PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Leveraging functional annotations in genetic risk prediction for human complex diseases

Yiming Hu[1☯], Qiongshi Lu[1☯], Ryan Powles[2], Xinwei Yao[3], Can Yang[4], Fang Fang[1], Xinran Xu[1], Hongyu Zhao[1,2,5,6*]

1 Department of Biostatistics, Yale School of Public Health, New Haven, CT, United States of America, **2** Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States of America, **3** Yale College, New Haven, CT, United States of America, **4** Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, **5** Department of Genetics, Yale University School of Medicine, New Haven, CT, United States of America, **6** Clinical Epidemiology Research Center (CERC), Veterans Affairs (VA) Cooperative Studies Program, VA Connecticut Healthcare System, West Haven, CT, United States of America

☯ These authors contributed equally to this work.
* hongyu.zhao@yale.edu

OPEN ACCESS

Citation: Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. (2017) Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput Biol 13(6): e1005589. https://doi.org/10.1371/journal.pcbi.1005589

Editor: Isidore Rigoutsos, Thomas Jefferson University, UNITED STATES

Received: October 21, 2016

Accepted: May 19, 2017

Published: June 8, 2017

Data Availability Statement: All the GWAS summary statistics are available online and can be accessed through: http://www.ibdgenetics.org, http://gameon.dfci.harvard.edu, http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/, http://diagram-consortium.org/downloads.html and https://www.immunobase.org/downloads/protected_data/GWAS_Data/. Individual level genotype data are available from dbGaP (accession numbers: phs000147, phs000383, phs000237 and phs000274) and WTCCC on EGA https://www.ebi.

Abstract

Genetic risk prediction is an important goal in human genetics research and precision medicine. Accurate prediction models will have great impacts on both disease prevention and early treatment strategies. Despite the identification of thousands of disease-associated genetic variants through genome wide association studies (GWAS), genetic risk prediction accuracy 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 the presence of linkage disequilibrium. In this paper, we introduce AnnoPred, a principled framework that leverages diverse types of genomic and epigenomic functional annotations in genetic risk prediction for complex diseases. AnnoPred is trained using GWAS summary statistics in a Bayesian framework in which we explicitly model various functional annotations and allow for linkage disequilibrium estimated from reference genotype data. Compared with state-of-the-art risk prediction methods, AnnoPred achieves consistently improved prediction accuracy in both extensive simulations and real data.

Author summary

Genetic risk prediction plays a significant role in precision medicine. Accurate prediction models could have great impact on disease prevention and early treatment strategies. For example, mutations in *BRCA1* and *BRCA2* have been used to evaluate women's breast cancer risk and as a guideline for early screening. However, genetic risk prediction models also present important challenges, including extreme high-dimensionality, limited access to and efficient computational methods for individual-level genotype data. To make use of rich GWAS summary statistics, we propose a novel method to address these challenges by integrating genomic functional annotations, which have been successfully applied in GWAS to generate biological insights. We demonstrate the improvement in accuracy in

ac.uk/ega/ (accession numbers: 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.

both extensive simulation studies and real data analysis of breast cancer, Crohn's disease, celiac disease, rheumatoid arthritis and type-II diabetes.

This is a *PLOS Computational Biology* Methods paper.

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 early treatment strategies [1]. Advancements in high-throughput genotyping technologies and imputation techniques have greatly accelerated discoveries in genome-wide association studies (GWAS) [2]. 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 [3–8], and polygenic risk scores (PRS) estimated using GWAS summary statistics [9, 10]. Despite the potential information loss in summary data, PRS-based approaches have been widely adopted in practice since the summary statistics for large-scale association studies are often easily accessible [11, 12] while individual-level data are more difficult to acquire, deposit, and process. 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) [13].

Explicit modeling and incorporation of external information, e.g. pleiotropy [7, 8] and LD [10], has been shown to effectively improve risk prediction accuracy. Recent advancements in integrative genomic functional annotation, coupled with the rich collection of summary statistics from GWAS, have enabled increase of statistical power in several different settings [14–16]. To our knowledge, the impact of functional annotations on performance of genetic risk prediction has not been systematically studied. Here, we introduce AnnoPred (available at https://github.com/yiminghu/AnnoPred), a principled framework that integrates GWAS summary statistics with various types of annotation data to improve risk prediction accuracy. We compare AnnoPred with state-of-the-art PRS-based approaches and demonstrate its consistent improvement in risk prediction performance using both simulations and real data of multiple complex diseases.

AnnoPred risk prediction framework has three main stages (**Methods**). First, we estimate GWAS signal enrichment in 61 different annotation categories, including functional genome predicted by GenoCanyon scores [17], GenoSkyline tissue-specific functionality scores of 7 tissue types [14], and 53 baseline annotations for diverse genomic features [18] for each trait analyzed. Second, we propose an empirical prior of SNP effect size based on annotation assignment and signal enrichment. In general, SNPs located in annotation categories that are highly enriched for GWAS signals receive a higher effect size prior. Finally, the empirical prior is adopted in a Bayesian framework in which marginal summary statistics and LD matrix estimated from a reference panel are jointly modeled to infer the posterior effect size of each SNP. AnnoPred PRS is defined by

$$PRS = \sum_{j=1}^{M} X_j E_A(\beta_j | \hat{\beta}, \hat{D})$$

where X_j and β_j are the standardized genotype and effect size of the jth SNP, respectively, $\hat{\beta}$ is the marginal estimate of β, $\hat{D}$ is the sample LD matrix, and $E_A(\beta_j|\hat{\beta}, \hat{D})$ denotes the posterior expectation of effect sizes under an empirical prior based on annotation assignment for all SNPs when adjusting for LD matrix estimated from a reference panel (**Methods**).

