Genetic Epidemiology 34: 702–715 (2010)

A Multipoint Method for Meta-Analysis of Genetic Association Studies

Pantelis G. Bagos* and Theodore D. Liakopoulos

Department of Computer Science and Biomedical Informatics, University of Central Greece, Lamia, Greece

Meta-analyses of genetic association studies are usually performed using a single polymorphism at a time, even though in many cases the individual studies report results from partially overlapping sets of polymorphisms. We present here a multipoint (or multilocus) method for multivariate meta-analysis of published population-based case-control association studies. The method is derived by extending the general method for multivariate meta-analysis and allows for multivariate modelling of log(odds ratios (OR)) derived from several polymorphisms that are in linkage disequilibrium (LD). The method is presented in a genetic model-free approach, although it can also be used by assuming a genetic model of inheritance beforehand. Furthermore, the method is presented in a unified framework and is easily applied to both discrete outcomes (using the OR), as well as to meta-analyses of a continuous outcome (using the mean difference). The main innovation of the method is the analytical calculation of the within-studies covariances between estimates derived from linked polymorphisms. The only requirement is that of an external estimate for the degree of pairwise LD between the polymorphisms under study, which can be obtained from the same published studies, from the literature or from HapMap. Thus, the method is quite simple and fast, it can be extended to an arbitrary set of polymorphisms and can be fitted in nearly all statistical packages (Stata, R/Splus and SAS). Applications in two already published meta-analyses provide encouraging results concerning the robustness and the usefulness of the method and we expect that it would be widely used in the future. *Genet. Epidemiol.* 34:702–715, 2010.

Key words: meta-analysis; multivariate methods; genetic association; linkage disequilibrium

*Correspondence to: Pantelis G. Bagos, Department of Computer Science and Biomedical Informatics, University of Central Greece, Papasiopoulou 2-4, Lamia 35100, Greece. E-mail: pbagos@ucg.gr
Received 2 July 2010; Accepted 5 July 2010
Published online 17 September 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/gepi.20531

INTRODUCTION

The explosion of genetic information and the continuously increasing number of published gene-disease association studies [Becker et al., 2004; Hirschhorn et al., 2002] made imperative the need of collecting and synthesizing the available data in a statistical procedure known as meta-analysis [Normand, 1999; Petiti, 1994; Trikalinos et al., 2008]. Meta-analysis constitutes a particular type of research, in which a set of original studies is synthesized and the potential diversity across them is explored using specific statistical methods [Glass, 1976; Greenland, 1998; Normand, 1999; Petiti, 1994]. Although meta-analysis was initially applied in the field of randomized clinical trials [Chalmers et al., 1987; Sacks et al., 1987], it is nowadays considered a valuable tool for the synthesis of observational studies [Stroup et al., 2000] as well as for gene-disease association. In the latter case, specialized methodology has been proposed for performing meta-analysis of population-based genetic-association studies [Bagos, 2008; Bagos and Nikolopoulos, 2007; Minelli et al., 2004, 2005; Thakkinstian et al., 2005; Trikalinos et al., 2008].

Currently, the statistical methodologies for performing meta-analysis of genetic association studies are oriented toward the ''one gene, one disease'' approach. That is, even in cases where there is available information concerning the association of multiple linked loci (polymorphisms of the same gene or from different ones), each association is considered and analyzed separately. Multivariate approaches for the meta-analysis of genetic association studies are oriented towards modelling separately the odds ratios (ORs) derived from the multiple genotypes in order to infer the genetic model of inheritance. Performing multivariate random effects meta-analysis utilizing several polymorphisms, we can obtain more accurate estimates and most importantly, we can ''borrow strength'' from external studies by pooling studies that report either one of the polymorphisms or both [Higgins and Whitehead, 1996]. The main problem is that the majority of individual studies do not report the combined genotypes for both polymorphisms. Were these available, an extension of the general multivariate method to include multiple markers would be possible, even though this has never been studied in the past (see *Methods*).

