nature genetics

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] & Ju-Hyun Park[1,3]

We report a new method to estimate the predictive performance of polygenic models for risk prediction and assess predictive performance for ten complex traits or common diseases. Using estimates of effect-size distribution and heritability derived from current studies, we project that although 45% of the variance of height has been attributed to SNPs, a model trained on one million people may only explain 33.4% of variance of the trait. Models based on current studies allow for identification of 3.0%, 1.1% and 7.0% of the populations at twofold or higher than average risk for type 2 diabetes, coronary artery disease and prostate cancer, respectively. Tripling of sample sizes could elevate these percentages to 18.8%, 6.1% and 12.2%, respectively. The utility of polygenic models for risk prediction will depend on achievable sample sizes for the training data set, the underlying genetic architecture and the inclusion of information on other risk factors, including family history.

© 2013 Nature America, Inc. All rights reserved.

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 prediction of risk 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-wide association studies (GWAS) 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].

Although 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 substantial 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 although about 200 loci identified through a large GWAS involving more than 100,000 subjects can explain only ~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]. There have also been similar studies for several other complex traits[17,19–21].

The gap between estimates of heritability based on known loci and those estimated owing 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 predict a trait value in an independent sample will be the sample size of the training data set based on which the prediction model can be built. Intuitively, as the sample size for the training data set increases, effects of underlying SNPs can be more precisely estimated. Corresponding to this, 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, body mass index (BMI), total cholesterol, high-density lipoprotein (HDL) and low-density lipoprotein (LDL), and the disease traits Crohn's disease, type 1 diabetes (T1D), type 2 diabetes (T2D), coronary artery disease (CAD) and prostate cancer. We used 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, that is, the total heritability of the traits attributable to additive effects of common SNPs.

The results provide several insights into the predictive ability of polygenic models based on existing GWAS, the marginal utility of an increase in sample size, the sample-size threshold beyond which

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Human and Human Services, Rockville, Maryland, USA. [2]Information Management System, Rockville, Maryland, USA. [3]Department of Statistics, Dongguk University–Seoul, Seoul, South Korea. Correspondence should be addressed to N.C. (chattern@mail.nih.gov).

Received 10 May 2012; accepted 8 February 2013; published online 3 March 2013; doi:10.1038/ng.2579

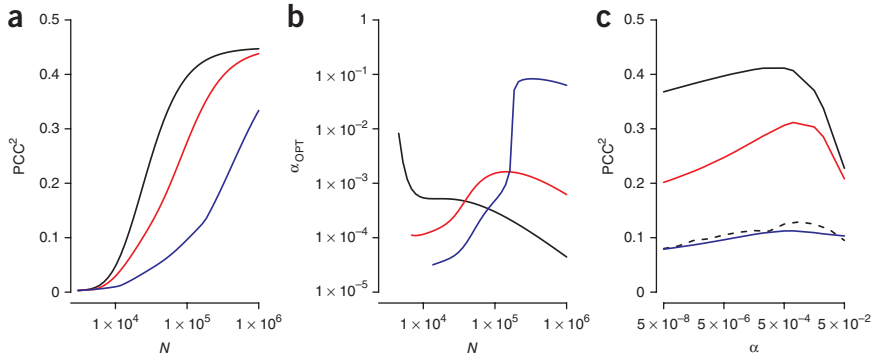


Figure 1 PCC for polygenic models and corresponding optimal significance level for SNP selection under three models for polygenic architectures for adult height. (**a**,**b**) Expected value of PCC^2 (**a**) and corresponding optimal significance level (α_{opt}; **b**) as a function of sample size (N). (**c**) PCC values reported in a predictive analysis of the GIANT study (dashed line) versus corresponding theoretical expected values under the three different models. 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. Effect-size distribution for susceptibility SNPs was 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).

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. The general theoretical framework we provide can be used to make projections for the predictive utility of different polygenic model–building strategies that may use alternate statistical algorithms and/or 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 deriving this formula, we assumed 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. Details of the underlying models and assumptions are available in Online Methods.

