Genetic association studies using disease liabilities from deep neural networks

Lu Yang[1, 4, *, +], **Marie C. Sadler**[1, 5, 6, +], **and Russ B. Altman**[1, 2, 3, 4, *]

[1]Deparment of Bioengineering, Stanford University, Stanford, CA, 94305, USA

[2]Department of Genetics, Stanford University, Stanford, CA, 94305, USA

[3]Department of Medicine, Stanford University, Stanford, CA, 94305, USA

[4]Department of Computer Science, Stanford University, Stanford, CA, 94305, USA

[5]University Center for Primary Care and Public Health, Lausanne, 1010, Switzerland

[6]Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland

[*]luy@stanford.edu, Russ.Altman@stanford.edu

[+]these authors contributed equally to this work

ABSTRACT

The case-control study is a widely used method for investigating the genetic landscape of binary traits. However, the health-related outcome or disease status of participants in long-term, prospective cohort studies such as the UK Biobank are subject to change. Here, we develop an approach for the genetic association study leveraging disease liabilities computed from a deep patient phenotyping framework (AI-based liability). Analyzing 44 common traits in 261,807 participants from the UK Biobank, we identified novel loci compared to the conventional case-control (CC) association studies. Our results showed that combining liability scores with CC status was more powerful than the CC-GWAS in detecting independent genetic loci across different diseases. This boost in statistical power was further reflected in increased SNP-based heritability estimates. Moreover, polygenic risk scores calculated from AI-based liabilities better identified newly diagnosed cases in the 2022 release of the UK Biobank that served as controls in the 2019 version (6.2% percentile rank increase on average). These findings demonstrate the utility of deep neural networks that are able to model disease liabilities from high-dimensional phenotypic data in large-scale population cohorts. Our pipeline of genome-wide association studies with disease liabilities can be applied to other biobanks with rich phenotype and genotype data.

Keyword: GWAS, complex traits, polygenic risk scores, AI-based genetic prediction, disease liability

Introduction

The advent of biobanks has greatly reduced the cost and difficulty of collecting large genome-wide data[1,2], and have enabled a more comprehensive understanding of the genetic basis of human health and disease[3]. The UK Biobank (UKBB) database has genotype data for about 500,000 participants and their socio-demographic, lifestyle and health-related phenotypes[4]. It provides a cohort of unparalleled size and scope, allowing the studies of human genetics research[5,6]. Over the past 15 years, genome-wide association studies (GWAS) of continuous and binary traits have provided new insights into the genetic architecture of a wide range of complex diseases[5]. However, while large-scale genotyping helps identify and localize common-to-rare variant signals, the proportion of phenotypic variance captured is limited[7]. In addition, recent studies have reported that the heritability estimates (h^2_{SNP}) based on pedigree studies are higher than the ones computed from SNP-based GWAS hits[8,9]. This gap also labeled as "missing heritability" is substantial and can be due to multiple factors, such as sample size, omitting rare and structural variants, as well as gene-environment interactions[10].

To unravel genetics' role in disease studies, researchers typically compare patients with the disease (the cases) to individuals who are not (the controls). However, patient cohorts may progress beyond a dichotomous setting of "diseased"and "non-diseased". Healthy patients can become diseased years down the line and GWAS studies explicitly modelling age-at-onset showed an increase in statistical power compared to case-control (CC) studies[11–13]. Time-to-event analyses are computationally expensive and alternative approaches have focused on integrating family history to capture the underlying disease liability. Genome-wide association study by proxy (GWAX) where individuals who reported cases within family members are included as cases themselves can identify new risk loci[14]. GWAX has been particularly popular in the study of Alzheimer's disease where 46,828 proxy-cases in addition to 2,447 diagnosed cases in the UKBB contributed to increased statistical power[15,16]. Going beyond a binary phenotype, more refined models have used family history to derive a continuous genetic liability phenotype where CC and family history configurations are differentiated more appropriately[17].

Although previous studies have identified thousands of disease-associated loci, researchers have yet to deeply explore the joint contribution of family history, clinical measures, and lifestyle factors to model disease liabilities. Machine learning techniques have advanced in recent years, allowing to reveal patterns in massive, high-dimensional datasets and capturing non-linear relationships[18,19]. Our study seeks to improve genetic prediction with deep learning-based estimates of disease liability probabilities that leverage joint semantic and structure-based embeddings of phenotypes in the UKBB. To the best of our knowledge, this is the first demonstration of leveraging deep neural networks to present disease phenotypes quantitatively for genetic studies of complex traits.

Recently, we developed a framework POPDx[20] to recognize a comprehensive set of phenotype labels in UK Biobank. Based on this framework, we present a methodology to improve the genetic prediction of complex traits using a continuous representation of the phenotypes based on individual-level data in the UKBB. Applied to 44 complex traits, our method improved statistical power in GWASs as demonstrated through higher heritability estimates, and higher polygenic risk score (PRS) for newly diagnosed cases. We also found many putative disease susceptibility variants for phenotypes such as essential hypertension, breast neoplasm, uterine leiomyoma, hypothyroidism, type 2 diabetes, alcoholism, myocardial infarction, and brain cancer compared to traditional CC studies.

Results

AI-based pipeline for genetic studies of complex diseases

We used a deep learning framework[20] to extract multiple layers of information, including electronic medical records, laboratory tests, physical measurements, answered questionnaires and interviews about the socio-demographic status, lifestyle, and family history of 392,246 individuals in the UK Biobank. The overview of our artificial intelligence (AI)-based genetic analysis is illustrated in Figure 1. First, we used a pre-trained deep learning framework (POPDx)[20] for patient phenotyping in UK Biobank. The performance of the patient phenotyping is shown in Table 1. The AI-based method transformed the binary Phecode labels of the patients into a continuous spectrum that represents the probability of disease assigned to each individual as shown in Figure 1A. Thus, we retrieved the output of POPDx as the disease liability scores for downstream genetic analysis instead of the conventional CC strategy of disease studies. In the next step, we conducted GWAS analyses for 44 common traits using the AI-based disease liability scores on 577,395 directly genotyped SNP markers. Then, we compared the disease-associated risk loci identified by the AI-based approach to the associations from the ascertained cases and controls routine (Figure 1B). In addition, we estimated the heritability scores (h^2_{SNP}) of these phenotypes using the summary statistics from the AI-based and CC GWAS. Lastly, we estimated PRS effects using the liability scale as well as the original CC status. We then focused on new cases that have emerged in a more recent release of the UKBB and compared their PRS percentiles obtained from the AI-based and CC phenotypes. Overall, we seek to quantify the extent to which AI-based phenotyping helps genetic prediction by finding new disease variants, improving heritability estimates, and enhancing PRS derivations.

New risk loci identified in GWAS

We leveraged AI-based disease liabilities to conduct GWAS analyses and discovered independent risk loci for a wide range of complex and polygenic traits. Three methods of patient characterization were used to identify disease-associated risk loci: (1) cases and controls defined by the in-patient diagnosis record of the UK Biobank

Table 1. AUROCs of Patient phenotyping using the POPDx framework

Phenotypes	Phecode	AUROC
Malignant neoplasm of female breast	174.11	0.9929
Varicose veins of lower extremity	454.1	0.9739
Valvular heart disease/ heart chambers	747.12	0.9661
Myocardial infarction	411.2	0.9605
Hematuria	593.0	0.9562
Excessive or frequent menstruation	626.12	0.9535
Coronary atherosclerosis	411.4	0.9525
Hypothyroidism NOS	244.4	0.9489
Uterine leiomyoma	218.1	0.9464
Hyperplasia of prostate	600.0	0.9440
Angina pectoris	411.3	0.9440
Peripheral enthesopathies and allied syndromes	726.0	0.9431
Cholelithiasis	574.1	0.9425
Asthma	495.0	0.9386
Ulceration of intestine	556.1	0.9340
Type 2 diabetes	250.2	0.9325
Benign neoplasm of colon	208.0	0.9281
Essential hypertension	401.1	0.9258
Cancer of brain	191.11	0.9196
Inguinal hernia	550.1	0.9167
Cataract	366.0	0.9059
Atrial fibrillation and flutter	427.2	0.9056
Osteoarthrosis, localized, primary	740.11	0.8984
Hypercholesterolemia	272.11	0.8970
Circulatory disease NEC	459.9	0.8926
Disorders of iris and ciliary body	379.5	0.8823
Uveitis, noninfectious or NOS	371.1	0.8798
Diverticulosis	562.1	0.8794
Benign neoplasm of skin	216.0	0.8700
Enthesopathy	726.1	0.8698
Obesity	278.1	0.8664
Transient cerebral ischemia	433.31	0.8645
Urinary tract infection	591.0	0.8637
Osteoarthritis; localized	740.1	0.8547
Tobacco use disorder	318.0	0.8539
Hemorrhoids	455.0	0.8511
Lung transplant	510.2	0.8376
Reflux esophagitis	530.14	0.8341
Gastritis and duodenitis	535.0	0.8333
Heart transplant/surgery	429.1	0.8317
Diaphragmatic hernia	550.2	0.8300
Abdominal pain	785.0	0.7981
Alcoholism	317.1	0.7793
Allergy/adverse effect of penicillin	960.2	0.7488

dataset (CC), (2) AI-based disease liabilities, (3) a combination of AI-based disease liabilities and case-control status (combined CC-liability). Figure 2 compares the AI-based method and the CC GWAS study using Manhattan plots of two common traits with highlighted risk loci. We identified 772 genome-wide significant SNPs in 114 distinct loci for essential hypertension, compared to 299 significant SNPs in 100 unique loci in the CC analysis (Figure 2A). Figure 2B depicts the GWAS analysis for hypothyroidism, one common thyroid disorder. In addition, we identified 1,932 genome-wide significant associations at 42 distinct loci using AI-based liabilities. The novel loci link the interferon-induced helicase gene (IFIH1), a helix-loop-helix protein gene (BHLHE40), ataxin 2 gene (ATXN2), protein tyrosine phosphatase N2 gene (PTPN2), and an adaptor protein-coding gene in the TLR signaling pathways (TICAM2) to hypothyroidism. Generally, our AI-based findings indicate an increased capability to discover additional genome-wide significant SNP associations in many complex traits (Figure 2C). In addition, the AI-based approach revealed new variants for 37 out of 44 phenotypes that the conventional analyses have missed under a similar experimental condition (sample size, etc.). The AI-based method associates fewer risk loci for specific phenotypes compared to CC-GWAS. However, it still assists in identifying additional risk loci, such as CHRDL2 (Chordin Like 2) and CABLES2 (Cdk5 and Abl enzyme substrate 2) associated with diverticulosis, a disease that affects the large intestine. CHRDL2 has previously been shown to be upregulated in colorectal cancer (CRC) tissues[21]. According to earlier research, CABLES2 has a crucial function in colorectal carcinogenesis[22].

