nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0193-x

Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits

Yan Zhang[1], Guanghao Qi[1], Ju-Hyun Park[2] and Nilanjan Chatterjee[1,3]*

We developed a likelihood-based approach for analyzing summary-level statistics and external linkage disequilibrium information to estimate effect-size distributions of common variants, characterized by the proportion of underlying susceptibility SNPs and a flexible normal-mixture model for their effects. Analysis of results available across 32 genome-wide association studies showed that, while all traits are highly polygenic, there is wide diversity in the degree and nature of polygenicity. Psychiatric diseases and traits related to mental health and ability appear to be most polygenic, involving a continuum of small effects. Most other traits, including major chronic diseases, involve clusters of SNPs that have distinct magnitudes of effects. We predict that the sample sizes needed to identify SNPs that explain most heritability found in genome-wide association studies will range from a few hundred thousand to multiple millions, depending on the underlying effect-size distributions of the traits. Accordingly, we project the risk-prediction ability of polygenic risk scores across a wide variety of diseases.

Sample sizes for genome-wide association studies (GWAS) for many complex traits now range between tens of thousands to hundreds of thousands due to the successes of the large consortia. These studies have led to discoveries of dozens and sometimes hundreds of common susceptibility SNPs for individual traits[1–3]. Although the effects of individual markers are modest, collectively they provide meaningful insights into underlying pathways and contribute to models for risk-stratification for some common diseases, such as breast cancer[4,5]. Further, existing GWAS for almost all traits indicate that common variants have the potential to explain much more heritability than that explained by SNPs, achieving the stringent genome-wide significance level[6–12]. It can be anticipated that sample sizes for many easily ascertainable traits and common diseases will continue to rise rapidly, allowing GWAS to reach their full potential. However, for rare diseases and difficult- or expensive-to-ascertain traits, it is not clear what is realistically achievable, given the practical limits of sample size and uncertainty about the best way to distribute resources as a community based on the likely yield for the traits. We and others have earlier shown that yields of future GWAS critically depend on underlying effect-size distributions[13–15].

Recently, the linkage disequilibrium (LD)-score regression method has become a popular approach for estimation of heritability and co-heritability using summary-level association statistics across GWAS[11,16,17]. This method relies on the observation that for highly polygenic traits, the association test statistics for GWAS markers are expected to be linearly related to their LD scores, a measure of the total amount of LD individual SNPs have with others in the genome; the slope of this linear relationship is determined by the degree of narrow-sense heritability of the trait that could be explained by the underlying reference panel of SNPs tagged by the GWAS markers.

We developed a likelihood-based framework that allows estimation of potentially complex effect-size distribution of a trait based on a single set of summary statistics that are widely available from GWAS consortia. We applied this method to analyze publicly available summary-level association statistics for 19 quantitative traits and 13 binary traits, to provide a large and comprehensive analysis of effect-size distributions underlying GWAS (see Supplementary Table 1 for the list of data sources). These applications provide detailed insights into the diversity of genetic architecture of complex traits, including numbers of underlying susceptibility SNPs and different clusters of effects that contribute to heritability of the traits. Using these estimated effect-size distributions, we then provide projections regarding potential of future GWAS, both in terms of their ability to identify susceptibility SNPs and to improve models for genetic risk prediction.

Results

Simulation studies. Simulation studies showed that when the number of components of underlying mixture models is correctly specified, our proposed method produced parameter estimates that trended toward underlying true values as sample size increased (Supplementary Table 2). There was generally, however, some downward bias in estimates of proportions of susceptibility SNPs (π_c and p) belonging to various non-null components and upward bias in corresponding variance-component parameters. For estimation of the overall effect-size distribution, the resulting biases manifested in underestimating the number of SNPs with extremely small effect sizes, with the magnitude of bias rapidly decreasing with increasing sample size (Supplementary Fig. 1). The bias was much more pronounced and did not disappear with increasing sample sizes when data were simulated under the three-component model (M3) but analyzed with the two-component model (M2). Standard errors for parameter estimates were distinctively larger when data were analyzed with M3 than M2, indicating additional uncertainty associated with fitting more complex models. Under both models, the proposed sandwich estimator produced slightly conservative estimates of the true standard errors across different sample

[1]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Statistics, Dongguk University, Seoul, Republic of Korea. [3]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA. *e-mail: nilanjan@jhu.edu

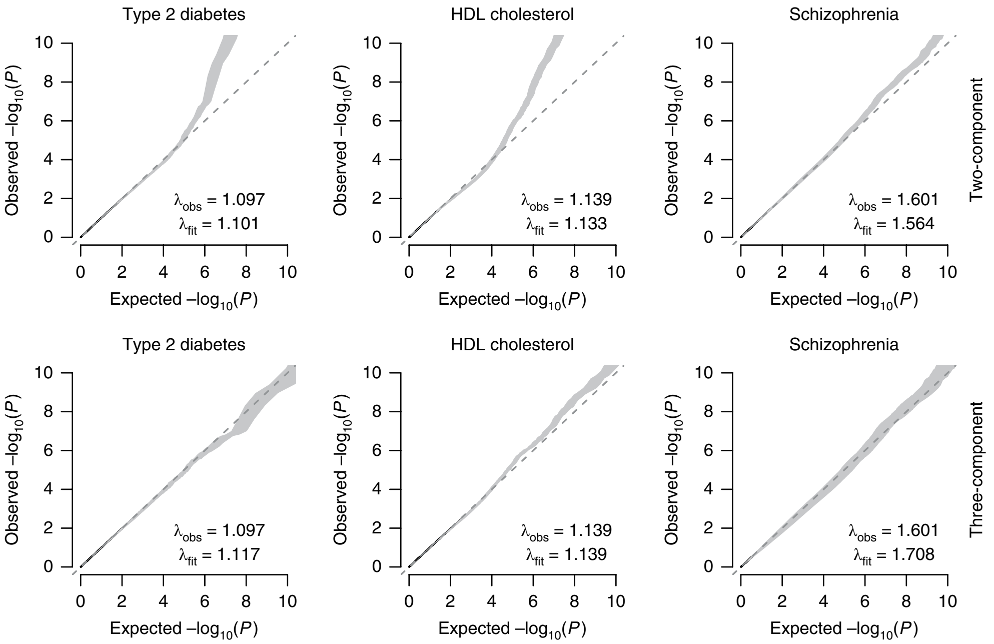

Fig. 1 | Q-Q plots comparing observed distributions of association statistics against those expected under the fitted models for three representative traits. Plots in top and bottom panels are generated under the two- and three-component models, respectively. Shaded regions mark 80% point-wise confidence intervals derived from 100 simulations (see Methods, Supplementary Note). λ_{obs} is the genomic control factor in the observed summary-level GWAS data; λ_{fit} is the mean genomic control factor in simulated data over 100 replications. While the more flexible three-component model provides a noticeably better fit for Type 2 diabetes and high-density lipoprotein (HDL) cholesterol, the simpler two-component model is adequate for schizophrenia. See Supplementary Figs. 5–9 for analogous plots for 29 additional complex traits.

sizes. Further, the modified Bayesian information criterion (BIC; Supplementary Note) appears to have allowed valid model selection, with its accuracy increasing with sample size (Supplementary Table 2). Very similar patterns were seen when we first simulated studies with individual level data and then generated summary-level statistics using standard GWAS analysis (Supplementary Table 3 and Supplementary Fig. 2).

For additional sensitivity analysis, we conducted simulation studies allowing true effect sizes to follow alternative distributions that did not conform to our model assumptions. We considered distributions that had either additional mixture components or that could not be represented in the normal-mixture form at all (Supplementary Fig. 3). We also allowed effect sizes to depend on local LD structures, as empirically evidenced by some recent studies[18] (Supplementary Fig. 4). In all these settings, M3 produced estimates of overall effect-size distribution that followed patterns fairly similar to those observed when the data were generated under this model itself. In particular, the model tended to underestimate the number of SNPs with extremely small effects, but the magnitude of bias diminished steadily with increasing sample sizes. Thus, it appears that, in these simulation settings, while the assumed three-component model was not correct it still provided reasonable approximations of the underlying effect-size distributions, and the observed pattern of bias in estimates was mainly driven by sample size rather than model mis-specification.

