Title: Across-cohort QC analyses of genome-wide association study summary statistics from complex traits

Authors: Guo-Bo Chen[1], Sang Hong Lee[1,2], Matthew R Robinson[1], Maciej Trzaskowski[1], Zhi-Xiang Zhu[3], Thomas W Winkler[4], Felix R Day[5], Damien C Croteau-Chonka[6,7], Andrew R Wood[8], Adam E Locke[9], Zoltán Kutalik[10-12], Ruth J F Loos[13-15], Timothy M Frayling[8], Joel N Hirschhorn[16-19], Jian Yang[1,21], Naomi R Wray[1], The Genetic Investigation of Anthropometric Traits (GIANT) Consortium[20], Peter M Visscher[1,21]

Affiliations:

[1] Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia

[2] School of Environmental and Rural Science, The University of New England, Armidale, New South Walsh, Australia

[3] SPLUS Game, Guangzhou, Guangdong, China

[4] Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

[5] Medical Research Council (MRC) Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK

[6] Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA

[7] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[8] Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK

[9] Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA

[10] Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland

[11] Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

[12] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[13] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[14] The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[15] The Genetics of Obesity and Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[16] Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

[17] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[18] Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, USA

[19] Division of Endocrinology, Boston Children's Hospital, Boston, Massachusetts, USA

[20] A full list of members is available in the **Supplementary Note**

[21] The University of Queensland Diamantina Institute, Translation Research Institute, Brisbane, Queensland, Australia

Correspondence should be addressed to

GBC (chen.guobo@foxmail.com) or PMV (peter.visscher@uq.edu.au)

Running title: Four metrics for GWAMA

Keywords: meta-analysis, overlapping samples, meta-PCA, Fst, PPSR, GIANT, Metabochip

Abstract

Genome-wide association studies (GWASs) have been successful in discovering replicable SNP-trait associations for many quantitative traits and common diseases in humans. Typically the effect sizes of SNP alleles are very small and this has led to large genome-wide association meta-analyses (GWAMA) to maximize statistical power. A trend towards ever-larger GWAMA is likely to continue, yet dealing with summary statistics from hundreds of cohorts increases logistical and quality control problems, including unknown sample overlap, and these can lead to both false positive and false negative findings. In this study we propose a new set of metrics and visualization tools for GWAMA, using summary statistics from cohort-level GWASs. We proposed a pair of methods in examining the concordance between demographic information and summary statistics. In method I, we use the population genetics F_{st} statistic to verify the genetic origin of each cohort and their geographic location, and demonstrate using GWAMA data from the GIANT Consortium that geographic locations of cohorts can be recovered and outlier cohorts can be detected. In method II, we conduct principal component analysis based on reported allele frequencies, and is able to recover the ancestral information for each cohort. In addition, we propose a new statistic that uses the reported allelic effect sizes and their standard errors to identify significant sample overlap or heterogeneity between pairs of cohorts. Finally, to quantify unknown sample overlap across all pairs of cohorts we propose a method that uses randomly generated genetic predictors that does not require the sharing of individual-level genotype data and does not breach individual privacy.

Introduction

Genome-wide association studies (GWASs) have been successful in discovering SNP-trait associations for complex traits[1]. To elucidate genetic architecture, which requires maximized statistical power for discovery of risk alleles of small effect, large genome-wide association meta-analyses (GWAMA) are tending towards ever-larger scale that may contain data from hundreds of cohorts. At the individual cohort level, GWAS analysis is often based on various genotyping chips and conducted with different protocols, such as different software tools and reference populations for imputation, inclusion of study specific covariates and association analyses using different methods and software. Although solid quality control analysis pipelines of GWAMA exist[2], these analyses focus on quality control (QC) for each cohort independently. With ever-increasing sizes of GWAMA there is a need for additional QC that goes beyond the cohort-by-cohort genotype-level analysis performed to date.

In this study, we propose a new set of QC metrics for GWAMA. In contrast to previous QC metrics, our approach explores the genetic and QC context of the all cohorts in GWAMA together rather than by treating them one at a time. These metrics include

(i) a genome-wide comparison of allele frequency differences across cohorts or against a common reference population

(ii) principal component analysis for reported allele frequencies

(iii) a pairwise cohort statistic that uses allele frequency or effect size concordance to detect the proportion of sample overlap or heterogeneity

(vi) an easy to implement analysis to pinpoint each between-cohort overlapping sample that does not require the sharing of individual-level genotype data.

All these applications assume that there is a central analysis hub where summary statistic data from GWAS are uploaded for each cohort. In addition, these metrics reveal information of interest other than merely QC.

Materials and Methods

Overview of materials

Cohort-level summary statistics. The GWAS height GWAS summary statistics were provided by the GIANT Consortium and were from 82 cohorts (174 separate files due to different ways a cohort was split into different sexes, different disease statuses) representing a total of 253,288 individuals, and nearly 2.5 million autosome SNPs imputed to the HapMap2 reference[3]. The Metabochip summary statistics were for body mass index (BMI) from 43 cohorts (120 files due to different ways a cohort was split into different sexes, different disease statuses) representing a total of 103,047 samples from multiple ethnicities with about 200,000 SNPs genotyped on customised chips[4,5]. For convenience, we consider each file a cohort. All the summary statistics have already been cleaned using established protocols for GWAS meta-analysis[2].

1000 Genomes Project samples. 1000 Genomes Project (1KG) reference samples[6] were used as the reference samples for calculating F_{st}. When assessing the global-level F_{st} measures, Yoruba represent African samples (YRI, 108 individuals), Han Chinese in Beijing represent East Asian samples (CHB, 103 individuals), and Utah Residents with Northern and Western European Ancestry represent European samples (CEU, 99 individuals) were employed as the reference panels. For calculating within-Europe F_{st}, CEU, Finnish (FIN, 99 individuals), and Tuscani (TSI, 107 individuals) were employed to represent northwest, northeast and southern Europeans, respectively. For analyses using a whole European panel, CEU, FIN, TSI, GBR (British, 91 individuals), and IBS (Iberian, 107 individuals) were pooled together as an “averaged” European reference.

WTCCC GWAS data. WTCCC GWAS data has 2,934 shared controls for 7 diseases with a total of 14,000 cases[7]. Individual GWAS was conducted for each disease using PLINK[8], and their summary statistics used to estimate λ_{meta} (see text below). WTCCC GWAS data were also used for demonstrating pseudo profile score regression (see text below).

Simulated cohort-level summary statistics. M independent loci were generated for cohort-level summary statistics. Each locus had allele frequency p_i, which was sampled from a uniform distribution ranging from 0.1 to 0.5, and had genetic effect b_i, sampled from a standard normal distribution $N(0,1)$. After rescaling, $\Sigma_{i=1}^{M} 2p_i(1-p_i)b_i^2 = h^2$. p and b were treated as true parameters. For a particular cohort with n samples, its $\tilde{p}_i \sim N(p_i, \frac{p_i(1-p_i)}{2n})$, $\tilde{b}_i \sim N(b_i, \frac{1}{2np_i(1-p_i)})$, and the sampling variance for $\tilde{b}_i$ is $\sigma^2_{\tilde{b}_i} = \frac{1}{2n_i p_i(1-p_i)}$. All cohorts were assumed to share common genetic architecture, and differences were only due to genetic drift, allele frequencies and sampling variance of genetic effects.

Overview of the methods

F_{st}-based genetic distance between cohorts. For a cohort, its F_{st} with reference cohorts, such as CEU, YRI, and CHB, is calculated. Given those three F_{st} values, the coordinate of this cohort can be uniquely projected into the reference equilateral that has CEU, YRI, and CHB at its corners.

Principal component analysis for cohort-level allele frequencies. A genetic relationship matrix for cohorts can be constructed based on received allele frequencies. Principal component analysis (PCA) can be implemented on the genetic relationship matrix. The projection of the cohorts into PCA space can reveal the genetic background and relative geographical distance between cohorts.

λ_{meta} for detecting overlapping samples. In concept, λ_{meta} resembles λ_{gc}, which indicates population stratification for a GWAS[9], but λ_{meta} measures the proportion of overlapping samples between a pair of cohorts. Based on reported genetic effects and their sampling variance, λ_{meta} can be constructed for a pair of cohorts and follows a chi-square distribution with 1 degree of freedom. λ_{meta} will be close to 1 when there is no overlapping samples, smaller than 1 when there are overlapping samples, and greater than 1 when there are heterogeneity between a pair of cohorts. For GWAMA over a single trait across method, we assume heterogeneity is zero.

Pseudo profile score regression for pinpointing overlapping samples/relatives. Pseudo profile score regression (PPSR) provides a framework for pinpointing the overlapping samples/relatives between cohort without sharing genotypes. Each GWAS analyst generates pseudo profile scores (PPS) for each sample on a set of loci, which are chosen by a GWAMA central analyst. If the similarity metric of PPS for a pair of cohorts reaches a similarity threshold, say 1 overlapping samples and 0.5 for first-degree relatives, then overlapping samples/relatives are found. PPSR can have a controlled type-I and type-II error rates in pinpointing overlapping samples, and also can reduce the comprise of privacy. PPSR is an enhanced version of Gencrypt[10], a previous method in pinpointing overlapping samples.

The technical details of these four methods can be found in the Supplementary notes.

Results

Population genetic quality control analysis using F_{st}

Allele frequency differentiation among populations reflects population characteristics such as demographic past and geographic locations[11,12]. In GWAMA only summary statistics such as allele frequencies are available to the central analysis hub, and so it is not possible to run principal component analysis for each cohort that requires individual-level data. Therefore it is difficult to quantify genetic distance between cohorts or to a reference in order to identify population outliers. Outlier cohorts can be due to real differences in ethnicity or mistakes in the primary analysis prior to uploading data to the GWAMA analysis hub. Gross differentiation in allele frequencies at specific SNPs between GWAMA cohorts and a reference (such as 1000 Genomes Project, denoted as 1KG)[6] are part of standard QC protocols[2] but checking for more differentiation than expected across the entire genome is not usually part of the QC pipeline. We propose that a genetic distance inferred from F_{st}, which reflects genetic distance between pairwise populations, is a useful additional QC statistic to detect cohorts that are population outliers. Using the relationship between F_{st} and principal components[13–15], our F_{st} Cartographer algorithm can be used to estimate the relative genetic distance between cohorts (**Supplementary notes, and Supplementary Fig. 1**).

We applied the F_{st} metric to the GIANT Consortium body mass index (BMI) Metabochip cohorts (55 male-only cohorts, 55 female-only cohort, and 10 mixed-sex cohorts; for convenience, we called each file a cohort), which were recruited from multiple ethnicities[4], such as Europeans, African Americans in The

Atherosclerosis Risk in Communities Study (ARIC) and cohorts from Jamaica (SPT), Pakistan (PROMISE), Philippines (CLHNS) and Seychelles (SEY). For each Metabochip cohort, we sampled 30,000 (see Online method for details) independent markers to calculate F_{st} values with each of three 1KG samples (CEU, CHB, and YRI, respectively). For validation of the method, we also calculated F_{st} values against the 1KG Japanese (JPT, Japanese in Tokyo, Japan), Indian (GIH, Gujarati Indian in Houston, US), Kenyan (LWK, Luhya in Webuye, Kenya) and European samples (IBS, Iberian populations, Spain; FIN, Finnish, Finland; TSI, Toscani, Italy, and GBR, British in England and Scortland, GBR), to see whether the known genetic origins of those cohorts can be recovered.

According to the origins of the samples, each Metabochip cohort showed a different genetic distance spectrum to the three reference populations (**Fig. 1a**). The JPT and Philippine cohorts had very small genetic distances to CHB, as expected, but large to CEU and YRI; however, the Pakistan cohorts showed much closer genetic distances to CEU than to CHB and YRI, indicating their demographic history. The cohorts sampled from Jamaica, Seychelles, Hawaii, and the African American ARIC cohort had small genetic distances to YRI, but large distances to CHB and CEU. For most European cohorts, as expected, the distances to CEU were very small compared with those to CHB and YRI. Given their relative distances to CEU, CHB, and YRI, using our F_{st} cartographer algorithm (**Supplementary notes, and Supplementary Fig. 1**), the cohorts were projected into a two-dimensional space, called F_{st} derived principal components (F_{PC}) space, constructed by YRI, CHB, and CEU as the reference populations (**Fig. 1b**). The allocation of the cohorts to the F_{PC} space resembles that of eigenvector 1 against eigenvector 2 in principal component analysis (PCA)[12], and is similar to those observed in PCA using individual-level GWAS data for populations of various ethnicities such as in 1KG samples[6]. Therefore, our method to place cohorts in geographical regions from GWAS summary statistics works well at a global-population scale.

We next investigated whether our genetic distance method works at a much finer geographic scale. It is known that using individual-level data, principal component analysis can mirror the geographic locations for European samples[11]. Here, we analyzed the 103 GIANT European-ancestry Metabochip cohorts (48 male-only cohorts, 47 female-only cohorts, and 8 mix-sex cohorts) for fine-scale F_{st} genetic distance measure by using the CEU, FIN, and TSI reference populations, which represent northwest, northeast, and southern European populations, respectively. For each of the GIANT European-ancestry Metabochip cohorts, F_{st} was calculated relative to each of these three reference populations and showed concordance with the known origin of the samples (**Fig. 1c**). For example, cohorts from Finland and Estonia were close to FIN but distant to TSI; cohorts from South Europe such as Italy and Greece had small genetic distance to TSI; and cohorts from West European nations had small genetic distance to CEU. Similarly, the projected origin for each European-ancestry Metabochip cohort resembles their geographic location within the European map as expected (**Fig. 1d**). Therefore, our QC measure based upon population differentiation also works at a fine scale.

We next applied the F_{st} genetic distance measures to 174 GIANT height GWAS cohorts (79 male-only cohorts, 76 female-only cohorts, and 19 mixed-sex cohorts; excluding Metabochip data), which were all of European ancestry imputed to the HapMap reference panel[3]. Given the three F_{st} values to CEU, FIN, and TSI (**Fig. 2a**), the geographic origin for each cohort can be inferred as for the GIANT BMI Metabochip data (**Supplementary notes**). The projected coordinates of each GWAS cohort matches its origin very well (**Fig. 2b**). For example, a Canadian cohort, the Quebec Family Study (QFS), was closely located to DESIR, a French cohort, consistent with the French genetic heritage of the QFS[16]. In addition, we also observe complexity due to mixed samples from different countries. For example, the DGI/Botnia study had samples recruited from Sweden and Finland, and its inferred geographic location is in between of the Swedish cohorts and Finnish cohorts[17]. We also note that for the MIGEN consortia cohorts, which are from Finland, Sweden, Spain and the US, the same allele frequencies were reported for all their sub-cohorts, and all cohorts were allocated to southern Europe (very closely located to 1KG IBS cohort; **Fig. 2b and Supplementary Fig. 2**). As the allele frequencies, used in QC steps to eliminate low quality loci, were not directly used in estimating genetic effects in the GWAMA, the reported allele frequencies in MIGEN have not impacted on the published GWAMA results[3].

Next, we show that F_{st} can detect populations that have a different demographic past. Using all 1KG European samples as the reference panel (that is, an “averaged” European reference panel), most cohorts in GIANT had $F_{st} < 0.005$ with this average, which agrees with previously reported results using individual level data from European nations[11]. A few cohorts showed large F_{st}, such as the AMISH cohort with $F_{st} = 0.018$, and the North Swedish Population Health Study (NSPHS)[18] with $F_{st} = 0.014$. Consistent with these results, both these populations are known to have been genetically isolated (**Supplementary Fig. 3**).

Principal component analysis for allele frequencies

It is well established that given individual-level data principal component analysis (PCA) can reveal the ancestral information for samples[12]. Given the same allele frequencies as used for F_{st}-based analysis above, we conducted PCA for allele frequencies, denoted as meta-PCA. In meta-PCA each cohort was analogously considered as an “individual”. For example, 120 Metabochip cohorts were considered as a sample of 120 “individuals”. Although the inferred ancestral information was for each cohort rather than any individuals, implementation of meta-PCA was the same as the conventional PCA (**Supplementary Notes**).

Meta-PCA was tested with 1KG samples over nearly 1 million SNPs. The cohort-level allele frequencies were calculated first for 26 1KG cohorts, and meta-PCA was conducted. The projected cohorts were consistent to their genetic origin (**Fig. 3**). In contrast, conventional PCA was also conducted on 1KG individual genotypes directly, and the mean coordinates for each cohort was then calculated. As illustrated in Fig. 3, these two techniques resulted in nearly identical projection for 1KG, and the correlation between cohort coordinates remained consistently high for the first eight eigenvectors, $R^2 > 0.8$. It indicated that meta-PCA could reveal genetic background for each cohort as precise as that based on individual-level data.

We applied meta-PCA to 120 Metabochip cohorts for nearly 34 thousand common SNPs between Metabochip and 1KG variants, with the inclusion of 10 1KG cohorts (East Asian: CHB, JPT; South Asian: GIH; European: CEU, FIN, GBR, IBS, TSI; African: LWK, YRI) as the reference cohorts. Consistent with demographic information, the inferred ancestral information of each cohort agreed well with demographic information. For example, PROMISE (Pakistan) located very close to GIH, CLHNS (Philippines) close to CHB and JPT, ARIC (African American) and SPT (Jamaican) close to YRI and LWK, and the European cohorts close to CEU and FIN (**Fig 4**).

We also applied meta-PCA to 174 GIANT height GWAS cohorts for nearly 1M SNPs, with the inclusion of 10 1KG reference cohorts. At the global-population level, the 174 cohorts were all allocated close to CEU and FIN, consistent with their reported demographic information (**Fig. 5**). For fine-scale inference, we conducted meta-PCA again but with the inclusion of the five European samples. As demonstrated, the resolution of the inferred relative location between European cohorts reflected their real geographical locations, as previously observed using individual-level data[11].

These results were consistent to what observed from F_{pc} as described in the last section, and also agreed well with demographic information. So, based on the reported allele frequencies, the demographic information could be examined by meta-PCA method.

$\boldsymbol{\lambda_{meta}}$ to detect pairwise cohort heterogeneity and sample overlap

For a single cohort GWAS, λ_{GC} provides a tool for assessing average trait-SNP associations in GWAS[9], and an value departing from 1 may indicate undesired phenomena such as population stratification. In this study, we use the summary statistics for a pair of cohorts to calculate λ_{meta}, a metric that examines heterogeneity from the concordance of reported effect sizes and sampling variance. We use 30,000 markers in linkage equilibrium along the genome between a pair of cohorts to estimate λ_{meta}.

For a SNP marker (i), given its reported estimated effect size (b_i) and sampling variance (σ_i^2) in a pair of cohorts 1 and 2, we can calculate a test statistic $T_i = \frac{(b_{1.i}-b_{2.i})^2}{\sigma_{1.i}^2+\sigma_{2.i}^2}$, the ratio between the squared difference of their reported effects to the sum of their reported sampling variances. Under the null hypothesis of no overlapping samples/heterogeneity, T follows a chi-square distribution with 1 degree of freedom

(**Supplementary notes**). $\lambda_{meta} = \frac{median(T)}{median(\chi_1^2)}$, the ratio between the median of the 30,000 T values and the median of a chi-square statistic with 1 degree of freedom (a value of 0.455), has an expected value of 1 for two independent GWAS summary statistics sets for the same trait. When there is heterogeneity between estimated genetic effects, the expectation is $\lambda_{meta} > 1$, and in contrast $\lambda_{meta} < 1$ if there are overlapping samples. In general, not only overlapping samples but also close relatives present in different cohorts can lead to correlated summary statistics generating $\lambda_{meta} < 1$ (**Supplementary notes**). However, unless the proportion of overlapping relatives is substantial and their phenotypic correlation is high, the correlation of the summary statistics due to the effective number of overlapping samples (n_o) is expected to be dominated by the same individuals contributing phenotypic and genetic information to different cohorts (**Supplementary Fig. 4**). Furthermore, if genomic control is applied to adjust the sampling variance[19] then λ_{meta} will be reduced relative to its value without genomic control (**Supplementary notes**).

