PLOS | GENETICS

RESEARCH ARTICLE

Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction

Yiming Hu[1], Qiongshi Lu[1], Wei Liu[2], Yuhua Zhang[3], Mo Li[1], Hongyu Zhao[1,4,5,6]*

1 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America, **2** Peking University, Beijing, China, **3** Shanghai Jiao Tong University, Shanghai, China, **4** Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **5** Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, United States of America, **6** Clinical Epidemiology Research Center (CERC), Veterans Affairs (VA) Cooperative Studies Program, VA Connecticut Healthcare System, West Haven, Connecticut, United States of America

* hongyu.zhao@yale.edu

Citation: Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H (2017) Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet 13(6): e1006836. https://doi.org/10.1371/journal.pgen.1006836

Editor: Xiaofeng Zhu, Case Western Reserve University, UNITED STATES

Received: November 8, 2016

Accepted: May 23, 2017

Published: June 9, 2017

Data Availability Statement: All the GWAS summary statistics are available online and can be accessed through: http://www.ibdgenetics.org, https://www.immunobase.org/downloads/protected_data/GWAS_Data/, http://www.ibdgenetics.org, http://diagram-consortium.org/downloads.html, http://www.reprogen.org/data_download.html, https://www.med.unc.edu/pgc/downloads, https://www.med.unc.edu/pgc/downloads, http://egg-consortium.org/birth-length.html, http://egg-consortium.org/birth-weight-2016.html, http://www.broadinstitute.org/collaboration/

Abstract

Accurate prediction of disease risk based on genetic factors is an important goal in human genetics research and precision medicine. Advanced prediction models will lead to more effective disease prevention and treatment strategies. Despite the identification of thousands of disease-associated genetic variants through genome-wide association studies (GWAS) in the past decade, accuracy of genetic risk prediction remains moderate for most diseases, which is largely due to the challenges in both identifying all the functionally relevant variants and accurately estimating their effect sizes. In this work, we introduce PleioPred, a principled framework that leverages pleiotropy and functional annotations in genetic risk prediction for complex diseases. PleioPred uses GWAS summary statistics as its input, and jointly models multiple genetically correlated diseases and a variety of external information including linkage disequilibrium and diverse functional annotations to increase the accuracy of risk prediction. Through comprehensive simulations and real data analyses on Crohn's disease, celiac disease and type-II diabetes, we demonstrate that our approach can substantially increase the accuracy of polygenic risk prediction and risk population stratification, i.e. PleioPred can significantly better separate type-II diabetes patients with early and late onset ages, illustrating its potential clinical application. Furthermore, we show that the increment in prediction accuracy is significantly correlated with the genetic correlation between the predicted and jointly modeled diseases.

Author summary

Genetic risk prediction plays a significant role in precision medicine. Accurate prediction models could have great impact on disease prevention and treatment strategies. However, prediction accuracies for most complex diseases remain moderate mainly due to the challenges in identifying and quantifying the effects of genetic variants from millions of markers, limited access to individual-level genotype data, and lack of efficient computational

giant/index.php/GIANT_consortium_data_files, http://egg-consortium.org/childhood-obesity.html, http://www.cardiogramplusc4d.org/downloads/, http://www.magicinvestigators.org/downloads/, http://csg.sph.umich.edu/abecasis/public/lipids2010/, http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files https://www.med.unc.edu/pgc/downloads, http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/, https://www.med.unc.edu/pgc/downloads. Individual level genotype data are available from dbGaP (accession numbers: phs000237, phs000274 and phs000674) and WTCCC (EGAD00000000001, EGAD00000000002, EGAD00000000007 and EGAD00001000401).

Funding: This study was supported in part by the National Institutes of Health (https://www.nih.gov/) grants R01 GM59507, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development (http://www.research.va.gov/programs/csp/), and the Yale World Scholars Program (http://bbs.yale.edu/training/initiatives/csc.aspx) sponsored by the China Scholarship Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

methods. Up to now, most methods have been focused on predicting disease risk using data from a single trait. With the discovery of genetic correlations among many complex diseases, incorporating data of genetically correlated diseases could have the potential to increase prediction accuracy. Current statistical methods are not able to fully exploit the richness of these kinds of data to take into account the shared genetic architecture. To make use of commonly available GWAS summary statistics, we propose a novel method to address these challenges by jointly modeling genetically correlated diseases and integrating genomic functional annotations. We demonstrate the substantial improvement in accuracy in both extensive simulation studies and real data analysis of Crohn's disease, celiac disease and type-II diabetes. Furthermore, we show that the increment in prediction accuracy is significantly correlated with the genetic correlation between the predicted and jointly modeled diseases.

Introduction

Achieving accurate disease risk prediction using genetic information is a major goal in human genetics research and precision medicine. Accurate prediction models will have great impacts on disease prevention and treatment strategies[1]. Various approaches that utilize genome-wide data in genetic risk prediction have been proposed, including machine-learning models trained on individual-level genotype and phenotype data[2–7], and polygenic risk scores (PRS) derived from genome-wide association study (GWAS) summary statistics [8, 9]. Despite the potential information loss in summary data, PRS-based approaches have been widely adopted in practice due to computational efficiency and the easy accessibility of GWAS summary level data[10, 11]. However, prediction accuracies for most complex diseases remain moderate, which is largely due to the challenges in both identifying all the functionally relevant variants and accurately estimating their effect sizes in the presence of linkage disequilibrium (LD) [12].

Integrating external information, e.g. pleiotropy [2, 3], LD [9], and functional annotations [13] has been shown to effectively address these challenges. Maier et al.[3] and Li et al.[2] showed that joint modeling of correlated traits could increase the prediction accuracy using individual level genotype data for psychiatric disorders and autoimmune diseases. Using summary level data, Hu et al.[13] proposed a single-trait risk prediction framework explicitly modeling LD and functional annotations, which consistently improves prediction accuracy for complex diseases. Furthermore, integrative genomic functional annotation, coupled with the rich collection of summary statistics from GWAS, have enabled increased statistical power in several different settings [14, 15]. Here, we introduce PleioPred (available at https://github.com/yiminghu/PleioPred), a principled framework that integrates GWAS summary statistics of genetically correlated diseases with various types of annotation data and reference genotype panels to improve risk prediction accuracy. Incorporating data from related traits and functional annotations increases the effective sample size and statistical power to detect functionally relevant variants, especially when diseases share similar genetic architecture. We compare PleioPred with state-of-the-art single-trait PRS-based approaches and demonstrate its consistent improvement in risk prediction performance using real data of multiple complex diseases.

We first apply PleioPred to Crohn's disease (CD), celiac disease (CEL) and type-II diabetes (T2D) by jointly modeling them with known correlated diseases (i.e. CD with Ulcerative Colitis (UC); CEL with UC; T2D with coronary artery disease (CAD)) and show a statistically significant improvement in prediction performance in independent validation cohort over

single-trait models. By comparing two-trait prediction model with and without functional annotations in both simulation and real data analysis, we demonstrate that functional annotation may further improve the performance of joint modeling. Furthermore, we show that PRS calculated from PleioPred can effectively partition T2D patients by their age of onset, indicating the potential clinical usage of our approach[16, 17]. Through jointly modeling T2D with a wide spectrum of diseases, we demonstrate that the increment in prediction accuracy is significantly correlated with the genetic correlations between T2D and the jointly modeled diseases.

Results

Methods overview

We propose a Bayesian framework to incorporate functional annotations and pleiotropy. We assume throughout the report that the phenotypes of two diseases $Y^{(1)}_{N_1\times 1}$, $Y^{(2)}_{N_2\times 1}$ and the genotypes $X_{N_1\times M}$, $Z_{N_2\times M}$ are standardized with mean zero and variance one. When phenotypes are binary, $Y^{(1)}_{N_1\times 1}$ and $Y^{(2)}_{N_2\times 1}$ denote disease liabilities instead [18, 19]. Here N_1 and N_2 denote the sample sizes for the two diseases and M is the number of markers. We assume a linear model with genotype matrices, effect sizes (β and γ) and random errors (ε and δ) mutually independent as follows

$$Y^{(1)}_{N_1\times 1} = X_{N_1\times M}\beta_{M\times 1} + \varepsilon_{N_1\times 1}$$

$$Y^{(2)}_{N_2\times 1} = Z_{N_2\times M}\gamma_{M\times 1} + \delta_{N_2\times 1}$$

We also assume that the effect sizes of different SNPs are independent. As for random errors, we assume that

$$\begin{pmatrix}\varepsilon\\ \delta\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix}(1-h_1^2)I_{N_1} & S\\ S^T & (1-h_2^2)I_{N_2}\end{bmatrix}\right)$$

$$S_{ij} = \begin{cases}\rho_e & \text{if individual i in study 1 and individual j in study 2 are the same}\\ 0 & \text{if individual i in study 1 and individual j in study 2 are different}\end{cases}$$

where h_1^2 and h_2^2 denote the heritability of two diseases and ρ_e measures the covariance within the overlapping individuals between two studies. Denote the LD matrix and marginal effect size estimator from GWAS as: $\widehat{D}_1 = \frac{1}{N_1}X^TX$, $\widehat{D}_2 = \frac{1}{N_2}Z^TZ$, $\tilde{\beta} = \frac{1}{N_1}X^TY^{(1)}$ and $\tilde{\gamma} = \frac{1}{N_2}Z^TY^{(2)}$. In practice, $\widehat{D}_1$ and $\widehat{D}_2$ can be estimated from a reference panel and we therefore denote the LD matrix as $\widehat{D}$ for convenience. Then following the derivation in Hu et al. [13], we can derive the conditional distribution of GWAS summary statistics as

$$\begin{matrix}\tilde{\beta}\\ \tilde{\gamma}\end{matrix}\Big|\begin{matrix}\beta\\ \gamma\end{matrix}, \widehat{D} \sim N\left(\begin{pmatrix}\widehat{D}\beta\\ \widehat{D}\gamma\end{pmatrix}, \begin{bmatrix}\frac{1-h_1^2}{N_1}\widehat{D} & \frac{N_S\rho_e}{N_1N_2}\widehat{D}\\ \frac{N_S\rho_e}{N_1N_2}\widehat{D} & \frac{1-h_2^2}{N_2}\widehat{D}\end{bmatrix}\right)$$

where N_s is the number of overlapping samples between the two studies. When N_s is relatively small, we can discard terms with $\frac{N_S\rho_e}{N_1N_2}$ to reduce the computation burden.

