS} > 0.6$ and $r^2_{EUR} > 0.6$) with GWAS variants; (2) the missense variant was included in the 95% credible set; and (3) *Q* values of the lead variants in the eQTL study were <0.05.

Genetic correlations between sex-specific GWASs. We estimated genetic correlations between our GWAS results by LDSC (version 1.0.0)[19] using East Asian linkage disequilibrium scores, which we presented in our previous study[23]. We excluded variants in the *HLA* region (chromosome 6: 26–34 Mb). We analyzed 20 diseases based on two criteria: (1) heritability was reliably estimated (heritability *z* score > 2; Supplementary Table 2); and (2) both male and female patients were included.

TF binding sites. We obtained 3,158 raw human chromatin immunoprecipitation sequencing data files in SRA format from the Gene Expression Omnibus database. We converted them to FASTQ format using the fastq-dump function of the SRA Toolkit (https://www.ncbi.nlm.nih.gov/sra/). We performed quality control of sequence reads using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We mapped these reads to the genome assembly GRCh37 using Bowtie 2 (version 2.2.5; http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) with default parameters. We called peaks using MACS (version 2.1; https://github.com/taoliu/MACS) with default parameters ($q < 0.01$) and defined them as TF binding sites. We excluded TF binding site tracks that did not have at least one binding region on every chromosome, and 2,868 genome-wide TF binding site tracks remained (Supplementary Table 14).

Stratified linkage disequilibrium score regression. We conducted S-LDSC[38] to partition heritability. For S-LDSC analysis of the sex-specific GWAS of asthma, we used 220 cell type-specific annotations used in previous articles[23,38]. For other S-LDSC analyses, we used the TF binding site tracks described in the previous paragraph. For all sites of TF binding, we empirically extended sites by 500 base pairs at both ends for this analysis. We computed annotation-specific linkage disequilibrium scores using the 1KG Phase3 (version 5) East Asian reference haplotypes[18]. We estimated heritability enrichment of binding sites of each TF, while controlling for the merged binding sites of all TFs and the 53 categories of the full baseline model available from the authors' website (https://data.broadinstitute.org/alkesgroup/LDSCORE/). We did not use the baselineLD model (version 2.1)[21] in this analysis to increase the power of detecting significant enrichment. We excluded variants in the *HLA* region (chromosome 6: 26–34 Mb). We analyzed 24 diseases whose heritability was reliably estimated (heritability *z* score > 2; Supplementary Table 2). We calculated the *P* value of the regression coefficient. For each trait, we calculated the FDR using the Benjamini–Hochberg method. We set a significance threshold at FDR < 0.05 for this analysis.

Visualization of TF binding sites. There was a complex correlation structure among the 2,868 TF binding site tracks used for S-LDSC analysis. In S-LDSC, we regressed GWAS chi-squared statistics on linkage disequilibrium scores of each TF binding site (TF linkage disequilibrium score); hence, we focused on correlations between TF linkage disequilibrium scores, not correlations between TF binding sites. We first performed PCA using all TF linkage disequilibrium scores. To classify them into mutually correlated TF groups, we performed *k*-means clustering ($k = 15$) using the top 15 principal components. We named each cluster using the most dominant TF in each cluster (Fig. 4). TF binding sites and their assigned cluster names are provided in Supplementary Table 14. We then performed UMAP[41] using the top 15 principal components to project all TF binding sites into a two-dimensional space. UMAP was conducted using the R package umap (version 0.2.0.0). Our workflow is illustrated in Supplementary Fig. 7.

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

Data availability

GWAS summary statistics of the 42 diseases are publicly available from our website (JENGER; http://jenger.riken.jp/en/) and the National Bioscience Database Center Human Database (https://humandbs.biosciencedbc.jp/en/) (research ID: hum0014) without any access restrictions. GWAS genotype data for case samples were deposited at the National Bioscience Database Center Human Database (research ID: hum0014).

References

53. Kuriyama, S. et al. The Tohoku Medical Megabank Project: design and mission. *J. Epidemiol.* **26**, 493–511 (2016).
54. Altshuler, D. M. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
55. 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).
56. Matoba, N. et al. GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. *Nat. Hum. Behav.* **4**, 308–316 (2020).
57. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
58. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
59. Mizuno, H. et al. Impact of atherosclerosis-related gene polymorphisms on mortality and recurrent events after myocardial infarction. *Atherosclerosis* **185**, 400–405 (2006).
60. Asanomi, Y. et al. A rare functional variant of SHARPIN attenuates the inflammatory response and associates with increased risk of late-onset Alzheimer's disease. *Mol. Med.* **25**, 20 (2019).
61. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

Acknowledgements

We acknowledge the staff of the BBJ for their outstanding assistance. We express our heartfelt gratitude to the Tohoku Medical Megabank Organization (Tohoku University), Iwate Tohoku Medical Megabank Organization (Iwate Medical University), Japan Public Health Center-based Prospective Study, Japan Multi-Institutional Collaborative Cohort Study and NCGG Biobank for their invaluable contributions to collecting control samples. We also express our gratitude to the OACIS for contributing to the replication study of CAD, and the National Cancer Center Hospital for contributing to the replication study of lung cancer. We also express our gratitude to E.K. and H.S. for kindly sharing their results of chromatin immunoprecipitation sequencing data analysis. We extend our appreciation to Y. Yukawa, Y. Yokoyama and other members of the Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences for great support. This research was supported by the Tailor-Made Medical Treatment Program (BBJ) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Agency for Medical Research and Development (AMED) under grant numbers JP17km0305002 (to M.Kubo)