Effect-size distribution for 32 complex traits. We analyzed summary-level results from GWAS for each trait using both M2 and M3, and here report the main results based on the best models selected by the modified BIC criterion (Supplementary Note). In general, M3 provided distinctly better fits for the observed distribution of *P* values (Fig. 1 and Supplementary Figs. 5–9). For most of the traits, it provided excellent to adequate fit to observed *P* values over a wide range, except at extreme tails of the distribution ($P < 10^{-10}$), indicating the presence of a small number of susceptibility SNPs whose underlying effects were ‘outliers’ with respect to the fitted effect-size distributions. Notably, for a subset of traits (consisting of psychiatric diseases and traits related to intelligence, cognitive ability, and educational attainment), M2 and M3 fit data equally well, indicating that the effect sizes for underlying susceptibility SNPs can be adequately modeled using a single normal distribution. For these traits, as expected, the BIC criterion selected M2 as the best fitting model.

Parameter estimates associated with the best-fitted (M2 or M3) model showed wide diversity in genetic architecture across the traits (Table 1, Fig. 2, and Supplementary Tables 4 and 5). Estimates of narrow-sense heritability from the fitted models were generally close to those reported by LD-score regression[17] (Supplementary Table 6). Estimates of the number of underlying susceptibility SNPs varied widely, sometimes even among traits with similar estimates of heritability. In general, anthropometric traits, psychiatric diseases, and traits related to intelligence, cognitive ability, and educational attainment were found to be most polygenic, each involving >10,000 underlying susceptibility SNPs. In contrast, some of the early growth traits, autoimmune disorders, and adult-onset chronic common diseases (for example, coronary artery disease, asthma, Alzheimer’s disease, type-2 diabetes) were less polygenic, although each still involved at least a few thousand underlying susceptibility SNPs. Consistent with results from simulation studies, we observed that the fitting of the two-component model generally provided substantially lower estimates for the number of susceptibility SNPs across most traits (Supplementary Tables 4 and 5).

For a majority of the traits, the three-component model detected distinct clusters of effects. For these traits, the average heritability explained per variant in one cluster (σ_1^2) was often ten-fold or higher than that (σ_2^2) in the other cluster (Supplementary Table 4). Although a small fraction (typically 0.6–11%) of the susceptibility SNPs belonged to clusters with larger effect sizes, the fraction

Table 1 | Estimated parameter values and standard errors from the best fitted (two- or three-component) model (M2 or M3) for effect size distributions across 32 traits

Trait	Sample size (in 1,000)	Total number of sSNPs[a] (SE[b]), in 1,000	Number of sSNPs in cluster 1[c] (SE), in 1,000	Heritability explained by cluster 1 (SE)	Heritability explained by cluster 2[d] (SE)	Total heritability[e] (SE)
Continuous Traits:						
Age at menarche	182	13.2 (1.4)	0.38 (0.06)	0.047 (0.006)	0.088 (0.007)	0.135 (0.008)
BMI	124	15.0 (1.5)	0.10 (0.03)	0.017 (0.005)	0.179 (0.011)	0.197 (0.010)
Height	134	9.5 (1.2)	0.90 (0.16)	0.132 (0.017)	0.192 (0.017)	0.324 (0.015)
Hip circumference	213	11.9 (1.4)	0.27 (0.11)	0.022 (0.006)	0.114 (0.009)	0.136 (0.008)
Waist circumference	232	12.8 (1.4)	0.20 (0.08)	0.016 (0.004)	0.102 (0.008)	0.118 (0.007)
Waist-to-hip ratio	212	9.2 (1.7)	0.20 (0.10)	0.013 (0.004)	0.071 (0.008)	0.084 (0.007)
HDL cholesterol	95	10.5 (1.5)	0.19 (0.05)	0.029 (0.005)	0.078 (0.011)	0.107 (0.010)
LDL cholesterol	95	9.3 (1.9)	0.13 (0.04)	0.029 (0.005)	0.082 (0.011)	0.111 (0.011)
Total cholesterol	95	6.4 (1.9)	0.16 (0.05)	0.036 (0.006)	0.086 (0.011)	0.122 (0.012)
Triglycerides	95	9.5 (1.4)	0.06 (0.02)	0.014 (0.003)	0.094 (0.010)	0.107 (0.010)
Child birth length	28	2.7 (1.1)	N/A[f]	N/A	N/A	0.149 (0.032)
Child birth weight	144	5.5 (1.9)	0.30 (0.31)	0.027 (0.017)	0.078 (0.017)	0.106 (0.009)
Childhood obesity	14	6.1 (2.2)	0.05 (0.02)	0.049 (0.020)	0.308 (0.055)	0.357 (0.054)
Infant head circumference	11	5.3 (6.8)	N/A	N/A	N/A	0.300 (0.073)
Childhood IQ[g]	12	51.0 (8.6)	N/A	N/A	N/A	0.238 (0.070)
Cognitive performance	107	11.2 (2.3)	N/A	N/A	N/A	0.117 (0.010)
Intelligence	78	14.9 (2.0)	N/A	N/A	N/A	0.224 (0.015)
Years of schooling	294	19.4 (2.2)	N/A	N/A	N/A	0.131 (0.006)
Neuroticism	161	5.4 (15.9)	0.60 (2.48)	0.005 (0.015)	0.008 (0.013)	0.013 (0.005)
Disease Traits:						
Alzheimer	17/37[h]	2.6 (1.9)	0.04 (0.04)	0.075 (0.038)	0.276 (0.075)	0.351 (0.075)
Asthma	10/16	1.6 (1.0)	0.03 (0.01)	0.136 (0.053)	0.336 (0.115)	0.471 (0.125)
Coronary artery disease	22/65	2.5 (0.9)	0.02 (0.01)	0.026 (0.014)	0.363 (0.065)	0.389 (0.066)
Type 2 diabetes	12/57	4.8 (2.4)	0.04 (0.04)	0.079 (0.039)	0.614 (0.104)	0.694 (0.105)
Crohn's disease	6/15	6.2 (2.1)	0.36 (0.09)	1.380 (0.211)	1.170 (0.267)	2.550 (0.286)
Inflammatory bowel disease	13/22	5.6 (2.2)	0.39 (0.07)	0.842 (0.117)	0.649 (0.140)	1.490 (0.153)
Ulcerative colitis	7/20	2.7 (1.6)	0.19 (0.09)	0.565 (0.181)	0.895 (0.217)	1.460 (0.219)
College completion	22/73	12.8 (3.1)	N/A	N/A	N/A	0.672 (0.060)
Rheumatoid arthritis	14/44	3.9 (1.5)	0.17 (0.06)	0.292 (0.062)	0.619 (0.087)	0.911 (0.093)
Autism spectrum disorder	18/28	10.4 (2.3)	N/A	N/A	N/A	0.896 (0.109)
Bipolar disorder	7/9	10.9 (3.6)	N/A	N/A	N/A	2.030 (0.232)
Major depressive disorder	60/113	30.4 (4.4)	N/A	N/A	N/A	0.367 (0.028)
Schizophrenia	34/43	19.3 (1.9)	N/A	N/A	N/A	2.100 (0.095)

All results are reported with respect to a reference panel of 1.07 million common SNPs included in the Hapmap3 panel after removal of MHC region. An r^2 threshold of 0.1 and LD-window size of 1 MB is used to define the set of reference SNPs the GWAS markers may tag. Results from M2 for all 32 traits and alternative r^2 threshold and LD-window size are shown in Supplementary Tables 5 and 13. [a]Susceptibility SNPs. [b]Standard errors. [c,d]Cluster 1 and cluster 2 refer to the two non-null components of M3, in which cluster 1 corresponds to the component with larger variance-component parameter. [e]Total heritability under M3 is defined as $h^2 = M\pi_c\{p_1\sigma_1^2 + (1-p_1)\sigma_2^2\} = h_1^2 + h_2^2$, where M is the total number of SNPs in the Hapmap3 panel, π_c is the proportion of susceptibility SNPs, p_1 is the proportion of SNPs in the first cluster among all the susceptibility SNPs, and σ_1^2 and σ_2^2 are the variance estimates corresponding to each cluster. This definition corresponds to heritability in the observed scale for continuous traits and in the log-odds-ratio scale for disease traits; similarly, total heritability under M2 is defined as $h^2 = M\pi_c\sigma^2$, where σ^2 is the variance estimate. [f]N/A implies that M2 was selected as the best-fitted model, and the respective parameters are not defined. [g]Intelligence quotient. [h]Number of cases/number of controls.

of heritability they explained was substantial, ranging between 7 and 57% (Table 1 and Supplementary Table 4). In contrast, for all psychiatric diseases and for traits related to intelligence, cognitive ability, and educational attainment, the estimates of two variance components collapsed to a single value or were close to each other, a phenomenon consistent with adequate fit for M2 for these traits (Fig. 1 and Supplementary Figs. 5–9). Comparison of the number of SNPs in the tail regions of effect-size distributions showed that some traits, such as early growth traits and inflammatory bowel diseases, had distinctly larger numbers of SNPs with moderate-to-large effects than other traits (Supplementary Table 7).

