Hum Genet (2014) 133:265–279
DOI 10.1007/s00439-013-1366-4

ORIGINAL INVESTIGATION

On individual genome-wide association studies and their meta-analysis

Yu-Fang Pei · Lei Zhang · Christopher J. Papasian · Yu-Ping Wang · Hong-Wen Deng

Received: 27 February 2013 / Accepted: 22 September 2013 / Published online: 11 October 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract Individual genome-wide association (GWA) studies and their meta-analyses represent two approaches for identifying genetic loci associated with complex diseases/traits. Inconsistent findings and non-replicability between individual GWA studies and meta-analyses are commonly observed, hence posing the critical question as to how to interpret their respective results properly. In this study, we performed a series of simulation studies to investigate and compare the statistical properties of the two approaches. Our results show that (1) as expected, meta-analysis of larger sample size is more powerful than individual GWA studies under the ideal setting of population homogeneity among individual studies; (2) under the realistic setting of heterogeneity among individual studies, detection of heterogeneity is usually difficult and meta-analysis (even with the random-effects model) may introduce elevated false positive and/or negative rates; (3) despite relatively small sample size, well-designed individual GWA study has the capacity to identify novel loci for complex traits; (4) replicability between meta-analysis and independent individual studies or between independent meta-analyses is limited, and thus inconsistent findings are not unexpected.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-013-1366-4) contains supplementary material, which is available to authorized users.

Y.-F. Pei · L. Zhang (✉) · H.-W. Deng
Center of System Biomedical Sciences, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China
e-mail: zlbio12@gmail.com

Y.-F. Pei
e-mail: pyf0419@gmail.com

Y.-F. Pei · L. Zhang · Y.-P. Wang · H.-W. Deng (✉)
Center for Bioinformatics and Genomics, Department of Biostatistics and Bioinformatics, School of Public Health and Tropical Medicine, Tulane University, 1440 Canal St., Suite 2001, New Orleans, LA 70112, USA
e-mail: hdeng2@tulane.edu

C. J. Papasian
Department of Basic Medical Science, University of Missouri-Kansas City, Kansas City, MO 64108, USA

Introduction

Genome-wide association (GWA) studies are commonly used to discover genetic loci responsible for complex diseases. To date, nearly 8,000 SNPs have been identified to be associated with a variety of complex diseases through over 1,400 GWA studies (Hindorff et al. 2009). This number is anticipated to increase dramatically with the maturation of next-generation sequencing technologies which permit direct mapping of causal SNPs. Nevertheless, for most traits/diseases studied, only a small portion of phenotypic variation or disease risk is accounted for by the genetic variants that have been identified thus far. The unidentified "missing" heritability may be due, in part, to genetic variants with minor phenotypic effects that escaped detection by individual GWA studies with insufficient statistical power (Altshuler and Daly 2007; Wang et al. 2005). In order to increase statistical power, meta-analysis of multiple GWA studies attempts to combine summary statistics from multiple individual studies. This enlarges the study sample size dramatically and may, therefore, provide a substantial increase in statistical power. A variety of GWA meta-analyses conducted to date have indeed proven their superiority over individual GWA studies in detecting many additional genetic loci underlying complex diseases/

traits (Bradfield et al. 2012; Elks et al. 2010; Franke et al. 2010; Kato et al. 2011).

As the number of meta-analyses of GWA studies increases, however, it has been observed that there are frequent inconsistencies between meta-analyses and individual studies or between different meta-analyses. This situation raises a critical question as to how to interpret association results derived from different studies. Large-scale meta-analyses have been generally used to analyze the value, and judge the robustness, of findings of individual GWA studies, because of the presumed increased power associated with meta-analyses (Richards et al. 2012). However, without a systematic investigation on the relative statistical properties of different studies such as individual GWA studies versus their meta-analyses, it may be premature to routinely accept the superiority of meta-analyses.

In this study, we assessed and compared the relative statistical properties of individual GWA studies and their meta-analyses. We explored possible mechanisms for discordant results between individual GWA studies and meta-analysis, and those between different meta-analyses. The results we present have significant implications for the design and, in particular, the interpretation of results for individual GWA studies and their meta-analyses.

Methods

We focused on a quantitative trait and performed simulation studies with a variety of contexts. Under each context, multiple individual studies of homogeneous or heterogeneous setting were simulated. Association analyses were performed on each individual study. Summary statistics from individual studies, e.g., regression coefficient, were then combined for a meta-analysis with fixed-effects (FE) or random-effects (RE) model.

Assessing the effect of different LD patterns and allele frequencies on heterogeneity

Between-study heterogeneity is a common issue in meta-analyses (Ioannidis et al. 2001), and it could be caused by a variety of factors, such as study-specific analytical approaches, unmatched study designs, different linkage disequilibrium (LD) patterns, different allele frequencies and/or effect sizes, and different patterns of interaction between genetic variants and environmental exposures (Evangelou and Ioannidis 2013; Glasziou and Sanders 2002; McCarthy et al. 2008). Here, we selected different LD patterns and allele frequencies that were of most interest and investigated their effects on between-study heterogeneity.

We simulated ten individual studies. For ease of illustration, all studies had an equal number of 2,000 subjects. Three scenarios were considered. In the first scenario, a causal SNP was simulated and tested with a homogeneous setting across the ten studies in terms of effect size and allele frequency. In the second scenario, a causal SNP was simulated with a heterogeneous setting in terms of allele frequency. Specifically, individual studies were drawn from three distinct populations with proportions 60 %:30 %:10 %. Allele frequency for each population was generated using the Balding-Nichols model (Balding and Nichols 1995). Briefly, it was drawn from a Beta distribution with parameters $p(1-F_{st})/F_{st}$ and $(1-p)(1-F_{st})/F_{st}$, where p was the ancestral allele frequency and F_{st} was a measure of genetic distance between populations (Wright 1950). We set a moderate level of F_{st} at 0.05 and set p to vary from 0.05 to 0.95, respectively. In the third scenario, a causal SNP was simulated but tested by another proxy SNP that was in LD with it. The minor allele frequencies (MAFs) of both causal and proxy SNPs were set at 0.1. A heterogeneous setting in terms of LD was assumed. Specifically, the LD measure r^2 between the causal and proxy SNPs varied among 0.3–0.5–0.8, 0.7–0.8–0.9, and 0.8–0.8–0.8, respectively, in the three populations. Under each scenario, we simulated 5,000 replicates and measured heterogeneity by I^2 index (see below for details) (Higgins and Thompson 2002; Higgins et al. 2003).

In what follows, we disregarded the detailed sources of heterogeneity effects and introduced a single parameter σ^2_{bp} to measure between-population heterogeneity, and σ^2_{wp} to describe the within-population heterogeneity (see below for detail).

Model for heterogeneous effects

Assuming there are s populations and there are n_i ($i = 1, 2,\ldots, s$ and $\sum n_i = k$) individual studies in the ith population, let θ_{ij} be the effect size for the jth study from the ith population,

$$\theta_{ij} = \theta_0 + u_i + w_{ij}, \quad (1)$$

where θ_0 is the mean effect size; $u_i \sim N\left(0, \sigma^2_{bp}\right)$ is the population specific random effect, and $w_{ij} \sim N\left(0, \sigma^2_{wp}\right)$ is the study-specific random effect. The variance $\tau^2 = \sigma^2_{wp} + \sigma^2_{bp}$ measures between-study heterogeneity effect.

Various heterogeneous effects could be simulated by setting different values for σ^2_{wp} and σ^2_{bp}: (1) $\sigma^2_{wp} = \sigma^2_{bp} = 0$, the component individual studies were all homogeneous; (2) $\sigma^2_{bp} = 0$ and $\sigma^2_{wp} > 0$, the component individual studies were sampled from the same population and within-population variation existed due to different phenotype measures, study designs and analytical approaches, etc.; (3) $\sigma^2_{bp} > 0$ and $\sigma^2_{wp} = 0$, the component individual studies

were sampled from different populations and all studies were homogenous with respect to study design, analytical approach and other study-level properties; between-study heterogeneity was entirely attributable to variance in LD levels, allele frequencies and other population-level properties; (4) $\sigma^2_{bp} > 0$ and $\sigma^2_{wp} > 0$, the component individual studies were sampled from different populations and between-study heterogeneity existed due to different study designs, phenotype measures, LD levels, allele frequencies, etc. In reality, the 4th scenario would most closely reflect GWA studies and meta-analyses in the field, as individual GWA studies included in meta-analyses generally come from populations that differ to varying extents ($\sigma^2_{bp} > 0$) and often also have different phenotyping approaches, study designs, and analytical approaches ($\sigma^2_{wp} > 0$).