Figure 3 illustrates the improvement of multiple-traits GWAS by combining diagnosed cases with AI-based disease liabilities of the controls. In this scenario, controls are assigned disease liability probabilities (range of [0,1]) whereas cases are assigned a probability of 1. Manhattan plots (Figure 3A) are shown to summarize GWAS results for angina pectoris and type 2 diabetes. We annotate a few genetic variants highly associated with these two traits, but some SNP-gene pairs are only found in the combined approach. Through a combination of cases with AI-based disease liabilities of the controls, we identified more genome-wide risk loci for 29 out of 44 phenotypes compared to the conventional GWAS analyses. ABCG8 (ATP-binding cassette subfamily G member 8) is an instance of a new susceptibility gene for angina pectoris identified by our combined approach. It has been previously reported that ABCG8 variants are risk factors for coronary heart disease[23]. ZC3H4 (zinc finger CCCH domain-containing protein 4) is a candidate gene associated with susceptibility to type 2 diabetes (T2D) identified via the combined approach alone. zinc finger protein. Previous research indicates ZC3H4 having an essential role in mediating non-coding transcription in human cells[24]. It was also previously reported the association ZC3H4 and uncontrolled eating[25]. Additional results from the GWAS studies are provided in the Supplementary Figures 1, 2 and Supplementary Table ST1.

We performed an in-depth investigation of our GWAS results (Supplementary Table GWAS Significant Loci). Most of the significant loci identified by the liability and CC-liability methods have been either reported in our CC analysis or other previous studies. Some SNPs are missed in the CC study but could be validated in the literature. For example, rs6544713 (ABCG8) and rs12190287 (TCF21) identified in our combined approach for coronary atherosclerosis were previously found to be associated with the risk for coronary artery disease[26–28]. For breast cancer (malignant neoplasm of female breast), rs13281615 (PCAT1) is highly associated in both AI-based and combined approach. It was not identified in our CC analysis but was reported in other breast cancer studies[29,30]. Some SNPs such as rs10889338 (DOCK7), rs12916 (HMGCR), rs10455872 (LPA), rs15285 (LPL), rs507666 (ABO), and rs11057830 (SCARB1) were identified to be associated with hypercholesterolemia using the CC-liability method. These SNPs were found to be significant in the other GWAS studies of the UKBB[31].

Increasing heritability estimates

Next, we calculated SNP-based heritability (h^2_{SNP}) estimates of 44 phenotypes using summary statistics from GWAS analysis. Heritability estimates were based on the BLD-LDAK model which assumes that the expected h^2_{SNP} depends on linkage disequilibrium (LD), minor allele frequency (MAF) and 66 functional annotations[32] (Figure 3, Supplementary Table ST2). Across the 44 traits, the average h^2_{SNP} based on 577,395 markers was 0.040 for the AI-based phenotype characterization which was significantly higher than the average of 0.030 for the CC analysis (paired two-sided t-test: p-value = 1.77e-4; Figure 3A). The combination of disease liability and known CC status yielded a h^2_{SNP} mean of 0.038 across traits which was not significantly different from using the liability scale alone (paired two-sided t-test: p-value = 0.27). Across the 44 traits, heritability estimates were higher for 32 traits when using disease liability scores compared to CC and 43 traits had higher h^2_{SNP} when considering the combined CC-liability scores (Figure 3A). On average, the heritability from the combined phenotype characterization was significantly higher than the CC analysis (paired two-sided t-test: p-value = 4.95e-09; Figure 3A).

We further compared the difference in h^2_{SNP} to the difference in identified independent loci between the liability scale and CC-GWAS. Both quantities highly correlate, with essential hypertension benefiting the most from the liability scale in terms of h^2_{SNP} and identified loci (Figure 4B). Comparing the CC-GWAS to the combined CC-liability phenotype definition resulted in similar results (Figure 4C). Using the CC-GWAS as baseline, the number of identified loci, however, was only significantly higher when considering the combined CC-liability phenotype (two-sided Wilcoxon signed rank test across 44 traits: p-value = 4.44e-4) and not with the liability scale alone (two-sided Wilcoxon signed rank test across 44 traits: p-value = 0.35). As an example, 26 loci were identified in the coronary atherosclerosis CC-GWAS, 20 in the liability-scale GWAS, but 30 in the combined CC-liability GWAS.

Recognizing new cases with AI empowered PRS

Placing individuals on a continuous spectrum instead of assigning them a binary CC value holds the promise of capturing who is most at risk for developing the disease (Figure 5A). To test this hypothesis, we identified the newly emerged cases in the latest release of UK Biobank diagnosis records (2022) compared to the previous release (2019; Figure 5B). First, we assessed whether liability scores were indeed higher for emerging cases and then tested whether the same holds for polygenic risk scores (PRSs). For the latter, we estimated predictor (i.e. SNP) effects for the CC and liability phenotype definitions separately, and calculated corresponding PRSs. Thus, each individual was assigned two PRSs - one calculated through CC phenotypes and the other through liability scores. Per phenotype, the corresponding PRS ranks of new cases were then compared side-by-side. Liability scores of new cases were higher for 43 out of the 44 traits when averaging among the individuals, and the difference was statistically significant for 37/44 traits at a Bonferroni-corrected threshold of p-value $< 0.05/44$ (two-sided t-test of liability scores of newly emerged cases vs consistent controls; Supplementary Table ST3). The same trend could be observed on the genetic level (Figure 5C). For 32/44 traits, newly emerged cases significantly ranked higher (i.e. had a higher genetic risk) when the PRS was calculated on the liability scale than with the CC definition (at a p-value $< 0.05/44$; Method section: Polygenic risk score analysis). When translating the improvement in genetic prediction to PRS percentiles, we found that across diseases new cases had on average a 6.2% percentile rank increase (Figure 5D, Supplementary Table ST4).

Discussion

We introduced a new framework for genetic association testing where the phenotypes are modelled as disease liabilities derived from deep neural networks. We demonstrated how integrating a variety of patient data—including family history, electronic medical records, socio-demographic characteristics, lifestyle surveys, physical measures, and laboratory examinations—can significantly contribute to boosting genetic association power and prediction performance. The deep neural networks were trained on the entire UKBB cohort, however, computed phenotype probabilities could differ from the true diagnosis status. Nevertheless, raw model outputs resulted in significantly higher genetic heritabilities (h^2_{SNP}) than using binary diagnosis values. For the 44 traits, the number of identified loci was not significantly higher in this setting than in the CC-GWAS, although it helped discover additional loci that the CC-GWAS has missed. The prediction accuracy of phenotype labels was generally very high (mean AUROC of 90%, Table 1). Although AUROC scores did not correlate with the difference in identified loci between liability and CC-GWAS (two-sided Pearson correlation test p-value equalled 0.87 calculated over the 44 traits), performance

increased when incorporating the true disease state into the modelled disease liabilities (combined CC-liability). With this approach, both heritability estimates and number of identified GWAS loci were significantly higher compared to the CC status analysis.

Overall, the liability scale performed well for traits such as essential hypertension, hypothyroidism, asthma, angina pectoris, obesity, myocardial infarction and breast cancer. On the other hand, diverticulosis and atrial fibrillation gained little from the disease liability representation of the phenotype data, and performance remained worse compared to CC analyses even with the CC-liability approach. We explored the longitudinal changes associated with patients' health states by comparing the 2022 release of UK Biobank data sets to an earlier release (Figure 5A, B). Our method provides an effective patient phenotyping algorithm that can differentiate the newly diseased population from the healthy individuals based on the liability scores. Accordingly, the PRS ranks of new cases from the AI-based liability approach are significantly higher than those computed through the prevailing CC method. Taken together, these results indicate the liability phenotypes better capture the underlying genetic architecture than a static dichotomous classification.

Although our AI-based liability and combined CC-liability approaches increased the statistical power of genetic prediction of common traits in the UK Biobank, they have several limitations. First, patient phenotyping is noisy, resulting in incorrect disease liabilities which in turn impairs the downstream genetic analysis. Second, our strategy is specific to the UKBB cohorts; in addition to family history and medical data, our AI-based model uses lifestyle, environmental exposures, and lab examinations to predict patient phenotypes. These data are not available in some observational studies. Third, for computational reasons, we limited the analysis to directly-genotyped SNPs which may explain why the heritability scores for some diseases are lower than those reported in other research studies[33,34]. We have not applied our approaches to other prospective cohort studies, however, the pipeline could be adapted to other biobanks. On a comparable scale, large population cohorts such as the "All of Us" program[35], China Kadoorie Biobank, and the US Million Veteran Program[36] can be promising resources to implement our genetic analyses. For all these, the assessment of AI-based disease liabilities may improve the power to detect genetic associations.