We first consider an infinitesimal model to account for a polygenic genetic architecture. We assume that the effect sizes follow a multivariate normal distribution:

$$\begin{pmatrix}\beta\\ \gamma\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{bmatrix} diag(\sigma_{1i}^2) & \rho_g * diag(\sigma_{1i}\sigma_{2i}) \\ \rho_g * diag(\sigma_{1i}\sigma_{2i}) & diag(\sigma_{2i}^2)\end{bmatrix}\right)$$

where σ_{1i}^2 and σ_{2i}^2 denote the variance of effect sizes of SNP i and $\rho_g := cor(\beta_i, \gamma_i)$, represents the genetic correlation between two diseases. This is equivalent to a multivariate random effects model with various variance components. Suppose that the whole genome is partitioned into K functional regions $A_1, \ldots, A_K$. We assume that the effect size of a SNP depends on the functional regions it falls in and the effect sizes are additive in the overlapping regions. To be specific, we have

$$var(\beta_i) = \sigma_{1i}^2 = \sum_{c:i\in A_c} \tau_{1c}$$

$$var(\gamma_i) = \sigma_{2i}^2 = \sum_{c:i\in A_c} \tau_{2c}$$

where τ_{jc} denotes the variance of the effect size of SNPs on disease j falling in A_c alone. In the random effects model, the variance of effect size can be interpreted as heritability and thus for convenience, we will use heritability of SNP i instead of the variance of effect size in the rest of the manuscript.

Details on parameter estimation are described in **Methods**. When all the parameters are specified, we can estimate the expectation of the effect sizes given the marginal effect size estimators of two diseases. The PRSs are defined as

$$PRS_1 = \sum_{j=1}^{M} X_j \mathbb{E}(\beta_j | \tilde{\beta}, \tilde{\gamma}, \widehat{D})$$

$$PRS_2 = \sum_{j=1}^{M} Z_j \mathbb{E}(\gamma_j | \tilde{\beta}, \tilde{\gamma}, \widehat{D})$$

Finally, we treat ρ_g as a tuning parameter and the posterior expectation of the effect sizes can be calculated in closed form (Methods).

In practice[9, 13], we note that a sparse model yields higher accuracy for most diseases. Moreover, the infinitesimal model assumption is relatively strong in some cases. For example, two related diseases may only share some causal variants and have no correlation among the effect sizes or the correlation structures may vary across the genome. We therefore propose a hierarchical Bayesian model with a more general assumption and we refer to this framework as the non-infinitesimal model. Under this model, we assume that the effect sizes follow a

mixture distribution.

$$\begin{pmatrix}\beta_i\\ \gamma_i\end{pmatrix}|\vec{p} \triangleq (p_{11},p_{10},p_{01},p_{00}) \sim p_{11}\begin{pmatrix}N(0,\frac{\sigma_{1i}^2}{p_{11}+p_{10}})\\ N(0,\frac{\sigma_{2i}^2}{p_{11}+p_{01}})\end{pmatrix}+p_{10}\begin{pmatrix}N(0,\frac{\sigma_{1i}^2}{p_{11}+p_{10}})\\ \delta_0\end{pmatrix}$$
$$+p_{01}\begin{pmatrix}\delta_0\\ N(0,\frac{\sigma_{2i}^2}{p_{11}+p_{01}})\end{pmatrix}+p_{00}\begin{pmatrix}\delta_0\\ \delta_0\end{pmatrix}$$

$$\vec{p} \sim Dirichlet(\alpha)$$

That is, the effect sizes of SNP *i* for the two diseases follow a mixture distribution with two independent normal distribution (when SNP *i* is causal in both diseases), joint normal and point mass (when SNP *i* is causal in only one diseases) and joint point mass (when SNP *i* is not causal in either disease) [20]. Although we do not have closed form solution for the posterior expectation of the effect sizes, we use Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution of the effect sizes to estimate the posterior expectation (Methods).

For both infinitesimal and non-infinitesimal models, we used a total of 61 different annotation categories, including functional genome predicted by GenoCanyon scores [14], GenoSkyline tissue-specific functionality scores of 7 tissue types [15], and 53 baseline annotations for diverse genomic features [21]. More specifically, GenoCanyon is a statistical framework to predict functional regions in the human genome through integrative analysis of ENCODE epigenomic data and multiple conservation metrics [14]. Later we further extended the framework and developed GenoSkyline, which aimed to predict tissue-specific functionality [15]. We smoothed GenoCanyon scores by a 10Kb window, a strategy previously shown to improve robustness of functionality prediction [22]. The smoothed GenoCanyon annotation and raw GenoSkyline annotations of seven tissue types were dichotomized based on a cutoff of 0.5. The regions with GenoCanyon or GenoSkyline scores greater than the cutoff are interpreted as non-tissue-specific or tissue-specific functional regions in the human genome. Such dichotomization has been previously shown to be robust against the cutoff choice [15].

We compare the prediction performance of eight methods, corresponding to infinitesimal and non-infinitesimal versions of single-trait and two-trait approaches with and without functional annotations. As shown in [9, 13], LDpred and AnnoPred outperform other state-of-the-art PRS methods, we therefore use these two approaches as the representative single-trait prediction methods.

- AnnoPred-inf/AnnoPred: single-trait prediction model with 61 functional annotations
- LDpred-inf/LDpred: single-trait prediction model without functional annotations, corresponding to a special case of AnnoPred when assuming only one annotation covering the whole genome
- PleioPred-anno-inf/PleioPred-anno: two-trait prediction model with 61 functional annotations
- PleioPred-inf/PleioPred: two-trait prediction model without functional annotations, corresponding to a special case of PleioPred-anno when assuming only one annotation covering the whole genome

All of these methods studied require a pre-specified tuning parameter except for PleioPred and PleioPred-anno. To select a suitable tuning parameter, we divided the independent testing

dataset (individual level genotype and phenotype data) into two equal parts (A and B, non-overlapping), and selected the tuning parameters by optimizing prediction accuracy on dataset A. We then evaluated prediction accuracy using the remaining half of testing data, i.e. dataset B. Finally, we repeated the analysis one more time by choosing the tuning parameter on dataset B while evaluating the prediction accuracy on dataset A. Results from these two separate analyses were averaged to quantify model performance. Ideally, the parameter should be tuned in an independent cohort and then evaluated in another independent cohort. However, it is very challenging to find two independent cohorts without any overlapping samples with the training GWAS and we therefore chose a cross-validation scheme. In real data analysis, tuning the parameter within the same cohort may lead to a little bit over-optimistic results due to possible shared confounders. However, the proposed non-infinitesimal models address this issue via a hierarchical Bayesian approach to avoid tuning parameter and thus result in more robust and generalizable estimation. Besides the methods discussed above, we have also compared the performance of proposed joint models with a recently developed multi-trait analysis tool (MTAG [23]). Following the Polygenic Prediction section in their bioRxiv preprint (page 8), we first applied MTAG to GWAS summary statistics to get the multi-trait adjusted p values and effect sizes and then used the generated summary statistics as input to LDpred. The AUC of LDpred with MTAG adjusted summary statistics and all other four methods are shown in S7 Table. Our method outperformed all other methods including MTAG. Notably, MTAG outperformed LDpred in Crohn's disease but its performance was even slightly worse than LDpred for celiac disease and type-II diabetes.

Simulations

We first performed simulations to demonstrate PleioPred's ability to improve risk prediction accuracy. We simulated traits from GERA (dbGaP access number phs000674.v1.p1) genotype data, which contains 61,172 individuals genotyped for 670,176 SNPs. More specifically, we randomly selected ~28,000 individuals as training set to calculate the summary statistics for disease 1 and another ~28,000 for disease 2. The remaining ~5000 individuals were used for testing. Throughout the simulation we used genotype data of chromosome 1 (50,279 SNPs) to generate phenotypes. We first generated two annotations and each annotation was simulated by randomly selecting 10% of the genome, denoted as A_1 and A_2. Denote the heritability of each trait as h_1^2 and h_2^2 (both 30%) and the number of causal variants as m_1 and m_2 (both 300). Causal variants were generated as follows: one third of causal variants were selected from A_1, one third from A_2 and the rest from $(A_1 \bigcup A_2)^C$, of which p of the causal variants was shared by both diseases (0.2 and 0.8). Effect sizes of causal variants were sampled from $N\left(0, \frac{h_1^2}{m_1}\right)$ and $N\left(0, \frac{h_2^2}{m_2}\right)$. We also randomly selected 5000 individuals and 10000 individuals from the training data of disease 1 and 2 respectively to calculate summary statistics in order to study the effect of unbalanced sample sizes on the increment of prediction accuracy.