and JP17km0305001 (to M.M., S.Nagayama, Y.D., Y.Miki, T.Katagiri, O.O., W.O., H.I., T.Yoshida, I.I., T.Takahashi, J.I. and K.M.), JST KAKENHI grants (18H02932 to S.I.) and the Research Program on Hepatitis from AMED (JP19fk0310109 and JP19fk0210020 to K.C.). The Tohoku Medical Megabank Organization (Tohoku University) was supported in part by MEXT-JST and AMED, the most recent grant numbers being JP19km0105001 and JP19km0105002 (to M.Y.). The Iwate Tohoku Medical Megabank Organization (Iwate Medical University) was supported in part by MEXT-JST and AMED, the most recent grant numbers being JP19km0105003 and JP19km0105004. The Japan Public Health Center Prospective Study has been supported by the National Cancer Center Research and Development Fund since 2011 (latest grant number: 29-A-4, to S.T.) and was supported by a Grant-in-Aid for Cancer Research from the Ministry of Health, Labour and Welfare of Japan from 1989–2010. The Japan Multi-Institutional Collaborative Cohort Study was supported by Grants-in-Aid for Scientific Research for Priority Areas of Cancer (17015018 to Nobuyuki Hamajima) and Innovative Areas (221S0001 to Hideo Tanaka) and by Japan Society for the Promotion of Science KAKENHI grants (CoBiA and 16H06277) from MEXT (to Hideo Tanaka and K.W.). The NCGG study was partly supported by AMED under grant number JP18kk0205009 (S.Niida) and JP20dk0207045 (K.O.). The OACIS was supported by AMED (JP19ek0210081 to Yasuhiko Sakata). The lung cancer study at the National Cancer Center Hospital was supported by the National Cancer Center Research and Development Fund (NCC Biobank), AMED (JP16ck0106096 to T.Kohno) and the Ministry of Health, Labour and Welfare program (H29-Gantaisaku-Ippann-025 to T.Kohno). The study at Fujita Health University was supported by AMED under grant numbers JP20dm0107097 (M.Ikeda and N.I.), JP20km0405201 (N.I.) and JP20km0405208 (M.Ikeda).

Author contributions

K. Ishigaki wrote the manuscript with critical input from S.R. and Y. Kamatani. K. Ishigaki conducted all of the bioinformatics analyses with the help of M.A., M. Kanai, A.T., S.S., N. Matoba, S.-K.L., Y.O., C. Terao, T.A., S.G., S.R. and Y. Kamatani. Y. Momozawa and M. Kubo performed the genotyping. H.S. and E.K. analyzed the chromatin immunoprecipitation sequencing data. K. Ito, S.K., K.O., S. Niida, Yasushi Sakata, Yasuhiko Sakata, T. Kohno and K. Shiraishi contributed to the replication studies in a Japanese population. S.-K.L., Y. Kochi, M. Horikoshi, Ken Suzuki, K. Ito, M. Hirata, K.M., S.I., I.K., T. Tanaka, H.N., A. Suzuki, T.H., M.T., K.C., D.M., M.M., S. Nagayama, Y.D., Y. Miki, T. Katagiri, O.O., W.O., H.I., T. Yoshida, I.I., T. Takahashi, C. Tanikawa, T.S., N. Sinozaki, S. Minami, H. Yamaguchi, S.A., Y.T., K. Yamaji, K. Takahashi, T.F., R.T., H. Yanai, A.M., Y. Koretsune, H.K., M. Higashiyama, S. Murayama, K. Yamamoto, Y. Murakami, Y.N., J.I., T. Yamauchi, T. Kadowaki, M. Kubo and Y. Kamatani contributed to management of the BBJ data. M. Ikeda and N.I. managed the other GWAS data. N. Minegishi, Kichiya Suzuki, K. Tanno, A.Shimizu, T. Yamaji, M. Iwasaki, N. Sawada, H.U., K. Tanaka, M.N., M.S., K.W., S.T. and M.Y. contributed to management of the cohort control data.

Competing interests

The authors declare no competing interests.

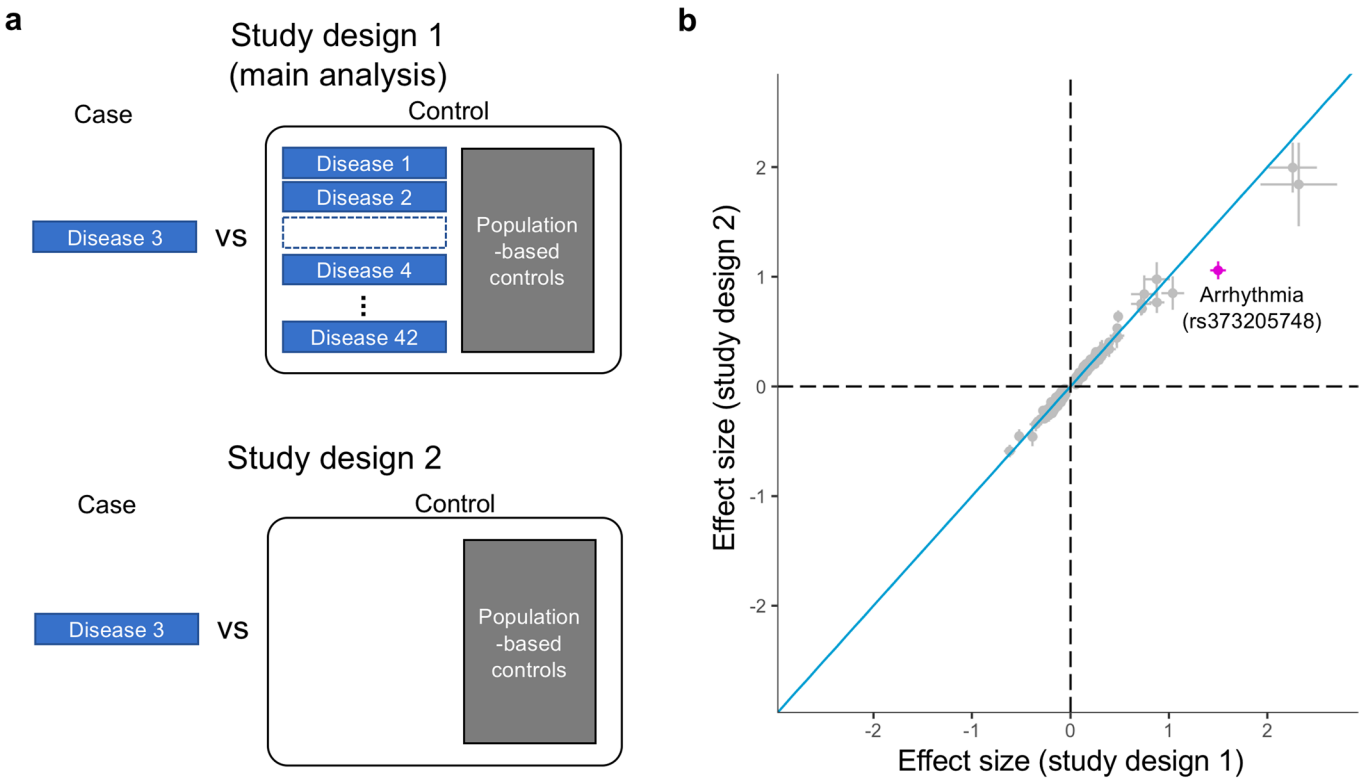
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0640-3.

