Published in final edited form as:

Nat Genet. 2013 April ; 45(4): 400–405e3. doi:10.1038/ng.2579.

Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies

Nilanjan Chatterjee[1], **Bill Wheeler**[2], **Joshua Sampson**[1], **Patricia Hartge**[1], **Stephen J. Chanock**[1], and **Ju-Hyun Park**[1,3]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Department of Human and Human Services, 6120 Executive Boulevard, Rockville Maryland 20852, USA

[2]Information management system, Rockville, Maryland 20852, USA

[3]Department of Statistics, Dongguk University-Seoul, Seoul 100715, South Korea

Abstract

We report a new model to project the predictive performance of polygenic models based on the number and distribution of effect sizes for the underlying susceptibility alleles and the size of the training dataset. Using estimates of effect-size distribution and heritability derived from current studies, we project that while 45% of the variance of height has been attributed to common tagging Single Nucleotide Polymorphisms (SNP), a model trained on one million people may only explain 33.4% of variance of the trait. Current studies can identify 3.0%, 1.1%, and 7.0%, of the populations who are at two-fold or higher than average risk for Type 2 diabetes, coronary artery disease and prostate cancer, respectively. Tripling of sample sizes could elevate the percentages to 18.8%, 6.1%, and 12.2%, respectively. The utility of future polygenic models will depend on achievable sample sizes, underlying genetic architecture and information on other risk-factors, including family history.

Introduction

For quite some time, many have predicted that the identification of heritable disease susceptibility markers, such as common genetic variants, could eventually lead to stable models for risk-prediction with important individual and public health implications[1]. Even for a trait such as breast cancer, which manifests a modest degree of familial aggregation, a polygenic model based on a comprehensive set of genetic variants could achieve sufficient discriminatory power and thus be applied in targeted screening programs[2]. To date, genome-

Corresponding author: Nilanjan Chatterjee, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Department of Human and Human Services, 6120 Executive Boulevard, Rockville Maryland 20852, USA. Phone: 301-402-7933, Fax: 301-402-0081, chattern@mail.nih.gov.

Author contributions

N.C. led the development of the statistical methods and drafted the manuscript. J.P. contributed to the development of the methods and performed the illustrative analyses. B.W implemented simulation studies. J.S, P.H and S.J.C contributed to designs of various analyses and interpretation of results. All authors participated in reviewing and revising the manuscript.

wide association studies (GWAS) now have identified thousands of common susceptibility variants for a wide spectrum of complex traits. Recent studies, however, indicate that for most individual traits the loci discovered so far, explain only a small fraction of heritability and thus, collectively have low predictive power[3–11].

While the phenomenon of "missing heritability"[12,13] can be due to many factors such as an overestimation of heritability itself, lack of knowledge of gene-gene and gene-environment interactions and contributions from rare variants, there is increasing recognition that a significant part of the heritability comes from a large number of common SNPs, each of which individually has too small of an effect to be detected at the stringent genome-wide significance level with current sample sizes[14–18]. Recent studies, for example, have indicated that while about 200 loci identified through a large GWAS involving more than 100,000 subjects can explain only approximately 10% of the variance of adult height[6], a set of common SNPs included in existing GWAS platforms can explain up to 45% of the variance of the same trait[16]. Similar studies for a number of other complex traits have indicated the presence of significant "hidden heritability" in GWAS[17,19–21].

The gap between estimates of heritability based on known loci and those estimated due to the comprehensive set of common susceptibility variants raises the possibility of substantially improving prediction performance of risk models by using a "polygenic" approach, one that includes many SNPs that do not reach the stringent threshold for genome-wide significance. A major factor that determines how well such a model can perform in predicting a trait value in an independent sample will be the sample size of the "training" dataset based on which the prediction model can be built. Intuitively, as the sample size for the training dataset increases, the effects of the underlying SNPs can be more precisely estimated. Correspondingly, the underlying true polygenic model, which harnesses the full predictive power associated with total heritability associated with the SNPs, will be more accurately approximated.

In this report, we measure the ability of models based on current as well as future GWAS to improve the prediction of individual traits. We develop a new theoretical framework that characterizes the relationship between sample-size and predictive performance of a polygenic model based on the number and distribution of effect-sizes for the underlying susceptibility SNPs and the optimal balance of type-I and type-II error associated with the underlying criterion of SNP selection. Based on this, we provide a realistic assessment of the predictive performance of a polygenic model for each of ten complex traits, namely, the quantitative traits height (HT), body mass index (BMI), total cholesterol (TC), HDL and LDL and the disease traits, Crohn's disease (CD), Type 1 diabetes (T1D), Type 2 diabetes (T2D), coronary artery disease (CAD) and prostate cancer (PrCA). We use a range of effect-size distributions that are consistent with both known discoveries, 412 in total, reported from the largest GWAS of these traits and recent estimates of the "narrow-sense" heritability, i.e. the total heritability of the traits attributable to additive effects of common SNPs.

The results disclose several insights into the predictive ability of existing GWAS, the marginal utility of further increase in sample size, the sample-size threshold beyond which the predictive ability of the models may reach a plateau, the optimal threshold for SNP selection and the joint utility of family history information and polygenic risks. Furthermore, the general theoretical framework we provide can be used to make projections for the predictive utility of different polygenic model building strategies that may utilize alternative statistical algorithms or/and could incorporate other types of effects, such as those due to gene-gene interactions and rare variants.

Results

Throughout, we assess the predictive performance of a model based on its “predictive correlation coefficient” (PCC), which, for a continuous outcome, is equivalent to the Pearson’s correlation coefficient between true and predicted outcomes for the underlying population of subjects. For a binary disease outcome, we show that PCC has a one-to-one mathematical correspondence to the area under the curve (AUC) statistics and other standard measures for discriminatory performance of risk models. In derivation of this formula, we assume a simple but popularly used[22] model building algorithm in which SNPs are first selected for inclusion in the model depending on whether the corresponding individual tests of association achieve a specified significance threshold (α) and then a polygenic score is built by weighing the selected SNPs based on their estimated regression coefficients. The details of the underlying models and assumptions can be found in **Methods** section.

The relationship between predictive performance of the model and the sample size (N) for the training data set is shown in the formula (1), which forms the basis of our analytical calculations **(Methods)**. Simulation studies confirm the accuracy of the expression (1) (Supplementary Figure 1). According to this formula, the predictive performance of a model depends upon: (i) the number of true susceptibility SNPs (M_1) compared to the total number of SNPs under study (M), (ii) the true effect-sizes (β_ms) for the underlying susceptibility SNPs, (iii) the chosen significance level (α) for SNP selection, (iv) the power of the underlying association test to reach that significance level, and (v) the expected value of the estimated regression coefficients and their squared values for the selected SNPs. The sample-size of the training dataset (N) influences both the power of the association test-statistics and the deviations of the estimated regression coefficients from their true values (see **Methods** for more details). Given an effect-size distribution, since the number of underlying susceptibility SNPs (M_1) determine the total variability of the trait explainable by the underlying model, expression (1) can be also re-written in terms of “narrow sense heritability” (h_g^2), which is defined for the purpose of this report to be the heritability of a trait due to additive effects of common tagging SNPs included on current, commercially available SNP microarrays (see formula 2). In all our subsequent analyses, we assume that genotyping platforms based on which most current GWA studies have been conducted to contain approximately on average M=200,000 independent SNPs.

