Journal of Human Genetics (2009) 54, 615–623
© 2009 The Japan Society of Human Genetics All rights reserved 1434-5161/09 $32.00
www.nature.com/jhg

REVIEW

Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse

Hirofumi Nakaoka[1,2] and Ituro Inoue[1]

Meta-analysis is a useful tool to increase the statistical power to detect gene–disease associations by combining results from the original and subsequent replication studies. Recently, consortium-based meta-analyses of several genome-wide association (GWA) data sets have discovered new susceptibility genes of common diseases. We reviewed the process and the methods of meta-analysis of genetic association studies. To conduct and report a transparent meta-analysis, the search strategy, the inclusion or exclusion criteria of studies and the statistical procedures should be fully described. Assessing consistency or heterogeneity of the associations across studies is an important aim of meta-analysis. Random effects model (REM) meta-analysis can incorporate between-study heterogeneity. We illustrated properties of test for and measures of between-study heterogeneity and the effect of between-study heterogeneity on conclusions of meta-analyses through simulations. Our simulation shows that the power of REM meta-analysis of GWA data sets (total case–control sample size: 5000–20 000) to detect a small genetic effect (odds ratio (OR)=1.4 under dominant model) decreases as between-study heterogeneity increases and then the mean of OR of the simulated meta-analyses passing the genome-wide significance threshold would be upwardly biased (*winner's curse* phenomenon). Addressing observed between-study heterogeneity may be challenging but give a new insight into the gene–disease association.

Journal of Human Genetics (2009) **54,** 615–623; doi:10.1038/jhg.2009.95; published online 23 October 2009

Keywords: genome-wide association study; heterogeneity; meta-analysis; winner's curse

INTRODUCTION

Population-based association studies provide a powerful approach to the identification of susceptibility genes underlying common diseases.[1,2] A very large amount of information about genetic variants in the human genome has been accumulated through the International Human Genome Sequencing Project and the International HapMap Project.[3–6] Combined with the establishment of high-throughput single-nucleotide polymorphism (SNP) typing systems, genome-wide association (GWA) studies have been widely applied.[7] Accordingly, gene–disease associations have been reported.

Replication studies were extensively implemented to establish the credibility of the initial positive findings. However, comprehensive reviews of the published literatures in the era of the candidate gene approach show that most of the initial positive associations were not reproduced in the subsequent replication studies.[8–13] These findings suggest that a large number of original findings were false-positive reports and another possibility is that most of the studies were underpowered to detect small genetic effect.[8,9] Furthermore, inconsistency or between-study heterogeneity of results of genetic associations can be observed regardless of whether the associations are true or not,[10,14] and it may be attributed to population stratification, genotyping errors, differences in the pattern of linkage disequilibrium (LD) structure and other factors.[15,16] In the era of GWA studies, this problem remains one of the most difficult issues of genetic association studies.[10,15,16] For example, the large-scale international study of Parkinson's disease failed to replicate 13 SNPs identified by the previous GWA study.[17]

In these circumstances, meta-analysis can be a useful tool to combine both statistically significant and nonsignificant results from individual studies on the same research question. In case–control study, the odds ratios (ORs) for individual studies are combined to calculate a summary OR. Meta-analysis improves the estimation of a summary OR and 95% confidence interval (CI) and increases the statistical power to detect gene–disease associations.[18] Therefore, conclusions from a meta-analysis are more robust than those from a single small study. In addition, meta-analysis is useful to investigate the consistency or heterogeneity of the associations across studies. Testing for and quantifying between-study heterogeneity is an

[1]Division of Molecular Life Science, School of Medicine, Tokai University, Isehara, Kanagawa, Japan and [2]The Japan Health Sciences Foundation, Chuo-ku, Tokyo, Japan
Correspondence: Professor I Inoue, Division of Molecular Life Science, Tokai University, School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan.
E-mail: ituro@is.icc.u-tokai.ac.jp

Received 6 August 2009; revised 4 September 2009; accepted 15 September 2009; published online 23 October 2009

important aim of meta-analyses to determine whether there are differences underlying the results of the study.[19,20] Addressing the observed between-study heterogeneity could generate a new insight into the gene–disease association.[20]

In this review, we begin with describing the process of meta-analysis of genetic association studies. The statistical backgrounds, methodological issues and sources of between-study heterogeneity of meta-analysis of genetic association studies are briefly reviewed. Finally, we present the results of our simulation study to illustrate the effect of between-study heterogeneity on conclusions of meta-analyses.

LITERATURE-BASED META-ANALYSIS

In a basic meta-analysis, data are retrospectively collected from published literatures to assess whether a gene–disease association of interest is true or not.[18] When planning a meta-analysis, it is important to define precise search strategy beforehand.[21] If relevant studies are excluded or inadequate studies are included, conclusions of the meta-analysis may be biased.[22] The literature search is conducted in databases such as PubMed and EMBASE. The HuGe Published Literature database (http://www.cdc.gov/genomics/hugenet/) is also useful, as it includes published literatures on genetic associations and other human genome epidemiology.[23] It is important to collect the largest possible number of studies; therefore, we should use appropriate key words. Once the search has been completed, bibliographies of retrieved articles should be examined for further relevant publications.

These processes make up the essential part of the methods section of a meta-analysis, because literature-based meta-analysis is subjected to bias caused by difficulty to identify and include all conducted and relevant studies,[13,24] and small difference in selected literatures may alter conclusions of meta-analyses on the same genetic association.[25] However, the essential features of the search strategy have not fully reported in most meta-analyses of genetic association studies.[26] In order to avoid such biases, it may be recommended to have two or more different researchers conducting the same search.[21] When conducting and reporting a literature-based meta-analysis, flowchart detailing the exclusion and inclusion criteria and the number of studies excluded and included at each step of the literature search is useful (Figure 1).

