Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies

Ronald de Vlaming[1,2], Aysu Okbay[1,2], Cornelius A. Rietveld[1,2], Magnus Johannesson[3], Patrik K.E. Magnusson[4], André G. Uitterlinden[1,5], Frank J.A. van Rooij[5], Albert Hofman[1,5,6], Patrick J.F. Groenen[1,7], A. Roy Thurik[1,2,8], and Philipp D. Koellinger[1,2,9]*

1 Erasmus University Rotterdam Institute for Behavior and Biology, Erasmus School of Economics, Rotterdam, the Netherlands.

2 Department of Applied Economics, Erasmus School of Economics, Rotterdam, the Netherlands.

3 Department of Economics, Stockholm School of Economics, Stockholm, Sweden.

4 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

5 Department of Epidemiology, Erasmus University Medical Center Rotterdam, Rotterdam, the Netherlands.

6 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, the United States of America.

7 Econometric Institute, Erasmus School of Economics, Rotterdam, the Netherlands.

8 Montpellier Business School, Montpellier, France.

9 Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam, the Netherlands.

* E-mail: koellinger@ese.eur.nl

Abstract

Large-scale GWAS results are typically obtained by meta-analyzing GWAS results from multiple studies spanning different regions and/or time periods. This approach averages the estimated effects of individual genetic variants across studies. In case genetic effects are heterogeneous across studies, the statistical power of a GWAS and the predictive accuracy of polygenic scores are attenuated, contributing to the so-called 'missing' heritability. However, a theoretical multi-study framework, relating statistical power and predictive accuracy to cross-study heterogeneity, is not available. We address this gap by developing an online Meta-GWAS Accuracy and Power calculator that accounts for the cross-study genetic correlation. This calculator enables to explore to what extent an imperfect cross-study genetic correlation (i.e., less than one) contributes to the missing heritability. By means of simulation studies, we show that under a wide range of genetic architectures, the statistical power and predictive accuracy inferred by this calculator are accurate. We use the calculator to assess recent GWAS efforts and show that the effect of cross-study genetic correlation on statistical power and predictive accuracy is substantial. Hence, cross-study genetic correlation explains a considerable part of the missing heritability. Therefore, *a priori* calculations of statistical power and predictive accuracy, accounting for heterogeneity in genetic effects across studies, are an important tool for adequately inferring whether an intended meta-analysis of GWAS results is likely to yield meaningful outcomes.

Author Summary

Large-scale genome-wide association studies are uncovering the genetic architecture of traits which are affected by many genetic variants. Such studies typically meta-analyze association results from multiple studies spanning different regions and/or time periods. These efforts do not yet capture a large share of the total proportion of trait variation attributable to genetic variation. The origins of this so-called ‘missing’ heritability have been strongly debated. One factor exacerbating the missing heritability is heterogeneity in the effects of genetic variants across studies. Its influence on statistical power to detect associated genetic variants and the accuracy of polygenic predictions, is poorly understood. In the current study, we derive the precise effects of heterogeneity in genetic effects across studies on both the statistical power to detect associated genetic variants as well as the accuracy of polygenic predictions. We provide an online calculator, available at `www.devlaming.eu`, which accounts for these effects. By means of this calculator, we show that imperfect genetic correlations between studies substantially decrease statistical power and predictive accuracy and, thereby, contribute to the missing heritability. We argue that researchers should account for genetic heterogeneity across studies, when assessing whether a proposed large-scale genome-wide association study is likely to yield meaningful results.

Introduction

Large-scale GWAS efforts are rapidly elucidating the genetic architecture of polygenic traits, including anthropometrics [1], [2], diseases [3], [4], [5], and behavioral and psychological outcomes [6], [7], [8]. These efforts have led to new biological insights, therapeutic targets, and individual-level polygenic scores (PGS), and help to understand the complex interplay between genes and environments in shaping individual outcomes [7], [9], [10]. However, GWAS results for polygenic traits do not yet account for a large part of the heritability [1], [2], [7], [8]. This dissonance, which is referred to as the 'missing' heritability, has received broad attention [11], [12], [13], [14], [15], [16], [17].

The missing heritability can be split into two parts. The first part, the 'still-missing' heritability [15], [16], [17], is defined as the difference between the estimate of heritability based on family data (h^2) and the SNP-based estimate (${h^2}_{\text{SNP}}$). The second part, the 'hiding' heritability [15], [16], [17], is defined as the difference between the ${h^2}_{\text{SNP}}$ and the estimate of heritability based on genetic variants that reach genome-wide significance in a GWAS (${h^2}_{\text{GWS}}$). Hence, $h^2 > {h^2}_{\text{SNP}} > {h^2}_{\text{GWS}}$ [15].

Four important factors have been proposed to explain the missing heritability. First, conventional genotyping is not sufficiently dense across the whole genome. Therefore, genotyping fails to capture rare variants that explain a non-negligible fraction of trait variation [18]. Second, gene–gene interactions inflate h^2, creating so-called 'phantom' heritability [14]. Third, sample sizes of GWAS efforts are not large enough to fully capture ${h^2}_{\text{SNP}}$ [18], [19]. Fourth, differences across strata (e.g., studies, ancestry groups, and sexes) in phenotypic measurement, in measurement accuracy, and in genetic effects, can all introduce additional noise and loss of signal [20], [21] and, hence, attenuate statistical power of a GWAS [17], [22], [23]. The first two factors lead primarily to still-missing heritability [14], [17], [18], while the latter two contribute foremost to hiding heritability [17], [18], [19], [23].

Recent work has demonstrated the feasibility of denser genotyping [24], [25] and larger GWAS samples [1], [2], [5], [26]. Hence, these two causes of the missing heritability can be amended. Moreover, the issue of phantom heritability is primarily of importance to the discussion about the still-missing heritability [14], [17]. In the current study, we focus on one important remaining factor in the hiding heritability discussion: heterogeneity of measures and/or heterogeneity of genetic effects across different strata, and, in particular, across studies.

Large-scale GWAS results are typically obtained by meta-analyzing GWAS results from multiple studies spanning different regions and/or time periods. This approach averages the estimated effects of individual genetic variants across studies. In case genetic effects are heterogeneous across studies (e.g., due to gene–environment interactions and heterogeneity in phenotypic measurement) at least three important quantities are decreased: the estimate of SNP-heritability [17], [20], the statistical power of a GWAS [17], [22], [23], and the predictive accuracy of the PGS [26]. The decrease in these quantities not only explains why heterogeneity contributes to the missing heritability but also shows that heterogeneity decreases the chances of a study to yield meaningful results [23], [27]. Therefore, the precise attenuation due to genetic heterogeneity should be well understood, in order to make an informed decision whether to pursue a proposed meta-analysis of GWAS results.

Others have already pointed at the issue of genetic heterogeneity across studies [21], [23], [28]. In particular, it has been shown theoretically that misclassification between two diseases tends to deflate heritability estimates and decrease statistical power to detect trait-associated SNPs [20]. In addition, empirical applications show that SNP-heritability estimates are attenuated when pooling across studies [21], [29]. Moreover, simulations have shown that phenotypic and genetic heterogeneity decrease statistical power [22]. Finally, a strong theoretical decrease in statistical power has been shown to exist under genetic heterogeneity of another sort, viz., when different intermediate phenotypes contribute to a single composite phenotype [17]. Finally, a theoretical reduction of PGS predictive accuracy has been shown for a scenario with one discovery study and one study used as hold-out sample for prediction [26]. Overall, findings from simulations, empirical work, and theory suggest attenuation due to genetic and phenotypic heterogeneity. Despite these efforts, a theoretical multi-study framework, relating statistical power and predictive accuracy to cross-study heterogeneity, is still absent.

In the current study, we address the absence of a general multi-study framework by developing a Meta-GWAS Accuracy and Power (MetaGAP) calculator that accounts for the cross-study genetic correlation (CGR). Moreover, by means of simulation studies, we show that under a wide range of genetic architectures, the statistical power and predictive accuracy inferred by this calculator are accurate. The calculator requires users to specify the number of studies, the sample size of each study, the SNP-based heritability per study, and the CGR. From these input parameters, the calculator infers the statistical power to detect associated SNPs and the predictive accuracy of the PGS in a meta-analysis of GWAS results from genetically and phenotypically heterogeneous studies. The MetaGAP calculator enables to explore to what extent an imperfect CGR (i.e., less than one) contributes to the hiding heritability.

As an empirical application of the proposed calculator, we estimate the SNP-based heritability and CGR of several polygenic traits across three distinct studies: the Rotterdam Study (RS), the Swedish Twin Registry (STR), and the Health and Retirement Study (HRS). For height, BMI, years of education, and self-rated health, we obtain point-estimates of CGR between 0.47 and 0.97, suggesting that even extremely large GWAS meta-analyses will fall short of explaining the full ${h^2}_{\text{SNP}}$ for these traits. Using the MetaGAP calculator, we quantify the expected number of hits and the predictive accuracy of the PGS in recent GWAS efforts for these traits. Our theoretical predictions align with empirical observations. Finally, by comparing these figures to the predicted number of hits and PGS accuracy under perfect CGRs, we show that there is considerable attenuation due to imperfect CGRs; even for height (CGR point-estimate of 0.97) the expected relative loss in the number of hits is 8% and the relative loss in PGS R^2 is 6%.

Importantly, the MetaGAP calculator has two desirable properties compared to other calculators. In other calculators one often needs to specify some true value of the SNP effect [30] (e.g., by taking the effect estimates of the most significant SNPs from an earlier GWAS, to which one first applies a 'winner's curse' correction [23]). Instead of requiring the input of an *a priori* unknown effect, our method incorporates a tacit assumption regarding the relation between allele frequency and effect size, such that each trait-affecting SNP has an equal R^2 with respect to the phenotype (e.g., [31]). Therefore, our method merely requires the $h^2{}_{\mathrm{SNP}}$ and the number of independent haplotype blocks harboring trait-affecting variation. The ratio of these two quantities fully specifies the proportion of phenotypic variance which can be explained by a 'representative' associated SNP. In addition, other calculators usually require not only the true effect of a SNP as input parameter but also the allele frequency [30], [32], [33]. By focusing on a 'representative' associated SNP, we also eliminate the allele frequency from our power calculator. In our simulations, we show that a violation of this equal-R^2 assumption hardly affects the quality of the predicted statistical power and PGS accuracy.

To summarize, the current study aims to formulate precise relations between genetic heterogeneity across studies on the one hand and statistical power and predictive accuracy, for a meta-analysis of GWAS results, on the other. This aim is achieved, and substantiated in the form of an online calculator, available at `www.devlaming.eu`, which accounts for the effect of genetic and phenotypic heterogeneity across studies. The calculator does not require *a priori* knowledge about the magnitude of the true association between the SNP and trait of interest. By means of this calculator, it can be shown to what degree CGR affects statistical power and predictive accuracy. By using the calculator to assess recent GWAS efforts, we show that the effect of CGR on statistical power and predictive accuracy is substantial. Hence, CGR explains a considerable part of the hiding heritability. Therefore, *a priori* calculations of statistical power and predictive accuracy, accounting for heterogeneity in genetic effects across studies, are an important tool for assessing whether an intended meta-analysis of GWAS results is likely to yield meaningful outcomes.

Materials and Methods

Definitions In our framework, we consider only the SNP-based heritability, as estimated based on the set of SNPs of interest. In line with others, we define the effective number of SNPs, S, as the number of haplotype blocks (i.e., independent chromosome segments) [34], where variation in each block is tagged by precisely one SNP. Hence, in our framework, there are S SNPs contributing to the polygenic score. Due to linkage disequilibrium this number is likely to be substantially lower than the total number of SNPs in the genome [35], and is inferred to lie between as little as 60,000 [15] and as much as 5 million [35]. In terms of trait-affecting variants, we consider a subset of M SNPs. Each SNP in this subset tags variation in a segment that bears a causal influence on the phenotype. We refer to M as the associated number of SNPs. We assume that the M associated SNPs capture the full SNP-based heritability for the trait of interest.

Power of a GWAS meta-analysis under heterogeneity

Generic expressions for the theoretical distribution of the Z statistic, resulting from a meta-analysis of GWAS results under imperfect CGRs, can be found in S1 Derivations Power. For intuition, we here present the specific case of a meta-analysis of results from two studies with CGR $\rho_{\mathbf{G}}$, with equal SNP-based heritability h^2_{SNP}, and equal sample sizes (i.e., N in Study 1 and N in Study 2). Under this scenario, we find that under high polygenicity, the Z statistic of an associated SNP k is normally distributed with mean zero and the following variance:

$$\operatorname{Var}(Z_k) \approx 1 + \frac{h^2_{\text{SNP}}}{M} N (1 + \rho_{\mathbf{G}}) . \tag{1}$$

The larger the variance in the Z statistic, the higher the probability of rejecting the null. The ratio of h^2_{SNP} and M can be regarded as the theoretical R^2 of each associated SNP with respect to the phenotype. Eq. 1 reveals that (i) when sample size increases, power increases, (ii) when h^2_{SNP} increases, the R^2 per associated SNP increases and therefore power increases, (iii) when the number of associated SNPs increases, the R^2 per associated SNP decreases and therefore power decreases, (iv) when the CGR is minus one, the studies perfectly cancel each other's genetic effects, thereby eliminating the power of the meta-analysis and reducing the distribution of the Z statistic for an associated SNP to a standard-normal distribution, yielding a strong disadvantage to meta-analyzing in this scenario, (v) when the CGR is zero the power of the meta-analysis is identical to the power obtained in each of the two studies when analyzed separately, yielding no strict advantage to meta-analyzing, and (vi) when the CGR is plus one the additional variance in the Z statistic relatively to the variance under the null is twice the additional variance one would have when analyzing the studies separately, yielding a strong advantage to meta-analyzing.

Others have focused on the highly related χ^2 statistics, defined as the squared Z statistics. In particular, it has been shown that the χ^2 statistics are influenced by linkage disequilibrium, population stratification, and polygenicity [36], [37], [38]. Although we focus on CGR and how it affects Z statistics rather than the χ^2 statistics, the factors that appear in our expressions of the variance of the GWAS Z statistics are highly similar to the factors that appear in work aiming to dissect the expected value of the GWAS χ^2 statistics. As an illustration of the similarity in expressions, consider the scenario where the CGR equals one between two samples of equal size. Based of Eq. 1, we then have that $\operatorname{Var}(Z_k) \approx 1 + N_{\text{total}} \frac{h^2_{\text{SNP}}}{M}$ for a trait-affecting haplotype block, where $N_{\text{total}} = 2N$. This expressions for the variance of the Z statistic of a trait-affecting haplotype block is completely equivalent to the expected χ^2 statistic from the linear regression analysis for a trait-affecting variant reported in Section 4.2 of the Supplementary Note to [37] as well as Equation 1 in [38] when assuming that confounding biases and linkage disequilibrium are absent. However, under a scenario with two or more studies with imperfect CGR, this overlap breaks down.