As for a model for complex trait, we first investigated the predictive performance of polygenic models for adult height. Figure 1 shows that the predictive accuracy of polygenic models greatly depends upon the distribution of effect sizes even when all distributions result in a total heritability of 45%[16]. The predictive performance of the model for all sample sizes is the highest when an exponential distribution underlies the effect-sizes. The performance of the model decreases substantially under a two-component, exponential-mixture model, which, compared to the exponential model, provides a much better fit to the observed effect sizes of the known SNPs by allowing for the presence of a larger number of SNPs, each with smaller effect (Supplementary Table 1). Finally, the performance of the model is the lowest under a three component exponential-mixture distribution, which allows an even larger number of SNPs with smaller effects and produces results that are most consistent with the observed discoveries in the GIANT study[6] (Supplementary Table 1). Our methods successfully reproduced results from a predictive analysis reported in the GIANT study in which distinct polygenic models were built with different significance thresholds for SNP selection and their predictive performance were empirically assessed using independently held out assessed. Our method, when applied to the three-component mixture exponential distribution at the given sample size of the GIANT study (N=130,000), provided an accurate approximation for the entire profile of the observed predictive performance of these polygenic models (Figure 1).

Expression (1) illustrates the trade-off between specificity and sensitivity of the SNP selection criterion on the predictive performance of the model. When a more liberal significance threshold (α) is chosen, then the value of the predictive correlation coefficient will increase through the power of the association tests, but will decrease as a function of the underlying type-I error (α). Figure 1 illustrates the optimal threshold for SNP selection that would maximize predictive performance of a model for adult height. Under both the two- and three- component mixture distributions for effect sizes, the optimal significance level initially increases, with sample size, it reaches a plateau and then remains constant or decreases slightly. In contrast, under the single exponential distribution that corresponds to stronger effect sizes, the optimal significance level becomes more stringent as sample size increases.

We next examined the potential predictive performance of polygenic models for a variety of traits that include both quantitative (BMI, TC, HDL, LDL) and qualitative phenotypes (CD, T1D, T2D, CAD, PrCA) that together demonstrate a spectrum of estimated heritability (Table 1). For most traits, we consider a range for the underlying effect-size distributions that are in accord with both reported discoveries from the largest GWAS and recent estimates of narrow-sense heritability (**Methods**, Supplementary Tables 2 and 3). For a few traits for which external estimates of h_g^2 are not available, we considered a range of its values within the limits of total heritability and effect-size distributions that can produce results consistent with the observed discoveries in the largest GWAS.

For all traits, the expected performance of the polygenic models built based on current GWAS (sample size=N) can be predicted fairly accurately (Figure 2 and Figure 3). Although it may be possible to improve the performance of these models by inclusion of SNPs that do not achieve strict genome-wide significance levels, the models are expected to have low to modest predictive power even after optimization of the SNP selection criterion (Table 2). As sample sizes of the future studies increase, the projected performances of the models will have a wider range reflecting the uncertainty associated with estimates of heritability. Nevertheless it is evident that only very large sample sizes can substantially improve the performance of the models, even in some of the best case scenarios. For prostate cancer (PrCA), for example, while a polygenic model built based on the current largest GWAS can be expected to achieve an AUC statistics of about 63%, in the future, a model built based on as many as triple that sample size is expected to increase the AUC statistic only in the range 64–70% (Figure 3). For all disease traits except coronary artery disease (CAD), it appears that the marginal utility of additional sample can be quite small after the size of GWAS reaches 100,000–200,000 subjects. In contrast, for CAD, BMI, and the lipid traits TC and LDL, the performance of predictive models may continue to improve gradually over a much wider range of sample sizes, reaching as high as 500,000 to one million subjects.

The predictive performance of a model strongly depends on the degree of heritability. For any given sample size, more accurate prediction is possible for more heritable traits, such as CD and T1D, than for relatively less heritable traits such as CAD, PrCA and T2D, which is in accord with classical estimates of heritability based on sibling and twin studies. Accordingly, the ability of the models to identify future cases in “high-risk” group varies (Table 3). For example, using models based on current GWAS, the proportion of future cases that could be identified among top 20 percentile of subjects with highest polygenic risk is 71% for T1D and about 32% for T2D. If the sample size for a future GWAS is tripled, then the corresponding proportions would be expected to increase to 75% and 48%, respectively. Among the three common chronic diseases, the proportion of the population that can be identified to have two-fold or higher risk than an average person ranged from 1.1% (CAD) to 7.0% (PrCA) for models built based on current sample sizes (Supplementary Table 4). If the sample size for future studies could be tripled, then these proportions could be elevate to 6.1% (CAD) and 18.8% (T2D).

For all diseases, family history (FH) information alone has low discriminatory ability. However, models including both FH and polygenic scores can perform substantially better than models using polygenic scores alone especially for rare highly familial conditional such as CD and T1D. Even if polygenic scores could be built in the future based on very large sample sizes (e.g. sample-size=5×N), FH is expected to remain an important variable for identifying high-risk subjects (Table 2 and 3).

Discussion

In summary, our analysis demonstrates that the predictive ability of polygenic models depends not only on total heritability, but also on the underlying effect-size distributions (Figure 1). The emerging effect size distributions from large GWAS suggest that although risk prediction models will continue to improve in the future as total sample size accumulates, the improvement will be slow and modest even when common SNPs account for a large proportion of heritability of the underlying traits (Figures 2 and 3). Our analysis also reveals that under the most likely effect-size distributions, the optimal significance threshold for selecting SNPs for prediction models in large GWAS can be more liberal than threshold standard (e.g. $p < 5 \times 10^{-8}$) used for discovery.

We observed that for less common, highly familial conditions, like T1D and CD, risk models including FH and optimal polygenic scores based on current GWAS can identify a large majority of cases by targeting a small group of high-risk individuals (e.g., subjects who fall in the highest quintile of risk). In contrast, for more common conditions with modest familial components, such as T2D, CAD and PrCA, risk models based on GWAS with current (N) or foreseeable sample sizes in the near future (e.g., triple in size, 3×N) can miss a large proportion (>50%) of cases by targeting a small group of high-risk individuals. For these common diseases, polygenic models using current GWAS can identify a small minority of the population with elevated risk. Based on our model, we suggest that it is necessary to augment sample size of current GWAS by at least three times to substantially increase the proportion of high-risk populations identified by polygenic models. Perhaps one day GWAS or sequencing would be carried out as part of standard clinical care and then such information together with electronic medical records (EMR) could be used to build polygenic models based on sufficiently large studies.

Consistent with a previous report[23], our analysis of T1D with and without contribution of the MHC region highlights the limited incremental discriminatory ability of polygenic scores for diseases that have established common and strong risk factors (Tables 2 and 3). Nevertheless, for most diseases, polygenic scores are expected to contribute substantially in addition to family history. One could also expect that in the foreseeable future, even crude family history information such as the presence or absence of the disease in any first degree relative, will remain an important contributing factor for predicting disease-risk in the general population. More detailed information on extended family history, including age-at-onset information, could further enhance predictive utility of these models especially for applications in high-risk family settings.

