European Journal of Human Genetics (2012) 20, 1174–1181
© 2012 Macmillan Publishers Limited All rights reserved 1018-4813/12
www.nature.com/ejhg

ARTICLE

Meta-analysis of genetic association studies under heterogeneity

Binod Neupane[1], Mark Loeb[1,2], Sonia S Anand[1] and Joseph Beyene*,[1]

In multi-cohort genetic association studies or meta-analysis, associations of genetic variants with complex traits across cohorts may be heterogeneous because of genuine genetic diversity or differential biases or errors. To detect the associations of genes with heterogeneous associations across cohorts, new global fixed-effect (FE) and random-effects (RE) meta-analytic methods have been recently proposed. These global methods had improved power over both traditional FE and RE methods under heterogeneity in limited simulation scenarios and data application, but their usefulness in a wide range of practical situations is not clear. We assessed the performance of these methods for both binary and quantitative traits in extensive simulations and applied them to a multi-cohort association study. We found that these new approaches have higher power to detect mostly the very small to small associations of common genetic variants when associations are highly heterogeneous across cohorts. They worked well when both the underlying and assumed genetic models are either multiplicative or dominant. But, they offered no clear advantage for less common variants unless heterogeneity was substantial. In conclusion, these new meta-analytic methods can be used to detect the association of genetic variants with high heterogeneity, which can then be subjected to further exploration, in multi-cohort association studies and meta-analyses.

European Journal of Human Genetics (2012) **20,** 1174–1181; doi:10.1038/ejhg.2012.75; published online 30 May 2012

Keywords: genome-wide and genetic association studies; single-nucleotide polymorphism; meta-analysis; study heterogeneity; statistical power; type I error rates

INTRODUCTION

Genetic associations of single-nucleotide polymorphisms (SNPs) that were identified by genome-wide association studies (GWAS) and successive replication efforts or meta-analyses as having robust associations with most complex diseases are of relatively small to modest magnitudes (odds ratios (ORs) <1.50).[1–3] Genetic association studies typically require a very large sample size for the desired power to detect associations of such magnitude, as a stringent significance level (usually $\alpha=5\times10^{-8}$ for genome-wide studies) is generally applied in order to minimize detection of false associations. To attain the required sample size, large-scale multi-team collaborative studies with participants recruited from distinct populations defined by country of origin, regional ancestry, ethnicity, or study center, or meta-analyses of individual studies are necessary.[4] Meta-analyses of genome-wide and/or replication studies have been successful in identifying novel genetic variants for complex diseases not previously identified by single studies.[5–8]

One important challenge that remains, however, is that multi-team collaborative studies or meta-analyses from distinct populations, hereafter called *cohorts*, are more likely to demonstrate inconsistent estimates of SNP associations across cohorts because of genuine diversity in genetic associations, or differential errors or biases across cohorts.[9–12] Between-cohort heterogeneity may result from the associations that truly exist in one, some or all cohorts with different magnitudes (eg, due to local gene-environment interactions, which might be further exaggerated by sampling variation), or which could be a false signal due to methodological errors and biases (eg, because of different linkage disequilibrium (LD) patterns of tagged markers with causal variants across cohorts, the phenotype of interest being correlated with other phenotype with which the SNP is correlated, population stratification, different study designs with differential ascertainment of phenotype across cohorts, differential genotyping errors) or merely by chance.[10,11] Therefore, heterogeneity could be a signal rather than a noise in genetic association studies. Even if the associations are modestly or highly heterogeneous across cohorts, association in some or all cohorts may be genuine and are of interest.

Traditionally, in meta-analyses of clinical trials and epidemiological studies, the fixed-effect (FE) approach[13] has been used when cohort-specific associations are more or less similar and the random-effects (RE) approach[14] has been used when heterogeneity is suspected, to test whether there exists an average effect of a treatment or an exposure. In genetic association studies, as heterogeneity may result for any reason, detecting an 'association' if it truly exists even in a single cohort, however, is of prime interest rather than detecting a non-null 'average effect' over cohorts.[15,16] Unfortunately, as between-cohort heterogeneity increases it needs even larger sample size to detect associations by using traditional meta-analytic approaches.[17] When heterogeneity is suspected, traditionally preferred approach, the random-effect method is less powerful in detecting a genuine association as it produces more conservative *P*-values than FE approach[18,19] and would be too conservative when a stringent significance level is used. Hence, even large multi-cohort studies

[1]Population Genomics Program, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; [2]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada
*Correspondence: Dr J Beyene, Department of Clinical Epidemiology and Biostatistics, Population Genomics Program, McMaster University, MDCL-3200, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5. Tel: +1 905 525 9140 Extn 21333; Fax: +1 905 528 2814; E-mail: beyene@mcmaster.ca
Received 1 November 2011; revised 8 March 2012; accepted 28 March 2012; published online 30 May 2012

or meta-analyses employing traditional approaches might fail to demonstrate associations for some genetic variants that do not have consistent associations across cohorts. So rather than increasing the sample size by including additional data from more cohorts or waiting until sufficient data are generated, it is more desirable to find statistical methods that have increased potential to detect associations in the presence of heterogeneity.

To overcome this limitation of traditional methods in detecting an association, Lebrec *et al*[15] recently proposed global methods as new sets of screening tools for associations in heterogeneous conditions for multi-cohort genetic association studies. The new FE global method aims to test if an association exists in at least one cohort and the new random-effect global method aims to test if the overall association or between-cohort variance of associations is non-zero. They argued that detecting a genuine association is important, so it's a matter of efficiency rather than principle in choosing which method to apply. In their simulation study, these global methods had higher power than both the traditional methods at nominal significance level when there was moderate to substantial between-cohort heterogeneity. More recently, Han and Eskin[16] compared the power of this new global RE method with traditional methods and found similar results in the presence of heterogeneity, suggesting that the new RE method can be used to discover genes with robust association in meta-analysis. However, Lebrec *et al*[15] reported results for a very simple scenario, and did not assess the comparative performance of these methods at more realistic scenarios or using real genomic data. Han and Eskin[16] did not assess the performance of the new FE method, which was shown to have higher power than the new RE method in the presence of high heterogeneity as seen in Lebrec *et al*'s[15] simulation. Earlier, Pereira *et al*[20] investigated the impacts of heterogeneity and genetic model mis-specification on power and other issues for traditional FE and RE methods in a simulation study. But to date, no comparative studies of both the newly proposed meta-analytic approaches have been carried out in a wide range of realistic scenarios. It is therefore not clear under which circumstances these new methods perform better or are of greater practical utility than traditional methods in screening for or discovering of genetic associations.