We estimated λ_{meta} from published GWAS summary statistics for a range of traits (other than BMI and height) and were able to find examples of both deflated and inflated λ_{meta}. First, we tested the λ_{meta} on data sets with known overlap. For example, GWAS summary statistics for schizophrenia were available in two phases: the first had 9,394 controls and 12,462 cases[20], and in the next phase about 18,000 Swedish samples were added[21]. Such a substantial overlap sample between these two sets of summary statistics led to the estimated value of λ_{meta} as low as 0.257 (**Supplementary Fig. 5**), consistent with this known overlap. In contrast, heterogeneity between data sets (represented by $\lambda_{meta} > 1$), was observed between GWAS summary statistics of rheumatoid arthritis from European and Asian studies[22], for which $\lambda_{meta} = 1.09$ (**Supplementary Fig. 6**). In addition, we note that the distribution of the empirical T-statistics deviates from expectation at the upper tail of the distribution, suggesting differences in effect size or linkage disequilibrium between these two ancestries.

Next, we estimated λ_{meta} from pairs of cohorts from the 174 GIANT height GWAS cohort[3]. We found no evidence for substantial sample overlap but do observe between-cohort heterogeneity, and technical artifacts. From the 174 GIANT height GWAS (supplied data files)[3], we calculated 15,051 cohort-pairwise λ_{meta} values, resulting in a bell-shape distribution (**Fig. 6a,b**) with the mean of 1.013 and the empirical standard deviation (S.D.) of 0.022, which was greater than theoretical S.D. of 0.014. The empirical mean and S.D can be used to construct a z-score test for each λ_{meta}. These results are consistent with a small amount of heterogeneity, which is not unexpected due to variation of actual (unknown) genetic architecture and analysis protocols. However, the mean is close to 1.0 and based upon this QC metric the results are consistent with stringent quality control and data cleaning. The minimum λ_{meta} value was around 0.88 (between SORBS MEN and SORBS WOMEN, **Fig. 3c**), with *p*-value < 1e-10 (testing for the difference from 1), and the maximum was 1.245 (between SardiNIA and WGHS, **Fig. 6d**), with *p*-value < 1e-10, leading to the most deflated and inflated λ_{meta} across GIANT height study cohorts; both were significant after correction for multiple testing. Illustrating λ_{meta} (**Fig. 6b**) highlighted that 20 cohorts from the MIGEN consortium showed substantially lower λ_{meta} with many other cohorts (right-bottom triangle in **Fig. 6b**) than the average, consistent with over-conservative models for statistical association analyses being used in these cohorts – which may be due to very small sample size (ranging from 36 to 320 for the 20 MIGEN cohorts, with an average sample size of 132). Consistent with this, cohorts from MIGEN also have many of their $\lambda_{GC} < 1$ (**Fig. 7a**). In contrast, the SardiNIA cohort (4,303 samples) showed heterogeneity with nearly all other cohorts (**Fig. 7b**), perhaps due to unknown artifacts or a slightly different genetic architecture for height as result of demographic history[23].

We investigated the relationship between $\bar{\lambda}_{meta}$ (the mean of all λ_{meta} values of a given cohort with each of the other 173 GIANT height cohorts) and λ_{GC} among the GIANT height cohorts. If there are no technical issues, such as inflated or deflated sampling variance for the estimated effects, we would expect to see: i) a correlation between λ_{GC} and sample size; ii) no correlation between $\bar{\lambda}_{meta}$ and sample size; iii) no correlation between $\bar{\lambda}_{meta}$ and λ_{GC} (**Supplementary Fig. S7**). Consistent with a previous study[24], for a polygenic trait such as height λ_{GC} of each cohort was related to its sample size (correlation of 0.235, p = 0.0018). In contrast, the correlation between $\bar{\lambda}_{meta}$ and sample size was of 0.116 (p = 0.127) (**Fig 7a,b**). Nevertheless, the correlation between the mean of $\bar{\lambda}_{meta}$ and λ_{GC} was 0.836 (p<10e-16) for 174 GIANT height cohorts (**Fig 7c**). We note that the 20 MIGEN cohorts had proportionally small λ_{GC} and $\bar{\lambda}_{meta}$, with very high correlation between them ($\rho = 0.98$); in contrast, the SardiNIA cohort, which had the largest λ_{GC}, showed the largest $\bar{\lambda}_{meta}$ (1.070 ± 0.049), standing out as a special case among the GIANT height cohorts. Assuming a polygenic model of $h^2 = 0.5$ over 30,000 independent loci, we simulated 174 cohorts using the actual size samples from the GIANT height cohorts (**Supplementary notes**), and observed an increased correlation ($R^2 = 0.78$) between $\bar{\lambda}_{meta}$ and λ_{GC} for simulated cohorts with sample sizes of the MIGEN cohorts **(Fig. 7d)**. Other effects, such as inflated/deflated sampling variance of the estimated genetic effects could also lead to correlation between $\bar{\lambda}_{meta}$ and λ_{GC} (**Supplementary Fig. S8**). In addition, we constructed a single MIGEN analysis by combining the 20 MIGEN cohorts using an inverse variance weighted meta-analysis[25], and calculated λ_{meta} between this combined MIGEN cohort and all 174 cohorts. As expected, the combined MIGEN had $\lambda_{meta} = 0.90 \pm 0.07$ with 20 MIGEN cohorts due to overlapping samples. In contrast, $\lambda_{meta} = 1.01 \pm 0.02$ with 154 other cohorts, was consistent with neither heterogeneity nor sample overlap. Given that the MIGEN (2,340 samples) and SardiNIA (4,303 samples) cohorts contributed less than 3% of the total sample size (253,288 samples from the GIANT height GWAS cohorts), any impact of unusual λ_{meta} values on the meta-analysis results is very small. Given no heterogeneity between a pair of cohorts, a deflated λ_{meta} reflects the effective number of overlapping samples (**Supplementary notes**). For example, the "combined MIGEN" had λ_{meta} values proportional to the sample size of each MIGEN cohort (**Fig 7e**).

The statistical power of detection of overlapping samples is maximized when a pair of cohorts has equal sample size (**Fig. 8a**), or in other words the confidence interval for null hypothesis of no overlapping samples depends on the sample sizes for a pair of cohorts. As a comparison, direct correlation that is estimated between the genetic effects for a pair of cohorts has been proposed to estimate overlapping samples[26,27], but it is confounded with genetic architecture, such as heritability underlying (**Table 1**). When there was heritability, the estimated correlation between genetic effects was biased and leads to incorrect overlapping samples for a pair of cohorts; when there was no heritability, the estimated correlation was correct and agreed well with the one estimated with λ_{meta}. As existence of heritability is one of the reasons that trigger GWAMA, so λ_{meta} is much proper in estimating overlapping samples between cohorts.

Another parameterization of λ_{meta} is to estimate it from differences in allele frequencies between a pair of cohorts instead of differences between estimated effect sizes (**Supplementary notes**). We show that λ_{meta} constructed on reported allele frequencies from genotyped loci from summary statistics can detect overlapping samples between two cohorts regardless of whether the GWAS is from quantitative traits or case-control data, even for pairs of different traits (**Supplementary notes and Supplementary notes**). For example, 2,934 common controls were shared across the WTCCC 7 diseases[7]. From the 21 pairwise λ_{meta}, we estimated a mean of the number of overlapping samples, assuming overlapping controls only (**Supplementary notes**), of $\hat{n}_o$ =2,708 (S.D. = 58.4), which was very close estimate to the actual number of overlapping samples (**Supplementary Fig. 8**). When constructing λ_{meta} on the reported genetic effects and their sampling variance, the estimated mean estimate of the number of shared controls was $\hat{n}_o$ = 2,127 (S.D. = 257.7), lower than that estimated from allele frequencies, which is likely due to real genetic heterogeneity between diseases (**Supplementary Fig. 8**). In practice, publically available summary statistics may not include sample specific allele frequencies, but may only be available with reference sample frequencies as a conservative strategy to prevent identification of individuals in a cohort.

Detection of overlapping samples using pseudo profile score regression

GWAMAs have grown in sample size and in the number of cohorts that participate, and this trend is likely to continue. The probability that a sample is represented in more than one meta-analysis study is also likely to increase, in particular when very large cohorts such as UK Biobank and 23andMe provide data to multiple studies. While the metric λ_{meta} can be transformed to give an estimate of n_o between cohorts for quantitative traits, it cannot give an estimate of overlapping samples in case-control studies due to the ratio of the cases and controls in each study (**Supplementary notes**). Sharing individual genotype data (or imputed genotypes) across the entire study would make it easy to detect identical or near-identical genotype samples (representing real duplicate samples from individuals who participated in the two studies or monozygotic twins). In fact, only a small number of common SNPs is needed to detect sample overlap, and if this is known then individuals could be removed and summary statistics regenerated or the meta-analysis

analysis itself can be adapted to correct for potential correlation due to n_o[28]. However, in many circumstances, individual cohorts are not permitted to share individual-level data, either by national law or by local ethical review board conditions. To get around this problem, Turchin and Hirshhorn[10] created a software tool, Gencrypt, which utilizes a security protocol known as one-way cryptographic hashes to allow overlapping participants to be identified without sharing individual-level data. To our knowledge, this encryption method has yet to be employed in meta-analysis studies. We propose an alternative approach, pseudo profile score regression (PPSR), which involves sharing of weighted linear combinations of SNP genotypes with the central meta-analysis hub. In essence, multiple random profile scores are generated for each individual in each cohort, using SNP weights supplied by the analysis hub, and the resulting scores are provided back to the analysis hub. PPSR works through three steps (**Supplementary notes and Supplementary Fig. 9**), and the purpose of PPSR is to estimate a relationship-like matrix of $n_i \times n_j$ dimension for a pair of cohorts, which have n_i and n_j individuals respectively. Each entry of the matrix is filled with genetic similarity for a pair of samples from each of the two cohorts, estimated via the PPSR.

We use WTCCC data as an illustration to detect 2,934 shared controls between any two of the diseases by PPSR. Among 330K unambiguous SNPs, which are not palindromic (A/T or G/C alleles), we randomly picked M = 100, 200, and 500 SNPs, to generate pseudo profile scores. It generated 21 cohort-pair comparisons, leading to the summation for 488,587,090 total individual-pair tests. To have an experiment-wise type I error rate = 0.01, type II error rate = 0.05 (power = 0.95) for detecting overlapping individuals, we needed to generated at least 57 pseudo profile scores (PPS). We generated scores $S = [s_1, s_2, s_3, \dots, s_{57}]$, where each s is a vector of M elements, sampled from a standard normal distribution (**Supplementary notes**). S is shared across 7 cohorts for generating pseudo-profile scores for each individual. In total 57 PPS were generated for each individual in each cohort. For a pair of cohorts, PPSR was conducted for each possible pair of individuals for any two cohorts over the generated pseudo-profile scores. Once the regression coefficient (b) was greater than the threshold, here $b = 0.95$, the pair of individuals was inferred to be having highly similar genotypes, implying that the individual was included in both cohorts (**Supplementary notes**).

When using 200 and 500 random SNPs, all the known 2,934 shared controls were detected from 21 cohort-pair-wise comparison; when using 100 randomly SNPs, on average 2,931 shared samples were identified, which is more accurate than using λ_{meta} constructed using either genetic effects or allele frequencies (**Fig. 8b**). In addition, for detected overlapping samples, there were no false positives observed – consistent with simulations that show the method was conservative in the controlling type I error rate (**Supplementary notes**). For comparison, we also used the Gencrypt to detect overlapping samples using the same set of SNPs as used in PPSR. Although Gencrypt guidelines suggest use of at least 20,000 random SNPs[10], selecting 500 random SNPs in the WTCCC cohorts also provided good accuracy with Gencrypt, and on average about 2,920 (99.6% of the shared controls) overlapping samples were detected, only slightly lower than PPSR. For example, for BP and CAD, Gencrypt detected 2,912 shared controls, but was unable to identify about 20 overlapping controls, due to missing data (on average 1% missing rate). Increasing the number of SNPs when using Gencrypt is likely to overcome the problem of missing data.

Furthermore, PPSR is able to detect pairs of relatives. For example, between the BD and CAD cohorts, two pairs of apparent first-degree relatives were detected (**Fig. 9a**). In order to find additional first-degree relatives between BD and CAD cohorts, at least 265 PPS were required to have a type I error rate of 0.01 and type II error rate of 0.05 (**Supplementary notes**) for a regression coefficient cutoff of 0.45, a threshold for first-degree relatives. As expected, all other individuals that did not show high relatedness did not reach the threshold of 0.45 of the PPS regression coefficient for first-degree relatives (**Fig. 9b**). Gencrypt did not detect any first-degree relatives.

The speed of PPSR depends on $n_i \times n_j$, the sample sizes for a pair of cohorts, and the number of PPS for each cohort; for the WTCCC data there are 21 cohort-pair comparisons, and each pair took about 20 minutes, on a computer with a 2.3 GHz CPU, given about $5{,}000 \times 5{,}000 = 25{,}000{,}000$ comparisons. The average sample size of GIANT is about 1,500, and takes about 2 minutes for each pair of cohorts. The two largest datasets are deCODE with 26,790 samples and WGHS with 23,100 samples, and PPSR to detect overlapping samples takes about 8.5 hours. As each pair of individuals is computationally an independent unit, analysis jobs can be parallelized on a cluster. Therefore, even for meta-analyses involving many large cohorts, the computation time is not a limiting factor.

PPSR for each individual uses very little personal information and can be minimized so that there is very low probability of decoding it. One way to attempt to decode the genotypes from PPS is to reverse the PPSR, so that the individual genotypes can be predicted in the regression (**Supplementary notes**). The individual-level genotypic information that can be recovered by an analyst, who knows the S matrix (the weights for generating PPS), is determined by the ratio between the number of markers (M) that generated PPS and the number of PPS (K). Therefore, inferred information on individual genotypes can be minimized and tailored to any specific ethics requirements. We suggest $\frac{M}{K} > 5 \sim 10$ to protect the privacy with sufficient accuracy (**Fig 9c**). Of note, if a meta-analysis is conducted within a research consortium, the application of PPSR is even safer because the exchange of information is between the consortium analysis hub and each cohort independently.

Discussion

In this study, we provide a set of metrics for monitoring and improving the quality of large-scale GWAMA based on summary statistics. These tools not only enrich the toolkit to analysts for GWAMA, but also provide informative summary and visualization for readers to understand the experimental design of GWAMA. As far as we know, no GWAMA to date has checked cohort-level outliers based upon population differentiation metrics or utilized estimated allelic effect sizes to identify and quantify sample overlap.

Using the F_{st} derived genetic distance measure, we can place all cohorts on an inferred geographic map and can easily identify cohorts that are genetic outliers or that have unexpected ancestry. In application, we should note that the F_{st} measure can identify unusual summary information, such as detected in the MIGEN cohorts from GIANT Consortium GWAMAs, in which the same allele frequencies were reported for all cohorts. Meta-PCA can also be used to infer the genetic background of cohorts. The high concordance between F_{pc} and meta-PCA indicates the both methods are robust. In practice, mete-PCA may be much easier to implement when there are many cohorts, such as GIANT height cohorts and Metabochip BMI cohorts, but the coordinates of a cohort may be slightly shifted with inclusion or exclusion of other cohorts.

There are limitation for both F_{pc} and meta-PCA. Firstly, the inference depends on the choice of reference cohorts. Meta-PCA is further upon the inclusion or exclusion of other cohorts. However, given the application of the data, we believe the impact will not influence the inference of the genetic background of cohorts in meta-analysis. Secondly, various mechanisms can give the identical projection in PCA[14]. The purpose of both methods is to find the discordance between demographic information and genetic information, or outliers. The projection is not attempt to discover the detailed demographic past that shapes a cohort.

Our third metric λ_{meta} provides information on sample overlap and heterogeneity between cohorts by utilizing the estimated allelic effect sizes and their standard errors. In most meta-analyses, the overall λ_{meta} is likely to be slightly greater than 1 solely due to unknown heterogeneity, slight as observed, in generating the phenotype and genotype data that cannot be accounted for by QC. The observed mean of λ_{meta} for the GIANT height GWAMA was 1.03 but with more variation than expected by chance. The strong correlation between λ_{GC} and λ_{meta} indicated the reported sampling of the reported data were systematically driven by analysis protocols. For cohorts with $\lambda_{GC} < 1$ and $\lambda_{meta} < 1$, it is likely that the GWAS modeling strategy employed for GWAS in the cohort was too conservative, for example MIGEN cohorts might have on average too small sample size for each cohort. Conversely, for cohorts with $\lambda_{GC} > 1$ and $\lambda_{meta} > 1$ results are too heterogeneous, perhaps reflecting systematically smaller sampling variances of the reported genetic effects. As the GWAMA often uses inverse-variance-weighted meta-analysis[25], such cohorts may lead to incorrect weights to the different cohorts in the meta-analysis, suggesting that the statistical analysis in meta-analyses can be improved by applying better weighting factors.

It is well-recognised that overlapping samples may inflate the type-I error rate of GWAMA and therefore lead to false positives. Although post-hoc correction of the test statistic is possible[26–28], stringent quality control ruling out overlapping samples makes the whole analysis easier and lowers the risk of false positives. A better solution would be to rule out shared samples at the start, for pairs of cohorts that show deflated λ_{meta}, and we propose PPSR to accomplish this.

In summary, to maximize the inference from multi-cohort GWAMA, accurate cohort-level information on allele frequencies, estimated effect sizes, and their sampling variance can be exploited to perform additional measures that are likely to lead to reduction in the number of false positives and increasing statistical power for gene discovery. All methods proposed are implemented in freely available software GEAR.

Acknowledgements

This work was funded by Australian National Health and Medical Research Council Project and Fellowship grants (1011506, 613601, 613602, 1078901, 1078037), grants GM 099568 from the National Institutes of Health and the Sylvia & Charles Viertel Charitable Foundation. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We also thank from High Performing Computing support from the Information Technology group at the Queensland Brain Institute, The University of Queensland.

Author contributions:

GBC and PMV designed the study. GBC, PMV and SHL derived the analytical results. GBC performed all analysis. CGB and ZXZ developed the software. GBC and PMV wrote the first draft of the paper. MRR, JY, NW discussed results and methods, and provided comments that improved earlier versions of the manuscript. Other authors provided cohort-level summary statistics and contributed to improving the study and manuscript.

Competing financial interests:

The authors declare no completing financial interests.

Web resources:

GEAR (GEnetic Analysis Repository): http://www.complextraitgenomics.com/
PGC: http://www.med.unc.edu/pgc/results
1000 Genomes Project: http://www.1000genomes.org/

References

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

2 Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* 2014; **9**: 1192–212.

3 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.

4 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.

5 Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS *et al.* The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLOS Genet* 2012; **8**: e1002793.

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

7 The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.

8 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M a R, Bender D *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–75.

9 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.

10 Turchin MC, Hirschhorn JN. Gencrypt: one-way cryptographic hashes to detect overlapping individuals across samples. *Bioinformatics* 2012; **28**: 886–8.

11 Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A *et al.* Genes mirror geography within Europe. *Nature* 2008; **456**: 98–101.

12 Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes*. Princeton University Press, 1996.

13 Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLOS Genet* 2006; **2**: e190.

14 McVean G. A genealogical interpretation of principal components analysis. *PLOS Genet* 2009; **5**: e1000686.

15 Bryc K, Bryc W, Silverstein JW. Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theor Popul Biol* 2013; **89**: 34–43.

16 Chaput J-P, Pérusse L, Després J-P, Tremblay A, Bouchard C. Findings from the Quebec Family Study on the Etiology of Obesity: Genetics and Environmental Highlights. *Curr Obes Rep* 2014; **3**: 54–66.

17 Diabetes Genetics Initiatives. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (80-)* 2007; **316**: 1331–1336.

18 Igl W, Johansson A, Gyllensten U. The Northern Swedish Population Health Study (NSPHS)--a paradigmatic study in a rural population combining community health and basic research. *Rural Remote Health* 2010; **11**: 1363.

19 de Bakker PIW, Ferreira M a R, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17**: R122–8.

20 Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans P a *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 2011; **43**: 969–76.

21 Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, Akterin S *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013; **45**: 1150–9.

22 Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014; **506**: 376–381.

23 Calò C, Melis A, Vona G, Piras I. Sardinian Population (Italy): a Genetic Review. *Int J Mod Anthropol* 2010; **1**: 39–64.

24 Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* 2011; **19**: 807–12.

25 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190–1.

26 Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K *et al.* A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLOS Genet* 2014; **10**: e1004198.

27 Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N *et al.* Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. *Am J Hum Genet* 2015; **96**: 21–36.

28 Lin D-Y, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* 2009; **85**: 862–72.

Figures & Tables

Figure 1 Recovery of cohort-level genetic background and inference of their geographic locations for GIANT BMI Metabochip cohorts using the F_{st} derived genetic distance measure.

Figure 2 Using the genetic distance spectrum to infer the geographic origins for GIANT height GWAS cohorts.