The relationship between predictive performance of the model and the sample size (N) for the training data set is shown in equation (1) in Online Methods, which forms the basis of our analytical calculations. Simulation studies confirmed the accuracy of this equation (**Supplementary Fig. 1**). According to this formula, the predictive performance of a model depends on (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_m values) 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 data set (N) influences both the power of the association test statistics and the deviations of the estimated regression coefficients from their true values (Online Methods). Given an effect-size distribution, because the number of underlying susceptibility SNPs (M_1) determine the total variability of the trait explainable by the underlying model, equation (1) can be rewritten 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 owing to additive effects of common tagging SNPs included on current, commercially available SNP microarrays (Equation (2) in Online Methods). In all our subsequent analyses, we assume that genotyping platforms based on which most current GWAS have been conducted to contain approximately on average $M = 200{,}000$ independent SNPs.

To model a complex trait, we first investigated the predictive performance of polygenic models for adult height. In **Figure 1** we show that the predictive accuracy of polygenic models greatly depends on the distribution of effect sizes even when all distributions result in a total heritability of 45% (ref. 16). Predictive performance of the model for all sample sizes was the highest when an exponential distribution underlies the effect sizes. Predictive performance of the model decreased substantially under a two-component, exponential-mixture model, which, compared to the exponential model, provided a much better fit to the observed effect sizes of the known SNPs by allowing for the presence of more SNPs, each with smaller effect (**Supplementary Table 1**). Finally, the performance of the model was 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 reproduced results from a predictive analysis reported in the GIANT study in which distinct polygenic models had been built with different significance thresholds for SNP selection, and their predictive performance was empirically

Table 1 Characteristics of ten complex traits and associated GWAS used in reported analysis

Trait	Height	BMI	TC	HDL	LDL	CD	T1D	T2D	PrCA	CAD
h_g^2	0.45	0.14	–	0.12	–	0.22	0.30	0.51	0.22	–
Effective sample-size for the largest GWAS	133,000	162,000	100,000	100,000	95,000	25,000	22,000	36,000	28,000	73,000
Number 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

TC, total cholesterol; CD, Crohn's disease; PrCA, prostate cancer. Estimates of h_g^2, that is, phenotype variability owing to total additive effects of common SNPs, for height, BMI, HDL, CD, T1D and T2D are from published studies[20,21,35] and h_g^2 for PrCA is based on internal analysis of a new GWAS at the National Cancer Institute involving ~5,000 cases and 5,000 controls genotyped on Illumina Omni 2.5M platform. For qualitative traits, estimates of h_g^2 are shown in the liability-threshold scale. Characteristics of largest GWAS and associated discoveries were obtained from published reports[6–8,10,36–39]. For each trait, an effect-size sample size was calculated for a single-stage study that has equivalent power as the original study, taking into account multistage genotyping and selective sampling by family history for PrCA. For height, sample size and reported discoveries correspond to only first stage of the GIANT study. The number of discoveries reported accounts for any genomic control adjustment used in the original study.

© 2013 Nature America, Inc. All rights reserved.