Results

We first performed simulations to demonstrate AnnoPred's ability to improve risk prediction accuracy. We compared AnnoPred with four popular PRS approaches (**Methods**), including PRS based on genome-wide significant SNPs (PRS_{sig}), PRS based on all SNPs in the dataset (PRS_{all}), PRS based on tuned cutoffs for p-values and LD pruning (PRS_{P+T}), and recently proposed LDpred [10]. Mean correlations between simulated and predicted traits were calculated from 100 replicates under different simulation settings (**Methods**). AnnoPred showed the best prediction performance in all settings when the causal SNPs are highly enriched in annotated regions (**Table 1**, **S2 Table** **and** **S2 Fig**). In general, performance of PRS_{sig}, PRS_{P+T}, LDpred, and AnnoPred all improved under a sparser genetic model and higher trait heritability. PRS_{all} showed comparable performance between sparse and polygenic models but its prediction accuracy was consistently worse than other methods. Sample size in the training set was also crucial for risk prediction accuracy. Increasing sample size could lead to continuous improvement in prediction accuracy under different settings (**Fig 1**).

To illustrate the improved risk prediction performance in real data, we applied AnnoPred to five human complex diseases—Crohn's disease (CD), breast cancer (BC), rheumatoid arthritis (RA), type-II diabetes (T2D), and celiac disease (CEL). We first estimated GWAS signal enrichment in different annotation categories (**Methods**). Enrichment pattern varies greatly across diseases (**Fig 2A**; **S1 Table**), reflecting the genetic basis of these complex phenotypes. Functional genome predicted by GenoCanyon was consistently and significantly enriched for all five diseases. Blood was strongly enriched for three immune diseases, namely CD ($P = 8.9\times10^{-12}$), CEL ($P = 7.0\times10^{-15}$), and RA ($P = 9.9\times10^{-6}$), while gastrointestinal (GI) tract was enriched in CD ($P = 2.6\times10^{-5}$) and CEL ($P = 1.4\times10^{-4}$), both of which have a known GI component. For BC, epithelium ($P = 7.4\times10^{-4}$), GI ($P = 5.9\times10^{-3}$), and muscle ($P = 6.1\times10^{-3}$) were significantly enriched. A few studies have shown that breast cancer could arise from epithelial cells [19, 20]. The connections between breast cancer and muscle as well as GI tract have also been previously suggested [21, 22]. In addition, studies have suggested that GI can be used as

Table 1. Mean correlation between simulated and predicted traits calculated from 100 replicates under different simulation settings. The highest mean correlations are highlighted in boldface. Standard deviations are shown in parentheses. Traits were simulated from WTCCC genotype data, which contain 15,918 individuals genotyped for 393,273 SNPs. In each setting, we used 70% of the data to calculate the training summary statistics and randomly divided the rest 30% into two parts for parameter tuning.

Training samples	Heritability	#Causal	PRS_{sig}	PRS_{all}	PRS_{P+T}	LDpred	AnnoPred
Half (~5K)	0.25	300	0.149(.028)	0.08(.021)	0.25(.028)	0.279(.025)	**0.286**(.024)
		3000	NA*	0.082(.016)	0.073(.020)	0.087(.019)	**0.096**(.020)
	0.5	300	0.304(.04)	0.16(.022)	0.48(.026)	0.502(.033)	**0.512**(.026)
		3000	NA*	0.157(.019)	0.157(.024)	0.195(.021)	**0.209**(.019)
Full (~10K)	0.25	300	0.217(.031)	0.11(.02)	0.332(.023)	0.35(.033)	**0.358**(.022)
		3000	NA*	0.11(.014)	0.107(.018)	0.136(.017)	**0.145**(.017)
	0.5	300	0.373(.036)	0.213(.023)	0.548(.024)	0.557(.047)	**0.566**(.034)
		3000	0.078(.023)	0.21(.019)	0.243(.021)	0.309(.021)	**0.324**(.019)

* NA means no SNP achieves genome-wide significance level (5e-8).

https://doi.org/10.1371/journal.pcbi.1005589.t001

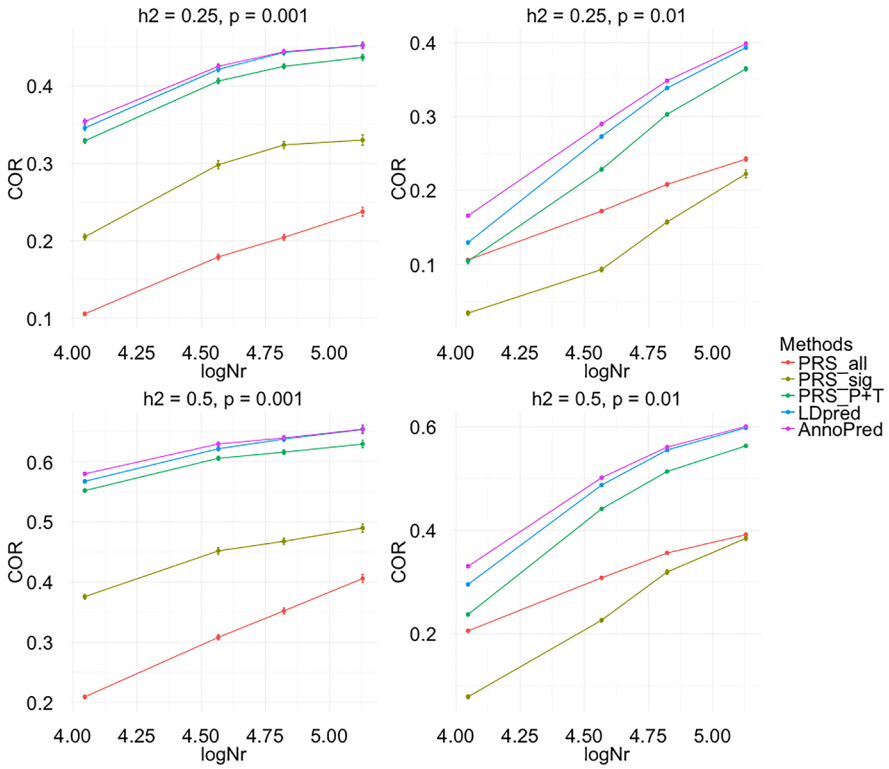

Fig 1. Evaluating the effect of sample size on prediction accuracy in simulation. Traits were simulated using SNPs of chromosome 1, chromosome 1 and 2, chromosome 1 to 4 and the whole genome while keeping the same proportion of causal variants and heritability to mimic the situation of increasing sample size. In the figure, logNr = $\log N\frac{M}{M_s}$, where N is the number of individuals, M is the total number of variants and M_s is the number of variants used in simulation. In total four settings were simulated for each effective sample size: $h^2 = 0.25$, $p = 0.001$; $h^2 = 0.25$, $p = 0.01$; $h^2 = 0.5$, $p = 0.001$; $h^2 = 0.5$, $p = 0.01$, where p represents the proportion of causal variants. Each dot represent the mean COR of 50 replicates in one simulation setting and error bar represents the standard error.