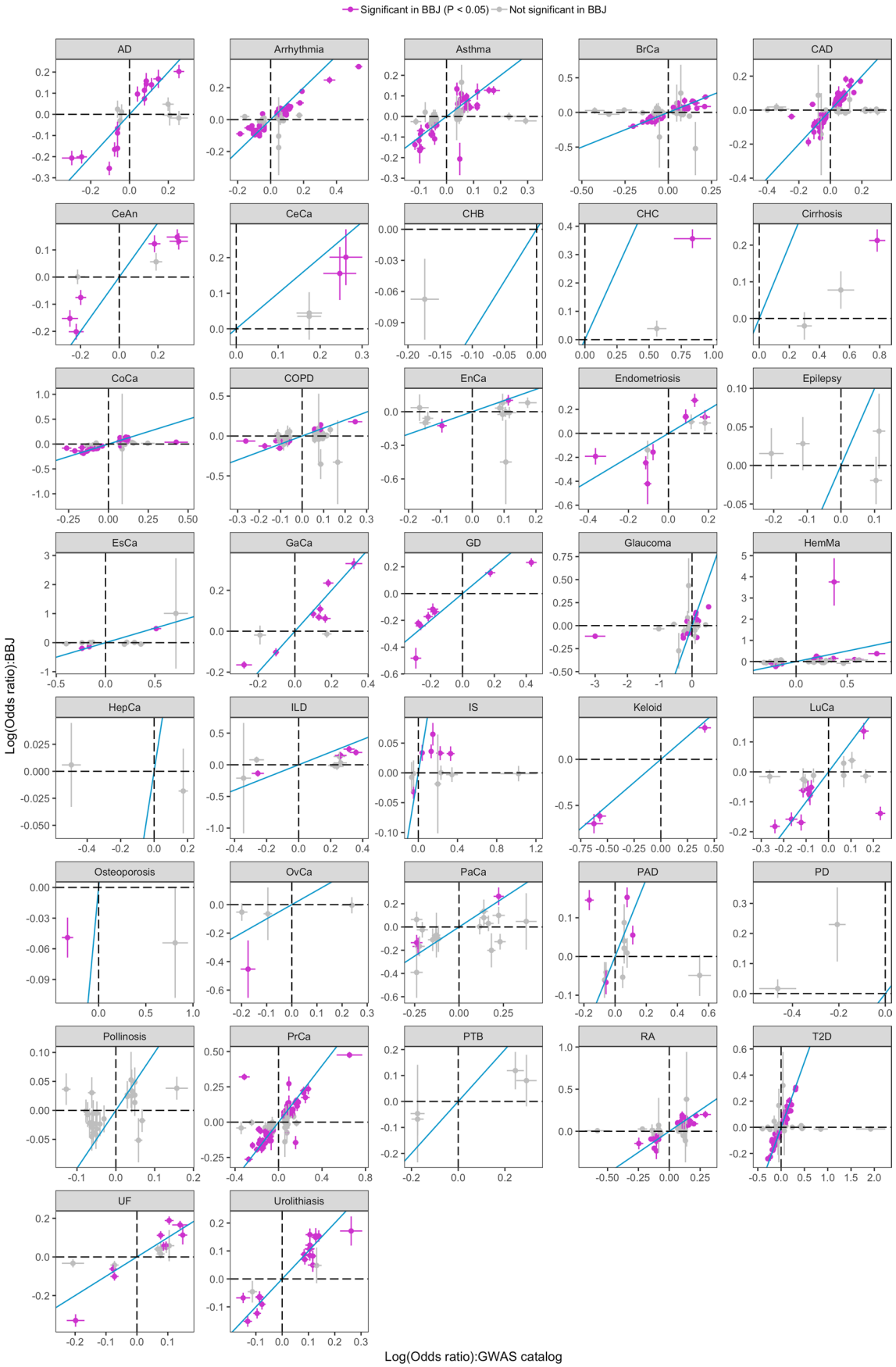
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0640-3.

Correspondence and requests for materials should be addressed to S.R., J.I., T.Y., T.K., M.K. or Y.K.