Correlations between simulated and predicted traits of disease 1 were calculated from 50 replicates under different simulation settings. PleioPred-anno showed the best prediction performance in all settings (Fig 1). The performance of the two-trait model improves as the proportion of shared causal variants increases. In the unbalanced case when the sample size of disease 1 is smaller than that of disease 2, we observed a larger increment in prediction accuracy, indicating that the benefit of integrating large GWAS of genetically correlated diseases and functional annotations when the sample size of disease of interest is moderate.

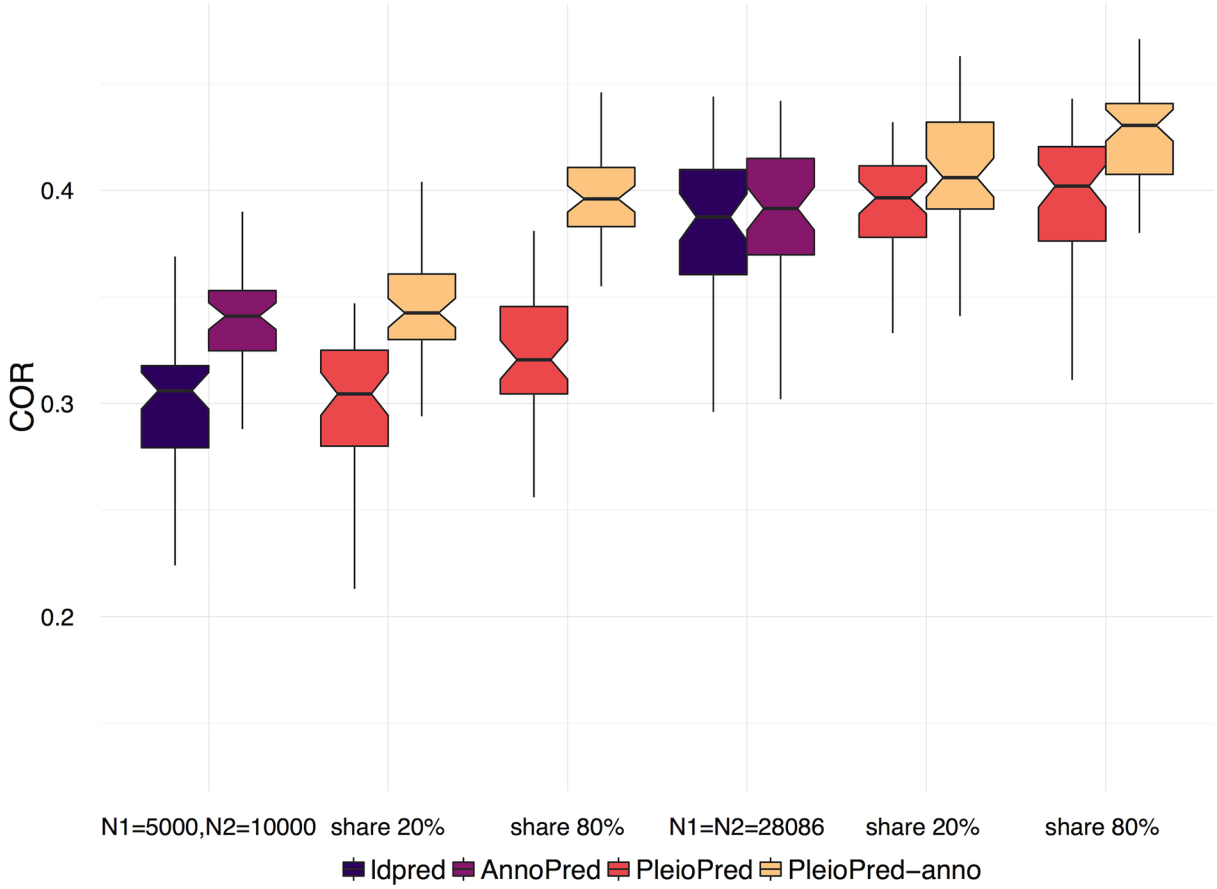

Fig 1. Prediction accuracy of non-infinitesimal models in simulated data. We trained the models with equal training sample sizes ($N_1 = N_2 = 28068$, right panel) and unequal training sizes ($N_1 = 5000$, $N_2 = 10000$, left panel). Prediction accuracy was measured by correlation between simulated traits and predicted PRS.

https://doi.org/10.1371/journal.pgen.1006836.g001

Real data analysis

To further illustrate the improvement in risk prediction accuracy, we first applied PleioPred to Crohn's disease (CD), celiac disease (CEL) and type-II diabetes (T2D). We jointly modeled CD with ulcerative colitis (UC), CEL with UC, and T2D with coronary artery disease (CAD). We trained PleioPred using publicly accessible GWAS summary statistics and evaluated risk prediction performance using individual-level genotype and phenotype data from cohorts independent from the training GWAS samples. The training summary statistics for the two autoimmune disease include the training summary statistics are from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC; CD: N_{case} = 6,333 and $N_{control}$ = 15,056, with samples from the Wellcome Trust Case Control Consortium (WTCCC) removed from the meta-analysis), a CEL GWAS with 4,533 cases and 10,750 controls [24], a UC GWAS from IIBDGC (N_{case} = 6,687 and $N_{control}$ = 19,718). For the validation data, we merged the CD cases from WTCCC (N_{case} = 1,829) and CEL cases from the National Institute of Diabetes and Digestive and Kidney Diseases study (NIDDK, N_{case} = 1,716) with healthy controls from the Resource for Genetic Epidemiology Research on Aging Cohort (GERA, $N_{control}$ = 5,488). For T2D, we trained the model on summary data from the Diabetes Genetics Replication

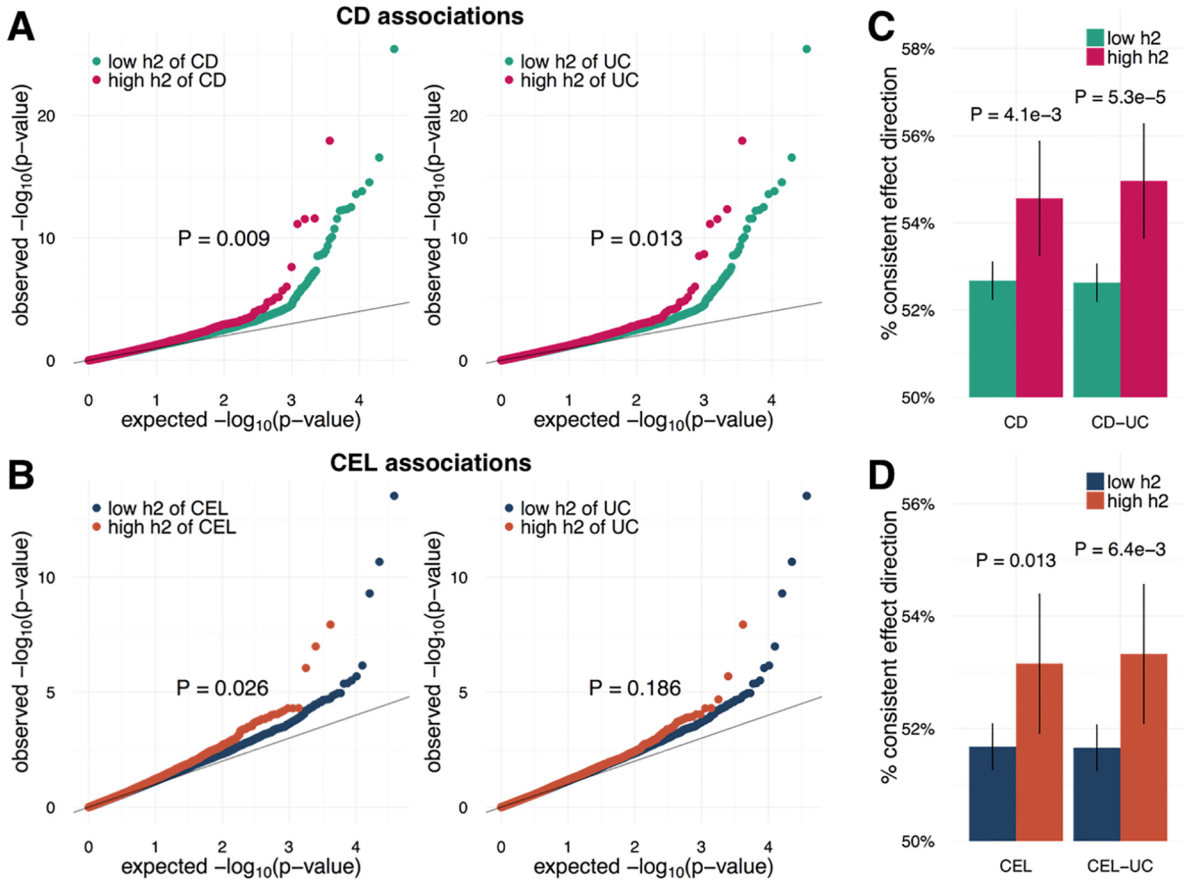

Fig 2. Evaluating effectiveness of annotations and per-SNP heritability. (A, B) Comparing signal strengths of SNPs with high and low heritability of related diseases in independent validation cohorts. Both SNPs with higher heritability of testing disease and related disease have significantly stronger associations across two independent and well-powered testing datasets (N>3,000, **(A)** Crohn's disease; **(B)** Celiac disease.). P-values were calculated using one-sided Kolmogorov-Smirnov test. **(C, D)** Comparing consistency of SNPs' effect direction between training and testing datasets. Each bar quantifies the proportion of SNPs with consistent effect directions. P-values were calculated using one-sided two-sample binomial test. **(C)** Crohn's disease; **(D)** Celiac disease.

https://doi.org/10.1371/journal.pgen.1006836.g002

and Meta-analysis study (DIAGRAM, N_{case} = 12,171 and $N_{control}$ = 56,862) [25] and the Coronary ARtery DIsease Genome wide Replication and Meta-analysis study (CARDIoGRAM, N_{case} = 22,233 and $N_{control}$ = 64,762)[26]. Samples from the Northwestern NUgene Project (N_{case} = 662 and $N_{control}$ = 517) [27] were used for validation. Details for each training GWAS summary statistics and independent testing cohorts are provided in S1 Text and S3 and S4 Tables.