The phenotypic value y_{ijk} of the kth individual in the jth study and the ith population is modeled as,

$$y_{ijk} = \theta_{ij} x_{ijk} + \varepsilon_{ijk}, \tag{2}$$

where x_{ijk} is genotype score encoded in an additive mode of inheritance, and $\varepsilon_{ijk} \sim N\left(0, \sigma^2_{\varepsilon}\right)$ is random error. Assuming independence between u_i and w_{ij}, phenotypic variance $\sigma^2_y = \mathrm{var}\left(\theta_0 x_{ijk}\right) + \mathrm{var}\left(u_i x_{ijk}\right) + \mathrm{var}\left(w_{ij} x_{ijk}\right) + \sigma^2_{\varepsilon}$. The heritability h^2 of the SNP is defined as,

$$h^2 = \mathrm{var}\left(\theta_0 x_{ijk}\right)/\sigma^2_y = 2\theta_0^2 q(1-q)/\sigma^2_y, \tag{3}$$

where q is the MAF of the causal SNP. The proportion of phenotypic variation explained by the between-population heterogeneity effect is

$$\mathrm{var}\left(u_i x_{ijk}\right)/\sigma^2_y = \left(\sigma^2_{bp}/\sigma^2_y\right)\left(2q - q^2\right), \tag{4}$$

and that explained by the within-population heterogeneity effect is

$$\mathrm{var}\left(w_{ij} x_{ijk}\right)/\sigma^2_y = \left(\sigma^2_{wp}/\sigma^2_y\right)\left(2q - q^2\right). \tag{5}$$

The total proportion of phenotypic variation explained by between-study heterogeneity is thus

$$\mathrm{var}\left(w_{ij} x_{ijk} + u_i x_{ijk}\right)/\sigma^2_y = \left(\left(\sigma^2_{bp} + \sigma^2_{wp}\right)/\sigma^2_y\right)\left(2q - q^2\right), \tag{6}$$

which depends on MAF q, σ^2_{bp}, σ^2_{wp}, and σ^2_y.

Meta-analysis

Two distinct models can be used to perform meta-analyses: FE model and RE model. Briefly, the FE model assumes that all individual studies have a common effect size and that differences between observed effect sizes are attributable merely to sampling error. The RE model, on the other hand, assumes that effect sizes are drawn from a distribution. Within the RE model, two sources of variance contribute to differences between the observed effect size estimates: sampling error and between-study heterogeneity. Assuming β_{ij} and v_{ij} are the observed effect size and its variance, we adopted the inverse variance weighted method to construct the test statistic for both FE and RE meta-analyses (de Bakker et al. 2008). The test statistic

$$z_{meta} = \frac{\hat{\beta}}{S\hat{E}}, \tag{7}$$

where $\hat{\beta} = \frac{\sum_i \sum_j \beta_{ij} w_{ij}}{\sum_i \sum_j w_{ij}}$ is the combined effect size across studies and $S\hat{E} = \sqrt{\frac{1}{\sum_i \sum_j w_{ij}}}$ is the corresponding standard error; the weight $w_{ij} = \frac{1}{v_{ij}}$ in the FE model and $w_{ij} = \frac{1}{v_{ij} + \hat{\tau}^2}$ in the RE model, where $\hat{\tau}^2$ is the estimate of between-study heterogeneity τ^2. It is a function of Cochran's Q (Cochran 1954),

$$\hat{\tau}^2 = \begin{cases} \dfrac{Q - (k-1)}{\sum_i \sum_j w_{ij} - \left(\sum_i \sum_j w_{ij}^2\right)/\left(\sum_i \sum_j w_{ij}\right)} & \text{for } Q > (k-1) \\ 0 & \text{for } Q \le (k-1) \end{cases}, \tag{8}$$

where $Q = \sum_i \sum_j w_{ij} \left(\beta_{ij} - \hat{\beta}\right)^2$ and $w_{ij} = \frac{1}{v_{ij}}$.

z_{meta} approximately follows a standard normal distribution under the null hypothesis, which is the basis for assessing its statistical significance.

Heterogeneity test

Cochran's Q test is widely used to determine whether the between-study heterogeneity $\tau^2 = 0$. Under the null hypothesis that the effect sizes are equal in all studies, the Q statistic follows a Chi square distribution with $k-1$ degrees of freedom. A significance level of $p < 0.1$ suggests the existence of significant heterogeneity (Cochran 1954).

I^2 index is another widely used measure for quantifying degree of heterogeneity (Higgins and Thompson 2002; Higgins et al. 2003),

$$I^2 = \begin{cases} \frac{Q-(k-1)}{Q} \times 100\,\% & \text{for} \quad Q > (k-1) \\ 0 & \text{for} \quad Q \le (k-1) \end{cases} \tag{9}$$

An I^2 index larger than 50 % indicates substantial heterogeneity (Kavvoura and Ioannidis 2008). We evaluated the performance of the heterogeneity test by summarizing the proportion of simulations with an I^2 index larger than 50 % or a p value for the Q test <0.1.

In the following, we investigated the relative statistical properties of individual studies and meta-analysis under

two simulation settings: single SNP test and genome wide scan.

Investigating the statistical properties of meta-analysis when testing a single SNP

Type-I error rate estimation

We simulated a total of k ($k = 5$, 10, 20, and 50) individual studies, each with 2,000 unrelated subjects. These individual studies were either from the same population ($\sigma^2_{bp} = 0$) or from three different populations with proportions 60 %:30 %:10 % (60 %:20 %:20 % for $k = 5$). The latter setting attempted to model the current situation in human genetics in which GWA studies are performed in different populations. θ_0 and σ^2_y were set at 0 and 100, respectively. Diverse levels of heterogeneity were simulated by setting different values for σ^2_{wp} and σ^2_{bp}. The allelic effect in each individual study may deviate from 0 because of study-specific and population-specific components. We tested the average effect size across all studied samples. The null hypothesis was thus defined as $\theta_0 = 0$. The statistically significant level was set at 0.05 and type-I error rates were estimated from 10,000 replicates.

Power estimation

Power was estimated by setting a non-null average effect size θ_0 according to a pre-specified value of h^2. The simulation setting was the same as that used for estimating type-I error rates. Four combinations of σ^2_{wp} (0, 0.6) and σ^2_{bp} (0, 0.1) were considered to represent different levels of heterogeneity effects. Causal SNPs were population-specific so their effect may only be present in some populations. The purpose of this scenario was to evaluate the performance of meta-analysis when combining samples with a true effect and samples without any effect, a situation that may occur, for example, with ethnic specific loci (Lei et al. 2009). The significance threshold was set at the genome-wide significance level (GWS, 5×10^{-8}) and powers were estimated on 10,000 simulated replicates.

Rare variants versus common variants

To compare the performance of meta-analysis for common and rare variants, we investigated type-I error rates and power for SNPs with different MAFs. We simulated ten studies under a heterogeneous setting ($\sigma^2_{wp} = 0.6$ and σ^2_{bp} varied from 0 to 1). Privacy issue is common in real applications, especially for sequencing based rare variants (de Bakker et al. 2008; Singh et al. 2013). As the focus of the present study was heterogeneity effects and statistical power of meta-analysis, we did not extensively study privacy issues and assumed that no private alleles were present. The MAF of the causal SNP was set at 0.005, 0.01, 0.1, 0.3, and 0.5. Type-I error rates and power were estimated at different h^2 values.

Effect of population stratification on meta-analysis

Our above simulations sampled subjects from a homogeneous population within each individual study. To simulate a more complicated situation, we introduced population structure within individual studies. Specifically, in each of the individual studies, subjects were sampled from an admixture of two ancestral populations (A and B). A total of 10,000 SNPs were simulated to constitute each subject's genotype profile. For each SNP, allele frequencies were generated using the Balding-Nichols model (Balding and Nichols 1995). Briefly, an ancestral allele frequency p was drawn from the uniform distribution U(0.1,0.9). Allele frequencies in populations A and B were then drawn from a beta distribution with parameters $p(1-F_{st})/F_{st}$ and $(1-p)(1-F_{st})/F_{st}$. We set F_{st} at 0.05 to simulate moderate population stratification. Each individual's genotype was then drawn from population A with probability $f \sim$ U(0.3, 0.7) and from population B with probability $1-f$. The heritability h^2 varied from 0.0 to 1.0 % and the difference in population phenotypic mean was set at 10. Principal component analysis (Price et al. 2006) was performed on these 10,000 SNPs to evaluate the ancestral backgrounds, and the top 10 principal components were used to adjust phenotypes in association analyses.

Meta-analysis versus mega-analysis

Meta-analysis refers to the analysis of the combination of association summary statistics, e.g., regression coefficients, from individual studies. With the availability of participant-level data, an alternative approach to meta-analysis is the analysis of the combined samples of individual participant-level samples, e.g., mega-analysis (Lin and Zeng 2010a, b). To investigate the relative performance of these two analyses, they were compared on a set of ten individual studies from three populations with proportions 60 %:30 %:10 %. Both type-I error rates and power estimates were evaluated.

Investigating the performance of individual study and meta-analysis in genome-wide scan

Under this simulation setting, we simulated 1,000 independent causal SNPs to mimic a genome-wide distribution. The average h^2 of these causal SNPs was set at 0.05 % so that the total heritability was approximately 50 %. Values

of h^2 were drawn from an inverse gamma distribution IG(15, 75) and then divided by 10,000. The empirical distribution of the generated values is plotted in Supplementary Fig. 1. For simplicity, all individual studies were drawn from a single population so that $\sigma^2_{bp} = 0$, while $\sigma^2_{wp} = 0.6$.