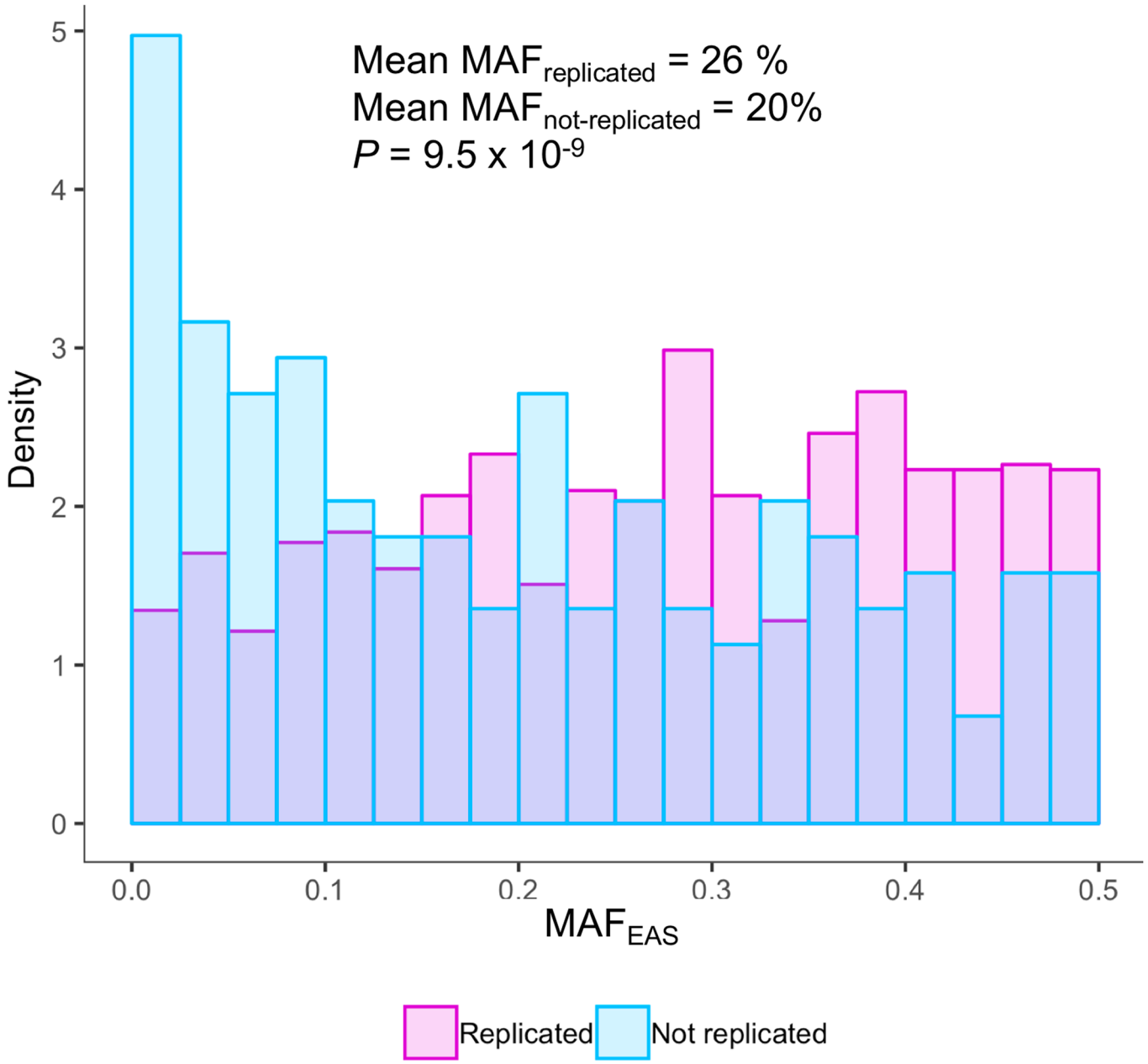
Reprints and permissions information is available at www.nature.com/reprints.



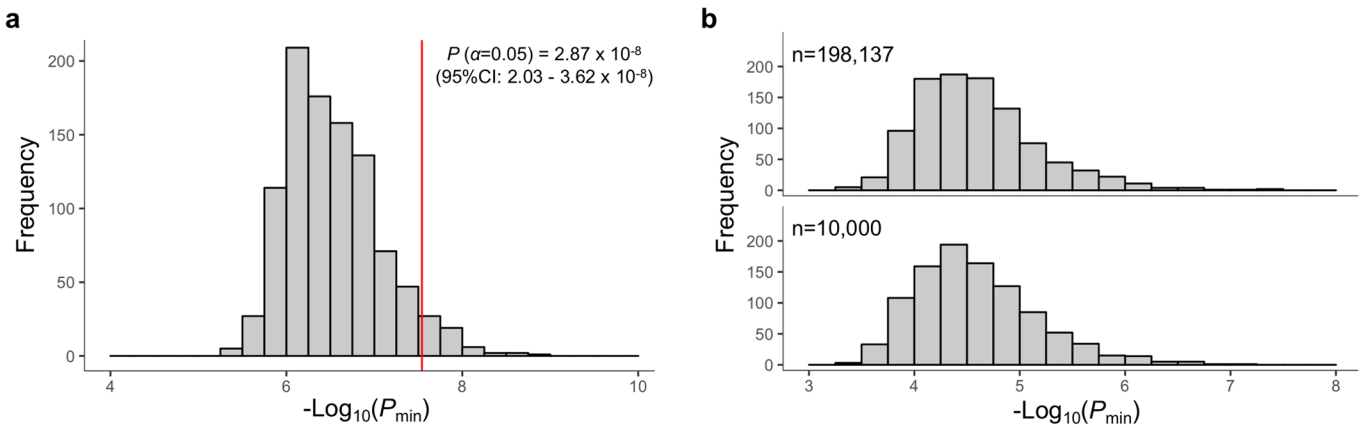
Extended Data Fig. 1 | Study design of this GWAS. a, Study designs in this GWAS. Study design 1 (top) was used in the main analysis. An example of study design 1 is provided; in GWAS of disease 3, we included all other patients (except those have related diseases) into control group. The definition of related diseases is provided in Supplementary Table 1. Study design 2 (bottom) was used to discuss the appropriateness of study design selection. **b**, Effect size estimates and S.E. at the 309 autosomal disease-associated variants detected in sex-combined analysis ($P < 5 \times 10^{-8}$). We compared the effect size estimates in study design 1 with those in study design 2. Heterogeneity between two studies was tested using Cochran's Q test. The identity line is shown in blue. The red dot (rs373205748 associated with arrhythmia) indicates a variant with significant heterogeneity in effect size estimates between two study designs ($P = 0.00012 < 0.05/309$).