Sample size requirement. The diversity of genetic architecture across the traits implies major differences in the future yield of GWAS (Fig. 3). In general, the number of discoveries is expected to rise rapidly across all traits in the foreseeable future. The degree of genetic variance they will explain will rise at a slower rate as the effect size explained per SNP will continue to diminish. For most

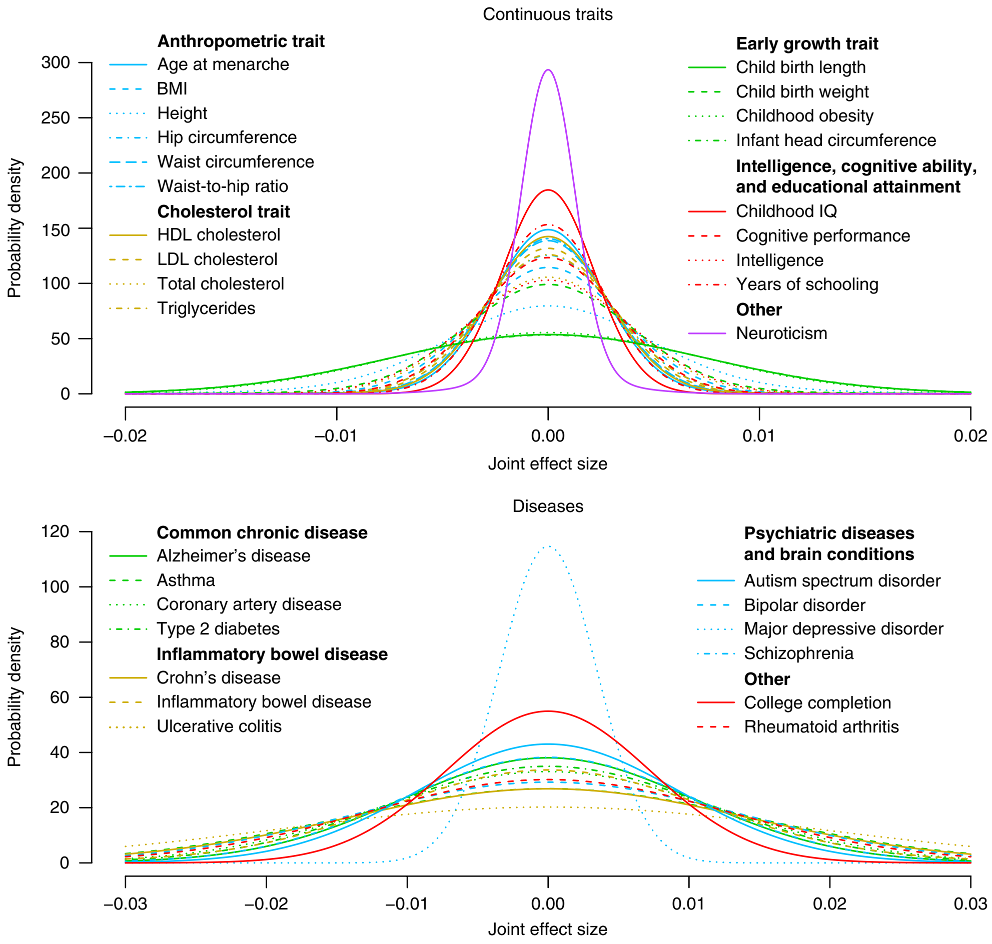

Fig. 2 | Estimated effect-size distributions for susceptibility SNPs based on the best fitted (M2 or M3) model for continuous (top) and binary traits (bottom). Distributions with fatter tails imply that the underlying traits have relatively greater numbers of susceptibility SNPs with larger effects. In general, traits related to mental health and ability have effect sizes with narrower trails despite larger estimates of heritability and associated numbers of susceptibility SNPs. See Supplementary Table 7 for a more detailed comparison of tail regions of effect-size distributions. LDL, low-density lipoprotein.

quantitative traits, the rate of increase in genetic variance explained is expected to diminish after sample size reaches 200,000–400,000. In contrast, for body mass index (BMI), hip circumference, waist circumference, waist-to-hip ratio, intelligence, cognitive ability, and educational attainment, the genetic variance explained is expected to increase at a steady rate until sample size reaches at least one million. The sample size needed to identify SNPs that can explain 80% of GWAS heritability is approximately 500,000 for some of the early growth traits, 1 million for adult height, between 2 and 4 million for various cholesterol and obesity related traits, and as high as 6 million for childhood intelligence quotient. For most disease traits, genetic variance explained is expected to rise either steadily or at an accelerated rate (in the case of highly polygenic psychiatric diseases) between sample sizes 50,000 and 300,000. The sample size needed to identify SNPs that can explain 80% of GWAS heritability is between 200,000 and 400,000 for inflammatory bowel diseases, around 600,000 for rheumatoid arthritis, between 500,000 and 1 million for most common adult-onset chronic diseases, between 0.7 and 1.5 million for most psychiatric diseases, and up to 10 million for major depressive disorder.

We evaluated the impact of sample size on estimates of heritability and effect-size distribution empirically for seven traits, accessing summary-level statistics from older and more recent GWAS with substantially different sample sizes (Supplementary Table 8). Estimates of heritability remain fairly stable for height, birth weight, and years of schooling. For all three traits, estimates of number of susceptibility SNPs increase in the more recent study, with the increase being most prominent for birth weight, for which the older study has relatively modest sample size ($n = 27,000$). These results are consistent with simulation studies in which we observed that the estimates of total number of susceptibility SNPs tended to increase toward true value as sample size increased (Supplementary Table 2). For BMI and all three psychiatric disease traits (autism spectrum disorder, major depressive disorder, and schizophrenia), estimates of heritability dropped substantially in the more recent study. The same trends were also seen when we applied LD-score regression to estimate heritability (also see LD-Hub[17] for the reported results of BMI from different studies). For these traits, the estimates of number of susceptibility SNPs remained fairly constant, and as a result, estimates of per-SNP heritability decreased. It is possible that such trend is due to the increasingly heterogeneous nature of phenotypes in larger studies, but more studies are needed to explore the phenomenon more broadly. Despite such discrepancies, across all seven traits, the models built based on the older studies were able to predict the number of loci reaching genome-wide significance in the newer study within or near the limits of uncertainty (Supplementary Table 9).