Meta-analysis of genetic association studies may be subjected to publication bias.[18,26] Publication bias tends to occur when small studies showing negative or nonsignificant results remain unpublished and may result in the overestimation of the genetic effect. If the presence of publication bias is suspected by statistical tests,[27,28] conclusions from the meta-analysis should be cautiously reported and the potential impact of the publication bias should be mentioned.[18]

The results obtained from the meta-analysis would be assessed by the following: (i) the size of the summary OR; (ii) the extent and possible cause of between-study heterogeneity; and (iii) the sufficiency and stability of the meta-analysis by using the cumulative and recursive cumulative meta-analysis approaches.[29–31] In the cumulative meta-analysis, studies are sorted chronologically and a summary OR is calculated when a new study is added.[29] As a result, we can present how the summary OR has shifted over time. The recursive cumulative meta-analysis is an extension of the cumulative meta-analysis, where the relative change in the summary OR by adding a new study is evaluated.[30,31]

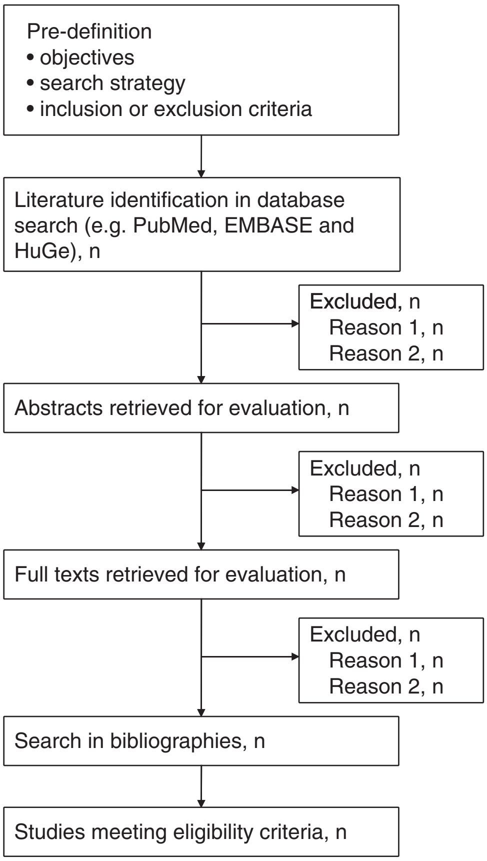
Pre-definition
• objectives
• search strategy
• inclusion or exclusion criteria
Literature identification in database search (e.g. PubMed, EMBASE and HuGe), n
Excluded, n
Reason 1, n
Reason 2, n
Abstracts retrieved for evaluation, n
Excluded, n
Reason 1, n
Reason 2, n
Full texts retrieved for evaluation, n
Excluded, n
Reason 1, n
Reason 2, n
Search in bibliographies, n
Studies meeting eligibility criteria, n

Figure 1 Flowchart detailing the exclusion and inclusion criteria and the number of studies excluded and included at each step of the literature search.

CONSORTIUM-BASED META-ANALYSIS

Consortium-based meta-analysis is the meta-analysis of individual patient data through the collaboration of consortium of investigators. Consortium-based meta-analysis attains increased attention,[32–34] because integration of several GWA data sets has been designed and new susceptibility genes have been discovered.[35–39] Although meta-analysis of GWA studies can be implemented using reported ORs and 95% CIs or *P*-values from different GWA studies, it is preferable to reanalyze several GWA data sets with individual patient data.[35] In the latter case, one can use imputation techniques for missing data when SNPs have been genotyped in some platforms but not in others.[40] Barrett *et al.*[39] conducted a meta-analysis of three GWA data sets for Crohn's disease that used different genotyping platforms using imputation methods. The combined GWA data sets included 635 547 SNPs in 3230 cases and 4829 controls. They used the GWA data sets at the screening stage. The power of the meta-analysis was reported to be 0.74 to detect associations with per allele OR of 1.2 and with risk allele frequency of 0.2 at the significance level of $P=1.0\times10^{-5}$. The meta-analysis of the GWA data sets and additional replication data sets confirmed 11 previously reported loci and identified genome-wide significant signals for novel 21 loci.

GENETIC ASSOCIATION STUDY-SPECIFIC METHODOLOGICAL ISSUES

There are methodological issues relevant to meta-analysis of genetic association studies: (i) assessment of Hardy–Weinberg equilibrium (HWE) and (ii) definition of genetic models.

Deviation from HWE in control samples is the most commonly used test for genotyping error.[41] However, the test for HWE has relatively low statistical power to detect genotyping error.[42]

Furthermore, SNPs that are not in HWE can be used for inference about genetic model of disease susceptibility at the locus.[43] Although there is no consensus how meta-analyses should handle the studies that are not in HWE, three strategies have been applied: including all studies regardless of departure from HWE,[44] performing sensitivity analyses in order to evaluate whether the genetic effects are different between subgroups of studies classified according to test for HWE[26,45–47] and excluding studies showing statistically significant departure from HWE.[18] Reporting the extent of departure from HWE measured by such as α,[48] the inbreeding coefficient,[49] and the disequilibrium parameter[50] is also useful.[44]

In a genetic association study, subjects are classified into three exposure groups (*AA*, *Aa* and *aa*). Let *A* be the susceptibility allele, there are several methods of dichotomizing these exposure groups for conducting a meta-analysis:[26] by comparing allele frequency, by assuming a specific mode of inheritance (recessive, dominance, complete overdominant or codominant) and by performing multiple pairwise comparisons. All these methods, with exception of the method performing multiple pairwise comparisons, assume a particular genetic model. When performing multiple pairwise comparisons or testing multiple genetic models, results of all analyses undertaken should be reported. In order to choose most likely genetic model describing the genetic architecture underlying a disease of interest, Minelli *et al.*[51] presented a 'genetic model free' approach. Their procedure is based on the estimation of the ratio (λ) of the log OR of *Aa* versus *aa* compared with the log OR of *AA* versus *aa*. λ will be 0 under a recessive model, 0.5 under a codominant model and 1 under a dominant model.

ESTIMATION OF A SUMMARY OR AND TEST FOR AND MEASURE OF BETWEEN-STUDY HETEROGENEITY

The statistical methods of combining the results of different studies are described. We consider a meta-analysis of k separate genetic association studies to estimate the genetic effect (θ) for dichotomous disease outcome quantified by log OR. Let θ_i and $\hat{\theta}_i$ be the true and observed log OR for *i*th case–control study, respectively (i=1, … ,k). Let v_i denote the variance of $\hat{\theta}_i$, the weight for *i*th study is given by w_i=1/v_i (that is, the inverse of the variance). OR for each study is given by $\mathrm{OR}_i = a_i d_i / b_i c_i$. $\hat{\theta}_i = \ln(\mathrm{OR_i})$. v_i is defined as $v_i = 1/a_i + 1/b_i + 1/c_i + 1/d_i$, where a_i and b_i correspond to numbers of affected individuals with and without the susceptible genotype, respectively, and c_i and d_i correspond to numbers of unaffected individuals with and without the susceptible genotype, respectively.