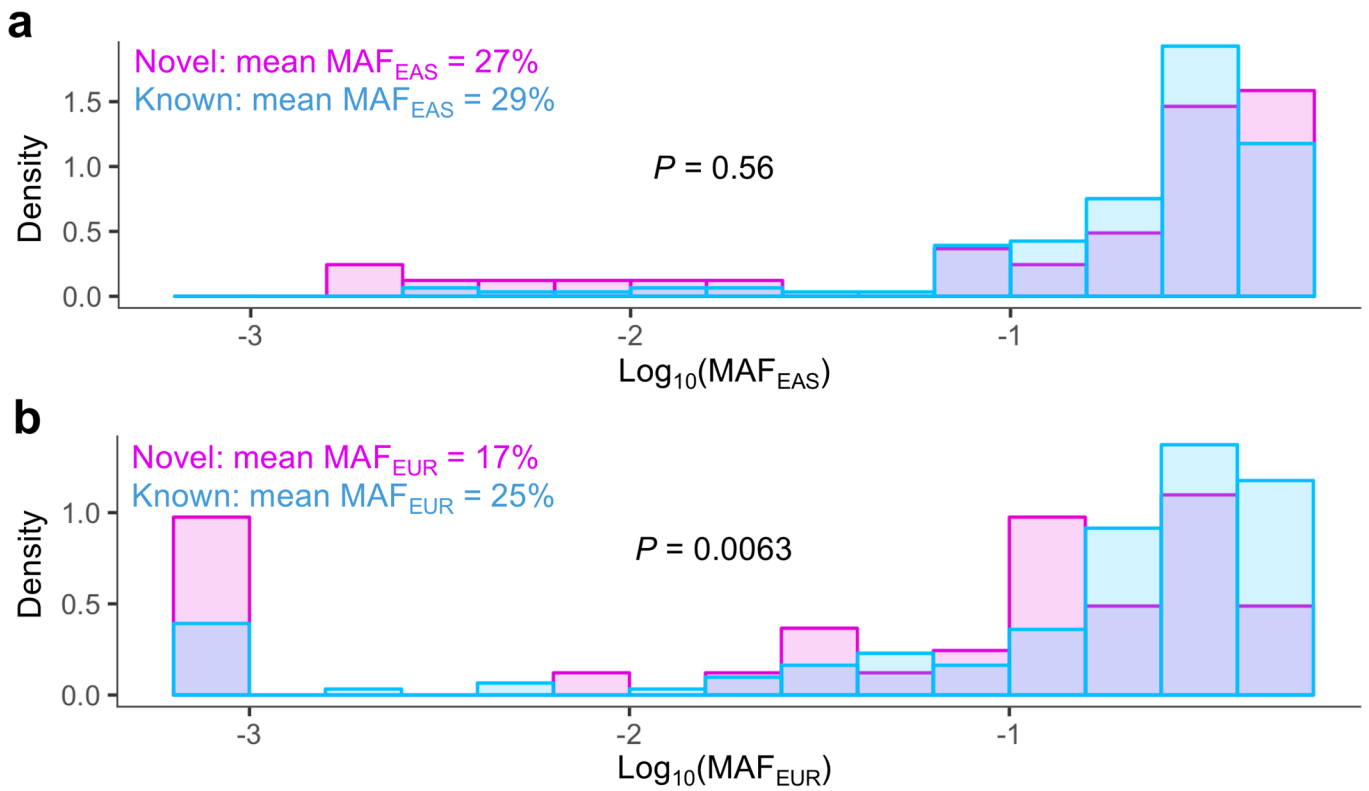
Extended Data Fig. 2 | Replication analysis of previous GWAS findings using this GWAS results. We compared effect sizes reported in the previous GWAS with those in this GWAS. Effect size and S.E. are shown. The identity line is shown in blue. The sample size of GWAS is provided in Table 1. We utilized a generalized linear mixed model in our GWAS.



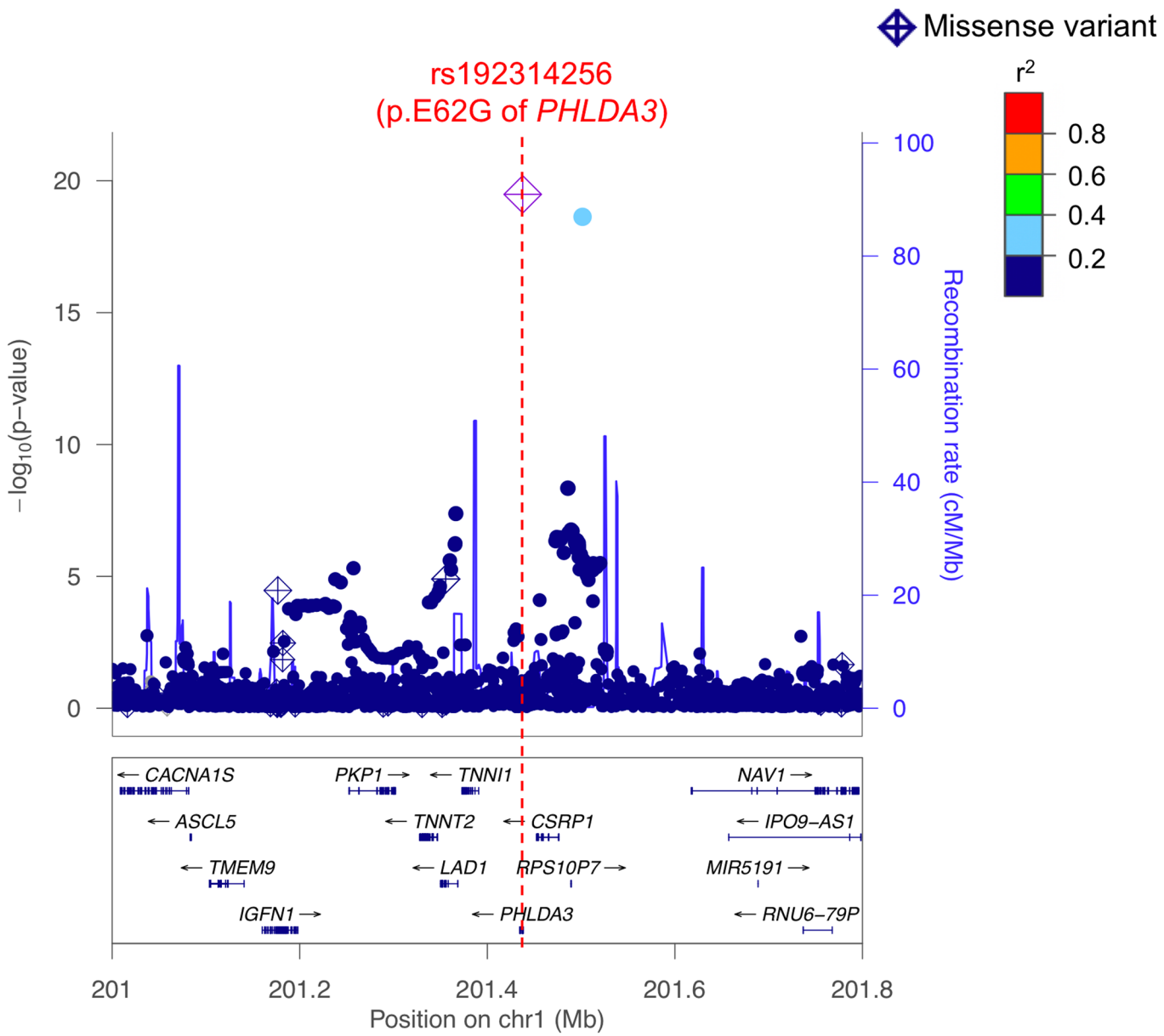
Extended Data Fig. 3 | Low allele frequency might contribute to replication failure. We first compared effect sizes reported in the previous GWAS with those in our GWAS (Supplementary Table 3 and Extended Data Fig. 2); 1,219 out of 1,396 previously reported risk alleles were replicated with the same effect direction (177 alleles were not replicated). We compared MAF of replicated variants (n = 1,219) and MAF of not replicated variants (n = 177). Mann-Whitney U test *P* value is provided (two-sided test).