https://doi.org/10.1371/journal.pcbi.1005589.g001

diagnostic and treatment target for type-II diabetes, Crohn's disease, and celiac disease [23–25]. Furthermore, the connection between immune system and Crohn's disease, celiac disease and rheumatoid arthritis have been extensively studied in literature [26–28]. Next, we evaluated the effectiveness of proposed empirical effect size prior in three diseases (i.e. CD, CEL, and RA) with well-powered testing cohorts (N>2,000). Interestingly, despite the highly variable enrichment results in training datasets, integrative effect size prior could effectively identify SNPs with large effect sizes and consistent effect directions in independent validation cohorts (**Fig 2B and 2C**).

Correlations between the calculated PRS and disease status (COR) for different approaches are summarized in **Table 2**. AnnoPred showed consistently improved prediction accuracy compared with all other methods across five diseases. Notably, PRS_{sig} and PRS_{all} showed suboptimal performance in these datasets, reaffirming the importance of modeling LD and other external information. 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 (**S11 Table**). From the test, AnnoPred with 61 annotations performed significantly better than LDpred (p = 1.2E-22 for CD, p = 0.045 for BC, p = 4.2E-7 for RA, p = 3.3E-4 for T2D and p = 1.3E-3 for CEL). Reversing the order of test (that is, comparing the likelihood of model using annotations with

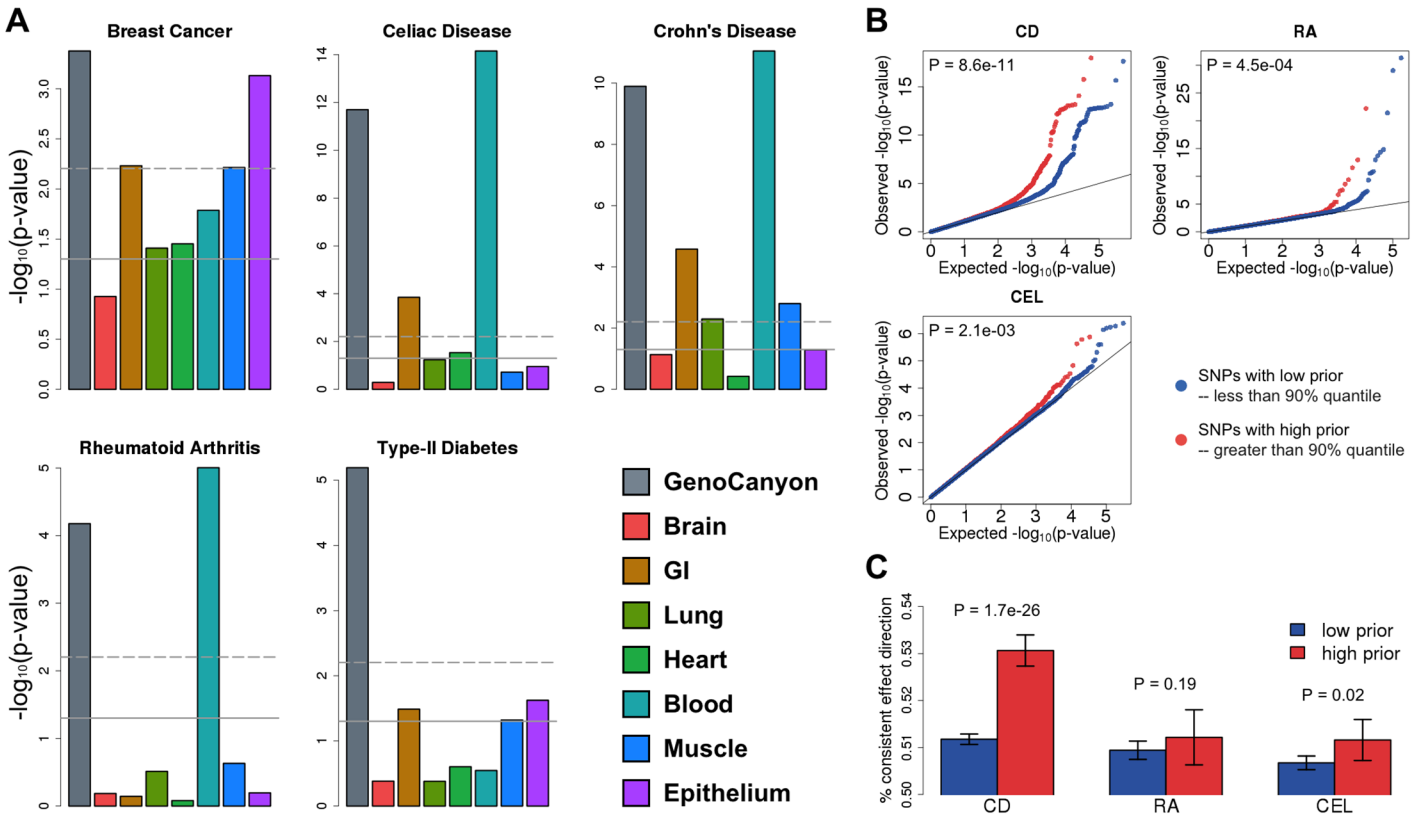

Fig 2. Evaluating effectiveness of annotations and empirical effect size prior. (A) GWAS signal enrichment across GenoCanyon and tissue-specific GenoSkyline annotations. The horizontal lines mark p-value cutoffs of 0.05 and Bonferroni corrected significance level. **(B)** Comparing signal strength of SNPs with high priors and low priors in independent validation cohorts. SNPs with higher priors have significantly stronger associations across three independent and well-powered testing datasets (N>2,000). P-values were calculated using one-sided Kolmogorov-Smirnov test. **(C)** 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.

https://doi.org/10.1371/journal.pcbi.1005589.g002

model using and not using annotations jointly) results in non-significant p-values for most tests (S11 Table), which further demonstrates that PRS incorporating functional annotations mostly encompasses the information of PRS without annotations. To test different methods' ability to stratify individuals with high risk, we compared the proportion of cases among testing samples with high PRS. AnnoPred outperformed all other methods in CD, CEL, RA, and T2D (S1 Fig). Next, we tested AnnoPred's performance using only the 53 baseline annotations and observed a substantial drop in prediction accuracy for all diseases (S3 Table). AnnoPred with GenoCanyon and GenoSkyline annotations only (nine annotation tracks in total) yields better performance than the 53 baseline annotations (S10 Table). For CD and T2D, by using

Table 2. CORs of different methods. The highest CORs are highlighted in boldface.