Predictive performance of polygenic risk scores. Using the inferred effect-size distributions and theoretical framework we developed earlier[15], we assessed the expected predictive performance of polygenic models when SNPs were included at optimal

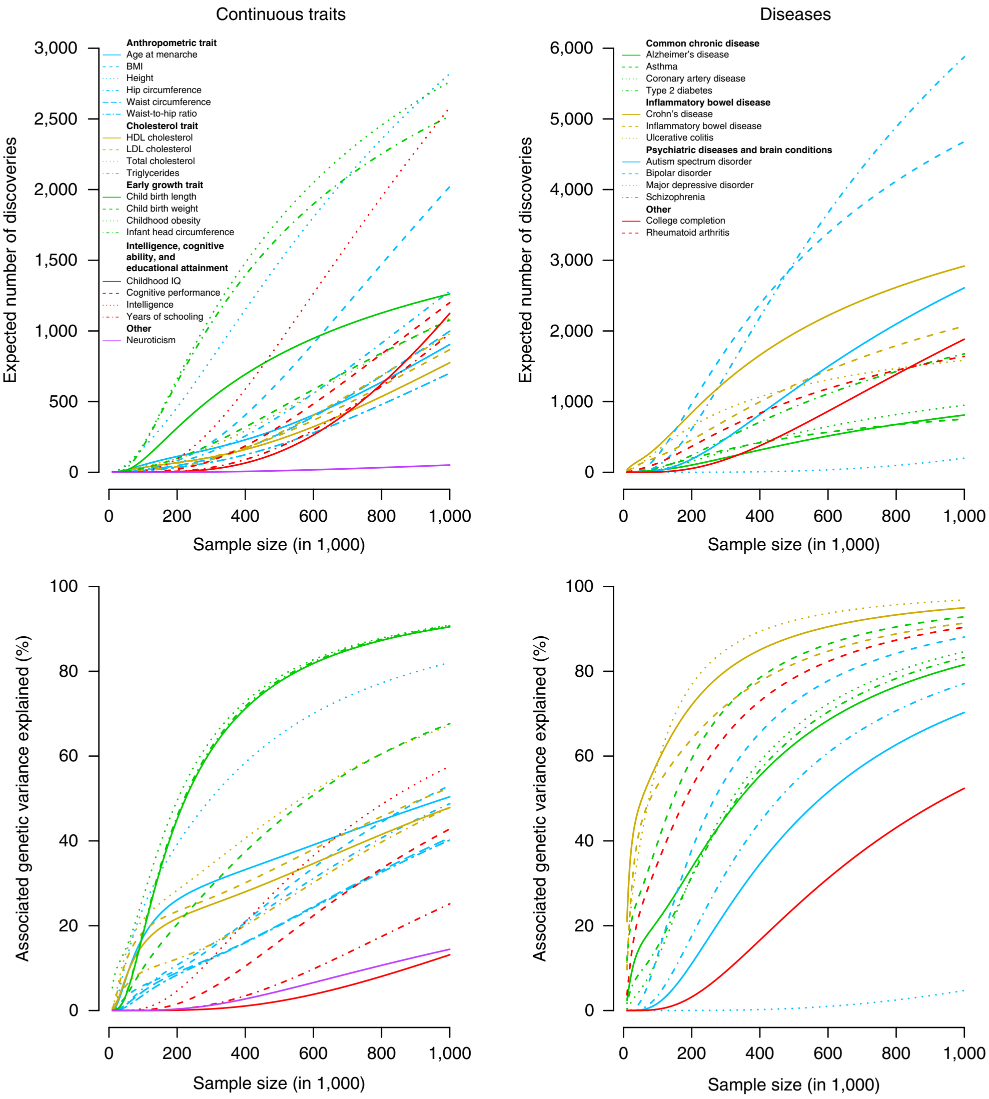

Fig. 3 | Projected number of discoveries (top) and corresponding percentage of GWAS heritability explained (bottom) based on the best-fitted (M2 or M3) model for effect-size distribution for continuous (left) and binary (right) traits. Results are based on power calculations for discovery at the genome-wide significance level ($P = 5 \times 10^{-8}$). Total estimates of GWAS heritability are used as the denominator in calculating the percentage of heritability that would be explained by SNPs reaching genome-wide significance (see Methods and Supplementary Note for details of formulas used for making these projections). The sample size for disease traits indicates total number of cases and controls, assuming a 1:1 ratio.

thresholds[19] and at the stringent genome-wide significance level ($P < 5 \times 10^{-8}$). The results showed two very distinct patterns. For psychiatric diseases, which include a continuum of highly polygenic effects, use of the optimal threshold for SNP selection was expected to lead to large improvements in performance of polygenic models in a wide range of sample sizes (Fig. 4 and Supplementary Fig. 10). For these traits, the optimal P-value threshold was expected to be highly liberal for relatively small studies (i.e., $n < 20{,}000$) and then become more stringent as sample size increases. In contrast, for all other diseases, which are less polygenic but include more SNPs with relatively large effects, the use of the optimal threshold was expected to lead to more modest benefits. For these diseases, the optimal threshold was expected to be highly stringent for studies with small sample sizes ($n < 10{,}000$), gradually become more liberal with intermediate sample sizes ($10{,}000 < n < 50{,}000$), and then slowly decrease for larger sample sizes.

False discovery rate and shrinkage estimation. We also assessed the potential implications for inferred effect-size distributions on subsequent analyses of GWAS to optimize SNP discovery and improve estimation of SNP effect sizes. We calculated local false discovery rates (FDR)[20,21] for each SNP based on observed z-statistics and the model for marginal effect sizes (see equation (2) in Methods). It was evident that a high degree of polygenicity for these traits implied the possibility of identifying large number of loci at fairly low FDR values (Supplementary Figs. 11 and 12). The posterior mean estimates for effect sizes for individual SNPs shrank heavily toward zero compared to their estimates available from GWAS, with the degree of shrinkage being highest for SNPs with intermediate effects and studies with the smallest sample sizes (Supplementary Figs. 13 and 14). Further, SNPs with largest effect sizes shrank more under the two-component model than under the three-component model because of the ability of the latter to accommodate SNPs with distinctly larger effects.

Discussion

Estimation of heritability based on SNP arrays has been a major focus of research for GWAS ever since the first application of the

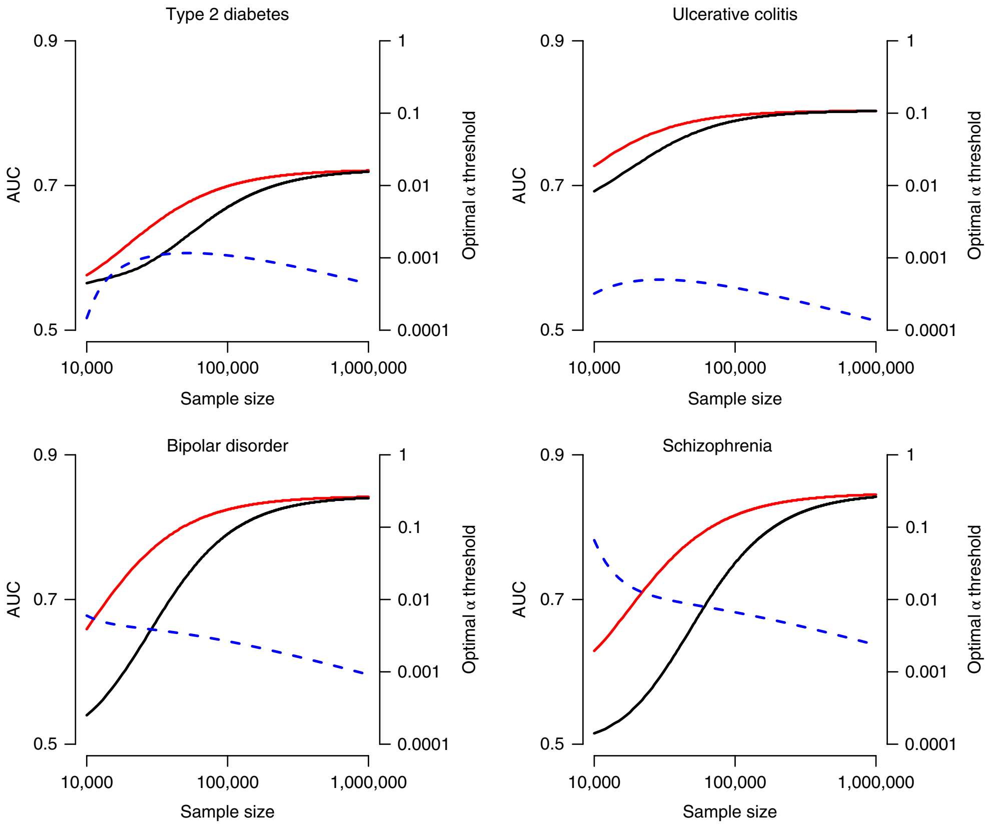

Fig. 4 | Expected area under the curve (AUC) for polygenic risk prediction models with SNPs included at the optimal significance (α) threshold (red solid line) and at the genome-wide significance level of 5×10^{-8} (black solid line). Optimum values for significance thresholds (shown in blue dashed line) are obtained based on the expected relationship of AUC with sample size and significance threshold under the best-fitted (M2 or M3) model for effect-size distributions. All calculations assume an analysis of 200,000 total independent sets of SNPs.

approach to analysis of human height[6,22–25]. While such estimates of heritability provide an understanding of the limits of GWAS, further assessment of effect-size distribution is critical for understanding how fast one can approach the limit as a function of sample size[14,15]. Although applied in limited settings, various approaches have been developed in the past for estimating effect-size distribution from GWAS. We have described a simple method for estimating distributions within the range of effects observed in an existing study based on the reported number of findings with different effect sizes and the power of the study for making discoveries at that effect size[13]. Several methods have been developed for inferring effect-size distribution by evaluating the predictive performance of a series of polygenic models on independent validation data sets[26,27]. A variety of Bayesian methods described for analysis of GWAS can also produce estimates of effect-size distribution based on the underlying 'prior' models[28–32]. Most recently, a few methods have been proposed for analyzing GWAS summary-level data under the two-component mixture model for estimating effect-size distribution[33–35].