Figure 3 Comparison between Meta-PCA and genotype PCA on 1KG.

Figure 4 Recovery of cohort-level genetic background for GIANT BMI Metabochip cohorts using meta-PCA.

Figure 5 The recovery of cohort-level genetic background using meta-PCA analysis for GWAS height cohorts.

Figure 6 λ_{meta} for the GIANT height GWAS cohorts.

Figure 7 $\bar{\lambda}_{meta}$ and λ_{gc} for GIANT height GWAS cohorts.

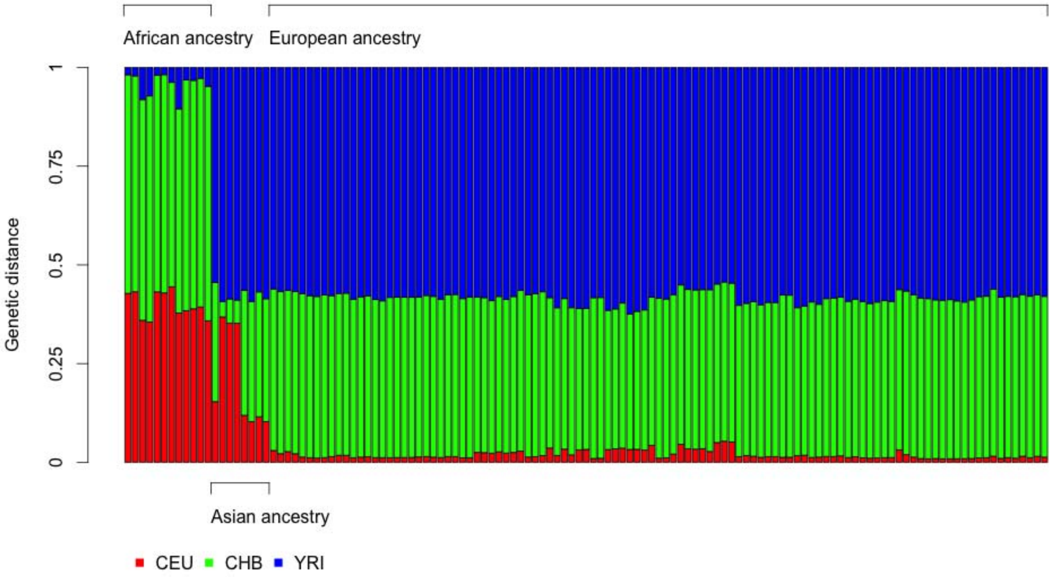
Figure 8 Pseudo profile score regression for the WTCCC 7 diseases.

Figure 9 PPSR coefficients for identifying shared controls/relatives between WTCCC BD and CAD cohorts.

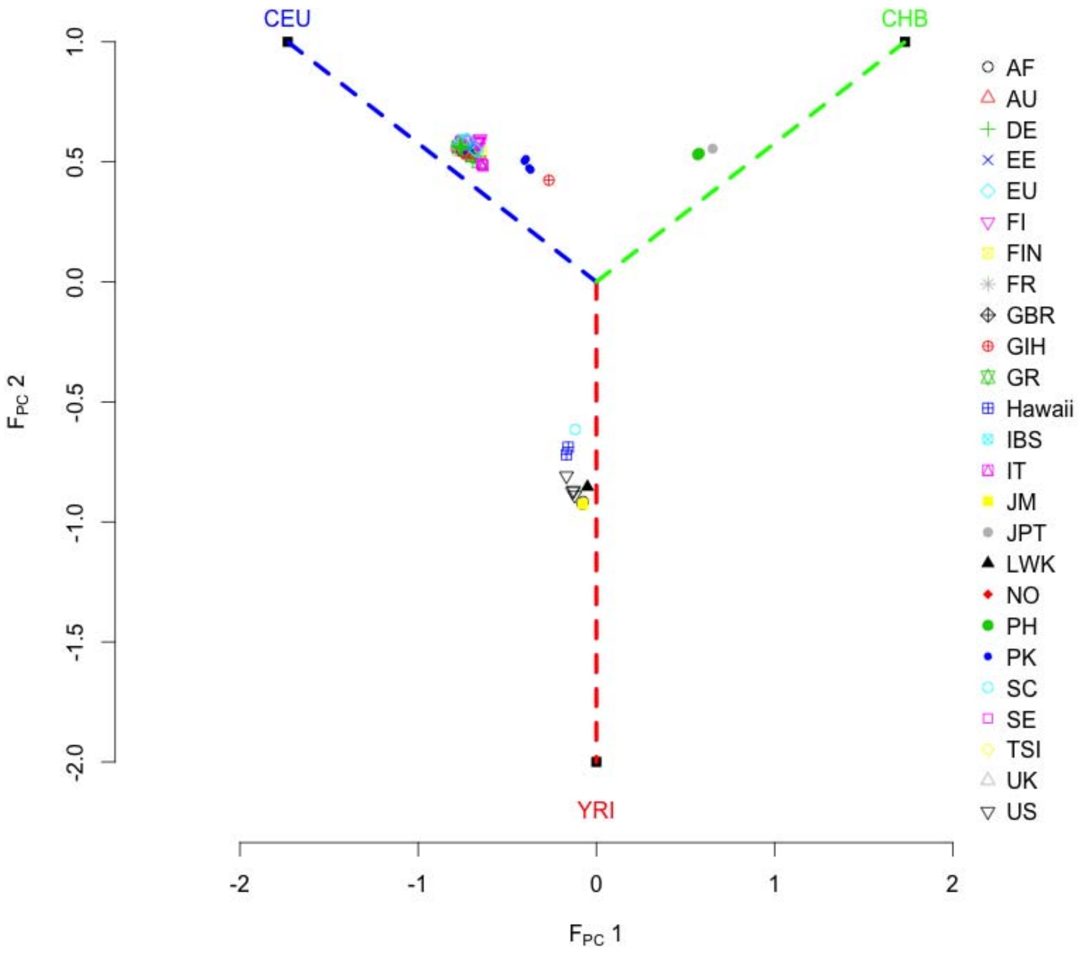
Table 1 The estimated correlation for a pair of cohorts via their summary statistics

Figure 1 Recovery of cohort-level genetic background and inference of their geographic locations for GIANT BMI Metabochip cohorts using the F_{st} derived genetic distance measure. (a) Genetic distance spectrum for all Metabochip cohorts to CEU, CHB and YRI. See Supplementary notes for more details. The origins of the cohorts are denoted on the horizontal axis. (**b)** Projection for Metabochip cohort into F_{PC} space defined by YRI, CHB, and CEU reference populations. The x- and y-axis represent relative distances derived from the genetic distance spectrum. Three dashed lines, blue for CEU, green for CHB, and red for YRI, partitioned the whole F_{PC} space to three genealogical subspaces. (**c)** The genetic distance spectrum for Metabochip European cohorts to CEU – Northwest Europeans, FIN – Northeast European, and TSI – Southern Europeans. The nationality of the cohorts are denoted on the horizontal axis. (**d)** The projection for Metabochip European cohorts to the F_{PC} space defined by CEU, FIN, and TSI reference populations. The whole space is further partitioned into three subspaces, CEU-TSI genealogical subspace (red and blue dashed lines), FIN-TSI genealogical subspace (green-blue dashed lines), and CEU-FIN genealogical subspace (red-green dashed lines), respectively. The open circles represent the mean of inferred geographic locations for the cohorts from the same country. Cohort/country codes: AF, African; AU, Australia; DE, Germany; EE, Estonia; EU, European Nations; FI, Finland; FIN, Fins in 1000 Genomes Project (1KG); FR, France; GBR, British in 1KG; GIB, Gujarati Indian in 1KG; GR, Greece; Hawaii, Hawaii in USA; IBS, Iberian Population in Spain in 1KG; IT, Italy; JM, Jamaica; JPT, Japanese in 1KG; LWK, Luhya in 1KG; NO, Norway; PH, the Philippines; PK, Pakistan; SC, Seychelles; SCT, Scotland; SE, Sweden; TSI, Tuscany in 1KG; UK, United Kingdom; US, United States of America.

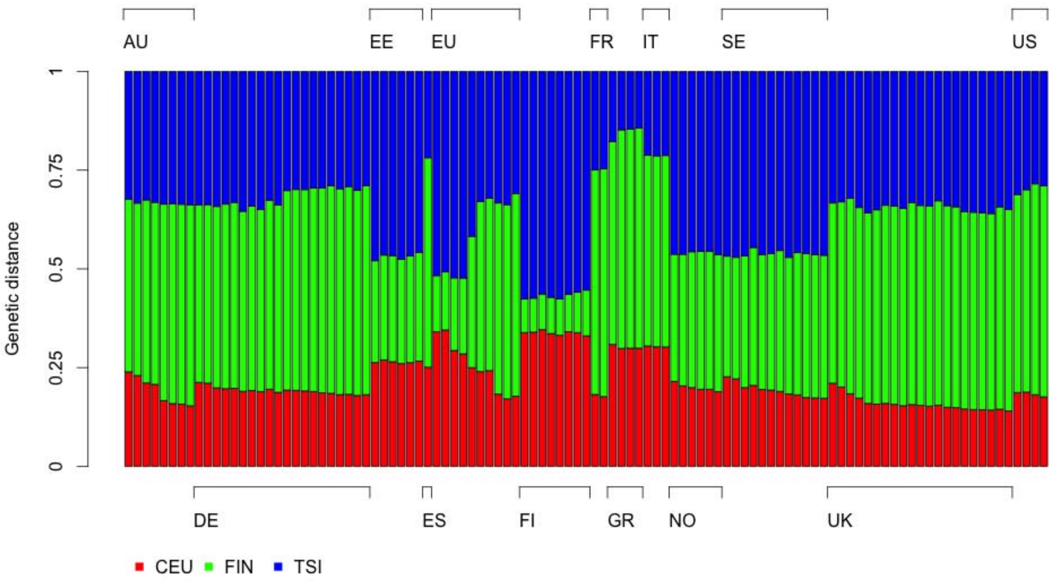
a)


African ancestry
European ancestry
Genetic distance
1
0.75
0.5
0.25
0
Asian ancestry
■ CEU ■ CHB ■ YRI

b)


CEU
CHB
YRI
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
F_{PC} 2
-2
-1
0
1
2
F_{PC} 1
AF
AU
DE
EE
EU
FI
FIN
FR
GBR
GIH
GR
Hawaii
IBS
IT
JM
JPT
LWK
NO
PH
PK
SC
SE
TSI
UK
US

c)


AU
EE
EU
FR
IT
SE
US
Genetic distance
1
0.75
0.5
0.25
0
DE
ES
FI
GR
NO
UK
🟥 CEU 🟩 FIN 🟦 TSI

d)

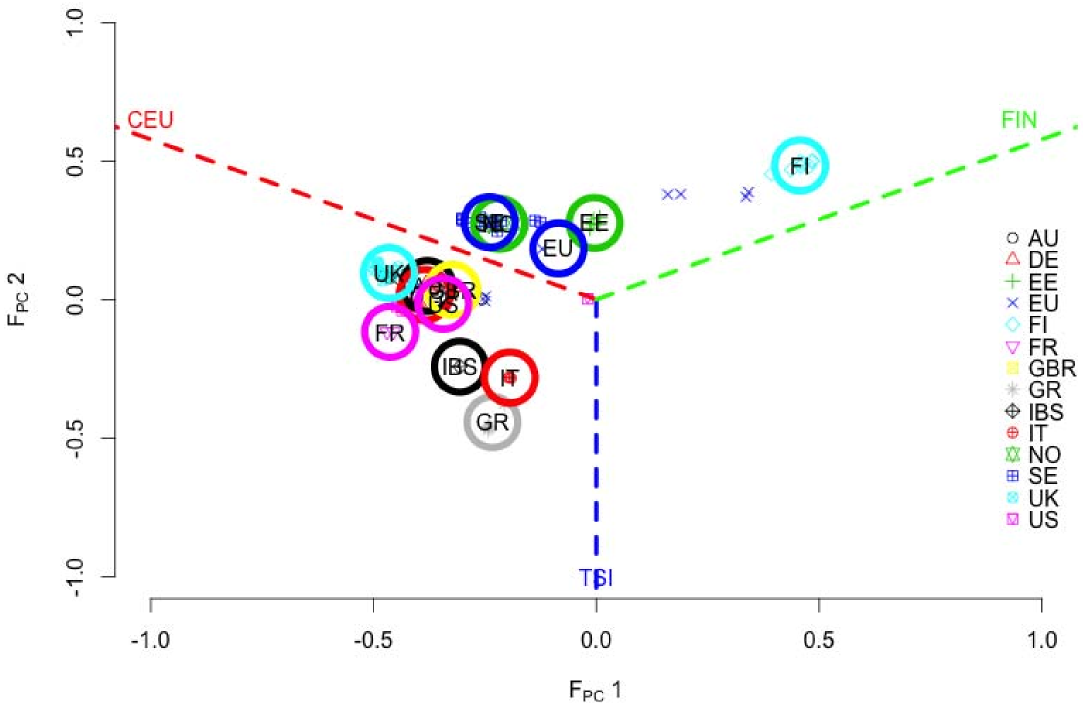
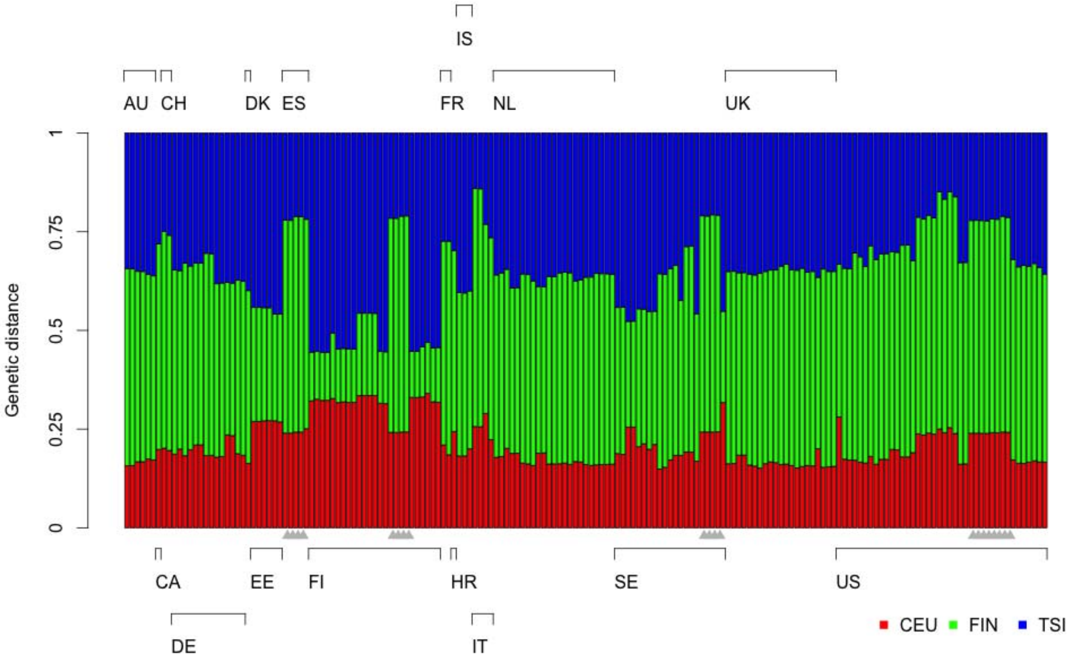

CEU
FIN
TSI
FI
SE
NO
EE
EU
UK
AU
GBR
DE
US
FR
IBS
IT
GR
○ AU
△ DE
+ EE
× EU
◇ FI
▽ FR
GBR
GR
IBS
IT
NO
SE
UK
US
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
F$_{PC}$ 2
F$_{PC}$ 1

Figure 2 Using the genetic distance spectrum to infer the geographic origins for GIANT height GWAS cohorts. (a) Each cohort has three F_{st} values by comparing with CEU, FIN, and TSI reference samples. The height of each bar represents its relative genetic distance to these three reference populations. The nationalities of the cohorts were denoted along the horizontal axis. The grey triangles along the x-axis indicate MIGEN cohorts. **(b)** Given the three F_{st} values, the location of each cohort can be mapped. The whole space was partitioned into three subspaces, CEU-TSI genealogical subspace (red and blue dashed lines), FIN-TSI genealogical subspace (green and blue dashed lines), and CEU-FIN genealogical subspace (red and green dashed lines). DGI (in the blue box) had samples from the Botnia study. Across MIGEN cohorts (denoted as red triangles in the red box), the same allele frequencies (likely calculated from a South European cohort) were presented for each cohort. Cohort/country codes: AU, Australia; CA, Canada; CH, Switzerland; DE, Germany; DK, Denmark; EE, Estonia; ES, Iberian Population in Spain in 1KG; FI, Finland; FR, France; GR, Greece; IT, Italy; IS, Iceland; NL, Netherlands; SE, Sweden; UK, United Kingdom; US, United States of America.

a)


IS
AU
CH
DK
ES
FR
NL
UK
Genetic distance
1
0.75
0.5
0.25
0
CA
EE
FI
HR
SE
US
DE
IT
🟥 CEU 🟩 FIN 🟦 TSI

b)

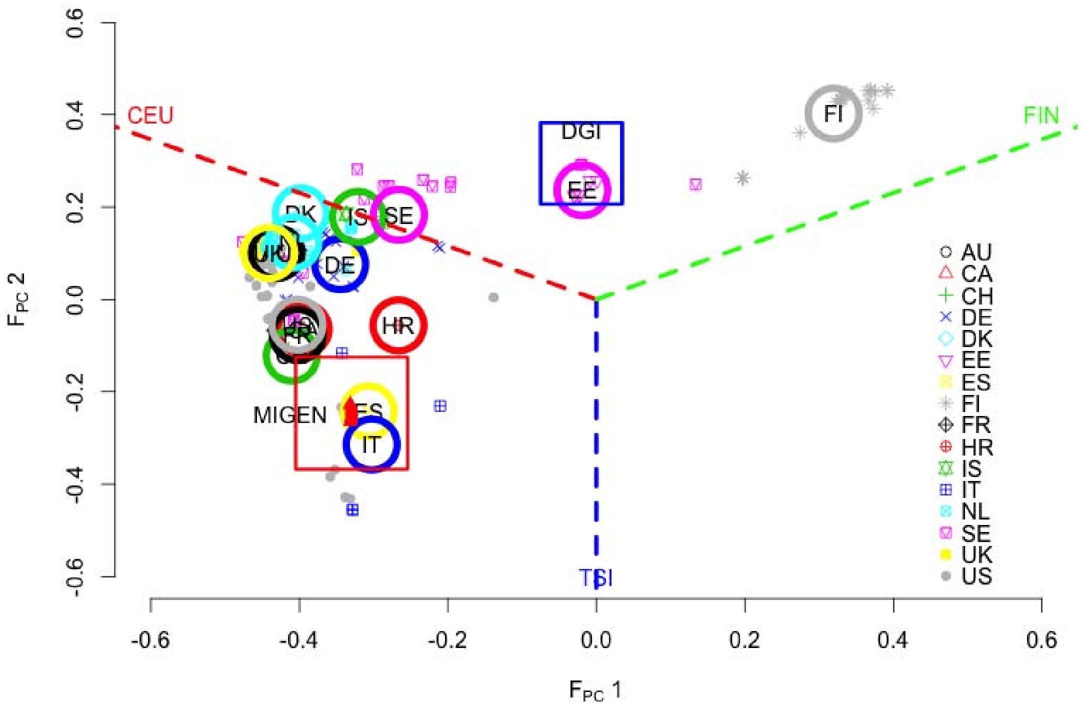

CEU
FIN
TSI
DGI
MIGEN
FI
EE
DK
IS
SE
UK
DE
HR
ES
IT
FR
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
F_{PC} 2
F_{PC} 1
○ AU
△ CA
+ CH
× DE
◇ DK
▽ EE
⊠ ES
✳ FI
⊕ FR
⊕ HR
✡ IS
⊞ IT
⊠ NL
⊠ SE
■ UK
● US

Figure 3 Comparison between Meta-PCA and genotype PCA on 1KG. Top left panel is the projection of cohorts based on cohort-level allele frequency for 1KG samples on the first two eigenvectors. Bottom left panel is conventional PCA based on individual genotypes on the first two eigenvectors. Top right panel is the projection by taking the mean of the 1KG individuals within each cohort. Bottom right panel is the correlation, measured in , between meta-PCA and genotype PCA for the first twenty eigenvectors.