Comparison of individual GWA studies and meta-analysis for identifying novel loci

In the presence of large-scale meta-analyses, the question 'what was the value of analyzing individual studies?' was of interest. Here we attempted to explore this question by evaluating the relative value of a smaller individual study compared with a larger meta-analysis. The individual GWA study was simulated with 10,000 unrelated individuals; this number was reasonable considering both discovery and replication phases (Styrkarsdottir et al. 2008; Xiong et al. 2009). Meta-analysis was simulated with ten individual studies, each with 2,000 unrelated subjects. The GWS level (5×10^{-8}) was used to declare positive results.

Comparisons were also made between meta-analysis and individual studies under equal sample size. Specifically, meta-analysis of five individual studies with 2,000 subjects each was compared with an individual GWA study with 10,000 subjects.

Power of identifying multiple loci and at extremely stringent levels

We noticed that many research consortia have reported extremely small *p*-values (such as 10^{-30}–10^{-50}) at multiple loci (Estrada et al. 2012; Speliotes et al. 2010). Given effect size and sample size, the question of how likely it was to generate extreme *p*-values was of interest. To answer this question, we evaluated the power of meta-analysis to simultaneously identify multiple loci and to identify loci at two extremely stringent levels (10^{-20} and 10^{-25}). The two thresholds were set arbitrarily in accordance with published studies using extremely small *p*-values.

Assessing power of meta-analysis to replicate individual GWA studies' findings

We simulated a collection of $k + 1$ ($k = 5$, 10, 20, and 50) individual studies, each with 2,000 unrelated individuals. The first GWA study was treated as the discovery study. Meta-analysis combining the remaining k studies was performed to replicate discovery findings. We performed 5,000 replicates to estimate power, which was defined as the proportion of replicates in which at least m SNPs identified by the discovery individual study were successfully replicated by the meta-analysis at the GWS level.

One phenomenon observed in the current genetic mapping era is that individual studies are often part of bigger meta-analysis consortia. Under these conditions, samples used for individual studies and meta-analyses are not independent. Nonetheless, there has been a trend toward regarding the results of meta-analysis as the standard for judging findings of individual studies, even when samples overlapped (Richards et al. 2012). To evaluate the replicability between GWA studies and meta-analysis when samples overlapped, we designed a simulation scenario in which individual study was part of meta-analysis samples.

Assessing replicability between meta-analyses

We simulated a collection of $2k$ ($k = 5$, 10, 20, and 50) individual studies, each with 2,000 unrelated individuals. Meta-analysis combining the 1st to kth GWA studies was considered as the discovery study. Meta-analysis combining the remaining k GWA studies was considered as the replication study. We repeated the simulation 5,000 times to estimate power, which was defined analogously.

We also evaluated the power of replicability between two meta-analyses with overlapped samples. Specifically, three individual studies were included into both meta-analyses samples.

Results

We first evaluated the potential effects of population-specific properties, including different allele frequencies and LD levels, on heterogeneity. Our results showed that differences in either allele frequency or LD levels could lead to significant heterogeneity (Supplementary Fig. 2). Larger variations in LD levels and lower ancestral MAF corresponded to higher levels of heterogeneity (see Supplementary notes for interpretation).

Performance of meta-analysis when testing a single SNP

Average I^2 estimates under different parameter settings

Changes in the heterogeneity measure I^2 with MAF are presented in Table 1. Generally, I^2 increased with increasing MAFs (see Supplementary notes for detailed interpretation). For example, when only within-population heterogeneity was present ($\sigma^2_{wp} = 0.6$ and $\sigma^2_{bp} = 0$), the average I^2 increased from 15.5 to 81.8 % as MAF increased from 0.005 to 0.5. Throughout the following

simulations, MAF was set at 0.1 for common variants, and at 0.005 for rare variants, unless otherwise stated. For ease of reference, major parameter settings are presented in Table 2.

When the average effect size θ_0 was the same in all populations, the level of between-study heterogeneity increased with increasing values of σ^2_{wp} and σ^2_{bp} (Supplementary Figs. 3A, 3B, and 3C). For example, when meta-analyzing 5 individual studies from a homogeneous population ($\sigma^2_{bp} = 0$), σ^2_{wp} values of 0.0 and 0.6 corresponded to average I^2 values of 14 and 53 %, respectively. When meta-analyzing five studies from different populations (e.g. $\sigma^2_{bp} = 0.1$), σ^2_{wp} values of 0.0 and 0.6 corresponded to average I^2 values of 20 and 55 %, respectively. Note that average I^2 always deviated from zero, even in an ideally homogeneous situation.

Variation of θ_o among different populations introduced additional heterogeneity, as illustrated by Supplementary Figs. 3D and 3E. For example, when only between-population variation existed ($\sigma^2_{bp} = 0.1$ and $\sigma^2_{wp} = 0$) and the tested SNP presented an effect in only the first population, the average I^2 could be as high as 91 % when the h^2 for the first population was 2 % (Supplementary Figs. 3D and 3E).

Table 1 The effects of MAF on the level of heterogeneity [average I^2 (%)]

MAF	σ^2_{bp}					
	0	0.2	0.4	0.6	0.8	1
0.005	15.5	15.9	16.9	17.4	18.4	19.1
0.01	19.2	20.8	22.5	23.9	25.7	26.8
0.1	60.7	64.8	67.9	69.7	71.8	73.8
0.3	78.9	81.5	83.6	85.1	85.9	86.8
0.5	81.8	84.2	85.7	87.1	88	88.8

σ^2_{bp} denotes between-population variance; within-population variance σ^2_{wp} was set at 0.6

Type-I error rate

When individual studies were ideally homogeneous ($\sigma^2_{wp} = 0$ and $\sigma^2_{bp} = 0$), both the FE and RE models had error rates close to the target level of 5 % (Fig. 1a, d). When different studies were from a homogeneous population ($\sigma^2_{bp} = 0$), but individual studies had within population heterogeneity ($\sigma^2_{wp} > 0$), the FE model had an inflated type-I error rate, regardless of the number of studies being meta-analyzed. The error rate could be as high as 36 % (Fig. 1a). This may be because the FE model tended to underestimate the variance of effect size when there was between-study heterogeneity. In this setting, the RE model was robust when k was large (e.g. $k > 10$). However, for small k's (e.g. $k = 5$) the type-I error rate for RE was not well controlled, and became increasingly inflated as σ^2_{wp} increased (Fig. 1d).

When individual studies were sampled from different populations, neither the FE nor the RE model could control type-I error rates (Fig. 1b, c, e, f), regardless of whether within-population variation existed or not, though the RE model was less liberal than the FE model. For example, when $\sigma^2_{bp} = 0.1$, $\sigma^2_{wp} = 0$, and $k = 10$ (corresponding to an average I^2 value of 18 %), the type-I error rates for the FE and RE models were 23 and 18 %, respectively. Type-I error rates increased with larger k's and increases in σ^2_{bp} and/or σ^2_{wp}. At an extreme setting in which $\sigma^2_{bp} = 1$, $\sigma^2_{wp} =$

Table 2 Major simulation parameters

Simulation settings	No. of studies	Sample sizes	MAF	σ^2_{bp}	σ^2_{wp}	h^2
Single SNP test						
Type-I error	5, 10, 20, 50	2,000	0.1	0 ~ 1	0 ~ 1	0
Power ($\sigma^2_{bp} = 0$)	5, 10, 20, 50	2,000	0.1	0	0, 0.6	0.1 ~ 2 %
Power ($\sigma^2_{bp} > 0$)	10	2,000	0.1	0.1	0, 0.6	0.1 ~ 2 %
Genome-wide scan						
Power	5, 10, 20, 50	2,000	0.1	0	0.6	0.02 ~ 0.12 %
Replicability	5, 10, 20, 50	2,000	0.1	0	0.6	0.02 ~ 0.12 %
Specific setting						
Rare versus common variants	10	2,000	0.005, 0.01, 0.1, 0.3, 0.5	0 ~ 1	0.6	0.1 ~ 1 %
Effects of population stratification	10	2,000	$F_{st} = 0.05$, $p = 0.1$	–	–	0.3
Meta versus Mega	10	2,000	0.1	0.1	0.6	0.1 ~ 1 %
Sample structure from real data	17	220 ~ 7,394	$F_{st} = 0.11$, $p = 0.1$	–	–	0.1 ~ 1 %

F_{st} is a measure of genetic distance between populations; p denotes ancestral allele frequency; σ^2_{bp} and σ^2_{wp} denote between- and within-population variances, respectively; h^2 represents heritability (phenotypic variation explained by the SNP)