The current analysis of effect-size distribution is unique in several ways. We provided comprehensive insights into effect-size distributions by analysis of GWAS summary-level statistics for a large variety of traits. We showed that a commonly used two-component model, which assumes that the effect sizes for underlying susceptibility SNPs can be described by a single normal distribution, can be inadequate for describing effect-size distribution across a large majority of the traits. Instead, a three-component model for effect-size distribution, which allowed a proportion of susceptibility SNPs to have distinctively larger effects than others, provided better fit to current GWAS for most traits and is thus likely to provide more accurate projections for future discoveries. In terms of methodology, the proposed approach, although closely related to some recent methods[34,35], has some unique aspects. We showed that under the commonly invoked assumption of independence of effect sizes and local LD patterns, the likelihood of summary statistics for individual GWAS markers depends on LD coefficients through the total LD score. The simplification allowed us to develop a robust, computationally tractable method for estimating parameters, as well as their standard errors, under the complex-mixture model for effect-size distribution based on an underlying composite likelihood inferential framework.

Our simulation studies showed some challenges in estimating the total number of susceptibility SNPs and other component specific parameters of mixture models due to identifiability issues. Because SNPs with vanishingly small effects are indistinguishable from null SNPs in GWAS of finite sample sizes, estimation of the total number of susceptibility SNPs required some degree of extrapolation through underlying parametric modeling assumptions. When the number of components of the mixture model was underspecified, the methods tended to classify the SNPs with the smallest effects as null and thus could substantially underestimate the total number of non-null SNPs. Allowing incorporation of additional mixture components can reduce bias to a large extent, but more complex models may not be well-identified in GWAS of modest sample size (for example, $n<25{,}000$), and some bias can persist even in much larger sample sizes (for example, $n=100{,}000$). Thus, overall it is quite likely that the complex traits we studied are even more polygenic than the current analysis suggested. Given the likelihood of such bias, it may be preferable to compare genetic architecture across traits in terms of the number of susceptibility SNPs that may have meaningfully large effects, such as an odds-ratio of 1.01 or larger, estimates of which are expected to be less sensitive to sample sizes of the underlying studies (Supplementary Table 7).

Our simulation studies also indicated that, despite likely biases in component-specific parameters, use of flexible mixture models could produce estimates of overall effect-size distributions with stable features. We observed that biases in mixing proportions and underlying variance component parameters tended to act in opposite directions. As a result, overall estimates of heritability were not generally very sensitive to underlying models used and sample sizes of GWAS. Moreover, estimates of numbers of SNPs with different effect sizes tended to be stable toward the tails of the distributions, and large bias was only seen at extremely small effect sizes, where current GWAS have virtually no power. Such estimates of effect-size distributions, in spite of potential bias for smallest effect sizes, can be useful for making projections for future GWAS results, up to substantial increases in sample size (Supplementary Fig. 15). In particular, when current sample sizes are relatively small (for example, 10,00–25,000), SNPs with the smallest inferred effect sizes are unlikely to contribute meaningfully to projected discovery in future GWAS with reasonable increases in sample size (for example, two- to five-fold), and thus the large bias in estimates of the numbers of underlying SNPs did not appear to make as big of an impact in these medium-term projections (Supplementary Fig. 15).

Our projections showed that a high degree of polygenicity of the traits implied that very large sample sizes, from hundreds of thousands to multiple millions, were required to identify SNPs explaining nearly all of the GWAS heritability. We provided empirical validation of the projections through analysis of data from older and more recent studies for several traits (Supplementary Table 9). Our projections were further validated by a very recent study[36], which reported that discoveries from GWAS analysis of ~700,000 individuals can explain about 24.6% and 5% of the phenotypic variance for height and BMI, respectively. Based on models we developed using GWAS involving about 125,000–135,000 individuals[37,38], we expected the latest study to lead to discoveries explaining about 24.0% and 7.7% of the phenotypic variances, respectively. Nevertheless, it is possible that in general our projections are somewhat optimistic, given that larger studies in the future, which may include increasingly heterogeneous samples, may show higher degrees of polygenicity of these traits. However, the current practice of using stringent genome-wide significance levels for identifying susceptibility SNPs is an extremely conservative approach. Instead, a more optimal strategy for discovery would be to select thresholds in a more adaptive fashion, taking into account underlying effect-size distributions while controlling for FDR[39,40].

A major utility of future GWAS could be improving performance of polygenic prediction models, as opposed to simply identifying susceptibility SNPs at high levels of significance[41–45]. Our projections showed that the use of optimal thresholds would lead to large benefit for psychiatric diseases, but more moderate benefits for others. In general, across all traits, we observed that the overall discriminatory performance of models, as measured by the area under the curve criterion, would be expected to rise very modestly after the sample size reaches around 100,000. However, larger sample sizes could still meaningfully improve the performance of models in terms of identifying individuals who are at the extremes of risk distribution. For type-2 diabetes, for example, a model built on GWAS with a sample size of 1 million instead of 100,000 individuals would be expected to identify an additional 0.2% (1.3% versus 1.1%) of the population who are at five-fold or higher risk than the average risk of the general population. Such an improved model could lead to intervention for an additional 2.0% of prospective cases (9.4 versus 7.4%; Supplementary Tables 10 and 11).

A limitation of the proposed method is that underlying definitions of tagging SNPs ($\mathcal{N}_k$) and LD score (ℓ_k) depend on the combination of r^2 threshold (r_T^2) and LD window size (WS). We conducted additional simulations involving individual-level data to explore the optimality of different combinations of r_T^2 and WS to produce the best model fit to the summary-level data, evaluated according to the proposed BIC statistics. We found that $r_T^2 = 0.1$ and $WS = 1$ MB, the combination used for our primary analysis, although not necessarily optimal (Supplementary Table 12), produced estimates of effect-size distributions comparable to those obtained from alternative combinations that produced a best fit across different settings (Supplementary Fig. 16). In a similar comparison for analysis of the real data sets, we found notable differences in estimates of the total numbers of susceptibility SNPs for some traits (Table 1 and Supplementary Table 13), but the estimates of effect-size distributions in the more stable tail regions were not generally affected (Supplementary Figs. 17 and 18).

We focused on estimating effect-size distributions based on widely available summary-level association statistics. Using individual level data, however, when available, may have certain advantages, including not requiring researchers to specify the r^2 threshold and LD-window size. We conducted limited simulation studies to compare the proposed method against BayesR[30], a method for estimating effect-size distribution using individual-level data within the normal-mixture model framework. As the running time of the BayesR method can be long for large data sets, we simulated data for only one chromosome for a study of sample size of about $n = 5,000$, but inflated the average effect size of the susceptibility SNPs to ensure adequate power. We observed that the proposed method and BayesR performed similarly for estimating the effect-size distribution in the tail region, but they could be biased in opposite directions for estimating the number of SNPs with extremely small effects (Supplementary Fig. 19). As discussed above, estimating numbers of SNPs with vanishingly small effects is an intrinsically challenging task, and all methods are expected to be somewhat sensitive to underlying modeling assumptions. One advantage of the proposed method is that it is expected to produce a lower bound for the number of susceptibility SNPs, and thus the resulting effect-size distribution can be used for conservative evaluation of local FDR and other empirical Bayes-type evaluations for the plausibility of associations[20].