Methods

UK Biobank data

The UKBB is a cohort with rich phenotype data including International Classification of Diseases, Tenth Revision (ICD-10) codes from which we could derive phecodes[4,37]. Phecodes are curated groups of ICD-10 codes for defining phenotypes[37,38]. We excluded individuals with no available ICD-10 codes in 2019 data version and used remaining

individuals (N = 392,246) to infer phecodes and calculate disease liability probabilities (section POPDx). Genetic data analyses were restricted to unrelated white British individuals that passed additional genetic quality control (QC) filters according to the UKBB (*used.in.pca.calculation* = 1 and *in.white.British.ancestry.subset* = 1 in Sample-QC v2 file). Furthermore, samples were excluded if the genetically inferred sex did not match the self-reported sex and if they had withdrawn their consent. Analyses were performed on autosomal genotype data provided by the UK Biobank with MAF > 0.01 (649,644 SNP markers, version 2). Further quality filters and overlapping with external datasets (i.e., SNP annotations) could reduce this number to 577,395 markers. All reported genetic analyses are for 261,807 unrelated white British individuals on *geq* 577,395 genetic markers. Sex-specific traits were analysed on 142,844 females and 118,963 males, respectively.

Phenotype data

ICD billing codes are commonly utilized to select individuals from large observational datasets such as the UK Biobank. Also, cases of multiple ICD codes are often put together as a group of patients to study a specific disease. Phecodes, combinations of relevant ICD codes used to describe phenotypes, have worked well in clinical research[38]. In this study, we retrieved 44 specific phecodes that associate with more than 10,000 patients to represent 44 common diseases from the 2019 download of phenotype data. To obtain diagnostic labels for all the individuals in the UK Biobank, we mapped International Classification of Diseases Tenth Revision (ICD-10) codes to phecodes[37,39,40]. From the 2019 downloads, we received phenotype data from the UK Biobank. Again, we obtained an updated release of phenotype data in early 2022. On average, participants have more ICD-10 codes in their records. For each of the 44 complex traits, we have identified the individuals that are no longer part of the control group. For PRS validation, the new cases are compared to the controls that do not have the diagnostic code for both data lookups.

Modelling phenotype probabilities

POPDx is an existing AI-based methodology that outputs the phenotype liability probabilities for each UK Biobank participant[20]. We trained the POPDx framework on the phenotypic and health-related patient data. The 1538 phecodes as gold standard labels correspond to 12,803 ICD-10 diagnosis codes from UKBB. We trained and tested POPDx on all the individuals based on the 2019 version of phenotype data.

GWAS calculation

We performed GWAS analyses on all 44 complex traits for three phenotype definitions: i) original case-control status, ii) POPDx probability, iii) combined case-POPDx phenotype. GWAS were performed using linear regression as implemented in the LDAK software (v5.2, `www.ldak.org`[41]). Prior regression, phenotypes were corrected for

sex, age, age_2, and PCs 1–40. The analysis was performed on 649,644 genetic markers with minor allele frequencies (MAF) ≥ 0.01.

We performed clumping on SNPs passing the genome-wide significance threshold of 5e-8 to obtain independent SNP associations ($r^2 < 0.05$ within a window of 1000kB). Clumping was performed using LDAK and as reference panel we used the genotypes of 20,000 randomly selected individuals among the included 261,807 samples.

Mapping SNPs to genes

We mapped the SNPs to genes under the guidance of a comprehensive reference annotation (GENCODE GRCh37 release 42)[42]. The GENCODE consortium provides public access to the most recent annotation of human reference genomes[42]. For each SNP, we searched through the genomic locations (genomic start and end positions) of all the protein-coding genes and non-coding RNA genes. The SNPs need to lie within the target genes to be annotated.

SNP-based heritability estimation

We calculated SNP heritability (h^2_{SNP}) of each trait for each of the three phenotype definitions (see GWAS calculation) using the BLD-LDAK model[32]. BLD-LDAK estimates h^2_{SNP} assuming that the expected h^2_{SNP} varies with linkage disequilibrium (LD) and minor allele frequency (MAF) while also taking into account 66 functional annotations. Estimates were calculated using SumHer (part of the LDAK software) and pre-computed tagging files designed for use with LDAK v5.2. Provided tagging files were computed on 2,000 white British individuals from the UK Biobank and we chose the set of non-ambiguous directly genotyped SNPs (MAF > 0.01, 577,525 markers of which 577,395 overlapped with our GWAS summmary statistics).

Polygenic risk score analysis

For each of the 261,807 individuals included in our analyses we calculated polygenic risk scores (PRS) for each trait. Per trait, two PRS were assigned to each person - the first based on original case-control status and the second based on POPDx probabilities. Since our aim was to compare which of the two phenotype definitions assigns a lower percentile rank (i.e., higher PRS) to emerging disease cases, estimation of per-predictor effect sizes and phenotype prediction were performed on the same individuals. Information as to whether an individual turns into a case is not available to either phenotype definition, and thus comparison at a later time point constitutes an independent validation analysis.

Estimation of per-predictor effect sizes was done on individual-level data (N = 261,807) following the LDAK-Bolt-Predict procedure as implemented in the LDAK software v5.2[33]. First, we estimated per-predictor heritabilities which requires the summary statistic of the trait and phenotype definition of interest (see GWAS calculation),

a SNP tagging file and corresponding heritability matrix. The SNP tagging file is calculated according to the assumed underlying heritability model which is the BLD-LDAK model in the LDAK-Bolt-Predict procedure. This model makes use of 65 functional annotations of which 64 can be downloaded. The last annotation, also called LDAK weightings, takes into account the LD structure and was calculated on the genetic data of a subset of 5,000 randomly selected individuals (as recommended to reduce computational demands). The tagging file and heritability matrix were then calculated with recommended settings (power = 0.25, window-kb = 1200) on the same subset of 5,000 individuals, and per-predictor heritabilities were derived by integrating the summary statistics data. Next, per-predictor effect sizes were estimated with the bolt predict command in a cross-validation (cv) fashion (suggested cv-proportion of 0.1). Effect size estimation was done using individual-level genetic data of the entire dataset (N = 261,807), covariate-adjusted phenotypes (see GWAS calculation) and the per-predictor heritabilities. The tagging file and heritability matrix were the same in each calculation, but per-predictor heritabilities and estimated effect sizes were specific to each trait and respective phenotype definition (i.e., POPDx probability or original case-control status). Finally, we estimated PRS for each of the 261,807 individuals by projecting their genetic data onto the estimated effect sizes.

Per trait, we compared the two phenotype definitions by concentrating on the newly emerged cases (control in the earlier release and case in the new release). To this end, we first ranked the individuals relative to their PRS in the entire dataset. Each new case would thus have two ranks, once calculated from POPDx probabilities and once from the original case-control classification. For each trait, the difference in ranks of the new cases were then compared through a two-sided Wilcoxon signed-rank test.

References

1. Orliac, E. J. *et al.* Improving gwas discovery and genomic prediction accuracy in biobank data. *Proc. Natl. Acad. Sci.* **119**, e2121279119 (2022).

2. Greely, H. T. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annu. Rev. Genomics Hum. Genet.* **8**, 343–364 (2007).

3. Zhou, W. *et al.* Global biobank meta-analysis initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).

4. Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

5. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in uk biobank. *Nat. genetics* **50**, 1593–1599 (2018).

6. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the uk biobank. *Nat. genetics* **53**, 942–948 (2021).

7. Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).

8. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of gwas discovery. *The Am. J. Hum. Genet.* **90**, 7–24 (2012).

9. Gusev, A. *et al.* Quantifying missing heritability at known gwas loci. *PLoS genetics* **9**, e1003993 (2013).

10. Génin, E. Missing heritability of complex diseases: case solved? *Hum. Genet.* **139**, 103–113 (2020).

11. Syed, H., Jorgensen, A. L. & Morris, A. P. Evaluation of methodology for the analysis of 'time-to-event'data in pharmacogenomic genome-wide association studies. *Pharmacogenomics* **17**, 907–915 (2016).

12. Staley, J. R. *et al.* A comparison of cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *Eur. J. Hum. Genet.* **25**, 854–862 (2017).

13. Ojavee, S. E. *et al.* Genomic architecture and prediction of censored time-to-event phenotypes with a bayesian genome-wide analysis. *Nat. communications* **12**, 1–17 (2021).

14. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case–control association mapping by proxy using family history of disease. *Nat. genetics* **49**, 325–331 (2017).

15. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nat. genetics* **51**, 404–413 (2019).

16. Bellenguez, C. *et al.* New insights into the genetic etiology of alzheimer's disease and related dementias. *Nat. genetics* **54**, 412–436 (2022).

17. Hujoel, M. L., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nat. genetics* **52**, 541–547 (2020).

18. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. communications* **11**, 1–9 (2020).

19. Thangaraj, P. M., Gisladottir, U. & Tatonetti, N. P. Medical data and machine learning improve power of stroke genome-wide association studies. *bioRxiv* (2020).

20. Yang, L., Wang, S. & Altman, R. Popdx: an automated framework for patient phenotyping across 392 246 individuals in the uk biobank study. *J. Am. Med. Informatics Assoc. JAMIA* ocac226–ocac226 (2022).

