

**New QC metrics for GWAS meta analysis: Supplementary Figures**  
**Across-cohort QC analyses of GWAS summary statistics from complex traits**

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

**Affiliations:**

<sup>1</sup> Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia

<sup>2</sup> School of Environmental and Rural Science, The University of New England, Armidale, New South Wales, Australia

<sup>3</sup> SPLUS Game, Guangzhou, Guangdong, China

<sup>4</sup> Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany

<sup>5</sup> Medical Research Council (MRC) Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK

<sup>6</sup> Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA

<sup>7</sup> Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

<sup>8</sup> Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK

<sup>9</sup> Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA

<sup>10</sup> Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland

<sup>11</sup> Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland

<sup>12</sup> Swiss Institute of Bioinformatics, Lausanne, Switzerland

<sup>13</sup> The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA

<sup>14</sup> The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA

<sup>15</sup> The Genetics of Obesity and Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, New York, USA

<sup>16</sup> Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

<sup>17</sup> Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

<sup>18</sup> Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, USA

<sup>19</sup> Division of Endocrinology, Boston Children's Hospital, Boston, Massachusetts, USA

<sup>20</sup> A full list of members is available in the **Supplementary Note**

<sup>21</sup> The University of Queensland Diamantina Institute, Translation Research Institute, Brisbane, Queensland, Australia

**Correspondence should be addressed to**

GBC (chengguobo@gmail.com) or PMV (peter.visscher@uq.edu.au)

## New QC metrics for GWAS meta analysis: Supplementary Figures

### Supplementary Figures

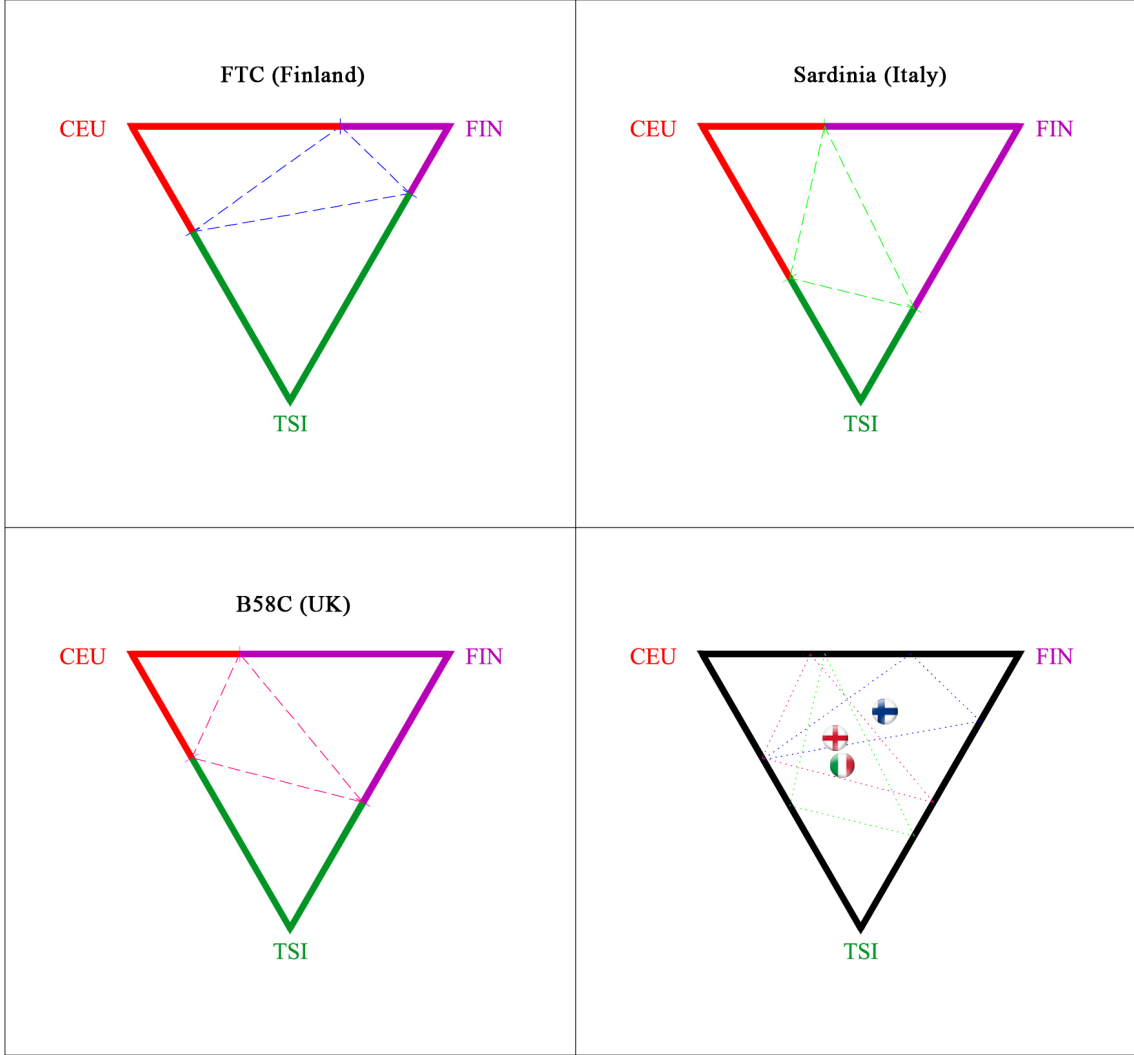
<b>Figure S1 Demonstration for Fst cartographer algorithm.....</b>	<b>3</b>
<b>Figure S2 Inference for geographic locations for GIANT height GWAS cohorts with the inclusion of MIGN. ....</b>	<b>4</b>
<b>Figure S3 Fst for GIANT height cohorts using 1000 Genome European as reference.....</b>	<b>5</b>
<b>Figure S4 Comparison between Meta-PCA and genotype PCA on 1000 Genome samples using nearly 1M SNPs.....</b>	<b>6</b>
<b>Figure S5 <math>\lambda_{\text{meta}}</math> and the effective number of overlapping samples (no).....</b>	<b>7</b>
<b>Figure S6 <math>\lambda_{\text{meta}}</math> for PGC schizophrenia summary statistics.....</b>	<b>8</b>
<b>Figure S7 <math>\lambda_{\text{meta}}</math> for rheumatoid arthritis summary statistics between Eueropean samples and Asian samples. ....</b>	<b>9</b>
<b>Figure S9 <math>\lambda_{\text{meta}}</math> and <math>\lambda_{\text{gc}}</math> for GIANT height GWAS cohorts. ....</b>	<b>11</b>
<b>Figure S10 Statistical power for detecting overlapping samples between a pair of cohorts given type I error rate of 0.05. ....</b>	<b>13</b>
<b>Figure S11 Using <math>\lambda_{\text{meta}}</math> constructed either on genetic effects or on allele frequency to estimate overlapping samples between WTCCC 7 diseases.....</b>	<b>14</b>
<b>Figure S12 Workflow for PPSR regression. ....</b>	<b>15</b>