In this study, our objective was to assess the performance of new and traditional meta-analytic methods with respect to type I error and statistical power through extensive simulations in a wide range of realistic applications. For instance, our simulation included scenarios such as: (1) a genetic variant is less common, (2) only few cohorts are available, (3) failure to adjust for important prognostic factors, and (4) assume a wrong genetic model. To determine the practical utility of these global methods in real data application, we applied these methods to West Nile virus infection complications data from a multi-center association study.

MATERIALS AND METHODS

Hypotheses and tests

Here, we briefly describe the methods given by Lebrec and colleagues (see Lebrec *et al*[15] for detailed descriptions). In a multi-cohort study or meta-analysis with k distinct cohorts for a binary phenotype, Y ($y=0$ for control, $y=1$ for case in a case–control study), suppose the information on the number of copies of the minor allele in a genotype, X ($x=0$, 1, 2), at an autosomal biallelic SNP locus and a set of covariates, Z, are available. Let the SNP effect in the ith cohort be $\beta_i=\ln(\mathrm{OR}_i)$ and its SE be s_i ($i=1,2,\ldots,k$). Then the multiplicative (log-additive) genetic model of phenotype risk in the ith cohort is

$$\mathrm{logit}\{P(Y=1\mid X,Z)\}=\alpha_i+\beta_i X+\gamma_i Z,$$

where β_i's are the parameters of interest whereas α_i's and γ_i's are nuisance parameters. Similarly for the quantitative phenotype, Y, the additive genetic model in the ith cohort is

$$y_i=\alpha_i+\beta_i X+\gamma_i Z+e_i,$$

where $e_i \sim N(0,\sigma^2)$ for all x and i. Different hypotheses and corresponding test statistics are given below.

FE and RE methods in traditional meta-analysis

FE level method (the traditional FE method). Under the FE assumption, effects are assumed to be similar across cohorts and hypothesis is tested for the average effect across all cohorts as: $H_0: \beta_1=\beta_2=\ldots=\beta_k=\beta=0$ *vs* $H_1: \beta_1=\beta_2=\ldots=\beta_k=\beta\neq 0$. The overall effect β is typically estimated as weighted average of cohort-specific effects using inverse variance weights as $\hat\beta=\sum_{i=1}^{k} w_i\hat\beta_i/\sum_{i=1}^{k} w_i$ and variance as $\mathrm{var}(\hat\beta)=1/\sum_{i=1}^{k} w_i$; where weight $w_i=1/s_i^2$ for cohort i. Then corresponding test statistic under H_0 is $T=\left(\sum_{i=1}^{k} w_i\hat\beta\right)^2/\sum_{i=1}^{k} w_i \sim \chi_1^2$ (asymptotically).

RE level method (the traditional RE method). Under the RE assumption, consider that cohort-specific effects, β_i's, represent a random sample from a grand normal population with overall mean μ (for example, $\mu=\ln(\mathrm{OR})$ with OR being the overall OR across cohorts for the binary trait, or the average mean difference across cohorts for the quantitative trait per one copy increase in number of the minor allele in a genotype under the multiplicative genetic model) and between-cohort variance τ^2; that is, $\beta_1,\beta_2,\ldots,\beta_k \sim N(\mu,\tau^2)$. Here, τ^2 represents the extent of heterogeneity in effects across cohorts. The overall effect is estimated as $\hat\mu=\sum_{i=1}^{k} w_i\hat\beta_i/\sum_{i=1}^{k} w_i$ and variance as $\mathrm{var}(\hat\mu)=1/\sum_{i=1}^{k} w_i$; where weight $w_i=1/(s_i^2+\hat\tau^2))$ for cohort i and $\hat\tau^2$ is the estimate of τ^2. Then the hypothesis is tested for overall effect across all cohorts as: H_0: $\mu=0$ *vs* H_1: $\mu\neq 0$. The Wald test, $T=\hat\mu^2/\mathrm{var}(\hat\mu) \sim \chi_1^2$ under H_0, is the standard test of the hypothesis.

New FE and RE global methods for multi-cohort association studies

FE global method (new FE method). Under the FE assumption, Lebrec *et al* proposed to test whether an association is present in any cohort: $H_0: \beta_1=\beta_2=\ldots=\beta_k=0$ *vs* $H_1: \beta_i\neq 0$ in at least one cohort i. The score test, $T=\sum_{i=1}^{k}(\hat\beta_i/s_i)^2 \sim \chi_k^2$ (asymptotically) under H_0, can be used to test the hypothesis.

RE global method (new RE method). The new RE model tests whether a non-null average association exists, or the between-cohort variance is non-zero (ie, a significant between-cohort heterogeneity is present), that is, H_0: $\mu=0$ and $\tau^2=0$ *vs* H_1: $\mu\neq 0$ or $\tau^2>0$. The likelihood ratio test, $T=2[l(\hat\mu,\hat\tau^2)-l(0,0)] \sim (\chi_1^2+\chi_2^2)/2$ (asymptotically) under H_0, can be used to test this hypothesis. Here, for large cohorts, $\hat\beta_i\mid\beta_i \sim N(\beta_i,s_i^2)$, but as $\beta_i \sim N(\mu,\tau^2)$, the approximate marginal distribution of the estimate in the ith cohort is $\hat\beta_i \sim N(\mu,s_i^2+\tau^2)$ with the corresponding log-likelihood $l_i(\mu,\tau^2)$ and the total log-likelihood $l(\mu,\tau^2)=\sum_{i=1}^{k} l_i(\mu,\tau^2)$.

Tests of heterogeneity

Cochran's Q statistic for test of heterogeneity[13] is obtained as $Q=\sum_{i=1}^{k} w_i(\hat\beta_i-\hat\beta)^2 \sim \chi_{k-1}^2$, where, $w_i=1/s_i^2$ ($i=1,2,\ldots,k$). The estimate of between-cohort variance can be obtained by the method of moment as $\tau^2=[Q-(k-1)]/\left[\sum_{i=1}^{k} w_i-\sum_{i=1}^{k} w_i^2/\sum_{i=1}^{k} w_i\right]$, which is 0 when $Q<k-1$. τ^2 can also be estimated by maximizing the profile likelihood, $\mathrm{pl}(\tau^2)=l(\hat\mu(\tau^2),\tau^2)$ by the method of maximum likelihood or restricted

maximum likelihood. I^2, an estimate of the degree of between-cohort heterogeneity due to factors other than chance is obtained as,[18,19] $I^2 = [(Q-(k-1))/Q] \times 100\ \%$, which is 0 when $Q < k-1$. In meta-analysis of clinical trials and epidemiological studies, heterogeneity is suspected if *P*-value <0.10 in Cochran's *Q*-test. Also, $25 \leq I^2 < 50$ and $I^2 \geq 50\%$ are considered evidences of modest and large heterogeneity, respectively.[11,19]