There are two commonly used procedures for combining $\hat{\theta}_i$s: 'fixed effects model' (FEM) and 'random effects model' (REM). FEM assumes that θ_is are homogeneous across studies (that is, $\theta_1 = \theta_2 = \ldots = \theta_k$) and all differences are due to chance. Inverse-variance, Mantel-Haenszel[52] and Peto's[53] methods are commonly used for FEM meta-analysis. Using the inverse-variance method for combining the results across studies, a summary log OR under FEM is calculated as a weighted average of the study estimates: $\hat{\theta}_{\mathrm{FEM}} = (\sum_{i=1}^{k} w_i \hat{\theta}_i)/(\sum_{i=1}^{k} w_i)$. The variance of $\hat{\theta}_{\mathrm{FEM}}$ is given by $v_{\mathrm{FEM}} = 1/\sum_{i=1}^{k} w_i$.

The assumption underlying FEM should be examined with the test for heterogeneity, Cochran's Q test.[54] Test statistics of Cochran's Q test is

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}_{\mathrm{FEM}} \right)^2$$

Under the null hypothesis of homogeneity (that is, $\theta_1 = \theta_2 = \ldots = \theta_k$), this statistics approximately follows a χ^2 distribution with $k-1$ degrees of freedom. Cochran's Q test has relatively low statistical power to detect between-study heterogeneity, especially when the number of studies is small;[55] therefore, the test is usually preformed at the significance level of 0.1.[56]

REM assumes that the genetic effects may vary across studies because of genuine difference and/or differential biases. The estimate of the between-study variance (τ^2) is included into the weight as $w_i^{'} = 1/(w_i^{-1} + \hat{\tau}^2)$. A summary log OR under REM are estimated as follows: $\hat{\theta}_{\mathrm{REM}} = (\sum_{i=1}^{k} w_i^{'} \hat{\theta}_i)/(\sum_{i=1}^{k} w_i^{'})$. The variance of $\hat{\theta}_{\mathrm{REM}}$ is approximated as $v_{\mathrm{REM}} = 1/\sum_{i=1}^{k} w_i^{'}$.

In DerSimonian and Laird[57] REM meta-analysis, the τ^2 is estimated as follows:

$$\hat{\tau}^2_{\mathrm{DL}} = \frac{Q - (k-1)}{\sum_{i=1}^{k} w_i - \left(\sum_{i=1}^{k} w_i^2 / \sum_{i=1}^{k} w_i \right)}$$

When $Q < k-1$, $\hat{\tau}^2_{DL}$ takes negative value. In practice, $\max\{0, \hat{\tau}^2_{DL}\}$ is used. Therefore, the precision of a summary log OR with REM ($1/v_{\mathrm{REM}}$) can never exceed that with FEM ($1/v_{\mathrm{FEM}}$).

The 95% CI for $\hat{\theta}$ is given by $\hat{\theta} \pm 1.96 \times \sqrt{v}$. Test statistic of test for the genetic effect is given by $Z = \hat{\theta}/\sqrt{v}$. Under the null hypothesis, Z follows a standard normal distribution.

Higgins and Thompson[58] proposed three criteria (H, R and I^2) for measure of heterogeneity, which have following desired characteristics: (i) dependence on the extent of heterogeneity, (ii) scale invariance (that is, comparison can be made across meta-analyses with different scales and different outcomes) and (iii) size invariance (that is, independence on the number of studies included). $H = \sqrt{Q/(k-1)}$ is the relative excess of Q to its degrees of freedom. Mittlbock and Heinzl[59] proposed $H^2_{\mathrm{M}} = \frac{Q-(k-1)}{k-1}$ as a modification of H. H^2_{M} is the proportion of between-study variance to within-study variance. In practice, $\max\{0, H^2_{\mathrm{M}}\}$ is used. H^2_{M} values over 1.0 indicate considerable heterogeneity.[59] $R = \sqrt{v_{\mathrm{REM}}/v_{\mathrm{FEM}}}$ is the ratio of the standard error of a summary effect with REM to the standard error with FEM. R represents the inflation of the CI for REM compared with FEM. H and R coincide when all studies have equal weight.[58] $I^2 = 100 \times \frac{Q-(k-1)}{Q}$. I^2 can take negative value, but $\max\{0, I^2\}$ is used in practice. I^2 represents the proportion of between-study variance to the total variation in study estimates and ranges from 0 to 100%. I^2 is most widely used for measure of heterogeneity. I^2 values over 50% indicate large heterogeneity.[58,60] Potential drawback of I^2 is that CIs are very large, especially when the number of studies is small.[61]

If heterogeneity is present or suspected by the statistical test or measures, there are several commonly used approaches: (i) performing sensitivity analysis by excluding one or more studies showing outlier effect size, (ii) stratifying the studies into homogeneous subgroups such as racial groups and applying FEM for each subgroup and (iii) implementing REM when observed heterogeneity could not be addressed. Some researchers recommend that the use of REM is preferable compared with FEM, because both models give similar summary effects when there is no between-study heterogeneity, FEM gives narrower CI for summary effect compared with REM when between-study heterogeneity exists and a negative result of test for heterogeneity does not always indicate homogeneity when the number of studies is small.[25]

SOURCE OF HETEROGENEITY

A number of reasons have been advanced for heterogeneity in the genetic effects across the results of various studies.[8,13,14,47] False-positive results in the initial studies and false-negative results in small replication studies are implicated as the most likely reasons for non-replications.[8–10,13,14] Inconsistency and between-study heterogeneity may be caused because of biases or genuine differences in the genetic effects across populations. We review briefly in this article.

Biases

Differential biases due to population stratification, misclassification of clinical outcome, genotyping error and overestimation of genetic effect in the first study can be sources of between-study heterogeneity.