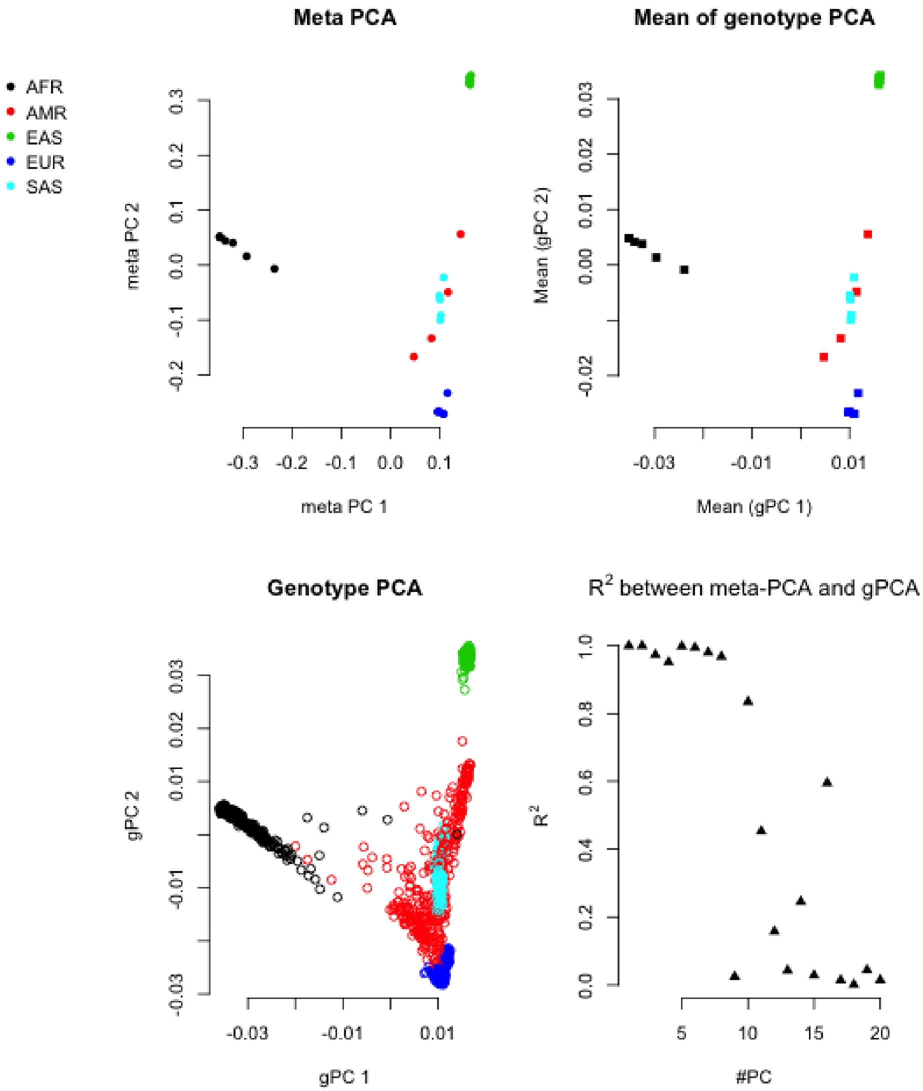


Figure 4 Recovery of cohort-level genetic background for GIANT BMI Metabochip cohorts using meta-PCA. The x-axis and y-axis represent the first two eigenvectors from meta-PCA. In meta-PCA, Metabochip cohorts could be classified into African ancestry (AFR), European ancestry (EAS), East Asian Ancestry (EAS), and South Asian Ancestry (SAS). The 1KG cohorts, yellow open circles, were added for comparison.

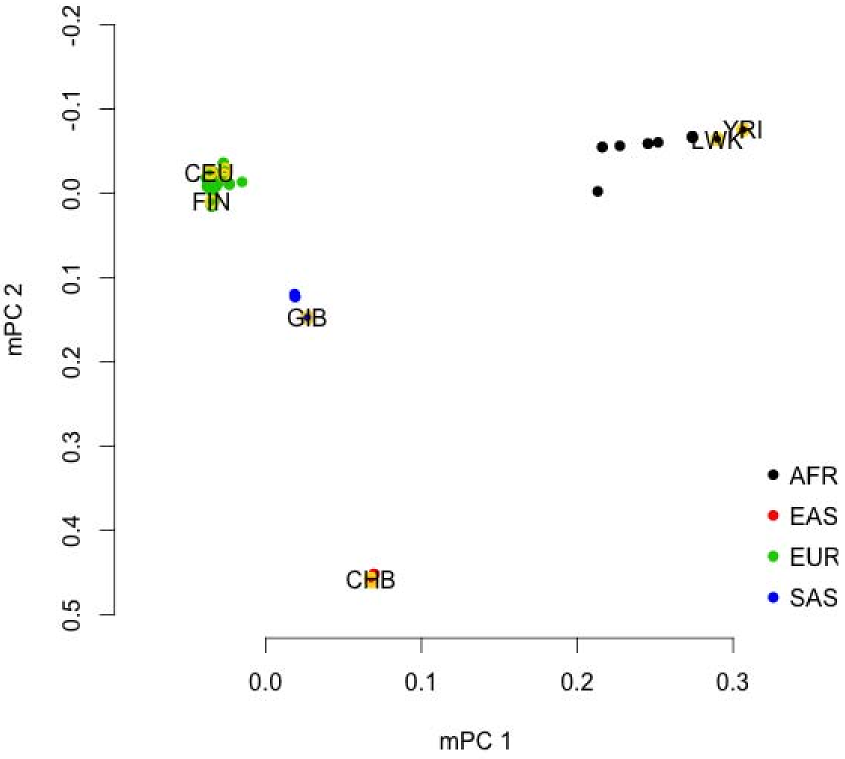
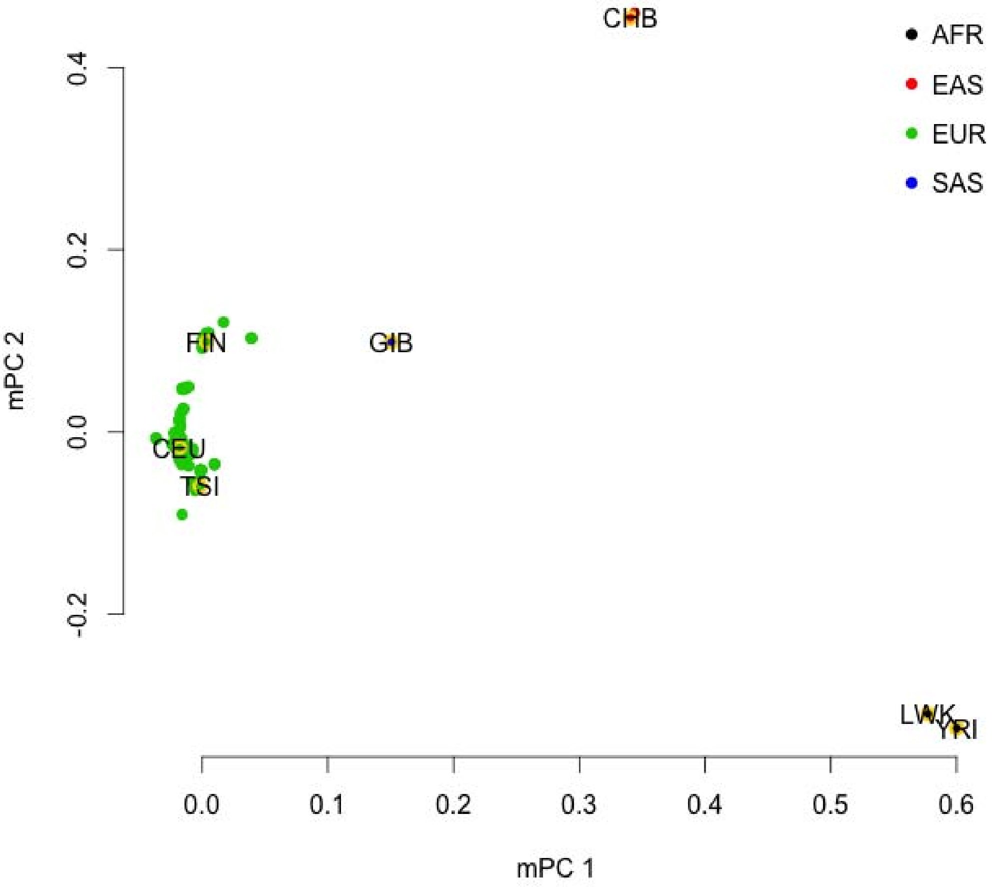


Figure 5 The recovery of cohort-level genetic background using meta-PCA analysis for GWAS height cohorts. The x-axis and y-axis represent the firs two eigenvectors inferred from meta-PCA. a) The genetic background inferred with the inclusion of 10 1KG reference populations. b) The genetic background and relative geographic location for 174 GIANT height cohorts. The large plot on top left was an overview of 174 cohorts, and the rest of plots were classified by the reported demographic information of cohorts. Within each country-level plot, the small black points represent one cohort, and the large open circle the mean coordinates for those cohorts from the same country.

a



b

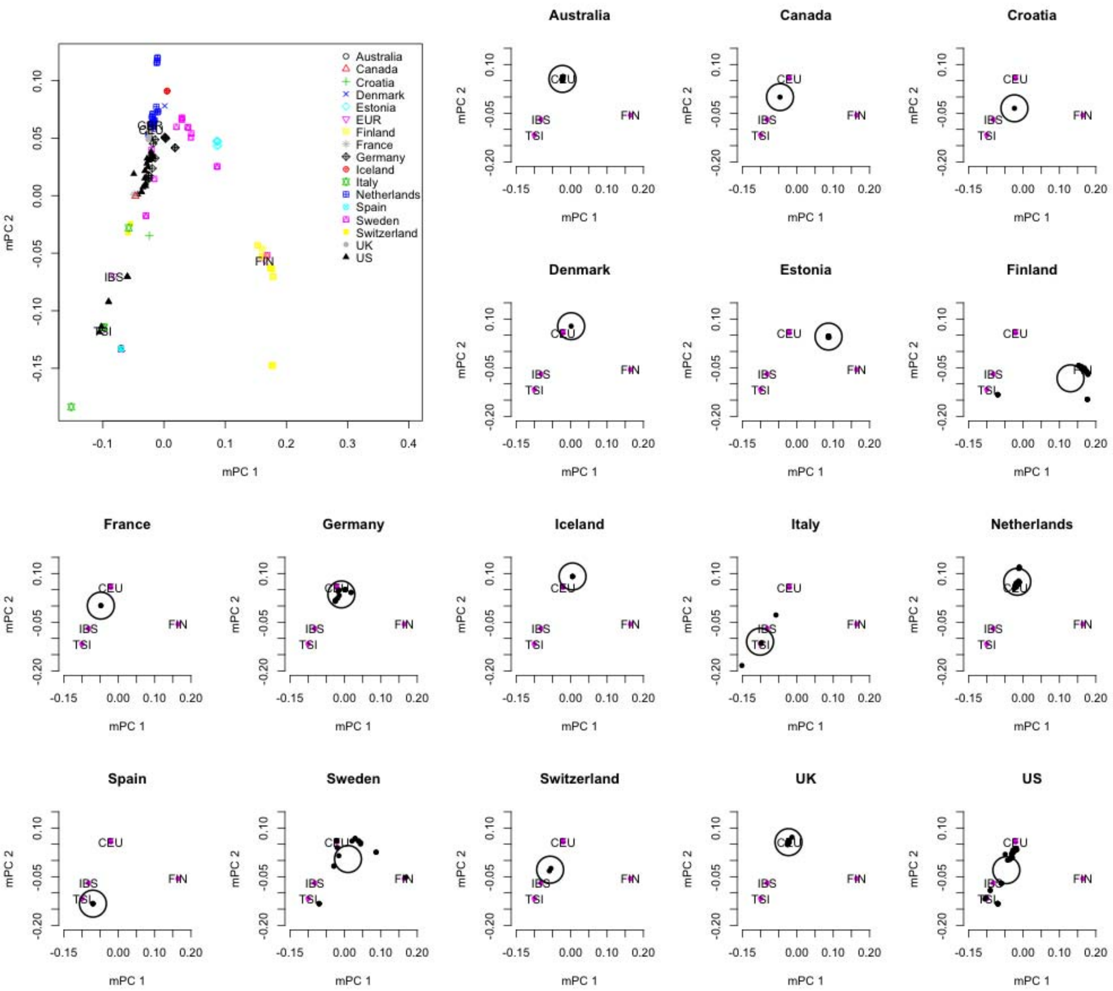
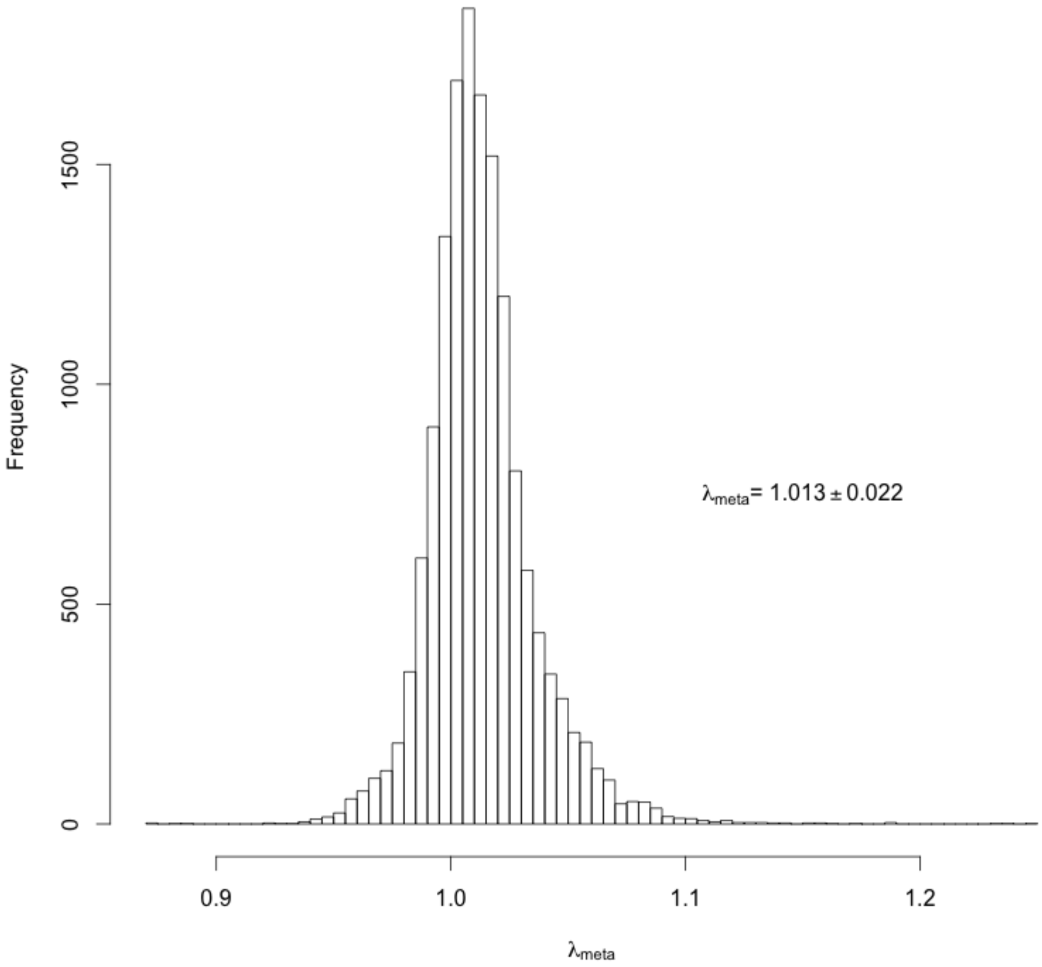

Australia
Canada
Croatia
Denmark
Estonia
EUR
Finland
France
Germany
Iceland
Italy
Netherlands
Spain
Sweden
Switzerland
UK
US
CEU
IBS
TSI
FIN
mPC 1
mPC 2
-0.1
0.0
0.1
0.2
0.3
0.4
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.15
0.00
0.10
0.20

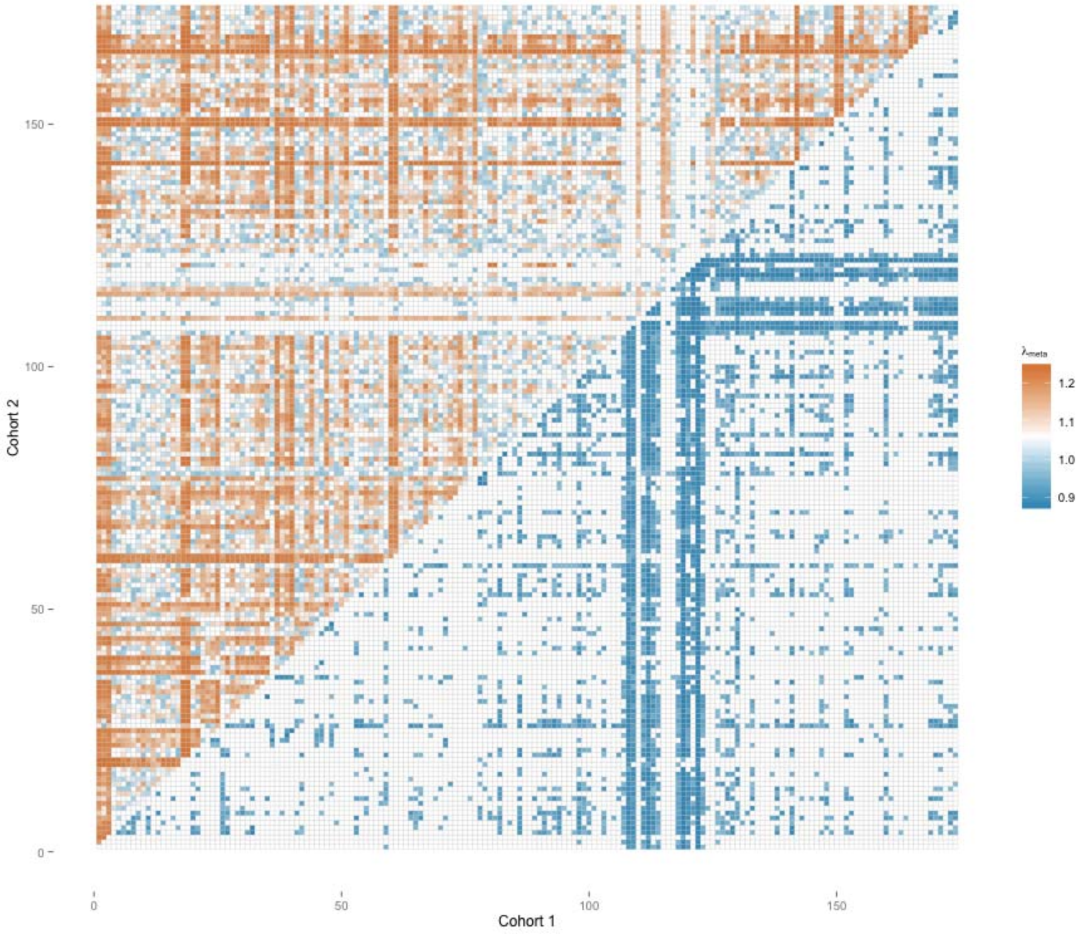
Figure 6 $\boldsymbol{\lambda_{meta}}$ for the GIANT height GWAS cohorts. Given 174 cohorts, there are 15,051 λ_{meta} values, which provide the overview of the quality control of the summary statistics.

(a) The distribution of λ_{meta} from 174 cohorts/files used in the GIANT height meta-analysis. The overall mean of 15,051 λ_{meta} is 1.013, and standard deviation is 0.022. **(b)** The heat map for λ_{meta}. Cohorts showed heterogeneity ($\lambda_{meta} > 1$) are illustrated on left-top triangle, and homogeneity ($\lambda_{meta} < 1$) on right-bottom triangle. **(c)** Illustration for homogeneity between two cohorts (SORBS MEN & WOMEN), $\lambda_{meta} = 0.876$. **(d)** Illustration of SARDINIA & WGHS, this pair of cohorts has $\lambda_{meta} = 1.245$. The grey band represents 95% confidence interval for λ_{meta}.