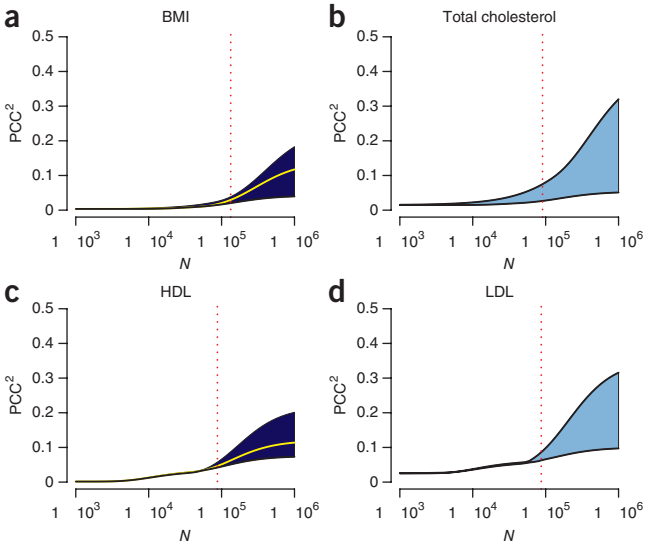
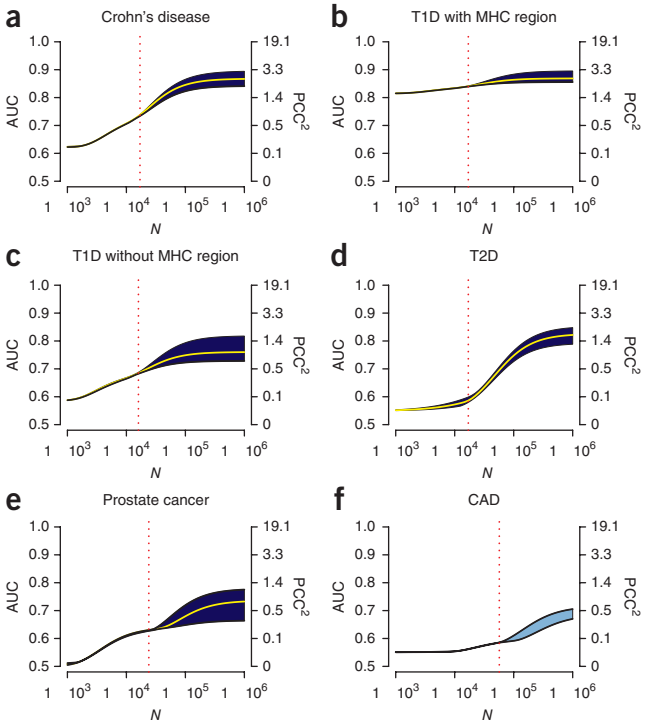


Figure 2 Expected PCC for polygenic models at optimal significance level for SNP selection for four quantitative traits. (**a**–**d**) For HDL and BMI, range of performance is shown corresponding to estimate of h_g^2 (yellow line) and associated 95% confidence interval (dark blue region). For LDL and total cholesterol, for which direct estimate of h_g^2 was not available, a range of values were chosen based on constraints imposed by the observed discoveries. For all traits, the underlying effect-size distribution was assumed to follow a mixture of three exponential distributions, which together with h_g^2 was calibrated to explain observed discoveries from the largest GWAS (Online Methods). Vertical dotted line corresponds to the sample size for the current largest genome-wide scans.

assessed using independently held out datasets. 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 (**Fig. 1**).

Equation (1) in Online Methods illustrates the tradeoff between specificity and sensitivity of the SNP selection criterion on the predictive performance of the model. With a more liberal significance threshold (α), the PCC value will increase through the power of the association tests but will decrease as a function of the underlying type I error (α). In **Figure 1** we illustrate 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 increased with an increase in sample size, then it plateaued and subsequently remained constant or decreased 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, total cholesterol, HDL and LDL) and qualitative phenotypes (Crohn's disease, T1D, T2D, CAD and prostate cancer) 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 h_g^2 (Online Methods and **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 (**Figs. 2** and **3**). Although it may be possible to improve the performance of these models including 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 will increase, the projected performance 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, for example, although a polygenic model built based on the current largest GWAS can be expected to achieve an AUC statistic of about 63%, in the future, a model built based on as many as three times that sample size is expected to yield an AUC statistic of only 64–70% (**Fig. 3**). For all disease traits except CAD, it appears that the marginal utility of additional samples can be quite small after the size of GWAS reaches 100,000–200,000 subjects. In contrast, for CAD, BMI, and the lipid traits total cholesterol and LDL, the performance of predictive models may continue to improve gradually over a much wider range of sample sizes, as high as 500,000 to one million subjects.

Figure 3 Expected AUC statistics at optimal significance level for SNP selection for five disease traits. (**a**–**f**) For all diseases except CAD, range of performance is shown corresponding to the 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 was not available, a range of its values were chosen based on constraints imposed by the observed discoveries. For all traits, the underlying effect-size distribution was assumed to follow a mixture of two- or three-exponential distributions, which together with h_g^2 was calibrated to explain observed discoveries from the largest GWAS (Online Methods). Vertical dotted line corresponds to the sample size for the current largest genome-wide scans.

© 2013 Nature America, Inc. All rights reserved.