The most common approach that takes into account the correlation of adjacent polymorphisms is the haplotype-based analysis [Liu et al., 2008; Schaid, 2004]. Meta-analysis using haplotypes can be easily performed using the multivariate framework of Bagos [2008]; however, the major problem persists; in order for the analysis to be performed, one needs the combined genotypes from each study for reconstructing or inferring the haplotypes,

usually applying an EM or EM-like algorithm [Marchini et al., 2006; Niu, 2004; Xu et al., 2004]. Of course, when individual studies do not report the combined genotypes, as it is generally the case, such an analysis cannot be performed. An alternative strategy for addressing simultaneously the effects of several and possibly correlated polymorphisms (markers) on a particular disease refers to the so-called multipoint (multilocus or multimarker) association methods [Devlin et al., 2003; Marchini et al., 2005; Shoemaker et al., 2001]. Various approaches have been proposed in the past, including seemingly unrelated regression [Verzilli et al., 2005], hierarchical models [Hung et al., 2004], principal components regression [Wang and Abbott, 2008], Markov chain methods [Browning, 2006] and Hidden Markov Models [Marchini et al., 2007]. A major obstacle in applying such methods in meta-analysis is, once again, the paucity of the individual patients' data (IPD) that are needed in order to specify the joint distribution of the genotypes from several linked polymorphisms. Recently, the International HapMap project [HapMap, 2003] has made widely available data concerning the haplotypic structure of the human populations. This structure, i.e. the degree of linkage disequilibrium (LD) [Devlin and Risch, 1995; Guo, 1997; Pritchard and Przeworski, 2001] between various polymorphisms, has been shown to be very useful in genetic association studies [Gu et al., 2008; Marchini et al., 2007; Wen and Nicolae, 2008; Zaitlen et al., 2007].

Recently, Bayesian methods have been proposed for the meta-analysis of published studies reporting various correlated markers, either involving a continuous trait [Verzilli et al., 2008], or a dichotomous outcome (case-control studies) [Newcombe et al., 2009]. In this work we present a simple, yet powerful approach for performing multivariate meta-analysis of published genetic association studies using the distinct polymorphisms in a frequentist framework. The model we propose is an extension of the framework for multivariate meta-analysis of genetic association studies [Bagos, 2008], which in turn is based on the general method for multivariate meta-analysis of van Houwelingen et al. [2002].

METHODS

DICHOTOMOUS OUTCOMES

Assume that we have two single nucleotide polymorphisms (SNPs) in LD, denoted by superscripts (1 and 2), with alleles denoted in both cases A and B. For instance, by A^1B^1 we denote the individuals having the AB genotype for polymorphism 1, whereas by A^2B^2 we denote the individuals having the AB genotype for polymorphism 2, and similarly for the other genotypes (AA, BB). In a meta-analysis setting involving a dichotomous outcome, we will usually have $i = 1,2,\ldots,k$ studies and from each one we will have two groups of individuals, cases denoted by subscript $j = 1$ and controls denoted by subscript $j = 0$. Assuming, as normally is the case, that the same set of patients and controls are genotyped, we see that $N_j = A^1A^1_{ji} + A^1B^1_{ji} + B^1B^1_{ji} = A^2A^2_{ji} + A^2B^2_{ji} + B^2B^2_{ji}$. In case the complete data are available, i.e. we have completely specified the joint distribution of genotypes as in the form of Table I, we can directly use the logistic regression framework provided by Bagos and Nikolopoulos [2007], which corresponds to the so-called analysis of IPD [Turner et al., 2000]. In the majority of situations though, only the marginal genotypes are reported from the published studies and it is more convenient to work with summary data. We proceed by modifying the multivariate framework proposed previously [Bagos, 2008], in order to be able to handle multiple loci. The logarithm of OR^1_{AB} (the OR of heterozygous vs. homozygous for the wild type concerning polymorphism 1) is given by:

TABLE I. The 3 × 3 table used to generate the combined genotypes for polymorphism 1 and polymorphism 2

		SNP 1 (Genotypes)		
		A^1A^1	A^1B^1	B^1B^1
SNP 2 (Genotypes)	A^2A^2	$A^1A^1\ A^2A^2_{ij}$	$A^1B^1\ A^2A^2_{ij}$	$B^1B^1\ A^2A^2_{ij}$
	A^2B^2	$A^1A^1\ A^2B^2_{ij}$	$A^1B^1\ A^2B^2_{ij}$	$B^1B^1\ A^2B^2_{ij}$
	B^2B^2	$A^1A^1\ B^2B^2_{ij}$	$A^1B^1\ B^2B^2_{ij}$	$B^1B^1\ B^2B^2_{ij}$
		$A^1A^1_{ij}$	$A^1B^1_{ij}$	$B^1B^1_{ij}$
				N_{ij}

The column and row totals of the table correspond to the marginal genotype counts for polymorphisms 1 and 2 respectively, whereas the interior cells correspond to the combined genotypes that are unobserved (i denotes the study and j the case/control status).

$$y_{1i} = \log\left(\frac{A^1B^1_{1i}A^1A^1_{0i}}{A^1A^1_{1i}A^1B^1_{0i}}\right) \tag{1}$$

with an approximate variance calculated by:

$$s^2_{1i} = 1/A^1A^1_{0i} + 1/A^1A^1_{1i} + 1/A^1B^1_{0i} + 1/A^1B^1_{1i}. \tag{2}$$

Under the same rationale, the logarithm of OR^2_{AB}(the OR of homozygous for the mutant allele vs. the homozygous for the wild type for polymorphism 1) is given by:

$$y_{2i} = \log\left(\frac{B^1B^1_{1i}A^1A^1_{0i}}{A^1A^1_{1i}B^1B^1_{0i}}\right) \tag{3}$$

with variance equal to:

$$s^2_{2i} = 1/A^1A^1_{0i} + 1/A^1A^1_{1i} + 1/B^1B^1_{0i} + 1/B^1B^1_{1i}. \tag{4}$$

In a similar fashion we can compute the corresponding ORs and the respective variances for polymorphism 2, just by changing the superscripts. Following the general framework for multivariate meta-analysis [Berkey et al., 1998; van Houwelingen et al., 2002], we denote by $\mathbf{y}_i$ the vector containing the four estimates and by $\boldsymbol{\beta}$, the vector of the overall means given by:

$$\mathbf{y}_i = \begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \tag{5}$$

In the multivariate random-effects meta-analysis, we assume that $\mathbf{y}_i$ is distributed following a multivariate normal distribution around the true means $\boldsymbol{\beta}$, according to the marginal model:

$$\mathbf{y}_i \sim \mathrm{MVN}(\boldsymbol{\beta}, \boldsymbol{\Sigma} + \mathrm{C}_i) \tag{6}$$

In the above model, we denote by $\mathbf{C}_i$ the within-studies covariance matrix:

$$\mathbf{C}_i = \begin{pmatrix} s_{1i}^2 & \rho_{W12} s_{1i} s_{2i} & \rho_{W13} s_{1i} s_{3i} & \rho_{W14} s_{1i} s_{4i} \\ \rho_{W12} s_{1i} s_{2i} & s_{2i}^2 & \rho_{W23} s_{2i} s_{3i} & \rho_{W24} s_{2i} s_{4i} \\ \rho_{W13} s_{1i} s_{3i} & \rho_{W23} s_{2i} s_{3i} & s_{3i}^2 & \rho_{W34} s_{4i} s_{3i} \\ \rho_{W14} s_{1i} s_{4i} & \rho_{W24} s_{2i} s_{4i} & \rho_{W34} s_{4i} s_{3i} & s_{4i}^2 \end{pmatrix} \quad (7)$$