Our analysis extends beyond prior reports[24–27] to project the predictive performance of polygenic models most of which relied on simulation studies. A previous report[25] had noted that predictive performance of models that include all GWAS SNPs in a polygenic score without SNP selection depends only on the sample size of the training dataset and h_g^2. More general theory shows that an algorithm which includes all SNPs in a model, i.e. uses the significance level of α=1, could be poor and the predictive performance of more efficient algorithms is expected to depend on the underlying effect-size distribution. Previous simulation studies often have relied on hypothetical effect size distributions. Here, we use the effect-size distributions that are implied by constraints imposed by both known discoveries reported from some of the largest GWAS to date and recent estimates of heritability to provide a realistic depiction of the future of genetic risk prediction.

Our results are generally consistent with a recent analysis[28] that used information on risk in monozygotic twins to examine the absolute limits of “personalized medicine” achievable by genome sequencing under the assumption that such technology can ultimately lead to an ideal model that can capture the full spectrum of genetic risk without possibility of any error. In this report, we provide much sharper bounds for what can be achieved in practice using current or future GWAS by taking into account the likely error associated with estimation of underlying risk that is inevitable because of constraints on sample sizes. Emerging effect-size distributions suggest that GWAS will require huge sample size to approach the ideal predictive power associated with additive effects of common SNPs.

Using a metric used by this report together with the assumption of independent susceptibility alleles across traits, for example, we can predict that while GWAS in principle can identify 55.1% of the population who might have two-fold or higher risk than average for at least one of the three common diseases, CAD, T2D and PrCA, the actual proportion which is achievable using current GWAS studies is only 10.7% and for future studies that triple e the sample size is 33.1%. If the susceptibility alleles across these traits are related, however, these proportions could become higher.

In this report, we have made projections based on a simple GWAS polygenic model building algorithm[6,22] after its optimization with respect to the criteria for SNP inclusion. The general framework we constructed (Supplementary Note), however, can be used to assess the likely performance of other, possibly even more efficient, model building strategies. Using this framework, for example, we project that an algorithm that utilizes LASSO-type [29] thresholds and can analyze all SNPs simultaneously, may outperform the standard GWAS polygenic model building algorithm. This may be particularly interesting for large sample sizes and highly heritable traits like height, but we also note that the gains are generally modest in scope (Supplementary Figure 2). Simultaneous modeling of correlated SNPs within small genomic regions can unmask allelic heterogeneity possibly adding to the overall predictive strength of the models[8,30]. Other strategies may include linear mixed modeling[16] and Bayesian methods[31,32] that can construct polygenic scores based on shrinkage estimates for SNP coefficients utilizing specific priors for the effect-size distribution. Although the absolute performance of different algorithms could be somewhat different across settings, the main results we highlight regarding the order of sample sizes required for improvement of risk prediction is intrinsically related to the underlying effect-sizes and are likely to be observed with other algorithms as well.

Our proposed theoretical framework can be used to speculate on the predictive performance of polygenic models that could be built based on rare variants. In an additional illustration (Supplementary Figure 3), under a model that allows large number of susceptibility loci each containing sets of low-penetrant rare variants, we assessed how polygenic models might perform if variants are included in a model as individual cofactors versus using a gene-collapsing strategy that has been advocated for improving power for association tests[33]. We observed that up to a certain range of sample sizes for the training dataset, models based on collapsed variables often can perform better, apparently due to the improved power for detection of the underlying susceptibility loci. For larger sample sizes, however, their performance might fall short compared to models based on individual variants as collapsed variables possibly including neutral variants can cause substantial dilution of effects for the susceptibility loci and the magnitude of such dilution may not diminish with increasing sample size for naive collapsing methods. In the future, it will be of great importance to determine the sample sizes at which such inflection point would occur for different traits depending on the underlying genetic architecture.

In this report, we use a flexible class of mixture-exponential models to specify effect-size distributions. One could specify effect-size distributions using alternative parametric models such as Weibull, Gamma or Beta distributions all of which can generate L-shaped distributions that appear to be natural for specification of effect-sizes of common SNPs.

Although the performance of polygenic models could differ widely in principle under different effect-size distributions, additional analyses (data not shown) indicate that when such alternative models were restricted so that they can also explain discoveries and estimates of heritabilities reported from current GWAS, each produced results that are qualitatively similar to what we report using the mixture of exponential distributions. For future studies of rare variants, however, the range of plausible models for effect-size distributions is substantial and thus evaluating the likely performance of polygenic models based on such variants remains challenging (Supplementary Figure 3).

In conclusion, we have used novel theory together with empirical observations from large GWAS to provide a comprehensive evaluation of future of polygenic risk models using common susceptibility SNPs. Although our analysis points toward the challenges for achieving high-discriminatory[34] power for polygenic risk models especially for common diseases, it is noteworthy that even models with modest discriminatory power can provide important stratification for absolute risk, thus providing a rationale for potential public health applications such as for weighing risks and benefits for a treatment or an intervention[34]. For most common disease, existing models based on established environmental risk factors, if any, also has modest discriminatory power and faces additional challenges for long-term risk prediction as risk-factor history, unlike susceptibility status, can change over lifetime of an individual. In the future, development of robust prediction models will need to integrate a spectrum of alleles, from rare to common variants, and other risk factors as well. The framework outlined in this paper could be used to identify challenges and opportunities for public health application as well as the required resources needed for development of such models.

Methods

Underlying polygenic model

We assume Y is the outcome variable of and $X_1,\ldots,X_M$ are a set of independent covariates that are potentially predictive of Y. Without loss of generality, we will assume all variables are standardized, so that $E(Y) = 0$ and $Var(Y) = 1$ and similarly $E(X_m) = 0$ and $Var(X_m) = 1$ for each m.

We assume that the true relationship between outcome and the set of covariates can be described by the underlying model ($\mathcal{M}$)

$$Y=\sum_{m=1}^{M_1}\beta_m X_m+\sum_{m=M_1+1}^{M}0\cdot X_m+\varepsilon,$$

where M_1 out the M covariates are truly predictive of Y. We also assume ε, the residual term, to be independently distributed of $X=(X_1,\ldots,X_M)$.

Measure of predictive performance of a model

Now suppose an "estimated" prediction model ($\hat{\mathscr{M}}$) is built base on "training" dataset of sample size N to predict Y using the formula

$$\hat{Y}=\sum_{m=1}^{M}\hat{\beta}_m\gamma_m X_m,$$

where γ_m is indicator of whether the variable is selected (γ_m = 1) or not (γ_m = 0) and $\hat{\beta}_m$ is the estimate of β_m for selected variables. We will denote λ to be a generic threshold parameter for the underlying model selection algorithm.