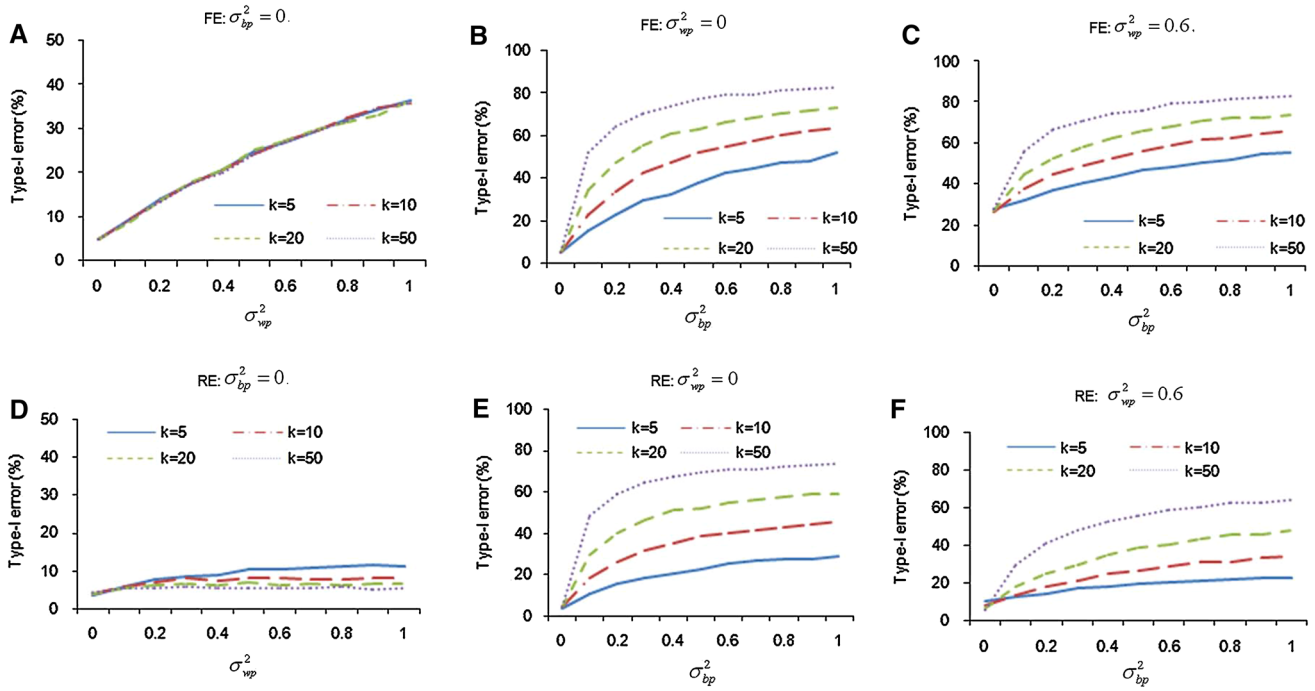


Fig. 1 Type-I error rates of meta-analyses. k is the number of studies included in the meta-analysis. **a**, **b**, and **c** Present the type-I error rates of the fixed-effects (FE) model, and **d**, **e**, and **f** present the type-I error rates of the random-effects (RE) model. σ^2_{wp} and σ^2_{bp} represent within- and between-population variances, respectively. For **a** and **d**, only within-population variation was present. For **b** and **e**, only between-population variation was present. For **c** and **f**, both within- and between-population variations were present. The statistically significant level was set at 0.05

0.6 and $k = 50$ (average $I^2 = 77$ %), type-I error rates were as high as 83 and 64 % for the FE and RE models, respectively (Fig. 1c, f).

The simulations for Type-I error rates discussed above analyzed common variants. Results for rare variants showed inflated type-I error rates as well. Notably, these error rates were less severe than those for common variants under the same parameter settings (Supplementary Fig. 4).

Power estimation

Under ideally homogeneous settings ($\sigma^2_{wp} = 0$ and $\sigma^2_{bp} = 0$), the FE model was more powerful than the RE model. As expected, both models were more powerful than individual studies due to enlarged sample sizes (Supplementary Figs. 5A and 5C). For example, the power for identifying a SNP with an h^2 of 0.3 % was 95, 63 and 0.1 % for the FE, RE models ($k = 10$) and analysis by an individual study.

Compared with the ideally homogeneous setting analyzed in the previous paragraph, a loss of power was observed for both the FE and RE models when only within-population heterogeneity effects existed ($\sigma^2_{bp} = 0$, $\sigma^2_{wp} > 0$) (Supplementary Figs. 5B and 5D). For example, when $\sigma^2_{wp} = 0.6$, the power to detect a SNP with an h^2 of 0.3 % decreased from 95 to 91 % and from 63 to 34 % for the FE and RE models, respectively. The severity of the power loss was greater for the RE model than for the FE model.

When only between-population variation existed ($\sigma^2_{bp} = 0.1$), the pattern of power change depended on the variation of θ_0. For these simulations, we used various combinations of ten individual studies of equal size from three populations, with proportions of 60 %:30 %:10 %. If the average effects of the SNP were equal in all populations (corresponding to an average I^2 value of 18 %, i.e., a level that was generally regarded to reflect no heterogeneity or non-detectable heterogeneity in meta-analyses), meta-analyzing samples from different populations had the potential to increase power. For example, when meta-analyzing the six studies from the first population, the power of the FE and RE models for identifying a SNP with an h^2 of 0.3 % was 63 and 52 %; the power increased to 93 and 81 % for the two models when meta-analyzing all ten studies (Fig. 2a). Compared with meta-analysis of multiple samples from a single population, all of which showed a causal effect, power decreased with the addition of samples from other populations without causal effect. For example, when the SNP was causal in only the first population ($h^2 = 0.3$ %), power of meta-analysis of six studies from this population was 63 and 52 % for the FE and RE models, respectively.

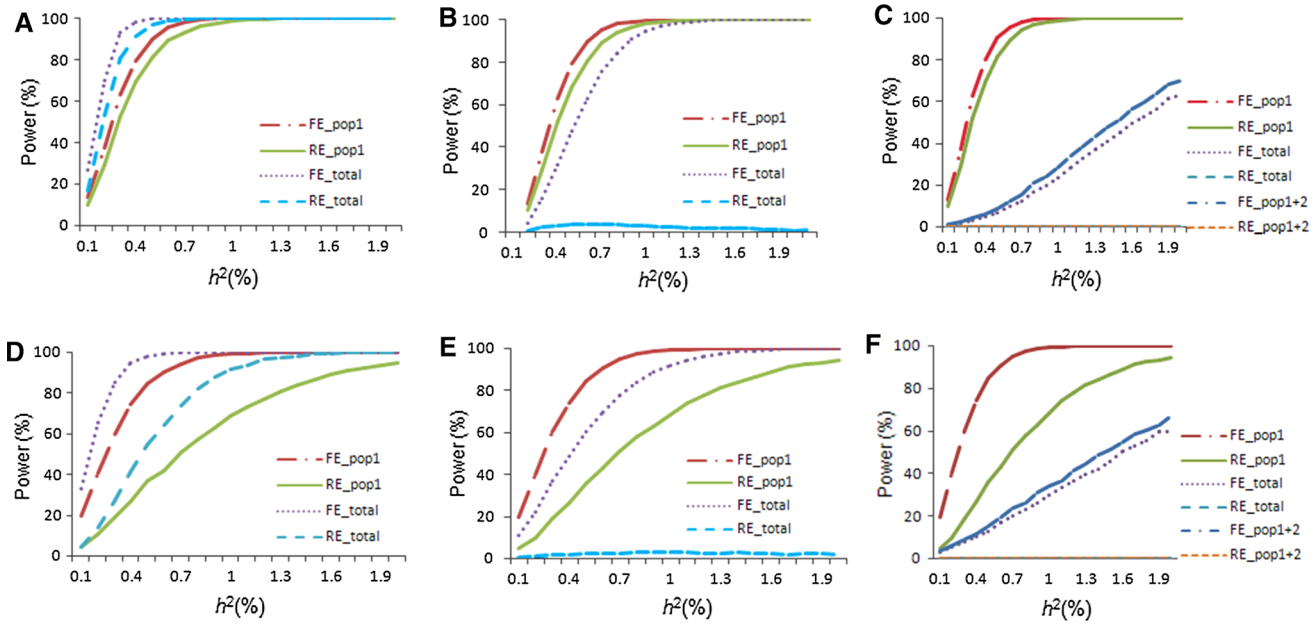


Fig. 2 Power estimates of meta-analyses including studies from different populations. "FE" and "RE" denote fixed- and random-effects meta-analyses, respectively. σ^2_{wp} and σ^2_{bp} represent within and between-population variances, respectively. Ten individual studies were simulated from three populations with proportions 60 %:30 %:10 %. The suffixes "pop1", "pop1+2", and "total" denote meta-analyses including studies from the first population, the first two populations, and all three populations, respectively (see Methods for details). h^2 on the *x*-axis represents heritability (phenotypic variation explained by the SNP). The statistically significant level was set at 5×10^{-8}. For **a**, **b**, and **c**, only between-population variation was present. For **d**, **e**, and **f**, both within- and between-population variations were present. For **a** and **d**, the SNP had the same effect in all populations in terms of both magnitudes and direction. For **b** and **e**, the SNP had an effect in only the first population. For **c** and **f**, the SNP produced effects with the same magnitude but in opposite directions in samples from the first and second populations, while the SNP had no effect on the third population

When adding four samples from the two other populations in which the SNP was neutral, power decreased to 32 and 3 %, respectively (Fig. 2b). Power loss was even more severe when combining samples with effects in opposite directions. For example, if the causal SNP produced effects of the same magnitude, but in opposite directions, in studies from the first and second populations, while that from the third population had no effect, including samples of the second population into the analysis decreased power to 4 % and <1 % for the FE and RE models, respectively. Finally, meta-analysis of all ten studies (corresponding to average I^2 value of 82 %) had powers of 3 and <1 % for the FE and RE models (Fig. 2c). Therefore, including heterogeneous studies from different populations with opposite directional effects, or without effect, may actually lead to a serious loss of power for both the RE and FE models.