Simulation study

For a SNP with effect $\ln(\mathrm{OR}_i) = \beta_i$, minor allele frequency (MAF) $= f_i$, and proportion of cases $= \pi_{0i}$ and satisfying Hardy–Weinberg Equilibrium (HWE) in the *i*th cohort ($i = 1,2,\ldots,k$), we generated the case or control status for a subject with genotype x (0,1,2), using multiplicative genetic model:

$$P(case \mid x) = \exp(\alpha_i + \beta_i x)/(1 + \exp(\alpha_i + \beta_i x)),$$

where, $\alpha_i = ln(\pi_{0i}/(1-\pi_{0i})) - 2ln(1 - f_i + f_i exp(\beta_i))$. Furthermore, to assess the impact of genetic model (mis)specification, we generated data under the dominant, recessive and multiplicative genetic model assumption. For quantitative trait, we generated the population data from

$$y_i = \alpha_i + \beta_i x + e_i,$$

where we used $\alpha_i = 0.5$ and $e_i \sim N(0,\ 1)$ for all $x(0,1,2)$ and i $(1,2,\ldots,k)$. We simulated $\beta_1, \beta_2, \ldots, \beta_k$ from $N(\mu,\ \tau^2)$. We ran 10 000 simulations for each combination of (μ, τ) under a variety of realistic scenarios listed in Table 1. For instance, we considered the overall association, μ, from null ($\mu = 0$) through small to modest in sizes ($\mu = (0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.25,$ and $0.30)$) with corresponding ORs, $exp(\mu)$: (1.00, 1.05, 1.11, 1.16, 1.22, 1.28, and 1.35) and for such effect sizes we considered between-cohort SD, τ, ranging from none ($\tau = 0$) through low ($\tau = 0.1$), moderate ($\tau = 0.2$), and substantial ($\tau = 0.3$) heterogeneity for a binary trait. We analyzed the binary data using logistic regression assuming multiplicative genetic risk effect per genotype. In a separate analysis, each of the dominant, recessive, and multiplicative genetic models was assumed while analyzing each of the data sets generated under each of these models using logistic regression. Data generated for quantitative traits were analyzed using linear regression assuming additive genetic risk. We assessed both the type I error rate and statistical power of these tests at nominal significance level, $\alpha = 0.05$ as well as more stringent significance levels, $\alpha = 5.0 \times 10^{-6}$ and 5.0×10^{-8}.

Application to real data

We used the West Nile virus infection severity data set,[21] where SNPs were genotyped by Illumina HumanNS-12 BeadChip, and subjects were recruited from seven study centers (cohorts) from Canada and the United States. We restricted the analysis to Caucasian population of Northern and Western European origin. Using PLINK: Whole genome data analysis tool set (http://pngu.mgh.harvard.edu/~purcell/plink/), we first applied standard quality control (QC) inclusion criteria: MAF $\geq 5\%$, genotyping error rate per SNP $< 5\%$, *P*-value for HWE exact test in control group $> 10^{-4}$, genotyping error rate per subject $<5\%$ for considering the SNPs or the subjects for analysis. Further, a SNP passing these criteria must have had MAF $\geq 1\%$ and HWE $P > 10^{-5}$ in an individual center for that particular center to be included in the meta-analysis for that SNP. Cryptic related subjects or those for which reported sex did not match to that in DNA sample were also discarded.[21] Then for each of the remaining SNPs, we obtained the estimates of β_i and its SE, s_i, in center i ($i = 1,2,\ldots,7$) using logistic regression assuming a multiplicative genetic risk model. We also re-estimated s_i applying genomic control to correct the center-specific *P*-values for any residual confounding due to population substructure. Then we applied all four meta-analytic methods to the center-specific aggregate data in R (http://www.r-project.org/) and compared their power based on association *P*-values for the respective I^2, heterogeneity *P*-value and τ estimated from the data. The significance level was adjusted for the multiple testing problem using the Bonferroni adjustment.

RESULTS

Simulation results

Type I error. Type I error (ie, when both $\mu = 0$ and $\tau = 0$) rates for all four tests at $\alpha = 0.05$ are presented in Supplementary Table 1. More data on the error rates can be found in Supplementary Table 2 at

Table 1 Parameters setting for different simulation scenarios

Parameters	*Assigned or assumed values/scenarios*
Number of simulations	10 000
Locus type	Biallelic; genotype X ($x = 0,\ 1,\ 2$) satisfying HWE criterion in each cohort
Binary and quantitative traits	Note: 10 000 simulations for each combination of parameters below (eg, 1400^+ scenarios for binary traits)
Total sample size, N	2000, 4000, 6000, 8000, 10 000
Number of cohorts (eg, studies), k	2, 3, 5, 7, 10 ($i = 1,2,\ldots,k$)
Sizes of the *i*th cohort, N_i	Average $N_i = N/k$, variable ($N_i \sim$ uniform(N/k–$N/2k$, $N/k + N/2k$) and adjusted for the total size so that $\sum_i N_i = N$
Minor allele frequency, f	Average $f = (0.05,\ 0.20)$, variable ($f_i \sim N$(mean $= f$, SD $= f/5$) from ($f - f/3$, $f + f/3$) and adjusted so that $\sum_i f_i/k = f$
Case proportion in each cohort, π_0	Average $\pi_0 = 0.50$, variable ($\pi_{0\mathrm{i}} \sim N$(mean $= \pi_0$, SD $= \pi_0/5$) from ($\pi_0 - \pi_0/3$, $\pi_{0\,+}\pi_0/3$) and adjusted so that $\sum_i \pi_{0i}/k = \pi_0$
Genetic model for data generation	Multiplicative for binary trait and additive for quantitative trait for each cohort
Genetic model for data analysis	Multiplicative for binary trait and additive for quantitative trait for each cohort
Average effect, μ	$\mu = (0,\ 0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.25,\ 0.30)$ with corresponding ORs, e^{μ}: (1.00, 1.05, 1.11, 1.16, 1.22, 1.28, 1.35)
Between-cohort SD, τ	$\tau = 0$ (no heterogeneity), 0.1 (low heterogeneity), 0.2 (moderate heterogeneity), 0.3 (substantial heterogeneity)
Cohort-specific effects, β_i	$\beta_1, \beta_2, \ldots, \beta_k$ are simulated from iid $N(\mu,\ \tau^2)$ for each combination of $(\mu,\ \tau)$ in each simulation
Data analysis	Logistic regression for binary trait, linear regression for quantitative trait in each cohort
Binary trait only (see Supplementary Table 1 for parameters eg, cohort sizes, proportion of cases, etc)	For each combination of $N = (2000,\ 4000)$, $k = (2,\ 5,\ 10)$ with $N_i = N/k$ (equal), MAF $= (0.05,\ 0.20)$ (equal), μ and τ (same as above for binary trait), we generated data with multiplicative underlying and assumed risk on genotype x. Additionally, for $N = 2000$ and $k = 5$, cohort-wise parameters were generated as $N_i \sim$ uniform(N/k–$N/3k$, $N/k + 3N/k$); $\pi_{0i} \sim$ uniform($0.33 \pm 0.33/2$); $f_i \sim$ uniform(0.20 ± 0.10) and then N_i's, f_i's, and π_{0i}'s were adjusted so that $\sum_i N_i = N$, $\sum_i f_i/k = f$ and $\sum_i \pi_{0i}/k = \pi_0$. N_i, π_{0i}, and f_i were all kept the same throughout the simulations (unlike above where they were all variable in each simulation for each combination of parameters). A binary independent covariate with 33% prevalence and OR $= 1.5$ was used for data generation, but was not adjusted in logistic regression
For assessment of impact of genetic model (mis)specification in data analysis (binary trait only)	For each combination of $N = 6000$; $k = 3$; $N_i = 2000$ (equal); $f_i = 0.20$ (equal); $\pi_{0i} = 0.50$ (equal); μ and τ (same as above for binary trait), we generated data under dominant ($x_{\mathrm{D}} = 0$ for $x = 0$ and 1 for $x = \{1,\ 2\}$), recessive ($x_{\mathrm{R}} = 0$ for $x = \{0,\ 1\}$ and 1 for $x = 2$), and multiplicative ($x_{\mathrm{M}} = 0,\ 1,\ 2$ for $x = 0,\ 1,\ 2$) risk, and each data set was analyzed assuming each of the three models using logistic regression