In order to compute statistical power in a multi-study setting, we first use the generic expression for the variance of the GWAS Z statistic derived in S1 Derivations Power to characterize the distribution of the Z statistic under the alternative hypothesis. We then use the inverse normal cumulative distribution function to quantify the probability of attaining genome-wide significance for an associated SNP. This probability we refer to as the 'power per associated SNP'. Moreover, given that we use SNPs tagging independent haplotype blocks, we calculate the probability of rejecting the null for at least one of the associated SNPs and the expected number of independent hits as follows:

$$\text{power to detect at least one SNP} = 1 - [1 - (\text{power per associated SNP})]^M \quad \text{and}$$

$$\mathbb{E}\left[\text{number of hits}\right] = M \cdot (\text{power per associated SNP}).$$

R^2 of a polygenic score under heterogeneity

In S2 Derivations Accuracy we derive a generic expression for the theoretical R^2 of a PGS in a hold-out sample, with SNP weights based on a meta-analysis of GWAS results under imperfect CGRs. We consider a PGS that includes all the SNPs that tag independent haplotype blocks (i.e., there is no SNP selection).

For intuition, we here present an approximation for prediction in a hold-out sample, with SNP weights based on a GWAS in a single discovery study with sample size N, where both studies have SNP-heritability ${h^2}_{\text{SNP}}$, and with CGR $\rho_{\mathbf{G}}$, between the studies. Under high polygenicity, the R^2 of the PGS in the hold-out sample is then given by the following expression:

$$R^2 \approx h^2_{\text{SNP}} \rho^2_{\mathbf{G}} \frac{h^2_{\text{SNP}}}{\frac{S}{N} + h^2_{\text{SNP}}}. \tag{2}$$

In case the CGR is one, and we consider the R^2 between the PGS and the genetic value (i.e., the genetic component of the phenotype) instead of the phenotype itself, the first two terms in Eq. 2 disappear, yielding an expression equivalent to Equation 1 in [34]. Assuming a CGR of one and that all SNPs are associated, Eq. 2 is equivalent to the expression in [26] for the R^2 between the PGS and the phenotype in the hold-out sample.

From Eq. 2, we deduce that (i) as the effective number of SNPs S increases, the R^2 of the PGS deteriorates (since every SNP-effect estimate contains noise, owing to imperfect inferences in finite samples), (ii) given the effective number of SNPs, under a polygenic architecture, the precise fraction of effective SNPs that is associated does not affect the R^2, (iii) R^2 is quadratically proportional to $\rho_{\mathbf{G}}$, implying a strong sensitivity to CGR, and (iv) as the sample size of the discovery study grows, the upper limit of the R^2 is given by $h^2_{\text{SNP}}\rho^2_{\mathbf{G}}$, implying that the full SNP-heritability in the hold-out sample cannot be entirely captured so long as CGR is imperfect.

Online power and R^2 calculator

An online version of the MetaGAP calculator can be found at `www.devlaming.eu`. This calculator computes the theoretical power per trait-affecting haplotype block, the power to detect at least one of these blocks, and the expected number of independent hits for a meta-analysis of GWAS results from C studies. In addition, it provides the expected R^2 of a PGS for a hold-out sample, including all GWAS SNPs, with SNP-weights based on the meta-analysis of the GWAS results from C studies. Calculations are based on the generic expressions for GWAS power derived in S1 Derivations Power and PGS R^2 derived in S2 Derivations Accuracy.

The calculator assumes a quantitative trait. Users need to specify either the average sample size per study or the sample size of each study separately. In addition, users need to specify either the average SNP-heritability across studies or the SNP-heritability per study. The SNP-heritability in the hold-out sample also needs to be provided. Users are required to enter the effective number of causal SNPs and the effect number of SNPs in total. The calculator assumes a fixed CGR between all pairs of studies included in the meta-analysis and a fixed CGR between the hold-out sample and each study in the meta-analysis. Hence, one needs to specify two CGR values: one for the CGR within the set of meta-analysis studies and one to specify the genetic overlap between the hold-out sample and the meta-analysis studies.

Finally, a more general version of the MetaGAP calculator is provided in the form of `MATLAB` code, also available at `www.devlaming.eu`. This code can be used in case one desires to specify a more versatile genetic-correlation matrix, where the CGR can differ between all pairs of studies. Therefore, this implementation requires the user to specify a full $(C+1)$-by-$(C+1)$ correlation matrix. Calculations in this code are fully in line with the generic expressions in S1 Derivations Power and S2 Derivations Accuracy.

Assessing validity of theoretical power and R^2

We simulate data for a wide range of genetic architectures in order to assess the validity of our theoretical framework. As we show in S3 Simulation Study, the theoretical expressions we derive for power and R^2 are accurate, even for data generating processes substantially different from the process we assume in our derivations. Our strongest assumption is that SNPs have equal R^2 with respect to the phenotype regardless of allele frequency. When we simulate data where this assumption fails and where allele frequencies are non-uniformly distributed, the root-mean-square prediction error of statistical power lies below 3% and that of PGS R^2 below 0.12%.

Estimating SNP-heritability and CGR

Using 1000 Genomes-imputed (1kG) data from the RS, STR, and HRS, we estimate SNP-based heritability and CGR respectively by means of univariate and bivariate genomic-relatedness-matrix restricted maximum likelihood (GREML) [31], [39] as implemented in GCTA [31]. In our analyses we consider the subset of HapMap3 SNPs available in the 1kG data. In S4 Data and Quality Control we report details on the genotype and phenotype data, as well as our quality control (QC) procedure. After QC we have a dataset, consisting of $\approx$ 1 million SNPs and $\approx$ 20,000 individuals, from which we infer h^2_{SNP} and CGR. In S5 GREML Estimation we provide details on the specifications of the models used for GREML estimation.

Results

Determinants of GWAS power and PGS R^2

Using the MetaGAP calculator, we assessed the theoretical power of a meta-analysis of GWAS results from genetically heterogeneous studies and the theoretical R^2 of the resulting PGS in a hold-out sample, for various numbers of studies and sample sizes, and different values of CGR and h^2_{SNP}.

Sample size and CGR Fig. 1 shows heat maps for the power per truly associated SNP and R^2 for a setting with 50 studies, for a trait with $h^2_{\text{SNP}} = 50\%$, for various combinations of total sample size and CGR. Increasing total sample size enhances both power and R^2. When the CGR is perfect, power and R^2 (relative to h^2_{SNP}) have a near-identical response to sample size. This similarity in response gets distorted when the CGR decreases. For instance, in the scenario of 100k SNPs of which a subset of 1k SNPs is causal with $h^2_{\text{SNP}} = 50\%$, in a sample of 50 studies with a total sample size of 10 million individuals, a CGR of one yields 94% power per causal SNP and an R^2 of 49%, which is 98% of the SNP-heritability, whereas for a CGR of 0.2 the power is still 87% per SNP, while the R^2 of the PGS is 8.5%, which is only 17% of h^2_{SNP}. Thus, R^2 is far more sensitive to an imperfect CGR than the meta-analytic power is. This finding is also supported by the approximations of power in in Eq. 1 and of PGS R^2 in Eq. 2; these expressions show that, for two discovery studies, the CGR has a linear effect on the variance of the meta-analysis Z statistic, whereas, for one discovery and one hold-out sample, the PGS R^2 is quadratically proportional to the CGR.

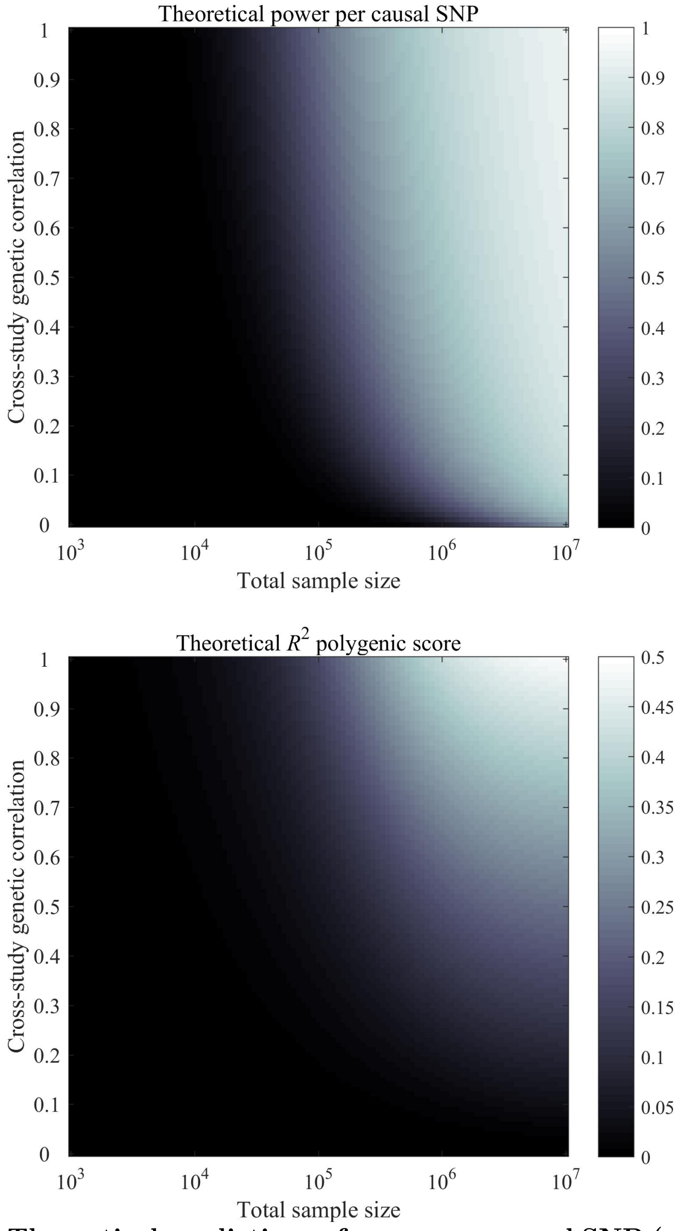

Figure 1. Theoretical predictions of power per causal SNP (upper panel) and out-of-sample R^2 of the PGS (lower panel), for total sample size (x-axis) and cross-study genetic correlation (y-axis). Factor levels: 50 studies, 100k independent SNPs, and heritability $h^2_{\text{SNP}} =$ 50% arising from a subset of 1k independent SNPs.

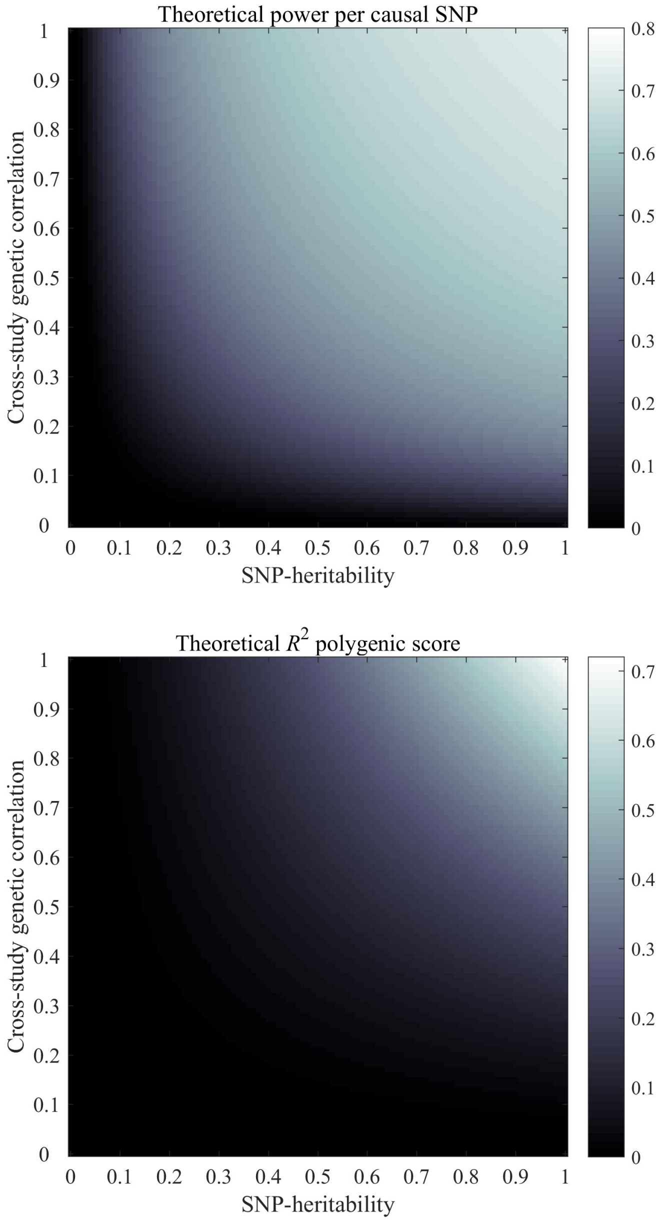

Figure 2. Theoretical predictions of power per causal SNP (upper panel) and out-of-sample R^2 of the PGS (lower panel), for a trait that across studies has SNP-heritability (x-axis) and cross-study genetic correlation (y-axis). Factor levels: 50 studies, sample size 5,000 individuals per study, 100k independent SNPs, and heritability arising from a subset of 1k independent SNPs.

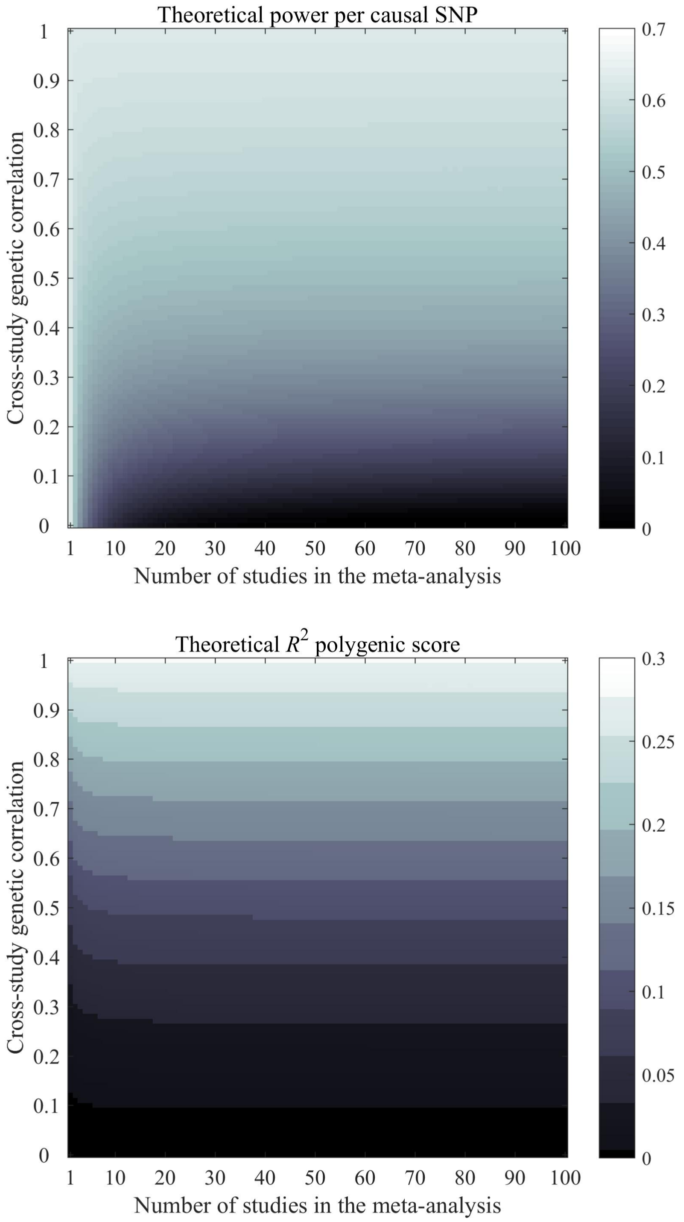

Figure 3. Theoretical predictions of power per causal SNP (upper panel) and out-of-sample R^2 of the PGS (lower panel), for a trait with GWAS results from the number of studies (x-axis) with cross-study genetic correlation (y-axis). Factor levels: total sample size 250,000 individuals, 100k independent SNPs, and heritability $h^2_{\text{SNP}} = 50\%$ arising from a subset of 1k independent SNPs. For R^2 a discontinuous color map is used to make salient details visible.