Table 2 Projected discriminatory performance (AUC statistic) for polygenic risk models

Trait	AUC with FH alone	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	0.612	17,000	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	16,000	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	22,000	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	24,000	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	57,000	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

Results are shown for models including SNPs at genome-wide significance level ($\alpha = 10^{-7}$) and at optimized significance threshold (α_{opt}). FH, presence of any family history in first-degree relatives. Prevalences of FH for CAD, prostate cancer (PrCA) and T2D are 0.14 (ref. 40), 0.07 (ref. 41) and 0.143 (ref. 42), respectively. Prevalence of FH for T1D and Crohn's disease (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 stages 1 and 2 combined. Results for T1D are shown with or without (in parentheses) contribution of the MHC region. For all diseases except CAD, AUC values are shown corresponding to point estimates of h_g^2 in **Table 1**. For CAD, for which direct estimate of h_g^2 was not available, a range of values were chosen based on constraints imposed by the observed discoveries. For all traits, the underlying effect-size distribution was assumed to follow a mixture of two- or three-exponential distributions, which together with h_g^2 was appropriately calibrated to explain observed discoveries from the largest GWAS to date.

Predictive performance of a model strongly depends on the extent of heritability of the trait. For any given sample size, more accurate prediction is possible for more heritable traits, such as Crohn's disease and T1D, than for less heritable traits such as CAD, prostate cancer 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 individuals likely to develop the disease among high-risk groups varies (**Table 3**). For example, using models based on current GWAS, the proportion of future cases that could be identified among top 20% 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 proportion would be expected to increase to 75% and 48%, respectively. For the three common chronic diseases, the proportion of the population that can be identified to have twofold or higher risk than an average person ranged from 1.1% (CAD) to 7.0% (prostate cancer) for models built based on current sample sizes (**Supplementary Table 4**). If the sample size in future studies could be tripled, then these proportions could be 6.1% (CAD) and 18.8% (T2D).

For all diseases, family history information alone provides low discriminatory ability. However, models that include both family history and polygenic scores can perform substantially better than models that use polygenic scores alone, especially for rare, highly familial conditions such as Crohn's disease and T1D. Even if polygenic scores could be built in the future based on very large sample sizes (for example, sample size = 5*N*), family history is expected to remain an important variable for identifying high-risk subjects (**Tables 2** and **3**).

DISCUSSION

Our analysis demonstrated that the predictive ability of polygenic models depends not only on total heritability but also on the underlying effect-size distributions. Effect-size distributions from large GWAS suggest that although risk prediction models will continue to improve as total sample size increases, the improvement will be slow and modest even when common SNPs account for a large proportion of heritability of the underlying traits. Our analysis also shows that under the most likely effect-size distributions, the optimal significance threshold for selecting SNPs for prediction models

Table 3 Proportion of cases followed among 20% of subjects with highest polygenic risk

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	17,000	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	16,000	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	22,000	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	24,000	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	57,000	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

Results are shown for models including SNPs at genome-wide significance level ($\alpha = 10^{-7}$) and at optimized significance threshold (α_{opt}). FH, presence of any family history in first-degree relatives. Prevalences of FH for CAD, prostate cancer (PrCA) and T2D are 0.14 (ref. 40), 0.07 (ref. 41), and 0.143 (ref. 42), respectively. Prevalence of FH for T1D and Crohn's disease (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 stages 1 and 2 combined. Results for T1D are shown with or without (in parentheses) 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 was not available, a range of values were chosen based on constraints imposed by observed discoveries. For all traits, the underlying effect-size distribution was assumed to follow a mixture of two- or three-exponential distributions, which together with h_g^2 was appropriately calibrated to explain observed discoveries from the largest GWAS to date.

© 2013 Nature America, Inc. All rights reserved.

in large GWAS can be more liberal than threshold standard (for example, $P < 5 \times 10^{-8}$) used for discovery.