When both within- and between-population variations existed ($\sigma^2_{bp} = 0.1$ and $\sigma^2_{wp} = 0.6$), the results were similar to those obtained when only between-population variation was present, although power loss was more severe due to the additional heterogeneity introduced by within-population variation (Fig. 2d–f).

The simulations for power estimates discussed above analyzed common variants. For rare variants, there was also a serious power loss for both the FE and RE models when adding samples with no causal effects (Supplementary Fig. 4).

To facilitate comprehension, we listed major results of type-I error rates and power estimates in Table 3.

Heterogeneity test

As expected, the power for detecting heterogeneity increased with increasing levels of heterogeneity and with larger *k*s (Fig. 3). Importantly, both *Q* test and I^2 measures were underpowered for detecting heterogeneity with small *k*s. For example, for $k = 5$ and $\sigma^2_{bp} = 0$, the power of detecting heterogeneity induced by σ^2_{wp} did not reach the recommended 80 % threshold (Cohen 1988). Even when $k = 10$, the 80 % threshold for both *Q* and I^2 was reached only when σ^2_{wp} exceeded 0.6. Considering that heterogeneity tests were found to be underpowered (Pereira et al. 2010), even when the number of studies increased to 20, it

Table 3 Type-I error rates and power estimates under the major parameter settings

Model	σ^2_{wp}	σ^2_{bp}	Type-I error (%)	Power (%)				
				$h^2 = 0.2$ %	$h^2 = 0.4$ %	$h^2 = 0.6$ %	$h^2 = 0.8$ %	$h^2 = 1$ %
FE	0	0	5.0	75.2	99.1	99.9	100.0	100.0
	0	0.1	23.1	70.2	98.3	99.9	100.0	100.0
	0.6	0	27.3	69.0	97.1	99.9	100.0	100.0
	0.6	0.1	38.0	22.0	48.9	69.4	83.8	91.7
RE	0	0	3.8	41.5	75.6	89.4	95.7	98.6
	0	0.1	17.9	52.9	91.1	98.8	99.7	99.9
	0.6	0	8.1	13.4	43.9	71.9	86.5	94.4
	0.6	0.1	13.6	1.0	2.0	2.1	2.7	2.9

"FE" and "RE" denote fixed-effects and random-effects models, respectively; σ^2_{bp} and σ^2_{wp} denote between- and within-population variances, respectively; h^2 represents heritability (phenotypic variation explained by the SNP). Ten individual GWA studies, each with 2,000 individuals, were included in the meta-analysis

is important to recognize failure to detect heterogeneity does not guarantee the absence of heterogeneity.

Effect of population stratification on meta-analysis

Correcting for population stratification within each individual study had a significant effect on the performance of subsequent meta-analyses in that heterogeneity, type-I error rates, and power were all decreased. For example, when this correction was applied, the proportion of replicates with $I^2 > 50$ % decreased from 20.1 to 4.1 %, and the proportion of replicates with $p(\mathrm{Q}) < 0.1$ decreased from 30.1 to 11.2 % (data not shown); power decreased from 97.7 to 71.0 % for the FE model, and from 89.8 to 67.8 % for the RE model; Type-I error rate decreased from 88.4 to 5.2 % for the FE model, and from 87.4 to 4.1 % for the RE model (data not shown).

Meta-analysis versus mega-analysis

When jointly analyzing data from multiple studies, one can either merge participant level data for analysis of a single mega-sample or combine summary statistics for a meta-analysis. In the current study, we compared the performance of meta-analysis and mega-analysis. Our results showed that the FE meta-analysis had an approximately equal performance to the mega-analysis. It is notable that this equivalence held even under heterogeneous setting. For example, when meta-analyzing ten heterogeneous studies ($\sigma^2_{wp} = 0.6$ and $\sigma^2_{bp} = 0.1$) for a SNP with $h^2 = 0.3$ %, the power of the FE meta-analysis and mega-analysis was 85.0 and 84.9 %, respectively, while the type-I error rates were both 37.0 % (data not shown).

Performance of meta-analysis and individual study in a genome-wide scan

Power to identify novel loci

We first compared the power of a smaller individual study with that of a larger meta-analysis. The power of individual GWA studies to identify any particular (pre-selected) locus was very limited (Fig. 4a). For example, the power of individual GWA studies to detect a particular SNP with an h^2 of 0.05 % was as low as <1 %. However, the power to identify any one of the 1,000 SNPs was much better and could increase to 74 %. This phenomenon may explain the reason that individual studies, despite limited sample sizes, have had reasonable success at identifying a number of true novel loci. Therefore, given the large number of yet unidentified loci, individual studies are still valuable in identifying at least some of the underlying causal loci.

Meta-analyses, under the conditions defined for Fig. 4b, had higher power than individual studies in identifying these SNPs; however, the power to identify any particular locus was still rather limited. To detect a SNP with an h^2 of 0.05 %, the power was <10 % for the FE model and <1 % for the RE model. On the other hand, the power to identify any one of the 1,000 SNPs achieved 100 % for both the FE and RE models (Fig. 4b). Importantly, the sample size for this comparison was twice as large for the meta-analysis as it was for the individual studies. Next, we compared the power of meta-analysis versus individual studies under the condition of equal sample size (Fig. 4c vs. a). Results showed that meta-analysis had increased power over analysis of individual studies. However, the type-I error rate was also inflated for meta-analysis (data not shown). Therefore, the increased power of meta-analysis compared with individual studies may actually be attributable to

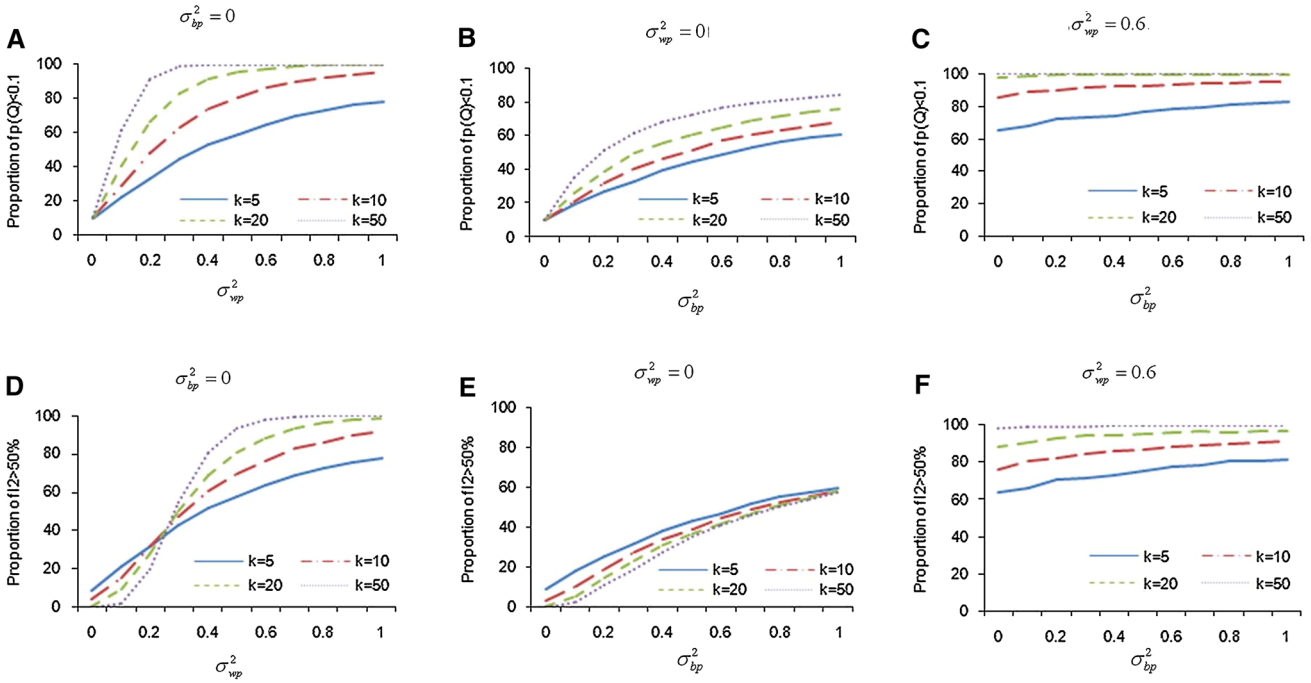


Fig. 3 Power estimates for heterogeneity tests. σ^2_{wp} and σ^2_{bp} represent within- and between-population variances, respectively. k is the number of studies included in the meta-analyses. **a**, **b** and **c** Present power estimates for the Q test, while **d**, **e**, and **f** present power estimates for I^2 index method. For **a** and **d**, only within-population variation was present. For **b** and **e**, only between-population variation was present. For **c** and **f**, both within and between-population variations were present

elevated type-I error rates. To address this, we compared measures of positive predictive value (PPV) (Altman and Bland 1994) which corrects for type-I error rates when estimating power. The results showed that meta-analysis actually had a lower PPV than individual GWA analysis. For example, the PPV for identifying a particular SNP with $h^2 = 0.03$ % was 100, 89, and 80 %, respectively, for individual study, the FE and RE meta-analyses (data not shown). The results were not unexpected given a heterogeneous setting for meta-analysis versus a homogeneous setting for individual studies.