The presence of population stratification tends to spurious associations. It can be caused when there are undetected genetically different subgroups within a study population and disease prevalence differs among these subgroups.[11,62] The effect of population stratification on the results of genetic association studies is debatable.[62–66] According to systematic reviews of meta-analyses of genetic association studies, it is not so much frequent that difference in racial or ethnic groups could explain heterogeneity.[9,67]

Inadequate assignment of cases and controls may cause misclassification bias. Although there is a possibility that misclassification of cases and controls would weaken the gene–disease association, the results of misclassification bias may be modest unless the trait is common.[13,32]

Ioannidis *et al.*[10] conducted a systematic review of 36 meta-analyses including a total of 370 genetic association studies. Statistically significant between-study heterogeneity was observed in 14 meta-analyses. Restricting to meta-analyses with at least 15 studies, 7 of 9 meta-analyses showed significant heterogeneity. In 25 or 26 meta-analyses, the first study showed more predisposing or protective OR than subsequent replication studies. Using cumulative meta-analysis plots, the authors depicted the process that strong associations claimed in the first study were regressed toward null associations, as subsequent replication studies were accumulated over time. Similar findings were reported in Lohmueller *et al.*[9] Associations passing predetermined thresholds of statistical significance tend to overestimate the size of the genetic effect, especially when the sample size of the study is small and the threshold is stringent in multiple testing situations.[68–74] Such an upward bias is called as *winner's curse* phenomenon.[9,69]

Genuine differences

Differences in the pattern of LD structure over chromosomal regions of interest across populations are implicated as a cause of between-study heterogeneity in the genetic effects. Zondervan and Cardon[75] show that marker allelic OR can vary according to the extent of LD between marker and true disease allele in terms of D' and according to mismatch between disease allele frequency and marker allele frequency. This issue may be especially pronounced in the GWA settings because the SNPs that most efficiently surrogate the other SNPs in a genomic region with high LD (that is, tag SNPs) rather than putative functional SNPs have been used to increase genome coverage. When the extent of LD between tag SNP and true disease allele varies across studied populations, the observed ORs could vary across studies.

Many common diseases are implicated to have a complex etiology involving multiple genetic and environmental factors including their interactions. Gene–disease associations can be modified when the gene–gene or gene–environment interaction exists. If these interactions are not identified and controlled for, the gene–disease associations would be heterogeneous across populations according to distribution of a genetic variant or prevalence of a particular environmental exposure. It is needed to conduct a consortium-based meta-analysis of individual patient data in large scale to account for gene–gene or gene–environment interactions.[47]

SIMULATION STUDY

We conducted a simulation study to illustrate (i) the power of Cochran's Q test, (ii) the properties of measures of between-study heterogeneity (I^2 and H^2_{M}) and (iii) the type I error rate and the power of meta-analysis for detecting the gene–disease association in the presence of between-study heterogeneity.

We consider meta-analysis of k case–control association studies to estimate the overall genetic effect (θ; log OR) of disease outcome. The exposure status (AA, Aa and aa) of subjects included in each case–control study are ascertained in the sampling manner outlined below.[70] The values $y \in \{1, 0\}$ are labels encoding case (1) or control (0). Let A denote the susceptibility allele, we assume the dominant model and then the SNP genotype predictor value x was designed as $1{=}AA$ or Aa, $0{=}aa$. Under the assumption of HWE, the frequency of x written as f_x is calculated based on the disease allele frequency f_{A}: $f_1{=}1-(1-f_{\mathrm{A}})^2$. The logistic regression model for ith study ($i = 1, 2, \ldots, k$) is produced as follows:

$$\log\left(\Pr(Y=1|x)/(1-\Pr(Y=1|x))\right) = \alpha_i + \theta_i x$$

where α_i is the intercept and θ_i is the log OR for ith study. θ_i is drawn from $N(\theta, \tau^2)$. τ^2 is the between-study variance. α_i can be calculated by using the equation for the prevalence of the disease $\pi = \sum_x \frac{\exp(\alpha_i+\theta_i x)}{1+\exp(\alpha_i+\theta_i x)} \times f_x$. The genotypes of case and control subjects are generated based on the conditional probabilities of x given by y as follows:

$$\Pr(X=x|Y=1) = \frac{f_x}{\pi} \times \frac{\exp(\alpha_i+\theta_i x)}{1+\exp(\alpha_i+\theta_i x)},$$

$$\Pr(X=x|Y=0) = \frac{f_x}{1-\pi} \times \frac{1}{1+\exp(\alpha_i+\theta_i x)}$$

For each study, the genotypes of case–control samples were generated and then the OR and its variance were calculated. Then, the ORs for k studies were combined by FEM and REM meta-analyses. Cochran's Q test was conducted and the I^2 and H^2_{M} were measured.

We considered simple five simulation scenarios of meta-analyses. The description of simulation scenarios is shown in Table 1. The scenarios I, II and III were designed to be same in sample size within each study but different in the number of included studies. In scenarios III, IV and V, numbers of studies were different but total number of case–control samples included in meta-analysis was fixed at 20 000. The pairs of scenarios I and V or II and IV were designed to have the same number of studies but differ in sample size within each study.

We examined 126 parameter combinations for each scenario. The between-study variance (τ^2) varied from 0.0 to 0.02 with increments of 0.001. The true summary OR ($\exp(\theta)$) was set to be 1.0, 1.4 or 2.0. The disease allele frequency f_{A} was assigned to be 0.1 or 0.3. The disease prevalence π was fixed at 0.01. The values of τ^2 were based on the literature values reported by Moonesinghe *et al.*[76] for the confirmed 10 loci in a meta-analysis of three GWA studies of type 2 diabetes.[77] Therefore, our simulation would reflect the possible range of between-study variance. For each scenario and parameter combination, 100 000 simulations were carried out.

Table 1 Description of five simulation scenarios of meta-analysis

Scenario	k	n*case*/n*control*
I	5	500/500
II	10	500/500
III	20	500/500
IV	10	1000/1000
V	5	2000/2000

k denotes the number of included studies and n_{case} and n_{control} are the number of cases and controls within each study, respectively.

The empirical power of Cochran's *Q* test was evaluated by the proportion of the simulation runs crossing the significance level of 0.1 when $\tau^2>0.0$. The top row of Figure 2 shows the powers of Cochran's *Q* test obtained with five scenarios as the function of τ^2 when the overall OR=1.0 and f_A=0.1 or 0.3. For each scenario, the power increased as τ^2 increased. Comparing among scenarios I, II and III, the power increased as the number of studies increased. When total number of case–control samples was fixed (that is, comparing among scenarios III, IV and V), the powers were similar but scenarios with smaller number of studies showed higher power