SNP-heritability and CGR Fig. 2 shows heat maps for the power per truly associated SNP and R^2 for a setting with 50 studies, with a total sample size of 250,000 individuals, for 1k causal SNPs and 100k SNPs in total, for various combinations of h^2_{SNP} and CGR. The figure shows a symmetric response of both power and R^2 to CGR and h^2_{SNP}. For instance, when $h^2_{\text{SNP}} = 25\%$ and CGR $= 0.5$ across all studies, the power is expected to be around 34% and the R^2 3.0%. When these numbers are interchanged (i.e., $h^2_{\text{SNP}} = 50\%$ and CGR $= 0.25$), similarly, the power is expected to be 35% and the R^2 2.9%. Hence, in terms of both R^2 and power, a low heritability can be compensated by a high CGR (e.g., by means of homogeneous measures across studies) and a low CGR can be compensated by high heritability.

When looking at two points with the same power (resp. R^2), any other point on a straight line between these points has a higher power (R^2), than at the end-points of the line. For instance, when both h^2_{SNP} and CGR lie at the midpoint between the 0.25 and 0.5 considered before (i.e., $h^2_{\text{SNP}} = 37.5\%$ and CGR $= 0.375$), the expected power is $37\% > 35\%$ and the expected R^2 $3.6\% > 3.0\%$.

When either CGR or heritability is (close to) zero, both power and R^2 are decimated in the multi-study setting. Hence, least power and R^2 can be found when h^2_{SNP}, CGR, or both are low. However, when both are moderately low but still substantially greater than zero, neither power nor R^2 are completely diminished.

Number of studies and CGR Fig. 3 shows heat maps for the power per truly associated SNP and R^2 for a trait with $h^2_{\text{SNP}} = 50\%$, 1k causal SNPs, 100k SNPs in total, and a fixed total sample size of 250,000 individuals. In this figure, various combinations of the number of studies and CGR are considered. The color map used for R^2 is discontinuous, in order to make salient details visible. Logically, when there is just one study for discovery, CGR does not affect power. However, even for two studies, the effect of CGR on power is quite pronounced. For instance, when CGR is a half, the power per causal SNP is 63% for one study, 58% for two studies, 51% for ten studies, and 50% for 100 studies. Thus, when the number of studies is low, increases in the number of studies make the effect of CGR on power more pronounced rapidly. When the number of studies is large, increases in the number of studies hardly make the effect of CGR on power more pronounced.

For a given number of studies, we observed that the effect CGR has on R^2 is stronger than the effect it has on power. This observation is in line with the approximated theoretical R^2 in Eq. 2, indicating that R^2 is quadratically proportional to CGR. However, an interesting observation is that this quadratic relation lessens as the number of studies grows large, despite the total sample size being fixed. For instance, at a CGR of a half, the R^2 in the hold-out sample is expected to be 6.9% when there is only one discovery study. However, the expected R^2 is 8.1% for two discovery studies, 9.3% for ten discovery studies, and 9.6% for 100 discovery studies. The reason for this pattern is that, in case of one discovery study, the PGS is influenced relatively strongly by the study-specific component of the genetic effects. This idiosyncrasy is not of relevance for the hold-out sample. As the number of studies increases, even though each study brings its own idiosyncratic contribution, each study also consistently conveys information about the part of the genetic architecture which is common across the studies. Now, since the idiosyncratic contributions from the studies are independent, they tend to average each other out, whereas the common underlying architecture gets more pronounced as the number of studies in the discovery increases, even if the total sample size is fixed.

SNP-heritability in the hold-out sample Fig. 4 shows a heat map for the PGS R^2 based on a meta-analysis of 50 studies with a total sample size of 250,000 individuals, with 1k causal SNPs and 100k SNPs in total, and a CGR of 0.8 between both the discovery studies and the hold-out sample. In the heat maps various combinations of h^2_{SNP} in the discovery samples and h^2_{SNP} in the hold-out sample are considered. The response of PGS R^2 to heritability in the discovery sample and the hold-out sample is quite symmetric, in the sense that a low h^2_{SNP} in the discovery samples and a high h^2_{SNP} in the hold-out sample yield a similar R^2 as a high h^2_{SNP} in the discovery sample and a low h^2_{SNP} in the hold-out sample. However, overall R^2 is slightly more sensitive to h^2_{SNP} in the hold-out sample than in the discovery samples. For instance, when SNP-heritability in the discovery samples is 50% and 25% in the hold-out sample, the expected R^2 is 10%, whereas in case the SNP-heritability is 25% in the discovery samples and 50% in the hold-out sample, the expected R^2 is 13%.

CGR between sets of studies Fig. 5 shows a heat map for the power per truly associated SNP in a setting where there are two sets consisting of 50 studies each. Within each set, the CGR is equal to one, whereas between sets the CGR is imperfect. Consider, for example, a scenario where one wants to meta-analyze GWAS results for height from a combination of two sets of studies; one set of studies consisting primarily of individuals of European ancestry and one set of studies with mostly people of Asian ancestry in it. Now, one would expect CGRs close to one between studies consisting primarily of individuals of European ancestry and the same for the CGRs between studies consisting primarily of people of Asian ancestry. However, the CGRs between those two sets of studies might be lower than one, though probably greater than zero.

As is shown in S1 Derivations Power, in case the CGR between the two sets is zero, meta-analyzing the two sets jointly is sub-optimal; the power of such a meta-analysis lies in between the power obtained by either of these sets when meta-analyzed separately. Since in Fig. 5 we considered two equally-powered sets, the power of a meta-analysis using both sets, under zero CGR between sets, is identical to the power obtained when meta-analyzing, for instance, only the first set. However, as CGR between sets increases so does power. For instance, when a total sample size of 250,000 individuals is spread across 2 clusters, each cluster consisting of 50 studies (i.e., sample size of 125,000 individuals per cluster and 2,500 individuals per study), under

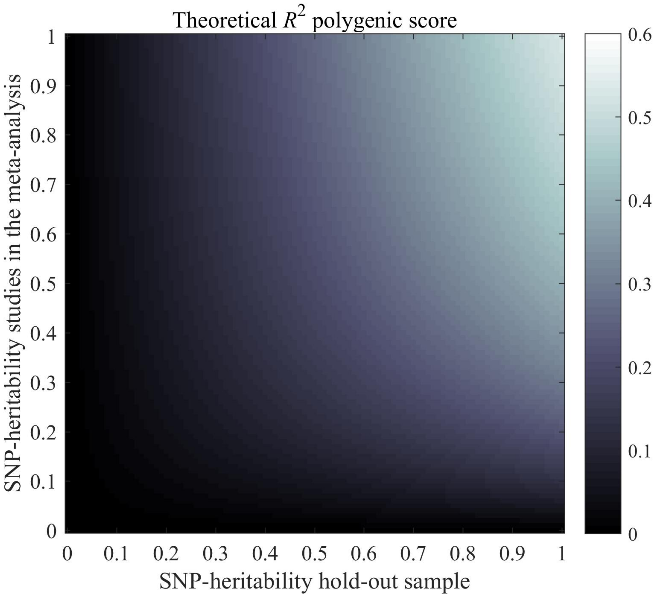

Figure 4. Theoretical predictions of out-of-sample R^2 of the PGS, for the SNP-heritability in the hold-out sample (x-axis) and the SNP-heritability in the discovery samples (y-axis). Factor levels: 50 studies, sample size 5,000 individuals per study, cross-study genetic correlation 0.8, 100k independent SNPs, and heritability arising from a subset of 1k independent SNPs.

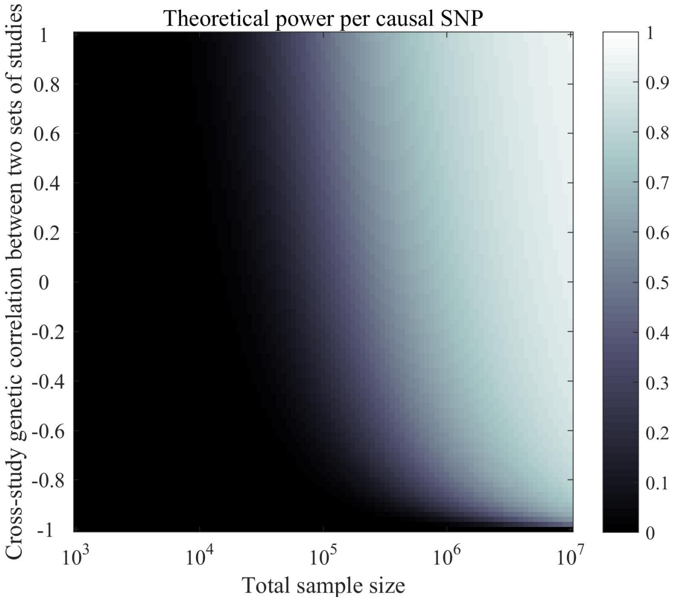

Figure 5. Theoretical predictions of power per causal SNP, for total sample size (x-axis) and CGR between two sets of studies (y-axis). Factor levels: 2 sets of 50 studies, CGR equal to 1 within both sets, 100k independent SNPs, and heritability $h^2_{\text{SNP}} = 50\%$ arising from a subset of 1k independent SNPs.

Table 1. GREML estimates of SNP-heritability (h^2_{SNP}) and genetic correlation across studies and sexes.

Phenotype	***N***	**Estimates SNP-heritability**[1,2]			**Estimates genetic correlation**[1]			
		pooled	**study**	**sexes**	**RS–STR**	**RS–HRS**	**STR–HRS**	**Females–Males**
Height	20,458	43.3% (1.8%) ***	44.9%	44.0%	0.976 (0.102) ***	0.954 (0.095) ***	0.967 (0.106) ***	0.981 (0.067) ***
BMI	20,449	20.9% (1.7%) ***	21.9%	22.8%	1.000 (0.269) ***	0.914 (0.172) ***	0.847 (0.246) ***	0.794 (0.122) *** †
EduYears	20,619	16.4% (1.7%) ***	18.2%	18.4%	0.690 (0.233) ***	0.659 (0.224) *** †	1.000 (0.263) ***	0.832 (0.162) ***
CurrCigt	20,686	18.2% (4.0%) ***	19.1%	24.2%	1.000 (0.643) ***	0.611 (0.448) *	1.000 (0.607) ***	0.543 (0.257) *** †
CurrDrinkFreq	20,072	7.0% (2.6%) ***	10.3%	8.3%	1.000 (0.666) ***	0.298 (0.670)	-0.056 (0.647)	1.000 (2.068) *
Self-rated health	19,184	10.3% (1.8%) ***	15.7%	9.5%	0.626 (0.439) **	0.363 (0.223) ** ††	0.447 (0.278) **	1.000 (0.349) ***

* h^2_{SNP} and/or genetic correlation > 0 at 10% sign. †genetic correlation < 1 at 10% sign. ‡genetic correlation < 0 at 10% sign.
** h^2_{SNP} and/or genetic correlation > 0 at 5% sign. ††genetic correlation < 1 at 5% sign. ‡‡genetic correlation < 0 at 5% sign.
*** h^2_{SNP} and/or genetic correlation > 0 at 1% sign. †††genetic correlation < 1 at 1% sign. ‡‡‡genetic correlation < 0 at 1% sign.

[1] Standard errors between parentheses.
[2] **pooled**: univariate estimate from pooled data, **study**: sample-size weighted average of univariate estimates across studies, and **sexes**: sample-size weighted average of univariate estimates across sexes.

$h^2_{\mathrm{SNP}} = 50\%$ due to 1k causal SNPs, a CGR of one within each cluster, and CGR of zero between clusters, the power is expected to be 49%, which is identical to the power of a meta-analysis of either the first or the second cluster. However, if the CGR between clusters is 0.5 instead of zero, the power goes up to 58%. In terms of the expected number of hits, this cross-ancestry meta-analysis yields an expected 82 additional hits, compared to a meta-analysis considering only one ancestry.

Alternatively, one could pool hits from two meta-analyses (e.g., in our example one in the European-ancestry set and one in the Asian-ancestry set). However, this would imply more independent tests being carried out, and, hence, the need for a stronger Bonferroni correction in order to keep the false-positive rate fixed, and, thus, a more stringent genome-wide significance threshold. Therefore, this route is likely to yield less statistical power than a meta-analysis of merely one of the set of two or a joint analysis of both sets.

Empirical results for SNP-based heritability and CGR

In Table 1 we report univariate GREML estimates of SNP-heritability and bivariate GREML estimates of genetic correlation for traits that attained a pooled sample size of at least 18,000 individuals, which gave us at least 50% power to detect a genetic correlation near one for a trait that has a SNP-heritability of 10% or more [40]. The smallest sample size is $N = 19{,}184$ for self-rated health. Details per phenotype (i.e., sample size, univariate estimates of SNP-heritability, and bivariate estimates of genetic correlation, stratified across studies and sexes, as well as cross-study and cross-sex averages) are provided in Tables 7-8 of S6 GREML Results.

The univariate estimates of SNP-heritability based on the pooled data assume perfect CGRs. Therefore, such estimates of SNP-heritability are downwards biased when based on data from multiple studies with imperfect CGRs. To circumvent this bias, we estimated SNP-heritability in each study separately, and focused on the sample-size-weighted cross-study average estimate of SNP-heritability.

For both height and BMI, we observed genetic correlations close to one across pairs of studies and between females and males. For years of schooling (*EduYears*) we found a CGR around 0.8 when averaged across pairs of studies. Similarly, the genetic correlation for *EduYears* in females and males lies around 0.8. The CGR of self-rated health is substantially below one across the pairs of studies, whilst the genetic correlation between females and males seems to lie around one. The reason for this difference in the genetic correlation between pairs of studies and between females and males may be due to the difference in the questionnaire across studies, discussed in S4 Data and Quality Control. The questionnaire differences can yield a low CGR, while not precluding the remaining genetic overlap for this measure across the three studies, to be highly similar for females and males. For *CurrCigt* and *CurrDrinkFreq*, the estimates of CGR and of genetic correlation between females and males are non-informative. For these two traits the standard errors of the genetic correlations estimates are large, mostly greater than 0.5. In addition, for *CurrDrinkFreq* there is strong volatility in the CGR estimate across pairs of studies.

Attenuation in power and R^2 due to imperfect CGR

Considering only the traits for which we obtained accurate estimates of CGR and SNP-heritability (i.e., with low standard errors), we used the MetaGAP calculator to predict the number of hits in a set of discovery samples and the PGS R^2 in a hold-out sample, in prominent GWAS efforts for these traits.