We evaluated the effectiveness of the per-SNP heritability estimated from functional annotations of the two autoimmune diseases (i.e. CD, CEL) with well-powered testing cohorts (N>3,000). Interestingly, not only the per-SNP heritability of the testing diseases (CD and CEL) but those of related diseases (UC) could effectively identify SNPs with large effect sizes (Fig 2A and 2B) and consistent effect directions in independent validation cohorts (Fig 2C and 2D), which shows that functional annotations can effectively prioritize shared causal variants between genetically correlated diseases.

Table 1. Mean CORs and Regression slopes of infinitesimal and non-infinitesimal methods in independent validation cohort of CE, CEL, and T2D. For two-trait prediction models, we jointly modeled CD with UC, CEL with UC and, T2D with CAD.

	COR[a]			Regression Slope[b]		
	CD	CEL	T2D	CD	CEL	T2D
ldpred-inf	0.196	0.072	0.137	0.454	0.168	1.99
AnnoPred-inf	0.219	0.098	0.145	0.572	0.255	2.15
PleioPred-inf	0.246	0.100	0.168	0.661	0.292	2.198
PleioPred-anno-inf	**0.248**	**0.122**	**0.184**	**0.739**	**0.400**	**2.333**
ldpred	0.247	0.120	0.217	0.873	0.661	2.83
AnnoPred	0.279	0.132	0.219	1.306	0.924	2.86
PleioPred	**0.307**	0.141	**0.225**	1.284	1.332	3.05
PleioPred-anno	0.297	**0.156**	0.22	**1.340**	**1.361**	**3.063**

[a] correlations between disease status and PRS;

[b] Regression slopes of logistic regression with case/control status as outcome and PRS as covariates, larger value indicates a larger increase in odds ratio when PRS increases by one unit.

https://doi.org/10.1371/journal.pgen.1006836.t001

Correlations between the calculated PRS and disease status (COR) for different approaches and area under the ROC curve (AUC) are summarized in Table 1 and S1 Table. In both infinitesimal and non-infinitesimal models, we observed that two-trait models consistently outperformed single-trait methods and incorporating functional annotations could further improve the prediction accuracy across different diseases. Furthermore, non-infinitesimal models achieved much better performance than infinitesimal models. We also fitted a logistic regression model with the case/control status as outcome and PRS as covariates and reported the corresponding slopes of PRSs, which measures the increase in odds ratio of getting disease with a unit change in PRS (Table 1) and further validated the advantage of integrating pleiotropy and functional annotations. A likelihood ratio test was used to test for the difference in the prediction accuracy between models comparing the likelihood of a logistic regression fitting PRS of one method to that of a logistic regression fitting PRS of two methods jointly (Table 2). From the test, PleioPred with 61 annotations performed significantly better than single-trait models (infinitesimal model: p = 1.4e-33 for CD, p = 1.6e-12 for CEL and p = 1.7e-3 for T2D; non-infinitesimal model: p = 5.2e-29 for CD, p = 2.8e-7 for CEL and p = 0.027 for T2D). Reversing the order of test (that is, comparing the likelihood of two-trait model with that of two-trait and single-trait model jointly or model using annotations with model using and not using annotations jointly) results in non-significant p-values for most tests (S2 Table), which further demonstrates that PRS incorporating functional annotations and pleiotropy mostly encompasses the information of PRS of single trait model. Besides CAD, we also jointly modeled T2D with a spectrum of traits, whose genetic correlations with T2D have been systematically studied [28], including age at menarche (AAM), autism spectrum (AUT), bipolar disorder (BIP), body mass index (BMI), birth length (BIL), birth weight (BIW), childhood obesity (CHO), fasting glucose (FG), HDL Cholesterol (HDL), height (HGT), major depressive disorder (MDD), rheumatoid arthritis (RA) and schizophrenia (SCZ). We estimated the genetic correlations between T2D and these traits using LDSC[21, 28] and showed that the increment in prediction accuracy is significantly correlated with the genetic correlation between T2D and the jointly modeled traits (P = 0.002; Fig 3 and S1 Fig).

Since COR only measures the global discriminating power of prediction method, it might not be the best evaluation metric for risk prediction approaches, with which it is of more use

Table 2. p-values from the likelihood ratio tests comparing different models.

		CD	CEL	T2D
x_1	x_2	p-values from LRT[a]		
ldpred-inf	**AnnoPred-inf**	**4.4e-15**	**2.8e-6**	**0.011**
ldpred-inf	PleioPred-inf	3.9e-34	2.3e-7	0.041
AnnoPred-inf	PleioPred-anno-inf	1.5e-18	4.9e-8	0.031
PleioPred-inf	PleioPred-anno-inf	1.8e-9	1.9e-8	0.017
ldpred-inf	**PleioPred-anno-inf**	**6.4e-31**	**1.6e-12**	**1.7e-3**
ldpred	**AnnoPred**	**1.3e-5**	**1.7e-5**	**0.066**
ldpred	PleioPred	9.3e-40	0.022	0.039
AnnoPred	PleioPred-anno	8.6e-13	5.7e-5	0.021
PleioPred	PleioPred-anno	7.7e-3	0.014	0.45
ldpred	**PleioPred-anno**	**5.2e-29**	**2.8e-7**	**0.027**

[a] Likelihood ratio = -2[logL(x_1)—logL(x_1 + x_2)], where logL(x_1) and logL(x_1 + x_2) is the log likelihood from a logistic regression with case/control status as outcome and x_1 and x_2 as covariates.

https://doi.org/10.1371/journal.pgen.1006836.t002

to stratify the population into clinically meaningful groups[1, 17, 29]. In order to test different methods' ability to stratify individuals with high risk, we compared the proportion of cases among testing samples with high PRS from non-infinitesimal models in CD and CEL. PleioPred-anno showed highest power in stratifying patients within the top risk population (Fig 4A). For T2D, we compared the distribution of the age of onset within risk groups stratified by different non-infinitesimal PRSs (Fig 4B). Onset ages of T2D are significantly lower among the individuals with higher two-trait PRS than those with higher single-trait PRS, which indicates that PRS of two-trait methods could effectively stratify the population with high absolute risk of T2D and demonstrates the potential clinical usage of the PleioPred and the advantage of joint modeling of related diseases over single-trait prediction methods.

In the non-infinitesimal two-trait model, the major contribution to improved performance came from pleiotropy. That is, the variants that are causal in both diseases would be prioritized and those are not causal or have smaller effect sizes in both diseases would be given lower effect size estimation. Therefore, incorporating a genetically correlated disease is equivalent to integrating a functional annotation and its effectiveness and power depend on the genetic correlation between two diseases. When the two diseases are very similar and share a large amount of causal and non-causal variants, adding less effective annotations may dilute the signals and lead to lower prediction accuracy. This aligns with our results in Tables 1 and 2, in which CD-UC and T2D-CAD have a rather high genetic correlation (0.427, 0.432 respectively) and PleioPred yields better performance. On the contrary, CEL-UC have a relatively lower genetic correlation (0.283) and PleioPred-anno yields the best prediction accuracy. We performed further analysis with T2D and 13 other correlated diseases (those used in Fig 3). We plot the prediction accuracy of PleioPred and PleioPred-anno against absolute genetic correlation and it can be seen that when the functional annotations are fixed, as the absolute genetic correlation increases, PleioPred tends to yield slightly better results (S2 Fig).