Disease/Trait	PRS_{sig}	PRS_{all}	PRS_{P+T}	LDpred	AnnoPred
Crohn's Disease	0.27	0.229	0.32	0.325	**0.343**
Breast Cancer	0.084	0.055	0.12	0.122	**0.137**
Rheumatoid Arthritis	0.204	0.114	0.248	0.282	**0.287**
Type-II Diabetes	0.165	0.156	0.204	0.202	**0.22**
Celiac Disease	0.11	0.136	0.18	0.197	**0.213**

https://doi.org/10.1371/journal.pcbi.1005589.t002

these 9 categories AnnoPred even achieved higher accuracy than the model with all 61 annotation tracks added. These results highlight the importance of annotation quality in genetic risk prediction, and also demonstrate GenoCanyon and GenoSkyline's ability to accurately identify functionality in the human genome. Since different diseases have various enrichment patterns, we also run AnnoPred with significantly enriched annotations (enrichment test p value less than 0.05) for each disease (**S10 Table**). In general, using only the significantly enriched annotations indeed improved the performance in most diseases.

Tissue specificity plays an important role in genetic risk prediction. Integrating more functional annotations with higher tissue and cell type specificity may further increase risk prediction accuracy, especially when the tissue type that is biologically relevant to the disease is not well characterized by the seven available tissue tracks in our current analyses. To explore how these factors will affect the AnnoPred model, we performed a few follow-up analyses. We have recently expanded our GenoSkyline annotations to more than 100 tissue and cell types from the Roadmap Epigenomics Project [29]. We investigated the performance of AnnoPred after integrating 66 annotation tracks representing a spectrum of adult tissue and cell types. As shown in S10 Table, incorporating more annotations into the model does not always further improve risk prediction accuracy compared with AnnoPred with fewer annotations in the model. This may be due to the overlap between functional regions (e.g. functional annotations for slightly different brain regions) when incorporating too many annotation tracks into the model, which will cause numerically unstable heritability estimates. This is because annotation-stratified LD score regression, the method we used to empirically estimate the informative prior for SNPs' effect sizes, is a multiple linear regression model that regresses SNP-level summary statistics against annotation-stratified LD scores. When two functional annotation tracks are similar, the corresponding LD scores will also be correlated by definition. It is well understood that if multi-collinearity (i.e. correlation among covariates) in multiple regression leads to numerically unstable estimates for regression coefficients [30] (the heritability parameters in our case).

In order to study the effect of highly associated SNPs (e.g. SNPs in MHC regions for immune traits), we repeated the analysis on CD, RA, BC and T2D after removing the SNPs in MHC region (chr6: 28,477,797–33,448,354 bp). Re-analysis of CEL was unnecessary since the training summary statistics of CEL does not contain any SNP in the MHC region. After removing SNPs in MHC regions, the prediction accuracies for RA drops dramatically for all methods and AnnoPred remained to be the method with the best performance (**S9 Table**). For the rest diseases, results varied little from the original analysis. Besides COR, we also included AUCs for all the analysis performed (**S2, S6, S9** and **S10 Tables**), all of which showed consistent patterns.

Due to distinct allele frequencies and LD structures across populations, risk prediction accuracy usually drops when the training and testing samples are from different populations. In order to investigate the robustness of AnnoPred against population heterogeneity, we applied AnnoPred to three non-European cohorts for breast cancer and type-II diabetes while training the model using summary statistics from European-based studies. The CORs and AUCs are summarized in **S6 and S7 Tables**. As expected, we observed a drop in prediction accuracy for all methods. However, AnnoPred still performed the best in all three trans-ethnic validation datasets.

Discussion

Our work demonstrates that functional annotations can effectively improve performance of genetic risk prediction. AnnoPred jointly analyzes diverse types of annotation data and GWAS

summary statistics to upweight SNPs with a higher likelihood of functionality, which lead to consistently better prediction accuracy for multiple complex diseases. Our method is not without limitation. First, despite the consistent improvement compared with existing PRS-based methods, accuracies 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 potentially be improved by new estimators for annotation-stratified heritability [19]. A few Bayesian models combining GWAS summary statistics with functional annotations have been proposed for the purpose of fine-mapping functional variants [16, 20, 21]. 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 AnnoPred a customizable and powerful tool. As GWAS sample size continues to grow, AnnoPred 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.

Methods

Annotation data

GenoCanyon is a statistical framework to predict functional regions in the human genome through integrative analysis of ENCODE epigenomic data and multiple conservation metrics [17]. Later we have further extended the model and developed GenoSkyline, which aimed to predict tissue-specific functionality [14]. In the AnnoPred model, we incorporated GenoCanyon general functionality scores, GenoSkyline tissue-specific functionality scores for seven tissue types (brain, gastrointestinal tract, lung, heart, blood, muscle, and epithelium), and 53 LDSC baseline annotations that covered a variety of genomic features [18] (**S1 Table**). 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 were 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 [14]. Notably, the AnnoPred framework allows users to specify their own choice of annotations.