Power to simultaneously identify multiple loci and to identify loci at stringent levels

Our simulation showed that the FE model was quite powerful for identifying multiple loci at the GWS level (Supplementary Fig. 6), while the RE model was more limited. For example, with 80 % power, the number of loci detected by the FE meta-analysis ($k = 10$) was as large as 106, while that for the RE model was only 8. At a more stringent level of significance, the power of both models was limited (Supplementary Fig. 7). At a level of 1×10^{-20}, for example, the power to detect at least one SNP was only 36 % for the FE model and 2 % for the RE model. At a level of 1×10^{-25}, the power of these models decreased to 3 and 0 %, respectively. Even if k increased to 20, the power to detect at least one SNP at the level 1×10^{-25} was only 63 % and <1 % for the FE and RE models, respectively. Therefore, at an extremely stringent level of significance (e.g. $<1 \times 10^{-25}$), the power to detect a SNP was rather small even with a large-scale meta-analysis; detecting multiple SNPs at this level of significance was even smaller or essentially unattainable, under the current simulation settings.

Power of meta-analysis to replicate findings from individual GWA studies

Our simulation demonstrated that the replicability between meta-analysis and an independent individual GWA study was quite limited. On average, individual GWA studies could identify seven of the 1,000 causal SNPs at the GWS level. The power of meta-analysis to replicate at least two of these seven SNPs was 20 % and <1 % for the FE and RE models, respectively (Fig. 5a). Therefore, even powerful meta-analyses had limited power to replicate findings from individual GWA studies. As expected, including more studies into meta-analysis increased its power of replicability (Supplementary Fig. 8), but even then, the power was rather small under the RE model.

When the individual study was included in the samples used for meta-analysis, the power of replicability for the

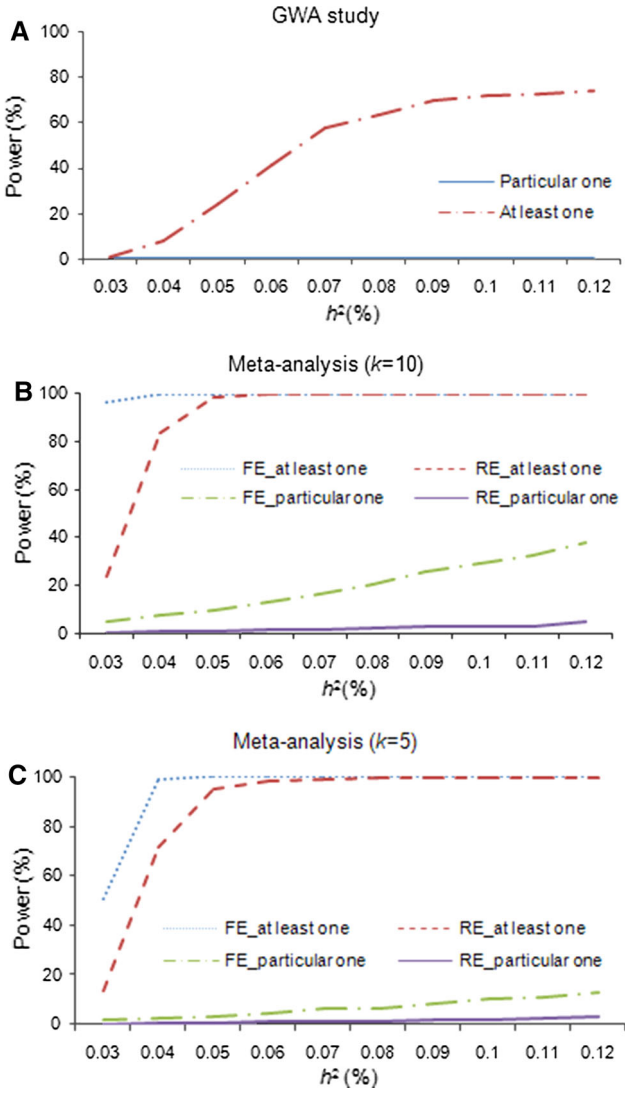


Fig. 4 Power estimates for identifying a particular SNP and at least one SNP on a genome-wide scan. h^2 on the *x*-axis represents phenotypic variation explained by the SNP. "FE" and "RE" denote fixed- and random-effects meta-analyses, respectively. For GWA studies (**a**), a total of 10,000 unrelated individuals were simulated. For meta-analyses, ten (**b**) and five (**c**) individual studies, each with 2,000 unrelated individuals were simulated. For "at least one" case, the x-axis corresponds to the h^2 threshold. The statistically significant level was set at 5×10^{-8}

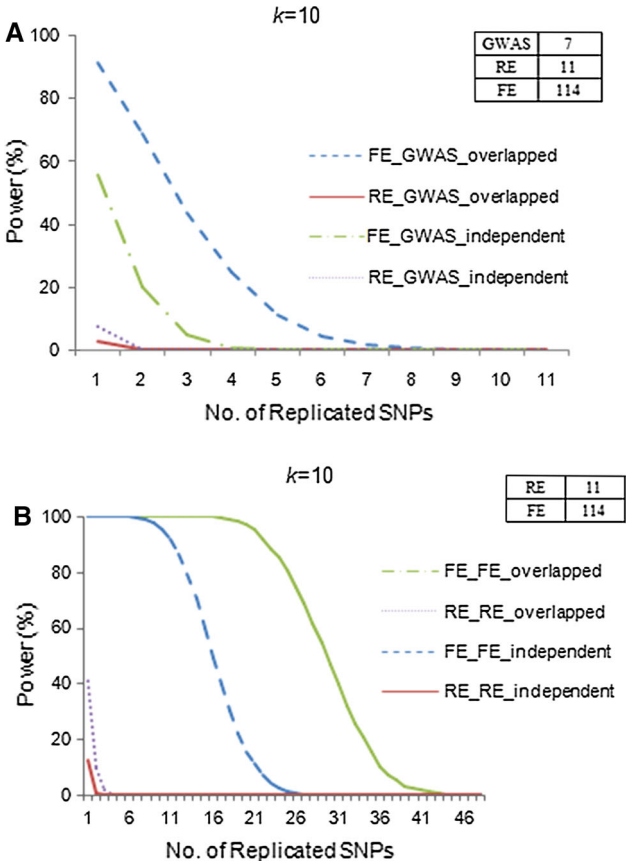


Fig. 5 Power for replication between individual GWA studies (GWAS) and meta-analysis. Tables on the top right of the figures give the average number of significant SNPs identified by each method. "FE" and "RE" denote the fixed- and random-effects meta-analyses, respectively. *k* is the number of studies included in meta-analyses. **a** Shows the power of meta-analysis to replicate findings from individual GWA studies at the genome-wide significance level ($p < 5 \times 10^{-8}$). "Overlapped" denotes that samples from GWA studies were included in the meta-analysis, whereas, "independent" denotes that samples from GWA studies and meta-analysis were independent. **b** Shows the replicability between different meta-analyses at the genome-wide significance level. For "overlapped" simulations, three studies were overlapped between two meta-analyses

FE model improved, while that for the RE model was still limited. For example, the power of the FE model ($k = 10$) to replicate at least two SNPs identified by a discovery GWA study increased to 69 %, while the power of the RE model was still <1 % (Fig. 5a). It is worth emphasizing the point that in such situations consistency of findings between individual GWA studies and meta-analysis did not confer independent replication because of the overlapped samples.

Replicability between meta-analyses

The replicability of results between two distinct meta-analyses was surprisingly low under the simulation conditions described for Fig. 5 in which FE meta-analysis ($k = 10$) identified an average of 114 SNPs at the GWS level. The power of a second independent FE meta-analysis to replicate at least 20 of these 114 SNPs was 15 %, and the power for replicating at least 30 SNPs decreased dramatically to <1 % (Fig. 5b). The RE model ($k = 10$) identified 11 SNPs under these conditions, and the power to replicate at least two of them in a second independent RE

model was <1 %. As expected, the power increased with larger k (Supplementary Fig. 9).

As seen in Fig. 5b, power of replication increased dramatically with overlapping samples. For example, the power to replicate 20 and 30 SNPs increased to 97 and 47 %, respectively, for the FE model. However, the successfully replicated SNPs still represented only a small proportion of the SNPs identified by the discovery meta-analysis. The results of our simulations help to explain the failure of published meta-analyses to replicate the findings of one another.

Discussion

In the present study, we performed comprehensive and comparative investigations to study the relative statistical properties of individual GWA studies and their meta-analyses. Using simulations under a variety of heterogeneity settings, we concluded that detection of heterogeneity was usually difficult and that either FE or RE models could introduce elevated false positive and/or negative rates. Regarding individual GWA studies and meta-analyses, well-designed individual GWA studies had the capacity to identify novel loci that were missed by meta-analyses for complex traits. Non-replicability between meta-analysis and individual GWA studies was not uncommon. Our study may be helpful for the design and interpretation of results from individual GWA studies and their meta-analyses.