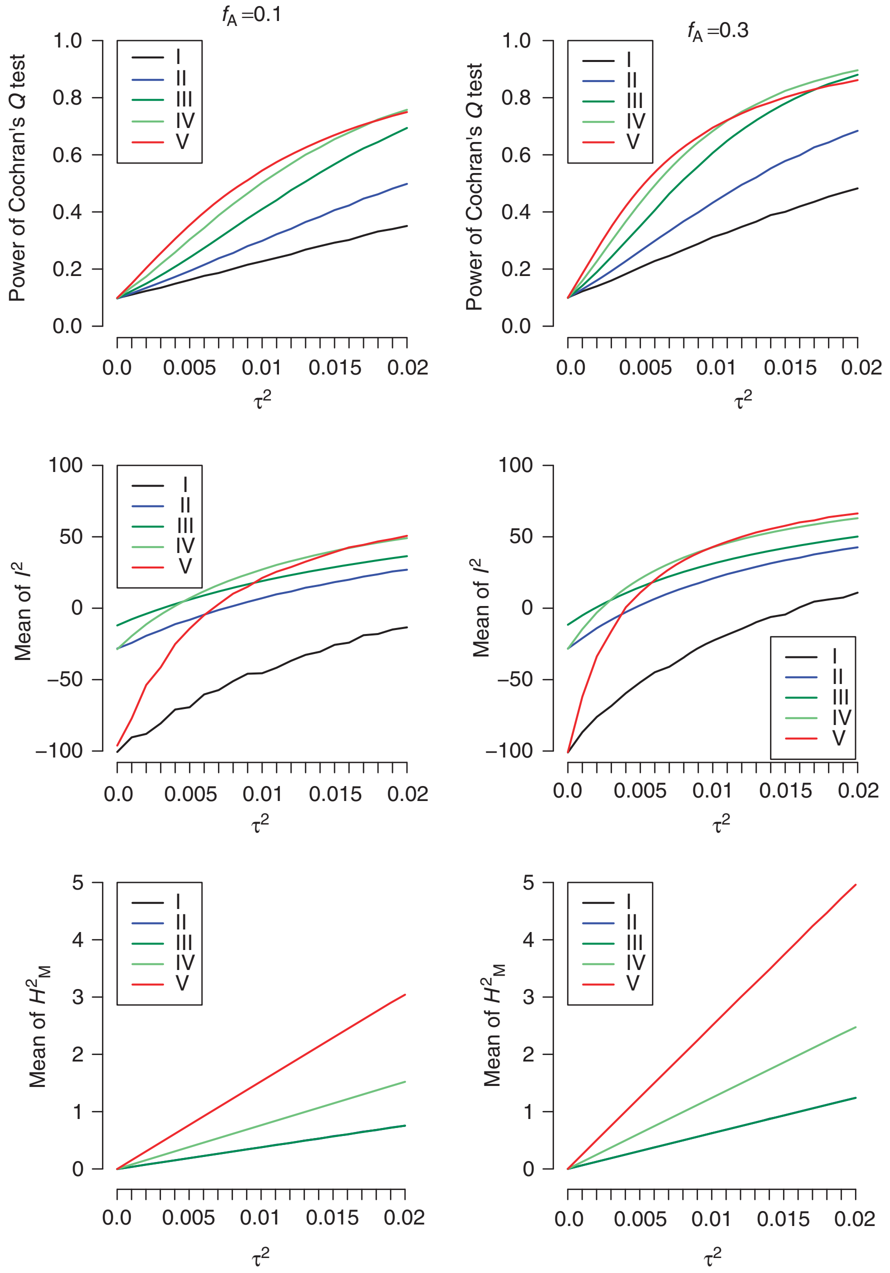

Figure 2 Behaviors of test for and measures of between-study heterogeneity for five simulation scenarios as the function of τ^2, the disease allele frequency f_A=0.1 or 0.3, and the overall odds ratio (OR)=1.0. The top row shows the power of the Cochran's *Q* test at the significance level of 0.1. The middle and bottom rows show the means of I^2 and H^2_M, respectively. The lines of H^2_M for scenarios I, II and III are overlapping. The description of each simulation scenario is in Table 1.

when τ^2 was small. When numbers of studies were identical (that is, two pairwise comparisons of scenarios I versus V or II versus IV), meta-analyses with larger sample size showed higher power for the same τ^2. The powers obtained with f_A=0.3 were higher than those with f_A=0.1. For most of our parameter settings, the powers of Cochran's Q test did not reach at 0.8, although the significance level was set to be 0.10.

The means of 100 000 simulated values for the measures of heterogeneity (I^2 and H_M^2) are shown as the function of τ^2 when the overall OR=1.0 and f_A=0.1 or 0.3 (the middle and bottom rows of Figure 2). In practice, max{0, I^2} and max{0, H_M^2} are used to restrict the ranges of these measures as positive. As the simulation study of Mittlbock and Heinzl,[59] unrestricted values of I^2 and H_M^2 were used to obtain unbiased distributions for these measures in this study. These two measures presented monotonic increases as τ^2 increased. I^2 and H_M^2 increased as the sample size per study increased (scenarios I versus V or II versus IV). The two measures obtained with f_A=0.3 were higher than those with f_A=0.1. These results indicate that I^2 and H_M^2 increased as within-study variance, $k/(\sum_{i=1}^{k} w_i)$, decreased. Comparing scenarios I, II and III shows the important difference between I^2 and H_M^2: whereas I^2 increased as the number of studies increased, H_M^2 did not change (the lines of H_M^2 for scenarios I, II and III are overlapping in the bottom rows of Figure 2). This suggests that H_M^2 may be a good indicator of comparing the extent of between-study heterogeneity across meta-analyses. Similar results and further discussion are provided by Mittlbock and Heinzl.[59] The 95% intervals of simulated I^2 and H_M^2 were large, especially when the number of studies is small (Supplementary Figure S1).

The type I error rate in meta-analysis was assessed as the proportion of the simulation runs showing significant summary OR at the significance level of 0.05 when the null hypothesis was true (that is, the true overall OR=1.0). Figure 3 shows the type I error rates of five scenarios when f_A=0.1 or 0.3. When there was no between-study variance (τ^2=0.0), the type I error rates under FEM were well controlled at 0.05, but REM showed slightly conservative results (the type I error rate ≈ 0.04). As τ^2 increased, the type I error rates under FEM rapidly inflated, but those under REM slightly increased. The type I error rates under both models for the same τ^2 increased when sample size per study was large or f_A=0.3. We should note that the use of FEM could increase the type I error rate even to the extent that the between-study heterogeneity could not be fully identified by Cochran's Q test and two measures I^2 and H_M^2. For example, in case of τ^2=0.005 and f_A=0.3, the type I error rate under FEM for five scenarios were 8.5–19.2% (Figure 3). For the parameter setting, the powers of Cochran's Q-test were 20.6–48.3%, the means of I^2 were −51.9 to 20.8% and the means of H_M^2 were 0.31–1.25 (Figure 2).