$\mu = 0$ and $\tau = 0$. The RE global (new RE) method resulted in the smallest type I error rates maintaining nominal significance level in all simulation scenarios. Other methods slightly exceeded nominal level in few simulation scenarios. At the more stringent $\alpha = 5.0 \times 10^{-6}$, no methods produced any significant associations.

Statistical power. The statistical power of the four methods in different scenarios at $\alpha = 0.05$, 5.0×10^{-6}, and 5.0×10^{-8} are presented in Supplementary Table 2 for both binary and quantitative traits. Some of the power comparisons are presented in Figures 1–3 and Supplementary Figures 1 to 8.

Power for a binary trait: At no heterogeneity ($\tau = 0$): both the traditional FE and RE methods had very similar power and higher than that of the global methods in all scenarios.

At low heterogeneity ($\tau = 0.1$): at the nominal significance level $\alpha = 0.05$ for a common variant (MAF ≈ 0.20), the traditional FE method was the most powerful in almost all scenarios for detecting modest associations (OR ≥ 1.20), followed by the traditional RE method. The new FE method performed as well or slightly better when there were fewer but larger cohorts in large studies (eg, when number of cohorts, $k = 3$ for the total sample size, $N = 8000$, or $k = 5$ for $N = 10\,000$) especially for smaller overall associations (OR < 1.20). But at more stringent significance levels, $\alpha = 5.0 \times 10^{-6}$ or smaller, there is no power advantage for global methods. For a less common variant (MAF ≈ 0.05), the new methods did not perform better even when $k = 2$ for $N = 10\,000$ at $\alpha = 0.05$.

At moderate heterogeneity ($\tau = 0.2$): for a common variant (MAF ≈ 0.20) at $\alpha = 0.05$, the new FE method had the highest power when fewer but larger cohorts were included ($k \leq 5$) while the new RE method had the better power when many smaller cohorts ($k \geq 7$) were available for the small or modest available total sample size ($N = 2000 \sim 4000$) (Figure 1 and Supplementary Figure 7). New FE almost always had better power when the overall associations were very small (OR < 1.20) (Figure 1 and Supplementary Figures 1 and 7). At $\alpha \leq 5.0 \times 10^{-6}$, for the given sample size each of the methods had some gain in power for fewer cohorts with larger sizes for $k \leq 7$, but such advantage tended to diminish or even altered for $N \geq 8000$ for larger k (Figure 2 and Supplementary Figure 3). For $N \leq 4000$, the new RE method performed better or similar to traditional FE but better than new FE for $k \geq 7$, while new FE performed the best for larger cohort sizes ($k < 7$). For $N \geq 6000$ the new methods generally performed better (where the new FE method had the highest power for $k \leq 5$ while the new RE had the highest power for larger $k \geq 7$) (Figure 2 and Supplementary Figures 5 and 7).

For a less common variant (MAF ≈ 0.05): even at $\alpha = 0.05$, all methods had considerably low power, and the gain in power for the new methods were not as prominent as that observed for more common variants (Supplementary Figures 2 and 7). For example, the new RE method did not perform better than traditional FE method and the advantage of the new FE method was not clear either when $N = 4000$ even when $k \leq 5$ (Supplementary Figure 2). For $\alpha = 5.0 \times 10^{-8}$, the power of all methods was very low for $N \leq 6000$. For $N \leq 8000$ with $k \geq 7$, traditional FE had the highest power where new