## New QC metrics for GWAS meta analysis: Supplementary Figures

**Figure S1 Demonstration for  $F_{st}$  cartographer algorithm.**

a) For example, for FTC cohort from Finland, its  $F_{st}$  to CEU=0.010,  $F_{st}$  to FIN = 0.0051,  $F_{st}$  to TSI=0.0157. So a point cutting the CEU-FIN edge into 0.010:0.0051 is found first, and similarly to find the points on another two edges. Connecting the three points made FTC triangle. The gravity of the FTC triangle is the inferred geographic coordinates for FTC.

The algorithm can be applied to b) Sardinia from Italy, and c) B58C cohort from England, to find their coordinates. d) A coarse map can be constructed.

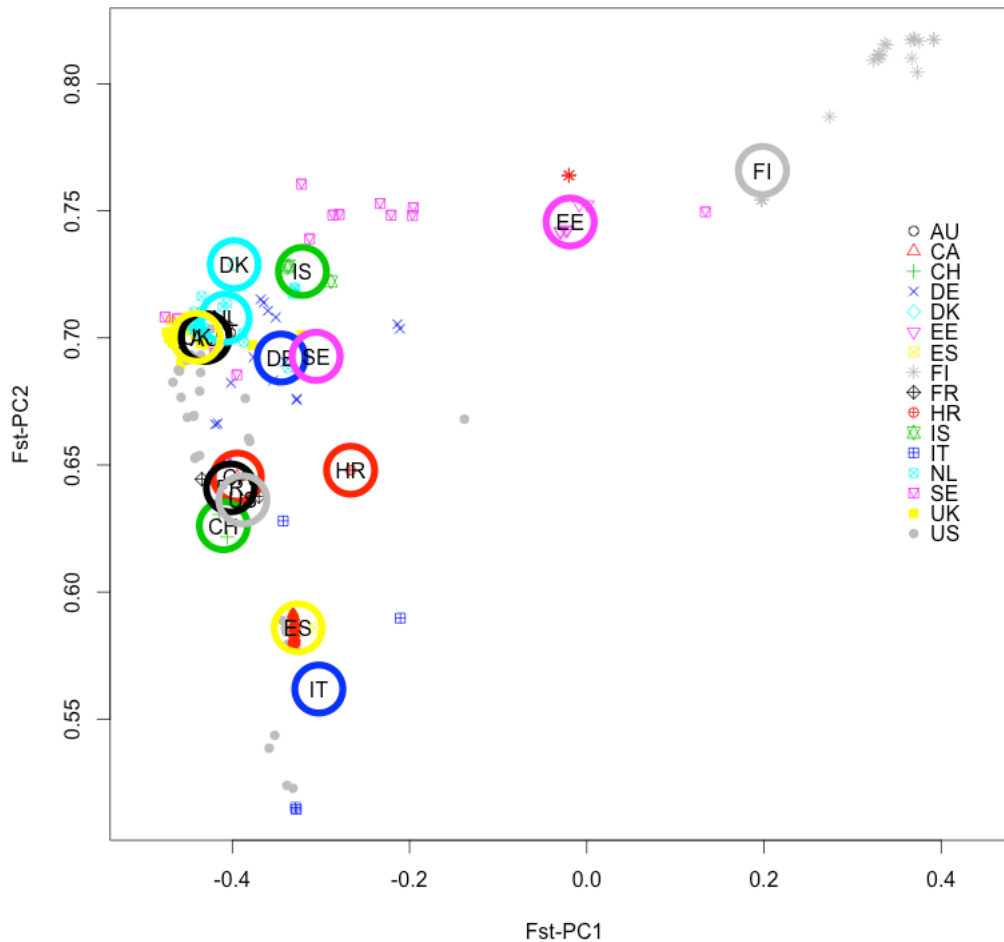


Cohort	Nation	$F_{st}$ to the reference samples		
		CEU	FIN	TSI
FTC.WOMEN	Finland	0.010	0.0051	0.0157
SARDINIA.ALL	Italy	0.0017	0.00286	0.00138
B58C-WTCCC	England	0.0021	0.0052	0.0040