The power of detecting a gene–disease association was evaluated as the proportion of simulation runs reaching the significance level of 5.7×10^{-7}, assuming the consortium-based meta-analysis of GWA data sets. As shown in Figure 3, applying FEM meta-analysis to heterogeneous genetic associations could lead to false-positive findings; therefore, we considered only REM when assessing the power of

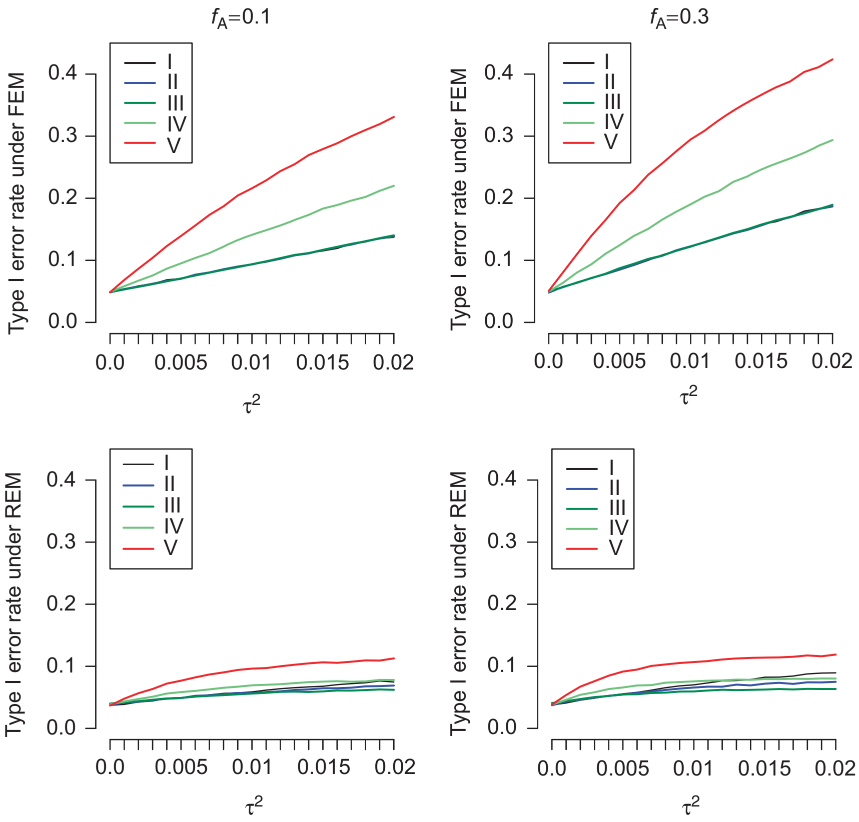

Figure 3 The type I error rate in fixed effects model (FEM) and random effects model (REM) meta-analyses at the significance level of 0.05 for five scenarios as the function of τ^2 and the disease allele frequency f_A=0.1 or 0.3. The top and bottom rows show the type I error rates when applying FEM and REM, respectively. The lines of the type I error rate under FEM for scenarios I, II and III are overlapping. The description of each simulation scenario is in Table 1.

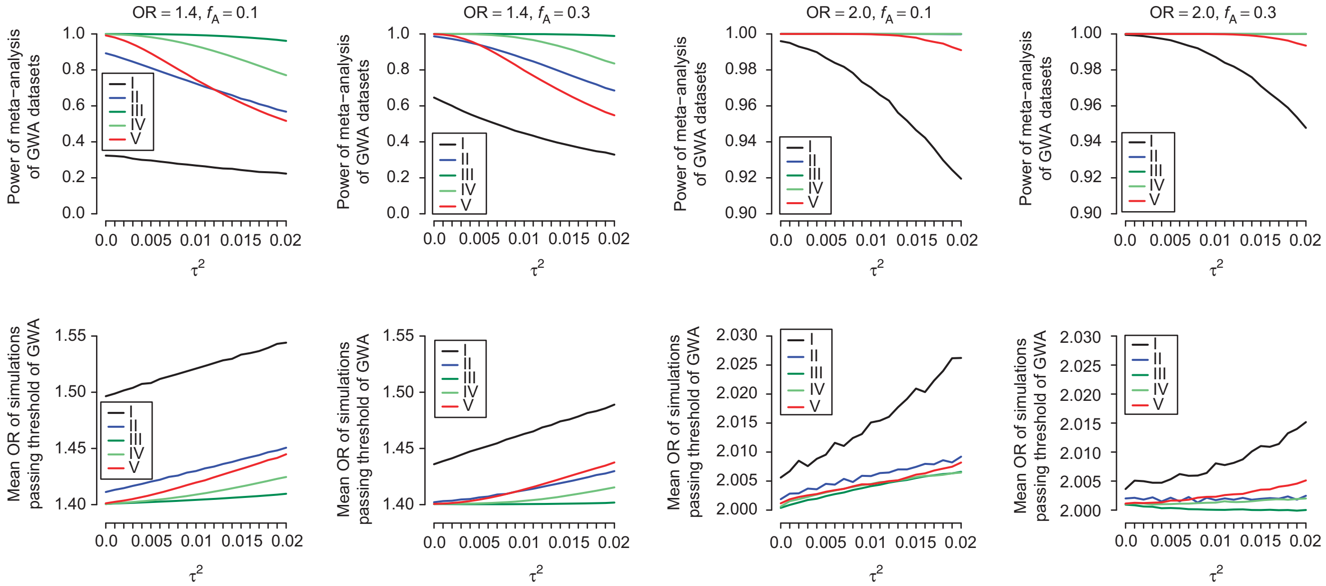

Figure 4 Simulations for the powers in random effects model (REM) meta-analyses of detecting a gene–disease association at the significance level of 5.7×10^{-7} (the top row) and the mean odds ratio (OR) of the simulations passing the threshold (the bottom row) as the function of τ^2, the disease allele frequency f_A=0.1 or 0.3, and the overall OR=1.4 or 2.0. When the overall OR=2.0, the lines of the powers for scenarios II, III and IV are overlapping. The description of each simulation scenario is in Table 1.