We observed that for less common, highly familial conditions, such as T1D and Crohn's disease, risk models that include family history and optimal polygenic scores based on current GWAS can identify a large majority of cases by targeting a small group of high-risk individuals (for example, subjects who fall in the highest quintile of risk). In contrast, for more common conditions with modest familial components, such as T2D, CAD and prostate cancer, risk models based on GWAS with current sample sizes (N) or foreseeable sample sizes in the near future (for example, $3N$) 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 data 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 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 major histocompatibility complex (MHC) region highlights the limited incremental discriminatory ability of polygenic scores for diseases that have established common and strong risk factors. 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 enhance the predictive utility of these models, especially for applications in high-risk families.

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 data set and h_g^2. More general theory shows that an algorithm that includes all SNPs in a model, that is, uses the significance level of $\alpha = 1$, 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 used 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 realistically depict 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 sizes to approach the ideal predictive power associated with additive effects of common SNPs. Using a metric used in this report together with the assumption of independent susceptibility alleles across traits, for example, we predict that although GWAS in principle can identify 55.1% of the population that might have twofold or higher risk than average for at least one of the three common diseases, CAD, T2D and prostate cancer, the actual proportion achievable using current GWAS data is only 10.7% and that tripling the sample size could increase this to 33.1%. If the susceptibility alleles across these traits are related, however, these proportions could be higher.

Here we 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 uses least absolute shrinkage and selection operator (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 such as height, but we also note that the gains are generally modest in scope (**Supplementary Fig. 2**). Simultaneous modeling of correlated SNPs in 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 using 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 to improve 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 Fig. 3**), under a model that allows large number of susceptibility loci each containing sets of low-penetrance 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 data set, models based on collapsed variables often can perform better, apparently because of 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; 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.

Here we used a flexible class of mixture-exponential models to specify effect-size distributions. One could specify effect-size distributions using alternate 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 models were restricted so that they can also explain discoveries and estimates of heritabilities reported from current GWAS, each produced results that

© 2013 Nature America, Inc. All rights reserved.

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 Fig. 3**).

In conclusion, we used a newly developed model together with empirical observations from large GWAS to comprehensively evaluate future polygenic risk models using common susceptibility SNPs. Although our analysis points to 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 have modest discriminatory power and face additional challenges for long-term risk prediction as risk-factor history, unlike susceptibility status, can change over the 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 to develop such models.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This research was supported by the intramural program of the US National Cancer Institute.

AUTHOR CONTRIBUTIONS

N.C. led the development of the statistical methods and drafted the manuscript. J.-H.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. N.C., B.W., J.S., P.H., S.J.C. and J.-H.P. reviewed and revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bowles Biesecker, B. & Marteau, T.M. The future of genetic counselling: an international perspective. *Nat. Genet.* **22**, 133–137 (1999).
2. Pharoah, P.D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**, 33–36 (2002).
3. van Hoek, M. *et al.* Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes* **57**, 3122–3128 (2008).
4. Pharoah, P.D., Antoniou, A.C., Easton, D.F. & Ponder, B.A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**, 2796–2803 (2008).
5. Wacholder, S. *et al.* Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362**, 986–993 (2010).
6. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
7. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
8. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
9. Jostins, L. & Barrett, J.C. Genetic risk prediction in complex disease. *Hum. Mol. Genet.* **20**, R182–R188 (2011).
10. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
11. Kraft, P. & Hunter, D.J. Genetic risk prediction–are we there yet? *N. Engl. J. Med.* **360**, 1701–1703 (2009).
12. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
13. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA* **109**, 1193–1198 (2012).
14. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
15. Park, J.H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. USA* **108**, 18026–18031 (2011).
16. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
17. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
18. Park, J.H. & Dunson, D.B. Bayesian generalized product partition model. *Statist. Sinica* **20**, 1203–1226 (2010).
19. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
20. Stahl, E.A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
21. Vattikuti, S., Guo, J. & Chow, C.C. Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. *PLoS Genet.* **8**, e1002637 (2012).
22. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
23. Clayton, D.G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet.* **5**, e1000540 (2009).
24. Wray, N.R., Goddard, M.E. & Visscher, P.M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
25. Daetwyler, H.D., Villanueva, B. & Woolliams, J.A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, e3395 (2008).
26. Janssens, A.C. *et al.* Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet. Med.* **8**, 395–400 (2006).
27. Mihaescu, R., Moonesinghe, R., Khoury, M.J. & Janssens, A.C. Predictive genetic testing for the identification of high-risk groups: a simulation study on the impact of predictive ability. *Genome Med.* **3**, 51 (2011).
28. Roberts,, N.J. *et al.* The predictive capacity of personal genome sequencing. *Sci Transl. Med.* **4**, 133ra58 (2012).
29. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
30. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
31. Goddard, M.E., Wray, N.R., Verbyla, K. & Visscher, P.M. Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* **24**, 517–529 (2009).
32. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5**, 1780–1815 (2011).
33. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
34. Gail, M.H. Personalized estimates of breast cancer risk in clinical practice and public health. *Stat. Med.* **30**, 1090–1104 (2011).
35. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
36. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
37. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
38. Eeles, R.A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41**, 1116–1121 (2009).
39. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
40. Scheuner, M.T. Genetic evaluation for coronary artery disease. *Genet. Med.* **5**, 269–285 (2003).
41. Mai, P.L., Wideroff, L., Greene, M.H. & Graubard, B.I. Prevalence of family history of breast, colorectal, prostate, and lung cancer in a population-based study. *Public Health Genomics* **13**, 495–503 (2010).
42. Annis, A.M., Caulder, M.S., Cook, M.L. & 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.* **2**, A19 (2005).