### New QC metrics for GWAS meta analysis: Supplementary Figures

**Figure S2 Inference for geographic locations for GIANT height GWAS cohorts with the inclusion of MIGN.**

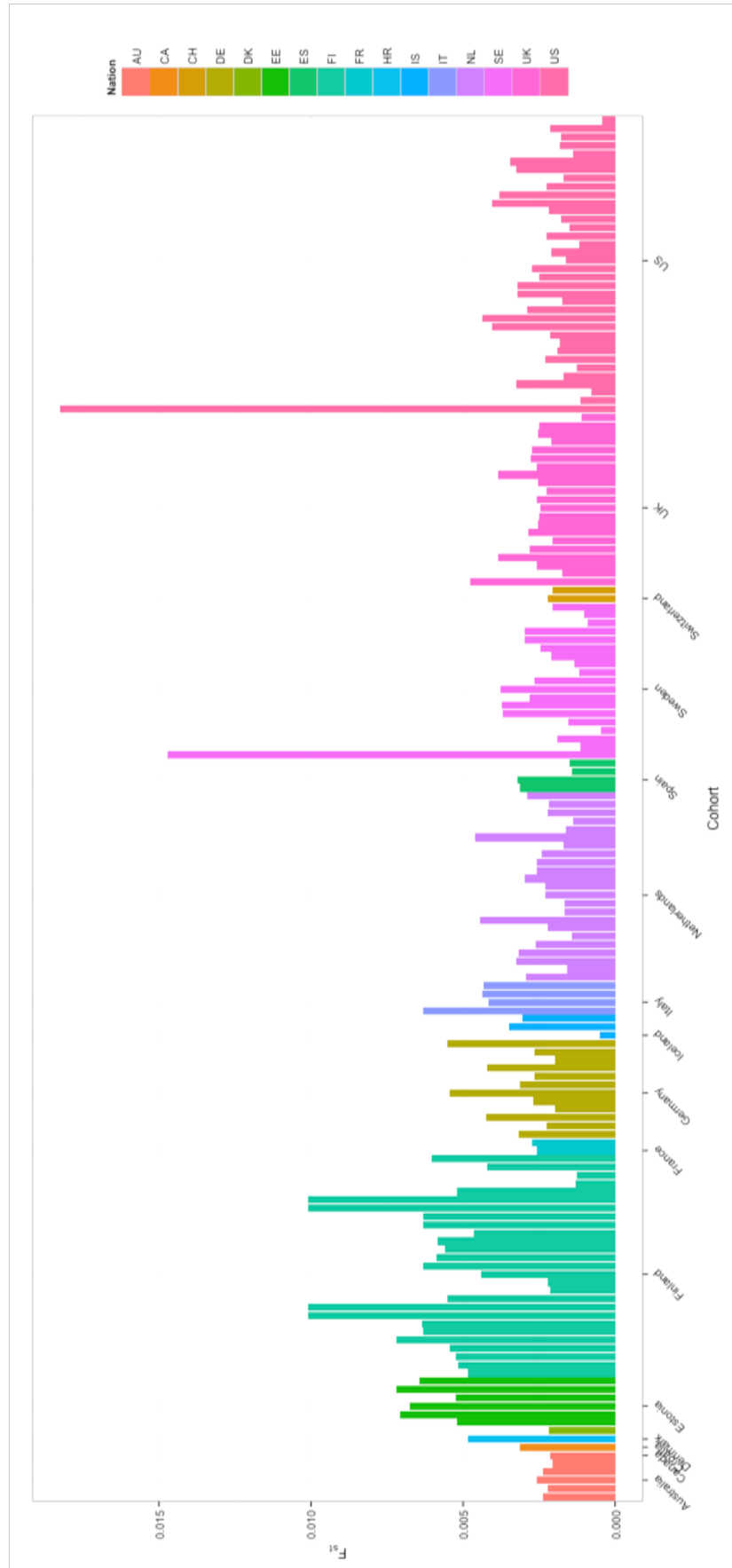
The open circles represent the mean of the inferred geographic locations for the cohorts from the same country. Red triangles (near ES circle – the 1000 Genomes Project IBS cohort) represent MIGN cohorts, which are from Sweden, Finland, USA, and Spain. However, they are all filled with the same allele frequencies in the summary statistics uploaded to the data sever.



## New QC metrics for GWAS meta analysis: Supplementary Figures

**Figure S3  $F_{st}$  for GIANT height cohorts using 1000 Genome European as reference.**

Using all 1000 Genomes Project European samples as the reference panel – an “averaged” European population,  $F_{st}$  is calculated for each cohort. Amish population (from US) and NSPHS population (from Sweden) shows high  $F_{st}$ . Cohorts from the same nations were arranged together on the x-axis.

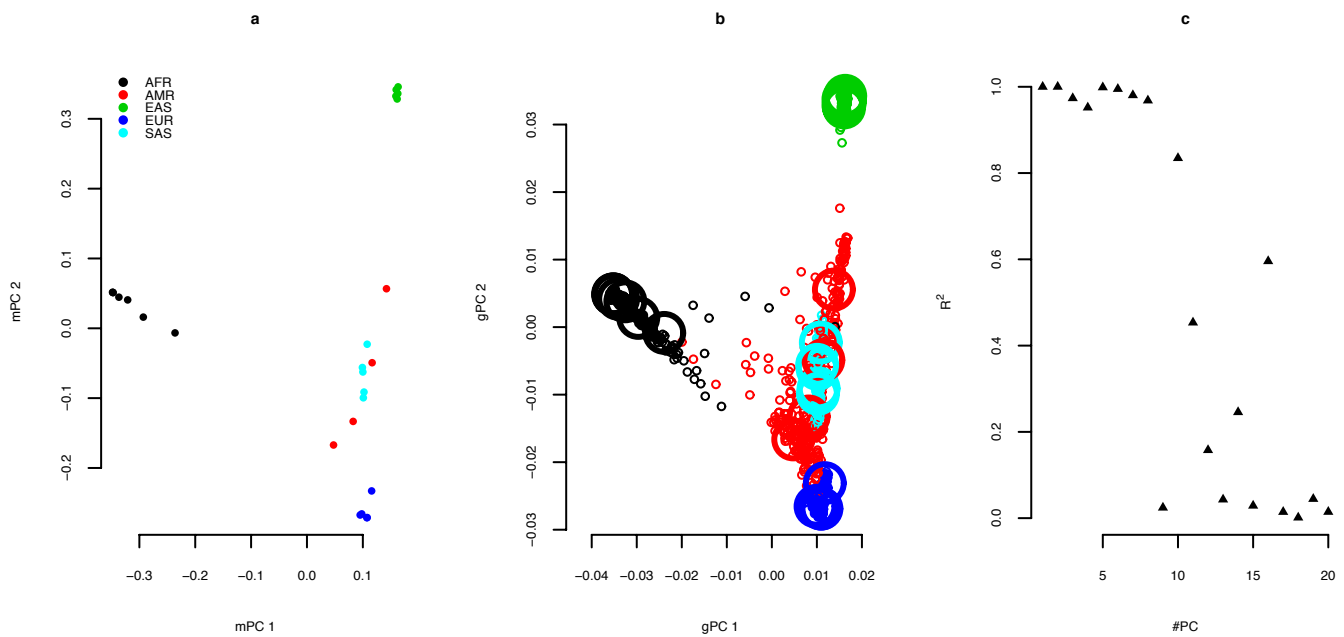


### New QC metrics for GWAS meta analysis: Supplementary Figures

**Figure S4 Comparison between Meta-PCA and genotype PCA on 1000 Genome samples using nearly 1M SNPs.**

The cohort-level allele frequencies were estimated for 26 1KG cohorts, and meta-PCA was conducted. **(a)** The projection of cohorts based on cohort-level allele frequency for 1KG samples on the first two eigenvectors. **(b)** Conventional PCA based on individual genotypes on the first two eigenvectors, the mean of the 1KG individuals within each cohort is represented with the big circles. **(c)** The correlation, measured in  $R^2$ , between meta-PCA and genotype PCA for the first twenty eigenvectors.