We will define the predictive correlation for the model $\hat{\mathscr{M}}$ to be

$$R_N(\hat{\mathscr{M}})=cor_{\varepsilon,X}(Y,\hat{Y})=\frac{\sum_{m=1}^{M_1}\beta_m\hat{\beta}_m\gamma_m}{\sqrt{\sum_{m=1}^{M}\hat{\beta}_m^2\gamma_m}},$$

where the subscript X and ε signify that the correlation coefficient is computed with respect to the distribution of X and ε in the the underlying population for which prediction is desired while the estimated model $\hat{\mathscr{M}}$ and its associated parameter estimates ($\hat{\beta}_m$s and γ_ms), are held fixed. The only source of variation of R_N($\hat{\mathscr{M}}$) is due to the randomness of the original training dataset based on which $\hat{\mathscr{M}}$ is built. For any fixed N and λ, the expected value of R_N($\hat{\mathscr{M}}$) can be approximated as (see Supplementary Note)

$$\mu_N(\lambda)=\frac{\sum_{m=1}^{M_1}\beta_m e_m(N,\lambda)p_m(N,\lambda)}{\sqrt{\sum_{m=1}^{M}\nu_m(N,\lambda)p_m(N,\lambda)}},$$

where $e_m(N, \lambda) = E_{N,\lambda}(\hat{\beta}_m|\gamma_m = 1)$, $p_m(N, \lambda) = Pr_{N,\lambda}(\gamma_m = 1)$ and $\nu_m(N,\lambda)=E_{N,\lambda}(\hat{\beta}_m^2|\gamma_m=1)$.

GWAS polygenic model building algorithm

Suppose in a GWAS study, independent SNPs are included in a prediction model depending on whether the corresponding marginal trend-test for association achieves a specified significance level α or not. Let Z_m denote the association test-statistics for the m-th SNP and $C_{\alpha/2}$ denote the critical level for a two-sided test at level α. For any SNP that achieves the required significance level, i.e. γ_m = 1, its corresponding coefficient in the prediction model could be taken as $\hat{\beta}_m$, i.e. the estimated regression coefficient from the marginal analysis of the SNP.

Based on general theory developed in Supplementary Note, we show that in the above setting the expected value of predictive correlation coefficient of the above polygenic model building algorithm over different GWAS datasets of sample size N can be written as

$$\mu_N(\alpha)=\frac{\sum_{m=1}^{M_1}\beta_m e_N(\beta_m)pow(N,\beta_m,\alpha)}{\sqrt{\sum_{m=1}^{M_1}\nu_N(\beta)pow(N,\beta_m,\alpha)+(M-M_1)\alpha\nu_N(0)}}, \quad (1)$$

where $pow(N, \beta_m, \alpha)$ denotes the power of the study of size N for detecting an effect-size of β_m at level α, $e_N(\beta_m) = E(\hat{\beta}_m||Z_m| > C_{\alpha/2})$, and $\nu_N(\beta_m)=E(\hat{\beta}_m^2||Z_m|>C_{\alpha/2})$. Based on the formula for $e_N(\beta_m)$ and $\nu_N(\beta_m)$ given in Suppementary Note, it is easy to see that as $N \to \infty$, $e_N(\beta_m) \to \beta_m$ and $\nu_N(\beta_m) \to \beta_m^2$. Thus, it follows that as $N \to \infty$,

$$\mu_N(\alpha) \to \mu_{max}(\alpha)=\mu_{max}=\sqrt{\sum_{m=1}^{M_1}\beta_m^2} \quad (2)$$

Since $\sum_{m=1}^{M_1}\beta_m^2$ is the variance of the trait due to the total additive effects of all susceptibility SNPs, $\mu_{max}=\sqrt{h_g^2}$, where h_g^2 is the total heritability in narrow sense.

Evaluation of AUC statistics and other performance measures for binary disease outcomes

Previously, several reports[2,43,44] have established the relationship between measures of discrminatory ability of risk models and the genetic variance explained by the true underlying polygenic score associated with a set of SNPs. To generalize such results when the polygenic score associated with a set of SNPs may be estimated with error, we assume that the true relationship between the risk of a binary disease outcome D and a set of covariates $X_1,\ldots,X_M$ is given by an underlying logistic model of the form

$$\text{logit}\{\text{pr}(D=1|X)\}=\alpha+\sum_{m=1}^{M_1}\beta_m X_m+\sum_{m=M_1+1}^{M}0\cdot X_m.$$

We assume that a risk-prediction model is built base on a training dataset of sample-size N using the formula

$$\text{logit}\{\text{pr}(D=1|X\}=\hat{\alpha}+\sum_{m=1}^{M}\hat{\beta}_m\gamma_m X_m,$$

where γ_m is indicator of whether the variable is selected ($\gamma_m = 1$) or not ($\gamma_m = 0$) and $\hat{\beta}_m$ is the estimate of β_m for selected variables. Let $\hat{U}=\sum_{m=1}^{M}\hat{\beta}_m\gamma_m X_m$ be the estimated risk for a person with covariate profile X in the underlying logistic scale. Without loss of generality, we assume each covariate X_m has been standardized with respect to its mean and variance of disease free population so that $E(X_m|D = 0) = 0$ and $Var(X_m|D = 0) = 1$. In the Supplementary Note, we show that the distribution of $\hat{U}$ in controls (D=0) and cases (D=1) for large M, M_1, and N can be approximated by normal distributions as

$$\Pr(\hat{U}|D{=}0){\sim}N(0, S_N^2)\text{ and }\Pr(\hat{U}|D{=}1){\sim}N(C_N, S_N^2),$$

where $S_N^2 = Var(\hat{U}|D{=}0) = \sum_{m=1}^{M} \hat{\beta}_m^2 \gamma_m$ and $C_N = Cov(\hat{U}, U|D{=}0) = \sum_{m=1}^{M_1} \beta_m \hat{\beta}_m \gamma_m$. It is noteworthy that while the characterization of the distributions of true risk (U) for cases and controls requires a single parameter, namely the variance of U[2,43,44], the characterizations for the corresponding distributions for estimated risk ($\hat{U}$) requires two parameters, namely the variance of $\hat{U}$ and its covariance with the true risk U.

Now, the area under the curve (AUC), i.e., the probability that value of risk-score will be greater for a randomly selected case than that of a randomly selected control, can be approximated as

$$\text{AUC}_N = \text{pr}(\hat{U}_1 > \hat{U}_0) = \Phi\left(\sqrt{0.5}R_N\right),$$

where $R_N = C_N/S_N$ is the predictive correlation measure defined earlier for continous outcome. Similarly, using above results, other measures of discreminatory performance of models, such as proportion of cases followed (PCF)[2], can be also characterized in terms of R_N.

In the Supplementary Note, we further show that the distribution of estimated risk $\hat{U}$ for subjects conditional on both his/her own disease status D and that of a relative D_R can be approximately characterized as:

$$\begin{gathered}\text{pr}(\hat{U}|D{=}0, D_R{=}0){\sim}N(0, S_N^2), \text{pr}(\hat{U}|D{=}0, D_R{=}1){\sim}N(k_R C_N, S_N^2)\\ \text{pr}(\hat{U}|D{=}1, D_R{=}0){\sim}N(C_N, S_N^2), \text{and pr}(\hat{U}|D{=}1, D_R{=}1){\sim}N\left((1{+}k_R)C_N, S_N^2\right)\end{gathered}$$

where $k_R = 2^{-R}$ is the *coefficient of relationship*. Based on these distributions, we further derive discriminatory ability of risk models that include both polygenic risk scores and family-history.