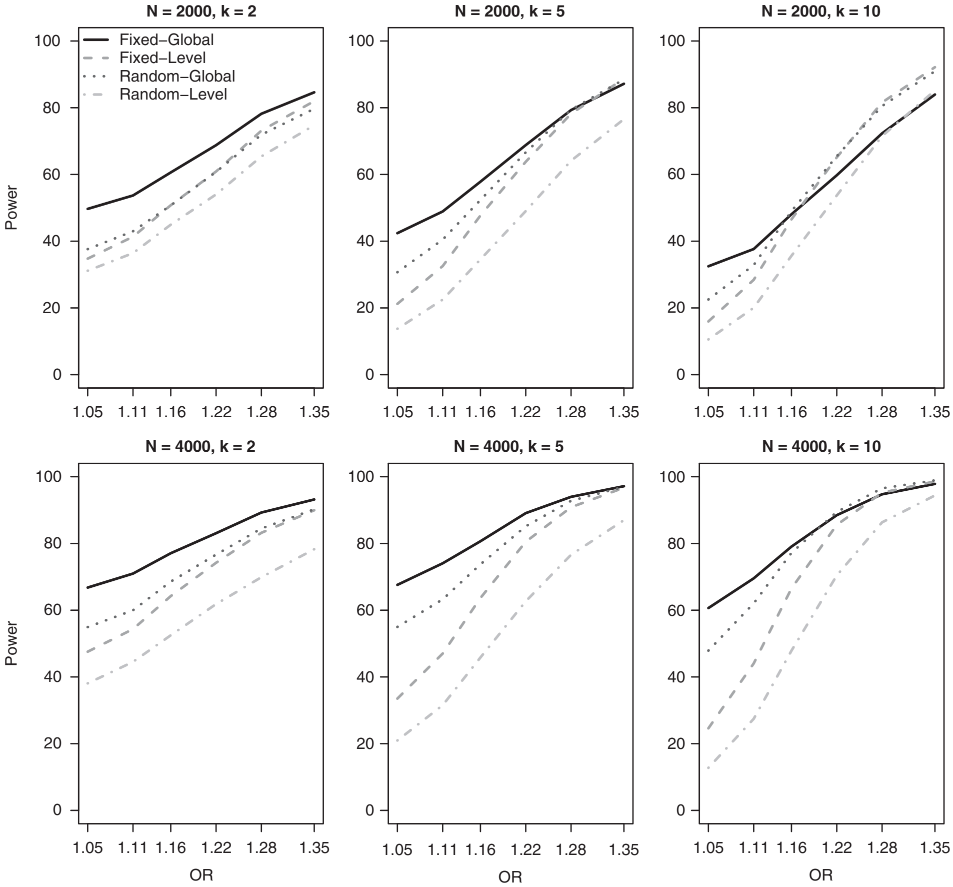

Figure 1 Power of four meta-analytic methods at $\tau = 0.2$ and $\alpha = 0.05$. Simulation scenario: binary trait; equal N_i (N/k), equal MAF (0.20), equal case–control ratio (1:1). N, total sample size; k, number of cohorts; N_i, cohort size; τ, between-cohort SD; OR, odds ratio; MAF, minor allele frequency.

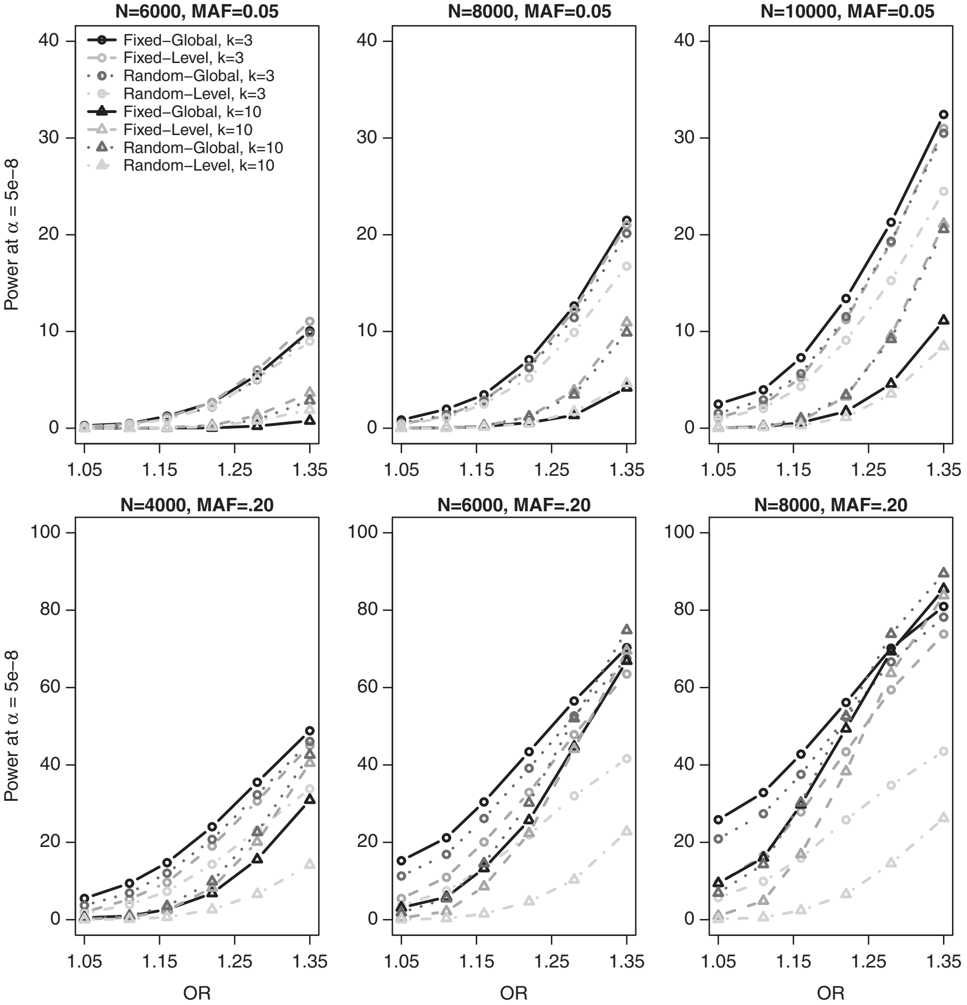

Figure 2 Power comparison at different total sample size and number of cohorts, and minor allele frequencies at $\tau = 0.2$ and $\alpha = 5 \times 10^{-8}$. Simulation scenario: binary trait; average $N_i = N/k$ (variable); MAF (variable), average case–control ratio = 1:1 (variable). *k*, number of cohorts; *N*, total sample size; N_i, cohort size; τ, between-cohort SD; OR, odds ratio; MAF, minor allele frequency.

RE performed better than new FE. New FE had similar or slightly better power than traditional FE for $N \geq 8000$ with $k \leq 3$ (Figure 2).

At substantial heterogeneity ($\tau = 0.3$): at $\alpha = 0.05$, the new FE generally outperformed all other methods for common or less common variant for any number of available cohorts (Supplementary Figures 2 and 3).

At $\alpha = 5.0 \times 10^{-8}$ for a common variant, the new FE method outperformed for any *k* for $N \geq 4000$ (Supplementary Figure 4), whereas new RE method performed similarly or better than new FE method for many cohorts of small sizes ($N = 2000$ with $k \geq 7$) in which situation traditional FE performs even better. For a less common variant, power was too low for $N \leq 4000$ for all methods to make any meaningful comparison (Supplementary Figure 4). For $N \geq 6000$, new RE generally outperformed when $k \geq 7$ while new FE outperformed when $k \leq 5$.

Power for a quantitative trait: At no or low heterogeneity, traditional methods generally performed better than the new methods. But under higher heterogeneity, the new methods in general performed quite well for quantitative traits (Supplementary Figures 4, 5 and 6). For example, at $\tau = 0.2$, quantitative trait analysis was more powerful than binary trait analysis, whereas the new global methods had higher power even for a *less common variant* even at a more stringent significance level and had considerable power advantage over traditional methods for a common variant. Even at $\tau = 0.1$ and $\alpha = 5.0 \times 10^{-8}$ even for $N = 2000$, the new FE method had similar or higher power than traditional methods when $k = 2$ and and the new RE method outperformed when $k = 10$ for a common variant. New global methods performed better even in presence of little heterogeneity for larger total sample sizes and almost always outperformed when the heterogeneity was substantial.