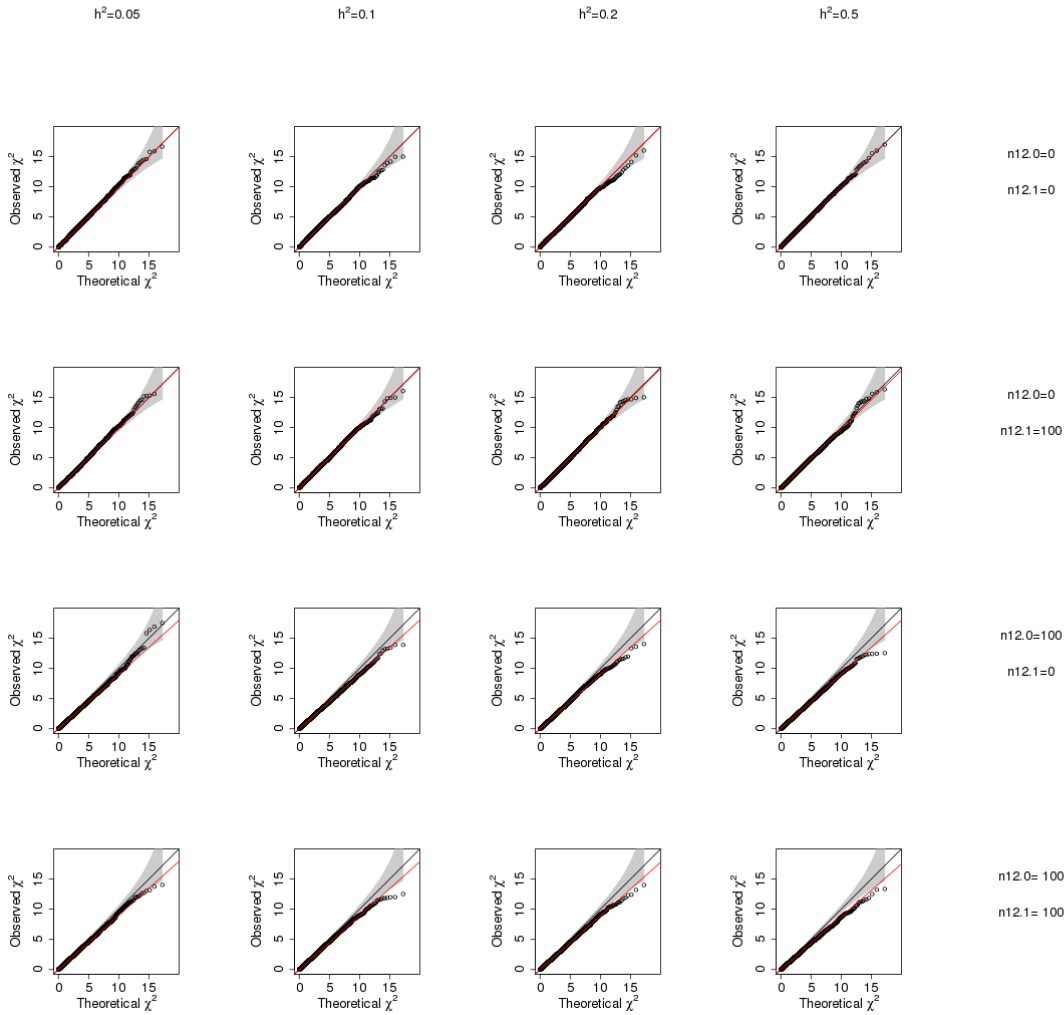
The projected cohorts were consistent with their genetic origins. In contrast, conventional PCA was also conducted on 1KG individual genotypes directly, and the mean coordinates for each cohort was then calculated. As illustrated, these two techniques resulted in nearly identical projection for 1KG, and the correlation between cohort coordinates remained consistently high,  $R^2 > 0.9$ , for the first eight eigenvectors.



## New QC metrics for GWAS meta analysis: Supplementary Figures

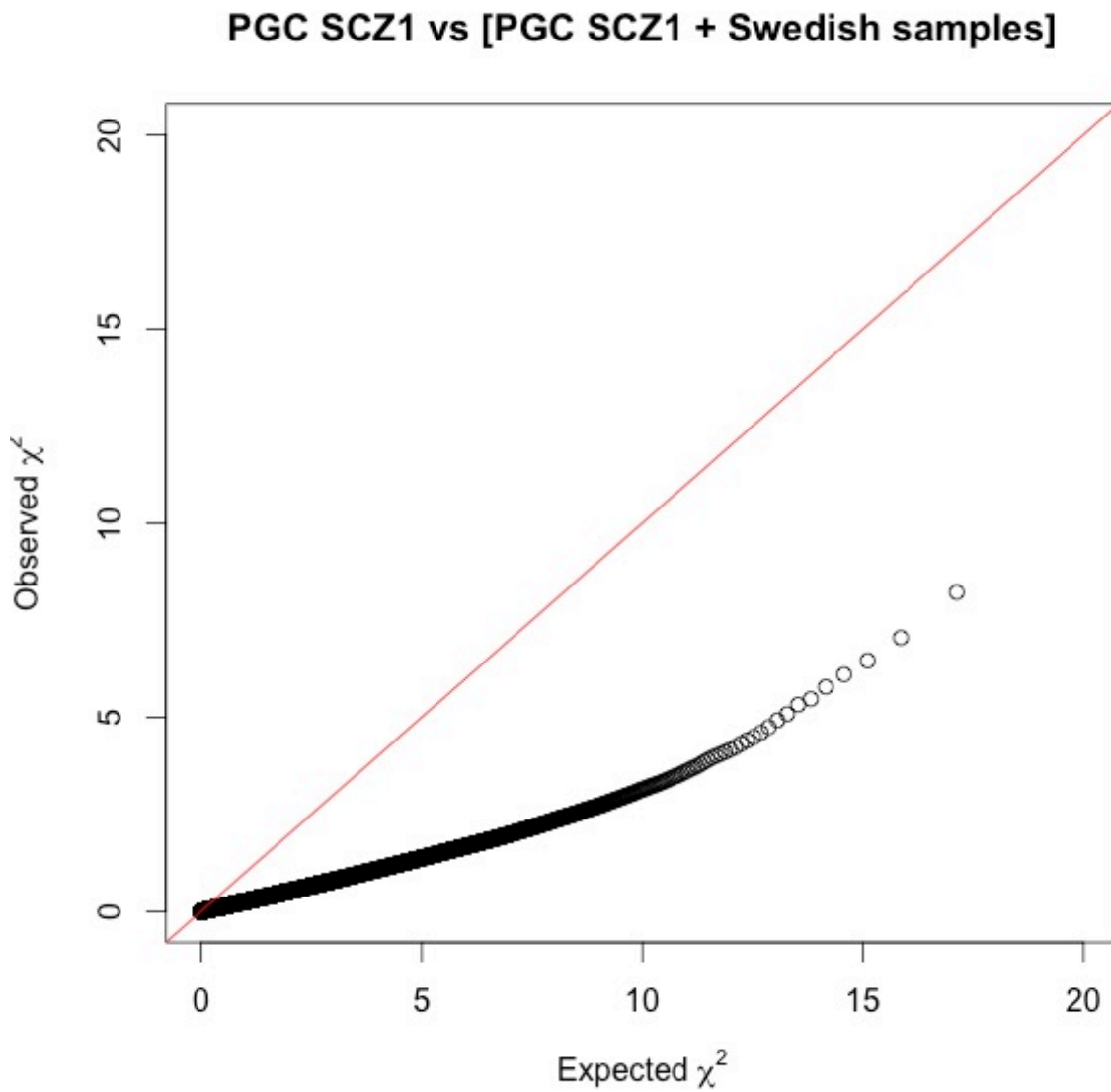
**Figure S5  $\lambda_{meta}$  and the effective number of overlapping samples ( $n_o$ ).**