Since we only had accurate estimates for height, BMI, *EduYears*, and self-rated health, we focused on these four phenotypes. For these traits, we computed sample-size-weighted average CGR estimates across the pairs of studies. Table 2 shows the number of hits and PGS R^2 reported in the most comprehensive GWAS efforts to date for the traits of interest, together with predictions from the MetaGAP calculator. We tried several values for the number of independent haplotype blocks (i.e., 100k, 150k, 200k, 250k) and for the number of trait-associated blocks (i.e., 10k, 15k, 20k, 25k). Overall, 250k blocks of which 20k trait-affecting yielded theoretical predictions in best agreement with the empirical observations; we acknowledge the potential for some overfitting (i.e., two free parameters set on the basis of 17 data points; 10 data points for the reported number of hits and 7 for PGS R^2).

For height – the trait with the lowest standard error in the estimates of h^2_{SNP} and CGR – the predictions of the number of hits and PGS R^2 for the two largest GWAS efforts are much in line with theoretical predictions. For the smaller GWAS of 13,665 individuals [42], our estimates seem somewhat conservative; 0 hits expected versus the 7 reported. However, in our framework, we assumed that each causal SNP has the same R^2. Provided there are some differences in R^2 between causal SNPs, especially in smaller samples, the first SNPs that are likely to reach genome-wide significance are the ones with a comparatively large R^2. This view

Table 2. Predicted and observed number of genome-wide-significant hits and PGS R^2, for large-scale GWAS efforts to date for height, BMI, *EduYears*, and self-rated health, assuming 250k effective SNPs (i.e., independent haplotype blocks) of which 20k trait-affecting, using averaged GREML estimates from Table 1 for setting SNP-heritability and CGR.

Phenotype	Main studies			Architecture		Number of hits				PGS R^2 using all SNPs			
	Study	N	C**	h^2_{SNP}	CGR	Study	Theory|CGR <1	=1	Atten-uation*	Study	Theory|CGR <1	=1	Atten-uation*
Height	Wood *et al.* (2014) [1]	253,288	79	44.9%	0.965	697	647.26	700.24	8%	13.5%	13.2%	14.0%	6%
	Allen *et al.* (2010) [41]	183,727	61	44.9%	0.965	180	292.03	320.77	9%	10.0%	10.5%	11.1%	6%
	Weedon *et al.* (2008) [42]	13,665	5	44.9%	0.965	7	0.00	0.00	*n.a.*	***2.9%	1.0%	1.1%	7%
BMI	Locke *et al.* (2015) [2]	339,224	125	21.9%	0.917	97	188.52	241.07	22%	6.5%	4.3%	5.0%	14%
	Speliotes *et al.* (2010) [43]	123,865	46	21.9%	0.917	19	5.48	7.64	28%	2.5%	1.8%	2.1%	15%
	Willer *et al.* (2008) [44]	32,387	15	21.9%	0.917	1	0.01	0.02	65%	*n.a.*	0.5%	0.6%	16%
EduYears	Okbay *et al.* (2016) [7]	405,072	65	18.2%	0.783	162	115.28	235.90	51%	*n.a.*	2.7%	4.1%	36%
	Okbay *et al.* (2016) [7]	293,723	64	18.2%	0.783	74	39.30	88.93	56%	3.2%	2.0%	3.2%	36%
	Rietveld *et al.* (2013) [45]	101,069	42	18.2%	0.783	1	0.63	1.64	62%	2.5%	0.8%	1.2%	38%
Self-rated health	Harris *et al.* (2015) [46]	111,749	1	15.7%	0.468	13	1.35	1.35	0%	*n.a.*	0.2%	1.0%	78%

* Attenuation measures the relatively loss in expected power and R^2 due to a CGR in accordance with averaged GREML estimates from Table 1.
** C denotes the number of studies in the meta-analysis; C is slightly subjective (e.g., RS I, II, and III can be considered as one study or as three).
*** Based on 20 SNPs.

is supported by the fact that a PGS based on merely 20 SNPs already explains 2.9% of the variation in height. Hence, for relatively small samples our theoretical predictions of power and R^2 tend to be somewhat conservative. In addition, for height the 10k SNPs with the lowest meta-analysis p-values can explain about 60% of the SNP-heritability [1]. If the SNPs tagging the remaining 40% each have similar predictive power as the SNPs tagging the first 60%, then the number of SNPs needed to capture the full h^2_{SNP} would lie around 10k/0.6=17k, which is somewhat lower than the 20k which yields the most accurate theoretical predictions. However, as indicated before, the SNPs which appear most prominent in a GWAS are likely to be the ones with a greater than average predictive power. Therefore, the remaining 40% of h^2_{SNP} is likely to be stemming for SNPs with somewhat lower predictive power, thereby potentially inflating the number of SNPs needed to fully capture h^2_{SNP}. Hence, 20k associated independent SNPs is not an unlikely number for height.

The notion of a GWAS first picking up the SNPs with a relatively high R^2 is also supported by the predicted and observed number of hits for the reported self-rated-health GWAS [46]; given a SNP-heritability estimate between 10% [46] and 16% (Table 2), according to our theoretical predictions, a GWAS in a sample of around 110k individuals is unlikely to yield even a single genome-wide significant hit. However, this GWAS has yielded 13 independent hits. This finding supports the view that some relatively-high-R^2 SNPs are present in the genome.

For BMI our predictions of PGS R^2 were quite in line with empirical results. However, for the number of hits, our predictions for the largest efforts seemed overly optimistic. We therefore suspect that the number of independent SNPs associated with BMI is higher than 20k; as a higher number of associated SNPs would reduce the GWAS power, while preserving PGS R^2, yielding good agreement with empirical observation. Nevertheless, given the limited number of data points, this strategy of setting the number of causal SNPs would increase the chance of overfitting.

For *EduYears* we observed that the reported number of hits is in between the expected number of hits when the CGR is set to the averaged GREML estimate of 0.783 and when the CGR is set to one. Given the standard errors in the CGR estimates for *EduYears*, the CGR might very well be somewhat greater than 0.783, which would yield a good fit with the reported number of hits. However, as with the number of truly associated SNPs for BMI, we can make no strong claims about a slightly higher CGR of *EduYears* due to the risk of overfitting.

Overall, our theoretical predictions of the number of hits and PGS R^2 are in moderate agreement with empirical observations, especially when bearing in mind that we are looking at a limited number of data points, making chance perturbations from expectation likely. In addition, regarding the number of hits, the listed studies are not identical in terms of the procedure to obtain the independent hits. Therefore, the numbers could have been slightly different, had the same pruning procedure been used across all reported studies. Such differences in procedures introduce an additional element of chance.

Regarding attenuation, we observed a substantial spread in the predicted number of hits and PGS R^2 when assuming either a CGR of one, or a CGR in accordance with empirical estimates, with traits with lower CGR suffering from stronger attenuation in power and predictive accuracy. In line with theory, R^2 falls sharply with CGR. For instance, for self-rated health, the estimate CGR of about 0.5, would – in expectation – yield a PGS that retains only 0.5^2=25% of the R^2 it would have had under a CGR of one. This is supported by the reported attenuation of roughly 80%.

Given our CGR estimates, we expect a relative loss in PGS R^2 of 6% for height, 14% for BMI, 36% for *EduYears*, and 78% for self-rated health, compared to the R^2 of a PGS under perfect CGRs (Table 2). This loss in R^2 is unlikely to be reduced by larger sample sizes and denser genotyping.

Somewhat contrary to expectation, the number of hits seems to respond even more strongly to CGR than PGS R^2. However, since in each study under consideration the average power per associated SNP is quite small, a small decrease in power per SNP in absolute terms can constitute a substantial decrease in relative terms. For instance, when one has 2% power per truly associated SNP, an absolute decrease of 1% – leaving 1% power – constitutes a relative decrease of 50% of power per causal SNP, and thereby a 50% decrease in the expected number of hits. This strong response shows, for example, in the case of *EduYears*, where the expected number of hits drop by about 37% when going from a CGR of one down to a CGR of 0.783.

Discussion

In this study we aimed to answer the question whether imperfect cross-study genetic correlations (CGRs) help to explain a part of the 'hiding' heritability for highly polygenic traits such as height. We showed that imperfect CGRs are indeed likely to contribute to the gap between the phenotypic variation accounted for by all SNPs jointly and by the leading GWAS efforts to date. We arrive at this conclusion in five steps.

First, we developed a Meta-GWAS Accuracy and Power (MetaGAP) calculator that accounts for the CGR. This online calculator relates the statistical power to detect associated SNPs and the R^2 of the polygenic score (PGS) in a hold-out sample to the number of studies, sample size and SNP-heritability per study, and the CGR. The underlying theory shows that there is a quadratic response of the PGS R^2 to CGR. Moreover, we showed that the power per associated SNP is also influenced by CGR, although – in absolute terms – not as strongly as the PGS R^2.

Second, we used simulations to demonstrate that our theory is robust to several violations of the assumptions about the underlying data-generating process, regarding the relation between allele frequency and effect size, as well as the distribution of allele frequencies. Further research needs to assess whether out theoretical predictions are also accurate under an even broader set of scenarios (e.g., when studying a binary trait or when studying a trait for which there are relatively many rare variants with relatively small effects).

Third, we used a sample of unrelated individuals from the Rotterdam Study, the Swedish Twin Registry, and the Health and Retirement Study, to estimate SNP-based heritability as well as the CGR for traits such as height and BMI. Although our CGR estimates have considerable standard errors, the estimates make it likely that for many polygenic traits the CGR is positive, albeit smaller than one.

Fourth, based on these empirical estimates of SNP-heritability and CGR for height, BMI, years of education, and self-rated health, we used the MetaGAP calculator to predict the number of expected hits and the expected PGS R^2 for the most prominent studies to date for these traits. We found that our predictions are in good agreement with empirical observations. Although our theory turned out to be somewhat conservative for smaller GWAS samples, for large-scale GWAS efforts our predictions were in line with the outcomes of these efforts.

Fifth, we used our theoretical model to assess statistical power and predictive accuracy for these GWAS efforts, had the CGR been one for the traits under consideration. Our estimates of power and predictive accuracy in this scenario indicated a strong decrease in the PGS R^2 and the expected number of hits, due to imperfect CGRs. Though these observations are in line with expectation for predictive accuracy, for statistical power the effect was larger than we anticipated. This finding can be explained, however, by the fact that though the absolute decrease in power per SNP is small, the relative decrease is large, since the statistical power per associated SNP is often low to begin with.

Overall, our study affirms that although PGS accuracy improves substantially with further increasing sample sizes, in the end PGS R^2 will continue to fall short of the full SNP-based heritability. Hence, this study contributes to the understanding of the hiding heritability reported in the GWAS literature.

Regarding the etiology of imperfect CGRs, the likely reasons are heterogeneous phenotype measures across studies, gene–environment interactions with underlying environmental factors differing across studies, and gene–gene interactions where the average effects differ across studies due to differences in allele frequencies. Our study is not able to disentangle these different causes; by estimating the CGR for different traits we merely quantify the joint effect these three candidates have on the respective traits.

However, in certain situations it is possible to disentangle the etiology of imperfect CGRs to some extent. For instance, in case one considers a specific phenotype that is usually studied by means of a commonly available but relatively heterogeneous and/or noisy measure, while there also exists a less readily available but more accurate and homogeneous measure. If one has access to both these measures in several studies, one can compare the CGR estimates for the more accurate measure and the CGR estimates for the less accurate but more commonly available measure. Such a comparison would help to get some sense of the relatively contribution of phenotype heterogeneity to imperfect CGR in the heterogeneous measure.

In considering how to properly address imperfect CGRs, it is important to note that having a small set of large studies, rather than a large set of small studies, does not by definition abate the problem of imperfect genetic correlations. Despite the fact that having less studies can help to reduce the effects of heterogeneous phenotype measures, larger studies are more likely to sample individuals from different environments. If gene–environment interactions do play a role, strong differences in environment between subsets of individuals in a study lead to imperfect genetic correlations within that study. The attenuation in power and accuracy resulting from the imperfect genetic correlations within studies may prove hard to address.

In addition to studying the reduction in power and predictive accuracy due to CGR, we used our theoretical framework to consider other factors influencing power and accuracy. We found that in terms of power, sample size trumps a lot, even a relatively low CGR. Moreover, we observed – in line with our theoretical framework – that PGS R^2 is far more sensitive to CGR than absolute power per SNP. Also, we found that low CGR can to some extent be leveraged by a high SNP-heritability and vice versa. However, it is better to have both at a moderate level than one extremely high and the other extremely low; if either is zero the meta-analysis approach will fail.

We observed – given a fixed total sample size – that the substantial effects of CGR on power and predictive accuracy arise even for as few as two studies. Moreover, the CGR-power and CGR-accuracy relations do not change much as the number of underlying studies keeps increasing. This finding is reassuring; given that some power in the meta-analysis is lost due to imperfect CGRs, whether the underlying data is then highly fractured into many small studies or into a few big ones does not really matter for predictive accuracy or statistical power.

For SNP-heritability in the discovery samples and in the hold-out samples, we found that the PGS accuracy is slightly more affected by SNP-heritability in the hold-out sample than in the discovery samples. Hence, when aiming at high PGS accuracy, we recommend to use the study with the highest SNP-heritability and the highest CGR with the discovery samples as hold-out sample.

In addition to the number of studies, sample size and SNP-heritability per study, and CGR, our theoretical model depends on the specification of the following two latent parameters: the number of independent haplotype blocks (i.e., the ‘effective number of SNPs’) and the number of blocks containing trait-affecting variation (i.e., the number of independent ‘causal’ SNPs). In our work, setting the independent number of blocks at 250k and the number of trait-affecting blocks at 20k for all traits yielded the most accurate predictions.

Regarding the response of PGS accuracy and statistical power to these two parameters, it is interesting to note that our equations point to strongly opposed responses. Since effect sizes tend to decrease with an increasing number of causal SNPs, the statistical power decreases as the number of causal SNPs increases. The PGS R^2, on the other hand, decreases with the effective number of SNPs, since each SNP in the prediction model contributes some noise. By applying SNP-selection methods in the construction of a PGS, one can reduce the number of SNPs entering the PGS, decreasing the amount of noise and improving R^2. However, such methods may also exclude associated regions, decreasing the amount of signal in the score and attenuating R^2. Hence, SNP-selection methods are only likely to improve PGS R^2 when the selection is based on sufficiently accurate inferences.

Finally, having shown the substantial effect of imperfect CGRs on GWAS power and PGS R^2, we believe that the online MetaGAP calculator will prove to be an important tool for assessing whether an intended meta-analysis of GWAS results from different studies, is likely to yield meaningful outcomes.

Supporting Information

S1 Derivations Power

In this section, we derive an expression for the power of a meta-analysis of GWAS results, under a design with many studies, with arbitrary sample sizes, SNP-based heritability, and cross-study genetic correlation (CGR).

First, the underlying assumptions are presented. Second, we write the GWAS Z statistics in terms of the true SNP effect and noise. Third, we incorporate cross-study genetic correlations by assuming a model with random SNP effects that are correlated imperfectly across studies. Using the Cholesky decomposition of the cross-study genetic correlation matrix, we write the correlated SNP effects in terms of a weighted sum of independent genetic factors. By means of this decomposition into independent factors, we derive the distribution of the Z statistic in a given study, as well as the distribution of the multi-study meta-analysis Z statistic. From the latter distribution we obtain a framework for performing multi-study power calculations.