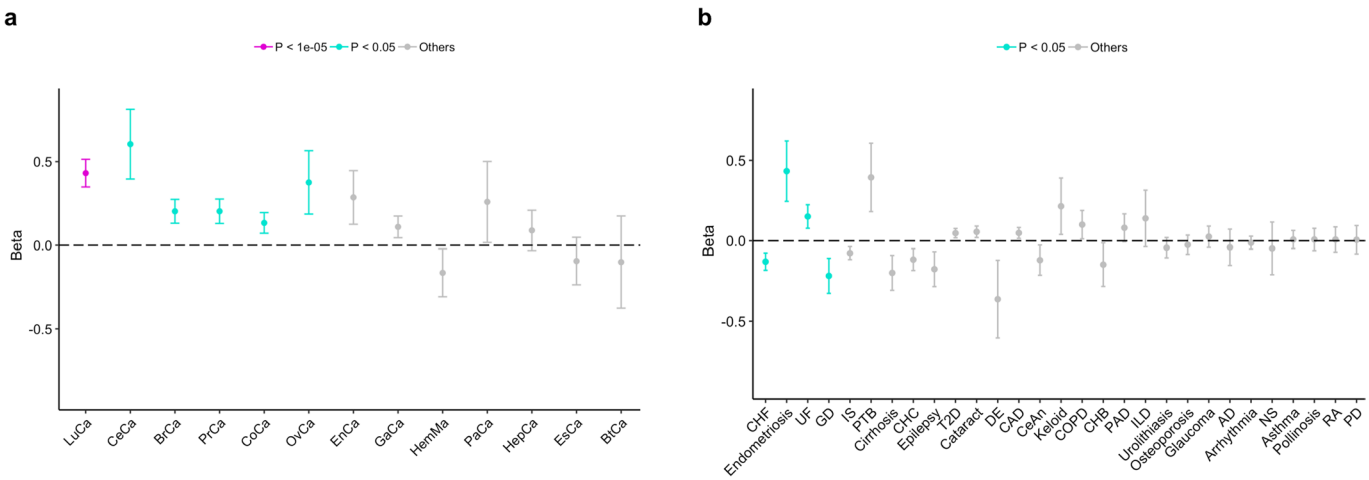
Extended Data Fig. 4 | Permutation test to estimate appropriate P value threshold to control type I errors. Using 1,000 simulated binary phenotypes with down-sampled samples (n = 10,000), we conducted GWAS utilizing the same strategy as used in the main analysis. **a**, The distribution of minimum *P* values in each phenotype (P_{min}). The 95-th percentile of P_{min} was 2.87×10^{-8}. The 95% confidence interval was estimated by 1,000 bootstraps. **b**, The distributions of P_{min} using all samples (n = 198,137) and those using 10,000 samples. To increase computational efficiency, we restricted this analysis to imputed genotype data in chromosome 22. For this analysis in **b**, we utilized Plink2.