Other limitations of the proposed method include the assumption that effect sizes are independent of allele frequencies and local LD patterns of SNPs. It has recently been shown that these simplified assumptions, which have been used implicitly or explicitly in many earlier methods, can lead to substantial underestimations of heritability[46]. Our simulation studies showed that modest violations of these assumptions, which we accounted for when considering evidence from recent empirical studies[18], were unlikely to lead to major bias in estimation of the overall effect-size distribution. Nevertheless, further studies are merited to extend the proposed method to model the dependence of effect sizes on various population genetic and functional genomic characteristics of the SNPs. Our ability to infer effect-size distribution for low-frequency and rare variants remains limited due to the insufficient power of current GWAS for identifying underlying susceptibility loci. Nevertheless, the limited number of findings from whole-exome and whole-genome sequencing studies to date suggests that these variants can contribute to heritability to a substantial extent only under highly polygenic models for underlying effect-size distribution[47].

To summarize, we proposed methods for statistical inference for effect-size distributions under flexible normal-mixture models using summary-level GWAS statistics. Applying these methods to a large number of GWAS identified wide diversity in genetic architecture of the underlying traits, with consequences for the yields of future GWAS in terms of both discovery and risk prediction.

URLs. Genetic effect-size distribution inference from summary-level data (GENESIS) software, https://github.com/yandorazhang/

GENESIS; 1000 GENOME Project phase 3 data, https://data.broadinstitute.org/alkesgroup/LDSCORE/; PLINK software, https://www.cog-genomics.org/plink2; LD-Hub, http://ldsc.broadinstitute.org/ldhub/; ReproGen Consortium (age at menarche), http://www.reprogen.org/data_download.html; GIANT Consortium (anthropometric traits) summary statistics, http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files; Global Lipids Genetics Consortium (cholesterol traits), http://csg.sph.umich.edu/abecasis/public/lipids2013/; EGG Consortium (early growth traits), http://egg-consortium.org/index.html; SSGAC (childhood intelligence quotient, cognitive performance, college completion, years of schooling), https://www.thessgac.org/data; CNCR (intelligence), http://ctg.cncr.nl/software/summary_statistics; Genetics of Personality Consortium (neuroticism), http://www.tweelingenregister.org/GPC/; IGAP (Alzheimer's disease), http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php; Gabriel Consortium (asthma), https://www.cng.fr/gabriel/results.html; CARDIoGRAMplusC4D Consortium (coronary artery disease), http://www.cardiogramplusc4d.org/data-downloads/; DIAGRAM (type 2 diabetes), http://www.diagram-consortium.org/downloads.html; IBDGC (inflammatory bowel disease), https://www.ibdgenetics.org/downloads.html; Rheumatoid arthritis summary statistics, http://plaza.umin.ac.jp/~yokada/datasource/software.htm; Psychiatric Genomics Consortium (psychiatric disease), https://www.med.unc.edu/pgc/results-and-downloads/downloads; GCTA software, http://cnsgenomics.com/software/gcta/#Download; BayesR software, https://github.com/syntheke/bayesR.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0193-x.

Received: 11 August 2017; Accepted: 18 May 2018;
Published online: 13 August 2018

References

1. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**(D1), D896–D901 (2017).
3. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
4. 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).
5. Garcia-Closas, M., Gunsoy, N. B. & Chatterjee, N. Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. *J. Natl. Cancer Inst.* **106**, dju305 (2014).
6. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
7. Chen, G.-B. et al. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum. Mol. Genet.* **23**, 4710–4720 (2014).
8. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet.* **8**, e1002637 (2012).
9. So, H. C., Gui, A. H. S., Cherny, S. S. & Sham, P. C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet. Epidemiol.* **35**, 310–317 (2011).
10. Sampson, J. N. et al. Analysis of heritability and shared heritability based on genome-wide association studies for thirteen cancer types. *J. Natl. Cancer Inst.* **107**, djv279 (2015).
11. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
12. Lee, S. H. et al. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
13. Park, J. H. et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
14. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
15. Chatterjee, N. et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013).
16. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
17. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
18. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
19. Purcell, S. M. et al. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
20. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, 275–294 (2017).
21. Efron, B. & Tibshirani, R. Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23**, 70–86 (2002).
22. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
23. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
24. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA* **111**, E5272–E5281 (2014).
25. So, H. C., Li, M. & Sham, P. C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet. Epidemiol.* **35**, 447–456 (2011).
26. Stahl, E. A. et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
27. Palla, L. & Dudbridge, F. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am. J. Hum. Genet.* **97**, 250–259 (2015).
28. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
29. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).
30. Moser, G. et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* **11**, e1004969 (2015).
31. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).
32. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
33. Thompson, W. K. et al. An empirical Bayes mixture model for effect size distributions in genome-wide association studies. *PLoS Genet.* **11**, e1005717 (2015).
34. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.* **11**, 1561–1592 (2017).
35. Holland, D. et al. Estimating phenotypic polygenicity and causal effect size variance from GWAS summary statistics while accounting for inflation due to cryptic relatedness. Preprint at *bioRxiv* https://doi.org/10.1101/133132 (2017).
36. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. Preprint at *bioRxiv* https://doi.org/10.1101/274654 (2018).
37. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
38. Speliotes, E. K. et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
40. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
41. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

42. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
43. Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* **18**, 3525–3531 (2009).
44. So, H.-C., Kwan, J. S. H., Cherny, S. S. & Sham, P. C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.* **88**, 548–565 (2011).
45. Kraft, P. & Hunter, D. J. Genetic risk prediction–are we there yet? *N. Engl. J. Med.* **360**, 1701–1703 (2009).
46. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
47. Derkach, A., Zhang, H. & Chatterjee, N. Power analysis for genetic association test (PAGEANT) provides insights to challenges for rare variant association studies. *Bioinformatics* **34**, 1506–1513 (2018).

Acknowledgements

The authors thank H. Zhang for assistance with computing, P. Kundu for data management for the UK Biobank data, and M. Chatterjee for editing of manuscript. The research was supported by Bloomberg Distinguished Professorship endowment. Some of the simulation studies were conducted using genotype data from the UK Biobank Resource accessed under Application Number 17712.

Author contributions

Y.Z., G.Q., and N.C. conceived the methods. Y.Z., G.Q., and J.-H.P. carried out all analyses. Y.Z. and N.C. wrote the manuscript. All authors reviewed the manuscripts.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0193-x.

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

Correspondence and requests for materials should be addressed to N.C.

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

Methods

Data and model. We assumed data were available on estimated regression coefficients $\widehat{\beta}_k$ and corresponding standard errors s_k for $k = 1, \ldots, K$ GWAS markers. We assumed that analysis was performed on a standardized scale such that both genotypes and phenotypes had unit variances. Typically, these 'summary-level' results were obtained from one-SNP-at-a-time 'marginal' analyses that did not account for correlations across SNPs. We assumed that the GWAS markers tagged a set of M SNPs in an underlying reference panel, with respect to which a joint linear model for association can be defined in the form $Y = \sum_{m=1}^{M} \beta_m^{(J)} \boldsymbol{G}_m + \epsilon$, where $\boldsymbol{Y}$ is a standardized $n \times 1$ phenotype vector of n individuals, $\boldsymbol{G}_m$ is an $n \times 1$ standardized genotype, and $\beta_m^{(J)}$ is an associated effect size for $m = 1, \ldots, M$ SNPs. The simple relationship $\beta_k = \sum_{m=1}^{M} \beta_m^{(J)} \rho_{km}$, between regression coefficients of SNPs from marginal and joint models, where ρ_{km} denotes correlations across SNP pairs, allows fitting of 'joint' models from estimates of marginal regression coefficients[48]. Essentially, all methods that aim to analyze summary-level statistics for making inference underlying a joint model implicitly assume that the relationship is approximately valid for logistic regression model for disease outcomes. We provide further theoretical justification underlying this approximation for rare diseases (Supplementary Note).