Heritability partition

We assume throughout the paper that both the phenotype $Y_{N\times 1}$ and the genotypes $X_{N\times M}$ are standardized with mean zero and variance one. We assume a linear model

$$Y_{N\times 1} = X_{N\times M}\beta_{M\times 1} + \boldsymbol{\varepsilon}_{N\times 1}$$

X, β and $\boldsymbol{\varepsilon}$ are mutually independent. We also assume that β is a random effect and effects of different SNPs are independent. A key idea in the AnnoPred framework is to utilize functional annotation information to accurately estimate SNPs' effect sizes. In order to achieve that, we first partition trait heritability by annotations using LD score regression [18]. Since genotypes are standardized, per-SNP heritability is defined as the variance of β_i for the i^{th} SNP, and is used to quantify SNP effect sizes. More specifically, assume there are $K + 1$ pre-defined annotation categories, denoted as $S_0, S_1, \ldots, S_K$ with S_0 representing the entire genome. Under an additive assumption for heritability in overlapped annotations, we have $\beta_i \sim N(0, \sum_{j:i\in S_j} \tau_j)$,

where $\tau_0, \tau_1, \ldots, \tau_K$, quantify the contribution to per-SNP heritability from each annotation category. Denote the estimated marginal effect size of the ith SNP as $\hat{\beta}_i = \frac{X_i^T Y}{N}$, then we have the following approximation

$$E(N\hat{\beta}_i^2) \approx (N-1)\sum_k \tau_k l(i,k) + 1$$

where $l(i,k)$ is the annotation-stratified LD score and N denotes the total sample size. Regression coefficients τ_k are estimated through weighted least squares. The estimated heritability of the ith SNP is then $\widehat{Var}(\beta_i) = \sum_{j:i\in S_j} \hat{\tau}_j$.

Empirical prior of effect size

Based on per-SNP heritability estimates, we propose two different priors for SNP effect sizes to add flexibility against different genetic architecture. For the first prior, we assume that SNP effect size follows a spike-and-slab distribution

$$\beta_i \sim p_0 N\left(0,\ \hat{\sigma}_i^2/_{p_0}\right) + (1-p_0)\delta_0$$

where p_0 is the proportion of causal SNPs in the dataset, and δ_0 is a Dirac function representing a point mass at zero. The empirical variance of each SNP, i.e. $\hat{\sigma}_i^2$, is determined by the annotation categories it falls in. More specifically, we assume $\hat{\sigma}_i^2 = c(\sum_{j:i\in S_j} \hat{\tau}_j)$, where c is a constant calculated from the following equation

$$\sum_i \hat{\sigma}_i^2 = \hat{H}^2.$$

We do not directly use $\sum_{j:i\in S_j} \hat{\tau}_j$ as the empirical variance prior because it is estimated in the context where all SNPs in the 1000 Genomes Project database are included in the model [18]. Such per-SNP heritability estimates cannot be extrapolated to the risk prediction context where many fewer SNPs are analyzed [23]. Therefore, we rescale the heritability estimates to better quantify each SNP's contribution toward chip heritability. Following [24], we use a summary statistics-based heritability estimator that approximates the Haseman-Elston estimator:

$$\hat{H}^2 = \frac{(\bar{\chi}^2 - 1)}{N\bar{l}}$$

where $\bar{\chi}^2$ and $\bar{l}$ denote the mean $N\hat{\beta}_i^2$ and mean non-stratified LD score, respectively.

In the first prior, we assumed the same proportion of causal SNPs but different effect sizes across annotation categories. We now describe the second prior that assumes different proportions of causal SNPs but the same effect size across annotation categories. To be specific, we assume the causal effect size to be $Var(\beta_{causal}) = V$, the total number of SNPs to be M_0, and the overall proportion of causal SNPs to be p_0. The total heritability H_0^2 can then be written as $H_0^2 = p_0 M_0 V$. For the ith SNP, use $T_i = (\bigcap_{j:i\in S_j} S_j) \cap (\bigcap_{k:i\notin S_k} S_k^c)$ to denote the collection of SNPs that share the same annotation assignment with the ith SNP, and let $M_{T_i} = |T_i|$, i.e. the number of SNPs in the set. Then, the total heritability of SNPs in T_i is $H_{T_i}^2 = p_{T_i} M_{T_i} V$, with p_{T_i} denoting the proportion of causal SNPs in T_i. Following these notations, we have

$$\beta_i \sim p_{T_i} N(0, V) + (1 - p_{T_i})\delta_0$$

where $V = \frac{H_0}{p_0 N_0}$ and $p_{T_i} = p_0 \frac{M_0 H_{T_i}^2}{M_{T_i} H_0^2}$. We use $\hat{H}^2$ to estimate H_0^2, and the following formula to

estimate $H^2_{T_i}$.

$$\hat{H}^2_{T_i} = \frac{\sum_{k\in T_i}\sum_{j:k\in S_j}\hat{\tau}_j}{\sum_{k=1}^{M_0}\sum_{j:k\in S_j}\hat{\tau}_j}\hat{H}^2$$

Finally, p_0 is treated as a tuning parameter for both prior functions in our analysis.

Calculation of posterior effect sizes

By Bayes' rule, the posterior distribution of β is:

$$f(\beta|\hat{\beta},\hat{D}) \propto f(\hat{\beta}|\beta,\hat{D})f(\beta)$$

where $\hat{D} = \frac{1}{N}X^TX$ is the sample correlation matrix and $\hat{\beta} = \frac{1}{N}X^TY$ is the marginal effect size estimates. Given β and $\hat{D}$, $\hat{\beta}$ follows a multivariate normal distribution asymptotically with the following mean and variance

$$E\left(\hat{\beta}|\beta,\hat{D}\right) = \frac{1}{N}\left[E(X^TX\beta|\beta,\hat{D}) + E(X^T\boldsymbol{\varepsilon}|\beta,\hat{D})\right] = \hat{D}\beta$$

$$Var\left(\hat{\beta}|\beta,\hat{D}\right) = Var\left(\frac{1}{N}X^T\boldsymbol{\varepsilon}|\beta,\hat{D}\right) = \frac{1}{N}\left(1-h_g^2\right)\hat{D}.$$