and by $\mathbf{\Sigma}$ the between-studies covariance matrix, given by:

$$\mathbf{\Sigma} = \begin{pmatrix} \tau_1^2 & \rho_{B12} \tau_1 \tau_2 & \rho_{B13} \tau_1 \tau_3 & \rho_{B14} \tau_1 \tau_4 \\ \rho_{B12} \tau_1 \tau_2 & \tau_2^2 & \rho_{B23} \tau_2 \tau_3 & \rho_{B24} \tau_2 \tau_4 \\ \rho_{B13} \tau_1 \tau_3 & \rho_{B23} \tau_2 \tau_3 & \tau_3^2 & \rho_{B43} \tau_3 \tau_4 \\ \rho_{B41} \tau_1 \tau_4 & \rho_{B42} \tau_2 \tau_4 & \rho_{B43} \tau_3 \tau_4 & \tau_4^2 \end{pmatrix} \quad (8)$$

The diagonal elements of $\mathbf{C}_i$ are the study-specific estimates of the variance that are assumed known, whereas the off-diagonal elements correspond to the pairwise within studies covariances, for instance $\rho_{w12} s_{1i} s_{2i} = \text{cov}(y_{1i}, y_{2i})$. Since the ORs derived from each polymorphism are correlated, their pairwise covariances will be given [Bagos, 2008] by:

$$\text{cov}(y_{1i}, y_{2i}) = 1/A^1 A_{1i}^1 + 1/A^1 A_{0i}^1 \quad (9)$$

$$\text{cov}(y_{3i}, y_{4i}) = 1/A^2 A_{1i}^2 + 1/A^2 A_{0i}^2 \quad (10)$$

Furthermore, the covariances between logORs derived from different polymorphisms need to be calculated. For instance, we need to calculate:

$$\text{cov}(y_{1i}, y_{3i}) = \text{cov}\left(\log\left(\frac{A^1 B_{1i}^1 A^1 A_{0i}^1}{A^1 A_{1i}^1 A^1 B_{0i}^1} \right), \ \log\left(\frac{A^2 B_{1i}^2 A^2 A_{0i}^2}{A^2 A_{1i}^2 A^2 B_{0i}^2} \right) \right) \quad (11)$$

and similarly for cov (y_{1i}, y_{4i}), cov (y_{2i}, y_{4i}) and cov (y_{2i}, y_{3i}). After some probability calculations using the properties of the covariance function, we obtain (Appendix A):

$$\begin{aligned} \text{cov}(y_{1i}, y_{3i}) = & \ \text{cov}(\log(A^1 B_{1i}^1), \log(A^2 B_{1i}^2)) \\ & + \text{cov}(\log(A^1 A_{0i}^1), \ \log(A^2 A_{0i}^2)) \\ & - \text{cov}(\log(A^1 B_{1i}^1), \ \log(A^2 A_{1i}^2)) \\ & - \text{cov}(\log(A^1 A_{0i}^1), \ \log(A^2 B_{0i}^2)) \\ & - \text{cov}(\log(A^1 A_{1i}^1), \ \log(A^2 B_{1i}^2)) \\ & - \text{cov}(\log(A^1 B_{0i}^1), \ \log(A^2 A_{0i}^2)) \\ & + \text{cov}(\log(A^1 A_{1i}^1), \ \log(A^2 A_{1i}^2)) \\ & + \text{cov}(\log(A^1 B_{0i}^1), \ \log(A^2 B_{0i}^2)) \end{aligned} \quad (12)$$