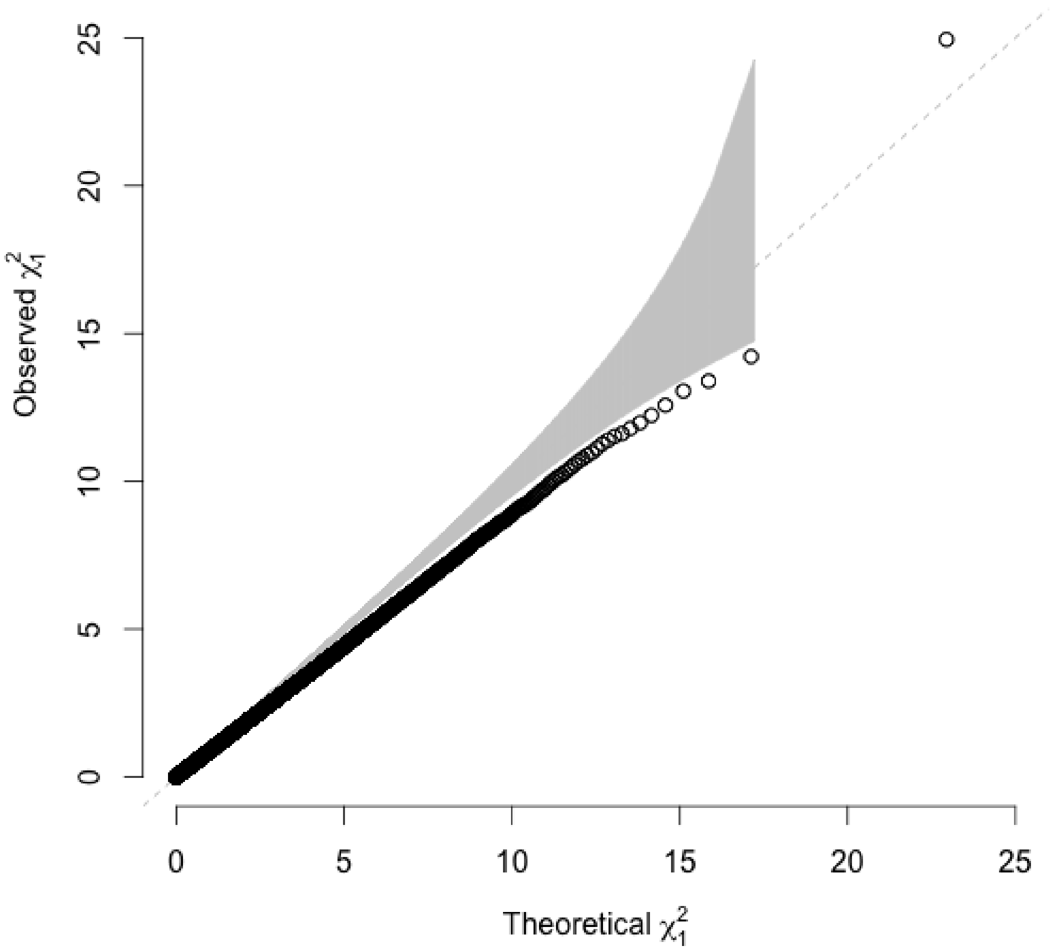
a)


Frequency
0
500
1000
1500
λ_{meta}= 1.013 ± 0.022
0.9
1.0
1.1
1.2
λ_{meta}

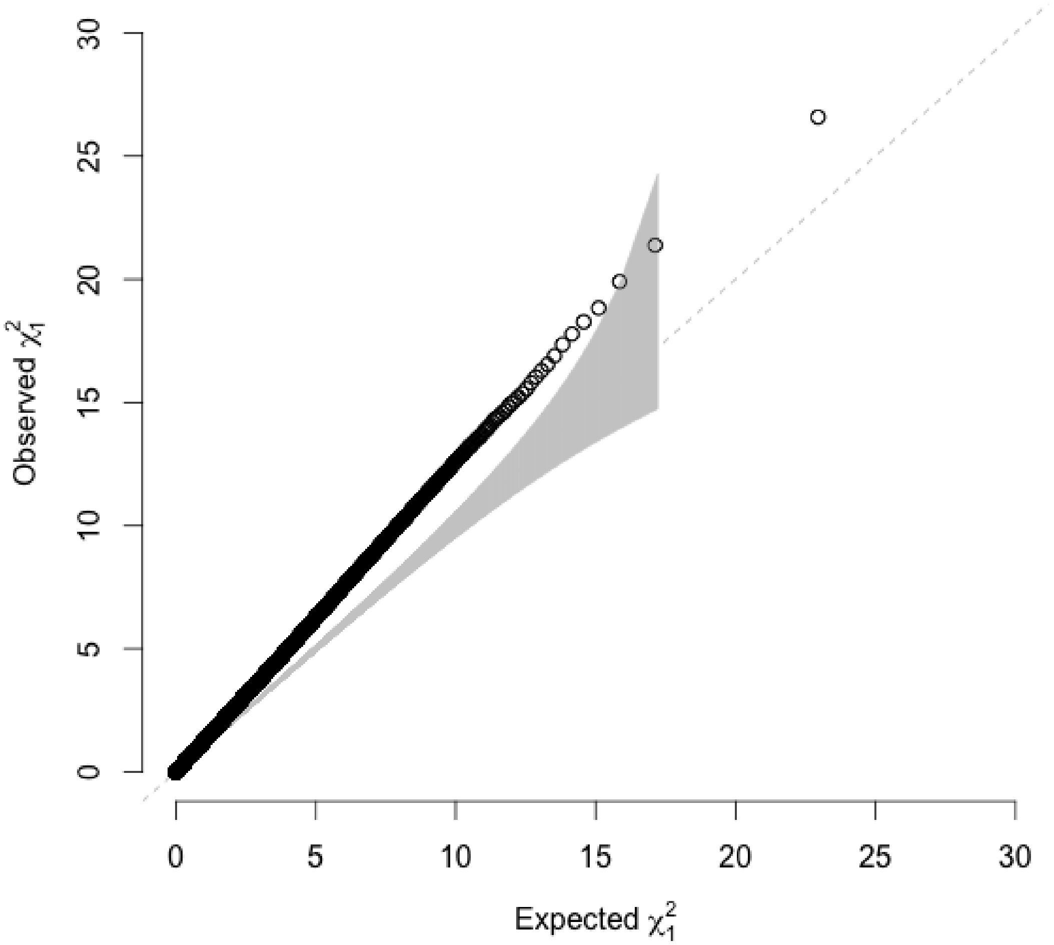
b)


150
100
50
0
Cohort 2
0
50
100
150
Cohort 1
λ_{meta}
1.2
1.1
1.0
0.9

c)



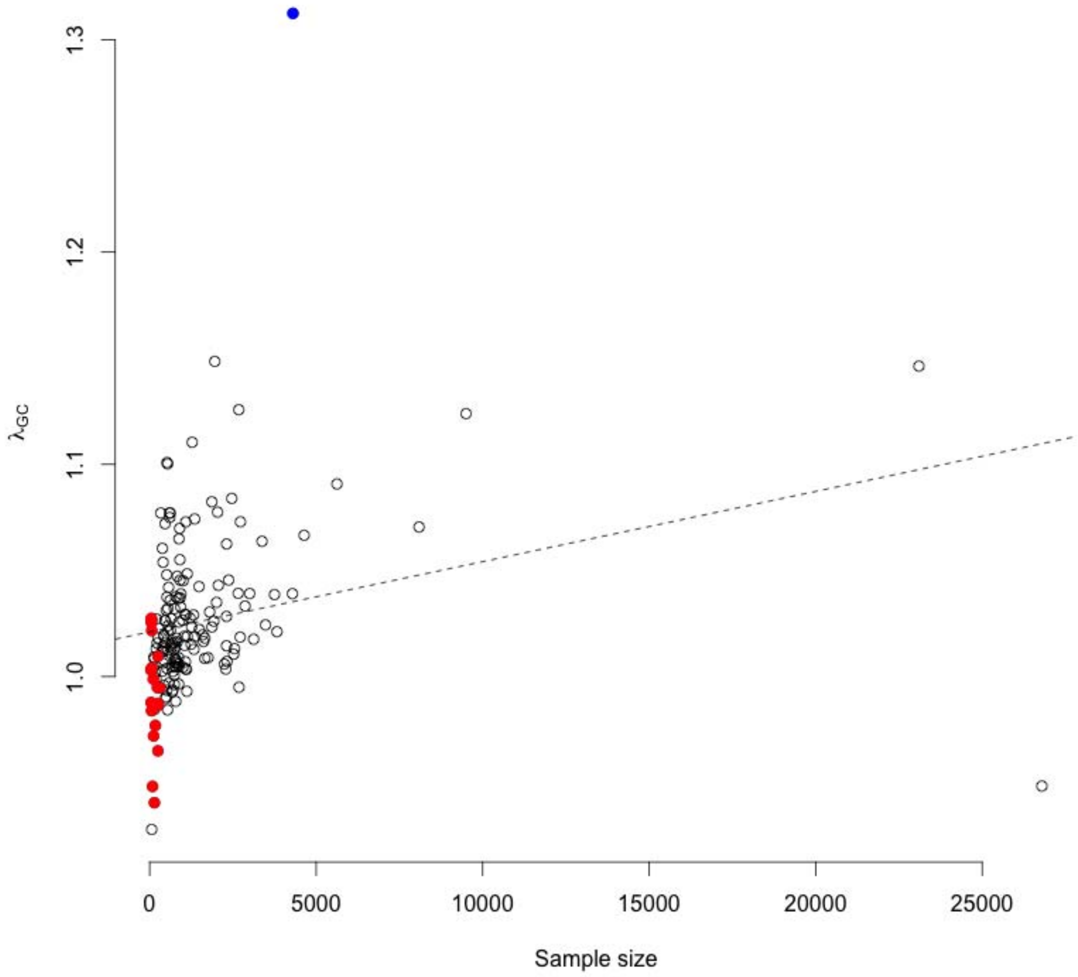
d)



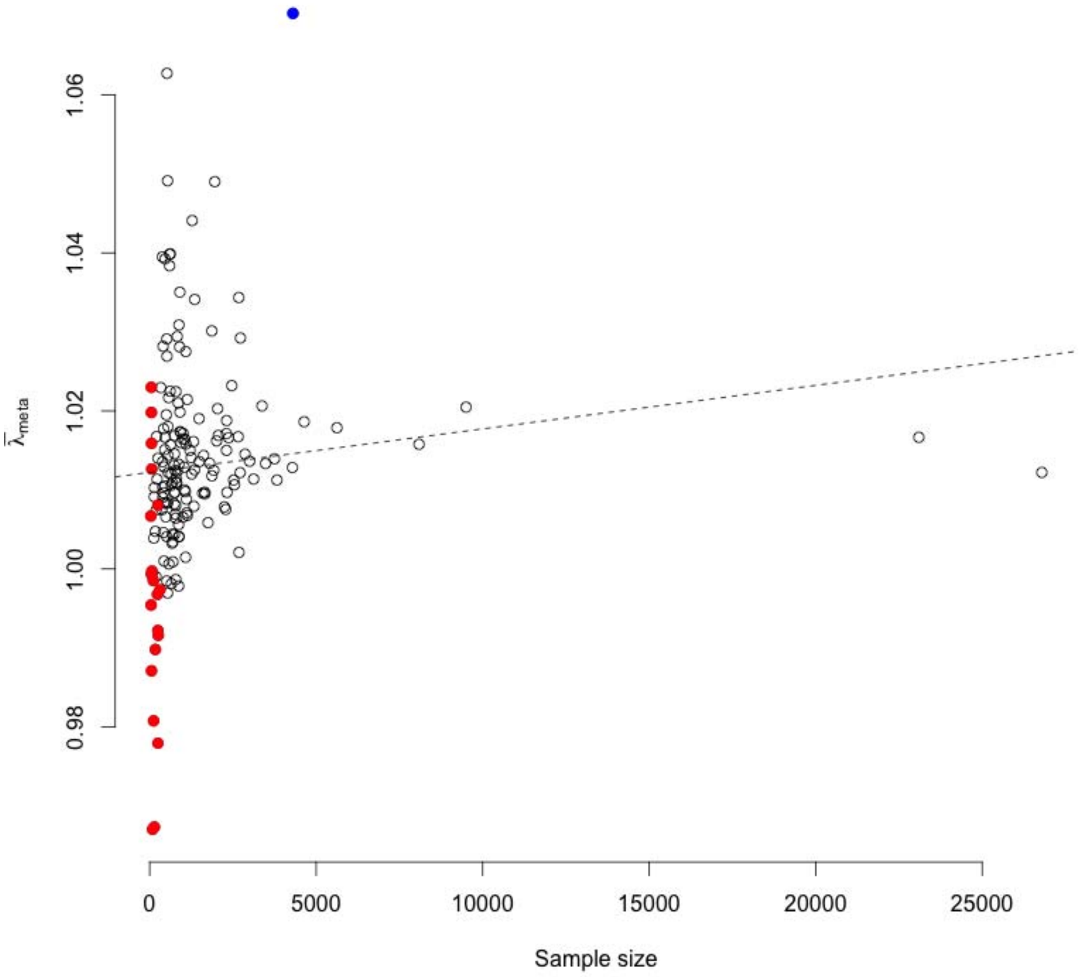
0
5
10
15
20
25
30
Observed χ^2_1
0
5
10
15
20
25
30
Expected χ^2_1

Figure 7 $\bar{\lambda}_{meta}$ and λ_{gc} for GIANT height GWAS cohorts. **(a)** Sample size of each cohort against λ_{GC}. The linear regression is presented as a dashed line, λ_{GC} = 1.021 + 0.0000033N, and $R^2 = 0.013$. **(b)** Sample size of each cohort against $\bar{\lambda}_{meta}$, which was the mean of a cohort's λ_{meta} over all other cohorts. The linear regression is presented as a dashed line, $\bar{\lambda}_{meta}$ = 1.012 + 0.00000055N (N = reported sample size), and $R^2 = 0.055$. **(c)** λ_{GC} against $\bar{\lambda}_{meta}$ for each cohort, showing a strong correlation, $R^2 = 0.70$. The black dash line indicates the regression slope for all 174 pairs: $\bar{\lambda}_{meta}$=0.7251+0.281λ_{GC}+e. The red dashed line indicates the regression slope for 20 pairs of MIGEN cohorts: $\bar{\lambda}_{meta}$=0.369+0.631λ_{GC}+e. The side of each circle is proportional to sampling size on logarithm scale. **(d)** Small sample size leads to a correlation between $\bar{\lambda}_{meta}$ and λ_{GC} using 174 GIANT height GWAS sample size. 30,000 independent loci, minor allele frequency ranged from 0.1~0.5, were simulated, and $h^2 = 0.5$. The red dashed line indicates the regression slope for 20 simulated MIGEN cohorts, $\bar{\lambda}_{meta} = 0.488 + 0.510\lambda_{GC}$+e ($R^2 = 0.78$). The side of each circle is proportional to sampling size on logarithm scale. **(e)** λ_{meta} for whole MIGEN to 174 cohorts. 20 MIGEN files were combined together to make "whole MIGEN" via meta-analysis, and the summary statistics were used to calculate λ_{meta} with 174 cohorts using 30,000 independent loci. As MIGEN cohorts were part of "whole MIGEN", their λ_{meta} were in general below 1. The dashed line is the mean of λ_{meta} of the "whole MIGEN". The subplot (red box) shows a strong correlation of 0.93 between λ_{meta} (for "whole MIGEN" vs each MIGEN cohort), and sample size of each MIGEN cohort.

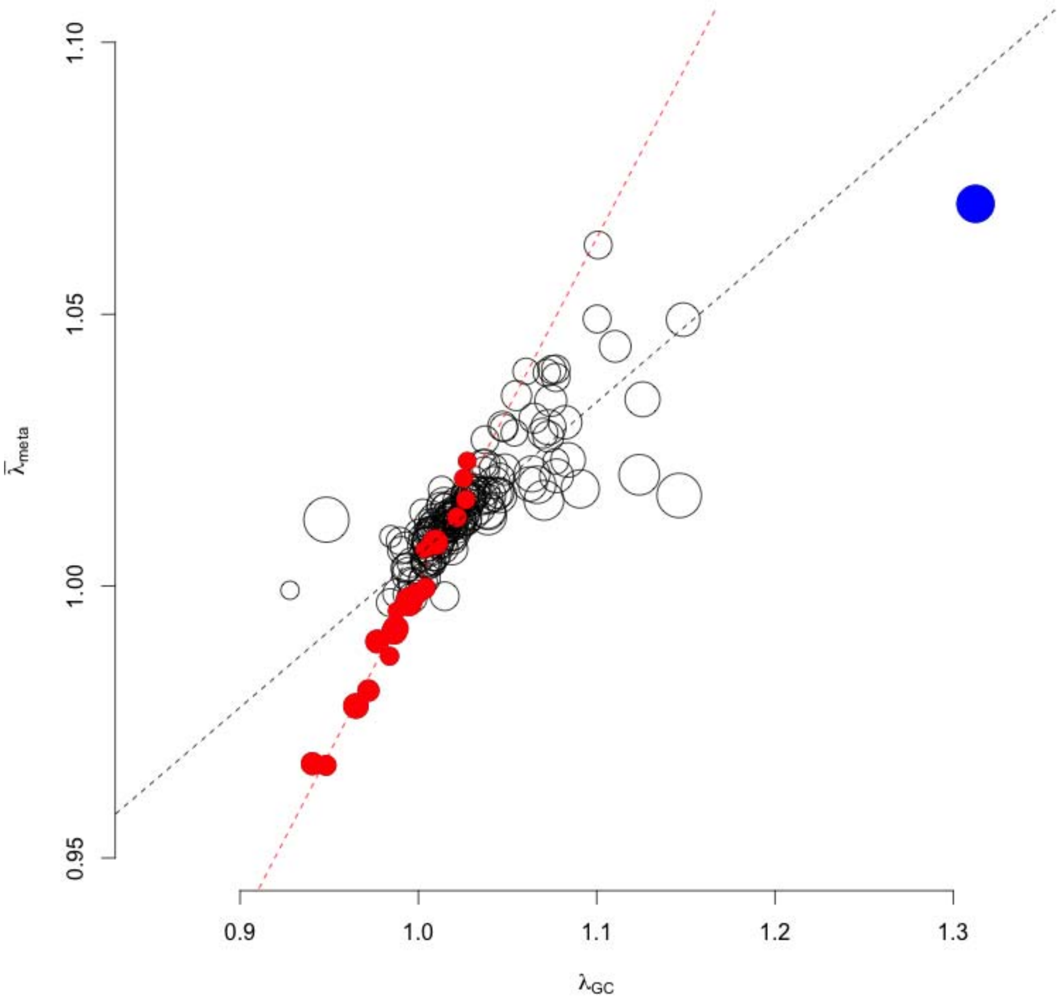
a)


1.3
1.2
1.1
1.0
λ_{GC}
0
5000
10000
15000
20000
25000
Sample size

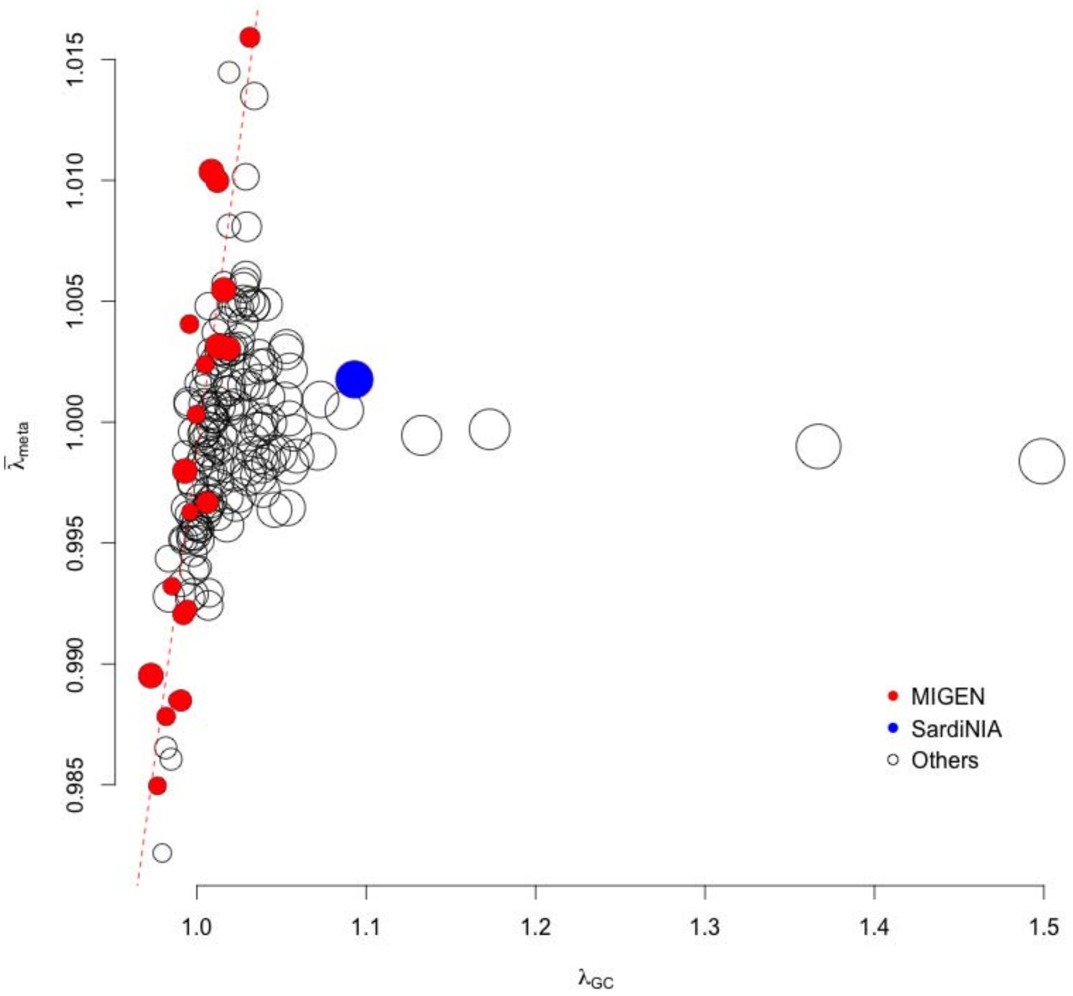
b)