Similar comparative results were observed for the power of these tests irrespective of whether the minor allele frequencies, the proportions of cases, and cohort sizes were similar or varied across cohorts, and if an important independent prognostic variable was not adjusted for in the analysis. The traditional FE outperformed traditional RE in all conditions. In general, when heterogeneity increased, the power of traditional meta-analytic approaches generally decreased while that of the new global approaches increased at $\alpha = 0.05$ (Supplementary Figure 7). Interestingly, the power of even the traditional methods, and in particular the FE method increased as heterogeneity increased for $\alpha \leq 5.0 \times 10^{-6}$ in situations where power is expected to be generally small (eg, when overall association was very small ($\text{OR} \leq 1.20$), or the total sample size was small ($N \leq 4000$ for common variant and $N \leq 8000$ for less common variant) (Supplementary Figure 7).

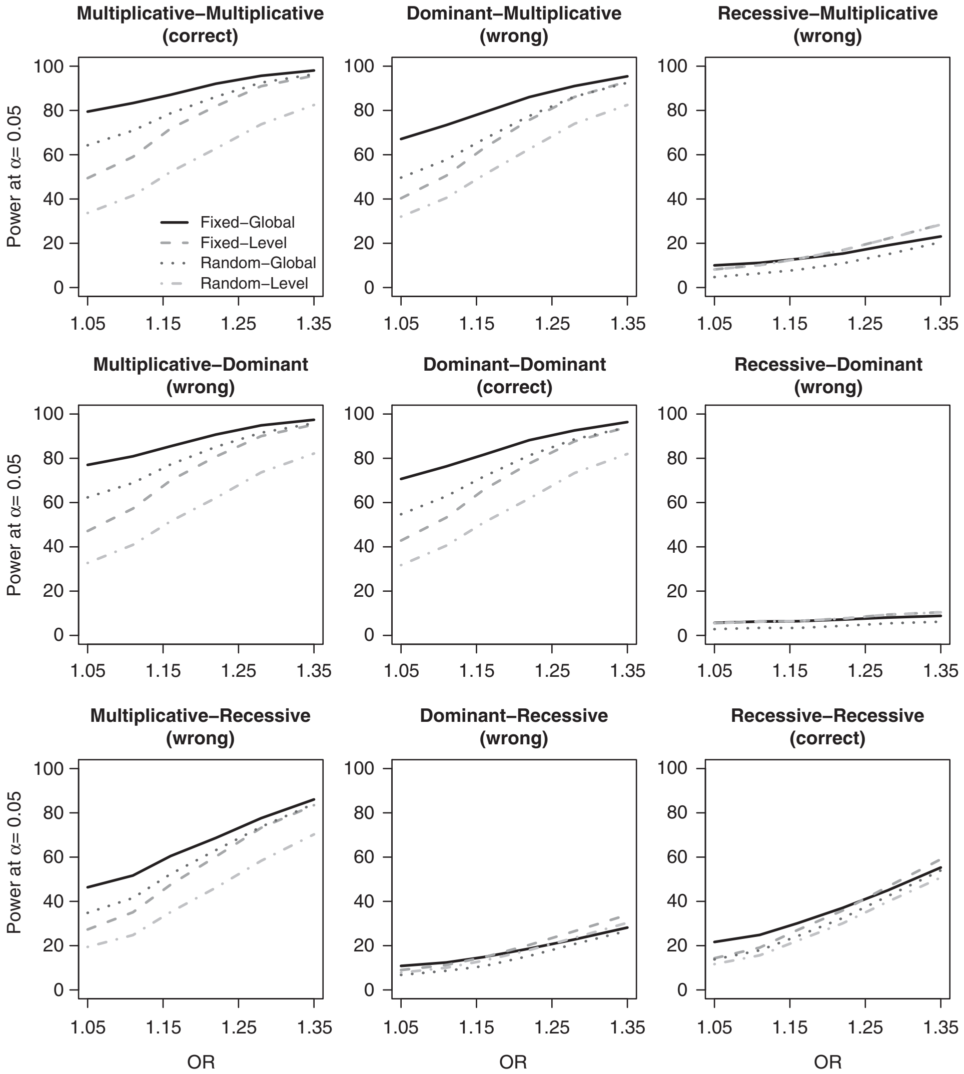

Figure 3 Power of four meta-analytic methods under each underlying and assumed genetic model at $\tau = 0.2$ and $\alpha = 0.05$. Simulation scenario: Binary trait, $N = 6000$, $k = 3$, $N_i = 2000$ (equal), MAF = 0.20 (equal), case–control ratio = 1:1 (equal). k, number of cohorts; N, total sample size; N_i, cohort size; τ, between-cohort SD; OR, odds ratio; MAF, minor allele frequency.

Impact of genetic model (mis)specification on power: At both $\alpha = 0.05$ and 5.0×10^{-6}, the new methods had better power than the traditional methods in the presence of moderate or substantial heterogeneity no matter whether correct or wrong multiplicative or dominant risk model was assumed when the underlying model was one of them (Figure 3 and Supplementary Figure 8). However, when the underlying or assumed model was recessive, all of these tests had considerably low power and the power was zero or almost zero at more stringent significance level where new methods (particularly new RE method) had the least power. However, the sample size ($N = 6000$, $k = 3$) was not sufficient to make any meaningful comparison under recessive risk model since only about 4% had the risk genotype for MAF = 0.20.