Among previous works, Lin et al. compare the performance of meta-analysis and mega-analysis and their results show, theoretically and numerically, that the FE model of meta-analysis is statistically as efficient as analysis of mega-samples, with or without the condition of equal effect size across individual studies (Lin and Zeng 2010a, b). The results of our simulations support their conclusions. Given the approximate equivalence of FE meta-analysis and mega-analysis, the replicability of mega-analysis should be equal to the FE models investigated in the present simulations.

A critical issue in any meta-analysis is the potential effects of heterogeneity among samples (Munafo and Flint 2004). In our simulation studies we modeled heterogeneity effects with two components: within-population and between-population variations. Between-population variation was of particular interest because of its likely existence in real applications, especially for studies involving more than one ethnic group (Torgerson et al. 2011). In the present study, we assumed that between-population heterogeneity only explained a small portion of the phenotypic variation (e.g. <0.19 %) under most scenarios. Our simulation results clearly indicated that between-population heterogeneity, even at these relatively miniscule levels, would significantly impact the results of a meta-analysis (e.g. this heterogeneity would introduce significantly inflated false positive/negative results). Heterogeneity in real data may greatly exceed that simulated in our study, especially when considering interactions between genetic variants and environmental exposures. Consequently, researchers should exercise caution when combining samples from different populations, or should recognize that meta-analyses may introduce false positive/negative results even for samples from the same population.

FE meta-analysis has been extensively investigated (Evangelou and Ioannidis 2013; Ioannidis et al. 2009), and these excellent reviews emphasize the caution necessary when using FE models. Pereira et al. (2009) investigated the properties of discovery GWA meta-analyses under heterogeneous settings. They concluded that FE models could maximize findings of discovery studies at the expense of elevated false discovery rates. The simulations used in our current study represent another attempt to quantitatively measure the impact of heterogeneity on power estimations and type-I error rates. Our results supported the conclusion that the FE model had a rather tenuous assumption (Evangelou and Ioannidis 2013; Ioannidis et al. 2009; Pereira et al. 2009); however, it is more commonly used than the RE model (Gogele et al. 2012), partly because of its improved statistical power.

Under a variety of realistic conditions, even the RE model could not control type-I error rates in the presence of heterogeneity, as demonstrated by our simulations. Consequently, further exploration of potential sources of heterogeneity becomes important (Ioannidis et al. 2007). The inferred source of heterogeneity can be used to correct for its effects through the mixed-effects (van Houwelingen et al. 2002) meta-analysis, Bayesian models (Morris 2011), or sub-group analysis (Bhattacharjee et al. 2012). Among other works (Ioannidis et al. 2007; Patsopoulos and Ioannidis 2010; Zeggini and Ioannidis 2009), Ioannidis et al. (2007) and Patsopoulos and Ioannidis (2010) provide excellent demonstrations for exploiting the source of heterogeneity to help draw causal inferences in the study of type 2 diabetes and rheumatoid arthritis association signals.

The majority of parameter settings in our simulations fell into the ranges of empirical applications conducted to date (Gogele et al. 2012) (Supplementary Table 1). One noticeable exception was the assumption of equal individual study sample sizes (2,000) in our simulation versus varied sample sizes in practice. Our assumption of equal sample sizes was made for ease of presentation, and we did not expect it to have a significant impact on analyses or conclusions. To verify this, we performed additional simulation studies with varying individual sample sizes. Specifically, we simulated individual studies with sample sizes exactly equal to those in a recently published meta-analysis (Estrada et al. 2012), which included 17 GWA studies with

sample sizes varying from 220 to 7,394 (Supplementary notes). The results of these simulations were similar to those attained with equal sample size (Supplementary Fig. 10).

Significant heterogeneity may also be caused by population stratification. In the present study, using a simple illustration, we found that correction for population stratification within each individual study could reduce heterogeneity at the cost of decreased power. One potential explanation for these findings could be that correction for phenotypes by ancestral backgrounds eliminated systematic differences in effect size between ancestral populations within each individual study, which in turn reduced heterogeneity effects in subsequent meta-analysis.

Though individual GWA studies virtually always have a smaller sample size than meta-analysis, they are also more likely to be homogeneous in subject background, study design, and analytical method. Consequently, individual GWA studies are still likely to provide some novel findings. Our simulations clearly demonstrated that well-designed individual GWA studies in homogeneous samples can identify additional novel loci for complex diseases which might escape detection in meta-analysis with larger samples. Consequently, it is critical to recognize that meta-analysis of GWA studies is not necessarily superior to an adequately powered GWA study (Munafo and Flint 2004).

Failure to replicate findings from other studies is common in the genetics field (Ioannidis 2007). Our previous study showed that replication between individual GWA studies is quite difficult in principle (Liu et al. 2008). The present simulation study extends this point to meta-analyses, emphasizing that meta-analyses introduce additional levels of complexity. Therefore, although meta-analysis can be a powerful approach for uncovering novel genetic components underlying complex diseases, it should not be used as a gold standard to evaluate and judge findings from individual studies or from other meta-analyses.

Acknowledgments The study was partially supported by grants from National Natural Science Foundation of China project (31100902), grants from NIH (P50AR055081, R01AG026564, R01AR050496, RC2DE020756, R01AR057049 and R03TW008221) and Edward G. Schlieder Endowed to Tulane University, and benefited from support of Shanghai Leading Academic Discipline Project (S30501) and young teacher startup fund from University of Shanghai for Science and Technology (slg11018 and slg11019). We thank the reviewers for their constructive comments and suggestions. The authors have declared that no competing interests exist.

References

Altman DG, Bland JM (1994) Diagnostic tests 2: predictive values. BMJ 309:102

Altshuler D, Daly M (2007) Guilt beyond a reasonable doubt. Nat Genet 39:813–815

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12

Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, Yeager M, Chung CC, Chanock SJ, Chatterjee N (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet 90:821–835

Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, Warrington NM, Hypponen E, Holst C, Valcarcel B, Thiering E, Salem RM, Schumacher FR, Cousminer DL, Sleiman PM, Zhao J, Berkowitz RI, Vimaleswaran KS, Jarick I, Pennell CE, Evans DM, St Pourcain B, Berry DJ, Mook-Kanamori DO, Hofman A, Rivadeneira F, Uitterlinden AG, van Duijn CM, van der Valk RJ, de Jongste JC, Postma DS, Boomsma DI, Gauderman WJ, Hassanein MT, Lindgren CM, Magi R, Boreham CA, Neville CE, Moreno LA, Elliott P, Pouta A, Hartikainen AL, Li M, Raitakari O, Lehtimaki T, Eriksson JG, Palotie A, Dallongeville J, Das S, Deloukas P, McMahon G, Ring SM, Kemp JP, Buxton JL, Blakemore AI, Bustamante M, Guxens M, Hirschhorn JN, Gillman MW, Kreiner-Moller E, Bisgaard H, Gilliland FD, Heinrich J, Wheeler E, Barroso I, O'Rahilly S, Meirhaeghe A, Sorensen TI, Power C, Palmer LJ, Hinney A, Widen E, Farooqi IS, McCarthy MI, Froguel P, Meyre D, Hebebrand J, Jarvelin MR, Jaddoe VW, Smith GD, Hakonarson H, Grant SF (2012) A genome-wide association meta-analysis identifies new childhood obesity loci. Nat Genet 44:526–531

Cochran W (1954) The combination of estimates from different experiments. Biometrics 10:101–129

Cohen J (1988) Statistical power analysis for the behavioral sciences. Academic Press, New York

de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17:R122–R128

Elks CE, Perry JR, Sulem P, Chasman DI, Franceschini N, He C, Lunetta KL, Visser JA, Byrne EM, Cousminer DL, Gudbjartsson DF, Esko T, Feenstra B, Hottenga JJ, Koller DL, Kutalik Z, Lin P, Mangino M, Marongiu M, McArdle PF, Smith AV, Stolk L, van Wingerden SH, Zhao JH, Albrecht E, Corre T, Ingelsson E, Hayward C, Magnusson PK, Smith EN, Ulivi S, Warrington NM, Zgaga L, Alavere H, Amin N, Aspelund T, Bandinelli S, Barroso I, Berenson GS, Bergmann S, Blackburn H, Boerwinkle E, Buring JE, Busonero F, Campbell H, Chanock SJ, Chen W, Cornelis MC, Couper D, Coviello AD, d'Adamo P, de Faire U, de Geus EJ, Deloukas P, Doring A, Smith GD, Easton DF, Eiriksdottir G, Emilsson V, Eriksson J, Ferrucci L, Folsom AR, Foroud T, Garcia M, Gasparini P, Geller F, Gieger C, Gudnason V, Hall P, Hankinson SE, Ferreli L, Heath AC, Hernandez DG, Hofman A, Hu FB, Illig T, Jarvelin MR, Johnson AD, Karasik D, Khaw KT, Kiel DP, Kilpelainen TO, Kolcic I, Kraft P, Launer LJ, Laven JS, Li S, Liu J, Levy D, Martin NG, McArdle WL, Melbye M, Mooser V, Murray JC, Murray SS, Nalls MA, Navarro P, Nelis M, Ness AR, Northstone K et al (2010) Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. Nat Genet 42:1077–1085

Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, Ntzani EE, Oei L, Albagha OM, Amin N, Kemp JP, Koller DL, Li G, Liu CT, Minster RL, Moayyeri A, Vandenput L, Willner D, Xiao SM, Yerges-Armstrong LM, Zheng HF, Alonso N, Eriksson J, Kammerer CM, Kaptoge SK, Leo PJ, Thorleifsson G, Wilson SG, Wilson JF, Aalto V, Alen M, Aragaki AK, Aspelund T, Center JR, Dailiana Z, Duggan DJ, Garcia M, Garcia-Giralt N, Giroux S, Hallmans G, Hocking LJ, Husted LB, Jameson KA, Khusainova R, Kim GS, Kooperberg C, Koromila T, Kruk M,

Laaksonen M, Lacroix AZ, Lee SH, Leung PC, Lewis JR, Masi L, Mencej-Bedrac S, Nguyen TV, Nogues X, Patel MS, Prezelj J, Rose LM, Scollen S, Siggeirsdottir K, Smith AV, Svensson O, Trompet S, Trummer O, van Schoor NM, Woo J, Zhu K, Balcells S, Brandi ML, Buckley BM, Cheng S, Christiansen C, Cooper C, Dedoussis G, Ford I, Frost M, Goltzman D, Gonzalez-Macias J, Kahonen M, Karlsson M, Khusnutdinova E, Koh JM, Kollia P, Langdahl BL, Leslie WD, Lips P, Ljunggren O, Lorenc RS, Marc J, Mellstrom D, Obermayer-Pietsch B, Olmos JM, Pettersson-Kymmer U, Reid DM, Riancho JA, Ridker PM, Rousseau F, Slagboom PE, Tang NL et al (2012) Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat Genet 44:491–501

Evangelou E, Ioannidis JP (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14:379–389

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Buning C, Cohen A, Colombel JF, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot JP, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panes J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PC, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, Parkes M (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42:1118–1125

Glasziou PP, Sanders SL (2002) Investigating causes of heterogeneity in systematic reviews. Stat Med 21:1503–1511

Gogele M, Minelli C, Thakkinstian A, Yurkiewich A, Pattaro C, Pramstaller PP, Little J, Attia J, Thompson JR (2012) Methods for meta-analyses of genome-wide association studies: critical assessment of empirical evidence. Am J Epidemiol 175:739–749

Higgins JP, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Stat Med 21:1539–1558

Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Ioannidis JP (2007) Non-replication and inconsistency in the genome-wide association setting. Hum Hered 64:203–213

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29:306–309

Ioannidis JP, Patsopoulos NA, Evangelou E (2007) Heterogeneity in meta-analyses of genome-wide association investigations. PLoS One 2:e841

Ioannidis JP, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 10:318–329

Kato N, Takeuchi F, Tabara Y, Kelly TN, Go MJ, Sim X, Tay WT, Chen CH, Zhang Y, Yamamoto K, Katsuya T, Yokota M, Kim YJ, Ong RT, Nabika T, Gu D, Chang LC, Kokubo Y, Huang W, Ohnaka K, Yamori Y, Nakashima E, Jaquish CE, Lee JY, Seielstad M, Isono M, Hixson JE, Chen YT, Miki T, Zhou X, Sugiyama T, Jeon JP, Liu JJ, Takayanagi R, Kim SS, Aung T, Sung YJ, Zhang X, Wong TY, Han BG, Kobayashi S, Ogihara T, Zhu D, Iwai N, Wu JY, Teo YY, Tai ES, Cho YS, He J (2011) Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. Nat Genet 43:531–538

Kavvoura FK, Ioannidis JP (2008) Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. Hum Genet 123:1–14

Lei SF, Yang TL, Tan LJ, Chen XD, Guo Y, Guo YF, Zhang L, Liu XG, Yan H, Pan F, Zhang ZX, Peng YM, Zhou Q, He LN, Zhu XZ, Cheng J, Liu YZ, Papasian CJ, Deng HW (2009) Genome-wide association scan for stature in Chinese: evidence for ethnic specific loci. Hum Genet 125:1–9

Lin DY, Zeng D (2010a) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol 34:60–66

Lin DY, Zeng D (2010b) On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika 97:321–332

Liu YJ, Papasian CJ, Liu JF, Hamilton J, Deng HW (2008) Is replication the gold standard for validating genome-wide association findings? PLoS One 3:e4037

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369

Morris AP (2011) Transethnic meta-analysis of genome wide association studies. Genet Epidemiol 35:809–822

Munafo MR, Flint J (2004) Meta-analysis of genetic association studies. Trends Genet 20:439–444

Patsopoulos NA, Ioannidis JP (2010) Susceptibility variants for rheumatoid arthritis in the TRAF1-C5 and 6q23 loci: a meta-analysis. Ann Rheum Dis 69:561–566

Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP (2009) Discovery properties of genome-wide association signals from cumulatively combined data sets. Am J Epidemiol 170: 1197–1206

Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JPA (2010) Critical interpretation of Cochran's Q test depends on power and prior assumptions about heterogeneity. Res Synth Methods 1:149–161

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

Richards JB, Zheng HF, Spector TD (2012) Genetics of osteoporosis from genome-wide association studies: advances and challenges. Nat Rev Genet 13:576–588

Singh AP, Zafer S, Pe'er I (2013) MetaSeq: privacy preserving meta-analysis of sequencing-based association studies. Pac Symp Biocomput 356–367

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Magi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segre AV, Estrada K, Liang L, Nemesh J, Park JH, Gustafsson S, Kilpelainen TO, Yang J, Bouatia-Naji N, Esko T, Feitosa MF, Kutalik Z, Mangino M, Raychaudhuri S, Scherag A, Smith AV, Welch R, Zhao JH, Aben KK, Absher DM, Amin N, Dixon AL, Fisher E, Glazer NL, Goddard ME, Heard-Costa NL, Hoesel V, Hottenga JJ, Johansson A, Johnson T, Ketkar S, Lamina C, Li S, Moffatt MF, Myers RH, Narisu N, Perry JR, Peters MJ, Preuss M, Ripatti S, Rivadeneira F, Sandholt C, Scott LJ, Timpson NJ, Tyrer JP, van Wingerden S, Watanabe RM, White CC, Wiklund

F, Barlassina C, Chasman DI, Cooper MN, Jansson JO, Lawrence RW, Pellikka N, Prokopenko I, Shi J, Thiering E, Alavere H, Alibrandi MT, Almgren P, Arnold AM, Aspelund T, Atwood LD, Balkau B, Balmforth AJ, Bennett AJ, Ben-Shlomo Y, Bergman RN, Bergmann S, Biebermann H, Blakemore AI, Boes T, Bonnycastle LL, Bornstein SR, Brown MJ, Buchanan TA et al (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42:937–948

Styrkarsdottir U, Halldorsson BV, Gretarsdottir S, Gudbjartsson DF, Walters GB, Ingvarsson T, Jonsdottir T, Saemundsdottir J, Center JR, Nguyen TV, Bagger Y, Gulcher JR, Eisman JA, Christiansen C, Sigurdsson G, Kong A, Thorsteinsdottir U, Stefansson K (2008) Multiple genetic loci for bone mineral density and fractures. N Engl J Med 358:2355–2365

Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, Himes BE, Levin AM, Mathias RA, Hancock DB, Baurley JW, Eng C, Stern DA, Celedon JC, Rafaels N, Capurso D, Conti DV, Roth LA, Soto-Quiros M, Togias A, Li X, Myers RA, Romieu I, Van Den Berg DJ, Hu D, Hansel NN, Hernandez RD, Israel E, Salam MT, Galanter J, Avila PC, Avila L, Rodriquez-Santana JR, Chapela R, Rodriguez-Cintron W, Diette GB, Adkinson NF, Abel RA, Ross KD, Shi M, Faruque MU, Dunston GM, Watson HR, Mantese VJ, Ezurum SC, Liang L, Ruczinski I, Ford JG, Huntsman S, Chung KF, Vora H, Li X, Calhoun WJ, Castro M, Sienra-Monge JJ, del Rio-Navarro B, Deichmann KA, Heinzmann A, Wenzel SE, Busse WW, Gern JE, Lemanske RF Jr, Beaty TH, Bleecker ER, Raby BA, Meyers DA, London SJ, Gilliland FD, Burchard EG, Martinez FD, Weiss ST, Williams LK, Barnes KC, Ober C, Nicolae DL (2011) Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet 43:887–892

van Houwelingen HC, Arends LR, Stijnen T (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 21:589–624

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Wright S (1950) Genetical structure of populations. Nature 166:247–249

Xiong DH, Liu XG, Guo YF, Tan LJ, Wang L, Sha BY, Tang ZH, Pan F, Yang TL, Chen XD, Lei SF, Yerges LM, Zhu XZ, Wheeler VW, Patrick AL, Bunker CH, Guo Y, Yan H, Pei YF, Zhang YP, Levy S, Papasian CJ, Xiao P, Lundberg YW, Recker RR, Liu YZ, Liu YJ, Zmuda JM, Deng HW (2009) Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups. Am J Hum Genet 84:388–398

Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10:191–201