1.06
1.04
1.02
1.00
0.98
$\bar{\lambda}_{meta}$
0
5000
10000
15000
20000
25000
Sample size

c)



d)


1.015
1.010
1.005
1.000
0.995
0.990
0.985
$\bar{\lambda}_{meta}$
1.0
1.1
1.2
1.3
1.4
1.5
λ_{GC}
● MIGEN
● SardiNIA
○ Others

e)

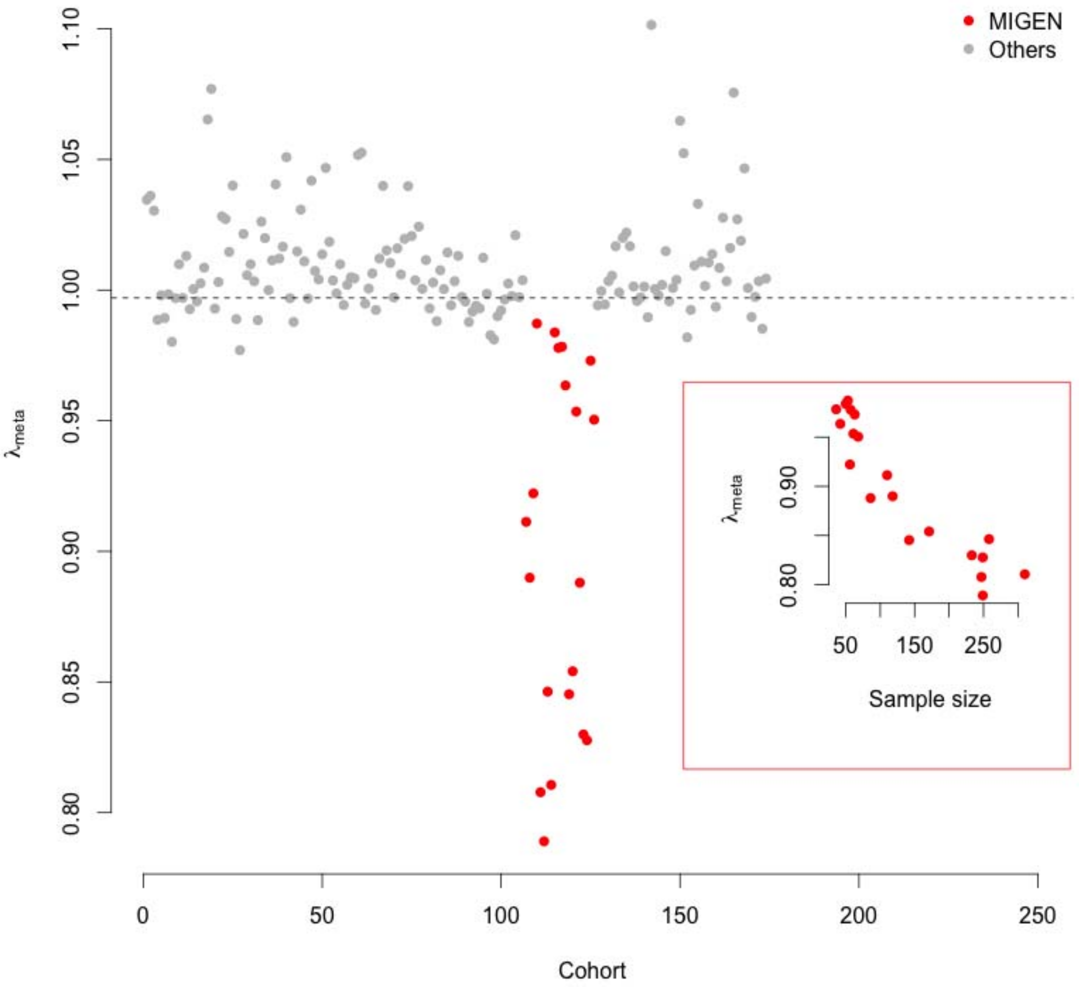
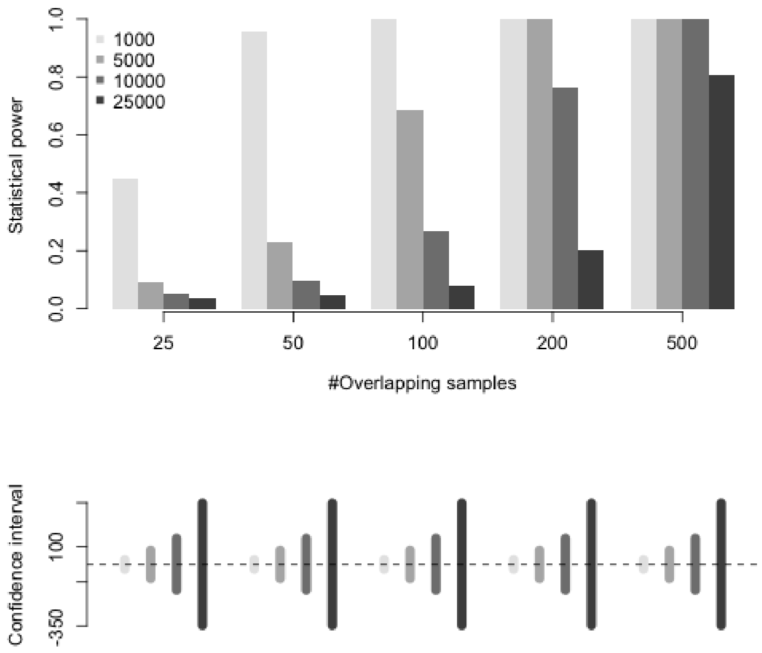

MIGEN
Others
λ_{meta}
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0
50
100
150
200
250
Cohort
λ_{meta}
0.90
0.80
50
150
250
Sample size

Figure 8 Pseudo profile score regression for the WTCCC 7 diseases. a) Statistical power for detecting overlapping samples between a pair of cohorts given type I error rate of 0.05. Top panel: The y-axis represents statistical power, and the x-axis the number of overlapping samples. Cohort 1 has 1,000, 5,000, 10,000, or 25,000 samples, and cohort 2 has 1,000 samples. The two cohorts have 25, 50, 100, 200, and 500 overlapping samples. Bottom panel: the corresponding 95% confidence interval is given for each scenario in the top panel. The statistical power is maximized when the two cohorts have the same sample size. b) Each cluster represents a pair of cohorts as denoted on the x-axis. Within each cluster, from left to right, the detected overlapping controls using λ_{meta} based either on effect size estimates or minor allele frequency (MAF), PPRS using 100, 200, and 500 markers. WTCCC cohort codes: BD for bipolar disorder, CAD for coronary artery disease, CD for Crohn's disease, HT for hypertension, RA for rheumatoid arthritis, T1D for type 1 diabetes, T2D for type 2 diabetes.

a)


Statistical power
0.0
0.2
0.4
0.6
0.8
1.0
1000
5000
10000
25000
25
50
100
200
500
#Overlapping samples
Confidence interval
-350
100

b)

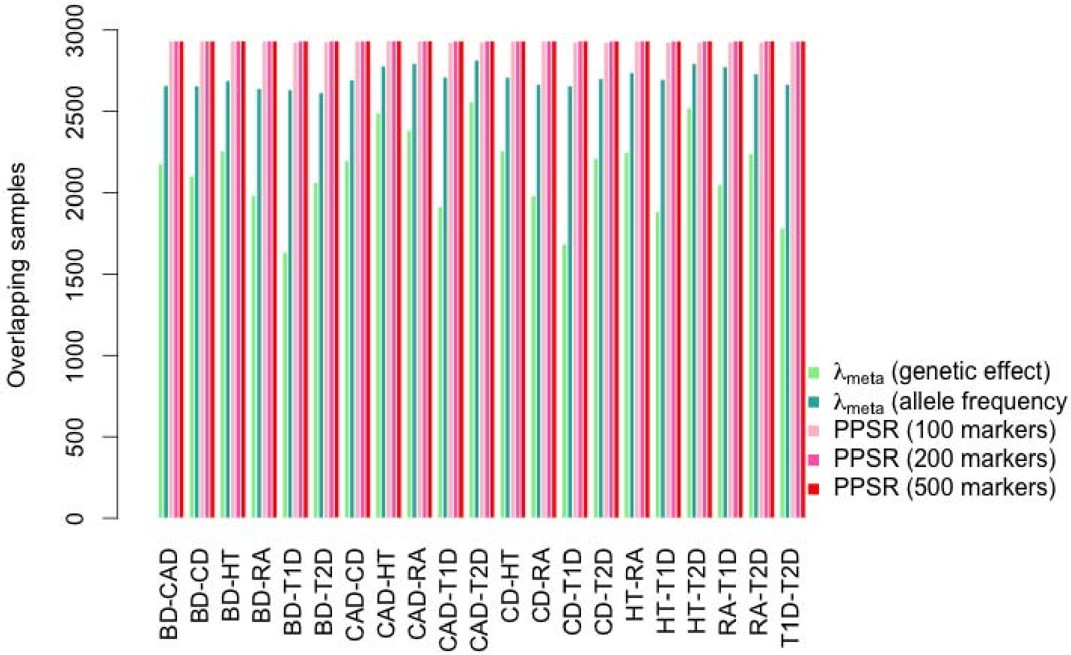
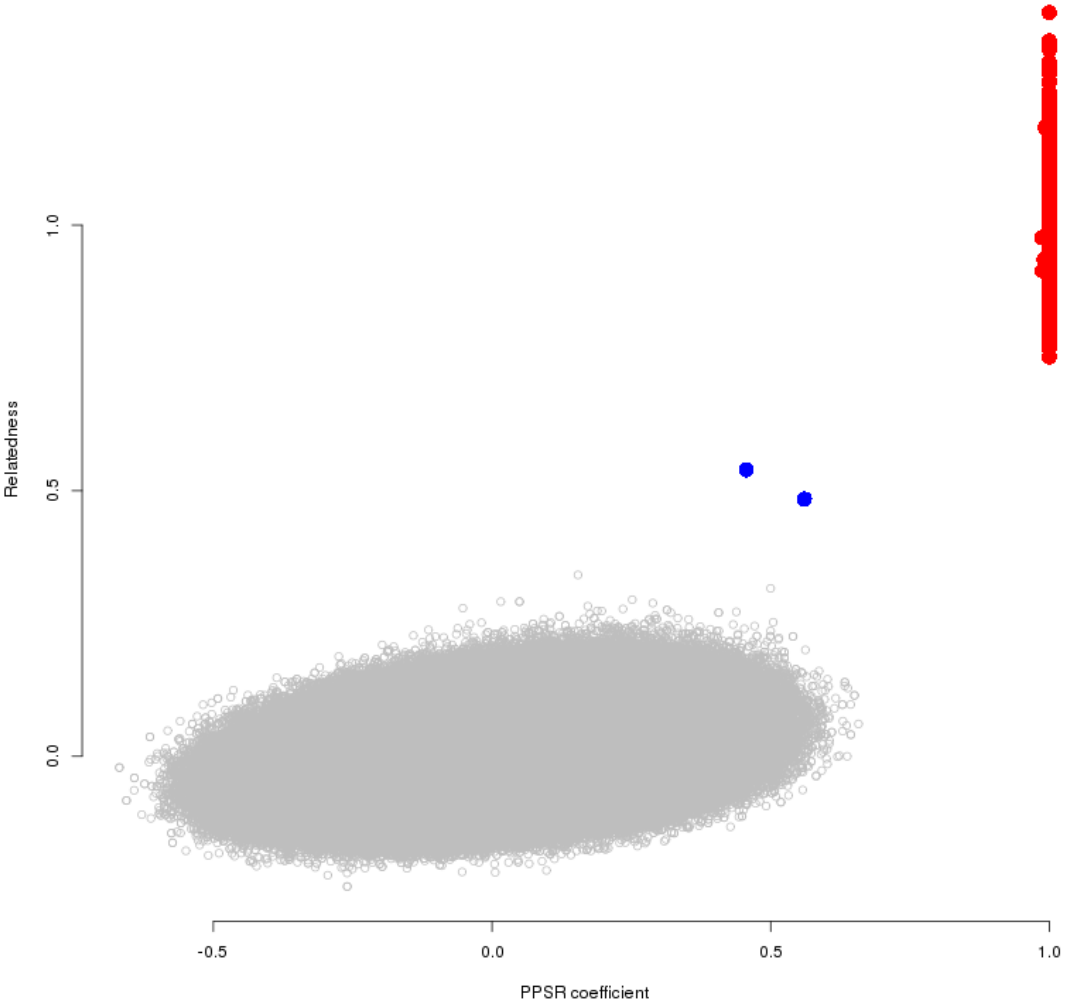

Overlapping samples
0
500
1000
1500
2000
2500
3000
BD-CAD
BD-CD
BD-HT
BD-RA
BD-T1D
BD-T2D
CAD-CD
CAD-HT
CAD-RA
CAD-T1D
CAD-T2D
CD-HT
CD-RA
CD-T1D
CD-T2D
HT-RA
HT-T1D
HT-T2D
RA-T1D
RA-T2D
T1D-T2D
λ_{meta} (genetic effect)
λ_{meta} (allele frequency
PPSR (100 markers)
PPSR (200 markers)
PPSR (500 markers)

Figure 9 PPSR coefficients for identifying shared controls/relatives between WTCCC BD and CAD cohorts. (a) Illustration for regression coefficients between WTCCC BD and CAD from 57 pseudo profile scores (PPS) generated from 500 markers. The x-axis is the PPSR regression coefficients and y-axis is real genetic relatedness (as calculated from individual level genotype data). The red points are the shared controls between two cohorts, and blue points are first-degree relatives. **(b)** The PPS regression coefficients for detecting overlapping first-degree relatives using 286 PPS generated from 500 markers. **(c)** Decoding genotypes from the PPS. Given the set of profile scores, one may run a GWAS-like analysis to infer the genotypes. The ratio between the number of markers (*M*) and number of pseudo profile scores (*K*) determines the potential discovery of individual-level information. The higher the ratio and, the higher the allele frequency, the less information can be recovered. From left to right, the profile scores generated using different number of markers. The y-axis is a R^2 metric representing the accuracy between the inferred genotypes and the real genotypes. From left to right panels 100, 200, 500, and 1000 SNPs were used to generate 10, 20, 50, and 1000 profiles scores. In each cluster, the three bars are inferred accuracy using different MAF spectrum alleles, given with the SE of the mean.

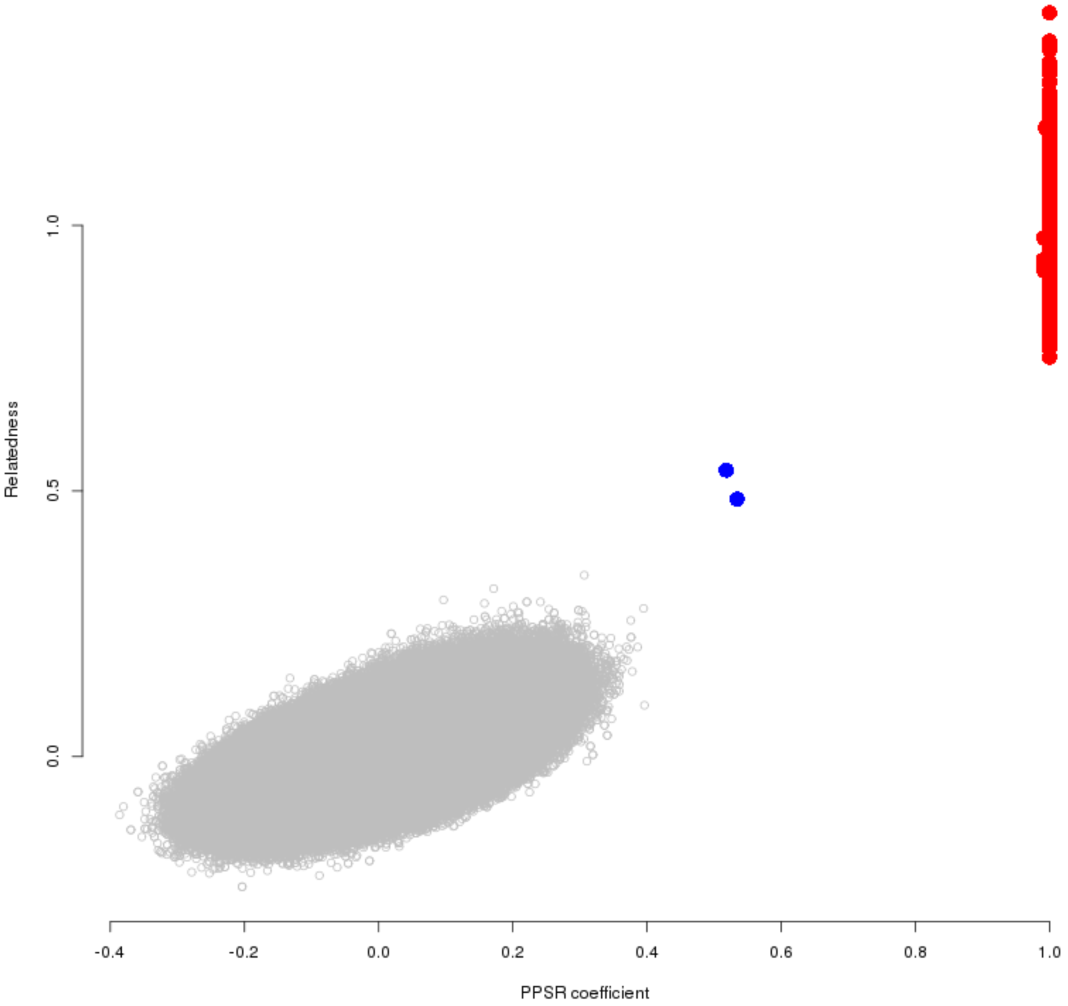
796 a)



797

798

b)



c)