Application to real data

In the West Nile virus infection severity data set,[21] 13 371 SNPs were genotyped in 1346 participants recruited from seven study centers (cohorts) across Canada and the United States. There were 488 cases with neuroinvasive disease (meningitis, encephalitis, acute flacid paralysis) and 858 controls (infected but did not have severe complications). After applying QCs criteria and restricting analysis to White population, 9051 SNPs with 441 cases and 815 controls were left for analysis. Five cases were further discarded from one center as it had only cases with no controls for comparison. The Bonferroni-adjusted significance level was set to 5.52×10^{-6}. However, it should be noted that this threshold is too conservative for association analysis as SNPs are not independent. There was no population substructure within each center except in center 2 for which genomic inflation factor, $\lambda = 1.057$. Correction for population substructure did not significantly alter the meta-analysis results. About 3.8%, 8.6%, and 14.8% SNPs had $\tau > 0.30$, 0.20, and 0.10, respectively. The estimates of heterogeneity was larger than the sizes of respective overall associations ($\tau > \log(\mathrm{OR})$) for about 13.6% SNPs, which had some center-specific associations in reverse directions. About 8.2% SNPs had $\tau > \log(\mathrm{OR})$ with very small average OR ($1 \leq \mathrm{OR} < 1.10$ or $1 \leq 1/\mathrm{OR} < 1.10$). About 3.4% and 8.2% SNPs had heterogeneity *P*-values < 0.05 and < 0.10, respectively, in Cochran's *Q*-test. About 16% of the SNPs had modest ($25\% \leq I^2 < 50\%$) and 6% had substantial ($I^2 \geq 50\%$) heterogeneity. Estimates of associations and their association *P*-values, and extent of heterogeneity for these tests are

Table 2 Estimates of ORs and *P*-values for few SNPs from meta-analysis of West Nile virus data set

			Fixed effect		*Random effects*		*Heterogeneity*[c]		
SNP[a]	*Minor–Major Allele (MAF)*	*OR (95% CI)*[b]	*Global*	*Level*	*Global*	*Level*	τ	I^2	P
rs2066786	A-G(0.43)	0.63 (0.52, 0.78)	7.87×10^{-4}	1.30×10^{-5}	4.37×10^{-5}	1.30×10^{-5}	0.00005	0	0.5469
rs2298771	C-T(0.32)	1.52 (1.22, 1.88)	1.13×10^{-3}	1.39×10^{-4}	4.22×10^{-4}	1.39×10^{-4}	0.00019	34.7	0.1761
rs3738573	G-C(0.35)	0.70 (0.57, 0.86)	2.13×10^{-2}	9.38×10^{-4}	2.57×10^{-3}	9.38×10^{-4}	0.00003	0	0.5605
rs7118900	A-G(0.18)	1.12 (0.87, 1.44)	1.04×10^{-3}	0.38768	0.01232	0.20985	0.60950	76.9	0.00062
rs2960306	A-C(0.36)	1.05 (0.85, 1.29)	1.70×10^{-3}	0.65600	0.09607	0.82623	0.60749	76.2	0.00081
rs3795498	A-G(0.27)	1.13 (0.90, 1.41)	1.21×10^{-3}	0.28529	0.00290	0.12648	0.41083	76.0	0.00086

Abbreviations: CI, confidence interval; MAF, minor allele frequency; OR, odds ratio; SNP, single-nucleotide polymorphism.
Note: results presented here were without applying genomic control. Results were very similar after applying genomic control, where genomic inflation factor, $\lambda = 1$ for centers 1, 4, 6, 7; $\lambda = 1.057$ for center 2, and $\lambda = 1.007$ for center 5.
[a] SNPs inclusion criteria for analysis: MAF$\geq$0.05, Genotyping error per SNP$<$0.05, HWE *P*-value $<$0.0001 in all centers, and MAF$\geq$0.01 and HWE P $<$0.00001 in individual center.
[b] Genotypic OR and its 95% CI; results using traditional fixed- effect method are reported here.
[c] τ (between-cohort SD of log(OR)) were estimated using maximum likelihood approach and *P* was heterogeneity *P*-value from Cochran *Q*-test.

presented in Table 2 for three most significant SNPs as seen in the FE level test and another three of the most heterogeneous SNPs as suggested by Cochran *Q*-test for illustration purpose. In Supplementary Table 3, these three SNPs with the most heterogeneous effects were further explored within each cohort. In this analysis, none of the methods yielded any SNPs that remained significant at the Bonferroni-adjusted level. The traditional FE method produced the smallest *P*-values in the test of association for those SNPs having small heterogeneity (eg, rs2066789). For SNPs with large heterogeneity *P*-values, the new FE method produced the smallest association *P*-values, followed by the new RE method, both sets of which were much lower than those derived using the traditional FE and RE methods.

DISCUSSION

The new RE global test produced the smallest type I error rates at nominal significance level. No method produced significant associations at a more stringent significant level, $\alpha = 5.0 \times 10^{-6}$. As expected, the traditional RE (level) method that is proposed for heterogeneous conditions performed the worst at high heterogeneity. This method assesses the average effect without utilizing the heterogeneity information and is overly conservative for genetic association studies where associations could be heterogeneous for genuine reasons.[15,16] New global methods work well for common variants, even if a wrong multiplicative or dominant genetic model is assumed when the underlying risk model was one of them, at high heterogeneity. At high heterogeneity, when fewer but larger individual cohort are available for the given small to modest total sample size (2000$\sim$4000 subjects), the new FE method performs quite well, but it may fail if cohort sizes are very small.[15] When there are many cohorts with smaller sizes, another global random method may be a more powerful choice than traditional methods. However, these global methods offer no clear advantage even as screening tools for less common genetic variants even at substantial heterogeneity in a small to modest sized study.

One concern is that these global methods and in particular the global FE method have a clear advantage in power over traditional methods mostly when overall associations are small (OR$<$1.20) but are highly heterogeneous across cohorts (ie, when $\tau \geq \mu = \ln(\text{OR})$). Can we use these global methods for gene discovery in meta-analysis, where even the new RE method might achieve significance at genome-wide level with much higher power even when overall OR = 1.0 or 1.05 at high heterogeneity? Although some degree of heterogeneity is expected because of genuine reasons, credibility of the association is questionable if very high heterogeneity is observed with such a small overall association. An observed association is unlikely to be robust, not even in a single cohort, if the associations of large magnitudes in individual cohorts are flip-flopped in opposite directions suggesting both protective and harmful effects of the same mutant gene in distinct populations, which would result in moderate or substantial heterogeneity. Such association could be more likely a spurious finding as a result of some undetected methodological error (eg, because of genotyping error) or chance variation.[22,23] In such studies, if the individual cohorts are well designed or large, real biases (eg, population stratification) are unlikely to alter the genuine association in the reverse direction with large magnitudes.[11] If associations are highly heterogeneous across cohorts but most of the larger associations are in the same direction, the average association is also likely to be of larger magnitude and the traditional FE method can perform equally well for large effect sizes. Further, this method also has some increased power at more stringent significance levels to detect associations of small magnitudes as heterogeneity increases, although the gain in power is smaller compared with global methods. Thus, it may also better control false positives at extreme heterogeneity conditions caused by errors or chances rather than genuine factors.