The covariances between log-counts, i.e. cov$(\log(A^1 B_{ji}^1), \log(A^2 B_{ji}^2))$, from different polymorphisms, are generally non-zero and we can calculate them using the delta method (Appendix A). For instance, we will have ($j = 0,1$):

$$\text{cov}\left[\log(A^1 B_{ji}^1), \ \log(A^2 B_{ji}^2) \right] = \frac{A^1 B_{ji}^1 A^2 B_{ji}^2}{(A^1 B_{ji}^1)\,(A^2 B_{ji}^2)} \quad (13)$$

where, by $A^1 B^1 A^2 B_{ji}^2$ we denote the number of individuals in the j group of ith study carrying the combined genotype (i.e. *AB* for both markers). Similar results will hold for the other terms (Appendix A). It is easy to show that the covariances will be equal to zero if and only if the two polymorphisms are independent (i.e. $r = 0$). Furthermore, similar results could be obtained if we choose to model the genetic model of inheritance (Appendix A).

Attention should be paid when the genotype counts for different polymorphisms do not sum to the same total for a given published study. This is observed often, since genotyping errors usually result in small discrepancies. In such a case, we can extrapolate the counts of the polymorphism with the smaller sum to match the larger one and proceed based on the extrapolated genotype counts. If we assume, for instance, that polymorphism 2 has a smaller total count compared to polymorphism 1 (i.e. $N_j = A^1 A_{ji}^1 + A^1 B_{ji}^1 + B^1 B_{ji}^1 > A^2 A_{ji}^2 + A^2 B_{ji}^2 + B^2 B_{ji}^2$), we can extrapolate to obtain $A^2 A_{0i}^2{}' + A^2 B_{0i}^2{}' + B^2 B_{0i}^2{}' N_0$. Then, using Equation (12) we can calculate cov(y'_{ki}, y'_{li}), and subsequently the correlation ρ_{wkl} cov$(y'_{ki}, y'_{li})/(s'_{ki}\, s'_{li})$. Note that, s'_{li} and s'_{ki} have been calculated from the extrapolated genotypes. Finally, the ''expected'' covariance will be given by cov$(y_{ki}, y_{li}) = (s_{li} s_{ki}/s'_{li}\, s'_{li})$ cov(y'_{ki}, y'_{li}). The resulting differences will be negligible, unless there are large discrepancies in the total sums; in such a case, the within studies covariance matrix $\mathbf{C}_i$ may not be positive definite. A non-positive definite covariance matrix can, in general, occur when there are missing values [Schwertman and Allen, 1979] and we expect to occur frequently especially when the analysis is performed on more than two polymorphisms. In order to transform a non-positive definite covariance matrix into positive definite, we used here a simple heuristic consisting of adding the negative of the smallest eigenvalue (which will be negative) plus a small constant (for instance 10^{-7}) to the diagonal elements. In matrix notation, if we denote the identity matrix by $\mathbf{I}$, $\varepsilon = 10^{-7}$ and by λ the smallest eigenvalue of $\mathbf{C}_i$, we will have:

$$\mathbf{C}_i' = \mathbf{C}_i + \mathbf{I}\,(\varepsilon - \lambda) \quad (14)$$

Several other correction techniques have been proposed in the literature [Higham, 2002; Rebonato and Jäckel, 1999; Schwertman and Allen, 1979] and their application in a meta-analysis setting should be investigated in the future.

A major advantage of the model of Equation (6) is that it can be used to estimate the OR for a combined genotype. For instance, after the model is fitted we can estimate the OR for persons carrying the $B^1 B^1 B^2 B^2$ vs. $A^1 A^1 A^2 A^2$ genotype by computing the quantity $\beta_2 + \beta_4$ (assuming no gene-gene interaction). Similarly, $\beta_1 + \beta_4$ will give the OR for $A^1 B^1 B^2 B^2$ vs. $A^1 A^1 A^2 A^2$ and so forth. A confidence interval can be easily constructed from the estimated variance-covariance matrix, using the delta method. This technique can be viewed as a more flexible alternative to the widely used haplotype analysis approach, since it allows different models of inheritance to be considered; by fitting the model of Equation (6), we can directly compute the ratio of the two logORs for each genotype (denoted by λ) along with the respective confidence interval that provides an estimate for the genetic model of inheritance [Bagos, 2008; Minelli et al., 2005].