However, $\hat{D}$ is usually non-invertible and has very high dimensions. We thus study the posterior distribution of a small chunk of $\hat{\beta}$ instead. Let $\hat{\beta}_b$ be the estimated marginal effect size of SNPs in a region b (e.g. a LD block) and the corresponding genotype matrix is X_b and sample correlation matrix is $\hat{D}_b$. Then the conditional mean and variance of $\hat{\beta}_b$ are

$$E\left(\hat{\beta}_b|\beta_b,\hat{D}_b\right) = \frac{1}{N}\left[E(X_b^TX\beta|\beta_b,\hat{D}_b) + E(X_b^T\boldsymbol{\varepsilon}|\beta_b,\hat{D}_b)\right] = \hat{D}_b\beta_b$$

$$\begin{aligned} Var(\hat{\beta}_b\ |\beta_b,\hat{D}_b) &= \frac{1}{N^2}var(X_b^TX_b\beta_b + X_b^T(X_{-b}\beta_{-b}+\boldsymbol{\varepsilon})|\beta_b,\hat{D}_b) \\ &= \frac{1}{N^2}var(X_b^T(X_{-b}\beta_{-b}+\boldsymbol{\varepsilon})|\beta_b,\hat{D}_b) \\ &= \frac{1}{N^2}X_b^Tvar(X_{-b}\beta_{-b}+\boldsymbol{\varepsilon}\ |\beta_b,\hat{D}_b)X_b \\ &= \frac{1}{N}(1-h_b^2)\hat{D}_b \end{aligned}$$

where $h_b^2 = \sum_{i\in b}\sigma_i^2$ is the heritability of SNPs in region b, and X_{-b} and β_{-b} denote the genotype matrix and effect sizes of SNPs not in region b. The conditional distribution of β_b is:

$$f(\beta_b\ |\hat{\beta}_b,\hat{D}_b) \propto N\left(\hat{D}_b\beta_b,\frac{1}{N}(1-h_b^2)\hat{D}_b\right)\prod_{i\in b}f(\beta_i)$$

$$\propto \begin{cases} N\left(\hat{D}_b\beta_b,\frac{1}{N}(1-h_b^2)\hat{D}_b\right)\prod_{i\in b}\left[p_0N\left(0,\ \sigma_i^2/_{p_0}\right)+(1-p_0)\delta_0\right], \text{under the first prior} \\ N\left(\hat{D}_b\beta_b,\frac{1}{N}(1-h_b^2)\hat{D}_b\right)\prod_{i\in b}[p_{T_i}N(0,V)+(1-p_{T_i})\delta_0], \text{under the second prior} \end{cases}$$

Although it is difficult to derive $E(\beta_b|\hat{\beta}_b, \hat{D}_b)$ from the joint conditional distribution of β_b, each element of β_b follows a mixed normal distribution conditioning on $\hat{\beta}_b$, $\hat{D}_b$, and all other elements in β_b. Therefore, we apply a Gibbs sampler to draw samples from $f(\beta_b|\hat{\beta}_b, \hat{D}_b)$ and use the sample mean as an approximation for $E(\beta_b|\hat{\beta}_b, \hat{D}_b)$. We further performed a sensitivity analysis on the choice of the size of block b (**S6 Fig**). Specifically, we ran AnnoPred on the data of Crohn's disease with different sizes of block and found that the results were robust to the sizes. In practice, the size of block b is specified by the total number of variants divided by 3,000.

Calculation of PRS

PRS is calculated using the following formula

$$PRS = \sum_{j=1}^{M} X_j E_A(\beta_j|\hat{\beta}, \hat{D}),$$

where E_A denotes the posterior expectation as described above. In practice, the individual-level genotype matrix is not available and we use the LD matrix estimated from a reference panel or the validation samples to substitute $\hat{D}$. We apply the same standard of choosing the size of b as described in [10]. Choices of prior and p_0 can be tuned in an independent cohort. For the data analysis described in this work, we adopted a cross-validation scheme to select tuning parameter due to the challenge in finding multiple independent cohorts without overlapping with the training GWAS summary statistics. The training datasets in our real data analyses and simulations are always fixed, i.e. GWAS summary statistics. We did not perform a classical cross-validation by using different subsets of the complete data to train and test our prediction model. The purpose of cross-validation in our study is purely parameter tuning. To select a suitable tuning parameter, we divide the independent testing dataset (individual level genotype and phenotype data) into two equal parts (A and B), and select the tuning parameters by optimizing prediction accuracy on dataset A. We then evaluate prediction accuracy using the remaining half of testing data, i.e. dataset B. Finally, we repeat 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 are averaged to quantify model performance. For T2D where multiple independent cohorts are available (phs000237 and phs000388), we used an independent cohort for parameter tuning and the other for evaluating performance (**S12 Table**). The results are consistent with the cross-validation.

Comparison with existing methods

We compared AnnoPred with several commonly used risk prediction methods based on summary data of association studies. PRS_{sig} and PRS_{all} were both calculated as the inner product of marginal effect size estimates and the corresponding genotypes. PRS_{all} used all the SNPs that are shared between training and testing datasets while PRS_{sig} only used SNPs with p-values below 5×10^{-8} in the training set. PRS_{P+T} used SNPs passing both LD pruning and p-value thresholding. The thresholds are tuned in an independent dataset over a grid (0, 0.1, 0.2, . . . 0.9 for LD; 1, 0.3, 0.1, 0.03, 0.01, 3E-3, 1E-3, 3E-4, 1E-4, 3E-5, 1E-5, 1E-6, 1E-7, 5E-8, 1E-8 for p-value). LDpred can be viewed as a special case of AnnoPred, assuming the whole genome as the only functional annotation. This is because when enrichment is constant (i.e. causal variants are uniformly distributed across the genome), per-SNP heritability estimates would be nearly constant and therefore results in similar performance to LDpred. We have performed an additional simulation to demonstrate this using WTCCC genotype data with ~15K individuals and ~330K variants. We randomly divided the genome into two parts (two annotations)

and uniformly selected causal SNPs. Then the traits were simulated in a similar way as other simulations in this paper. We estimated per-SNP heritability using LDSC in the two annotation categories, respectively. We ran the procedure for 100 times and the distributions of estimated per-SNP heritability in both regions are summarized in the figure below (the dashed line denotes the true per-SNP heritability, added as S4 Fig in the manuscript), which indicates that the per-SNP heritability estimates are uniform across the genome under constant enrichment. Therefore, AnnoPred would be mathematically equivalent with LDpred with enrichment is constant. We downloaded python code for PRS_{P+T} and LDpred from Github (https://github.com/bvilhjal/ldpred). All the tuning parameters were tuned through cross-validation as we did for AnnoPred. Besides all these PRSs, we also compared AnnoPred with a evaluating method used in [5], which uses 1E-1, 1E-2,. . ., 1E-5 as p-value threshold to select SNPs and report the accuracy for the best performed threshold (**S4 and S5 Tables**).