21. Sun, J. *et al.* Overexpression of colorectal cancer oncogene chrdl2 predicts a poor prognosis. *Oncotarget* **8**, 11489 (2017).

22. Guo, X. *et al.* Identifying novel susceptibility genes for colorectal cancer risk from a transcriptome-wide association study of 125,478 subjects. *Gastroenterology* **160**, 1164–1178 (2021).

23. Wu, G. *et al.* Abcg5/8 variants are associated with susceptibility to coronary heart disease. *Mol. Medicine Reports* **9**, 2512–2520 (2014).

24. Estell, C., Davidson, L., Steketee, P. C., Monier, A. & West, S. Zc3h4 restricts non-coding transcription in human cells. *Elife* **10**, e67305 (2021).

25. Cornelis, M. C. *et al.* Obesity susceptibility loci and uncontrolled eating, emotional eating and cognitive restraint behaviors in men and women. *Obesity* **22**, E135–E141 (2014).

26. Webb, T. R. *et al.* Systematic evaluation of pleiotropy identifies 6 further loci associated with coronary artery disease. *J. Am. Coll. Cardiol.* **69**, 823–836 (2017).

27. Miller, C. L. *et al.* Coronary heart disease-associated variation in tcf21 disrupts a mir-224 binding site and mirna-mediated regulation. *PLoS genetics* **10**, e1004263 (2014).

28. Wang, Y. *et al.* Genetic variants associated with myocardial infarction and the risk factors in chinese population. *PloS one* **9**, e86332 (2014).

29. Gong, W.-F. *et al.* Single nucleotide polymorphism 8q24 rs13281615 and risk of breast cancer: meta-analysis of more than 100,000 cases. *PLoS One* **8**, e60108 (2013).

30. Zhang, Z. *et al.* Frequent mutation of rs13281615 and its association with pvt1 expression and cell proliferation in breast cancer. *J. genetics genomics* **41**, 187–195 (2014).

31. Cortes, A., Albers, P. K., Dendrou, C. A., Fugger, L. & McVean, G. Identifying cross-disease components of genetic risk across hospital data in the uk biobank. *Nat. genetics* **52**, 126–134 (2020).

32. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).

33. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. communications* **12**, 1–9 (2021).

34. Heritability of $> 4,000$ traits & disorders in uk biobank. http://www.nealelab.is/uk-biobank/, DOI: 10.5281/zenodo.7186871. Accessed: 2022-12-01.

35. All of Us Research Program Investigators. The "all of us" research program. *New Engl. J. Medicine* **381**, 668–676 (2019).

36. Gaziano, J. M. *et al.* Million veteran program: A mega-biobank to study genetic influences on health and disease. *J. clinical epidemiology* **70**, 214–223 (2016).

37. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-wide association studies as a tool to advance precision medicine. *Annu. review genomics human genetics* **17**, 353 (2016).

38. Bastarache, L. Using phecodes for research with the electronic health record: from phewas to phers. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).

39. Wu, P. *et al.* Mapping icd-10 and icd-10-cm codes to phecodes: workflow development and initial evaluation. *JMIR medical informatics* **7**, e14325 (2019).

40. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and icd-9-cm codes for phenome-wide association studies in the electronic health record. *PloS one* **12**, e0175508 (2017).

41. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide snps. *The Am. J. Hum. Genet.* **91**, 1011–1021 (2012).

42. Frankish, A. *et al.* Gencode 2021. *Nucleic acids research* **49**, D916–D923 (2021).

Funding

This work was supported by NIH GM102365, Agilent and the Chan-Zuckerberg Biohub. MCS was supported by the Swiss National Science Foundation (310030-189147) and would like to thank the Fulbright Program for funding her research stay at Stanford University.

Acknowledgements

The authors would like to thank Dr. Martin Jinye Zhang for the helpful discussion regarding human genome annotation.

Author contributions statement

LY, MCS, and RBA conceived the original idea. LY and MCS carried out the experiments, and conducted the analyses. RBA supervised the project. All authors revised and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Figures

Figure 1. Overview of our AI-based genetic analysis. (A) Patient phenotyping in UK Biobank is the first step in this study's workflow. After extracting the data, we preprocess diverse patient information such as laboratory results, medication intake, answered questionnaires, medical records, and family history. This probability distribution for each disease of interest is obtained by using an AI-based framework to train the one-hot encoding matrix of patient embeddings with the corresponding binary (case-control) diagnostic labels. The outputs of the patient phenotyping framework are continuous values between 0 and 1, representing the probabilities of complex traits that belong to an individual (disease liability). (B) Three different methods are used to conduct genome-wide association studies of 44 complex traits, including (i) case and control status of the patient (CC), (ii) AI-based disease liabilities, (iii) combined CC and AI-based disease liabilities. SNP-heritability estimates of 44 traits are computed based on the GWAS summary statistics. (C) We identify new cases of diseases based on the 2019 and 2022 releases of phenotypic data and then compare the polygenic risk scores of these newly identified cases calculated from the AI-based liabilities and case-control status of the patients. Figure created with `BioRender.com`.

Figure 2. GWAS analysis using the AI-based disease liabilities. Mirrored Manhattan plots of GWAS results for (A) essential hypertension and (B) hypothyroidism. The CC-GWAS results are shown in the bottom panel with annotated risk loci. The GWAS analyses conducted through the AI-based disease liability are shown in the top panel with highlighted risk loci that are not significant in the CC-GWAS. Chromosome position is plotted in genomic order on the x-axis. The y-axis indicates the strength of the association (P values). The dashed horizontal line in red denotes the genome-wide significance threshold of 5×10^{-8}. (C) Comparing the number of novel risk loci identified in the AI-based approach with the case-control (CC) associations across 44 complex traits. The top panel illustrates the difference in risk loci counts in descending order (blue to red). The lower panel shows the number of independent variants only significant in the AI-based approach.

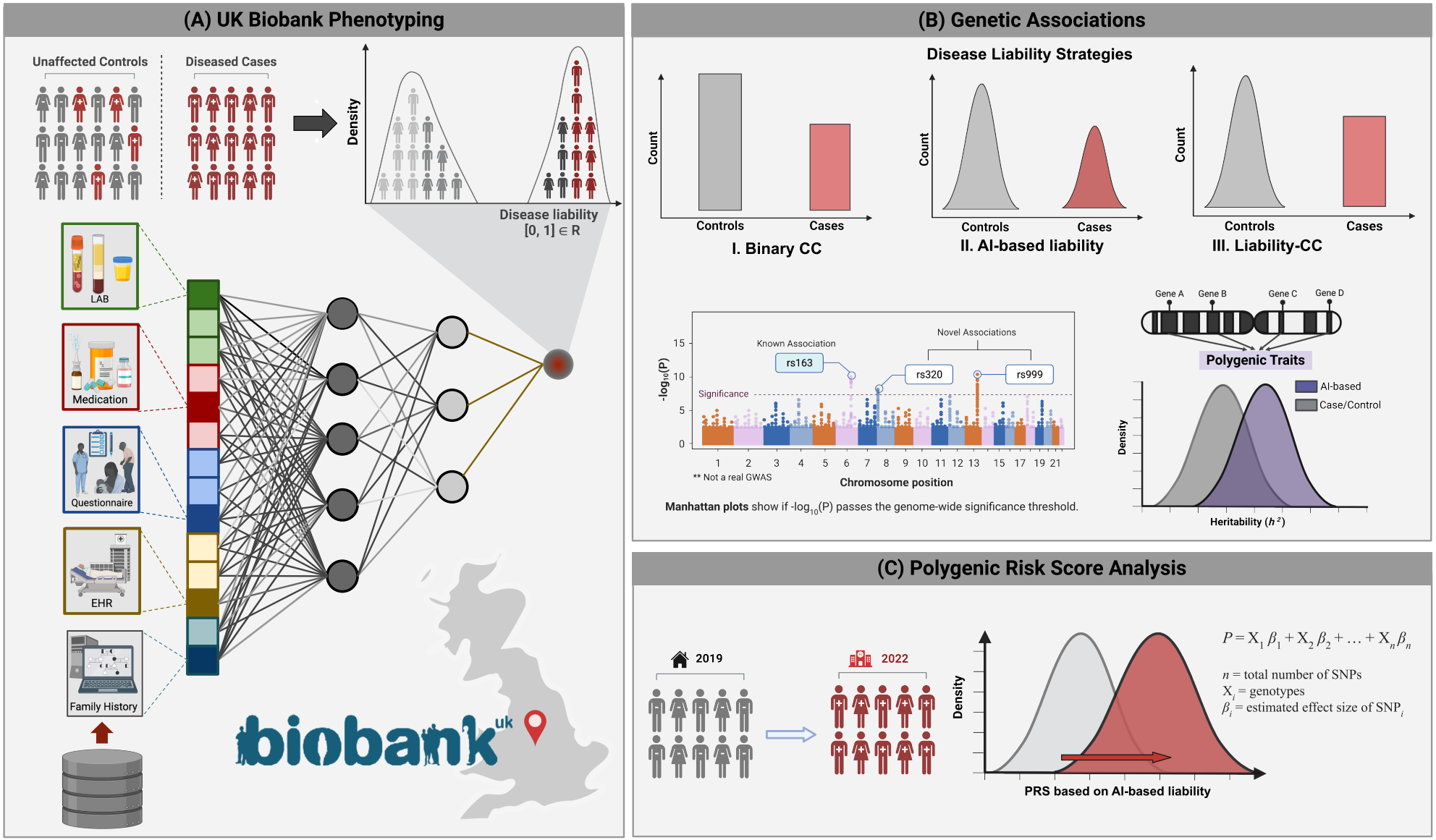
Figure 3. GWAS analysis using the combination of cases and AI-based disease liabilities of the controls. Mirrored Manhattan plots of GWAS results for (A) angina pectoris and (B) type 2 diabetes. The CC-GWAS results are shown in the lower panel with annotated risk loci. The GWAS analyses conducted via the liability-CC method are shown in the upper panel with highlighted risk loci that are not significant in the CC-GWAS. Chromosome position is plotted in genomic order on the x-axis. The y-axis indicates the strength of the association (P values). The dashed horizontal line in red denotes the genome-wide significance threshold of 5×10^{-8}. (C) Comparing the number of novel risk loci identified in the combined approach with the case-control (CC) associations across 44 complex traits. The upper panel illustrates the difference in risk loci counts in descending order (blue to red). The lower panel shows the number of independent variants only significant in the combined approach.