Two cohorts, with 1000 individuals each and 30,000 independent loci are simulated. The genetic effects are estimated using single-marker association. Using these 30,000 summary statistics, the  $\lambda_{meta}$  is computed for each locus and contrasted to the null distribution. Each vertical panel represents different heritability, and each horizontal panel represents different combination of overlapping samples. The red line, with its expected slope of  $(1 - \frac{n_o}{n_1+n_2})$ , in each plot represents the expected distribution of the sampled loci, and gray area represents the 95% confidence interval for  $\lambda_{meta}$  under the null distribution. The overlapping first-degree relatives cause correlation between summary statistics when heritability is large (first two horizontal panels). The overlapping samples always cause reduced  $\lambda_{meta}$  (the last two horizontal panels).



**Figure S6  $\lambda_{meta}$  for PGC schizophrenia summary statistics.**

There is substantial sample overlap between PGC1 and PGZ+Swedish, and led to  $\lambda_{meta} = 0.257$ .

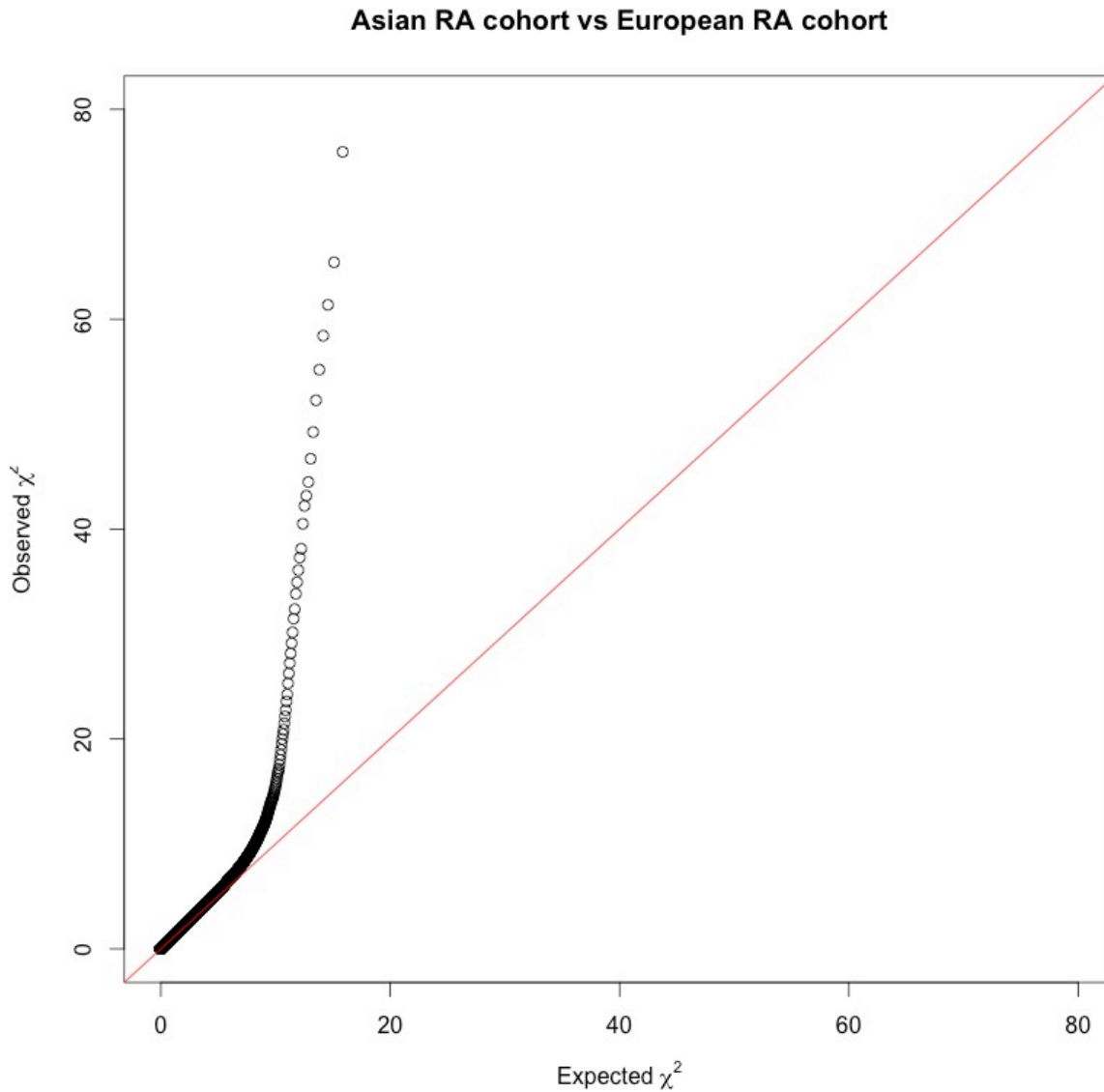




### New QC metrics for GWAS meta analysis: Supplementary Figures

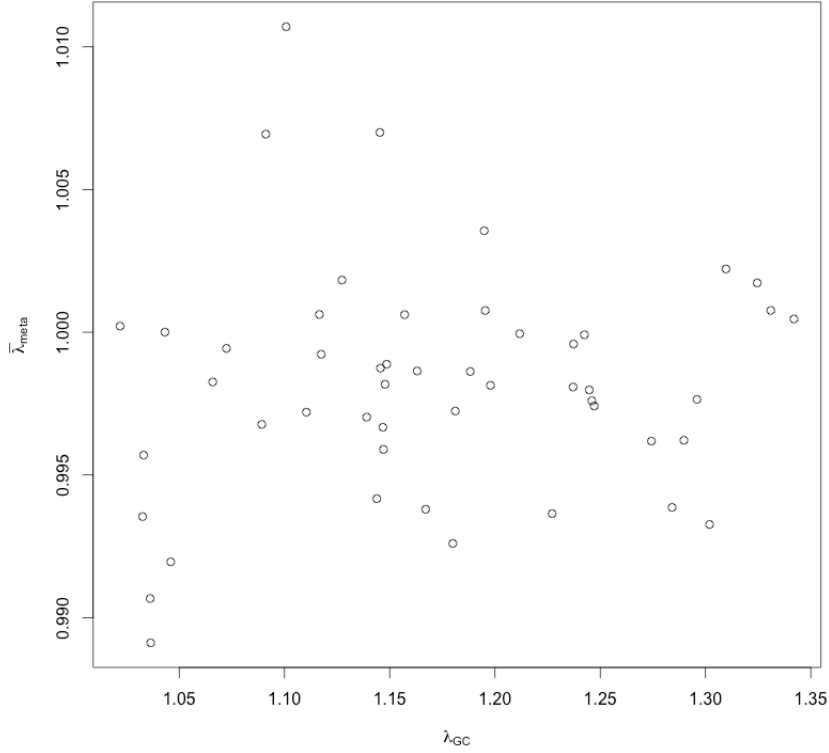
**Figure S7  $\lambda_{meta}$  for rheumatoid arthritis summary statistics between European samples and Asian samples.**

Genetic architecture may lead to inflated  $\lambda_{meta} = 1.07$  from RA GWAS summary statistics.



### New QC metrics for GWAS meta analysis: Supplementary Figures

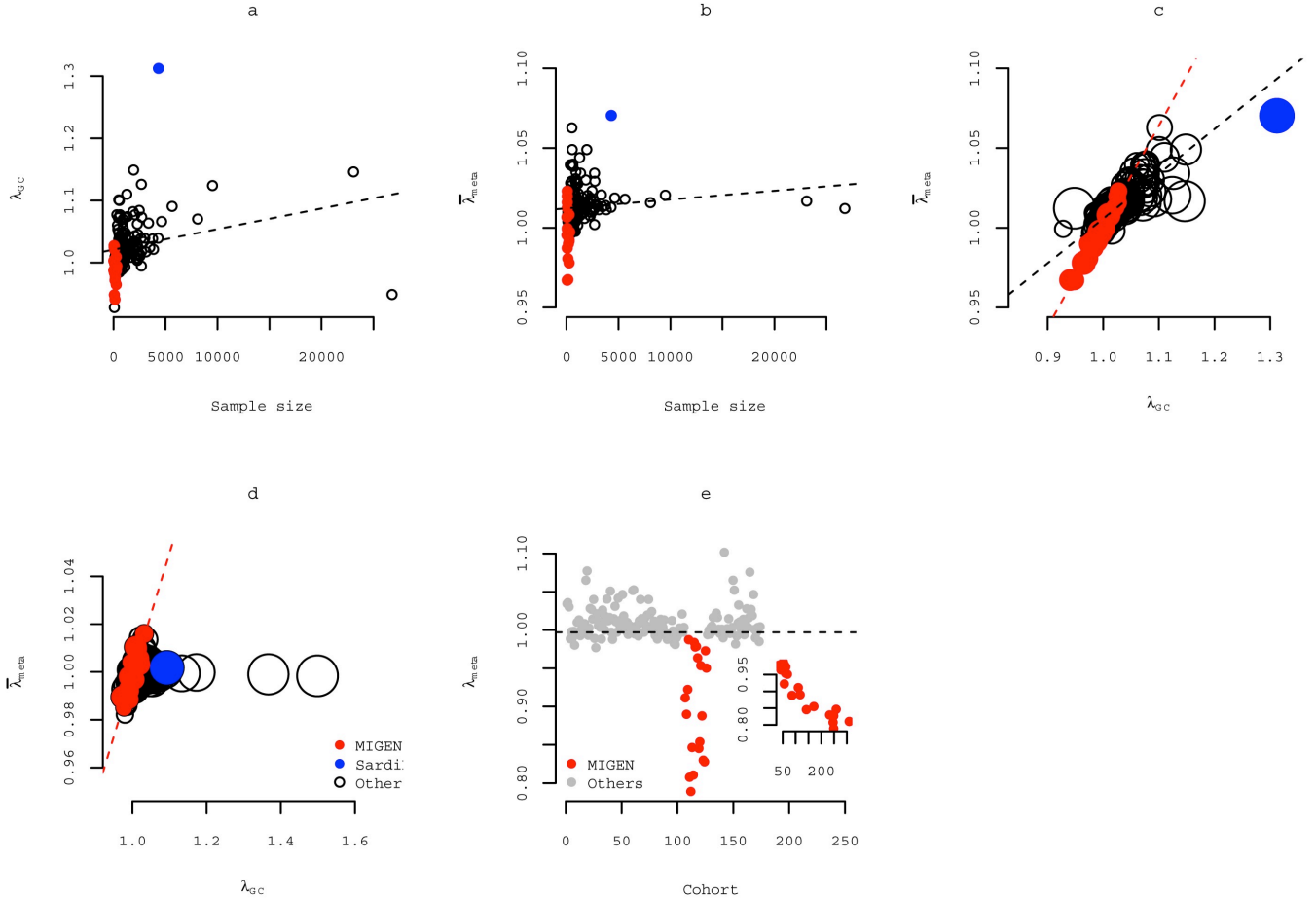
**Figure S8 No correlation between  $\lambda_{meta}$  and  $\bar{\lambda}_{GC}$  for summary statistics when there are no technical errors.** For a single cohort,  $\lambda_{GC} = \frac{b_1^2}{\omega \sigma_1^2} / 0.456$ , in which  $\omega$  is an unknown factor (technical artifact) of value around 1. When  $\omega = 1$ , there is no technical error, the behaviour of  $\lambda_{GC}$  as previous described (Yang et al., 2011): proportional to sample size under polygenic model.