We assumed that the regression coefficients in the joint model were independently and identically distributed (i.i.d.) according to a mixture distribution in the form

$$\beta_m^{(J)} \sim \pi_c \sum_{h=1}^{H} p_h N(0, \sigma_h^2) + (1 - \pi_c)\delta_0 \tag{1}$$

where δ_0 is the Dirac delta function indicating that a fraction, $1 - \pi_c$, of the SNPs have no association with the trait and that the effect-size distribution for non-null SNPs is symmetric and modal around zero. The model allows distinct clusters of non-null effects through incorporating different variance component parameters (σ_h^2, $h = 1, \ldots, H$). In our application, we considered fitting two-component (M2) or three-component (M3) models, which allow the distribution of effects for non-null SNPs to follow either a single normal distribution ($H = 1$) or a mixture of two normal distributions ($H = 2$). The latter model allows two distinct variance component parameters, thereby allowing a fraction (p_1) of SNPs to have distinctly larger effects ($\sigma_1^2 > \sigma_2^2$).

Composite likelihood estimation. In principle, a joint likelihood for summary-level association statistics across GWAS markers can be derived using the relationship $\beta_k = \sum_{m=1}^{M} \beta_m^{(J)} \rho_{km}$ and the fact that $\widehat{\boldsymbol{\beta}} | \boldsymbol{\beta}^{(J)}$ is expected to follow a multivariate normal distribution[34]. In general, however, the computation of this likelihood for genome-wide analysis of millions of SNPs can be complex. We showed that under an assumption of independence of LD patterns and the probability of SNPs belonging to different mixture components in eq. (1), the distribution of marginal effects for individual SNPs can be approximated by another mixture form (see Supplementary Note):

$$\beta_k | \boldsymbol{\theta}^\star \sim \sum_{(N_k^{(0)}, \ldots, N_k^{(H)})} \Pr_{\boldsymbol{\xi}}(N_k^{(0)}, \ldots, N_k^{(H)}) \times N\left(0, \sum_{h=0}^{H} \frac{N_k^{(h)}}{N_k^\star} \sigma_h^2 \ell_k\right) \tag{2}$$

where $\boldsymbol{\theta}^\star = (\pi_c, p_1, \ldots, p_{H-1}, \sigma_1^2, \ldots, \sigma_H^2)$ denotes the unknown parameters in eq. (1); $N_k^\star$ is the total number of SNPs in the reference panel in a 'neighborhood' ($\mathcal{N}_k$) that may be 'tagged' by marker k; $N_k^{(h)}$ for $h = 1, \ldots, H$ are latent variables indicating the number of SNPs in $\mathcal{N}_k$ that have underlying effects from the hth component of the mixture distribution (see eq. (1)); $\ell_k = \sum_{m \in \mathcal{N}_k} \rho_{km}^2$ is the LD-score for the kth GWAS marker associated with $N_k^\star$ SNPs in the reference panel; $\boldsymbol{\xi} = (\pi_c, p_1, \ldots, p_{H-1})$, and $\sigma_0^2 = 0$. In the above formula, the mixing probability $\Pr_{\boldsymbol{\xi}}(N_k^{(0)}, \ldots, N_k^{(H)})$ can be calculated based on the standard multinomial distribution, with total counts defined by $N_k^\star = N_k^{(0)} + N_k^{(1)} + \cdots + N_k^{(H)}$ and cell probabilities given by $(1 - \pi_c, \pi_{c\times} p_1, \ldots, \pi_c \times p_H)$. Intuitively, eq. (2) implies that the distribution of marginal effects of the GWAS markers is given by mixtures of mean-zero normal distributions with variance component parameters determined by the product of the LD-score and the weighted sum of the variance component parameters of the original mixture model for joint effects (see eq. (1)); these weights, defined by the number of different types of underlying effects a GWAS marker tags, are expected to follow a multinomial distribution in general and a binomial distribution in special cases when $H = 1$.

By exploiting the fact that $\widehat{\beta}_k | \beta_k \sim N(\beta_k, a + s_k^2)$, which, similar to LD-score regression, incorporates an additional variance inflation factor a that accounts for systematic bias in variance estimates due to effects such as population-stratification or cryptic relatedness[11], we can express the likelihood for an individual GWAS marker as

$$L(\boldsymbol{\theta}; \widehat{\beta}_k) \approx \sum_{(N_k^{(0)}, \ldots, N_k^{(H)})} \Pr_{\boldsymbol{\xi}}(N_k^{(0)}, \ldots, N_k^{(H)}) \times N\left(0, \sum_{h=0}^{H} \frac{N_k^{(h)}}{N_k^\star} \sigma_h^2 \ell_k + a + s_k^2\right) \tag{3}$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}^*, a)$ denoting the unknown parameters in eq. (1), appended with the extra variance parameter a. In computing eq. (3), we exploited the fact that the number of underlying susceptibility SNPs $(N_k^{(1)} + \ldots + N_k^{(H)})$ that may be tagged by an individual GWAS marker were likely to be small, for example, ≤10, and so the number of terms in the mixture can be dramatically truncated to increase the speed of computation. To combine information across all markers, we formed a composite likelihood in the form $CL = \prod_{k=1}^{K} L(\boldsymbol{\theta}; \widehat{\beta}_k)$, which ignores correlations in $\widehat{\beta}_k$ across $k = 1, \ldots, K$. Following the theory of composite likelihood methods[49,50], it is evident that such a composite likelihood approach will produce consistent, i.e., asymptotically unbiased, estimates of $\boldsymbol{\theta}$ as long as eq. (3) is a valid likelihood for the summary-statistics of the individual markers. We maximized the likelihood using an expectation-maximization algorithm, where in each M-step, the mixing proportions $(\pi_c, p_1, \ldots, p_{H-1})$ were estimated in closed form and the variance component parameters were estimated by the numerical optimization of weighted univariate normal-likelihoods (see Supplementary Note).

Variance calculations. We obtained estimates of standard errors for parameter estimates based on a sandwich variance estimator associated with the composite likelihood (see eq. (3)). Let $l_k(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}; \widehat{\beta}_k)$ and $U_k(\boldsymbol{\theta}) = \partial l_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ denote the log-likelihood and score function, respectively, associated with the kth GWAS marker, and let $\bar{U}_k(\boldsymbol{\theta}) = \sum_{k' \in \mathbb{N}_k} U_{k'}(\boldsymbol{\theta})$ be the total likelihood score across all GWAS markers that are in the neighborhood $\mathbb{N}_k$ of the kth marker, including itself. Note that, unlike the calculation of the total LD-score, which involves SNPs in the underlying reference panel, the total likelihood score is computed only with respect to the set of markers that are included in the GWAS study itself. Further, we defined $I(\boldsymbol{\theta}) = -E\{\sum_{k=1}^{K} \partial^2 l_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T\}$ as the total information matrix associated with the composite likelihood. Using techniques parallel to those develop for generalized estimating equations (GEE) for the analysis of time-series data[51,52], we proposed a sandwich variance estimator accounting for 'banded' correlation structure across SNPs in the form

$$\mathrm{var}(\boldsymbol{\theta}) = [I(\boldsymbol{\theta})]^{-1} E\left[\sum_{k=1}^{K} U_k(\boldsymbol{\theta}) \bar{U}_k^T(\boldsymbol{\theta})\right] [I(\boldsymbol{\theta})]^{-1} \tag{4}$$

which itself can be estimated by plugging in the estimated parameter values $\widehat{\boldsymbol{\theta}}$ in lieu of $\boldsymbol{\theta}$. The estimator accounts for correlation across the GWAS markers by calculating empirical variance–covariance matrices across the likelihood scores within sets of correlated markers defined by physical distance and the same LD-threshold used to define the LD-scores. The estimator was expected to produce valid estimates of standard errors for the parameter estimates even when the underlying model is mis-specified.

Calculation of LD-score (ℓ_k) and number of tagged SNPs ($N_k^\star$). To implement the proposed method, we estimated the number of underlying SNPs in the reference panel tagged by the GWAS markers and the corresponding LD-scores. As we analyzed GWAS of primarily Caucasian studies, we used the genotype data from the 1,000 GENOME project Phase 3 study[53] involving 489 individuals of European origin. We extracted ~1.07 million common SNPs that were included in the Hapmap3 SNPs from the 1000 GENOME data as our reference panel. We evaluated all LD-scores and the number of tagged SNPs based on this reference panel. We defined the tagging SNPs for the kth GWAS marker as the SNPs in the reference panel that were within 1 Mb distance and had an estimated LD coefficient r^2 with this GWAS marker above a fixed threshold (for example, $r^2 \geq 0.1$). We estimated $N_k^\star$ by the total number of such tagging SNPs for the kth GWAS marker. Then we calculated the corresponding LD-score by summing up the squared LD coefficients for the $N_k^\star$ tagging SNPs. We evaluated the sensitivity of our results with respect to variation in the r^2 threshold. In calculating the LD-score, we employed the same bias-correction adjustment used in the LD-score regression[11].