The complete data that we initially need are shown in Table I. If the combined genotypes are known (which is rarely the case) we can use the counts to derive the covariances directly (Equations (12–16)). This is equivalent to the analysis using the logistic regression model described above. In the majority of the published studies though, the combined genotypes are not given; instead, only the marginal genotypes are reported. Some studies

may additionally report a measure of LD (usually r); however, even in the absence of this, the Hapmap project can be utilized in the majority of situations, since the majority of genes being meta-analyzed are likely to be included in the database. It is important to notice here that the summary data multivariate meta-analysis framework described above can be easily extended to three or more SNPs, since we only need the pairwise correlations between polymorphisms. On the contrary, the IPD approach requires the full joint distribution of all SNPs included in the model.

CONTINUOUS OUTCOMES

In case we have a continuous outcome (Y), such as systolic blood pressure (SBP) or any other continuous trait, each study will provide us with the mean and standard deviation for the outcome for each genotype, along with the number of persons having that genotype. The general multivariate framework (Equation (6)) allows direct modeling of the pairwise differences of the outcomes arising from individuals carrying the two mutant genotypes, from the reference genotype. For instance, for the first polymorphism we will have (denoting the genotypes by A^1B^1, B^1B^1, A^1A^1):

$$y_{1i} = Y_{A^1B_i^1} - Y_{A^1A_i^1} \text{ and } y_{2i} = Y_{B^1B_i^1} - Y_{A^1A_i^1} \quad (15)$$

It is straightforward to show that the two differences are distributed normally with mean equal to the difference of the means and variance equal to the sum of the variances:

$$y_{1i} \sim N\left(Y_{A^1B_i^1} - Y_{A^1A_i^1}, \frac{sd^2_{A^1B_i^1}}{A^1B_i^1} + \frac{sd^2_{A^1A_i^1}}{A^1A_i^1}\right) \text{ and}$$

$$y_{2i} \sim N\left(Y_{B^1B_i^1} - Y_{A^1A_i^1}, \frac{sd^2_{B^1B_i^1}}{B^1B_i^1} + \frac{sd^2_{A^1A_i^1}}{A^1A_i^1}\right) \quad (16)$$

Similar equations also hold for the second polymorphism. Hence, it is obvious that the model can be directly applied in order to perform multivariate modeling of the four pairwise differences. The two differences derived from each polymorphism are correlated, since they are both estimating a difference from a baseline category. By simple probability calculations [Bagos, 2008], we get the covariance as follows:

$$\text{cov}(y_{1i}, y_{2i}) = \text{var}(Y_{A^1A_i^1}) = \frac{sd^2_{A^1A_i^1}}{A^1A_i^1} \quad (17)$$

Similarly to the discrete outcome measures, the covariance is equal to the amount of variance attributed to the baseline group. A completely analogous expression also holds for the second polymorphism, i.e. for $\text{cov}(y_{3i},y_{4i})$. The pairwise covariances for different polymorphisms, i.e. $\text{cov}(y_{2i},y_{4i})$, can be calculated by expressing the mean value of the continuous outcome Y for the persons carrying a particular marginal genotype (A^1A^1), as a weighted average of the mean values of the persons carrying all composite genotypes that include this marginal ($A^1A^1A^2A^2$, $A^1A^1A^2B^2$ and $A^1A^1B^2B^2$) with weights given by the combined genotype counts that we have already computed (see Appendix A). For instance, we can show that (Appendix B):

$$\begin{aligned}\text{cov}(y_{2i}, y_{4i}) = &\frac{(B^1B^1B^2B_i^2)}{(B^1B_i^1)\,(B^2B_i^2)} sd^2_{B^1B^1B^2B_i^2} \\ &+ \frac{(A^1A^1A^2A_i^2)}{(A^1A_i^1)\,(A^2A_i^2)} sd^2_{A^1A^1A^2A_i^2} \\ &- \frac{(B^1B^1A^2A_i^2)}{(B^1B_i^1)\,(A^2A_i^2)} sd^2_{B^1B^1A^2A_i^2} \\ &- \frac{(A^1A^1B^2B_i^2)}{(A^1A_i^1)\,(B^2B_i^2)} sd^2_{A^1A^1B^2B_i^2}\end{aligned} \quad (18)$$