## New QC metrics for GWAS meta analysis: Supplementary Figures

**Figure S9  $\bar{\lambda}_{meta}$  and  $\lambda_{gc}$  for GIANT height GWAS cohorts.**

**(a)** Sample size of each cohort against  $\lambda_{GC}$ . The linear regression is presented as a dashed line,  $\lambda_{GC} = 1.021 + 0.0000033N$  ( $N$  is reported sample size), and  $R^2 = 0.013$ . **(b)** Sample size of each cohort against  $\bar{\lambda}_{meta}$ , which was the mean of a cohort's  $\lambda_{meta}$  over all other cohorts. The linear regression is presented as a dashed line,  $\bar{\lambda}_{meta} = 1.012 + 0.00000055N$ , and  $R^2 = 0.055$ . **(c)**  $\lambda_{GC}$  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:  $\bar{\lambda}_{meta} = 0.7251 + 0.281\lambda_{GC} + e$ . The red dashed line indicates the regression slope for 20 pairs of MIGN cohorts:  $\bar{\lambda}_{meta} = 0.369 + 0.631\lambda_{GC} + e$ . 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  $\lambda_{GC}$  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 MIGN 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)**  $\lambda_{meta}$  for whole MIGN to 174 cohorts. 20 MIGN files were combined together to make “whole MIGN” via meta-analysis, and the summary statistics were used to calculate  $\lambda_{meta}$  with 174 cohorts using 30,000 independent loci. As MIGN cohorts were part of “whole MIGN”, their  $\lambda_{meta}$  were in general below 1. The dashed line is the mean of  $\lambda_{meta}$  of the “whole MIGN”. The subplot (red box) shows a strong correlation of 0.93 between  $\lambda_{meta}$  (for “whole MIGN” vs each MIGN cohort), and sample size of each MIGN cohort.



We investigated the relationship between  $\bar{\lambda}_{meta}$  (the mean of all  $\lambda_{meta}$  values of a given cohort with each of the other 173 GIANT height cohorts) and  $\lambda_{GC}$  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  $\lambda_{GC}$  and sample size; ii) no correlation between  $\bar{\lambda}_{meta}$  and sample size; iii) no correlation between  $\bar{\lambda}_{meta}$  and  $\lambda_{GC}$ . Consistent with a previous study<sup>25</sup>, for a polygenic trait such as height  $\lambda_{GC}$  of each cohort was related to its sample size (correlation of 0.235,  $p = 0.0018$ ). In contrast, the

### New QC metrics for GWAS meta analysis: Supplementary Figures

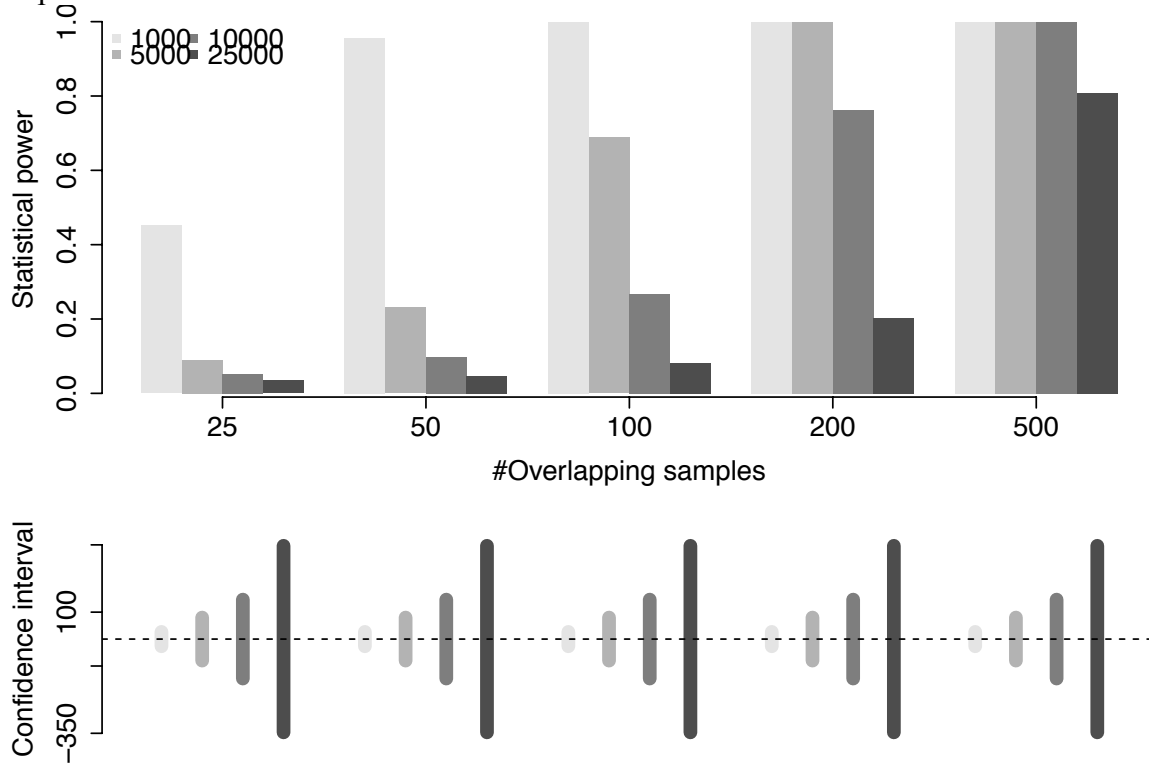
correlation between  $\bar{\lambda}_{meta}$  and sample size was of 0.116 ( $p = 0.127$ ) (**Figure S9a,b**). Nevertheless, the correlation between the mean of  $\bar{\lambda}_{meta}$  and  $\lambda_{GC}$  was 0.836 ( $p < 10e-16$ ) for 174 GIANT height cohorts (**Figure S9c**). We note that the 20 MIGEN cohorts had proportionally small  $\lambda_{GC}$  and  $\bar{\lambda}_{meta}$ , with very high correlation between them ( $\rho = 0.98$ ); in contrast, the SardiNIA cohort, which had the largest  $\lambda_{GC}$ , showed the largest  $\bar{\lambda}_{meta}$  ( $1.070 \pm 0.049$ ), 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, and observed an increased correlation ( $R^2 = 0.78$ ) between  $\bar{\lambda}_{meta}$  and  $\lambda_{GC}$  for simulated cohorts with sample sizes of the MIGEN cohorts (**Figure S9d**). Other effects, such as inflated/deflated sampling variance of the estimated genetic effects could also lead to correlation between  $\bar{\lambda}_{meta}$  and  $\lambda_{GC}$  (**Supplementary Figure S8**). In addition, we constructed a single MIGEN analysis by combining the 20 MIGEN cohorts using an inverse variance weighted meta-analysis<sup>26</sup>, and calculated  $\lambda_{meta}$  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  $\lambda_{meta}$  values on the meta-analysis results is very small.