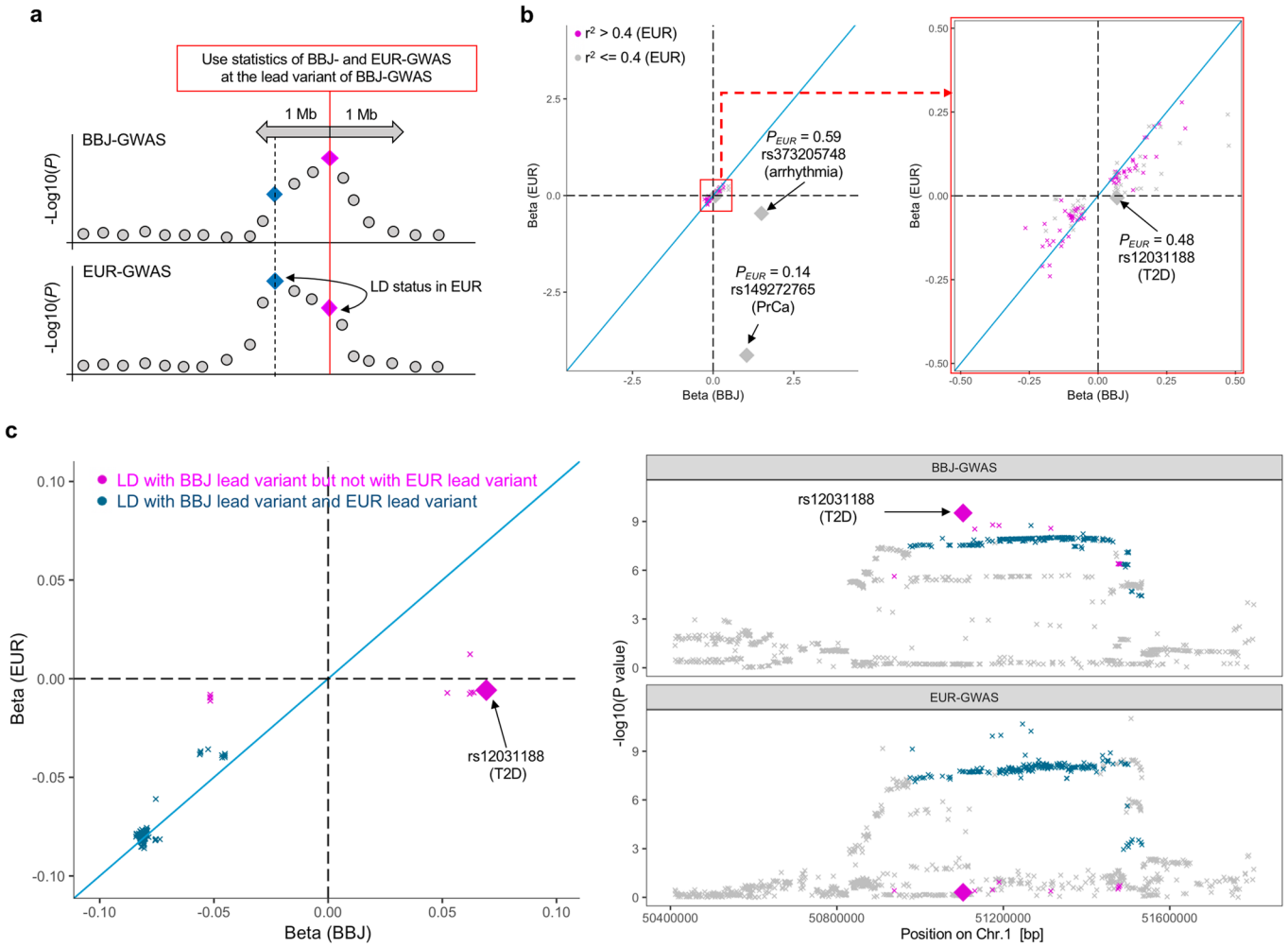
Extended Data Fig. 5 | Allele frequency comparison between novel and known disease-associated variants. MAF comparison at disease-associated variants at novel (n = 41) and known loci (n = 153) with suggestive significance ($P < 5 \times 10^{-8}$) (**a**, East Asian populations; **b**, European populations in 1KG phase3). For known loci, we restricted this analysis to loci where the closest reported variants were discovered by GWAS in European populations. Mann-Whitney U test *P* value is provided (two-sided test).



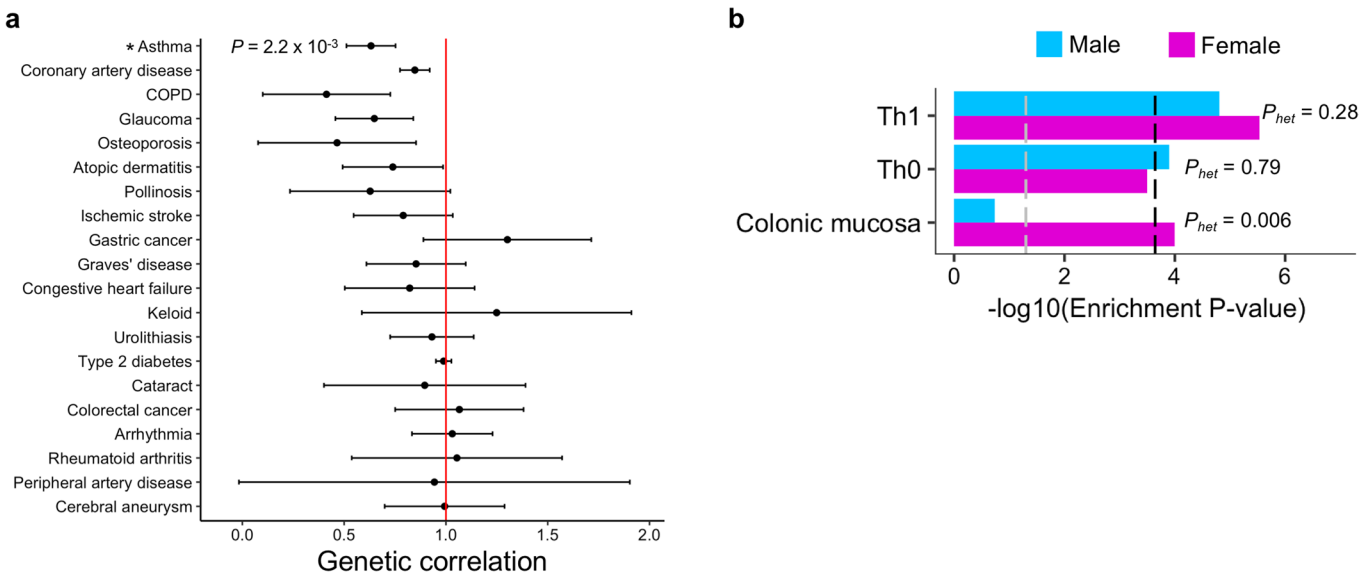
Extended Data Fig. 6 | A novel association which can be explained by an East Asian-specific missense variant. A regional association plot for keloid (812 cases vs 211,641 controls) at the *PHLDA3* region is provided. We utilized a generalized linear mixed model in our GWAS.