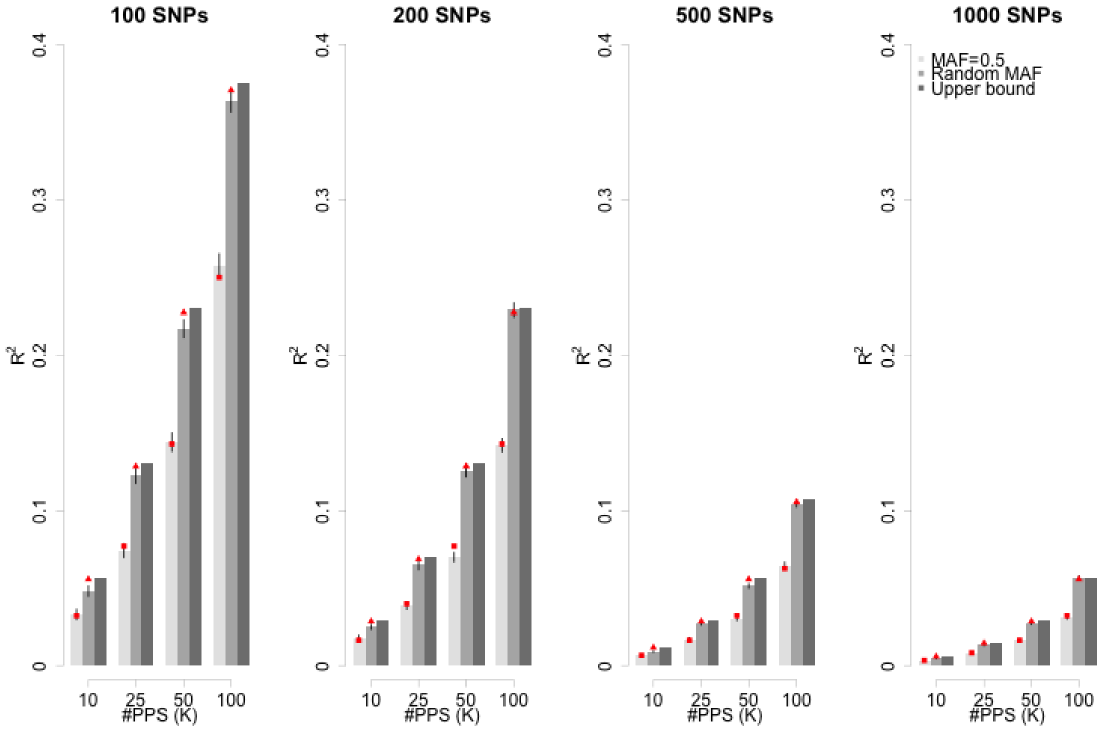

100 SNPs
200 SNPs
500 SNPs
1000 SNPs
R^2
0
0.1
0.2
0.3
0.4
10
25
50
100
#PPS (K)
MAF=0.5
Random MAF
Upper bound

Table 1 The estimated correlation for a pair of cohorts via their summary statistics

n_1	n_2	$n_{1,2}$	h^2	M	Q	$\gamma_{1,2} = \frac{n_{1,2}}{\sqrt{n_1 n_2}}$	$\hat{\rho}_{1,2} \pm \mathrm{SD}$	$\hat{\gamma}_{1,2} \pm \mathrm{SD}$
1,000	1,000	100	0.25	30,000	1,000	0.1	0.1072±0.0064	0.101±0.0093
1,000	2,000	100	0.25	30,000	1,000	0.0707	0.0814±0.0054	0.0709±0.0088
1,000	5,000	100	0.25	30,000	1,000	0.0447	0.0615±0.0055	0.0425±0.0096
1,000	10,000	100	0.25	30,000	1,000	0.0316	0.0556±0.0063	0.0325±0.0099
1,000	1,000	1	0.25	30,000	1,000	0.001	0.0092±0.0056	0.0017±0.0093
1,000	2,000	1	0.25	30,000	1,000	0.0007	0.0126±0.0053	0.0006±0.0079
1,000	5,000	1	0.25	30,000	1,000	0.000447	0.0189±0.0060	0.0016±0.0090
1,000	10,000	1	0.25	30,000	1,000	0.000316	0.0259±0.0059	0.0008±0.0092
1,000	1,000	100	0	30,000	1,000	0.1	0.0996±0.0052	0.094±0.0085
1,000	2,000	100	0	30,000	1,000	0.0707	0.0704±0.0048	0.0712±0.0097
1,000	5,000	100	0	30,000	1,000	0.0447	0.0453±0.0057	0.0441±0.0090
1,000	10,000	100	0	30,000	1,000	0.0316	0.0335±0.0057	0.0325±0.0079

* Q is the number of QTLs among M simulated loci. We also tried $Q = 100$, the results were nearly identical.

$\gamma_{1,2}$ represents the true correlation due to overlapping samples

$\hat{\rho}_{1,2}$ represents the estimated correlation estimated via the method proposed by Bolormaa et al[26], and Zhu et al[27]

$\hat{\gamma}_{1,2}$ represents the estimated correlation estimate via λ_{meta}, $\hat{\gamma}_{1,2} = \frac{1-\hat{\lambda}_{meta}}{\frac{2\sqrt{n_1 n_2}}{n_1+n_2}}$.

Table of contents

Method I: F_{st} derived genetic distance

F_{st} is a measure of genetic differentiation between populations. It is usually estimated using individual-level genotype data from multiple samples in two or more populations[1]. Here, we calculate F_{st} using summary data on allele frequencies, which implicitly assumes Hardy-Weinberg equilibrium genotype frequencies within populations. We use summary statistic calculated F_{st} as a metric for quality control for each cohort. If the allele frequencies reported for a cohort depart genome-wide from its expectation based on known ancestry due to technical artifacts, then we may observe an unexpected F_{st} value when comparing to a reference panel of know ancestry.

We calculate F_{st} between each cohort and a reference panel, choosing the appropriate reference sample depending on the purpose of the analysis. For the inference of global-level diversity, we chose YRI, CHB, and CEU as the reference panels. For the inference of within-Europe diversity, we chose CEU, FIN, and TSI as the reference panels. As the different allele frequencies across three samples reflected the real diversity among these reference panels, we did not apply any exclusion criteria on the reference allele frequency. Nevertheless, as GIANT height GWAS samples were imputed to the HapMap panel, the majority of SNPs matched to the 1KG reference samples comprised common SNPs. After ranking the calculated F_{st} in ascending order for all matched SNPs, we sampled 30,000 F_{st} evenly along the ordered F_{st}. These 30,000 markers are quasi-independent and evenly distributed across the genomes. The mean of the 30,000 F_{st} was employed to represent the F_{st} measure between a cohort and a reference panel. The sampled 30,000 markers may differ from one pair of cohorts to another pair, but as tested resample 30,000 markers caused ignorable changes of the mean of F_{st}. Another reason we chose 30,000 markers is that there are around 30,000 quasi-independent markers for GWAS data as observed in empirical data and expected from theory[2,3].

In this study, F_{st} is calculated from the allele frequencies estimated from cohorts, provided as summary statistics. F_{st} is treated as a data statistic for measuring allele frequency differentiation. In general the interpretation of F_{st} can vary with context[4].

$$F_{st} = \frac{\frac{r}{(r-1)\sum_{i=1}^{r} n_i}\left[\sum_{i=1}^{r} n_i (p_i - \bar{p})^2\right]}{\bar{p}(1-\bar{p})} \quad \textbf{(Equation 1)}$$

with p_i the estimated reference allele frequency in population i from a sample of n_i alleles, $\bar{p}$ is the weighted average frequency in the entire sample, and r is the number of populations.

Here, we only compared each cohort to the 1KG reference panel, so $r = 2$ and the equation becomes

$$F_{st} = \frac{\frac{2}{n_{1,2}}\left[\sum_{i=1}^{2} n_i (p_i - \bar{p})^2\right]}{\bar{p}(1-\bar{p})} \qquad \textbf{(Equation 2)}$$

in which $n_{1,2} = n_1 + n_2$, and $\bar{p} = \frac{n_1}{n_{1,2}} p_1 + \frac{n_2}{n_{1,2}} p_2$ is the mean allele frequency. Alternative estimators for F_{st} are possible, and a comprehensive comparison of different F_{st} estimators was recently reported[5].

If the two cohorts are not that different in terms of their allele frequencies, for example, the cohorts from European nations, $p_i \approx p_j \approx \bar{p}$,

$$E(F_{st}) \approx \frac{1}{n_{i,j}} + \frac{2 n_i n_j}{n_{i,j}^2} \frac{\left[E(p_i) - E(p_j)\right]^2}{\bar{p}(1-\bar{p})} \qquad \textbf{(Equation 3)}$$

At the right side of the equation, the first term represents the sampling variance for allele frequency for a pair of cohorts, and the second term represents the allele frequency difference due to divergence from a common ancestor. The estimated F_{st} is influenced by sample size, and $F_{st} \geq \frac{1}{n_{i,j}}$, which is the sampling variance of F_{st} for a pair of cohorts[1]. As each 1KG reference population has a sample size around 100, there is no disproportionate impact of sample size in calculating F_{st}.

***F_{st}* Cartographer algorithm.** The purpose of using the F_{st} Cartographer algorithm is to find the coordinates of a cohort given its F_{st} to the reference populations. The algorithm can be expressed in Cartesian geometry. Given three reference populations, a target cohort has three F_{st} measures, F_1, F_2, and F_3, respectively. Given a Cartesian coordinate system, the coordinate for these three reference populations are (a_1, b_1), (a_2, b_2), and (a_3, b_3), respectively. The algorithm tries to find the coordinates $(x, y)_{E_{i,j}}$ on each the edge $(E_{i,j})$ that connects reference populations i and j

$$(x, y)_{E_{i,j}} = \left[(a_j - a_i)\frac{F_i}{F_i + F_j} + a_i, (b_j - b_i)\frac{F_i}{F_i + F_j} + b_i\right] \qquad \textbf{(Equation 4)}$$

The coordinates of the gravity of triangle, $(x, y)_G$, that connects $E_{1,2}$, $E_{1,3}$, and $E_{2,3}$ are

$$(x, y)_G = \left(\frac{x_{E_{1,2}} + x_{E_{1,3}} + x_{E_{2,3}}}{3}, \frac{y_{E_{1,2}} + y_{E_{1,3}} + y_{E_{2,3}}}{3}\right) \qquad \textbf{(Equation 5)}$$

Inference of cohort origins at the global level. To assess genetic background, for each cohort we calculated its F_{st} values using CEU, CHB, and YRI as the reference panel, respectively. We denote these three F_{st} values as F_{CEU}, F_{CHB}, and F_{YRI}. These values reflect genetic distances between a cohort and the reference panels - the greater the value the further the genetic distance. We developed an algorithm called F_{st} cartographer, which can map a cohort to global genetic variation as previously observed using individual level data from principal component analysis[6]. The steps in the algorithm are as follows (**Supplementary Fig. 1**):

Create the coordinates for the reference samples. Without loss of generality, these three reference populations form an equilateral triangle, and we set the length of each edge to unity. For example, the coordinates CEU, CHB, and YRI are ($-\sqrt{3}$, 1), ($\sqrt{3}$, 1), and (0, -2), respectively, and connecting the coordinates of the three reference populations formed an equilateral triangle – the reference space. The gravity of this equilateral triangle is the origin of the Cartesian space. The choice for the coordinates for the reference population is arbitrary.

Step 1 Create a cohort triangle using Equation 4. Finding a point the distances of that to both ends, which represent two populations, is proportional to the ratio of the F_{st} values of the cohort to these two reference populations. Similarly, find the points on the other two edges. For example, Finland Twin Cohort (FTC) had $F_{CEU} = 0.0102$, $F_{YRI} = 0.153$, and $F_{CHB} = 0.099$. On the CEU-YRI edge, a point split the length to 0.0102:0.153, was $(-1.72, 0.98)_{E_{1,2}}$; On the CEU-CHB edge into 0.0102:0.099, was $(-1.70, 1)_{E_{1,3}}$; and on the YRI-CHB edge into 0.153:0.099, was $(1.05, -0.18)_{E_{2,3}}$. Connecting the three coordinates created a "FTC" triangle inside the reference triangle.

Step 2 Find the gravity of the cohort triangle using Equation 5. The gravity of the "FTC" triangle had its coordinates of $(-0.79, 0.60)_{G_{FTC}}$, which is inferred as the geographic coordinates for FTC in F_{PC} space. It had relative distances of 1.03, 2.55, and 2.72 to CEU, CHB, and YRI, respectively. The shorter the distance, the closer the genetic background is.

Step 3 Repeat Steps 1, and 2 until the gravity of each cohort is found.

Plots of the coordinates for each cohort will show the relative distance of each cohort to the reference samples. If a cohort has equal distances to three reference populations, its gravity will be close to the origin of the reference triangle.

Method II: Principal component analysis for cohort-level allele frequencies

PCA has been widely used in genetics[7] and recently proposed for controlling population stratification for GWAS[8,9]. We provide a new method that uses cohort-level allele frequencies, often provided as summary statistics in meta-analysis. We call the new method as meta-PCA.

Meta-PCA is based on a $G = (C + K) \times M$ matrix, which includes K reference populations and C cohorts of question on M markers. In G, the m^{th} column represents the reported reference allele frequencies for the m^{th} marker for (K+C) cohorts. The kernel correlation matrix for PCA is constructed on $\Sigma = G_s \times G_s^T$, in which G_s is the standardization for G for each locus (on each column of G). Compared with individual-level data PCA, in the context of meta-PCA each cohort can be viewed as an individual in the conventional sense. Given Σ matrix, all the implementation is the same as the individual-data PCA.

There are efforts in establishing genetic interpretation for PCA [8,10–12]. The interpretation of meta-PCA could be approached by F_{pc} as described in the last section.

Method III: The detection of overlapping samples with λ_{meta}

Inference of cohort origins at the within-Europe level. To assess genetic background, for each cohort we calculated its F_{st} values using CEU, FIN, and TSI as the reference panel, with coordinates $(-\sqrt{3}$, 1), ($\sqrt{3}$, 1), and (0, -2), respectively. For FTC, it had F_{st} values of 0.0102, 0.0052, and 0.0157, to CEU, FIN, and TSI, respectively. Using the F_{st} Cartographer algorithm, the gravity of the FTC triangle had its coordinates of $(0.274, 0.361)_{G_{FTC}}$. It had relative distances of 2.10, 1.59, and 2.42, to CEU, FIN, and TSI, respectively.

Genealogical subspace. Furthermore, we partition the F_{PC} space into three subspaces. For example, given coordinates of $(-\sqrt{3}$, 1), ($\sqrt{3}$, 1), and (0, -2), for CEU, FIN, and TSI, respectively, connecting the origin and the coordinates for any two reference populations created a subspace, which is defined as a genealogical subspace. We had three genealogical subspaces: CEU-FIN genealogical subspace, CEU-TSI genealogical subspace, and FIN-TSI genealogical subspace, respectively. If a cohort is located inside a subspace, it indicates that this cohort may be derived from these two reference populations that creates the genealogical subspace.

For European cohorts, the coordinates calculated from F_{st} Cartographer algorithm mirror the origins of geographic locations of the cohorts, similar, but less refined, to what has been observed in previous studies using individual level data for European samples[13,14].

Effective number of overlapping samples ($\boldsymbol{n_o}$). If a pair of cohorts has overlapping samples, it leads to a correlation of the estimated genetic effects for each locus. In the recent literature, two kinds of correlation due to overlapping samples were introduced. The first one was defined by directly calculating correlation between all estimated test statistics, $r = cor(Z_1, Z_2)$, in which Z is a vector of M matched loci between two cohorts [15,16]. The second one was defined on the correlation for single locus given overlapping samples, as introduced by Lin and Sullivan[17]. We used the second definition, and then extended the correlation due to any relatives, a generalization of Lin and Sullivan.

For a pair of cohorts of sample sizes n_1 and n_2 ($n_1 \geq n_2$), for M matched loci which have GWAS summary statistics, for example additive effects and their standard errors. For the m^{th} locus, estimated association effect sizes are $b_{1.m}$ and $b_{2.m}$ with sampling variance $\sigma^2_{b_{1.m}}$ and $\sigma^2_{b_{2.m}}$, respectively. b_1 is assumed to be drawn from a normal distribution $N(b_{1.m}, \sigma^2_{b_{1.m}})$, and

$b_2 \sim N(b_{2.m}, \sigma^2_{b_{2.m}})$. In cohort 1, $w_{1|k} = \frac{n_{12|k}}{n_1}$ is proportion of samples with a k^{th}-degree relatives in cohort 2 with $n_{12|k}$ the number of relatives of kth degree relatives shared between the samples; the phenotypic variance is assumed to be the same across the cohorts for a quantitative trait. For a locus, the genetic effect is estimated by linear regression $y_1 = a + b_1 x + e$ in cohort 1 (the index for the locus is dropped for convenience). If the sampling variance of a locus is assumed to be the same for any subset of samples

$$b_1 = \Sigma_{k=0}^{K} w_{1|k} \frac{\left[E\left(y_{1|k}x_k\right) - E\left(y_{1|k}\right)E\left(x_k\right)\right]}{var(x_m)} = \Sigma_{k=0}^{K} w_{1|k} b_{1|k}$$

The standard error of b_m is $\sigma_{b_1} = \sqrt{\frac{\left(1-h^2_{b_1}\right)\sigma^2_{y_1}}{n_1}} \approx \sqrt{\frac{\sigma^2_{y_1}}{n_1}}$, in which $h^2_{b_1}$ is the proportion of phenotypic variance explained by the locus and $\sigma^2_{y_1}$ is the phenotypic variance of the trait. The sampling variance for $\sigma_{b_{1.k}} = \sqrt{\frac{\sigma^2_{y_1}}{w_{1|k}n_1}}$. This decomposition of the genetic effect can be applied to cohort 2. Consequently, the covariance between b_1 and b_2 for the locus is

$$cov(b_1, b_2) = cov\left(\Sigma_{k=0}^{K} w_{1|k} b_{1.k}, \Sigma_{k=0}^{K} w_{2|k} b_{2.k}\right) = \Sigma_{k=0}^{K} w_{1|k} w_{2|k} cov\left(b_{1|k}, b_{2|k}\right)$$

in which $cov\left(b_{1|k}, b_{2|k}\right) = \rho_k \theta_k \sigma_{b_1|k} \sigma_{b_2|k}$ is the covariance between the genetic effects estimated in two cohorts due to the k-degree relatives. ρ_k is the phenotypic correlation for the k-degree relatives, and θ_k is the genetic relatedness for the k-degree relatives. $\theta_k = \left(\frac{1}{2}\right)^k$ is the coefficient of identity for descent. For duplicated samples, $\rho_0 = h^2 + \rho_{e|0}$, in which h^2 is the heritability, and $\rho_{e|0}$, the environmental correlation to be close 1 for overlapping samples; for other relatives ($k \geq 1$), $\rho_k \approx \theta_k h^2$.

Correlation between the estimated genetic effects. The covariance can be generalized as $cov(b_1, b_2) = \Sigma_{k=0}^{K} \sqrt{w_{1|k} w_{2|k}} \rho_k \theta_k \sqrt{\frac{\sigma^2_{y_1}}{n_1} \frac{\sigma^2_{y_2}}{n_2}}$. After adjustment by the sampling variance, the correlation between b_1 and b_2 is

$\rho_{b_1,b_2} = \frac{cov(b_1,b_2)}{\sigma_{b_1}\sigma_{b_2}} = \Sigma_{k=0}^{K} \rho_k \theta_k \sqrt{w_{1|k} w_{2|k}} = \frac{\Sigma_{k=0}^{K} \rho_k \theta_k n_{12|k}}{\sqrt{n_1 n_2}} = \frac{n_o}{\sqrt{n_1 n_2}}$ **(Equation 6)**

in which $n_o = \Sigma_{k=0}^{K} \rho_k \theta_k n_{12|k}$, is the effective number of overlapping samples averaged over all relative pairs that are across the two cohorts. As the variance explained by each locus is small, and after further weighted by θ_k, the contribution from overlapping relative is small. When ignoring the first and higher degree relatives n_o equals the contribution from overlapping samples. This is consistent with the results from Lin and Sullivan[17], who considered overlapping samples only. So, the correlation at any single locus is largely determined by the overlapping samples ($n_{12|0}$) for summary statistics.

$$\rho_{b_1,b_2} = \frac{n_o}{\sqrt{n_1 n_2}} \approx \frac{n_{12|0}}{\sqrt{n_1 n_2}}$$ **(Equation 7)**

So, in the text hereafter, n_e indicates overlapping samples only, otherwise specified.

Correlation for case-control studies. The theory above is based on a quantitative trait, but it holds approximately true for case-control studies if a locus is from the null distribution of no association with the disease. Given $n_{12.ctrl}$ overlapping controls and $n_{12.cs}$ overlapping cases, for a locus associated with disease its correlation of the regression coefficient is $\rho_{b_1,b_2} = \frac{n_{12.ctrl}\sqrt{R_1 R_2} + n_{12.cs}\frac{1}{\sqrt{R_1 R_2}}}{\sqrt{n_1 n_2}}$ as indicated by Lin and Sullivan[17], in which R_i is the ratio between cases and controls in the i^{th} cohort. When it is balanced case-control design – $R. = 1$, $\rho_{b_1,b_2} = \frac{n_{12.ctrl} + n_{12.cs}}{\sqrt{n_1 n_2}} = \frac{n_o}{\sqrt{n_1 n_2}}$ resembles the correlation for quantitative traits. However, it should be noticed that for case control data, n_e is confounded with the number of overlapping cases and controls.