Therefore, any perceived advantage of especially the new FE global method in high heterogeneity may not directly translate in to practice if the purpose is to achieve significance for discovery of genes with robust associations even in at least one cohort rather than just screening for the potential association in multi-cohort GWA studies or meta-analysis. For example, in our example data set, the SNPs explored in the Supplementary Table 3 had very high heterogeneity. They had similar MAFs with no genotyping errors or deviation from HWE across cohorts, whereas the study was conducted on subjects of genetically similar backgrounds and the same study protocol was followed across centers. Then what might have caused so much heterogeneity for these SNPs? Here, cohorts were defined based on geographic locations and might not be very distinct in terms of genetic and environmental factors. Furthermore, the total sample size and individual cohorts sizes were quite small and the case–control ratio was quite variable across centers (as the disease complications under study was quite uncommon in north America, it was a challenge to obtain a sufficient number of case and control subjects in each center during the time frame of the study). Therefore, the most plausible explanation for the substantial heterogeneity is the likely sampling variation. In practice, inclusion of such SNPs in analysis might just inflate the overall heterogeneity distributions and hence warrant tougher adjustment for population structure than is necessary for other SNPs that are more genuine candidates for analysis. However, we were too cautious to filter out such SNPs during the QC phase, because our purpose was to assess the utility of these newly proposed methods not only as tools to achieve significance at more stringent (adjusted or genome-wide) level to identify new genes associated with diseases, but also as screening tools to achieve significance in such level in the presence of heterogeneity so that they could be carried forward for further scrutiny. If there was a genuine small association of such a SNP in some of the cohorts because of, say, gene-local environment interaction while its association was reversed in some other cohorts by chance, then we would have missed an opportunity to test such a SNP had we filtered it out before analysis. In our example, for SNPs with very small overall associations with large heterogeneity, which could have been filtered out before analysis, the new FE method produced quite strong association *P*-values, whereas the new RE method suggested less impressive association and is less likely to lead to any unnecessary follow-up of such SNPs.

In recent years large-scale multi-cohort association studies have been carried out collaboratively for many complex diseases. For example, the INTERHEART Study[24] assessed the associations of different genetic variants with myocardial infarction risk factors in over 8000 individuals from five ethnic populations. Many SNPs may be expected to display modestly or highly heterogeneous associations for myriad reasons in such studies in genetically and environmentally distinct cohorts. Substantial heterogeneity is likely for some variants even in genetically close populations. For example, in a meta-analysis of three GWA studies of type 2 diabetes in the northern European population,[10] some SNP had an I^2 as high as 77%. Multi-cohort GWA studies or meta-analyses are, in practice, likely to be much bigger in size, and include large individual cohorts than the data set we used. Hence, any strong association in a single cohort might justify a further exploration as genotyping errors or chance might not be the only explanations for such large association in a cohort. In such studies, these new approaches may be useful in screening genetic variants to assess association in the presence of high heterogeneity and prioritize them for further scrutiny. Simple exploration across cohorts can identify methodological and chance errors or biases causing heterogeneity; and if heterogeneity is still unexplained, pathway-based analysis could provide better insights about the role of genes and environments causing the heterogeneity.[15] If this suggests the presence of some genuine associations in some cohorts, future replication or fine mapping studies in the cohort can resolve the issue. Then a genuine variant is more likely to be identified in the investigation process.

In considering the potential and pitfalls of these new global methods, there are some important questions that require further research and discussion: Are these new methods useful in practice in small to moderate sized genetic association studies to also validate or discover new genes in the presence of high heterogeneity? Do they work well in meta-analyses of independent research studies that might have employed different genotyping platforms or even recruited subjects with different ethnic backgrounds having different LD patterns, in which case an analyst might have to impute untagged markers? Also, for the genetic variants with heterogeneous associations across cohorts, the pathway-based three-point mixture model seems to be a promising tool to resolve the heterogeneity in specific cohorts.[15] Although the method might not be feasible for meta-analysis of GWA studies, the prospect of the method could be further explored for certain sets of SNPs that are known to belong to biologically defined pathways.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

JB would like to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR) (grant number 84392). JB holds the John D Cameron Endowed Chair in the Genetic Determinants of Chronic Diseases at McMaster University. We are grateful to two anonymous reviewers whose comments were very helpful and improved our manuscript. We would also like to thank Dr Russell de Souza for helpful comments.

1 Manolio TA, Brooks LD, Collins FS: A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008; **118**: 1590–1605.

2 Khoury MJ, Little J, Gwinn M, Ioannidis JP: On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *Int J Epidemiol* 2007; **36**: 439–445.

3 Wray NR, Goddard ME, Visscher PM: Prediction of individual genetic risk of complex disease. *Curr Opin Genet Dev* 2008; **18**: 257–263.

4 Seminara D, Khoury MJ, O'Brien TR *et al*: The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* 2007; **18**: 1–8.

5 Zeggini E, Scott LJ, Saxena R *et al*: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638–645.

6 Cooper JD, Smyth DJ, Smiles AM *et al*: Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 2008; **40**: 1399–1401.

7 Houlston RS, Webb E, Broderick P *et al*: Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 2008; **40**: 1426–1435.

8 Willer CJ, Sanna S, Jackson AU *et al*: Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genet* 2008; **40**: 161–169.

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

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

11 Ioannidis JP, Boffetta P, Little J *et al*: Assessment of cumulative evidence on genetic associations: interim guidelines. *Int J Epidemiol* 2008; **37**: 120–132.

12 Kavvoura FK, Ioannidis JPA: Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Human Genet* 2008; **123**: 1–14.

13 Cochran WG: The combination of estimates from different experiments. *Biometrics* 1954; **10**: 101–129.

14 DerSimonian R, Laird N: Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**: 177–188.

15 Lebrec JJ, Stijnen T, van Houwelingen HC: Dealing with heterogeneity between cohorts in genomewide SNP association studies. *Stat Appl Genet Mol Biol* 2010; **9**: 8.

16 Han B, Eskin E: Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 2011; **88**: 586–98.

17 Moonesinghe R, Khoury MJ, Liu T, Ioannidis JP: Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc Natl Acad Sci USA* 2008; **105**: 617–622.

18 Higgins JP, Thompson SG: Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**: 1539–58.

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

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

21 Loeb M, Eskandarian S, Rupp M *et al*: Genetic variants and susceptibility to neurological complications following west nile virus infection. *J Infect Dis* 2011; **204**: 1031–1037.

22 Lin PI, Vance JM, Pericak-Vance MA, Martin ER: No gene is an island: the flip-flop phenomenon. *Am J Hum Genet* 2007; **80**: 531–8.

23 Clarke GM, Cardon LR: Aspects of observing and claiming allele flips in association studies. *Genet Epidemiol* 2010; **34**: 266–74.

24 Anand SS, Xie C, Pare G *et al*: Genetic variants associated with myocardial infarction risk factors in over 8000 individuals from five ethnic groups: The INTERHEART Genetics Study. *Circ Cardiovasc Genet* 2009; **2**: 16–25.

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