It is important to note that models which incorporate random SNP effects have been widely used, for instance, to estimate variance components [31] and genetic correlations across traits and samples [39], to control for cryptic relatedness and population structure in a GWAS [37], and to distill the constituents of genomic inflation [36], [38]. Hence, the novelty in our work lies not in using random SNP-effect models to incorporate imperfect genetic correlations across studies. Instead the novelty lies in the subsequent step, viz., to use such models in order to perform power calculations under the presence of imperfect CGRs.

Assumptions We derive an expression of statistical power for a quantitative trait in sample-size weighted meta-analysis [47]. In order to arrive at a tractable expression of statistical power, we make the following assumptions.

1. When considering a given SNP in the GWAS, any phenotypic variance due to other SNPs gets absorbed by the normally, independent, and identically distributed residual term (which is what happens when studying a sample of unrelated individuals, which is in line with assumptions underlying most GWAS packages, except for family-based and mixed-linear-model-type GWAS software). This assumption keeps the algebra simple at the cost of a small loss in generality.

2. The genome can be divided into independent haplotype blocks, where for each block we have precisely one SNP that tags the variation within this block. By means of this assumption, we can ignore linkage disequilibrium, thereby strongly reducing the complexity of our derivations. In addition, we assume that the effects of trait-affecting haplotype blocks are independent.

3. The SNP-effect-sizes are inversely related to SNP variance (i.e., rare variants have larger effects than common variants, such that the expected R^2 of each causal SNP, with respect to the phenotype, is equal regardless of allele frequency. In S3 Simulation Study we show that violations of this assumption hardly affect results). This assumption makes it possible to compute statistical power without having to specify the allele frequency and an *a priori* unknown effect size. Under this assumption, SNP-heritability and the number of trait-affecting haplotype blocks replace a SNP-specific effect size and allele frequency.

4. The regressors (i.e., SNP data) in the meta-analysis studies are fixed (i.e., non-stochastic)—this assumption is equivalent to conditioning on the genotype data (in S3 Simulation Study we show that violations of this assumption do not affect results). This assumption also keeps the algebra simple at the cost of a small loss in generality.

Single-SNP model Here, we write the GWAS Z statistic in a given study for a given SNP, as a function of the true effect and noise. This decomposition into true effect and noise helps to derive the distribution of the Z statistic.

For studies $j = 1, \ldots, C$ and SNPs $k = 1, \ldots, S$, let the model for a quantitative trait with a single SNP as predictor (Assumption 1) for the mean-centered phenotype $\mathbf{y}_j$ be given by

$$\mathbf{y}_j = \mathbf{x}_{jk}\beta_{jk} + \boldsymbol{\varepsilon}_j, \tag{3}$$

$$\boldsymbol{\varepsilon}_j \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{\boldsymbol{\varepsilon}_j}\mathbf{I}_{N_j}\right) \tag{4}$$

where $\mathbf{x}_{jk}$ denotes the mean-centered genotype vector of SNP k in study j, scaled such that $(\mathbf{x}_{jk}^\top\mathbf{x}_{jk})/N_j = 1$. In Eq. 3, β_{jk} is the effect of SNP k in study j. In Eq. 4, $\boldsymbol{\varepsilon}_j$ is the residual and $\mathbf{I}_{N_j}$ the $N_j \times N_j$ identity matrix, where N_j denotes the sample size of study j.

The GWAS estimate of β_{jk} for a quantitative trait is usually obtained by applying OLS. Hence, it can be written as

$$\begin{aligned}\widehat{\beta}_{jk} &= \left(\frac{1}{N_j}\mathbf{x}_{jk}^\top\mathbf{x}_{jk}\right)^{-1}\frac{1}{N_j}\mathbf{x}_{jk}^\top\mathbf{y}_j \qquad (5)\\ &= \frac{1}{N_j}\mathbf{x}_{jk}^\top\mathbf{y}_j \qquad (6)\\ &= \frac{1}{N_j}\mathbf{x}_{jk}^\top\mathbf{x}_{jk}\beta_{jk} + \frac{1}{N_j}\mathbf{x}_{jk}^\top\boldsymbol{\varepsilon}_j \qquad (7)\\ &= \beta_{jk} + \frac{1}{N_j}\mathbf{x}_{jk}^\top\boldsymbol{\varepsilon}_j. \qquad (8)\end{aligned}$$

Using standard results from regression theory assuming fixed regressors (Assumption 4) and the aforementioned scaling of the genotype vector, the theoretical variance of the OLS-estimate of the SNP effect is given by

$$\begin{aligned}\mathrm{Var}\left(\widehat{\beta}_{jk}\right) &= \sigma^2_{\boldsymbol{\varepsilon}_j}\left(\mathbf{x}_{jk}^\top\mathbf{x}_{jk}\right)^{-1}\\ &= \frac{\sigma^2_{\boldsymbol{\varepsilon}_j}}{N_j}.\end{aligned}$$

Therefore, the standard error of the OLS estimate is given by

$$\text{s.e.}\left(\widehat{\beta}_{jk}\right) = \frac{\sigma_{\boldsymbol{\varepsilon}_j}}{\sqrt{N_j}}. \tag{9}$$

By taking the ratio of Eq. 8 and 9 we obtain the Z statistic (instead of the commonly used and highly similar t-test statistics) for SNP k in study j. That is,

$$\begin{align} Z_{jk} &= \frac{\widehat{\beta}_{jk}}{\text{s.e.}\left(\widehat{\beta}_{jk}\right)} \tag{10} \\ &= \frac{\sqrt{N_j}}{\sigma_{\boldsymbol{\varepsilon}_j}}\beta_{jk} + \frac{\mathbf{x}_{jk}^{\top}\boldsymbol{\varepsilon}_j}{\sigma_{\boldsymbol{\varepsilon}_j}\sqrt{N_j}}. \tag{11} \end{align}$$

Let v_{jk} denote the last term in the right-hand side of Eq. 11. Under the aforementioned scaling of the regressor and the distribution of $\boldsymbol{\varepsilon}_j$, it follows from standard properties of the multivariate normal distribution that $v_{jk} \sim \mathcal{N}(0, 1)$.

Modelling cross-study genetic correlation Here, we incorporate cross-study genetic correlations by considering a model with random SNP effects, correlated across studies. In order to simplify further derivations, we use a Cholesky decomposition to write correlated SNP effects in terms of independent underlying factors. Using this independent-factor representation, we derive the distribution of a GWAS Z statistic, in terms of the study-specific noise and contributions of the underlying genetic factors.

Genetic correlation can be conceptualized as the correlation between SNP effects across different strata (e.g., across populations, studies, age groups, etc.). Taking studies as 'strata', a group of C studies has $C \times C$ genetic correlation matrix, denoted by $\boldsymbol{P}_{\mathbf{G}}$.

When effects are normally distributed, a given correlation structure between effects is most straightforwardly obtained by constructing the Cholesky decomposition of the correlation matrix, and multiplying independent standard-normal random variables by this decomposition. An interpretation of this decomposition is that it provides a set of weights that transform a set of independent underlying genetic factors into correlated genetic effects.

First, we formalize how to transform independent standard-normal random variables into correlated normal random variables. Let $\mathbf{\Gamma}_{\mathbf{G}}$ be the lower-triangular Cholesky decomposition of the genetic correlation matrix, such that $\mathbf{\Gamma}_{\mathbf{G}}\mathbf{\Gamma}_{\mathbf{G}}{}^{\top} = \boldsymbol{P}_{\mathbf{G}}$, let $\mathcal{M}$ denote the set of M causal SNPs, let $\mathbf{E}$ be an C-by-M matrix of independent standard normal draws from different genetic factors (rows) for the different causal SNPs (columns), and let $\boldsymbol{\eta}_k$ be the column of $\mathbf{E}$ corresponding to causal SNP k. Then

$$\boldsymbol{\eta}_k = \begin{pmatrix} \eta_{1k} \\ \vdots \\ \eta_{Ck} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_C\right), \tag{12}$$

where $\boldsymbol{\eta}_k$ is independent of $\boldsymbol{\eta}_l$ for $l \neq k$ (Assumption 2). Now, for SNP k in the set of causal SNPs, we can define the vector of effects across studies for the given SNP, such that it has correlation matrix $\boldsymbol{P}_{\mathbf{G}}$, as follows:

$$\boldsymbol{\beta}_k = \begin{pmatrix} \beta_{1k} \\ \vdots \\ \beta_{Ck} \end{pmatrix} = \text{diag}\left(\sigma_{\beta_1}, \ldots \sigma_{\beta_C}\right) \boldsymbol{\Gamma}_{\mathbf{G}} \boldsymbol{\eta}_k, \tag{13}$$

where diag() is a diagonal matrix with specified elements as diagonal entries, and

$$\sigma_{\beta_j} = \sqrt{\frac{h_j^2 \sigma_{\mathbf{y}_j}^2}{M}}, \tag{14}$$

with h_j^2 (resp. $\sigma_{\mathbf{y}_j}^2$) denoting the SNP-heritability (phenotypic variance) in study j. Under this design of study-specific SNP effects, we attain a CGR structure in line with $\boldsymbol{P}_{\mathbf{G}}$ and the desired study-specific SNP-heritabilities.

Using this approach for constructing correlated SNP effects, we can write the effect of SNP k in study j (i.e., β_{jk}) as a linear combination of the independent underlying $\mathcal{N}(0,1)$ distributed random variables. That is,

$$\beta_{jk} = \sigma_{\beta_j} \sum_{i=1}^{j} \gamma_{ji} \eta_{ik}, \tag{15}$$

where γ_{ji} denotes element in row j column i of $\boldsymbol{\Gamma}$ and η_{ik} the i-th element of $\boldsymbol{\eta}_k$. Given our scaling of SNPs, the R^2 of each causal SNP in study j is given by $\sigma_{\beta_j}^2$, regardless of the allele frequency of the SNP of interest (Assumption 3).

We can now write the GWAS Z statistic for a given SNP in a given study, as a linear combination of independent random variables. For SNP k in the set of P non-causal SNPs, denoted by $\mathcal{P}$ (such that $\mathcal{M} \cap \mathcal{P} = \varnothing$), we have for all studies j that $\beta_{jk} = 0$. By substituting β in Eq. 11 according to Eq. 15 for causal SNPs and the preceding equality for non-causal SNPs, we obtain the following expression for the Z

statistic of SNP k in study j:

$$Z_{jk} = \begin{cases} \upsilon_{jk} + \sqrt{N_j}\frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} \sum_{i=1}^{j} \gamma_{ji}\eta_{ik} & \text{for} \quad k \in \mathcal{M}, \text{ and} \\ \upsilon_{jk} & \text{for} \quad k \in \mathcal{P}. \end{cases} \tag{16}$$

Distribution meta-analysis Z statistic Here, we derive the distribution of the meta-analysis Z statistic and reduce the number of input parameters by appropriate substitutions. Finally, for intuition, we present distribution of Z statistics from a meta-analysis of GWAS results from two studies.

For any SNP k in the set $\mathcal{S} = \mathcal{M} \cup \mathcal{P}$ of $S = M + P$ causal and non-causal SNPs, we use the sample-size-weighted meta-analysis Z statistic [47], defined as follows:

$$Z_k = \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N}} Z_{jk}, \tag{17}$$

where $N = N_1 + \ldots + N_C$ denotes the total sample size. Plugging Eq. 16 for $k \in \mathcal{M}$ into Eq. 17, yields an expression for the meta-analysis Z statistic in terms of independent random variables. That is,

$$Z_k = \begin{cases} \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N}} \upsilon_{jk} + \sum_{j=1}^{C} \sum_{i=1}^{j} \frac{N_j}{\sqrt{N}} \frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} \gamma_{ji}\eta_{ik} & \text{for} \quad k \in \mathcal{M}, \text{ and} \\ \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N}} \upsilon_{jk} & \text{for} \quad k \in \mathcal{P}. \end{cases} \tag{18}$$

As the υ_{jk} terms in the preceding expression are independent standard-normal draws, it follows that

$$\upsilon_k = \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N}} \upsilon_{jk} \sim \mathcal{N}(0, 1). \tag{19}$$

In Eq. 18 we have a double sum over random variables. However, by changing the order of summation, this double sum can be rewritten as follows:

$$\sum_{j=1}^{C} \sum_{i=1}^{j} \frac{N_j}{\sqrt{N}} \frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} \gamma_{ji}\eta_{ik} = \sum_{i=1}^{C} \eta_{ik} \sum_{j=i}^{C} \frac{N_j}{\sqrt{N}} \frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} \gamma_{ji}. \tag{20}$$

Therefore, we can rewrite Eq. 18 as follows:

$$Z_k = \begin{cases} \upsilon_k + \sum_{i=1}^{C} \eta_{ik} \sum_{j=i}^{C} \frac{N_j}{\sqrt{N}} \frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} \gamma_{ji} & \text{for} \quad k \in \mathcal{M}, \text{ and} \\ \upsilon_k & \text{for} \quad k \in \mathcal{P}, \end{cases} \tag{21}$$

where the inner sum yields the weight for the random variable of interest.

Exploiting the fact that η_{ik} and υ_k are independent standard-normal draws, the variance of the sum of terms is equal to the sum of the variance of the respective terms. Hence, we have that

$$Z_k \sim \begin{cases} \mathcal{N}(0, 1+d) & \text{for} \quad k \in \mathcal{M}, \text{ and} \\ \mathcal{N}(0,1) & \text{for} \quad k \in \mathcal{P}, \end{cases} \tag{22}$$

where

$$d = \sum_{i=1}^{C} \left(\sum_{j=i}^{C} \frac{N_j}{\sqrt{N}} \frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} \gamma_{ji} \right)^2 \tag{23}$$

$$= \frac{1}{N} \sum_{i=1}^{C} \left(\sum_{j=i}^{C} N_j \frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} \gamma_{ji} \right)^2 \tag{24}$$

The quantity d we refer to as the 'power parameter'. Since this parameter is a sum of squares, it is non-negative. The greater the power parameter is, the more statistical power the meta-analysis of GWAS results has. Note that in case $\sigma_{\beta_j} = 0$ for all j (i.e., the trait is not heritable in any study), that $d = 0$, and hence the meta-analysis Z statistic reverts to a standard-normal test statistic, which matches the distribution under the null. However, as σ_{β_j} increases, d becomes larger, yielding a meta-analysis with more statistical power.

Given SNP-based heritability, phenotypic variation, and the number of causal variants, we have that the effect size per causal SNP in a study is given by $\sigma^2_{\beta_j} = \frac{h^2_j \sigma^2_{\mathbf{y}_j}}{M}$, and the residual variance, absorbing the variance due to the omitted $M-1$ SNPs (Assumption 1), is given by $\sigma^2_{\boldsymbol{\varepsilon}_j} = \sigma^2_{\mathbf{y}_j} - \sigma^2_{\beta_j}$. Using these expressions, we can write the following ratio, appearing in Eq. 24, as a function of only heritability and the number of causal SNPs.

$$\frac{\sigma^2_{\beta_j}}{\sigma^2_{\boldsymbol{\varepsilon}_j}} = \frac{\frac{h^2_j \sigma^2_{\mathbf{y}_j}}{M}}{\sigma^2_{\mathbf{y}_j} - \frac{h^2_j \sigma^2_{\mathbf{y}_j}}{M}} \tag{25}$$

$$= \frac{h^2_j}{M - h^2_j}. \tag{26}$$

Plugging the square root of this last expression into Eq. 24 yields

$$d = \frac{1}{N} \sum_{i=1}^{C} \left(\sum_{j=i}^{C} N_j \sqrt{\frac{h^2_j}{M - h^2_j}} \gamma_{ji} \right)^2 \tag{27}$$

This expression for the power parameter shows that it is not affected by scaling due to phenotypic variance; the parameter is only affected by the cross-study genetic correlation matrix, the SNP-based heritability per study, and the sample size per study.