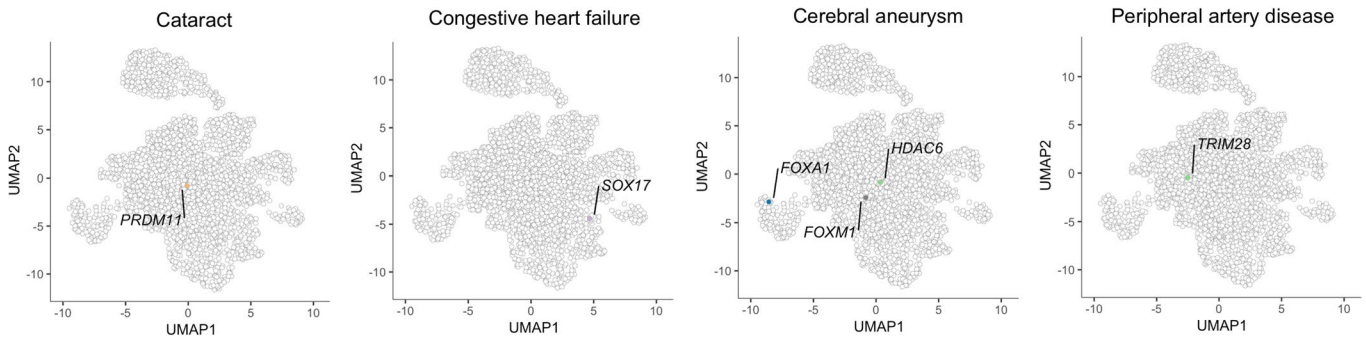
Extended Data Fig. 7 | The association of p.V326A of POT1 for all diseases in this GWAS. Effect size and S.E. are provided for neoplastic diseases (**a**) and non-neoplastic diseases (**b**). The sample size of GWAS is provided in Table 1. We utilized a generalized linear mixed model in our GWAS.



Extended Data Fig. 8 | Comparison of allelic directions between this GWAS and previous European GWAS at known loci. a, Schematic explanations how we compared statistics between BBJ-GWAS and GWAS conducted in European populations (EUR-GWAS). We utilized two inclusion criteria of known loci: (i) EUR-GWAS has significant associations ($P < 5 \times 10^{-8}$) within 1Mb from the BBJ-lead variants and (ii) the BBJ-lead variant is in LD with the lead variant in the European-GWAS ($r^2 > 0.4$ in European samples in 1KG phase3). The first criterion was added to exclude loci where EUR-GWAS has insufficient power (112 known loci remained after applying the first criterion). The second criterion was added because EUR-GWAS statistics at the BBJ-lead variant is not representing those at the EUR-lead variant when they are not in LD. **b**, effect sizes of BBJ- and EUR-GWAS at the BBJ-lead variants. All variants which passed the first criterion were used (n = 112). Variants which passed the second criterion are shown in red (n = 65). Since two variants have extremely large effect size, we provided two plots in different scales. The three variants with the opposite effect directions are marked by large dots, and their details are also provided. **c**, Regional association of T2D around rs12031188. Variants in LD ($r^2 > 0.4$) with BBJ-lead variant (rs12031188) but not with EUR-lead variant are shown in red; Variants in LD ($r^2 > 0.4$) with both lead variants are shown in blue. East Asians and Europeans in 1KG phase3 were used for LD calculation of the BBJ- and the EUR-lead variant, respectively.



Extended Data Fig. 9 | Genetic correlations between male- and female-specific GWAS. a. Genetic correlations between male- and female-specific GWAS. Estimates of genetic correlation and standard errors are provided. *: genetic correlation was significantly different from one (two-sided t test $P = 2.2 \times 10^{-3} < 0.05/20$). **b**. The results of S-LDSC analysis based on sex-specific GWAS of asthma using 220 cell-type specific annotations. Significant annotations in either male or female asthma were shown ($P < 0.05/220$). Heterogeneity was tested by Cochran's Q test, and its P values (P_{het}) were also provided. Black dashed line indicates P value = 0.05/220; grey dashed line indicates P value = 0.05.



Extended Data Fig. 10 | S-LDSC results of four diseases in our GWAS. The results of S-LDSC were plotted on the UMAP space. The significant results (FDR<0.05) were highlighted by cluster-specific colors (the same colors as used in Fig. 4). The names of the top five most significant TFs were also shown on the plot. The results of diseases with less than five significant TF binding site tracks were shown.

nature research

Corresponding author(s): Soumya Raychaudhuri

Last updated by author(s): Apr 23, 2020

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.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	We used publicly available softwares for the analysis. All softwares and their versions were described in the method section. The softwares include SHAPEIT2 (v2.778), minimac3 (v2.0.1), SAIGE (v0.29.4.2), Plink (v2.0), LocusZoom (v1.2), LDSC (v1.0.0), ANNOVAR, SRA Toolkit, Bowtie2 (v2.2.5), and MACS (v2.1).

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. 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

GWAS summary statistics of the 40 diseases are publicly available at our website (JENGER; http://jenger.riken.jp/en/) and the National Bioscience Database Center Human Database (https://humandbs.biosciencedbc.jp/en/; Research ID: hum0014) without any access restrictions. GWAS results for two diseases (breast cancer and coronary artery disease) will be available when this manuscript is accepted. GWAS genotype data for case samples were already deposited at the NBDC Human Database (Research ID: hum0014).

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	We used all available data without sample size calculation because it is expected to detect new loci even after exceeding sample size of millions. We haven't achieved it yet and it will not be realized soon in the complex disease GWAS setting. Current standard practice of GWAS should be to use as much subjects as possible. Therefore, we believe the sample size determination is not applicable.
Data exclusions	We excluded the samples and variants based on the standard quality control procedure in GWAS which is pre-established. Detailed information was described in the method section.
Replication	We did not perform replication GWAS because to perform replication GWAS of all of 42 diseases in Japanese populations was not practical.
Randomization	Randomization is not relevant for our study because we used all of the recruited data and this is a retrospective case-control study.
Blinding	Blinding was not relevant in our study because it is a retrospective case-control study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Population characteristics of BioBank Japan were provided in Supplementary Table 1 and in our previous reports (see below). BioBank Japan is a biobank that collaboratively collects DNA and serum samples from 12 medical institutions in Japan and recruited approximately 200,000 patients with the diagnosis of at least one of 47 diseases. 1: Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. 27, S2–S8 (2017). 2: 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).
Recruitment	BioBank Japan is a bank that collaboratively recruited samples from 12 medical institutions in Japan. All diagnoses are based on the decision of physicians. Details were described in our previous reports (see below). 1: Nagai, A. et al. Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. 27, S2–S8 (2017). 2: 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).
Ethics oversight	All participating studies obtained informed consent from all participants by following the protocols approved by their institutional ethical committees. We obtained approval from ethics committees of RIKEN Center for Integrative Medical Sciences, and the Institute of Medical Sciences, The University of Tokyo.

Note that full information on the approval of the study protocol must also be provided in the manuscript.