Across all traits, we first extracted association statistics available from the underlying studies for the intersection of the available GWAS markers and the set of the Hapmap 3 SNPs. As most studies provide results after imputation, association statistics were available for large majority of the Hapmap 3 SNPs across these studies. We then followed filtering steps similar to those used in LD-Hub[17] to select GWAS markers to standardize the summary-level data sets. In particular, we removed SNPs that had sample sizes less than 0.67 times the 90th percentile of sample sizes, that were within the major histocompatibility complex (MHC) region (i.e., SNPs between 26 Mb and 34 Mb on chromosome six), or that had extremely large effects ($\chi^2 > 80$). Finally, we filtered SNPs to the ~1.07 million Hapmap3 SNPs with 1000 GENOME MAF ≥ 0.05.

Future projection. Given the parameter estimates ($\widehat{\boldsymbol{\theta}}$) of the underlying effect-size model, we projected the number of expected discoveries and associated heritability explained in future studies based on analytic formula. Let ND_α denote the random variable indicating the number of susceptibility SNPs to be discovered at the genome-wide significance level $\alpha = 5 \times 10^{-8}$ for a GWAS study with sample size n. We approximated the expected number of discoveries as

$$E(ND_\alpha) = \sum_{m=1}^{M} Pr\left(\sqrt{n}\ |\widehat{\beta}_m^{(J)}| > c_{\frac{\alpha}{2}}|\beta_m^{(J)}\right)$$
$$\approx M\widehat{\pi}_c \int_\beta pow_\alpha(\beta) \sum_{h=1}^{H} \widehat{p}_h N(0, \widehat{\sigma}_h^2)\, d\beta$$

where $pow(\beta) = 1 - \Phi\left(c_{\frac{\alpha}{2}} - \sqrt{n}\beta\right) + \Phi\left(-c_{\frac{\alpha}{2}} - \sqrt{n}\beta\right)$, in which $\Phi(\cdot)$ is the standard normal cumulative density function and $c_\alpha = \Phi^{-1}(1-\alpha)$ is the αth quantile for the standard normal distribution.

Following similar arguments, the expected value of the proportion of genetic variance explained by susceptibility SNPs reaching genome-wide significance can be written as

$$E(GV_\alpha) \approx M\widehat{\pi}_c h^{-2} \int_\beta \beta^2 pow_\alpha(\beta) \sum_{h=1}^{H} \widehat{p}_h N(0, \widehat{\sigma}_h^2)\, d\beta$$

We note that the above two formulas involve the sample size of the GWAS study through the power function $pow_\alpha(\beta)$.

Simulation studies. We used a simulation scheme to generate summary-level association statistics for GWAS without generating individual-level data. We simulated summary-statistics based on the model

$$\widehat{\beta}_k = \beta_k + \nu_k + e_k$$

where ν_k is assumed to be i.i.d. following a normal distribution, with mean zero and variance a and $\tilde{\boldsymbol{e}} = (e_1, \ldots, e_K)$ is assumed to follow a multivariate normal distribution with mean zero and variance–covariance matrix $\boldsymbol{R}/n$, with n denoting the sample size for GWAS and $\boldsymbol{R}$ denoting the symmetric matrix of LD coefficients across the GWAS. Above, the error term ν_k accounts for possible overdispersion in summary statistics by an underlying constant factor a, and e_K accounts for standard estimation error due to the finite sample size of the study. In each simulation, we first generated valued for $\beta_m^{(J)}$ and $m = 1, \ldots, M$ based on eq. (1) and then generated values for β_k and $k = 1, \ldots, K$ based on the transformation $\beta_k = \sum_{m \in \mathcal{N}_k} \beta_m^{(J)} \rho_{km}$. For simulating $\tilde{\boldsymbol{e}}$, we observed that summary-level association statistics in a GWAS are expected to follow the same multivariate distribution as $\tilde{\boldsymbol{e}}$ when the underlying phenotype has no association with any of the markers. Thus, we simulated null phenotypes for the samples in our reference 1000 GENOME data set and calculated association statistics $\tilde{\boldsymbol{u}} = (u_1, \ldots, u$K) for the GWAS markers. We then defined $\tilde{\boldsymbol{e}} = \sqrt{n_{ref}/n} \times \tilde{\boldsymbol{u}}$ to account for the difference in sample sizes between the reference panel and the GWAS. We also assessed the validity of the proposed scheme by simulating studies with individual-level data first and then generating summary statistics using traditional GWAS analysis in selected settings (see Supplementary Note).

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

Data availability. GENESIS software and a tutorial example on how to use it are available online (https://github.com/yandorazhang/GENESIS). Links to the 1000 GENOME Project Phase 3 data and all publicly available summary statistics are provided (see "URLs").

References

48. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
49. Lindsay, B. G. Composite likelihood methods. *Contemp. Math.* **80**, 221–239 (1988).
50. Varin, C., Reid, N. & Firth, D. An overview of composite likelihood methods. *Stat. Sin.* **21**, 5–42 (2011).
51. Heagerty, P. J. & Lumley, T. Window subsampling of estimating functions with application to regression models. *J. Am. Stat. Assoc.* **95**, 197–211 (2000).
52. Lumley, T. & Heagerty, P. Weighted empirical adaptive variance estimators for correlated data regression. *J. R. Stat. Soc. Series B Stat. Methodol.* **61**, 459–477 (1999).
53. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

nature research

Corresponding author(s): Nilanjan Chatterjee

Reporting Summary

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

Statistical parameters

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

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software code was used
Data analysis	Software code available from: https://github.com/yandorazhang/GENESIS

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

Data

Policy information about availability of data

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

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

Software tool for GENetic Effect-Size distribution Inference from Summary-level data (GENESIS), https://github.com/yandorazhang/GENESIS;
1000 GENOME Project Phase 3 data, https://data.broadinstitute.org/alkesgroup/LDSCORE/; PLINK software, https://www.cog-genomics.org/plink2;
LD-Hub, http://ldsc.broadinstitute.org/ldhub/;

ReproGen Consortium (age at menarche), http://www.reprogen.org/data_download.html; GIANT Consortium (anthropometric traits) summary statistics, http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files; Global Lipids Genetics Consortium (cholesterol traits), http://csg.sph.umich.edu/abecasis/public/lipids2013/;
EGG Consortium (early growth traits), http://egg-consortium.org/index.html; SSGAC (childhood IQ, cognitive performance, College completion, years of schooling), https://www.thessgac.org/data;
CNCR (Intelligence), http://ctg.cncr.nl/software/summary_statistics;
Genetics of Personality Consortium (neuroticism), http://www.tweelingenregister.org/GPC/;
IGAP (Alzheimer's disease), http://web.pasteurlille.fr/en/recherche/u744/igap/igap_download.php;
Gabriel Consortium (Asthma), https://www.cng.fr/gabriel/results.html;
CARDIoGRAMplusC4D Consortium (coronary artery disease), http://www.cardiogramplusc4d.org/data-downloads/;
DIAGRAM (Type 2 diabetes), http://www.diagram-consortium.org/downloads.html;
IIBDGC (Inflammatory bowel disease), https://www.ibdgenetics.org/downloads.html; rheumatoid arthritis summary statistics, http://plaza.umin.ac.jp/~yokada/datasource/software.htm;
Psychiatric Genomics Consortium (Psychiatric disease), https://www.med.unc.edu/pgc/results-and-downloads/downloads;
GCTA software, http://cnsgenomics.com/software/gcta/#Download;
BayesR software, https://github.com/syntheke/bayesR.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We used pubicly available data that had pre-determined sample sizes
Data exclusions	We used summary-level data and thus has no ability to remove subjects from analysis. We filtered SNPs following guidelines pre-established by LD-score regression. The filtering steps are described in Methods.
Replication	The final software code for analysis is made publicly available through GitHub repository.. This code can be used to analyze the publicly available datasets to reproduce our data analysis results.
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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