Discussion

Our work demonstrates that pleiotropy and functional annotations can effectively improve the performance of genetic risk prediction. PleioPred jointly analyzes genetically correlated diseases and diverse types of annotation data with GWAS summary statistics to upweight

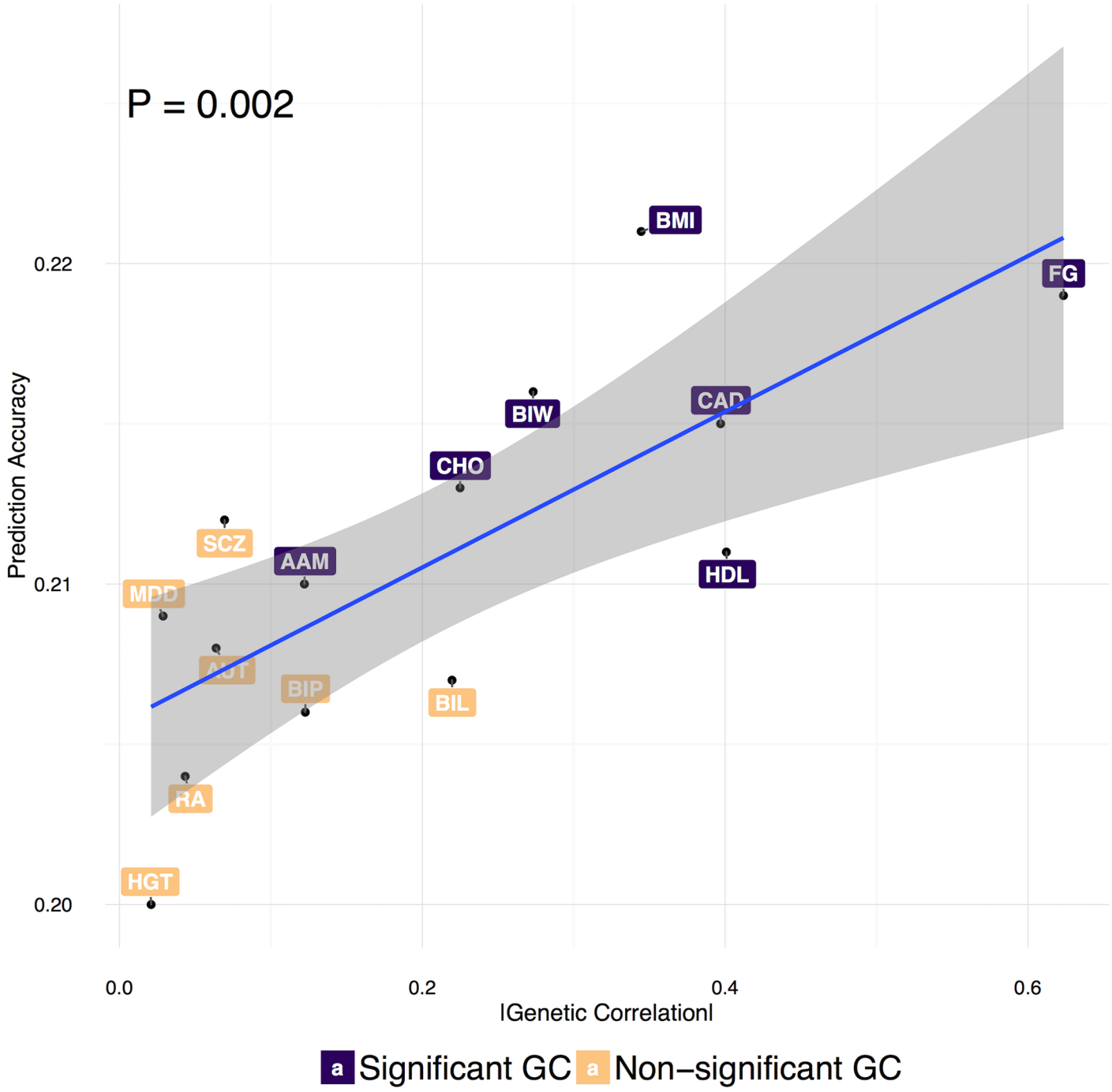

Fig 3. Prediction accuracy of the PleioPred-anno on T2D when jointly modeled with additional traits. Genetic correlations were estimated using LDSC[28] and the significant correlations were labeled in purple. P-value and confidence region indicates the significant correlation between prediction accuracy and genetic correlation. The similar pattern was observed in infinitesimal and non-infinitesimal models without annotations (S1 Fig). AAM: age at menarche, AUT: autism spectrum, BIP: bipolar disorder, BMI: body mass index, BIL: birth length, BIW: birth weight, CHO: childhood obesity, CAD: coronary artery disease, FG: fasting glucose, HDL: HDL Cholesterol, MDD: major depressive disorder, RA: rheumatoid arthritis, and SCZ: schizophrenia.

https://doi.org/10.1371/journal.pgen.1006836.g003

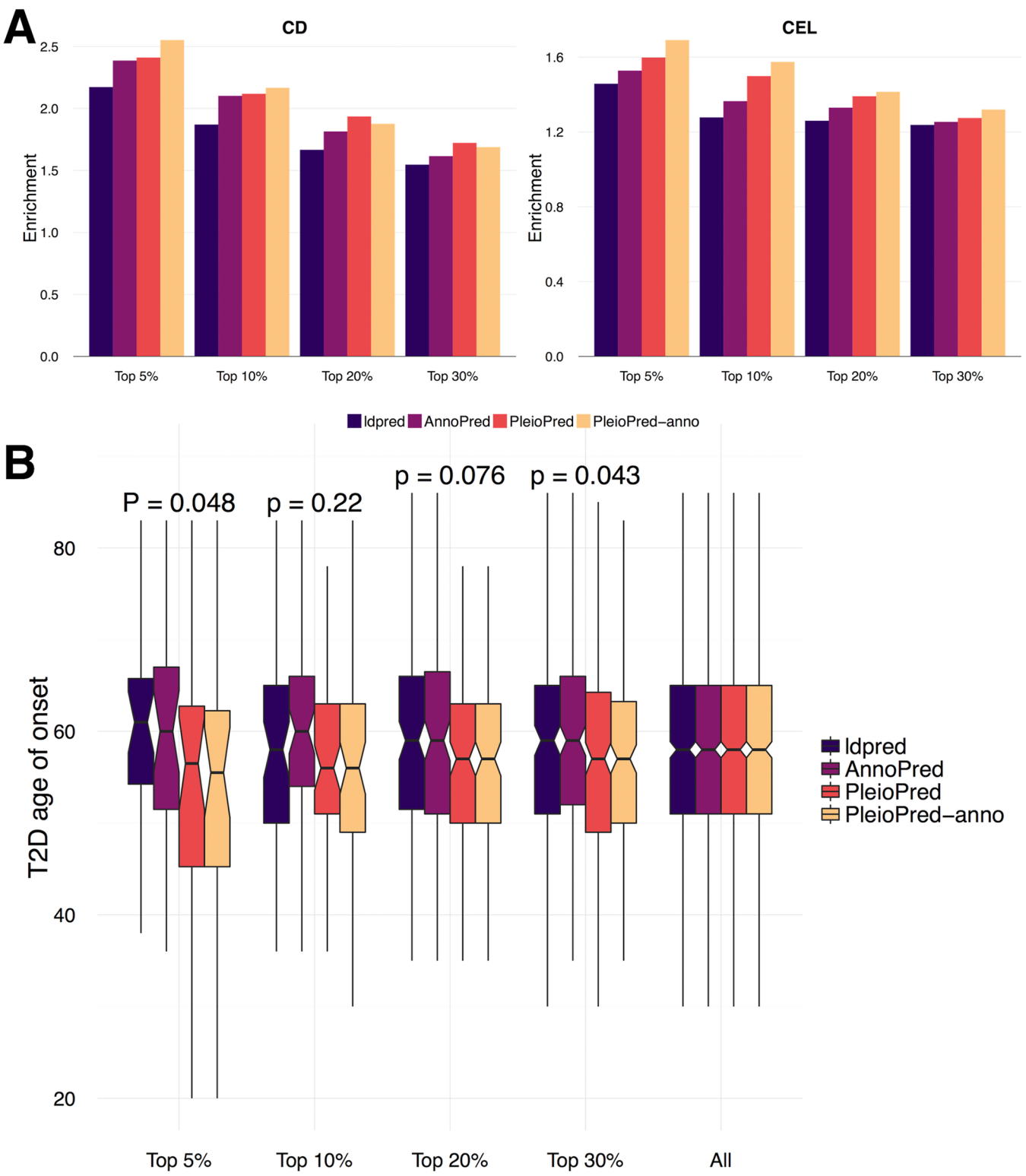

Fig 4. Comparing non-infinitesimal methods in different standards. (A) Enrichment of proportion of cases in testing samples with high PRS (top 1%, 5%, 10%, 20% and 30% risk groups stratified by PRS) in CD and CEL. **(B)** Distribution of age of onset of T2D in testing samples with high PRS (top 5%, 10%, 20% and 30% risk groups stratified by PRS) in T2D. P-values were calculated using Wilcoxon rank test comparing the two-trait models with the one-trait models. The last column represents the overall age of onset in testing samples.

https://doi.org/10.1371/journal.pgen.1006836.g004

causal SNPs shared between diseases and with a higher likelihood of functionality, which lead to consistently better prediction accuracy for multiple complex diseases. Besides prediction accuracy, PleioPred can better stratify population into different risk groups and has greater potential in clinical usage. Our method is not without limitation. First, despite consistent improvement compared with existing PRS-based methods, AUCs for most diseases remain moderate. In order to effectively stratify risk groups for clinical usage, our model remains to be further calibrated using large cohorts with measured environmental and clinical risk factors [1]. Second, accurate estimation of GWAS signal enrichment and SNP effect sizes requires a large sample size for the training dataset. This could be potentially improved by better estimators for annotation-stratified heritability in the future [30]. Third, it is non-trivial to foresee whether PleioPred or PleioPred-anno would work better for a given pair of diseases. According to our observation in real data analysis, PleioPred would eventually outperform PleioPred-anno with an increasing genetic correlation. The threshold at which the change happens could be learned with a validation dataset in practice. The proposed framework can be easily customized and extended to incorporate more than two diseases, which could potentially further increase the prediction accuracy. However, it is worth noting that computation burden and the difficulty in model fitting also increases with the number of diseases. Furthermore, many GWAS have shared control samples, which may result in duplicated information and noise in the training samples. A few Bayesian models combining GWAS summary statistics with functional annotations have been proposed for the purpose of fine-mapping functional variants [31–33]. Whether these models could be adapted to benefit risk prediction accuracy remains to be investigated in the future. Importantly, the rich collection of publicly available integrative annotation data, in conjunction with the increasing accessibility of GWAS summary statistics, makes PleioPred a customizable and powerful tool. As GWAS sample size continues to grow, PleioPred has the potential to achieve even better prediction accuracy and become widely adopted as a summary of genetic contribution in clinical applications of risk prediction. Although more and more GWAS summary results are becoming available [34], in order to evaluate the prediction accuracy, a cohort independent with both training GWAS samples is required, which is very challenging to find. We will apply the proposed methods to a wide range of diseases when independent validation data become available in the future.