In case the number of studies is two, with sample size N in Study 1 and N in Study 2, SNP-heritability h^2_{SNP}, and a genetic correlation $\rho_{\mathbf{G}}$ between the two studies, we have that the meta-analysis Z statistic, of a trait-affecting SNP k, is normally distributed with mean zero and

$$\operatorname{Var}\left(Z_{k,C=2}\right) = 1 + \frac{h^2_{\text{SNP}}}{M - h^2_{\text{SNP}}} N \left(1 + \rho_{\mathbf{G}}\right).$$

Bearing in mind that the number of causal SNPs $M \gg 1$ under a highly polygenic model, while $h^2 \in [0,1]$, we have that under high polygenicity $M - h^2_{\text{SNP}} \approx M$. Hence, an easy approximation of the variance of Z_k is given by

$$\operatorname{Var}\left(Z_{k,C=2,\text{high polygenicity}}\right) \approx 1 + \frac{h^2_{\text{SNP}}}{M} N \left(1 + \rho_{\mathbf{G}}\right).$$

In the scenario where the cross-study genetic correlations equals one, we have that $\operatorname{Var}\left(Z_k\right) \approx 1 + N_{\text{total}} \frac{h^2_{\text{SNP}}}{M}$ for a trait-affecting haplotype block and $\operatorname{Var}\left(Z_k\right) = 1$ for a non-causal haplotype block, where $N_{\text{total}} = 2N$. These expressions are equivalent to the expected χ^2 statistics from linear regression analysis reported in Section 4.2 of the Supplementary Note to [37], as well as Equation 1 in [38] when assuming that confounding biases and linkage disequilibrium are absent.

Adding genetically uncorrelated studies to the meta-analysis Here, we consider what happens to statistical power of a meta-analysis of GWAS results from several sets of studies, with genetic correlations between the studies within each set, but with no genetic correlation between the different sets. We first consider a scenario with one set consisting of $C-1$ studies and one other set consisting of only one study. We then generalize to a setting with multiple sets, each set containing at least one study. We show that the power parameter for a meta-analysis of several sets of studies with no genetic correlations between sets, can be written as a sample-size weighted sum of the power parameters within the respective sets.

In case one has $C-1$ studies with associated CGR matrix, the associated Cholesky decomposition denoted by $\mathbf{\Gamma}_{(C)}$, and an additional study indexed by C, which is genetically uncorrelated to the $C-1$ other studies, then the $C \times C$ Cholesky decomposition of the full CGR matrix is given by

$$\mathbf{\Gamma_G} = \begin{pmatrix} \mathbf{\Gamma}_{(C)} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{pmatrix}, \tag{28}$$

where $\mathbf{0}$ denotes a column vector of zeros.

Now, the quantity d in Eq. 27 can be decomposed as follows.

$$d = \frac{1}{N}\sum_{i=1}^{C-1}\left(\sum_{j=i}^{C-1} N_j\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)^2 + \frac{1}{N}\left(N_C\sqrt{\frac{h_C^2}{M-h_C^2}}\right)^2 \tag{29}$$

$$= \frac{N_{(C)}}{N}\frac{1}{N_{(C)}}\sum_{i=1}^{C-1}\left(\sum_{j=i}^{C-1} N_j\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)^2 + \frac{N_C}{N}\frac{1}{N_C}\left(N_C\sqrt{\frac{h_C^2}{M-h_C^2}}\right)^2 \tag{30}$$

$$= \frac{N_{(C)}}{N}d_{(C)} + \frac{N_C}{N}d_C, \tag{31}$$

where d_C denotes the power parameter in Eq. 27 had only study C (with sample-size N_C) be considered, and $d_{(C)}$ the power parameter in Eq. 27 had only the first $C-1$ studies (with total corresponding sample-size $N_{(C)}$) be considered. Hence, the power parameter in this scenario is the sample-size-weighted sum of the power parameter of the first $C-1$ studies jointly and the power parameter of the last study.

Eq. 31 can be generalized, to reflect a situation where there are P disjoint sets of studies, denoted by $\mathcal{C}_1,\ldots,\mathcal{C}_P$, with genetic correlation within each set, but no genetic correlation between the sets. In this scenario, the power parameter d in Eq. 27 for a joint meta-analysis of all sets is given by

$$d_{\mathcal{C}_1\cup\mathcal{C}_2\cup\ldots\cup\mathcal{C}_P} = \sum_{p=1}^{P}\frac{N_{\mathcal{C}_p}}{N}d_{\mathcal{C}_p}, \tag{32}$$

where $N_{\mathcal{C}_p}$ denotes the total sample size in study-set $\mathcal{C}_p$ and $d_{\mathcal{C}_p}$ the power parameter in Eq. 27 for the meta-analysis of all studies in set $\mathcal{C}_p$, and N the total sample size when aggregating over all study sets. This equation states that power parameter for a meta-analysis of several sets of studies with CGR within each set, but no CGR between sets, is a weighted average of the power parameters in the underlying sets.

The implication of Eq. 32 is simple yet powerful; when several sets of studies with genetic correlation within each set, but no genetic correlation between sets, are considered for meta-analysis, one should not meta-analyze sets $\mathcal{C}_1,\ldots\mathcal{C}_P$ jointly, but rather meta-analyze only the set of studies which has the largest power parameter according to Eq. 27.

Only when $d_{\mathcal{C}_1\cup\mathcal{C}_2\cup\ldots\cup\mathcal{C}_P} > \max\{d_{\mathcal{C}_1},\ldots,d_{\mathcal{C}_P}\}$, does the meta-analysis of all sets jointly have more statistical power than a meta-analysis based on only one set of studies.

S2 Derivations Accuracy

This section extends the theoretical framework for meta-analytic power. Derivations are based on the same assumptions as in S1 Derivations Power. We consider the predictive accuracy of the polygenic score (PGS) including all S independent SNPs, with SNP-weights based on the meta-analysis results from the set of C

study, in a hold-out sample indexed as 'study' $C+1$. In this hold-out sample, we focus exclusively on the theoretical R^2 of the PGS; instead of considering N_{C+1} realizations of the stochastic processes underlying the genotypes and treating these as fixed explanatory variables, we treat the phenotype, the PGS, and the underlying genotypes as random variables, and use probability theory to derive R^2. The hold-out sample is also allowed a study-specific SNP-based heritability, h^2_{C+1}, and genetic-correlations with the other C studies (thus extending both the CGR matrix and its Cholesky decomposition to $(C+1)\times(C+1)$ matrices).

First, we write the phenotype in hold-out sample as a function of noise and the independent genetic factors discussed in the preceding section. Second, we derive an expression for the PGS as a function of the genetic factors. Third, using this representation we derive the theoretical covariance between the PGS and the phenotype. Fourth, using the theoretical variances and covariance, we obtain an expression for the theoretical R^2.

Polygenic model Here, we derive an expression for the phenotype in the hold-out study as a function of independent genetic factors and an expression for the phenotypic variance.

Aggregating across causal SNP set $\mathcal{M}$ and the noise, the phenotype in study $C+1$ can be written as follows:

$$Y_{C+1} = \sum_{k\in\mathcal{M}} X_{C+1,k}\beta_{C+1,k} + \varepsilon_{C+1}, \tag{33}$$

where, analogous to Eq. 15,

$$\beta_{C+1,k} = \sigma_{\beta_{C+1,k}} \sum_{i=1}^{C+1} \gamma_{C+1,i}\eta_{ik}, \tag{34}$$

where η_{ik} now indicates the i-th element of the now $(C+1)$-dimensional vector of independent normal draws, $\boldsymbol{\eta}_k$, and where $\gamma_{C+1,i}$ describes an element of the Cholesky decomposition $\boldsymbol{\Gamma}_{\mathbf{G}}$ of the $(C+1)\times(C+1)$ cross-study genetic correlation matrix, incorporating the hold-out sample. Hence, the phenotype can be written as

$$Y_{C+1} = \varepsilon_{C+1} + \sum_{k\in\mathcal{M}} \left(X_{C+1,k}\sigma_{\beta_{C+1,k}} \sum_{i=1}^{C+1} \gamma_{C+1,i}\eta_{ik} \right). \tag{35}$$

Analogous to the scaling of SNPs in S1 Derivations Power here, with genotypes treated as random variables, we assume

$$\begin{aligned} \mathbb{E}\left[X_{C+1,k}\right] = 0 \text{ and } \operatorname{Var}\left(X_{C+1,k}\right) = 1, \text{ for } k \in \mathcal{S}, \text{ and} \\ \operatorname{Cov}\left(X_{C+1,k}, X_{C+1,l}\right) = 0 \text{ for } k \neq l. \end{aligned}$$

Consequently, the phenotypic variance in the hold-out sample is given by

$$\operatorname{Var}\left(Y_{C+1}\right) = M\sigma^2_{\beta_{C+1}} + \sigma^2_{\boldsymbol{\varepsilon}_{C+1}}. \tag{36}$$

Polygenic score Here, we derive an expression for the PGS as a function of independent genetic factors, an expression for the PGS variance, and its covariance with the phenotype in the hold-out sample.

Since each SNP in each study in the meta-analysis has been scaled such that its dot product equals the sample size of that study, by analogy of the standard error of the SNP effect estimate in a single study, the standard-error of the meta-analytic effect estimate $\widehat{\beta}_{meta}$ for study $C+1$ can be approximated by

$$\text{s.e.}\left(\widehat{\beta}_{meta}\right) \propto \frac{1}{\sqrt{N}} \propto 1.$$

Hence, the meta-analytic effect estimate is proportional to the meta-analysis Z statistic. Since any scalar multiple of the PGS will not affect its R^2 with respect to the phenotype, the Z statistics of the meta-analysis can be applied as SNP weights directly. Therefore, the PGS in the hold-out sample, including all SNPs, is given by

$$\widehat{Y}_{C+1} = \sum_{k \in \mathcal{S}} X_{C+1,k} Z_k. \tag{37}$$

Plugging the expression for Z_k from Eq. 21 into Eq. 37, and substitution of terms by means of the square root of Eq. 26, the PGS is given by

$$\widehat{Y}_{C+1} = \left(\sum_{k \in \mathcal{S}} X_{C+1,k} \upsilon_k\right) + \left(\sum_{k \in \mathcal{M}} X_{C+1,k} \sum_{i=1}^{C} \eta_{ik} \sum_{j=i}^{C} \frac{N_j}{\sqrt{N}} \sqrt{\frac{h_j^2}{M - h_j^2}} \gamma_{ji}\right). \tag{38}$$

Exploiting the fact that η_{ik}, υ_k, and $X_{C+1,k}$ are all independent random variables, with mean zero and variance one, we find that the variance of the PGS is given by

$$\operatorname{Var}\left(\widehat{Y}_{C+1}\right) = S + M \sum_{i=1}^{C} \left(\sum_{j=i}^{C} \frac{N_j}{\sqrt{N}} \sqrt{\frac{h_j^2}{M - h_j^2}} \gamma_{ji}\right)^2. \tag{39}$$

Again exploiting independence, zero mean, and unit variance of the respective terms, the covariance between the PGS and the phenotype is given by

$$\begin{aligned}
\operatorname{Cov}\left(Y_{C+1},\widehat{Y}_{C+1}\right) &= \mathbb{E}\left[Y_{C+1}\widehat{Y}_{C+1}\right] && (40)\\
&= \begin{matrix}\mathbb{E}\left[\left(\sum_{k\in\mathcal{M}} X_{C+1,k}\sigma_{\beta_{C+1,k}} \sum_{i=1}^{C+1} \gamma_{C+1,i}\eta_{ik}\right)\dots\right.\\ \left.\cdot\left(\sum_{k\in\mathcal{M}} X_{C+1,k}\sum_{i=1}^{C}\eta_{ik}\sum_{j=i}^{C}\frac{N_j}{\sqrt{N}}\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)\right]\end{matrix} && (41)\\
&= \mathbb{E}\left[\left(\sum_{k\in\mathcal{M}} X_{C+1,k}^2\sigma_{\beta_{C+1,k}}\left(\sum_{i=1}^{C}\gamma_{C+1,i}\eta_{ik}^2\sum_{j=i}^{C}\frac{N_j}{\sqrt{N}}\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)\right)\right] && (42)\\
&= \sigma_{\beta_{C+1,k}} M\left(\sum_{i=1}^{C}\sum_{j=i}^{C}\frac{N_j}{\sqrt{N}}\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{C+1,i}\gamma_{ji}\right). && (43)
\end{aligned}$$

Theoretical R^2 Here, we derive the theoretical R^2 between the PGS and the phenotype in a hold-out study. For intuition, we present the theoretical R^2 for a scenario with one study for discovery and one study as hold-out sample.

By combining Eq. 36, 39, and 43, the R^2, defined as the squared correlation of the outcome and the PGS in the hold-out sample, is now given by

$$\begin{aligned}
R^2\left(Y_{C+1},\widehat{Y}_{C+1}\right) &= \frac{\left(\operatorname{Cov}\left(Y_{C+1},\widehat{Y}_{C+1}\right)\right)^2}{\operatorname{Var}\left(Y_{C+1}\right)\operatorname{Var}\left(\widehat{Y}_{C+1}\right)}\\
&= \frac{\sigma_{\beta_{C+1,k}}^2 M^2\left(\sum_{i=1}^{C}\sum_{j=i}^{C}\frac{N_j}{\sqrt{N}}\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{C+1,i}\gamma_{ji}\right)^2}{\left(M\sigma_{\beta_{C+1}}^2+\sigma_{\boldsymbol{\varepsilon}_{C+1}}^2\right)\left(S+M\sum_{i=1}^{C}\left(\sum_{j=i}^{C}\frac{N_j}{\sqrt{N}}\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)^2\right)}.
\end{aligned}$$

This expression can be simplified as follows:

$$R^2\left(Y_{C+1},\widehat{Y}_{C+1}\right) = h_{C+1}^2\frac{n}{\frac{S}{M}+d}, \tag{44}$$

where d is the meta-analysis power parameter given in Eq. 27 and numerator n is given by

$$n = \frac{1}{N}\left(\sum_{i=1}^{C}\sum_{j=i}^{C}N_j\sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{C+1,i}\gamma_{ji}\right)^2, \tag{45}$$

where N is the total sample size in the meta-analysis.

The expression for R^2 in Eq. 44 is such that, in addition to the parameters needed for the power calculation, one only needs the genetic correlation between the hold-out sample and the meta-analysis samples and the heritability in the hold-out sample.