Given that many large-scale GWAS summary statistics have included almost all available cohorts for a disease of interest, it is challenging to find independent datasets with individual-level genotype and phenotype information and sufficient sample sizes. We were able to identify ideal validation datasets for the five diseases we analyzed in this paper. The performance of different methods on more traits shall be evaluated when we get access to more data in the future.

Simulation settings

We simulated traits from WTCCC genotype data, which contain 15,918 individuals genotyped for 393,273 SNPs after filtering variants with missing rate above 1% and individuals with genetic relatedness above 0.05. We first generated two annotations and each annotation was simulated by randomly selecting 10% of the genome, denoted as A_1 and A_2, which we assume are known when applying AnnoPred. Denote the heritability of the trait as h_g^2 (25% or 50%) and the number of causal variants as m (300 or 3,000). Causal variants were generated as follows: $^m/_3$ causal variants were selected from A_1, $^m/_3$ from A_2 and the rest from $(A_1 \cup A_2)^C$ corresponding to a high enrichment of signals in A_1 and A_2. Effect sizes of causal variants were sampled from $N\left(0, \frac{h_g^2}{m}\right)$. For each simulation, we used 70% of the data to calculate the training summary statistics and randomly divided the rest 30% into two parts for parameter tuning. We also randomly selected half of the training data to calculate summary statistics in order to study the effect of sample size on prediction accuracy.

In order to evaluate the improvement in accuracy, we performed a permutation test to compare the CORs of AnnoPred and LDpred. Suppose the CORs of LDpred and AnnoPred in simulations are $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, respectively. And the hypothesis we want to test is

$$H_0 : \mu_x = \mu_y \leftrightarrow H_1 : \mu_x \neq \mu_y$$

where μ_x and μ_y represent the population mean of accuracies of LDpred and AnnoPred. We used $|\bar{x} - \bar{y}|$ as the test statistics and the p value can be calculated as $p = Pr(|\bar{x} - \bar{y}) > |\bar{x}_{obs} - \bar{y}_{obs}||H_0)$, in which $\bar{x} - \bar{y}$ represents the random variable and $\bar{x}_{obs} - \bar{y}_{obs}$ represents the actually observed values. We used permutation to approximate the distribution of $(\bar{x} - \bar{y})$ when H_0 is true. Specifically, we first pooled $x_i's$ and $y_i's$ together. Then $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ and $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n$ were sampled from the pooled data for $N = 10^6$ times and we calculated $(\bar{\tilde{x}} - \bar{\tilde{y}})$ for each $\tilde{x}_i's$ and $\tilde{y}_i's$ sampled, which formed the empirical distribution of $(\bar{x} - \bar{y})$ under H_0. And the p value could be approximated by $\hat{p} = \frac{\sum_{k=1}^{N} I\{|\bar{\tilde{x}}_k - \bar{\tilde{y}}_k| > |\bar{x}_{obs} - \bar{y}_{obs}|\}}{N}$, in which $\bar{\tilde{x}}_k - \bar{\tilde{y}}_k$ represents the sampled test statistic of the kth permutation.

To further study the effect of sample size on prediction performance, we simulated traits using SNPs of chromosome 1, chromosomes 1 and 2, chromosomes 1 to 4 and the whole genome while keeping the same proportion of causal variants and heritability to mimic the situation of increasing sample size. The corresponding relative sample sizes ($N\frac{M}{M_s}$, where N is the number of individuals, M is the total number of variants and M_s is the number of variants used in simulation) for the four scenarios are ~135K, ~67K, 37K and ~11K. For each effective sample size, we simulated traits under four settings: $h^2 = 0.25$, $p = 0.001$; $h^2 = 0.25$, $p = 0.01$; $h^2 = 0.5$, $p = 0.001$; $h^2 = 0.5$, $p = 0.01$, where p represents the proportion of causal variants (Fig 1).

Ethics statement

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

Data access

GWAS summary statistics and validation data

We trained AnnoPred using publicly accessible GWAS summary statistics and evaluated risk prediction performance using individual-level genotype and phenotype data from cohorts independent from the training samples. Only SNPs shared between training and testing datasets were kept in our analyses. Details for each training and testing dataset are provided in S1 Text and S8 Table.

For Crohn's disease, we trained the model using summary statistics from International Inflammatory Bowel Disease Genetics Consortium (IIBDGC; N_{case} = 6,333 and $N_{control}$ = 15,056) [25]. Samples from the Wellcome Trust Case Control Consortium (WTCCC) were removed from the meta-analysis and used as the validation dataset (N_{case} = 1,689 and $N_{control}$ = 2,891) [26]. For breast cancer, we trained the model using summary statistics from Genetic Associations and Mechanisms in Oncology (GAME-ON) study (N_{case} = 16,003 and $N_{control}$ = 41,335) [27], and tested the performance using samples from the Cancer Genetic Markers of Susceptibility (CGEMS) study (N_{case} = 966 and $N_{control}$ = 70) [28]. Shared samples between CGEMS and GAME-ON were removed. We used samples from the CIDR-GWAS of breast cancer for trans-ethnic analysis (N_{case} = 1,666 and $N_{control}$ = 2,038) [29]. For rheumatoid arthritis, we used summary statistics from a meta-analysis with 5,539 cases and 20,169 controls to train the model [30]. WTCCC samples were removed from the meta-analysis and used for validation (N_{case} = 1,829 and $N_{control}$ = 2,892) [26]. For type-II diabetes, the training dataset is Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium GWAS with 12,171 cases and 56,862 controls [31]. We used samples from Northwestern NUgene Project for validation (N_{case} = 662 and $N_{control}$ = 517) [32]. Samples from Institute for Personalized Medicine (IPM) eMERGE project are used for trans-ethnic analysis (African American: N_{case} = 517 and $N_{control}$ = 213; Hispanic: N_{case} = 477 and $N_{control}$ = 102) [33]. The training dataset for celiac disease is from a GWAS with 4,533 cases and 10,750 controls [34]. Samples in the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) celiac disease study were used for validation (N_{case} = 1,716 and $N_{control}$ = 530) [35].