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

Underlying polygenic model. We assume Y is the outcome variable 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 ($\mathcal{M}$)

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

where M_1 out of 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 ($\widehat{\mathcal{M}}$) is built based on a 'training' data set 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 ($\gamma_m = 1$) or not ($\gamma_m = 0$) 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 define the predictive correlation for the model $\widehat{\mathcal{M}}$ to be

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

where the subscript X and ε signify that the correlation coefficient is computed with respect to the distribution of X and ε in the underlying population for which prediction is desired while the estimated model $\widehat{\mathcal{M}}$ and its associated parameter estimates ($\hat{\beta}_m$ and γ_m, $m = 1, 2, \ldots, M$) are held fixed. The only source of variation of $R_N(\widehat{\mathcal{M}})$ is due to the randomness of the original training data set based on which $\widehat{\mathcal{M}}$ is built. For any fixed N and λ, the expected value of $R_N(\widehat{\mathcal{M}})$ 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 mth 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, that is, $\gamma_m = 1$, its corresponding coefficient in the prediction model could be taken as $\hat{\beta}_m$, that is, the estimated regression coefficient from the marginal analysis of the SNP.

Based on general theory developed in the **Supplementary Note**, we show that in the above setting the expected value of the PCC of the above polygenic model–building algorithm over different GWAS data sets of sample size N can be written as

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

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,

$$e_N(\beta_m) = E\left(\hat{\beta}_m \,\middle|\, |Z_m| > C_{\frac{\alpha}{2}}\right)$$

and

$$\nu_N(\beta_m) = E\left(\hat{\beta}_m^2 \,\middle|\, |Z_m| > C_{\frac{\alpha}{2}}\right).$$

Based on the formula for $e_N(\beta_m)$ and $\nu_N(\beta_m)$ given in the **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 $N \to \infty$,

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

Because $\sum_{m=1}^{M_1} \beta_m^2$ is the variance of the trait owing 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 discriminatory 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

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

We assume that a risk-prediction model is built based on a training data set 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 an indicator of whether the variable is selected (γ_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

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

where

$$S_N^2 = Var(\hat{U} \,|\, D = 0) = \sum_{m=1}^{M} \hat{\beta}_m^2 \gamma_m$$

and

$$C_N = Cov(\hat{U}, U \,|\, D = 0) = \sum_{m=1}^{M_1} \beta_m \hat{\beta}_m \gamma_m$$

© 2013 Nature America, Inc. All rights reserved.

It is noteworthy that although 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 $\widehat{U}$ requires two parameters, namely the variance of $\widehat{U}$ and its covariance with the true risk U.