Estimation of effect-size distribution

We extend our previous methods[14,15,45] to obtain realistic estimates of effect-size distribution for all underlying susceptibility SNPs for individual traits by combining information from both known discoveries from largest GWAS and estimates of h_g^2 that have recently become available for most of the traits we studied. The major steps are: (1) identify the largest GWAS, termed the "current study", for each of the traits and list "observed susceptibility SNPs" that are discovered through these studies; (2) following the design of the discovery studies (Supplementary Table 2), compute the power to detect SNPs with given effect-sizes; (3) obtain an estimate effect size distribution by fitting parametric mixture-exponential distribution to observed susceptibility SNPs after accounting for statistical power for their discovery and (4) incorporate an additional mixture component to the effect-size distribution that can allow a larger number of SNPs with very small effects so that the overall distribution can explain both estimate of heritability due to common variants (h_g^2) and the number of observed discoveries and genetic variances explained in current studies. Below we describe the details for each step.

In step 1, for each trait, we identified the largest GWAS to date (Supplementary Table 2) and constructed a list of observed susceptibility SNPs that could be considered to have been "detected" from this study. All independent SNPs that reach genome-wide significance according to specified criteria for these studies are included in the list of known susceptibility SNPs. Some studies used multistage designs and did not follow-up previously established susceptibility SNPs beyond the first stage. We included such previously established SNPs in our list if they reached the required threshold for follow-up in the first stage of the current study, on the assumption that these SNPs would have reached genome-wide significance had they been followed up like all other SNPs meeting the same criterion. For each observed susceptibility SNP, we obtained the effect-size as $es = \psi^2 \times 2f(1-f)$ where ψ is linear or logistic regression coefficient depending on quantitative or qualitative traits and f is the allele frequency. In the GWAS context, a covariate X in a polygenic model is the number of risk alleles associated with a SNP and thus following the notation in the main text where a covariate X is assumed to be standardized, it follows that $\beta=\psi\sqrt{2f(1-f)}$ and $es = \beta^2$. To minimize bias from the winner's curse, we estimated effect-sizes by excluding discovery stage data whenever replication phase data were available. Otherwise, we corrected for possible bias using statistical techniques[46].

In step 2, we evaluated power for detection for each susceptibility SNP at their observed effect-sizes following the exact design of the original discovery studies (see Supplementary Table 2).

In step 3, we obtained estimate of effect-size distribution by fitting a parametric model to the effect-sizes for observed susceptibility SNPs. In our previous work[14,15,45], we have described non-parametric methods for estimating effect-size distribution within the range of effect-sizes for observed susceptibility SNPs. In this report, we considered use of parametric models that can be used to describe distribution of effect-sizes beyond the range of known discoveries. Specifically, we used the class of mixture of exponential distributions that allows specification of effect-size distribution in a flexible, weakly parametric fashion. The model is very natural as it allows for increasingly large number of susceptibility SNPs with decreasingly smaller effects, a common pattern that is emerging from GWAS. Mathematically, we assumed that the distribution of effect-sizes for all underlying susceptibility SNPs are given by

$$f(es|\boldsymbol{\theta})=\sum_{h=1}^{H}p_h g\left(es|\lambda_h\right),$$

where θ =(p_1, …, p_H, λ_1, …, λ_H) with p_h being the mixture weight for the h-th component, h = 1, …, H and $g(es|\lambda_h)$ is an exponential distribution with mean $1/\lambda_h$. Noting that the set of K observed susceptibility SNPs can be viewed as a random sample from the set of all underlying susceptibility SNPs with probability of sampling for each SNP is proportional to its power for discovery, we constructed a likelihood as

$$L(\boldsymbol{\theta})=\frac{\{\prod_{i=1}^{K}f(es_i|\theta)pow_{study}(es_i|N,\alpha)\}}{\{\int f(es|\boldsymbol{\theta})pow_{study}(es|N,\alpha)\mathrm{des}\}^K},$$

where $pow_{study}(es_i|N, \alpha)$ is the power to detect a SNP with effect size *es* in the current GWAS of size N at a significance level of α. We use Bayesian methods to estimate the parameters of the mixture model based on the above likelihood and non-informative priors for the parameter vectors **p**=(p_1, …, p_H) and λ =(λ_1, …, λ_H). Specifically, we assumed a discrete Dirichlet distribution for **p** that leads to uniform prior for each of the p_h, h = 1, .., H marginally. We assumed λ_h, h = 1, …, H to be independently distributed each following a Gamma distribution with shape and scale parameters a = 0.5 and b = 2×10^4, respectively. Posterior means for all parameters were obtained based on MCMC algorithms. For each trait, among several fitted mixture models with varying H (up to 3), we selected the best mixture model on the basis of the DIC model selection criterion[47]. For all traits except PrCA and CAD, a two-component (H = 2) mixture model was found to be the best fitted distribution. For PrCA and CAD, a single exponential distribution (H = 1) was found to be adequate.

In step 4, we incorporated an additional mixture component to the effect-size distribution estimated in step 3 so that the overall distribution can be used to describe the effect-sizes for all SNPs that contribute to narrow-sense heritability $h_g^2=\sum_{m=1}^{M_1}\beta_m^2$. We observed that if we had assumed that the parametric effect-size distribution estimated based on known loci can be extrapolated to describe the effect-sizes for all susceptibility loci explaining h_g^2, then the expected number of discoveries and the corresponding heritabilities explained in the current GWAS will be substantially larger than those empirically observed in these studies (Supplementary Table 1). Thus it is very likely that the true effect-size distribution for all susceptibility SNPs contributing to narrow-sense heritability is more skewed toward smaller effects. To obtain a properly calibrated effect-size distribution for all susceptibility SNPs, we thus added an additional mixture component to the fitted effect-size distribution that we estimated based on known loci. We assumed

$$f(es|\boldsymbol{\theta}){=}p_{H+1}f(es|\lambda_{H+1}){+}(1{-}p_{H+1})\sum_{h=1}^{H}\hat{p}_h g(es|\hat{\lambda}_h),$$

where the summation in the right side corresponds to the fitted mixture model based on known loci. For any given value of h_g^2, we found the value of parameters p_{H+1} and λ_{H+1} for the additional component by equating the expected and observed number of discoveries and the corresponding heritability explained in the current largest GWAS by solving the equations

$$M_{obs}{=}\sum\nolimits_{m=1}^{M_1}1\,\left(|Z_m|>C_{\frac{\alpha}{2}}\right)\approx M_1\int pow_{study}(es|N,\alpha)f(es|\theta)des \quad (3)$$

and

$$GV_{obs}{=}\sum\nolimits_{m=1}^{M_1}\beta_m^2 1\,\left(|Z_m|>C_{\frac{\alpha}{2}}\right)\approx M_1\int es\,pow_{study}(es|N,\alpha)f(es|\theta)des \quad (4)$$

where α is the genome-wide significance level used for discovery and M_1 is defined by

$$h_g^2{=}\sum_{m=1}^{M_1}\beta_m^2\approx M_1\int es\,f(es|\boldsymbol{\theta})des.$$

We solved for p_{H+1} and λ_{H+1} by performing a grid-search within the ranges $0.01\leq p_{H+1}\leq 0.99$ and $\hat{\lambda}_H\leq\lambda_{H+1}\leq 20\times\hat{\lambda}_H$ where the latter constraint is imposed to allow the mean of the new component to be smaller than that of the smallest component of the fitted distribution by a factor of up to 20-fold. For traits for which estimates of h_g^2 and associated confidence intervals were available, values of h_g^2 were chosen to be at their point estimates (Tables 2 and 3) or varied within the range of their CIs (Figures 2 and 3) and for each such value of h_g^2 a corresponding effect-size distribution was obtained by solving the above equations. For TC, LDL and CAD, for which direct estimates of h_g^2 were not available, we varied the value of h_g^2 to be within 20–80% of the range of total heritability of these traits that are available from family studies. For CAD, however, the range of h_g^2 for which solutions could be found for the equations (3) and (4) were severely restricted. In particular, it appears that the limited number of findings (21 SNPs) from the very large existing GWAS (N=75,000) of this trait automatically imposes major constraint on the upper bound of h_g^2, at least for the class of effect-size distributions we considered.