Theory for $\boldsymbol{\lambda_{meta}}$. For the summary statistics between a pair of cohorts for the m^{th} locus, we can construct a statistic

$$T_m = \frac{(b_{1.m} - b_{2.m})^2}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}}} = \left[\frac{(b_{1.m} - b_{2.m})^2}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}} - 2\rho_{1,2}\sigma_{b_{1.m}}\sigma_{b_{2.m}}}\right] \times \left[\frac{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}} - 2\rho_{1,2}\sigma_{b_{1.m}}\sigma_{b_{2.m}}}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}}}\right]$$ **(Equation 8)**

in which $\rho_{1,2}$ is the correlation between $b_{1,m}$ and $b_{2,m}$.

$$E(T_m) = \left\{\frac{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}} - 2\rho_{1,2}\sigma_{b_{1.m}}\sigma_{b_{2.m}}}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}} - 2\rho_{1,2}\sigma_{b_{1.m}}\sigma_{b_{2.m}}} + \frac{[E(b_{1.m}) - E(b_{2.m})]^2}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}} - 2\rho_{b_1,b_2}\sigma_{b_{1.m}}\sigma_{b_{2.m}}}\right\}\left\{\frac{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}}}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}}} -\right.$$

ρ1,22σb1.mσb2.mσb1.m2+σb2.m2=(1+H)(1−ρ1,2κ) **(Equation 9)**

in which $H = \frac{[E(b_{1.m}) - E(b_{2.m})]^2}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}} - 2\rho_{1,2}\sigma_{b_{1.m}}\sigma_{b_{2.m}}}$, $\kappa = \frac{2\sigma_{b_{1.m}}\sigma_{b_{2.m}}}{\sigma^2_{b_{1.m}} + \sigma^2_{b_{2.m}}} = \frac{2\sqrt{n_1 n_2}}{n_1 + n_2}$, and $\rho_{1,2} = \frac{n_o}{\sqrt{n_1 n_2}}$, as defined in Equation 7, is the correlation for this locus due to overlapping samples between this pair of cohorts. Of note, $\rho_{1,2}$ is same for each locus regardless of a null locus or a locus associated to genetic effects. For convenience, the subscript b was dropped in the text hereafter.

Under the null hypothesis of no heterogeneity ($H = 0$) and no correlation ($\rho_{1,2} = 0$), $T_0 \sim \chi^2_1$, a standard 1-degree-of-freedom chi-square distribution. $\rho_{1,2} = \frac{n_o}{\sqrt{n_1 n_2}}$, in which n_o is the effective number of overlapping samples. Of note, since the majority of markers are likely sampled from the null distribution or have very small effect sizes, we can approximate $E(b_{1.m}) = 0$ and $E(b_{2.m}) = 0$, and therefore $H \approx 0$ for most marker pairs between a pair of cohorts. For the m^{th} marker that is in linkage disequilibrium with causal variants, $E(b_{1.m}) = \sum_{j=1}^{J_1} \beta_{1.j} \ell_{1.j}$, in which J_1 is the number of causal variants in linkage disequilibrium with the m^{th} marker for cohort 1, $\beta_{1.j}$ is the j^{th} causal variants in linkage disequilibrium with the m^{th} marker, and $\ell_{1.j}$ is the LD correlation between the m^{th} marker and the j^{th} causal variant[18]. Similarly for $E(b_{2.m}) = \sum_{j=1}^{J_2} \beta_{2.j} \ell_{2.j}$. If the cohorts are from the same ethnicity, the difference in the LD correlation can be ignored, for example for samples from cohorts with European ancestry. So, under a polygenic model H is expected to be zero, or close to zero.

The T statistic is calculated for each matched SNP between a pair of cohorts. After ordering all T values, we evenly sample 30,000 independent markers from the order statistic of all T values. Each pair of cohorts may sample T values based on 30,000 markers different from another pair of cohorts.

$$\lambda_{meta} = \frac{median(T)}{median(\chi_1^2)} = 1 - \frac{2n_o}{n_1+n_2}$$ **(Equation 10)**

in which $median(\chi_1^2) = 0.4549$. Under the null hypothesis of no heterogeneity and overlapping samples ($n_o = 0$), plotting the ordered T against its corresponding quartiles from χ_1^2, will be along the diagonal, leading to $\lambda_{meta} = 1$. Heterogeneity between two cohorts, equivalent to a “negative” number of overlapping samples, will drive $\lambda_{meta} > 1$, and overlapping samples will make $\lambda_{meta} < 1$. The distribution of λ_{meta} can be assessed via the beta distribution, and λ_{meta} follows asymptotically a normal distribution $N(1, 0.0136)$ given 30,000 independent markers.

Factors that influence $\boldsymbol{\lambda_{meta}}$. A number of factors will influence the λ_{meta}. 1) Sample overlap, including close relatives across cohorts, reduces the value of λ_{meta} (**Supplementary Fig. 2**) Conservative modeling, such as inclusion of covariates in the association model that are genetically correlated with the phenotype or the ‘genomic control’ approach (adjusting the sampling variance with λ_{GC}, $z = \frac{b}{\sqrt{\lambda_{GC}}\sigma}$), will inflate the sampling variance, and deflate λ_{meta}. 3) Genetic heterogeneity, which can be caused by differences in genetic architecture or methodological difference, will inflate λ_{meta}. 4) As characterized by Equation 10, the lower bound (cohort 2 is completed included in cohort 1, given $n_1 > n_2 = n_o$) of λ_{meta} is $1 - \frac{2}{\frac{n_1}{n_2}+1}$, upon the ratio of the samples sizes of the two cohorts.

Estimating overlapping samples. As shown in Equation 10, λ_{meta} is a linear function of n_o, hence the statistical power to detect overlapping samples is equivalent to asking how λ_{meta} departs from the null distribution. Assuming $H = 0$, the overlapping samples can be estimated as $\hat{n}_o = (1 - \hat{\lambda}_{meta})\frac{(n_1+n_2)}{2}$, and $\sigma_{\hat{n}_o} = \frac{n_1+n_2}{2} \times 0.0136 \approx 0.0068(n_1 + n_2)$ given 30,000 independent markers. Hence, using summary statistics only the proportion of overlapping samples can be estimated for quantitative traits. Given the type I error rate of 0.05 ($\alpha = 0.05$), the statistical power for detecting $\tilde{n}_o$ overlapping samples between two cohorts is $p = \Phi^{-1}(T, \tilde{n}_o, \sigma_{\tilde{n}_o})$, in which Φ^{-1} represents the accumulation power function of a normal distribution with the mean of $\tilde{n}_o$ and standard deviation of $\sigma_{\tilde{n}_o}$. The statistical power is determined by T, the threshold for significance, $\tilde{n}_0$, the real overlapping samples, and $\sigma_{\tilde{n}_o}$, the standard deviation of the null hypothesis that there is no overlapping samples. Without loss of generality, $T = 1.96\sigma_{\tilde{n}_o} \approx 0.13(n_1 + n_2)$ given $\alpha = 0.05$. The 95% confidence interval is $[-0.13(n_1 + n_2), 0.13(n_1 + n_2)]$. The statistical power is maximized when $n_1 = n_2$, i.e. when a pair of cohorts has the same sample size.

For case-control studies, as $\hat{n}_o = n_{12.ctrl}\sqrt{R_1R_2} + n_{12.cs}\frac{1}{\sqrt{R_1R_2}}$, the estimate cannot distinguish between overlapping cases and overlapping controls; when $R_1 = 1$ and $R_2 = 1$ (balanced case-control design for both cohorts), $\hat{n}_o = n_{12.ctrl} + n_{12.cs}$, indicating the overall overlapping samples between two cohorts, summed across cases and controls. If we know that only controls (cases) were shared between two cohorts, then $\hat{n}_o = n_{12.ctrl}\sqrt{R_1R_2}$ ($\hat{n}_o = n_{12,cs}\frac{1}{\sqrt{R_1R_2}}$), so then an estimate of n_o indicates the number of overlapping controls (cases). Therefore, quantifying overlapping samples for case-control studies is more difficult than that for quantitative traits.

Method IV: Pseudo profile score regression (PPSR)

PPSR resembles the previously proposed Gencrypt method[19], but PPSR is more powerful in detecting various degree of relatives and more robust to missing data and imputation errors. For each individual, the PPS can be generated as below

$A_i = S \times G_i$ **(Equation 11)**

in which A_i is the PPS for the i^{th} individual, S is a $K \times M$ score matrix, and G_i is vector for the genotypes for the chosen M loci.

In detail,

$$\begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{iK} \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & s_{22} & \cdots & s_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{K1} & s_{K2} & \cdots & s_{KM} \end{bmatrix} \begin{bmatrix} g_{i1} \\ g_{i2} \\ \vdots \\ g_{iM} \end{bmatrix}$$

in which a_{ik} is the k^{th} profile score for the i^{th} individual, s_{km} is the additive effect at the m^{th} locus (m from 1 to M) for the k^{th} profile score, and g_{im} is the standardized genotype at the k^{th} locus for the i^{th} individual. Each s, the pseudo genetic effect, follows a standard normal distribution $N(0,1)$; each pseudo genetic effect is independent to another. For each PPS, $var(a_i) = \Sigma_{m=1}^{M} var(g_{im} s_{k.}) = \Sigma_{m=1}^{M} g_{im}^2 var(s_{k.}) = M$, in which $s_{k.}$ is the k^{th} column for the S matrix, and on average each locus explains $\frac{1}{M}$ of the variation. For an individual a pair of PPS, say a_{l1} and a_{l1}, has $cov(a_{i1}, a_{i2}) = \sum_{m=1}^{M} g_{im}^2 cov\left(s_{l_1 m}, s_{l_2 m}\right) = 0$.

Each PPS can be seen as a trait with $h^2 = 1$ because it does not have any sampling variance. For a pair of individuals, individual i and individual j, when both A_i and A_j have been standardized, their covariance for the k^{th} PPS $cov\left(a_{i1}, a_{j1}\right) = \theta h^2$, in which θ is the relatedness scores in terms of identity by state[20]. Depending on the relatedness between a pair of individuals, $\theta = 1$ for monozygous twins or to a duplicated sample, $\theta = 0.5$ for first-degree relatives such as parent and offspring or full sibs. In general, for r^{th}-degree of relatives, $E(\theta_r) = 0.5^r$.

The theory presented above provides a theoretical basis for detecting overlapping samples using PPS other than sharing individual level genotypes. Assuming that each individual has K independent PPS (A_i having K elements), for individual i and j, we can regress A_i on A_j,

$A_i = \mu + bA_j + e_{ij}$ **(Equation 12)** in which μ is the grand mean, b is the regression coefficient, and e_{ij} is the residual. $E(b) = \frac{cov(A_i, A_j)}{var(A_j)} = \theta_r$. $E(b) = 0$ if individual i is not correlated with individual j, $E(b) = 0.5$ for first-degree relatives, and $E(b) = 1$ if individual i and j are genetically same, say an overlapping sample or the homozygous twins. The sampling variance of b is $\sigma_b^2 = \frac{\sigma_{A_i}^2 - \sigma_{A_j}^2 \theta_r^2}{\sigma_{A_j}^2 K} = \frac{1-\theta_r^2}{K}$. Under the null distribution for no related or overlapping samples, $b \sim N(0, \frac{1}{K})$. The residual e_{ij} accounts the discordant genotypes, including missing genotypes and genotyping or imputation errors. For current GWAS data, after quality control, the discordant rate is often smaller than 1%.

If now we have C cohorts for which the individual genotypes of which cannot be disclosed to the central analysis hub, overlap between cohorts can be identified if PPS are supplied. By regressing their PPS to each other the overlapping individuals could be detected if $b \approx \theta_r$. Assuming there are N_c samples in each cohort, a total of $N = \sum_{c_1=1}^{C} \sum_{c_2>c1}^{C} N_{c_1} \times N_{c_2}$ regressions need to be carried out as defined in Equation 12. If we want to control the experiment-wise type I error rate α under the null hypothesis and type II error rate β (with power$= 1 - \beta$) for $b = \theta_r$, the required number of pseudo profile scores for each individual is

$$K \geq \left(\frac{z_{(1-\beta)}\sqrt{1-b^2} + z_{(1-\alpha)}}{b}\right)^2 \qquad \textbf{(Equation 13)}$$

in which $z_{(1-\beta)}$ and $z_{(1-\alpha)}$ are z scores under the given p-values at the subscripts. To accommodate technical errors, such as missing genotypes and genotype error, a cutoff of 0.95 for b is adopted for detection of overlapping samples, and $0.4 \sim 0.45$ for detecting first-degree relative.

The standardization of genotypes can either use the allele frequency from each cohort, or from a reference sample. Throughout the study, we used the allele frequency calculated from WTCCC bipolar disorder cohort as the reference, and using it as an approximation to standardize genotypes for all cohorts in comparison.

Workflow for PPSR. Given the statistical method for detecting overlapping samples as described above, the whole workflow for detecting can be split into three steps (**Supplementary Fig. 9**).

In step 1, the required type I and type II error rates are defined and from that the required number of pseudo profiles to be generated. The GWAMA central analyst selects consensus SNP markers across cohorts, and determines additive effects matrix S that will be used to generate pseudo profile scores for each cohort. In order to avoid strand issues, the loci having palindromic loci (A/T alleles or G/C alleles) are excluded.

In step 2, each cohort generates PPSR for each individual with the set of consensus markers and the marker weights received from the GWAMA coordinator. After generate the PPS, they send them back to the coordinator. This will be a file that contains N rows and K columns with pseudo-profile scores.

In step 3, the coordinator runs PPSR for each sample in a cohort on each PPS generated for another cohort. The final product of running PPSR is to generate a $n_i \times n_j$ matrix for a pair of cohorts, which have n_i and n_j samples respectively. For each pair of individuals in comparison, we take the one from cohort i as the response variable and from cohort j as the predictor variable in PPSR. In principle, swapping the response variable and the predictor variable do not affect the performance of PPSR. Each entry, the regression coefficient of PPSR, in the $n_i \times n_j$ matrix represents genetic similarity for these pair of individuals in comparison. Once the regression coefficients are above the threshold, it indicates there are samples duplicated. The central analyst can then request each cohort that is implicated in containing samples that are also in other cohorts to drop those samples, without revealing where the duplication occurred.

Privacy issues when using PPSR. As the exchange of the PPS is within a meta-analysis facility, it is not as vulnerable as that of releasing the GWAS summary to the public domain as discussed in previous studies[21–23]. However, as PPS are generated from genotypes, it is worth to consider whether the PPS will reveal individual genotype information, or can be decoded from PPS. As a demonstration for the principle-of-proof, we consider to reverse Equation 11 to estimate genotypes. We consider the case where the additive effect matrix in Equation 11 is known, otherwise it is nearly impossible to recover genotype information. Given the workflow of PPSR, the analysts who coordinate the meta-analysis know the additive effect matrix, S in Equation 11, and receive PPS from each cohort have the information to decode genotypes that are employed to generate PPS.

After reversing Equation 11, using the standard regression method, the genotype in each locus can be estimated as

$A_i = \mu + g_{im} \times s_{.m} + e$ **(Equation 14)**

In detail,

$$\begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{iK} \end{bmatrix} = \mu + g_m \begin{bmatrix} s_{im} \\ s_{im} \\ \vdots \\ s_{im} \end{bmatrix} + e$$

in which $s_{.m}$ is the m^{th} column in the additive effects matrix in Equation 11. Although $E(g_{im}) = g_{im}$, which is an unbiased estimate of the genotype, its sampling variance is $\sigma_{g_{im}} = \sqrt{\frac{\Sigma_{m=1}^{M}[1-(1-p_m)^2]\sigma_{s.m}^2}{K}}$. The sampling variance can be further written as $\sigma_{g_{im}} = \sqrt{\frac{M}{K}E(\mathcal{P}_m)}$ because $\sigma_{s.m}^2 = 1$ and $[1-(1-p_m)^2]$ is denoted as $\mathcal{P}_m$. The greater the ratio between $\frac{M}{K}$ and $E(\mathcal{P}_m)$, the larger the sampling variance, and consequently the lower probability to construct the real genotype.

Without loss of generality, the accuracy of the estimated $\hat{g}$, a continuous variable, and g, a discrete variable with values of 2, 1, and 0, can be measure using the squared correlation $(R^2)^{24}$,

$R^2 = \frac{E[\sigma_{\hat{g}}^2]}{E[\sigma_{\hat{g}}^2]+\frac{M}{K}E[g^2]}$ **(Equation 15)**

in which $E(g^2)$ and $E(\sigma_g^2)$ are:

$$E(g^2) = \tilde{p}_{AA}x_{AA}^2 + \tilde{p}_{Aa}x_{Aa}^2 + \tilde{p}_{aa}x_{aa}^2$$

$$E\left(\sigma_g^2\right) = \tilde{p}_{AA}(x)(x_{AA} - 2p)^2 + \tilde{p}_{Aa}(x_{Aa} - 2p)^2 + \tilde{p}_{aa}(x_{aa} - 2p)^2$$

$x_{AA} = 2$, $x_{Aa} = 1$, and $x_{aa} = 0$ if A is the reference allele, and $\tilde{p}_{AA}$, $\tilde{p}_{Aa}$, and $\tilde{p}_{aa}$ are weighted frequency given the distribution of g. $f = \tilde{p}_{AA} + 0.5\tilde{p}_{Aa}$.

When the reference allele frequency follows a uniform distribution between (a_1, a_2), assuming that the loci follow Hardy-Weinberg proportions, $p_{AA} = p^2$, $p_{Aa} = 2pq$, and $p_{aa} = q^2$, in which p follows a uniform distribution between a_1 and a_2 and $q = 1 - p$.

$$p_{AA} = \int_{a_1}^{a_2} p^2 = \frac{1}{3}p^3|_{a_1}^{a_2} = \frac{1}{3}(a_2^3 - a_1^3)$$

$$p_{Aa} = \int_{a_1}^{a_2} 2pq = \left(p^2 - \frac{2}{3}p^3\right)|_{a_1}^{a_2} = (a_2^2 - a_1^2) - \frac{2}{3}(a_2^3 - a_1^3)$$

$$p_{aa} = \int_{1-a_2}^{1-a_1} q^2 = \frac{1}{3}q^3|_{1-a_2}^{1-a_1} = \frac{1}{3}[(1-a_1)^3 - (1-a_2)^3]$$

and $\tilde{p}_{AA} = \frac{p_{AA}}{p_{AA}+p_{Aa}+p_{aa}}$, $\tilde{p}_{Aa} = \frac{p_{Aa}}{p_{AA}+p_{Aa}+p_{aa}}$, and $\tilde{p}_{aa} = \frac{p_{aa}}{p_{AA}+p_{Aa}+p_{aa}}$.

If the reference allele frequency follows a uniform distribution between (0, 0.5), $R^2 = \frac{\frac{5}{12}}{\frac{5}{12}+\frac{2M}{3K}} = \frac{5}{5+8\frac{M}{K}}$.

Given M loci with MAF of 0.5, the expected frequencies for AA, Aa, and aa are $\tilde{p}_{AA} = 0.25$, $\tilde{p}_{Aa} = 0.5$, $\tilde{p}_{aa} = 0.25$, and $f = 0.5$. $E(g^2) = 1.5$, and $E\left(\sigma_g^2\right) = 0.5$. Plugging them in to the Equation 13 leads to $R^2 = \frac{0.5}{0.5+1.5\frac{M}{K}} = \frac{1}{1+3\frac{M}{K}}$.

Equation 13 can be rewritten as $R^2 = \frac{1}{1+\varphi\frac{M}{K}}$, in which $\varphi = 3$ if MAF is 0.5, and $\varphi = 1.6$ if MAF in nearly from a uniform distribution. From Equation 13, it is easy to calculate the ratio between the number of markers and the number of PPS given a controlled R^2,

$\frac{M}{K} \geq \frac{1-R^2}{\varphi R^2}$ **(Equation 16)**

For uniform distribution of MAF, if $R^2 \leq 0.1$ is set as the threshold, $\frac{M}{K} \geq 5.4$; if $R^2 \leq 0.05$, $\frac{M}{K} \geq 11.4$, and if $R^2 \leq 0.01$, $\frac{M}{K} \geq 59.4$. In general, the higher the ratio between M and K, the less information can be inferred. We suggest $\frac{M}{K} \geq 5 \sim 10$ may be sufficient.