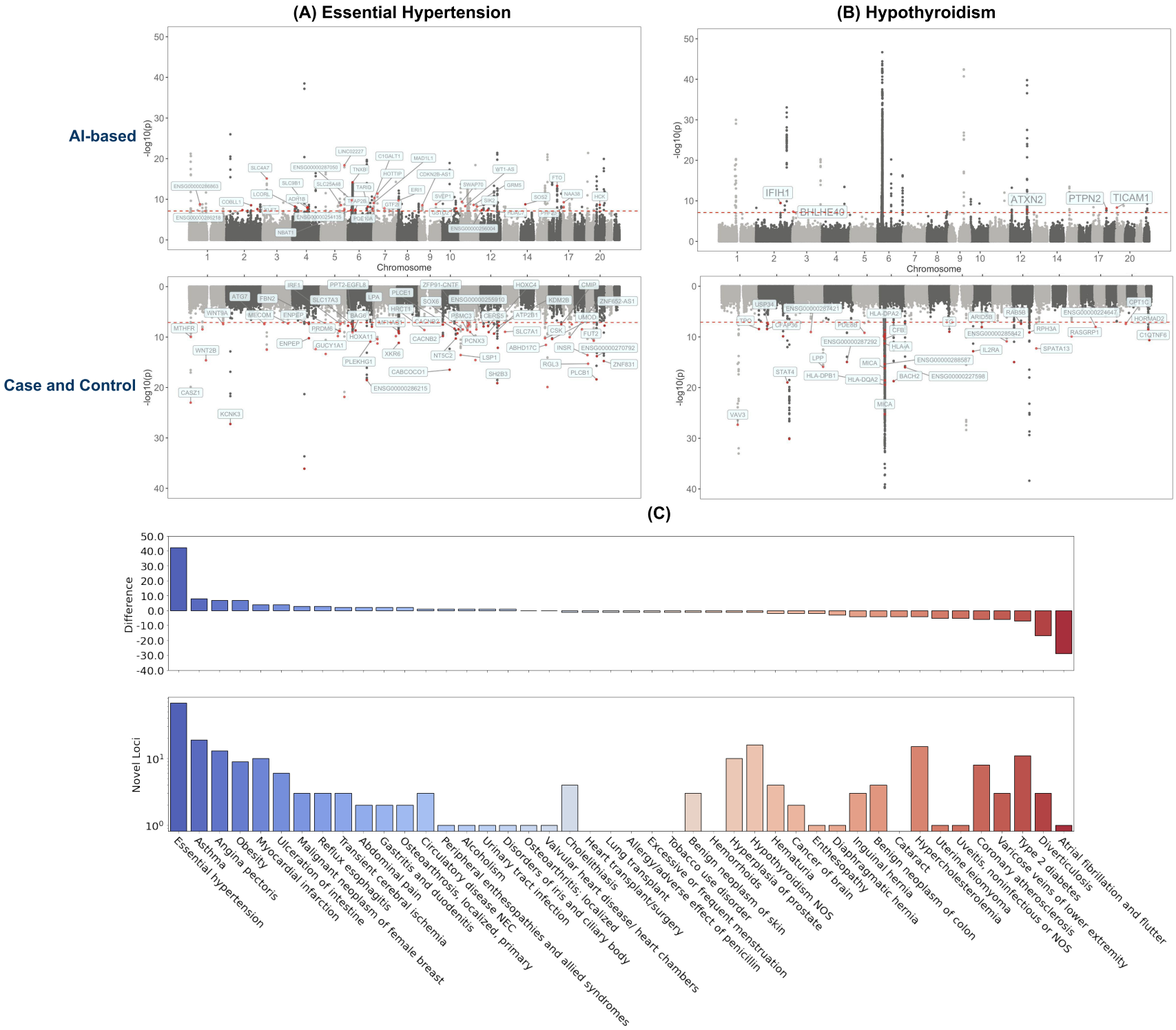
Figure 4. Heritability estimates (h^2_{SNP}) in comparison with identified GWAS loci. (A) Bar plot representing SNP-heritability estimates utilizing three different strategies, including cases-control (blue), combined (orange), and AI-based (grey). (B) We plot the increase in number of independent loci (x-axis) against the increase in heritability estimates by the AI-based approach in comparison with the case-control method. (C) We plot the increase in number of independent loci (x-axis) against the increase in heritability estimates by the combined approach in comparison with the case-control method.

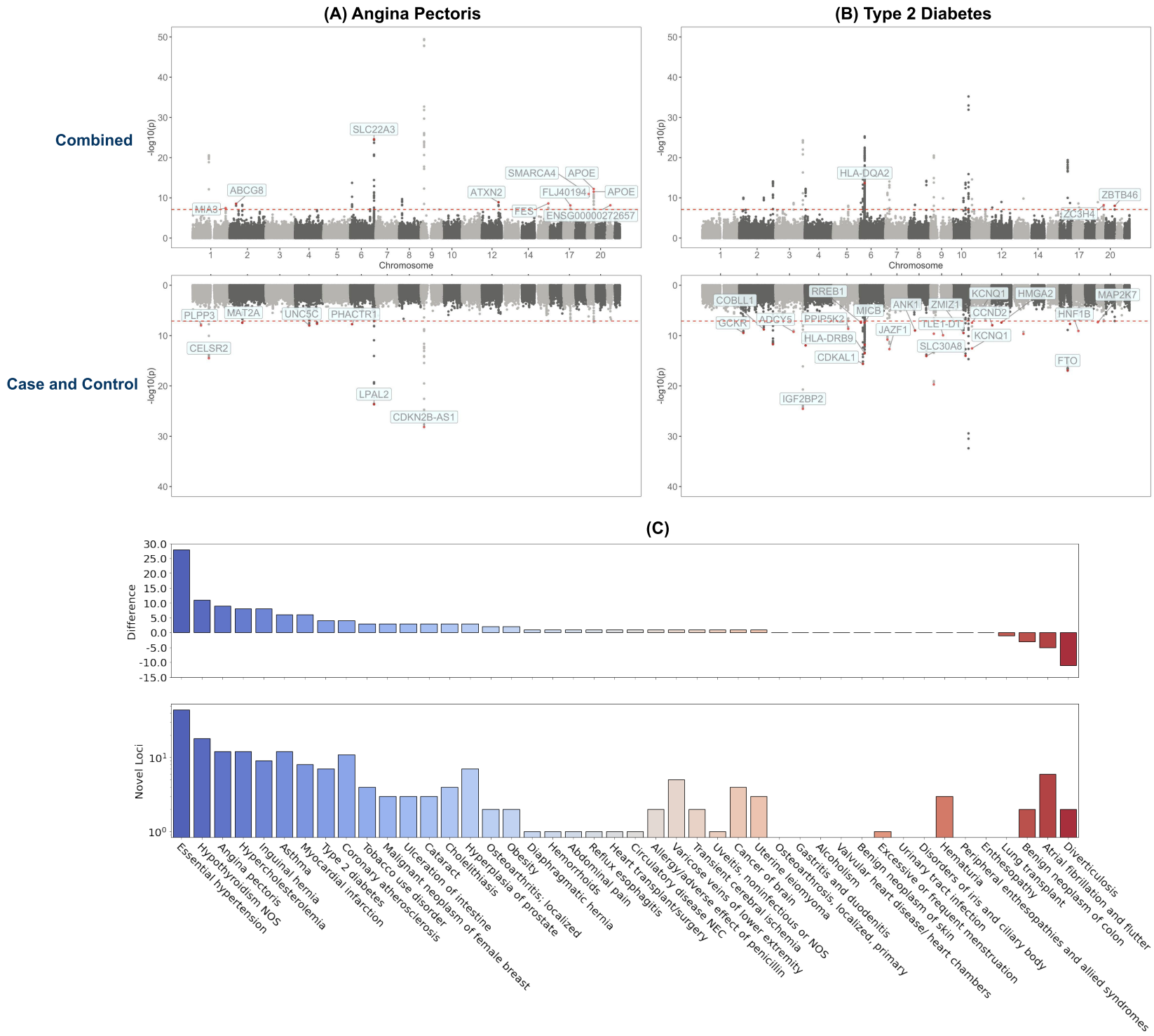
Figure 5. PRS analysis applied on newly diagnosed cases in the UKBB. (A) Bar plot showing the difference in case count per trait between the 2019 and 2022 release of the UKBB. (B) Scatter plot comparing the absolute case count per trait between the two releases. (C) Scatter plot comparing the difference in liability scores between newly diagnosed cases and persistent controls (x-axis) and the significance in PRS difference for new cases between CC and liability estimates (y-axis). The PRS difference results from two-sided Wilcoxon signed-rank tests (see Methods) and is plotted as a positive value if liability PRS allowed to rank newly diagnosed cases higher than CC PRS. (D) Difference in PRS percentile per trait between liability and CC PRS calculations. PRS percentiles were calculated relative to the entire dataset and the difference displayed here was calculated on newly diagnosed cases.