Characteristics of largest GWAS and associated discoveries are obtained from published reports[6–8,10,36–39]. For each trait, an effect-size sample size is calculated for a single-stage study that has equivalent power as the original study taking into accounting multi-stage genotyping and selective sampling by family history for PrCA. For HT, sample size and reported discoveries correspond to only first-stage of the GIANT study.

The number of discoveries reported takes into account any genomic control adjustment used in the original study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Bowles Biesecker B, Marteau TM. The future of genetic counselling: an international perspective. Nat Genet. 1999; 22:133–7. [PubMed: 10369253]
2. Pharoah PD, et al. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. 2002; 31:33–6. [PubMed: 11984562]
3. van Hoek M, et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes. 2008; 57:3122–8. [PubMed: 18694974]
4. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008; 358:2796–803. [PubMed: 18579814]
5. Wacholder S, et al. Performance of common genetic variants in breast-cancer risk models. N Engl J Med. 2010; 362:986–93. [PubMed: 20237344]
6. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–8. [PubMed: 20881960]
7. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42:937–48. [PubMed: 20935630]
8. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–13. [PubMed: 20686565]
9. Jostins L, Barrett JC. Genetic risk prediction in complex disease. Hum Mol Genet. 2011; 20:R182–8. [PubMed: 21873261]
10. Franke A, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010; 42:1118–25. [PubMed: 21102463]
11. Kraft P, Hunter DJ. Genetic risk prediction--are we there yet? N Engl J Med. 2009; 360:1701–3. [PubMed: 19369656]
12. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–53. [PubMed: 19812666]
13. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012; 109:1193–8. [PubMed: 22223662]
14. Park JH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010; 42:570–5. [PubMed: 20562874]
15. Park JH, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A. 2011; 108:18026–31. [PubMed: 22003128]
16. Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42:565–9. [PubMed: 20562875]
17. Yang J, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43:519–25. [PubMed: 21552263]
18. Park JH, Dunson DB. Bayesian Generalized Product Partition Model. Statistica Sinica. 2010; 20:1203–1226.
19. Lee SH, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet. 2012; 44:247–250. [PubMed: 22344220]

20. Stahl EA, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet. 2012; 44:483–9. [PubMed: 22446960]

21. Vattikuti S, Guo J, Chow CC. Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. PLoS Genet. 2012; 8:e1002637. [PubMed: 22479213]

22. Purcell SM, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–52. [PubMed: 19571811]

23. Clayton DG. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet. 2009; 5:e1000540. [PubMed: 19584936]

24. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007; 17:1520–8. [PubMed: 17785532]

25. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One. 2008; 3:e3395. [PubMed: 18852893]

26. Janssens AC, et al. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med. 2006; 8:395–400. [PubMed: 16845271]

27. Mihaescu R, Moonesinghe R, Khoury MJ, Janssens AC. Predictive genetic testing for the identification of high-risk groups: a simulation study on the impact of predictive ability. Genome Med. 2011; 3:51. [PubMed: 21797996]

28. Roberts NJ, et al. The Predictive Capacity of Personal Genome Sequencing. Sci Transl Med. 2012; 4:133ra58.

29. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B-Statistical Methodology. 1996; 58:267–288.

30. Yang J, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012; 44:369–75. S1–3. [PubMed: 22426310]

31. Goddard ME, Wray NR, Verbyla K, Visscher PM. Estimating effects and making predictions from genome-wide marker data. Statistical Science. 2009; 24:517–529.

32. Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Annals of Applied Statistics. 2011; 5:1780–1815.

33. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–21. [PubMed: 18691683]

34. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. Stat Med. 2011; 30:1090–104. [PubMed: 21337591]

35. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011; 88:294–305. [PubMed: 21376301]

36. Barrett JC, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009; 41:703–7. [PubMed: 19430480]

37. Voight BF, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010; 42:579–89. [PubMed: 20581827]

38. Eeles RA, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009; 41:1116–21. [PubMed: 19767753]

39. Schunkert H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011; 43:333–8. [PubMed: 21378990]

40. Scheuner MT. Genetic evaluation for coronary artery disease. Genet Med. 2003; 5:269–85. [PubMed: 12865756]

41. Mai PL, Wideroff L, Greene MH, Graubard BI. Prevalence of family history of breast, colorectal, prostate, and lung cancer in a population-based study. Public Health Genomics. 2010; 13:495–503. [PubMed: 20389042]

42. Annis AM, Caulder MS, Cook ML, Duquette D. Family history, diabetes, and other demographic and risk factors among participants of the National Health and Nutrition Examination Survey 1999–2002. Prev Chronic Dis. 2005; 2:A19. [PubMed: 15888230]

43. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010; 6:e1000864. [PubMed: 20195508]

44. So HC, Kwan JS, Cherny SS, Sham PC. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet. 2011; 88:548–65. [PubMed: 21529750]

45. Park JH, Gail MH, Greene MH, Chatterjee N. Potential Usefulness of Single Nucleotide Polymorphisms to Identify Persons at High Cancer Risk: An Evaluation of Seven Common Cancers. J Clin Oncol. 201210.1200/JCO.2011.40.1943

46. Ghosh A, Zou F, Wright FA. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. Am J Hum Genet. 2008; 82:1064–74. [PubMed: 18423522]

47. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology. 2002; 64:583–616.

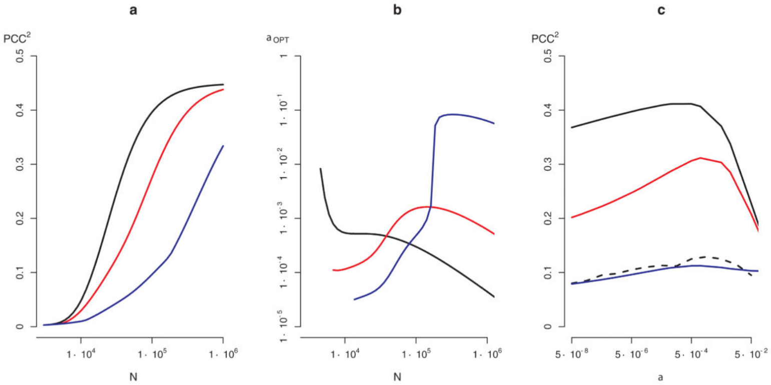

Figure 1. Predictive correlation coefficient (PCC) for polygenic models and corresponding optimal significance level for SNP selection under three models for polygenic architectures for adult height

Each model assumes a total of 45% of phenotypic variance of adult height can be explained by common SNPs included in standard GWAS platforms involving M=200,000 independent SNPs. The effect size distribution for susceptibility SNPs are assumed to follow an exponential distribution (black line), a mixture of two exponential distributions (red line) or a mixture of three exponential distributions (blue line). Panel (a) and (b) show expected value of squared PCC and corresponding optimal significance level (α_{opt}), respectively, as a function of sample size (N). Panel (c) compares PCC values reported in a predictive analysis of the GIANT study (dashed line) with corresponding theoretical expected values under the three different models.