Software availability

AnnoPred software and source code are freely available online at https://github.com/yiminghu/AnnoPred.

Supporting information

S1 Fig. Enrichment of proportion of cases in the top 5% testing samples with high PRS.
(TIFF)

S2 Fig. Boxplots of the simulation results in Table 1, p-values of the permutation tests (Methods) quantify the improvement of AnnoPred over PRS without incorporating functional annotations.
(TIFF)

S3 Fig. Heritability enrichment across GenoCanyon and tissue-specific GenoSkyline annotations. The horizontal line marks no enrichment.
(TIFF)

S4 Fig. Per-SNP heritability estimation under constant enrichment in simulation. Dashed line marks the true per-SNP heritability.
(TIFF)

S5 Fig. Proportion of SNPs in GenoCanyon and tissue-specific GenoSkyline annotations.
(TIFF)

S6 Fig. Prediction accuracy of AnnoPred on Crohn's disease data using different LD radiuses.
(TIFF)

S7 Fig. Comparing signal strength of SNPs with high priors and low priors in independent validation cohorts with underpowered sample size (<2000). (A) Breast cancer **(B)** Type-II diabetes **(C)** Comparing consistency of SNPs' effect direction between training and testing datasets. Each bar quantifies the proportion of SNPs with consistent effect directions. The association tests and effect size estimation on the testing data are underpowered due to the limited sample size.
(TIFF)

S1 Table. GWAS signal enrichment across 61 annotation categories.
(XLSX)

S2 Table. AUCs of different methods. The highest AUCs are highlighted in boldface.
(XLSX)

S3 Table. Comparison of the complete model and AnnoPred with baseline annotations. The highest AUCs are highlighted in boldface.
(XLSX)

S4 Table. Comparison of the AnnoPred with method used in (Speed and Balding 2014) for evaluation in real data analysis. The highest AUCs are highlighted in boldface.
(XLSX)

S5 Table. Comparison of the AnnoPred with method used in (Speed and Balding 2014) for evaluation in simulation. The highest correlations are highlighted in boldface.
(XLSX)

S6 Table. AUCs for trans-ethnic analyses. The highest AUCs are highlighted in boldface.
(XLSX)

S7 Table. CORs for trans-ethnic analyses. The highest CORs are highlighted in boldface.
(XLSX)

S8 Table. URLs for training and testing datasets.
(XLSX)

S9 Table. Prediction accuracies after removing SNPs in MHC regions. The highest CORs/AUCs are highlighted in boldface.
(XLSX)

S10 Table. Prediction accuracies of AnnoPred when different annotations used. The highest CORs/AUCs are highlighted in boldface.
(XLSX)

S11 Table. p-values from the likelihood ratio tests comparing different models.
(XLSX)

S12 Table. Prediction accuracies on T2D when tuning the parameter in an independent cohort.
(XLSX)

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

Acknowledgments

We sincerely thank DIAGRAM, GAME-ON, 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. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355. 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: CY.

Formal analysis: YH QL HZ.

Methodology: YH QL HZ.

Software: YH RP XY.

Validation: YH QL FF XX CY.

Writing – original draft: YH QL 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. https://doi.org/10.1038/nrg.2016.27 PMID: 27140283
2. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews genetics. 2008; 9(5):356–69. https://doi.org/10.1038/nrg2344 PMID: 18398418
3. 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

4. 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
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. 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
7. 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
8. 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
9. 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
10. 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
11. 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.
12. 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
13. 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.
14. 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
15. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics. 2014; 95(5):535–52. https://doi.org/10.1016/j.ajhg.2014.10.004 PMID: 25439723
16. 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
17. 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
18. 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.
19. Zhou X. A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-wide Association Studies. bioRxiv. 2016:042846.
20. 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
21. 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.
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. 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. doi: 10.1101/036574.
24. Bulik-Sullivan B. Relationship between LD Score and Haseman-Elston Regression. bioRxiv. 2015. doi: 10.1101/018283.

25. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010; 42(12):1118–25. https://doi.org/10.1038/ng.717 PMID: 21102463; PubMed Central PMCID: PMCPMC3299551.

26. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447 (7145):661–78. https://doi.org/10.1038/nature05911 PMID: 17554300

27. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics. 2013; 45(4):353–61. https://doi.org/10.1038/ng.2563 PMID: 23535729

28. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics. 2007; 39(7):870–4. https://doi.org/10.1038/ng2075 PMID: 17529973

29. Zheng Y, Ogundiran TO, Falusi AG, Nathanson KL, John EM, Hennis AJ, et al. Fine mapping of breast cancer genome-wide association studies loci in women of African ancestry identifies novel susceptibility markers. Carcinogenesis. 2013:bgt090.

30. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nature genetics. 2010; 42 (6):508–14. https://doi.org/10.1038/ng.582 PMID: 20453842

31. 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

32. 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.

33. Tayo BO, Teil M, Tong L, Qin H, Khitrov G, Zhang W, et al. Genetic background of patients from a university medical center in Manhattan: implications for personalized medicine. PLoS One. 2011; 6(5): e19166. https://doi.org/10.1371/journal.pone.0019166 PMID: 21573225; PubMed Central PMCID: PMCPMC3087725.

34. 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

35. Garner C, Ahn R, Ding YC, Steele L, Stoven S, Green PH, et al. Genome-wide association study of celiac disease in North America confirms FRMD4B as new celiac locus. PLoS One. 2014; 9(7): e101428. Epub 2014/07/08. https://doi.org/10.1371/journal.pone.0101428 PMID: 24999842; PubMed Central PMCID: PMCPmc4084811.