**Simulated cohort-level summary statistics for this figure.**  $M$  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,  $\sum_{i=1}^M 2p_i(1-p_i)b_i^2 = h^2$ .  $p$  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  $\sigma_{\tilde{b}_i}^2 = \frac{1}{2np_i(1-p_i)}$ . 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.

### New QC metrics for GWAS meta analysis: Supplementary Figures

**Figure S10 Statistical power for detecting overlapping samples between a pair of cohorts given type I error rate of 0.05.**

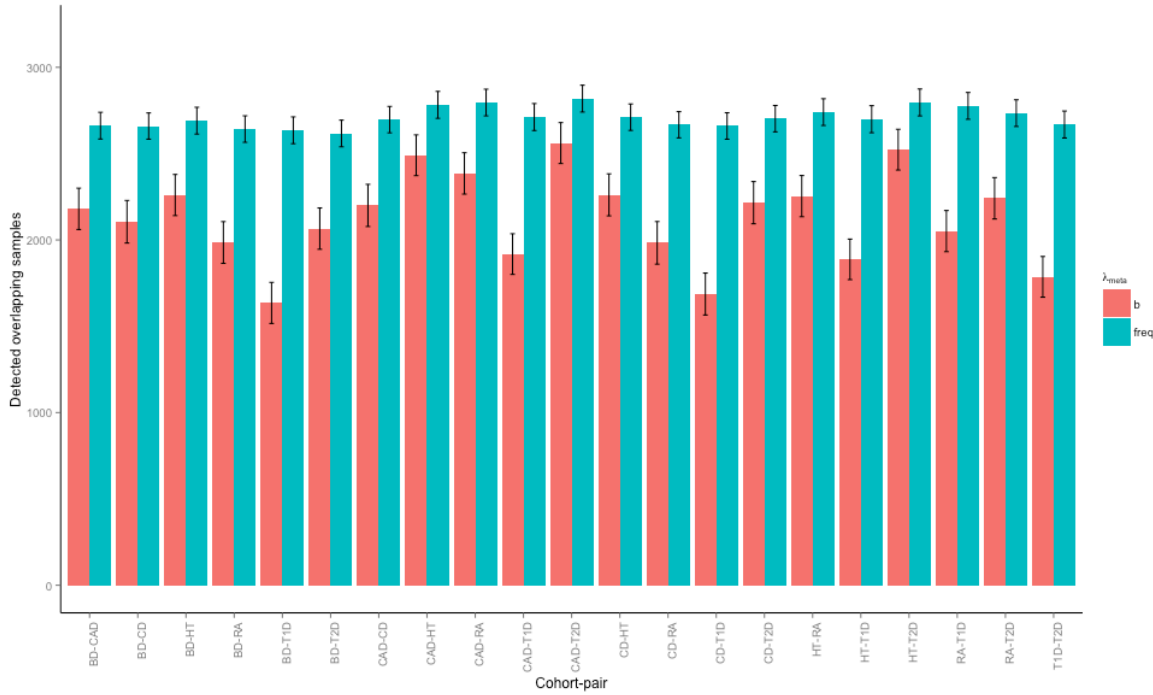
The y-axis represents statistical power, and the x-axis the number of overlapping samples. Cohort 1 has 1,000, 5,000, 10,000, or 25,000 samples, and cohort 2 has 1,000 samples. The two cohorts have 25, 50, 100, 200, and 500 overlapping samples. Bottom panel: the corresponding 95% confidence interval is given for each scenario in the top panel. The statistical power is maximized when the two cohorts have the same sample size.



## New QC metrics for GWAS meta analysis: Supplementary Figures

**Figure S11 Using  $\lambda_{meta}$  constructed either on genetic effects or on allele frequency to estimate overlapping samples between WTCCC 7 diseases.**

$\lambda_{meta}$  can be constructed on reported genetic effects (red bars), and alternatively can be constructed on allele frequency (blue bars). Both can be used to detecting overlapping samples. The x-axis indicate the pair of cohorts in WTCCC, and the y-axis represent the estimated overlapping samples based on  $\lambda_{meta}$ , which is estimated over 30,000 markers. 95% confidence interval is represented on top of each bar. The mean of the estimated overlapping samples using  $\lambda_{meta}$  on genetic effects is  $2127.38 \pm 257.73$ , for  $\lambda_{meta}$  on MAF is  $2707.99 \pm 58.41$ .



## New QC metrics for GWAS meta analysis: Supplementary Figures

### Figure S12 Workflow for PPSR regression.

**Step 1:** determine the number of pseudo profile scores. Given experiment-wise type I error rate = 0.01, type II error rate = 0.05 (power = 0.95). K pseudo profile scores should be generated using M markers, which guarantees the privacy of individual genotypes. **Step 2:** generate profile scores for each cohort. The meta-analysis center generates a KXM matrix for pseudo genetic effects. In total K profile scores were generated for each individual in each cohort. **Step 3:** PPSR method for detecting overlapping individuals using profile scores. For a pair of cohorts, PPS regression was conducted for each possible pair of individual for any two cohorts over the generated pseudo-profile scores. Once the regression coefficient was greater than the threshold, here  $b=0.95$ , the pair of individual was inferred to be having highly similar genotypes, which may indicate the individual has been included in both cohorts.