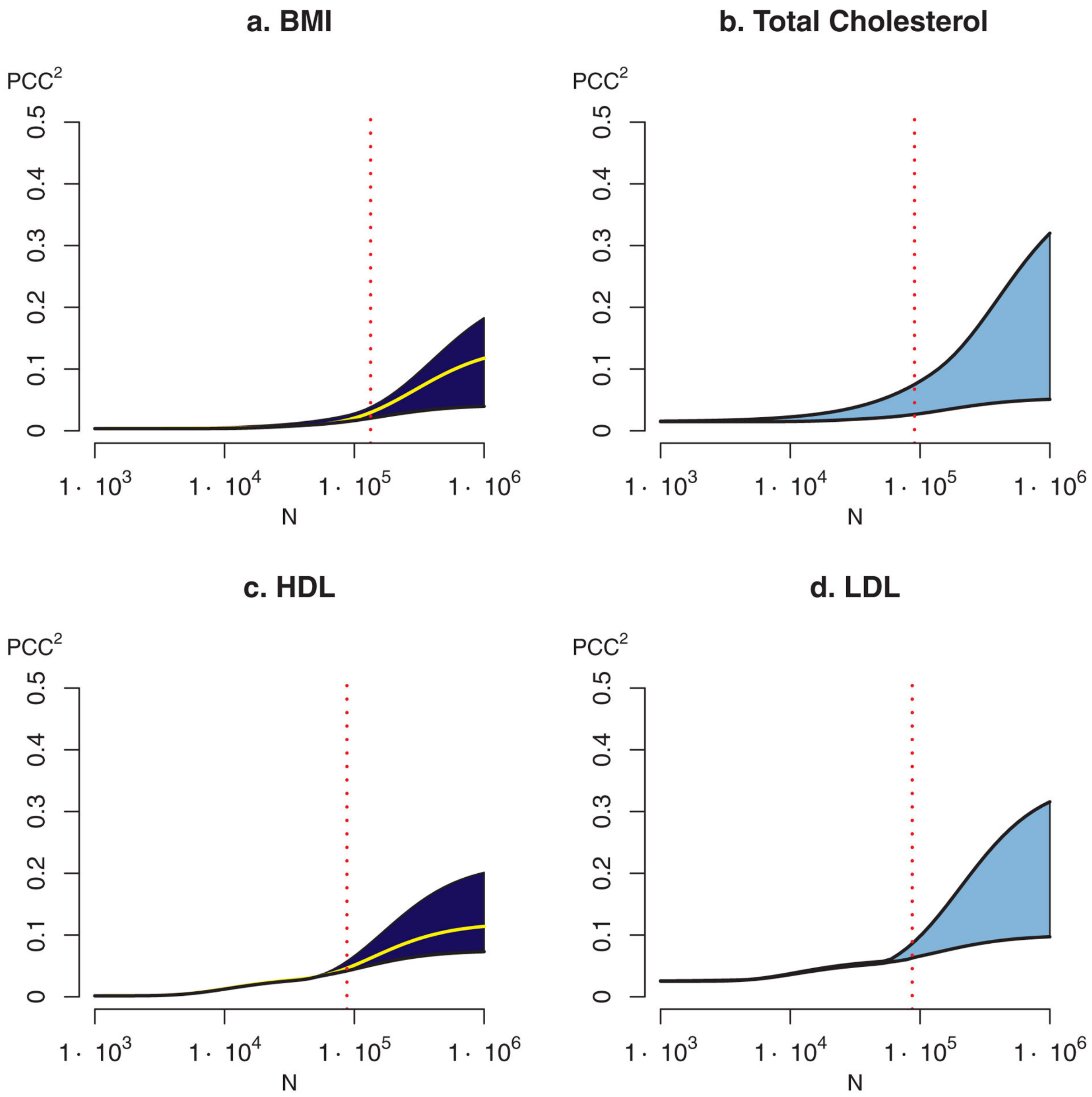

Figure 2.

Expected predictive correlation coefficient (PCC) for polygenic models at optimal significance level for SNP selection for four quantitative traits.

For HDL and BMI, range of performance is shown corresponding to estimate of h_g^2 (yellow line) and associated 95% confidence interval (dark blue region). For LDL and TC, for which direct estimate of h_g^2 is not available, a range of values are chosen based on constraints imposed by the observed discoveries. For all traits, the underlying effect-size distribution is assumed to follow a mixture of three exponential distributions, which together with h_g^2 is calibrated to explain observed discoveries from the largest GWAS (see **Methods**).

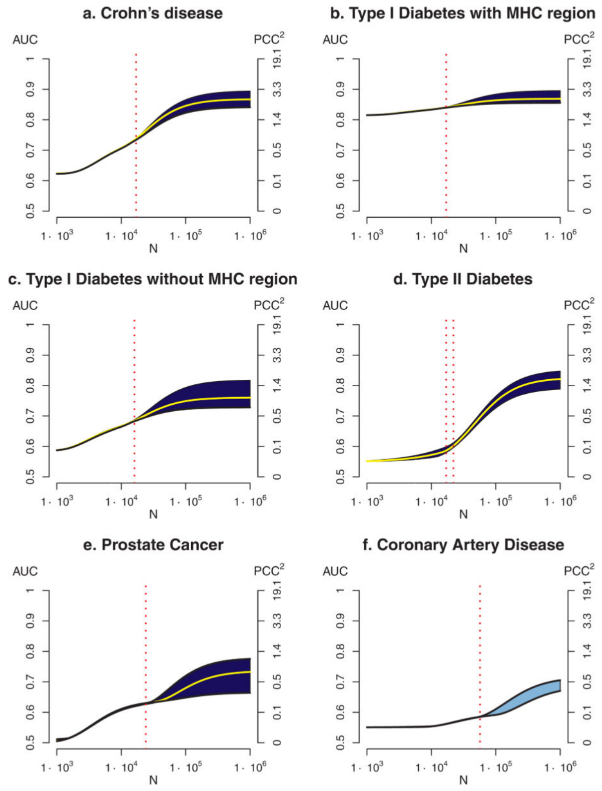

Figure 3. Expected AUC statistics at optimal significance level for SNP selection for five disease traits.

For all diseases except CAD, range of performance is shown corresponding to estimate of h_g^2 (yellow line) and associated 95% confidence intervals (dark blue region). For CAD, for which direct estimate of h_g^2 is not available, a range of its values are chosen based on constraints imposed by the observed discoveries. For all traits, the underlying effect-size distribution is assumed to follow a mixture of two or three exponential distribution, which together with h_g^2 is calibrated to explain observed discoveries from the largest GWAS (see **Methods**).