Methods

Conditional distribution of marginal effect size estimators

Assume the phenotypes of two diseases $Y^{(1)}_{N_1\times 1}$, $Y^{(2)}_{N_2\times 1}$ and the genotypes $X_{N_1\times M}$, $Z_{N_2\times M}$ are standardized with mean zero and variance one. Here N_1 and N_2 denote the sample sizes for the two diseases and M is the number of markers. We further assume a linear model with genotype matrices, effect sizes (β and γ) and random errors (ε and δ) mutually independent.

$$Y^{(1)}_{N_1\times 1} = X_{N_1\times M}\beta_{M\times 1} + \varepsilon_{N_1\times 1}$$

$$Y^{(2)}_{N_2\times 1} = Z_{N_2\times M}\gamma_{M\times 1} + \delta_{N_2\times 1}$$

Assume that the effect sizes of different SNPs are independent. As for random errors, we

assume that

$$\begin{pmatrix} \varepsilon \\ \delta \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} (1-h_1^2)I_{N_1} & S \\ S^T & (1-h_2^2)I_{N_2} \end{bmatrix}\right)$$

$$S_{ij} = \begin{cases} \rho_e & \text{if individual i in study 1 and individual j in study 2 are the same} \\ 0 & \text{if individual i in study 1 and individual j in study 2 are different} \end{cases}$$

where h_1^2 and h_2^2 denote the heritability of two diseases and ρ_e measures the covariance within the overlapping individuals between two studies. Denote the LD matrix and marginal effect size estimator from GWAS as: $\widehat{D}_1 = \frac{1}{N_1}X^TX$, $\widehat{D}_2 = \frac{1}{N_2}Z^TZ$, $\tilde{\beta} = \frac{1}{N_1}X^TY^{(1)}$ and $\tilde{\gamma} = \frac{1}{N_2}Z^TY^{(2)}$. In practice, $\widehat{D}_1$ and $\widehat{D}_2$ can be estimated from a reference panel and we therefore denote the LD matrix as $\widehat{D}$ for convenience. Then following the derivation in [13], we can derive the conditional distribution of GWAS summary statistics as

$$\begin{matrix} \tilde{\beta} \\ \tilde{\gamma} \end{matrix} \Big| \begin{matrix} \beta \\ \gamma \end{matrix}, \widehat{D} \sim N\left(\begin{pmatrix} \widehat{D}\beta \\ \widehat{D}\gamma \end{pmatrix}, \begin{bmatrix} \frac{1-h_1^2}{N_1}\widehat{D} & \frac{N_S\rho_e}{N_1N_2}\widehat{D} \\ \frac{N_S\rho_e}{N_1N_2}\widehat{D} & \frac{1-h_2^2}{N_2}\widehat{D} \end{bmatrix}\right)$$

where N_s is the number of overlapping samples between the two studies. When N_s is relatively small, we can discard terms with $\frac{N_S\rho_e}{N_1N_2}$ to reduce the computation burden. In practice, we usually ignore the overlap between samples mainly due to four reasons: 1) it is usually challenging to estimate the parameter ρ_e and obtain the exact number of overlapping samples. 2) The off-diagonal term $\frac{N_S\rho_e}{N_1N_2}$ is much smaller comparing to the diagonal terms ($\frac{N_S}{N_1N_2} \sim \frac{1}{N_1^2}$). Even in the case of complete overlap where $\frac{N_S\rho_e}{N_1N_2} = \frac{\rho_e}{N_1}$, ρ_e is still at the magnitude of $(1-h_1^2)(1-h_2^2)$. 3) sensitivity analysis through simulations indicated that the method is very robust to overlapping samples (S6 Table). 4) In practice, ρ_e can be estimated via LDSC if N_S is known. However, including the covariance matrix of $\tilde{\beta}$ and $\tilde{\gamma}$ can significantly increase the computational cost and thus increase the variability of estimation.

Infinitesimal model

Assume that the effect sizes follow a multivariate normal distribution:

$$\begin{pmatrix} \beta \\ \gamma \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} diag(\sigma_{1i}^2) & \rho_g * diag(\sigma_{1i}\sigma_{2i}) \\ \rho_g * diag(\sigma_{1i}\sigma_{2i}) & diag(\sigma_{2i}^2) \end{bmatrix}\right)$$

where σ_{1i}^2 and σ_{2i}^2 denote the variance of effect sizes of SNP i and $\rho_g := cor(\beta_i, \gamma_i)$, representing the genetic correlation between two diseases. Suppose that the whole genome is partitioned into K functional regions $A_1, \ldots, A_K$. Specific annotations used in PleioPred were described previously (Results). We assume the effect size of a SNP depends on the functional regions it

falls in and the effect sizes are additive in the overlapping regions:

$$var(\beta_i) = \sigma_{1i}^2 = \sum_{c:i\in A_c} \tau_{1c}$$

$$var(\gamma_i) = \sigma_{2i}^2 = \sum_{c:i\in A_c} \tau_{2c}$$

where τ_{jc} denotes the variance of the effect size of SNPs on disease j falling in A_c alone.

For parameter estimation, we applied a two-stage approach: first,σ_{1i}^2 and σ_{2i}^2 are estimated using annotation stratified LD score regression (LDSC)[21], which is essentially a method of moments estimator since LDSC utilizes the relationship between the second moment of marginal estimators and variance components of each functional region.

$$\mathbb{E}N_1\tilde{\beta}_i^2 \approx N_1 \sum_c \tau_{1c} l(i,c) + 1$$

$$\mathbb{E}N_2\tilde{\gamma}_i^2 \approx N_2 \sum_c \tau_{2c} l(i,c) + 1$$

$$l(i,c) = \sum_{k\in A_c} (\mathbb{E}X_i X_k)^2$$

Specifically for each disease, we use $\widehat{\sigma}_{ji}^2 = C_j(\sum_{c:\ i\in A_c} \widehat{\tau}_{jc})$ to specify the per-SNP heritability for disease j where C_j is a constant calculated from the following equation

$$\sum_i \widehat{\sigma}_{ji}^2 = \widehat{H}_j^2$$

We do not directly use $\sum_{c:\ i\in A_c} \widehat{\tau}_{jc}$ as the per-SNP heritability because it is estimated in the context where all SNPs in the 1000 Genomes database are included in the model [21]. Such per-SNP heritability estimates cannot be extrapolated to the risk prediction context where many fewer SNPs are analyzed [35]. Therefore, we rescale the heritability estimates to better quantify each SNP's contribution towards chip heritability. Following [36], we use a summary statistics-based heritability estimator that approximates the Haseman-Elston estimator:

$$\widehat{H}_j^2 = \frac{(\bar{\chi}_j^2 - 1)}{N_j \bar{l}}$$

where $\bar{\chi}_j^2$ and $\bar{l}$ denote the mean squared marginal estimators ($N_1\tilde{\beta}_i^2$ and $N_2\tilde{\gamma}_i^2$ for diseases 1 and 2) and the mean non-stratified LD score, respectively.

In the GWAS setting, $\widehat{D}$ are usually non-invertible and have very high dimensions. We thus study the posterior distribution of a small chunk of marginal effect size estimators instead. Let $\tilde{\beta}_b$ and $\tilde{\gamma}_b$ be the estimated marginal effect sizes of SNPs in a region b (e.g. an LD block) and the corresponding genotype matrices are X_b and Z_b and sample correlation matrices is $\widehat{D}_b$, respectively. Then the conditional distribution of the marginal effect size estimators is

(assuming no overlapping individuals or omitting the off-diagonal terms)

$$\begin{pmatrix}\tilde{\beta}_b \\ \tilde{\gamma}_b\end{pmatrix} \Big| \begin{pmatrix}\beta_b \\ \gamma_b\end{pmatrix}, \widehat{D}_b \sim N\left(\begin{pmatrix}\widehat{D}_b\beta_b \\ \widehat{D}_b\gamma_b\end{pmatrix}, \begin{bmatrix}\frac{1-h_{1b}^2}{N_1}\widehat{D}_b & 0 \\ 0 & \frac{1-h_{2b}^2}{N_2}\widehat{D}_b\end{bmatrix}\right)$$

$h_{1b}^2 = \sum_{i \in b}\sigma_{1i}^2$ and $h_{2b}^2 = \sum_{i \in b}\sigma_{2i}^2$ denote the heritability of SNPs in region b for the two diseases, which are usually close to zero since the region b is relatively small and can be safely rounded to zero in calculation. We choose the size of b using the standard described in [9].