meta-analysis. The top row of Figure 4 shows the result, assuming the dominant model and f_A=0.1 or 0.3. When the true overall OR=1.4, the power for each scenario gradually decreased as τ^2 increased. Comparing scenarios III, IV and V, the decreases in the power for the same τ^2 were larger in the scenarios with large sample size per study. While the values of v_{FEM} for scenarios III, IV and V were not different, the values of v_{REM} for scenarios III, IV and V varied when between-study heterogeneity was present. For the same τ^2 (>0), the following inequality was true: v_{REM} for scenario V$>$$v_{\mathrm{REM}}$ for scenario IV$>$$v_{\mathrm{REM}}$ for scenario III. When $\theta\neq0$, the mean of the distribution of the Z-test under REM is $\lambda=\theta/\sqrt{v_{\mathrm{REM}}}$. The power of detecting gene-disease association of effect size of θ is[78]

$$\mathrm{Power}=1-\Phi(C_{\alpha/2}-\lambda)+\Phi(-C_{\alpha/2}-\lambda)$$

where Φ is the cumulative distribution function of the standard normal and $C_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution. Along with the inequality described above, the decrease in the power for the same τ^2 is larger in the scenarios with large sample size per study when the total sample sizes are equal across scenarios. When the overall OR was set to be 2.0, the powers did not so much decrease in the simulated range of τ^2. Furthermore, we calculated the mean OR of the simulations passing the genome-wide significance threshold (P-value$<5.7\times10^{-7}$). The estimates of mean OR were upwardly biased, especially in scenarios whose powers of detecting gene–disease associations were low (the bottom row of Figure 4). On the other hand, if the meta-analyses were sufficiently powered (for example, the true overall OR=2.0), upward biases were not so pronounced in the simulated range of τ^2.

Our simulation suggests that the power of meta-analysis of GWA data sets to detect small genetic effect would decrease due to between-study heterogeneity ($\tau^2\sim0.02$). As a result, the discovered gene–disease association could have inflated effect (*winner's curse* phenomenon). Such a *winner's curse* phenomenon can be seen even to the extent that the between-study heterogeneity could not be fully identified. Similar results were obtained when different genetic models (that is, recessive and additive in log-odds scale models) were examined (data not shown).

CONCLUSION

We reviewed the process and the methods of meta-analysis of genetic association studies. To conduct and report a transparent meta-analysis, the search strategy, the inclusion or exclusion criteria of studies and the statistical procedures should be fully described. Assessment of HWE and determination of genetic model are methodological issues relevant to meta-analysis of genetic association studies.

In genetic association studies of common disease, effect size of consistently replicated gene–disease associations were found to be small (OR=1.2–1.5);[15] therefore, meta-analysis of GWA data sets is the most important approach to increase the power to detect such gene–disease associations.[35]

Our simulation shows that the power of REM meta-analysis of GWA data sets to detect a small genetic effect could decrease due to between-study heterogeneity and then the mean OR of the simulated meta-analyses that passing the genome-wide significance threshold would be upwardly biased. Recently, Moonesinghe *et al.*[76] show that the required sample size in meta-analysis to detect an overall association with adequate power at a significant level increases as between-study heterogeneity increases and when the between-study heterogeneity exceeds a threshold, meta-analysis cannot reach the power regardless of how large included studies are. At the same time, empirical evaluation of published meta-analyses[61] and our simulation study show the uncertainty of estimated between-study heterogeneity is large unless many studies are combined.

These findings suggest that when a meta-analysis of GWA data sets shows association signals reaching genome-wide significance with small between-study heterogeneity, the result should be cautiously reported and further replication studies by institutions other than GWA teams are required.[35] Moreover, when a large number of data sets are available, challenges to explain and reduce the observed between-study heterogeneity may become important.[74,76] The knowledge about the potential causes of between-study heterogeneity may help. Such post-GWA research will enable us to map the causative variant finely[79] or to detect polymorphisms associated with clinically important subtypes of diseases.[80]

1 Lander, E. S. The new genomics: global views of biology. *Science* **274,** 536–539 (1996).

2 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).

3 The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).

4 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).

5 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

6 Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

7 Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).

8 Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4,** 45–61 (2002).

9 Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33,** 177–182 (2003).

10 Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat. Genet.* **29,** 306–309 (2001).

11 Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nat. Rev. Genet.* **2,** 91–99 (2001).

12 Freely associating. *Nat. Genet.* **22,** 1–2 (1999).

13 Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* **361,** 865–872 (2003).

14 Ioannidis, J. P. Non-replication and inconsistency in the genome-wide association setting. *Hum. Hered.* **64,** 203–213 (2007).

15 Khoury, M. J., Little, J., Gwinn, M. & Ioannidis, J. P. On the synthesis and interpretation of consistent but weak gene–disease associations in the era of genome-wide association studies. *Int. J. Epidemiol.* **36,** 439–445 (2007).

16 NCI-NHGRI Working Group on Replication in Association StudiesChanock, S. J. Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J. *et al.* Replicating genotype–phenotype associations. *Nature* **447,** 655–660 (2007).

17 Elbaz, A., Nelson, L. M., Payami, H., Ioannidis, J. P., Fiske, B. K., Annesi, G. *et al.* Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: a large-scale international study. *Lancet Neurol.* **5,** 917–923 (2006).

18 Munafo, M. R. & Flint, J. Meta-analysis of genetic association studies. *Trends Genet.* **20,** 439–444 (2004).

19 Lau, J., Ioannidis, J. P. & Schmid, C. H. Summing up evidence: one answer is not always enough. *Lancet* **351,** 123–127 (1998).

20 Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* **2,** e841 (2007).

21 Sagoo, G. S., Little, J. & Higgins, J. P. Systematic reviews of genetic association studies. Human Genome Epidemiology Network. *PLoS Med.* **6,** e28 (2009).

22 Egger, M. & Smith, G. D. Bias in location and selection of studies. *BMJ* **316,** 61–66 (1998).

23 Lin, B. K., Clyne, M., Walsh, M., Gomez, O., Yu, W., Gwinn, M. *et al.* Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. *Am. J. Epidemiol.* **164,** 1–4 (2006).

24 Tang, J. L. Selection bias in meta-analyses of gene–disease associations. *PLoS Med.* **2,** e409 (2005).

25 Kavvoura, F. K. & Ioannidis, J. P. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum. Genet.* **123,** 1–14 (2008).

26 Attia, J., Thakkinstian, A. & D'Este, C. Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. *J. Clin. Epidemiol.* **56,** 297–303 (2003).

27 Begg, C. B. & Mazumdar, M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50,** 1088–1101 (1994).

28 Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315,** 629–634 (1997).