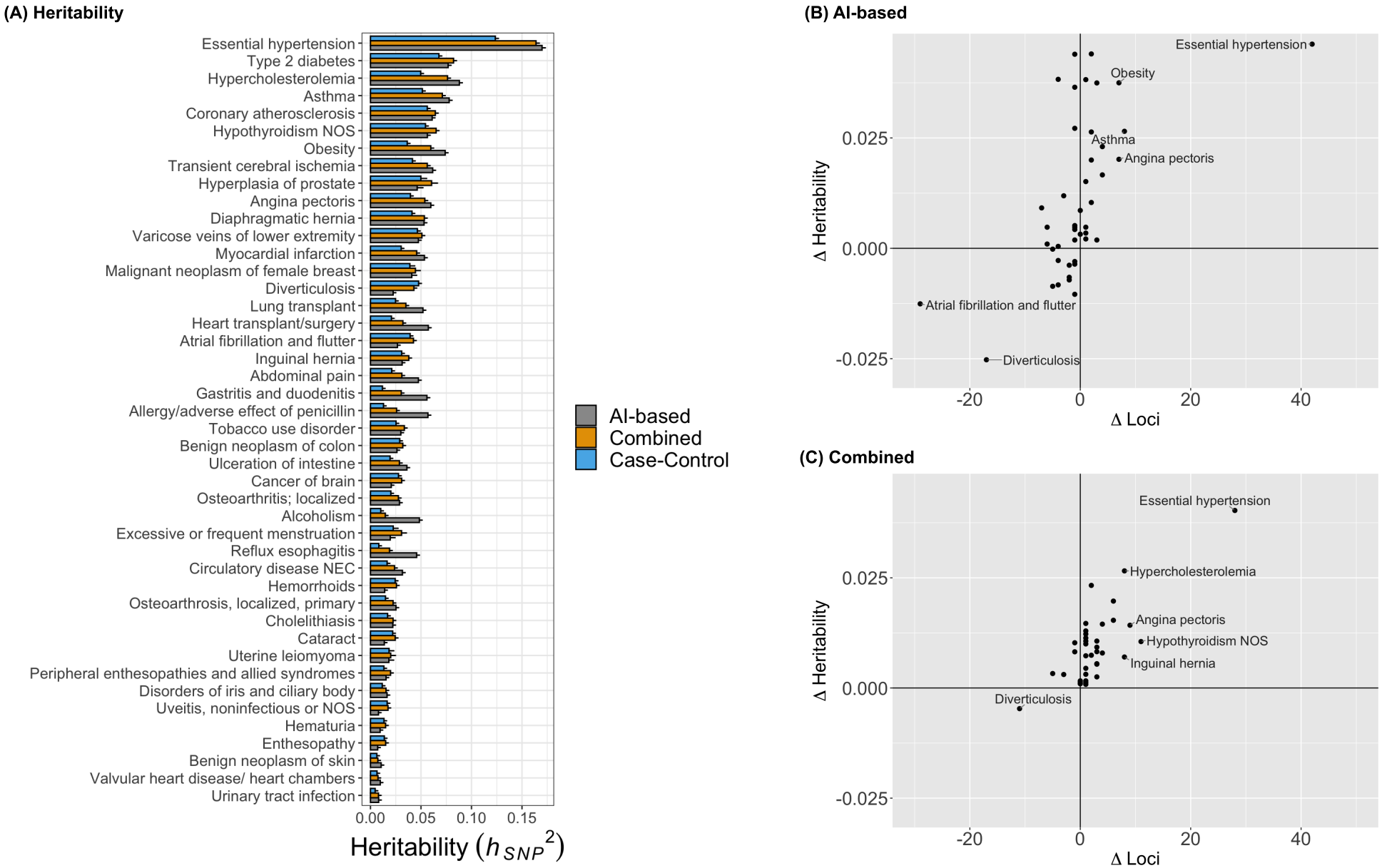
Supplementary Figure 1. Mirrored Manhattan plots of GWAS analysis using the AI-based disease liabilities. Mirrored Manhattan plots of GWAS results for (A) cancer of Brain, (B) obesity, (C) disorders of iris and ciliary body, (D) myocardial infarction, (E) transient cerebral ischemia, and (F) ulceration of intestine. The CC-GWAS results are shown in the bottom panel with annotated risk loci. The GWAS analyses conducted through the AI-based disease liability are shown in the top panel with highlighted risk loci that are not significant in the CC-GWAS. Chromosome position is plotted in genomic order on the x-axis. The y-axis indicates the strength of the association (P values). The dashed horizontal line in red denotes the genome-wide significance threshold of 5×10^{-8}.

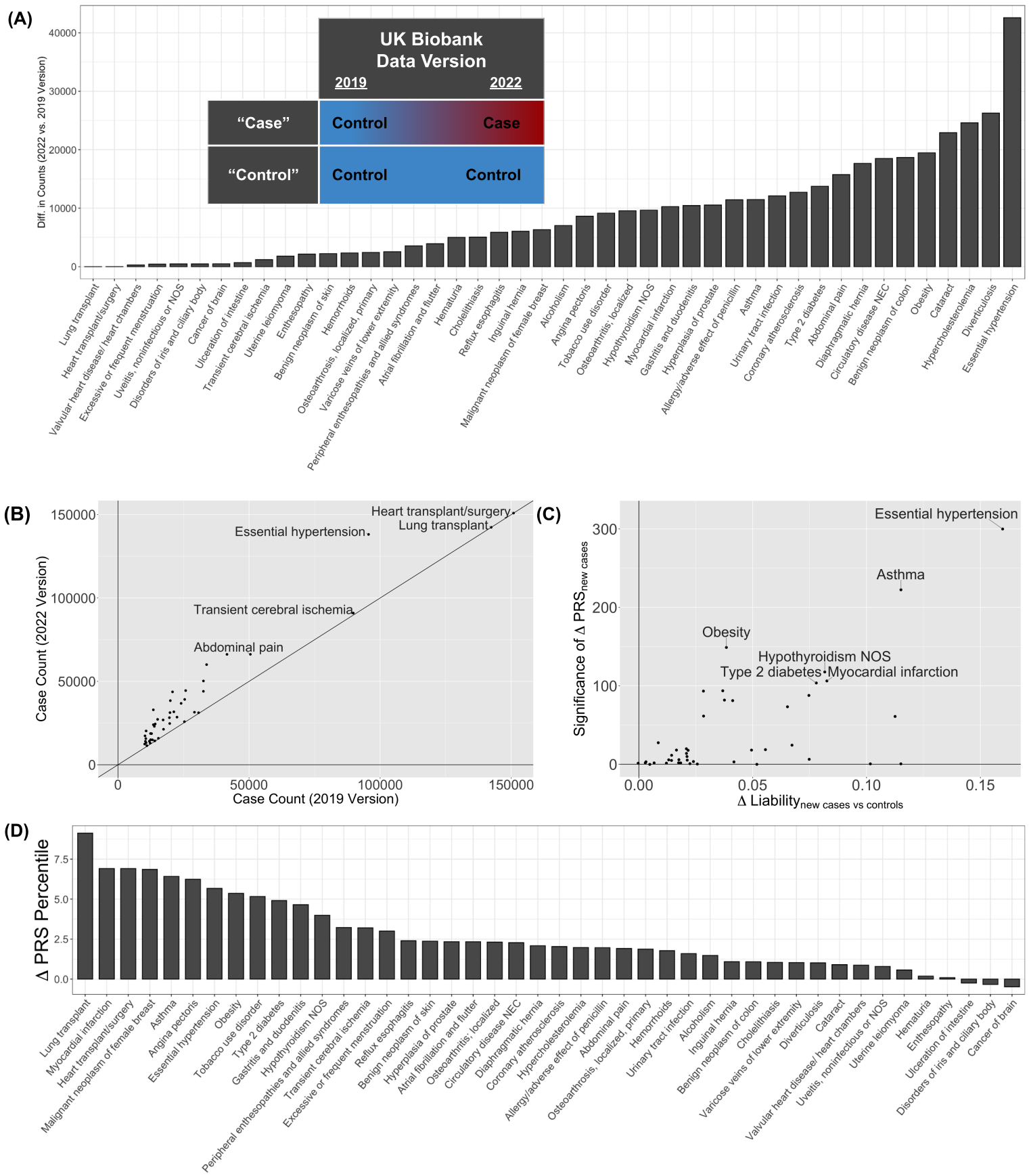
Supplementary Figure 2. Mirrored Manhattan plots of GWAS analysis using the combination of cases and AI-based disease liabilities of the controls. Mirrored Manhattan plots of GWAS results for (A) uterine Leiomyoma, (B) hypercholesterolemia, (C) uveitis, (D) coronary atherosclerosis, (E) varicose veins of lower extremity, and (F) inguinal hernia. The CC-GWAS results are shown in the lower panel with annotated risk loci. The GWAS analyses conducted via the liability-CC method are shown in the upper panel with highlighted risk loci that are not significant in the CC-GWAS. Chromosome position is plotted in genomic order on the x-axis. The y-axis indicates the strength of the association (P values). The dashed horizontal line in red denotes the genome-wide significance threshold of 5×10^{-8}.