In case the number of studies for discovery is one (i.e., $C = 1$), with a total sample size N, and with a genetic correlation $\rho_{\mathbf{G}}$ between the hold-out and discovery sample, we have that

$$R^2_{C=1} = h_2^2 \rho_{\mathbf{G}}^2 \frac{\frac{N h_1^2}{M - h_1^2}}{\frac{S}{M} + \frac{N h_1^2}{M - h_1^2}}.$$

As in S1 Derivations Power, we have that under high polygenicity $M - h_1^2 \approx M$. Therefore, an easy approximation of R^2 in this scenario is given by

$$R^2_{C=1,\text{high polygenicity}} \approx h_2^2 \rho_{\mathbf{G}}^2 \frac{h_1^2}{\frac{S}{N} + h_1^2}.$$

When $\rho_{\mathbf{G}}^2 = 1$, $S{=}M$, and $h_1^2 = h_2^2$, we obtain a known expression for PGS R^2 in terms of sample size, heritability, and the number of SNPs [26]. In case $\rho_{\mathbf{G}}^2 = 1$ and we consider the R^2 between the PGS and genetic value (i.e., the genetic component of the phenotype), both $\rho_{\mathbf{G}}^2$ and h_2^2 can be ignored, thereby making the last expression equivalent to Equation 1 in [34].

S3 Simulation Study

Using a set of three simulation studies, we assess the accuracy of the Meta-GWAS Accuracy and Power (MetaGAP) calculator, which is based on the expressions for GWAS power and PGS R^2 derived in S1 Derivations Power and S2 Derivations Accuracy respectively. Since the calculator is based on specific assumptions regarding the data-generating process, an important question is whether the calculator still provides accurate predictions of power and R^2 when the underlying assumptions are violated.

Hence, each simulation study has a different underlying data-generating process. The first study, Simulation 1, assumes that rare variants have larger effects than common variants to such an extent that each causal SNP, regardless of allele frequency, is expected to have the same R^2 with respect to the phenotype. However, the second study, Simulation 2, assumes that common variants have effects of the same magnitude as rare variants (leading a common causal variant to explain a larger proportion of the phenotypic variation that a rare causal variant). Finally, the third study, Simulation 3, also allows for differential R^2 between SNPs and, in addition, does not assume that SNP allele frequencies are uniformly distributed. Instead, the third study assumes that there are more variants in the lower minor allele frequency bins than in the higher minor allele frequency bins.

For each simulation study there are 100 independent runs. In each run data is simulated for $C = 3$ distinct

Table 3. Design of Simulations 1–3.

Data-generating process*	Notation	Simulation 1	Simulation 2	Simulation 3
#studies for meta-analysis	$C =$	3	idem	idem
Index prediction sample	$C+1 =$	4	idem	idem
Sample size per study	$\{N_1, N_2, N_3, N_4\} =$	{20k; 15k; 10k; 1k}	idem	idem
# Effective SNPs	$\lvert\mathcal{S}\rvert = S =$	100k	idem	idem
# Effective causal SNPs**	$\lvert\mathcal{M}\rvert = M =$	1k	idem	idem
SNP-based heritability***	$h^2_{\text{SNP}} \in$	$\{0\%, 1\%, \ldots, 100\%\}$	idem	idem
CGR***	$\rho_{\mathbf{G}} \in$	$\{0, 0.01, \ldots, 1\}$	idem	idem
Allele frequency SNP $k \in \mathcal{S}$	$f_k \sim$	$\mathcal{U}(0.05, 0.95)$	$\mathcal{U}(0.05, 0.95)$	*Beta*(0.35,0.35)****
Genotype k, individual i, study j	$G_{jik} \sim$	$\mathit{Binom}(2, f_k)$	idem	idem
Effect SNP $k \notin \mathcal{M}$, j	$\beta_{jk} =$	0	idem	idem
Effect SNP $k \in \mathcal{M}$, j	$\beta_{jk} \sim$	$\mathcal{N}(0,1)$	idem	idem
Correlation SNP effect k, $j \neq h$	$\text{corr}(\beta_{jk}, \beta_{hk}) =$	$\rho_{\mathbf{G}}$	idem	idem
Residual i, j	$\varepsilon_{ji} \sim$	$\mathcal{N}\left(0, 1-h^2_{\text{SNP}}\right)$	idem	idem
Genetic value i, j	$GV_{ji} =$	$\sum_{k \in \mathcal{M}} \frac{G_{jik}-2f_k}{\sqrt{2f_k(1-f_k)}} \beta_{jk}$	$\sum_{k \in \mathcal{M}} (G_{jik} - 2f_k)\beta_{jk}$	$\sum_{k \in \mathcal{M}} (G_{jik} - 2f_k)\beta_{jk}$
GV coefficient j	$c_j =$	$\sqrt{h^2_{\text{SNP}}/M}$	$\sqrt{\frac{h^2_{\text{SNP}}}{\frac{1}{N_j}\sum_{i=1}^{N_j} GV^2_{ji}}}$	$\sqrt{\frac{h^2_{\text{SNP}}}{\frac{1}{N_j}\sum_{i=1}^{N_j} GV^2_{ji}}}$
Phenotype i, j	$Y_{ji} =$	$GV_{ji}c_j + \varepsilon_{ji}$	idem	idem
Number of runs	$R =$	100	idem	idem

* Correlations between all random quantities are zero unless otherwise specified

** Set of effective causal SNPs $\mathcal{M} \subset \mathcal{S}$, the set of effective SNPs

*** For each combination of values of h^2_{SNP} and CGR simulation analyses are performed

**** The *Beta*(0.35,0.35) distribution fits the empirical distribution of allele frequencies well

samples for discovery as well as a fourth sample used as hold-out sample for prediction. The sample sizes of the respective studies are given by $N_1 = 20,000$, $N_2 = 15,000$, $N_3 = 10,000$, and $N_4 = 1,000$, where N_4 denotes the sample size of the hold-out sample. An 11×11 grid of equispaced values of $h^2 \in [0,1]$ and $\rho_{\mathbf{G}} \in [0,1]$ is considered. In all simulations there are $S = 100,000$ independent SNPs of which $M = 1,000$ have a causal influence. A detailed description of the data-generating process in each simulation study can be found in Table 3.

For every run, data is simulated in accordance with the underlying data-generating process. Next, a GWAS is carried out in each of the three discovery samples. GWAS results are then meta-analyzed using sample-size weighting. The fraction of causal SNPs reaching genome-wide significance in the meta-analysis is the estimate of statistical power per SNP. The squared correlation between the meta-analysis-based PGS for the hold-out sample and the corresponding phenotype is the estimate of the PGS R^2.

Final estimates of power per causal SNPs and PGS R^2 are obtained by averaging the estimates across the runs. Fig. 6 shows the resulting estimates of power per causal SNP in the meta-analysis and the R^2 of the PGS for Simulations 1–3. In addition, the figure reports the power per causal SNP and R^2 predicted by the theoretical model. Inspection of these figures shows that there is no qualitative difference between the plots. Moreover, when computing the root-mean-square error (RMSE) between the theoretical predictions and the simulation-based estimates of power and R^2, even for the most extreme departure from the assumptions

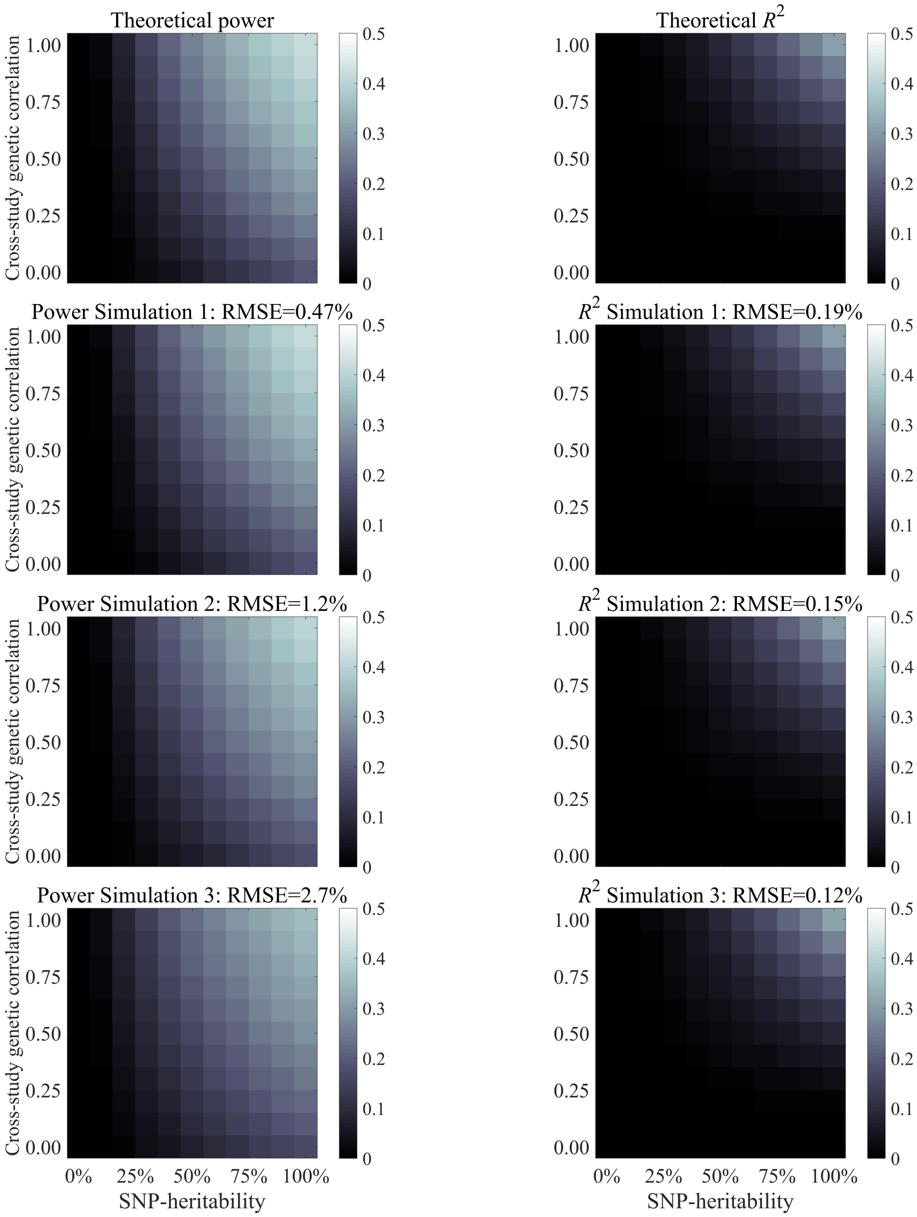

Figure 6. Power and polygenic score R^2 plots, with in each plot h^2 on the x-axis and cross-study genetic correlation on the y-axis. The first row shows predictions from the theoretical model. Subsequent rows show estimates based on respective simulation studies. The first column shows power per causal SNP. The second column the R^2 of a polygenic score in a hold-out sample. Above each plot, the root-mean-square error (RMSE) is reported for the difference between predictions from the theoretical model and the simulation-based estimates.

underlying our theory (Simulation 3), the RMSE in power remains below 3% and the RMSE in R^2 of the PGS below 0.2%.

S4 Data and Quality Control

Genotype data In the bivariate and univariate genomic-relatedness-matrix restricted maximum likelihood (GREML) analyses we use genotype data from the Rotterdam Study (RS; Ergo waves 1-4 sample denoted by RS-I, Ergo Plus sample denoted by RS-II, and Ergo Jong sample denoted by RS-III), the Swedish Twin Registry (STR; TwinGene sample), and the Health and Retirement Study (HRS). For each study, details on the genotyping platform, quality control (QC) prior to imputation, the reference sample used for imputation, and imputation software, are listed in Table 4.

To increase the overlap of SNPs across studies, we use genotypes imputed on the basis of the 1000 Genomes, Phase 1, Version 3 reference panel [48]. We only consider the subset of HapMap3 SNPs available in the 1kG data. By using this subset we substantially reduce the computational burden of the analyses, while preserving overlap between the SNP-sets in the studies and still having a sufficiently dense set of both common and more rare SNPs (# SNPs after QC $\approx$ 1 million).

Table 4. Genotyping and imputation

Study	Genotyping platform	SNP exclusions			Subject exclusions*	Imputation**
		MAF <	Call rate <	HWE p-val. <	Call rate <	Software
RS-I	Illumina 550K	0%	97.5%	10^{-7}	97.5%	MaCH/Minimac
RS-II	Illumina 550K	0%	97.5%	10^{-7}	97.5%	MaCH/Minimac
RS-III	Illumina 610K	0%	97.5%	10^{-7}	97.5%	MaCH/Minimac
STR	HumanOmniExpress 12v1A	1%	97.0%	10^{-7}	97.0%	MaCH/Minimac
HRS	Illumina Omni2.5	1%	98.0%	10^{-4}	98.0%	IMPUTE2

* Individuals are also excluded on the basis of sex mismatch, close relatives, duplicates and ancestry outliers (STR excepted), or autosomal heterozygosity outliers (HRS excepted)
** All samples have been imputed against the 1000Genomes, Phase 1, Version 3 haplotypes of all ancestries.

Quality control Prior to QC, we extract HapMap3 SNPs (source: `http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/hapmap3_r3/plink_format/`, accessed: December 11, 2014) from the imputed genotype data of each study and convert the allele dosages to best-guess `PLINK` [49], [50] binary files by rounding dosages using `GCTA` [31]. Subsequently, we perform QC on the best-guess genotypes in two stages. In the first stage, we clean and harmonize the imputed genotype data at the study level. The cleaned and harmonized study genotypes are then merged into a pooled dataset. The second round of QC is aimed at cleaning the pooled dataset, on the basis of the samples for which the phenotype is available. Hence, the first QC stage is phenotype-independent, whereas the second stage depends on the phenotype of interest.

In the first QC stage (prior to merging), we filter out the following markers and individuals:

1. SNPs with imputation accuracy below 70%.
2. Non-autosomal SNPs.
3. SNPs with minor-allele frequency below 1%.
4. SNPs with Hardy-Weinberg-Equilibrium p-value below 1%.
5. SNPs with missingness greater than 5%.
6. Individuals with missingness greater than 5%.
7. SNPs that are not present in all studies.
8. SNPs whose alleles cannot be aligned across studies.

Prior to the first QC stage, we apply the following two additional steps in HRS:

1. Switch alleles to address a strand-flip error due to incorrect annotation.
2. Drop individuals of non-European ancestry.

After the first round of QC, a set of roughly 1 million overlapping SNPs, available for about 30,000 individuals is left. Panel I in Table 5 shows, for each study, the number of SNPs and individuals before and after the first round of QC.

The second QC stage, applied to the pooled data set, comprises the following steps:

1. Keep only individuals for whom the phenotype of interest and all corresponding control variables are available.
2. Drop SNPs with a minor-allele frequency below 1%.
3. Drop SNPs with Hardy-Weinberg-Equilibrium p-value below 1%.
4. Drop SNPs with missingness greater than 5%.
5. Drop individuals with missingness greater than 5%.
6. Keep only one individual per pair of individuals with a genomic relatedness greater than 0.025.

Since the data in STR consists of twins and having highly related individuals can bias estimates of SNP-based heritability due to environment-sharing, we randomly select only one individual per twin pair after Step 1 in the second QC stage.

Panel II in Table 5 shows the sample size and the number of SNPs in the pooled dataset for each phenotype. We only consider phenotypes that attain a sample size of at least 18,000 individuals after all QC steps. The lowest sample size after QC is 19,184 for self-rated-health and the highest is 20,686 for *CurrCigt*. For all phenotypes, the number of SNPs is slightly greater than one million.