Finally, we treat ρ_e as a tuning parameter and the posterior expectation of the effect sizes can be calculated as:

$$\mathbb{E}\left(\begin{pmatrix}\beta_b \\ \gamma_b\end{pmatrix} \Big| \begin{pmatrix}\tilde{\beta}_b \\ \tilde{\gamma}_b\end{pmatrix}, \widehat{D}_b\right) \approx \left(\begin{bmatrix}N_1\widehat{D}_b & 0 \\ 0 & N_2\widehat{D}_b\end{bmatrix} + \begin{bmatrix}diag(\sigma_{1i}^2) & \rho_g * diag(\sigma_{1i}\sigma_{2i}) \\ \rho_g * diag(\sigma_{1i}\sigma_{2i}) & diag(\sigma_{2i}^2)\end{bmatrix}^{-1}\right)^{-1}\begin{pmatrix}N_1\tilde{\beta}_b \\ N_2\tilde{\gamma}_b\end{pmatrix}$$

Non-infinitesimal model

In practice[9, 13], we note that a sparse model yields a higher accuracy for most diseases. Moreover, the infinitesimal model assumption is relatively strong in some cases. For example, two related diseases may only share some causal variants and have no correlation among the effect sizes or the correlation structures may vary across the genome. We therefore propose a hierarchical Bayesian model with a more general assumption and we refer to this framework as the non-infinitesimal model. Under this model, we assume that the effect sizes follow a mixed distribution.

$$\begin{pmatrix}\beta_i \\ \gamma_i\end{pmatrix} | \vec{p} \triangleq (p_{11}, p_{10}, p_{01}, p_{00}) \sim p_{11}\begin{pmatrix}N(0, \frac{\sigma_{1i}^2}{p_{11}+p_{10}}) \\ N(0, \frac{\sigma_{2i}^2}{p_{11}+p_{01}})\end{pmatrix} + p_{10}\begin{pmatrix}N(0, \frac{\sigma_{1i}^2}{p_{11}+p_{10}}) \\ \delta_0\end{pmatrix} + p_{01}\begin{pmatrix}\delta_0 \\ N(0, \frac{\sigma_{2i}^2}{p_{11}+p_{01}})\end{pmatrix} + p_{00}\begin{pmatrix}\delta_0 \\ \delta_0\end{pmatrix}$$

$$\vec{p} \sim Dirichlet(\alpha)$$

That is, the effect sizes of SNP i to two diseases follow a mixed distribution with normal (when SNP i is causal in both diseases), joint normal and point mass (when SNP i is causal in only one diseases) and joint point mass (when SNP i is not causal in either disease). Although we do not have closed form solution for the posterior expectation of the effect sizes, we can use Gibbs sampler to sample from the posterior distribution of the effect sizes to estimate the

posterior expectation. The joint posterior distribution of β_i and γ_i given $\tilde{\beta}, \tilde{\gamma}, \beta_{-i}, \gamma_{-i}$ and $\vec{p}$ is

$$
\begin{aligned}
&f(\beta_i, \gamma_i | \tilde{\beta}, \tilde{\gamma}, \beta_{-i},\ \gamma_{-i},\ \vec{p}, \widehat{D}, \sigma_{1i}^2, \sigma_{2i}^2) \\
&\propto p_{11} \sqrt{\frac{1}{N_1 \frac{\sigma_{1i}^2}{p_{11}+p_{10}} + 1}} \sqrt{\frac{1}{N_2 \frac{\sigma_{2i}^2}{p_{11}+p_{01}} + 1}} \begin{pmatrix} N(C_1\Delta_1, \frac{C_1}{N_1}) \\ N(C_2\Delta_2, \frac{C_2}{N_2}) \end{pmatrix} \\
&+ p_{10} \sqrt{\frac{1}{N_1 \frac{\sigma_{1i}^2}{p_{11}+p_{10}} + 1}} \exp\left(-\frac{N_1}{2} C_1 \Delta_1^2\right) \begin{pmatrix} N\left(C_1\Delta_1, \frac{C_1}{N_1}\right) \\ \delta_0 \end{pmatrix} \\
&+ p_{01} \sqrt{\frac{1}{N_2 \frac{\sigma_{2i}^2}{p_{11}+p_{01}} + 1}} \exp\left(-\frac{N_2}{2} C_2 \Delta_2^2\right) \begin{pmatrix} \delta_0 \\ N\left(C_2\Delta_2, \frac{C_2}{N_2}\right) \end{pmatrix} \\
&+ p_{00} \exp\left(-\frac{N_1}{2} C_1 \Delta_1^2\right) \exp\left(-\frac{N_2}{2} C_2 \Delta_2^2\right) \begin{pmatrix} \delta_0 \\ \delta_0 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&\beta_i, \gamma_i | \tilde{\beta}, \tilde{\gamma}, \beta_{-i},\ \gamma_{-i},\ \vec{p}, \widehat{D}, \sigma_{1i}^2, \sigma_{2i}^2 \\
&\sim \tilde{p}_{11} \begin{pmatrix} N(C_1\Delta_1, \frac{C_1}{N_1}) \\ N(C_2\Delta_2, \frac{C_2}{N_2}) \end{pmatrix} + \tilde{p}_{10} \begin{pmatrix} N(C_1\Delta_1, \frac{C_1}{N_1}) \\ \delta_0 \end{pmatrix} + \tilde{p}_{01} \begin{pmatrix} \delta_0 \\ N(C_2\Delta_2, \frac{C_2}{N_2}) \end{pmatrix} + \tilde{p}_{00} \begin{pmatrix} \delta_0 \\ \delta_0 \end{pmatrix}
\end{aligned}
$$

$$
C_1 \triangleq \frac{N_1}{N_1 + \frac{p_{11}+p_{10}}{\sigma_{1i}^2}}, C_2 \triangleq \frac{N_2}{N_2 + \frac{p_{11}+p_{01}}{\sigma_{2i}^2}},\ \ \Delta_1 \triangleq \widehat{\beta}_i - \sum_{j \neq i} \widehat{D}_{ij}\beta_j,\ \ \Delta_2 \triangleq \widehat{\gamma}_i - \sum_{j \neq i} \widehat{D}_{ij}\gamma_j
$$

$$
\tilde{p}_{11} = p_{11} \sqrt{\frac{1}{N_1 \frac{\sigma_{1i}^2}{p_{11}+p_{10}} + 1}} \sqrt{\frac{1}{N_2 \frac{\sigma_{2i}^2}{p_{11}+p_{01}} + 1}} / p_{sum}
$$

$$
\tilde{p}_{10} = p_{10} \sqrt{\frac{1}{N_1 \frac{\sigma_{1i}^2}{p_{11}+p_{10}} + 1}} \exp\left(-\frac{N_1}{2} C_1 \Delta_1^2\right) / p_{sum}
$$

$$
\tilde{p}_{01} = p_{01} \sqrt{\frac{1}{N_2 \frac{\sigma_{2i}^2}{p_{11}+p_{01}} + 1}} \exp\left(-\frac{N_2}{2} C_2 \Delta_2^2\right) / p_{sum}
$$

$$\tilde{p}_{00} = p_{00}\exp\left(-\frac{N_1}{2}C_1\Delta_1^2\right)\exp\left(-\frac{N_2}{2}C_2\Delta_2^2\right)/p_{sum}$$

$$p_{sum} = p_{11}\sqrt{\frac{1}{N_1\frac{\sigma_{1i}^2}{p_{11}+p_{10}}+1}}\sqrt{\frac{1}{N_2\frac{\sigma_{2i}^2}{p_{11}+p_{01}}+1}}$$
$$+p_{10}\sqrt{\frac{1}{N_1\frac{\sigma_{1i}^2}{p_{11}+p_{10}}+1}}\exp\left(-\frac{N_1}{2}C_1\Delta_1^2\right)$$
$$+p_{01}\sqrt{\frac{1}{N_2\frac{\sigma_{2i}^2}{p_{11}+p_{01}}+1}}\exp\left(-\frac{N_2}{2}C_2\Delta_2^2\right)$$
$$+p_{00}\exp\left(-\frac{N_1}{2}C_1\Delta_1^2\right)\exp\left(-\frac{N_2}{2}C_2\Delta_2^2\right)$$

The posterior distribution of $\vec{p}$ is rather complicated and we therefore applied a Metropolis Hastings method to sample $\vec{p}$ and use the following proposing distribution.

$$\vec{p} \sim Dirichlet((\alpha_1 + d_{11},\ \alpha_2 + d_{10},\ \alpha_3 + d_{01},\ \alpha_4 + d_{00}))$$

in which d_{11} represents the number of SNPs that are causal in both diseases, d_{10} and d_{01} represent the number of SNPs that are causal in only one disease and d_{00} denotes the number of non-causal SNPs from previous sampling step. To ensure convergence, we shrink the posterior probability of being causal if the estimation of heritability at current step of either disease is larger than the heritability estimated from the GWAS summary statistics. That is, $(\tilde{p}_{11}, \tilde{p}_{10}, \tilde{p}_{01})$ are shrunken by a factor $c = \min\left(1, \frac{\widehat{h_1^2}}{\sum_j \widehat{\beta^2}_{(i),j}}, \frac{\widehat{h_2^2}}{\sum_j \widehat{\gamma^2}_{(i),j}}\right)$, where $\widehat{\beta}_{(i),j}$ and $\widehat{\gamma}_{(i),j}$ are the sampled effect size of SNP j in the ith iteration. And simulations showed the algorithm yields fast convergence and high accuracy in estimation (S5 Table). An important advantage about the non-infinitesimal approach is that it has no tuning parameters and thus more computationally efficient. Furthermore, by imposing a Bayesian shrinkage, we can better select functionally relevant variants and tune down the unrelated information.

The running time mainly depends on the number of SNPs and iterations in MCMC steps used in prediction and for a typical GWAS dataset with 400,000 SNPs, it usually takes approximately two hours to finish 250 iterations in MCMC (which already leads to good convergence). And we recommend using at least one thousand unrelated individuals with the same ancestry for which summary statistics datasets are obtained from following the same guideline of [9].