Table 1

Characteristics of ten complex traits and associated GWAS used in reported analysis.

Trait	HT	BMI	TC	HDL	LDL	CD	T1D	T2D	PrCA	CAD
Narrow sense heritability (h_g^2)	0.45	0.14	-	0.12	-	0.22	0.30	0.51	0.22	-
Effective sample-size for the largest GWAS	133K	162K	100K	100K	95K	25K	22K	36K	28K	73K
No. of detected SNPs	108	31	45	35	36	64	30	22	20	21
Heritability explained by detected SNPs	0.066	0.014	0.063	0.046	0.059	0.066	0.053	0.034	0.061	0.024

HT, height; BMI, body mass index; TC, total cholesterol; CD, Crohn’s disease; T1D, Type 1 diabetes; T2D, Type 2 diabetes; PrCA, prostate cancer; CAD, coronary artery disease.

Estimates of narrow sense heritability (h_g^2), i.e. phenotype variability due total additive effects of common SNPs, for HT, BMI, HDL, CD, T1D and T2D are taken from published studies[20,21,35] and that for PrCA is obtained based on internal analysis of a new NCI GWAS involving approximately 5000 cases and 5000 controls genotyped on Illumina Omni 2.5M platform. For qualitative traits, estimates are shown in the liability-threshold scale.

Table 2

Projected discriminatory performance (AUC statistic) for polygenic risk models including SNPs at genome-wide significance level (α=10^{-7}) and at optimized significance threshold (α_{OPT}). Results for T1D are shown with or without (in parenthesis) contribution of the MHC region. For all diseases except CAD, AUC values are shown corresponding to point estimates of h_g^2 shown in Table 1. For CAD, for which direct estimate of h_g^2 is not available, a range of values are chosen based on constraints imposed by the observed discoveries. For all traits, the underlying effect-size distribution is assumed to follow a mixture of two or three exponential distribution, which together with h_g^2 is appropriately calibrated to explain observed discoveries from the largest GWAS to date.

Trait	AUC with FH alone	Current Sample Size (N)	Model	N		3×N		5×N		10×N	
				α=10^{-7}	α_{OPT}	α=10^{-7}	α_{OPT}	α=10^{-7}	α_{OPT}	α=10^{-7}	α_{OPT}
CD	0.612	17K	SNPs	0.71	0.74	0.77	0.82	0.81	0.84	0.84	0.86
			SNPs+FH	0.79	0.81	0.83	0.87	0.86	0.89	0.89	0.90
T1D	0.533	16K	SNPs	0.84 (0.67)	0.84 (0.69)	0.85 (0.71)	0.86 (0.73)	0.86 (0.73)	0.86 (0.75)	0.86 (0.75)	0.87 (0.75)
			SNPs+FH	0.94 (0.70)	0.94 (0.71)	0.95 (0.74)	0.96 (0.76)	0.96 (0.76)	0.96 (0.77)	0.96 (0.77)	0.96 (0.78)
T2D	0.595	22K	SNPs	0.57	0.60	0.62	0.71	0.67	0.76	0.74	0.79
			SNPs+FH	0.63	0.66	0.67	0.74	0.71	0.78	0.77	0.81
PrCA	0.552	24K	SNPs	0.63	0.63	0.64	0.66	0.66	0.69	0.69	0.71
			SNPs+FH	0.65	0.66	0.66	0.68	0.68	0.71	0.71	0.73
CAD	0.601	57K	SNPs	0.58	0.59	0.59–0.60	0.62–0.64	0.61–0.62	0.64–0.67	0.64–0.66	0.67–0.69
			SNPs+FH	0.65	0.65	0.66	0.67–0.69	0.66–0.68	0.69–0.71	0.68–0.71	0.71–0.73

FH, presence of any family history in first-degree relatives. Prevalences of FH for CAD, PrCA and T2D are 0.14 (ref 40), 0.07 (ref 41), and 0.143 (ref 42), respectively. Prevalence of FH for T1D and CD are taken to be 0.005 and 0.01 which are the same as the disease prevalence[35].

For all diseases, except PrCA the current sample size is shown for the first-stage of the respective largest GWAS. For PrCA, where a large number of SNPs were followed to stage-2, an effective sample size is shown for stage-1 and stage-2 combined.

Table 3

Proportion of cases followed (PCF) among 20% of subjects with highest polygenic risk including SNPs at genome-wide significance level ($\alpha=10^{-7}$) and at optimized significance threshold (α_{OPT}). Results for T1D are shown with or without (in parenthesis) contribution of the MHC region. For all diseases except CAD, AUC values are shown corresponding to point estimates of h_g^2 available from GWAS studies. For CAD, for which direct estimate of h_g^2 is not available, a range of values are chosen based on constraints imposed by observed discoveries. For all traits, the underlying effect-size distribution is assumed to follow a mixture of two or three exponential distribution, which together with h_g^2 is appropriately calibrated to explain observed discoveries from the largest GWAS to date.

Trait	Current Sample Size (N)	Model	N		3×N		5×N		10×N	
			$\alpha=10^{-7}$	α_{OPT}	$\alpha=10^{-7}$	α_{OPT}	$\alpha=10^{-7}$	α_{OPT}	$\alpha=10^{-7}$	α_{OPT}
CD	17K	SNPs	0.48	0.52	0.58	0.65	0.62	0.72	0.72	0.75
		SNPs+FH	0.61	0.65	0.70	0.77	0.75	0.80	0.81	0.83
T1D	16K	SNPs	0.71 (0.42)	0.71 (0.44)	0.73 (0.48)	0.75 (0.51)	0.75 (0.51)	0.76 (0.54)	0.76 (0.54)	0.77 (0.55)
		SNPs+FH	0.91 (0.46)	0.92 (0.48)	0.94 (0.52)	0.95 (0.56)	0.95 (0.56)	0.95 (0.58)	0.95 (0.59)	0.96 (0.60)
T2D	22K	SNPs	0.28	0.32	0.34	0.48	0.41	0.55	0.52	0.63
		SNPs+FH	0.40	0.42	0.43	0.54	0.48	0.60	0.57	0.66
PrCA	24K	SNPs	0.35	0.35	0.37	0.40	0.39	0.44	0.44	0.48
		SNPs+FH	0.40	0.40	0.41	0.44	0.43	0.47	0.47	0.51
CAD	57K	SNPs	0.29	0.30	0.31	0.34–0.37	0.32–0.34	0.38–0.41	0.36–0.40	0.42–0.45
		SNPs+FH	0.42	0.42	0.42–0.43	0.44–0.46	0.43–0.44	0.46–0.49	0.46–0.48	0.49–0.52

FH, presence of any family history in first-degree relatives. Prevalences of FH for CAD, PrCA and T2D are 0.14 (ref 40), 0.07 (ref 41), and 0.143 (ref 42), respectively. Prevalence of FH for T1D and CD are taken to be 0.005 and 0.01 which are the same as the disease prevalence[35].

For all diseases, except PrCA the current sample size is shown for the first-stage of the respective largest GWAS. For PrCA, where a large number of SNPs were followed to stage-2, an effective sample size is shown for stage-1 and stage-2 combined.