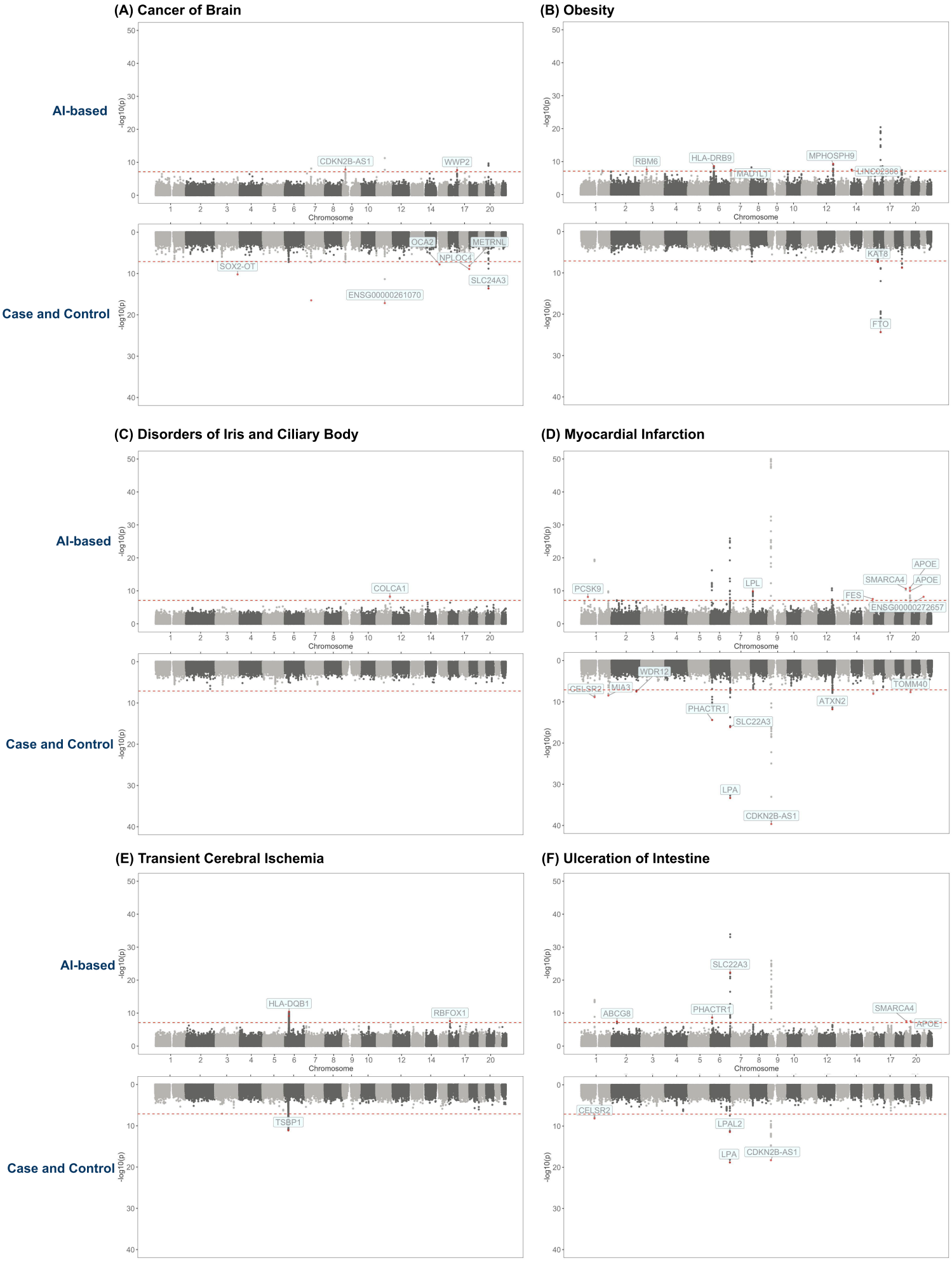

(A) UK Biobank Phenotyping
Unaffected Controls
Diseased Cases
Density
Disease liability
[0, 1] ∈ R
LAB
Medication
Questionnaire
EHR
Family History
biobank uk
(B) Genetic Associations
Disease Liability Strategies
Count
Controls
Cases
I. Binary CC
II. AI-based liability
III. Liability-CC
Known Association
Novel Associations
rs163
rs320
rs999
Significance
-log10(P)
** Not a real GWAS
Chromosome position
Manhattan plots show if -log10(P) passes the genome-wide significance threshold.
Gene A
Gene B
Gene C
Gene D
Polygenic Traits
AI-based
Case/Control
Density
Heritability (h²)
(C) Polygenic Risk Score Analysis
2019
2022
Density
PRS based on AI-based liability
P = X1 β1 + X2 β2 + … + Xn βn
n = total number of SNPs
Xi = genotypes
βi = estimated effect size of SNPi

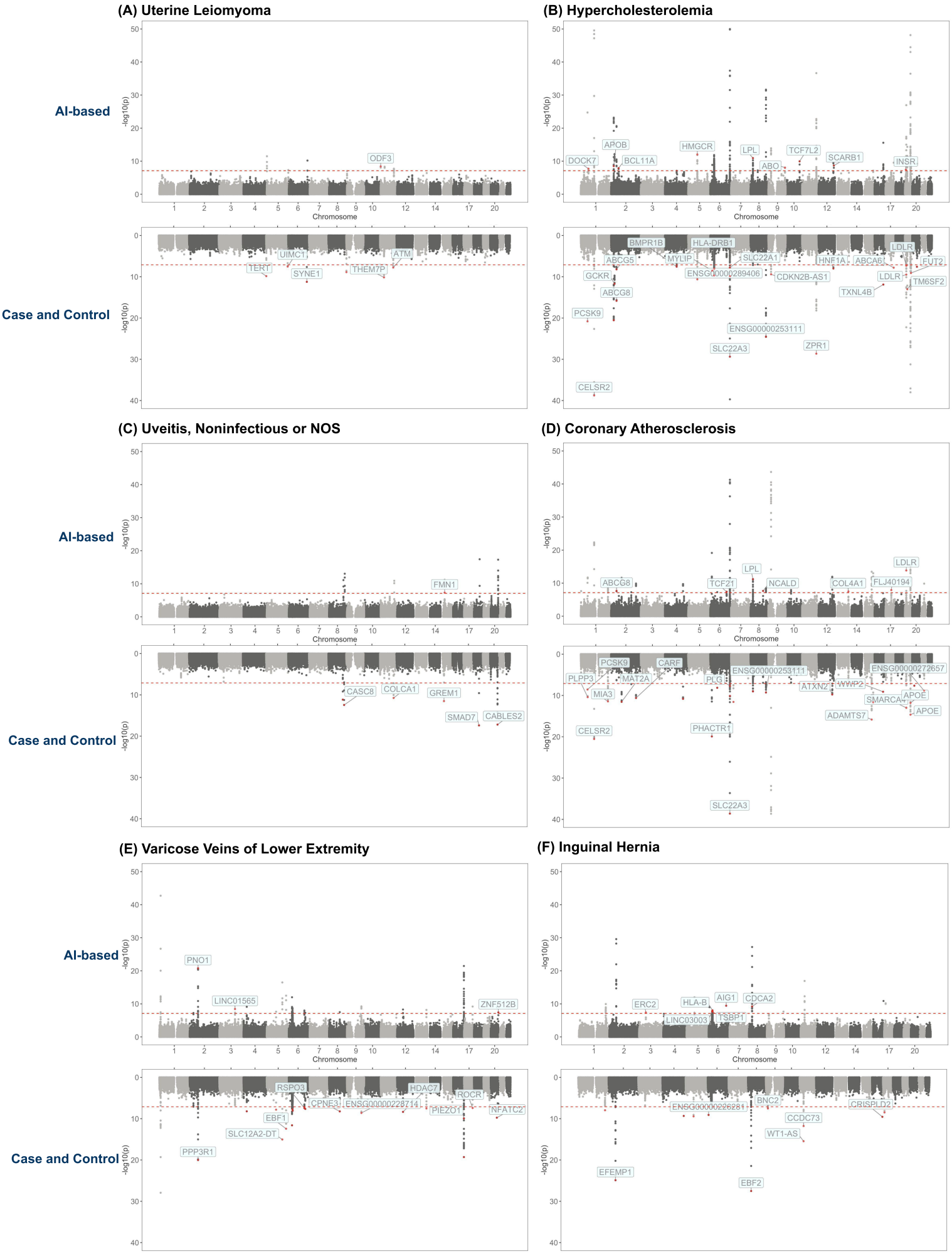

(A) Essential Hypertension
(B) Hypothyroidism
AI-based
Case and Control
-log10(p)
Chromosome
IFIH1
BHLHE40
ATXN2
PTPN2
TICAM1
(C)
Difference
Novel Loci
Essential hypertension
Asthma
Angina pectoris
Obesity
Myocardial infarction
Ulceration of intestine
Malignant neoplasm of female breast
Reflux esophagitis
Transient cerebral ischemia
Abdominal pain
Gastritis and duodenitis
Osteoarthrosis, localized, primary
Circulatory disease NEC
Peripheral enthesopathies and allied syndromes
Alcoholism
Urinary tract infection
Disorders of iris and ciliary body
Osteoarthritis; localized
Valvular heart disease/ heart chambers
Cholelithiasis
Heart transplant/surgery
Lung transplant
Allergy/adverse effect of penicillin
Excessive or frequent menstruation
Tobacco use disorder
Benign neoplasm of skin
Hemorrhoids
Hyperplasia of prostate
Hypothyroidism NOS
Hematuria
Cancer of brain
Enthesopathy
Diaphragmatic hernia
Inguinal hernia
Benign neoplasm of colon
Cataract
Hypercholesterolemia
Uterine leiomyoma
Uveitis, noninfectious or NOS
Coronary atherosclerosis
Varicose veins of lower extremity
Type 2 diabetes
Diverticulosis
Atrial fibrillation and flutter