Ethics statement

The study was approved by YALE UNIVERSITY HUMAN INVESTIGATION COMMITTEE with approval number 100 FR1 and 100 FR27.

Software availability

PleioPred software: https://github.com/yiminghu/PleioPred

AnnoPred software: https://github.com/yiminghu/AnnoPred
GenoCanyon: http://genocanyon.med.yale.edu/
GenoSkyline: http://genocanyon.med.yale.edu/GenoSkyline

Supplemental data

Supplemental data include two figures and six tables and detailed description of GWAS summary statistics and validation cohorts.

Supporting information

S1 Fig. Prediction accuracy of the PleioPred-inf, PleioPred-anno-inf and PleioPred on T2D when jointly modeled with a wide spectrum of diseases. Genetic correlations were estimated using LDSC[28] and significant correlations were labeled in purple. P value and confidence region indicates the significant correlation between increment in prediction accuracy and genetic correlation. AAM: age at menarche, AUT: autism spectrum, BIP: bipolar disorder, BMI: body mass index, BIL: birth length, BIW: birth weight, CHO: childhood obesity, CAD: coronary artery disease, FG: fasting glucose, HDL: HDL Cholesterol, MDD: major depressive disorder, RA: rheumatoid arthritis and SCZ: schizophrenia.
(TIFF)

S2 Fig. Prediction accuracy of the PleioPred and PleioPred-anno on T2D when jointly modeled with a wide spectrum of diseases. Genetic correlations were estimated using LDSC [28]. (AAM: age at menarche, gc (genetic correlation) = 0.1221; AUT: autism spectrum, gc = 0.0638; BIP: bipolar disorder, gc = 0.1227; BMI: body mass index, gc = 0.3445; BIL: birth length, gc = 0.2196; BIW: birth weight, gc = 0.2732; CHO: childhood obesity, gc = 0.2249; CAD: coronary artery disease, gc = 0.432; FG: fasting glucose, gc = 0.6234; HDL: HDL Cholesterol, gc = 0.4008; MDD: major depressive disorder, gc = 0.0288; RA: rheumatoid arthritis, gc = 0.0434; and SCZ: schizophrenia, gc = 0.0694).
(TIFF)

S1 Table. Mean AUCs of infinitesimal and non-infinitesimal methods in independent validation cohort of CE, CEL, and T2D. For two-trait prediction models, we jointly modeled CD with UC, CEL with UC, and T2D with CAD.
(XLSX)

S2 Table. p-values from the likelihood ratio tests comparing different models.
(XLSX)

S3 Table. URLs of GWAS summary statistics.
(XLSX)

S4 Table. URLs of validation data.
(XLSX)

S5 Table. Accuracy of parameter estimation in simulations using the proposed MCMC method. Data were generated from real genotype data of chromosome 1 with 29,596 individuals for both traits. We random selected 300 out of 41,334 SNPs as causal variants with 1/3 shared between two traits. We simulated in total 6 scenarios corresponding to different heritability of two traits. In each setting we use the maximum of mean squared error (MAX_MSE) of effect sizes of 41,334 SNPs to evaluate the estimation accuracy.
(XLSX)

S6 Table. Influence of overlapped individuals in training samples. Data were generated from real genotype data of chromosome 1 with 29,596 individuals for both traits. We random selected 300 out of 41,334 SNPs as causal variants with 1/3 shared causal variants. N_s: the number of overlapping individuals between diseases; rho_e: the covariance between random errors of two traits on the same individuals (see Methods for details); MAX_MSE1 and MAX_MSE2: the maximum of mean squared error of effect sizes of 41,334 SNPs in two traits respectively (100 replications, used for evaluating estimation accuracy).
(XLSX)

S7 Table. Mean AUCs of MTAG compared with other methods in real data analysis.
(XLSX)

S1 Text. Details on GWAS summary statistics and validation data.
(DOCX)

Acknowledgments

We sincerely thank DIAGRAM, ReproGen, EGG, PGC, GLGC, CARDIoGRAM, MAGIC, IIBDGC, and ImmunoBase for making their GWAS summary data publicly accessible. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We also thank Dr. Bjarni J. Vilhjálmsson for sharing his codes. And finally we thank Jina Li for her insightful suggestions and support.

Author Contributions

Conceptualization: YH QL HZ.

Data curation: YH QL WL YZ ML.

Formal analysis: YH QL WL YZ.

Funding acquisition: HZ.

Investigation: YH QL HZ.

Methodology: YH QL HZ.

Software: YH.

Writing – original draft: YH HZ.

Writing – review & editing: YH QL WL YZ ML HZ.

References

1. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;advance online publication. http://www.nature.com/nrg/journal/vaop/ncurrent/abs/nrg.2016.27.html-supplementary-information.
2. Li C, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. Human genetics. 2014; 133(5):639–50. https://doi.org/10.1007/s00439-013-1401-5 PMID: 24337655
3. Maier R, Moser G, Chen G-B, Ripke S, Coryell W, Potash JB, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. The American Journal of Human Genetics. 2015; 96(2):283–94. https://doi.org/10.1016/j.ajhg.2014.12.006 PMID: 25640677

4. Minnier J, Yuan M, Liu JS, Cai T. Risk classification with an adaptive naive bayes kernel machine model. Journal of the American Statistical Association. 2015; 110(509):393–404. https://doi.org/10.1080/01621459.2014.908778 PMID: 26236061
5. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome research. 2014; 24(9):1550–7. https://doi.org/10.1101/gr.169375.113 PMID: 24963154
6. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. The American Journal of Human Genetics. 2013; 92(6):1008–12. https://doi.org/10.1016/j.ajhg.2013.05.002 PMID: 23731541
7. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013; 9(2):e1003264. https://doi.org/10.1371/journal.pgen.1003264 PMID: 23408905
8. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748–52. https://doi.org/10.1038/nature08185 PMID: 19571811
9. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. The American Journal of Human Genetics. 2015; 97(4):576–92. https://doi.org/10.1016/j.ajhg.2015.09.001 PMID: 26430803
10. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. Journal of the National Cancer Institute. 2015; 107(5):djv036. https://doi.org/10.1093/jnci/djv036 PMID: 25855707
11. Ripke S, Neale BM, Corvin A, Walters JT, Farh K-H, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511(7510):421. https://doi.org/10.1038/nature13595 PMID: 25056061
12. Schrodi SJ, Mukherjee S, Shan Y, Tromp G, Sninsky JJ, Callear AP, et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. Front Genet. 2014; 5(162):1–18.
13. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging Functional Annotations in Genetic Risk Prediction for Human Complex Diseases. PLoS Comput Biol. 2017 Jun 8; 13(6):e1005589. https://doi.org/10.1371/journal.pcbi.1005589 [Epub ahead of print].
14. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. Sci Rep. 2015; 5. https://doi.org/10.1038/srep10576 PMID: 26015273
15. Lu Q, Powles RL, Wang Q, He BJ, Zhao H. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. PLoS Genet. 2016; 12(4):e1005947. https://doi.org/10.1371/journal.pgen.1005947 PMID: 27058395
16. Lall K, Magi R, Morris A, Metspalu A, Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. Genet Med. 2016. https://doi.org/10.1038/gim.2016.103 PMID: 27513194.
17. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA Oncol. 2016; 2(10):1295–302. https://doi.org/10.1001/jamaoncol.2016.1025 PMID: 27228256.
18. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011; 88(3):294–305. https://doi.org/10.1016/j.ajhg.2011.02.002 PMID: 21376301;
19. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012; 44(4):369–75, S1–3. https://doi.org/10.1038/ng.2213 PMID: 22426310;
20. Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet. 2014; 10(11):e1004787. https://doi.org/10.1371/journal.pgen.1004787 PMID: 25393678;
21. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics. 2015.
22. Lu Q, Yao X, Hu Y, Zhao H. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics. 2016; 32(4):542–8. https://doi.org/10.1093/bioinformatics/btv610 PMID: 26504140
23. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. MTAG: Multi-Trait Analysis of GWAS. bioRxiv. 2017:118810.

24. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nature genetics. 2010; 42(4):295–302. https://doi.org/10.1038/ng.543 PMID: 20190752

25. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics. 2012; 44(9):981. https://doi.org/10.1038/ng.2383 PMID: 22885922

26. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011; 43(4):333–8. https://doi.org/10.1038/ng.784 PMID: 21378990;

27. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011; 4(1):13.

28. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015; 47(11):1236–41. https://doi.org/10.1038/ng.3406 PMID: 26414676;

29. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010; 42(2):105–16. https://doi.org/10.1038/ng.520 PMID: 20081858;

30. Zhou X. A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-wide Association Studies. bioRxiv. 2016:042846.

31. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics. 2014; 94(4):559–73. https://doi.org/10.1016/j.ajhg.2014.03.004 PMID: 24702953

32. Kichaev G, Pasaniuc B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. The American Journal of Human Genetics. 2015; 97(2):260–71. https://doi.org/10.1016/j.ajhg.2015.06.007 PMID: 26189819

33. Li Y, Kellis M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Research. 2016:gkw627.

34. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics. 2016.

35. Yang J, Lee SH, Wray NR, Goddard ME, Visscher PM. Commentary on "Limitations of GCTA as a solution to the missing heritability problem". bioRxiv. 2016. https://doi.org/10.1101/036574

36. Bulik-Sullivan B. Relationship between LD Score and Haseman-Elston Regression. bioRxiv. 2015. https://doi.org/10.1101/018283