The AUC, that is, 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

$$\mathrm{AUC}_N = \mathrm{pr}(\widehat{U}_1 > \widehat{U}_0) = \Phi(\sqrt{0.5}R_N)$$

where $R_N = \frac{C_N}{S_N}$ is the predictive correlation measure defined earlier for continuous outcome. Similarly, using above results, other measures of discriminatory performance of models, such as proportion of cases followed (PCF)[2], can be also characterized in terms of R_N (**Supplementary Note**).

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

$$\mathrm{pr}(\widehat{U} \mid D=0, D_R=0) \sim N(0, S_N^2)$$

$$\mathrm{pr}(\widehat{U} \mid D=0, D_R=1) \sim N(k_R C_N, S_N^2)$$

$$\mathrm{pr}(\widehat{U} \mid D=1, D_R=0) \sim N(C_N, S_N^2) \text{ and}$$

$$\mathrm{pr}(\widehat{U} \mid D=1, D_R=1) \sim N\left((1+k_R)C_N, S_N^2\right)$$

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

Estimation of effect-size distribution. We extended 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: (i) identify the largest GWAS, termed the 'current study', for each of the traits and list 'observed susceptibility SNPs' that are discovered through these studies; (ii) following the design of the discovery studies (**Supplementary Table 2**), compute the power to detect SNPs with given effect sizes; (iii) 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 (iv) 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 owing 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 (i), 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 (ii), we evaluated power for detection for each susceptibility SNP at their observed effect sizes following the exact design of the original discovery studies (**Supplementary Table 2**).

In step (iii), 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 nonparametric methods for estimating effect-size distribution in the range of effect sizes for observed susceptibility SNPs. In this report, we considered the 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 is given by

$$f(es \mid \theta) = \sum_{h=1}^{H} p_h g(es \mid \lambda_h)$$

where $\theta = (p_1, \ldots, p_H, \lambda_1, \ldots, \lambda_H)$, with p_h being the mixture weight for the h^{th} component, $h = 1, \ldots, H$ 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 proportional to its power for discovery, we constructed a likelihood as

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

where $pow_{\text{study}}(es_i|N,\alpha)$ 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 used Bayesian methods to estimate the parameters of the mixture model based on the above likelihood and non-informative priors for the parameter vectors $\mathbf{p} = (p_1, \ldots, p_H)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_H)$. Specifically, we assumed a discrete Dirichlet distribution for $\mathbf{p}$ that leads to uniform prior for each of the p_h, $h = 1, \ldots, H$ marginally. We assumed λ_h, $h = 1, \ldots, H$ to be independently distributed each following a gamma distribution with shape and scale parameters $a = 0.5$ and $b = 2 \times 10^4$, respectively. Posterior means for all parameters were obtained based on Markov chain Monte Carlo 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 deviance information criterion[47]. For all traits except prostate cancer (PrCA) and CAD, a two-component ($H = 2$) mixture model was the best fitted distribution. For PrCA and CAD, a single exponential distribution ($H = 1$) was adequate.

In step (iv), we incorporated an additional mixture component to the effect-size distribution estimated in step (iii) so that the overall distribution can be used to describe the effect sizes for all SNPs that contribute to $h_g^2 = \sum_{m=1}^{M_1} \beta_m^2$. 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

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2579

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

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

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

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

and

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

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

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

We solved for p_{H+1} and λ_{H+1} by performing a grid-search within the ranges $0.01 \le p_{H+1} \le 0.99$ and $\hat{\lambda}_H \le \lambda_{H+1} \le 20 \times \hat{\lambda}_H$, 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 confidence intervals (**Figs. 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 total cholesterol (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.

43. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
44. So, H.C., Kwan, J.S., Cherny, S.S. & Sham, P.C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.* **88**, 548–565 (2011).
45. Park, J.H., Gail, M.H., Greene, M.H. & Chatterjee, N. Potential usefulness of single nucleotide polymorphisms to identify persons at high cancer risk: an evaluation of seven common cancers. *J. Clin. Oncol.* **30**, 2157–2162 (2012).
46. Ghosh, A., Zou, F. & Wright, F.A. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am. J. Hum. Genet.* **82**, 1064–1074 (2008).
47. Spiegelhalter, D.J., Best, N.G., Carlin, B.R. & van der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 583–616 (2002).

© 2013 Nature America, Inc. All rights reserved.