(A) Angina Pectoris
(B) Type 2 Diabetes
Combined
Case and Control
-log10(p)
Chromosome
SLC22A3
ABCG8
MIA3
SMARCA4
APOE
ATXN2
FLJ40194
APOE
FES
ENSG00000272657
HLA-DQA2
ZBTB46
ZC3H4
PLPP3
MAT2A
UNC5C
PHACTR1
CELSR2
LPAL2
CDKN2B-AS1
COBLL1
RREB1
KCNQ1
HMGA2
MAP2K7
MICB
ANK1
ZMIZ1
CCND2
HNF1B
GCKR
ADCY5
PPIP5K2
JAZF1
TLET-DT
KCNQ1
HLA-DRB9
SLC30A8
CDKAL1
FTO
IGF2BP2
(C)
Difference
Novel Loci
Essential hypertension
Hypothyroidism NOS
Angina pectoris
Hypercholesterolemia
Inguinal hernia
Asthma
Myocardial infarction
Type 2 diabetes
Coronary atherosclerosis
Tobacco use disorder
Malignant neoplasm of female breast
Ulceration of intestine
Cataract
Cholelithiasis
Hyperplasia of prostate
Osteoarthritis; localized
Obesity
Diaphragmatic hernia
Hemorrhoids
Abdominal pain
Reflux esophagitis
Heart transplant/surgery
Circulatory disease NEC
Allergy/adverse effect of penicillin
Varicose veins of lower extremity
Transient cerebral ischemia
Uveitis, noninfectious or NOS
Cancer of brain
Uterine leiomyoma
Osteoarthrosis, localized, primary
Gastritis and duodenitis
Alcoholism
Valvular heart disease/ heart chambers
Benign neoplasm of skin
Excessive or frequent menstruation
Urinary tract infection
Disorders of iris and ciliary body
Hematuria
Peripheral enthesopathy
Enthesopathies and allied syndromes
Lung transplant
Benign neoplasm of colon
Atrial fibrillation and flutter
Diverticulosis


(A) Heritability
Essential hypertension
Type 2 diabetes
Hypercholesterolemia
Asthma
Coronary atherosclerosis
Hypothyroidism NOS
Obesity
Transient cerebral ischemia
Hyperplasia of prostate
Angina pectoris
Diaphragmatic hernia
Varicose veins of lower extremity
Myocardial infarction
Malignant neoplasm of female breast
Diverticulosis
Lung transplant
Heart transplant/surgery
Atrial fibrillation and flutter
Inguinal hernia
Abdominal pain
Gastritis and duodenitis
Allergy/adverse effect of penicillin
Tobacco use disorder
Benign neoplasm of colon
Ulceration of intestine
Cancer of brain
Osteoarthritis; localized
Alcoholism
Excessive or frequent menstruation
Reflux esophagitis
Circulatory disease NEC
Hemorrhoids
Osteoarthrosis, localized, primary
Cholelithiasis
Cataract
Uterine leiomyoma
Peripheral enthesopathies and allied syndromes
Disorders of iris and ciliary body
Uveitis, noninfectious or NOS
Hematuria
Enthesopathy
Benign neoplasm of skin
Valvular heart disease/ heart chambers
Urinary tract infection
0.00
0.05
0.10
0.15
Heritability (h_{SNP}^2)
AI-based
Combined
Case-Control
(B) AI-based
Essential hypertension
Obesity
Asthma
Angina pectoris
Atrial fibrillation and flutter
Diverticulosis
0.025
0.000
-0.025
Δ Heritability
-20
0
20
40
Δ Loci
(C) Combined
Essential hypertension
Hypercholesterolemia
Angina pectoris
Hypothyroidism NOS
Inguinal hernia
Diverticulosis
0.025
0.000
-0.025
Δ Heritability
-20
0
20
40
Δ Loci


(A)
UK Biobank
Data Version
2019
2022
"Case"
Control
Case
"Control"
Control
Control
Diff. in Counts (2022 vs. 2019 Version)
40000
30000
20000
10000
0
Lung transplant
Heart transplant/surgery
Valvular heart disease/ heart chambers
Excessive or frequent menstruation
Uveitis, noninfectious or NOS
Disorders of iris and ciliary body
Cancer of brain
Ulceration of intestine
Transient cerebral ischemia
Uterine leiomyoma
Enthesopathy
Hemorrhoids
Benign neoplasm of skin
Osteoarthrosis, localized, primary
Varicose veins of lower extremity
Peripheral enthesopathies and allied syndromes
Atrial fibrillation and flutter
Hematuria
Cholelithiasis
Reflux esophagitis
Inguinal hernia
Malignant neoplasm of female breast
Alcoholism
Angina pectoris
Tobacco use disorder
Osteoarthritis; localized
Hypothyroidism NOS
Myocardial infarction
Gastritis and duodenitis
Hyperplasia of prostate
Allergy/adverse effect of penicillin
Asthma
Urinary tract infection
Coronary atherosclerosis
Type 2 diabetes
Abdominal pain
Diaphragmatic hernia
Circulatory disease NEC
Benign neoplasm of colon
Obesity
Cataract
Hypercholesterolemia
Diverticulosis
Essential hypertension
(B)
Case Count (2022 Version)
150000
100000
50000
0
Heart transplant/surgery
Lung transplant
Essential hypertension
Transient cerebral ischemia
Abdominal pain
0
50000
100000
150000
Case Count (2019 Version)
(C)
Significance of Δ PRS$_{new\ cases}$
300
200
100
0
Essential hypertension
Asthma
Obesity
Hypothyroidism NOS
Type 2 diabetes
Myocardial infarction
0.00
0.05
0.10
0.15
Δ Liability$_{new\ cases\ vs\ controls}$
(D)
Δ PRS Percentile
7.5
5.0
2.5
0.0
Lung transplant
Myocardial infarction
Heart transplant/surgery
Malignant neoplasm of female breast
Asthma
Angina pectoris
Essential hypertension
Obesity
Tobacco use disorder
Type 2 diabetes
Gastritis and duodenitis
Hypothyroidism NOS
Peripheral enthesopathies and allied syndromes
Transient cerebral ischemia
Excessive or frequent menstruation
Reflux esophagitis
Benign neoplasm of skin
Hyperplasia of prostate
Atrial fibrillation and flutter
Osteoarthritis; localized
Circulatory disease NEC
Diaphragmatic hernia
Coronary atherosclerosis
Hypercholesterolemia
Allergy/adverse effect of penicillin
Abdominal pain
Osteoarthrosis, localized, primary
Hemorrhoids
Urinary tract infection
Alcoholism
Inguinal hernia
Benign neoplasm of colon
Cholelithiasis
Varicose veins of lower extremity
Diverticulosis
Cataract
Valvular heart disease/ heart chambers
Uveitis, noninfectious or NOS
Uterine leiomyoma
Hematuria
Enthesopathy
Ulceration of intestine
Disorders of iris and ciliary body
Cancer of brain


(A) Cancer of Brain
(B) Obesity
(C) Disorders of Iris and Ciliary Body
(D) Myocardial Infarction
(E) Transient Cerebral Ischemia
(F) Ulceration of Intestine
AI-based
Case and Control
-log10(p)
Chromosome
CDKN2B-AS1
WWP2
SOX2-OT
OCA2
METRNL
NPLOC4
SLC24A3
ENSG00000261070
RBM6
HLA-DRB9
MAD1L1
MPHOSPH9
LINC02386
KAT8
FTO
COLCA1
PCSK9
LPL
APOE
SMARCA4
FES
ENSG00000272657
WDR12
CELSR2
MIA3
PHACTR1
SLC22A3
ATXN2
TOMM40
LPA
CDKN2B-AS1
HLA-DQB1
RBFOX1
TSBP1
ABCG8
LPAL2


(A) Uterine Leiomyoma
(B) Hypercholesterolemia
(C) Uveitis, Noninfectious or NOS
(D) Coronary Atherosclerosis
(E) Varicose Veins of Lower Extremity
(F) Inguinal Hernia
AI-based
Case and Control
-log10(p)
Chromosome
ODF3
APOB
DOCK7
BCL11A
HMGCR
LPL
ABO
TCF7L2
SCARB1
INSR
UIMC1
TERT
SYNE1
THEM7P
ATM
BMPR1B
HLA-DRB1
ABCG5
MYLIP
SLC22A1
ENSG00000289406
CDKN2B-AS1
HNF1A
ABCA6
LDLR
FUT2
TM6SF2
TXNL4B
GCKR
ABCG8
PCSK9
ENSG00000253111
SLC22A3
ZPR1
CELSR2
FMN1
CASC8
COLCA1
GREM1
SMAD7
CABLES2
TCF21
NCALD
COL4A1
FLJ40194
PLPP3
MAT2A
CARF
PLG
ATXN2
WWP2
SMARCA4
APOE
ADAMTS7
MIA3
PHACTR1
ENSG00000272657
PNO1
LINC01565
ZNF512B
RSPO3
CPNE3
ENSG00000228714
HDAC7
ROCR
PIEZO1
NFATC2
EBF1
SLC12A2-DT
PPP3R1
ERC2
HLA-B
AIG1
CDCA2
LINC03003
TSBP1
BNC2
ENSG00000226281
CRISPLD2
CCDC73
WT1-AS
EFEMP1
EBF2