29 Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F. & Chalmers, T. C. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N. Engl. J. Med.* **327,** 248–254 (1992).

30 Ioannidis, J. P., Contopoulos-Ioannidis, D. G. & Lau, J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. *J. Clin. Epidemiol.* **52,** 281–291 (1999).

31 Ioannidis, J. & Lau, J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proc. Natl Acad. Sci. USA* **98,** 831–836 (2001).

32 McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9,** 356–369 (2008).

33 Seminara, D., Khoury, M. J., O'Brien, T. R., Manolio, T., Gwinn, M. L., Little, J. *et al.* The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* **18,** 1–8 (2007).

34 Ioannidis, J. P., Bernstein, J., Boffetta, P., Danesh, J., Dolan, S., Hartge, P. *et al.* A network of investigator networks in human genome epidemiology. *Am. J. Epidemiol.* **162,** 302–304 (2005).

35 Zeggini, E. & Ioannidis, J. P. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10,** 191–201 (2009).

36 Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).

37 Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* **447,** 661–678 (2007).

38 Evangelou, E., Maraganore, D. M. & Ioannidis, J. P. Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PLoS ONE* **2,** e196 (2007).

39 Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40,** 955–962 (2008).

40 Browning, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* **124,** 439–450 (2008).

41 Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A. *et al.* Detection of genotyping errors by Hardy–Weinberg equilibrium testing. *Eur. J. Hum. Genet.* **12,** 395–399 (2004).

42 Cox, D. G. & Kraft, P. Quantification of the power of Hardy–Weinberg equilibrium testing to detect genotyping error. *Hum. Hered.* **61,** 10–14 (2006).

43 Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76,** 967–986 (2005).

44 Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A. & Attia, J. How should we use information about HWE in the meta-analyses of genetic association studies? *Int. J. Epidemiol.* **37,** 136–146 (2008).

45 Zintzaras, E. & Lau, J. Synthesis of genetic association studies for pertinent gene–disease associations requires appropriate methodological and statistical approaches. *J. Clin. Epidemiol.* **61,** 634–645 (2008).

46 Thakkinstian, A., McElduff, P., D'Este, C., Duffy, D. & Attia, J. A method for meta-analysis of molecular association studies. *Stat. Med.* **24,** 1291–1306 (2005).

47 Salanti, G., Sanderson, S. & Higgins, J. P. Obstacles and opportunities in meta-analysis of genetic association studies. *Genet. Med.* **7,** 13–20 (2005).

48 Lindley, D. Statistical inference concerning Hardy–Weinberg equilibrium. *Bayesian Stat.* **3,** 307–326 (1988).

49 Weir, B. S. in *Genetic Data Analysis II: Methods for Discrete Population Genetic Data* (Sinauer Associates, Sunderland, 1996).

50 Hernandez, J. L. & Weir, B. S. A disequilibrium coefficient approach to Hardy–Weinberg testing. *Biometrics* **45,** 53–70 (1989).

51 Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A. & Attia, J. The choice of a genetic model in the meta-analysis of molecular association studies. *Int. J. Epidemiol.* **34,** 1319–1328 (2005).

52 Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22,** 719–748 (1959).

53 Yusuf, S., Peto, R., Lewis, J., Collins, R. & Sleight, P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog. Cardiovasc. Dis.* **27,** 335–371 (1985).

54 Cochran, W. G. The combination of estimates from different experiments. *Biometrics* **10,** 101–129 (1954).

55 Hardy, R. J. & Thompson, S. G. Detecting and describing heterogeneity in meta-analysis. *Stat. Med.* **17,** 841–856 (1998).

56 Petitti, D. B. Approaches to heterogeneity in meta-analysis. *Stat. Med.* **20,** 3625–3633 (2001).

57 DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control Clin. Trials* **7,** 177–188 (1986).

58 Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21,** 1539–1558 (2002).

59 Mittlbock, M. & Heinzl, H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat. Med.* **25,** 4321–4333 (2006).

60 Higgins, J. P., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327,** 557–560 (2003).

61 Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* **335,** 914–916 (2007).

62 Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361,** 598–604 (2003).

63 Wacholder, S., Rothman, N. & Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J. Natl. Cancer Inst.* **92,** 1151–1158 (2000).

64 Wacholder, S., Rothman, N. & Caporaso, N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol. Biomarkers Prev.* **11,** 513–520 (2002).

65 Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36,** 512–517 (2004).

66 Thomas, D. C. & Witte, J. S. Point: population stratification: a problem for case–control studies of candidate-gene associations? *Cancer Epidemiol. Biomarkers Prev.* **11,** 505–512 (2002).

67 Ioannidis, J. P., Ntzani, E. E. & Trikalinos, T. A. 'Racial' differences in genetic effects for complex diseases. *Nat. Genet.* **36,** 1312–1318 (2004).

68 Garner, C. Upward bias in odds ratio estimates from genome-wide association studies. *Genet. Epidemiol.* **31,** 288–295 (2007).

69 Zollner, S. & Pritchard, J. K. Overcoming the winner's curse: estimating penetrance parameters from case–control data. *Am. J. Hum. Genet.* **80,** 605–615 (2007).

70 Ghosh, A., Zou, F. & Wright, F. A. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am. J. Hum. Genet.* **82,** 1064–1074 (2008).

71 Ioannidis, J. P. Why most discovered true associations are inflated. *Epidemiology* **19,** 640–648 (2008).

72 Kraft, P. Curses—winner's and otherwise—in genetic epidemiology. *Epidemiology* **19,** 649–651 (2008); discussion 657–658.

73 Yu, K., Chatterjee, N., Wheeler, W., Li, Q., Wang, S., Rothman, N. *et al.* Flexible design for following up positive findings. *Am. J. Hum. Genet.* **81,** 540–551 (2007).

74 Ioannidis, J. P., Thomas, G. & Daly, M. J. Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10,** 318–329 (2009).

75 Zondervan, K. T. & Cardon, L. R. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5,** 89–100 (2004).

76 Moonesinghe, R., Khoury, M. J., Liu, T. & Ioannidis, J. P. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc. Natl Acad. Sci. USA* **105,** 617–622 (2008).

77 Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).

78 Hedges, L. V. & Pigott, T. D. The power of statistical tests in meta-analysis. *Psychol. Methods* **6,** 203–217 (2001).

79 Helgason, A., Palsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S. *et al.* Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39,** 218–225 (2007).

80 Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D. A. *et al.* Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet.* **4,** e1000054 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)