Table 5. Number of individuals and SNPs before and after quality control (QC) at the study level (Panel I) and at the pooled level (Panel II).

Panel I: study-level QC

Study	*N*		# SNPs	
	pre-QC	post-QC	pre-QC	post-QC
RS-I	6,291	6,291	31,337,615	1,062,589
RS-II	2,157	2,157	31,337,615	1,062,589
RS-III	3,048	3,048	31,337,615	1,062,589
STR	9,617	9,617	31,326,389	1,062,589
HRS	12,454	8,652	21,632,048	1,062,589
Total		29,765		1,062,589

Panel II: pooled-level QC

Phenotype	*N*		# SNPs	
	pre-QC	post-QC	pre-QC	post-QC
Height	29,765	20,458	1,062,589	1,052,572
BMI	29,765	20,449	1,062,589	1,052,600
EduYears	29,765	20,619	1,062,589	1,052,626
CurrCigt	29,765	20,686	1,062,589	1,052,524
CurrDrinkFreq	29,765	20,072	1,062,589	1,052,958
Self-rated health	29,765	19,184	1,062,589	1,053,190

Phenotype data For HRS, we use the RAND HRS data, version N, to obtain the phenotypes of interest. These data consist of measurements from eleven waves. RS-I consists of four data waves (Ergo 1-4). In both HRS and RS-I, data for some phenotypes are only available in a subset of the waves. RS-II, RS-III and STR do not have multiple measures over time for the phenotypes considered in this study. Table 6 describes how the phenotypes are constructed in each of the five studies.

As Table 6 shows, height, BMI, *EduYears*, and *CurrCigt* are measured quite consistently across waves. The self-rated health phenotype is also measured quite consistently, although in RS respondents are asked about health compared to members of the same age group, whereas a more absolute question is posed in STR and HRS. The drinking measure *CurrFreqDrink* is also measured somewhat heterogeneously; the threshold for what we treat as 'frequent drinking' is determined solely by how fine-grained the drinking frequency measure is in the respective studies.

Table 6. Study-level phenotype measures.

Phenotype	Survey instrument in				
	RS-I	RS-II	RS-III	STR	HRS
Years of education (*EduYears*)	Constructed in line with [45] in all studies.				
Height	Median height across waves 1-4.	Height	Height	Height	Median height across waves 1-11
BMI	Median BMI across waves 1-4.	BMI	BMI	BMI	Median BMI across waves 1-11
Currently smoking cigarettes (*CurrCigt*)	1 if stated to be a current smoker of cigarettes in the latest available measurement across waves 1-4.	1 if stated to be a current cigarette smoker.	Same as RS-II.	1 if stated to be a current cigarette smoker.	1 if responded positively to “currently smokes cigarettes?” in the latest available measurement across waves 1-11.
Currently drinking frequently (*CurrDrinkFreq*)	1 if indicated to “drink one or more alcoholic beverages per week” in the latest available measurement across waves 1-4.	1 if indicated to “drink one or more alcoholic beverages per week”.	1 if indicated to “have drunk at least two alcoholic beverages a month during the the past year.”	1 if indicated to “have drunk at least two alcoholic beverages in the past month”.	1 if indicated to “drink alcohol once per week or more” in the latest available measurement across waves 3-11.
Self-rated health	Only available in wave 1: “How is your general health compared to members of your age group?” Response categories reverse-coded such that 0=worse, 1=same, and 2=better.	Same as RS-I.	*n.a.*	Rate their general health. Response categories re-coded such that 0=bad, 1=not so good, 2=average, 3=good, 4=excellent.	Mode of the 4-point self-reported health measure in HRS across waves 1-11. Responses reverse-coded such that 0=poor, 1=fair, 2=good, 3=very good, and 4=excellent.

S5 GREML Estimation

Height, BMI, *EduYears*, and self-rated health are treated as quantitative traits. *CurrCigt* and *CurrDrinkFreq* are treated as binary outcomes. In each study, (after aggregating across waves, if applicable) we regress quantitative phenotypes on age, squared age, sex, and an intercept. The residuals from the regression are standardized to have a sample-mean equal to zero and variance equal to one. For both binary and quantitative traits, the aforementioned covariates are also included in the GREML estimation. In addition, in bivariate GREML and pooled GREML estimation (i.e., considering multiple studies jointly), the intercept is replaced by indicator variables for the respective studies, capturing study-specific fixed effects. Finally, 20 principal components from the phenotype-specific genomic-relatedness matrix are added to the set of control variables in the GREML estimation, in order to correct for population stratification [51].

S6 GREML Results

Details per phenotype on sample size, univariate estimates of SNP-heritability, and bivariate estimates of genetic correlation, stratified across studies, and cross-study averages, are provided in Table 7. Results stratified across sexes are listed in Table 8.

Table 7. GREML estimates of SNP-heritability (h^2_{SNP}) and genetic correlation (ρ_{G}) across studies.

Phenotype	***N***				**Univariate estimates h^2_{SNP}[1]**				**Bivariate estimates ρ_{G}[1]**			
	RS	**STR**	**HRS**	**Total**	**RS**	**STR**	**HRS**	**Average[2]**	**RS–STR**	**RS–HRS**	**STR–HRS**	**Average[3]**
Height	6,780	5,342	8,336	20,458	48.9% (4.9%) ***	50.8% (6.0%) ***	37.9% (4.1%) ***	44.9%	0.976 (0.102) ***	0.954 (0.095) ***	0.967 (0.106) ***	0.965
BMI	6,775	5,341	8,333	20,449	28.9% (4.9%) ***	16.4% (6.1%) ***	19.6% (4.1%) ***	21.9%	1.000 (0.269) ***	0.914 (0.172) ***	0.847 (0.246) ***	0.917
EduYears	6,735	5,543	8,341	20,619	17.5% (4.8%) ***	20.6% (5.8%) ***	17.3% (4.0%) ***	18.2%	0.690 (0.233) ***	0.659 (0.224) ***†	1.000 (0.263) ***	0.783
CurrCigt	6,803	5,579	8,304	20,686	17.8% (10.1%) **	18.7% (13.8%) *	20.4% (11.2%) **	19.1%	1.000 (0.643) ***	0.611 (0.448) *	1.000 (0.607) ***	0.858
CurrDrinkFreq	6,172	5,564	8,336	20,072	13.5% (8.7%) *	14.1% (9.5%) *	5.3% (6.3%)	10.3%	1.000 (0.666) ***	0.298 (0.670)	-0.056 (0.647)	0.381
Self-rated health	5,264	5,577	8,343	19,184	13.5% (6.2%) **	9.4% (5.7%) **	21.3% (4.0%) ***	15.7%	0.626 (0.439) **	0.363 (0.223) **††	0.447 (0.278) **	0.468

* h^2_{SNP} and/or genetic correlation > 0 at 10% sign. †genetic correlation < 1 at 10% sign. ‡genetic correlation < 0 at 10% sign.
** h^2_{SNP} and/or genetic correlation > 0 at 5% sign. ††genetic correlation < 1 at 5% sign. ‡‡genetic correlation < 0 at 5% sign.
*** h^2_{SNP} and/or genetic correlation > 0 at 1% sign. †††genetic correlation < 1 at 1% sign. ‡‡‡genetic correlation < 0 at 1% sign.

[1] Standard errors between parentheses.
[2] Sample-size weighted average of univariate estimates across studies.
[3] Sample-size weighted average of bivariate estimates across pairs of studies.

Table 8. GREML estimates of SNP-heritability (h^2_{SNP}) and genetic correlation (ρ_{G}) across sexes.

Phenotype	***N***			Estimates h^2_{SNP} [1]			Estimate ρ_{G} [1]
	Females	**Males**	**Total**	**Females**	**Males**	**Average** [2]	**Females–Males**
Height	11,553	8,905	20,458	43.2% (3.0%) ***	45.1% (3.8%) ***	44.0%	0.981 (0.067) ***
BMI	11,542	8,907	20,449	22.1% (2.9%) ***	23.8% (3.8%) ***	22.8%	0.794 (0.122) *** †
EduYears	11,653	8,966	20,619	18.1% (2.9%) ***	18.9% (3.7%) ***	18.4%	0.832 (0.162) ***
CurrCigt	11,706	8,980	20,686	22.3% (7.1%) ***	26.7% (9.1%) ***	24.2%	0.543 (0.257) *** †
CurrDrinkFreq	11,312	8,760	20,072	14.1% (4.6%) ***	0.9% (6.0%)	8.3%	1.000 (2.068) *
Self-rated health	10,866	8,318	19,184	8.6% (3.1%) ***	10.8% (4.0%) ***	9.5%	1.000 (0.349) ***

* h^2_{SNP} and/or genetic correlation > 0 at 10% sign. †genetic correlation < 1 at 10% sign. ‡genetic correlation < 0 at 10% sign.
** h^2_{SNP} and/or genetic correlation > 0 at 5% sign. ††genetic correlation < 1 at 5% sign. ‡‡genetic correlation < 0 at 5% sign.
*** h^2_{SNP} and/or genetic correlation > 0 at 1% sign. †††genetic correlation < 1 at 1% sign. ‡‡‡genetic correlation < 0 at 1% sign.

[1] Standard errors between parentheses.
[2] Sample-size weighted average of univariate estimates across studies.

Acknowledgments

RS (Rotterdam Study) The generation and management of GWAS genotype data for the Rotterdam Study is supported by the Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012). This study is funded by the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) project nr. 050-060-810. We thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera and Marjolein Peters for their help in creating the GWAS database, and Karol Estrada and Maksim V. Struchalin for their support in creation and analysis of imputed data. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. Researchers who wish to use data of the Rotterdam Study must obtain approval from the Rotterdam Study Management Team. They are advised to contact the PI of the Rotterdam Study, Dr. Albert Hofman (a.hofman@erasmusmc.nl).

STR (Swedish Twin Registry) The Jan Wallander and Tom Hedelius Foundation (P2012-0002:1), the Ragnar Söderberg Foundation (E9/11), The Swedish Research Council (421-2013-1061), the Ministry for Higher Education, The Swedish Research Council (M-2205-1112), GenomEUtwin (EU/QLRT-2001-01254; QLG2-CT-2002-01254), NIH DK U01-066134, The Swedish Foundation for Strategic Research (SSF). Researchers interested in using STR data must obtain approval from the Swedish Ethical Review Board and from the Steering Committee of the Swedish Twin Registry. Researchers using the data are required to follow the terms of an Assistance Agreement containing a number of clauses designed to ensure protection of privacy and compliance with relevant laws. For further information, contact Patrik Magnusson (Patrik.magnusson@ki.se).

HRS (Health and Retirement Study) The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. The genotyping was funded separately by the National Institute on Aging (RC2 AG036495, RC4 AG039029). Our genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the Genetics Coordinating Center at the University of Washington. Genotype data can be accessed via the database of Genotypes and Phenotypes (dbGaP, `http://www.ncbi.nlm.nih.gov/gap`, accession number phs000428.v1.p1). Researchers who wish to link genetic data with other HRS measures that are not in dbGaP, such as educational attainment, must apply for access from HRS. See the HRS website (`http://hrsonline.isr.umich.edu/gwas`) for details.

RAND HRS RAND HRS Data, Version N. Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration. Santa Monica, CA (September 2014). Researchers who wish to use the RAND HRS data need to register via the RAND website (`http://www.rand.org/labor/aging/dataprod/hrs-data.html`).

Dutch national e-infrastructure This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

Individual acknowledgements Niels Rietveld acknowledges funding from the Netherlands Organisation for Scientific Research (NWO Veni grant 016.165.004). Philipp Koellinger acknowledges funding from the European Research Council (ERC consolidator grant 647648 EdGe). All authors acknowledge the valuable feedback provided by Peter M. Visscher, Jian Yang, David A. Cesarini, and Daniel J. Benjamin.

Author contributions

Theoretic framework, online power calculator, and simulations: RdV; data preparation: RdV, AO, CAR, MJ, PM, AGU, FJAvR, AH; cross-study SNP-heritability and genetic correlation estimates: RdV, AO; writing first draft of manuscript: RdV, AO, PK; overseeing the study: PK, ART. All co-authors contributed to the writing of the paper.

References

1. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46:1173–1186.

2. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197–206.

3. Eeles RA, Kote-Jarai Z, Al Olama AA, Giles GG, Guy M, Severi G, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009;41:1116–1121.

4. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478:103–109.

5. Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–427.

6. Rietveld CA, Cesarini D, Benjamin DJ, Koellinger PD, De Neve JE, Tiemeier H, et al. Molecular genetics and subjective well-being. Proc Natl Acad Sci USA. 2013;110:9692–9697.

7. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016;forthcoming.

8. Okbay A, Baselmans BM, De Neve JE, Turley P, Nivard MG, Fontana MA, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. Nat Genet. 2016;forthcoming.

9. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.

10. Benjamin DJ, Cesarini D, Chabris CF, Glaeser EL, Laibson DI, Guðnason V, et al. The Promises and Pitfalls of Genoeconomics. Annu Rev Econom. 2012;4:627–662.

11. Maher B. The case of the missing heritability. Nature. 2008;456:18–21.

12. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–753.

13. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11:446–450.

14. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci USA. 2012;109:1193–1198.

15. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14:507–515.

16. Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. Nat Rev Genet. 2014;15:765–776.

17. Wray NR, Maier R. Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability. Curr Epidemiol Rep. 2014;1:220–227.

18. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42:565–569.

19. 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.

20. Wray NR, Lee SH, Kendler KS. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. Eur J Hum Genet. 2012;20:668–674.

21. Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet. 2013;45:984–994.

22. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet. 2013;93:42–53.

23. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15:335–346.

24. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: Designing rare variant association studies. Proc Natl Acad Sci USA. 2014;111:E455–E464.

25. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015;47:1114–1120.

26. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLOS Genet. 2013;9:e1003348.

27. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17:1520–1528.

28. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141:210–217.

29. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, Goddard ME, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet. 2012;44:247–250.

30. Evans DM, Purcell SM. Power calculations in genetic studies. Cold Spring Harb Protoc. 2012;2012:pdb.top069559.

31. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.

32. Purcell SM, Cherny SS, Sham PC, et al. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003;19:149–150.

33. Menashe I, Rosenberg PS, Chen BE. PGA: power calculator for case-control genetic association analyses. BMC Genet. 2008;9:36.

34. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLOS ONE. 2008;3:e3395.

35. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum Genet. 2012;131:747–756.

36. Yang J, Weedon MN, Purcell SM, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011;19:807–812.

37. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46:100–106.

38. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291–295.

39. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28:2540–2542.

40. Visscher PM, Hemani G, Vinkhuyzen AA, Chen GB, Lee SH, Wray NR, et al. Statistical power to detect genetic (co) variance of complex traits using SNP data in unrelated samples. PLOS Genet. 2014;10:e1004269.

41. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467:832–838.

42. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008;40:575–583.

43. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42:937–948.

44. Willer CJ, Speliotes EK, Loos RJF, Li S, Lindgren CM, Heid IM, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet. 2008;41:25–34.

45. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013;340:1467–1471.

46. Harris SE, Hagenaars SP, Davies G, Hill WD, Liewald DC, Ritchie SJ, et al. Molecular genetic contributions to self-rated health. bioRxiv. 2015;p. 029504.

47. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190–2191.

48. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.

49. Purcell SM, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–575.

50. Chang CC, Chow CC, Tellier L, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4.

51. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–909.