Similar formulae can be obtained for the other covariances (Appendix B). Similarly to the discrete outcome case, we need the number of persons carrying the combined genotypes. In addition to that, we also need the standard deviations of the continuous outcome Y for the persons belonging to different combined genotype groups. The latter is difficult to be calculated, but we can safely approximate it by averaging the standard deviations of the groups representing the two marginal genotypes. For instance, $sd^2_{A^1A^1A^2A^2}$ can be approximated by $sd^2_{A^1A^1}/2 + sd^2_{A^2A^2}/2$. The assumption of equal variances is reasonable; furthermore, the covariance is influenced mostly by the combined genotypes.

The previously discussed issues concerning the within studies covariance matrix being non-positive definite apply in this case as well. Similar to the discrete outcomes, using the model of Equation (6) we can estimate the weighted mean difference for the continuous outcome for a combined genotype (i.e. the difference $B^1B^1B^2B^2$ vs. $A^1A^1A^2A^2$). Moreover, if we had available IPD the model would be equivalent to an extension of the linear mixed model described previously [Bagos, 2008]. Finally, we have to note that by assuming equal standard deviations, the covariance would be zero if and only if the polymorphisms are independent ($r = 0$).

CALCULATION OF THE COMBINED GENOTYPES

As we have mentioned, in the majority of the published studies the combined genotypes are not given. Some studies may additionally report a measure of LD (usually r or D') [Guo, 1997]; however, even if they do not, the Hapmap project [HapMap, 2003] or the relevant literature can be used in the majority of situations. In order to compute the combined genotypes, we will initially proceed by estimating the haplotype counts. Table II

TABLE II. The 2 × 2 table corresponding to the haplotype counts

		SNP 1 (Alleles)		
		A^1	B^1	
SNP 2 (Alleles)	A^2	$A^1\ A^2_{ij}$	$B^1\ A^2_{ij}$	$2A^2A^2_{ij} + A^2B^2_{ij}$
	B^2	$A^1\ B^2_{ij}$	$B^1\ B^2_{ij}$	$2B^2B^2_{ij} + A^2B^2_{ij}$
		$2A^1A^1_{ij} + A^1B^1_{ij}$	$2B^1B^1_{ij} + A^1B^1_{ij}$	$2N_{ij}$

The column and row totals of the table correspond to the marginal allele counts for polymorphisms 1 and 2 respectively, whereas the interior cells correspond to the haplotypes that are unobserved (i denotes the study, j the case/control status).

denotes the cross-tabulation of alleles for polymorphisms 1 and 2. It is a 2×2 contingency table with known marginals. Such contingency table formulations have been used for haplotype estimation from pooled genotype data [Xu et al., 2008]. However, it is known that if a measure of association is given, the interior cells of the table are uniquely identified [Agresti, 2002]. Then, by assuming r is given, the unique solution can be found analytically by solving a system of five equations in four unknowns (Appendix C). Accordingly, we can calculate the probabilities of observing the four haplotypes and given the haplotype probabilities, we can also calculate the probabilities of the combined genotypes under Hardy-Weinberg Equilibrium (HWE, Appendix C). In some cases the assumption of HWE could be unrealistic; however, the influence on the overall estimates of the meta-analysis is negligible since it affects only slightly the covariances between the logORs. Moreover, since we usually assume that cases and controls would have the same degree of LD between the polymorphism, we implicitly assume that there is no gene-gene interaction.

Care should be taken when the individual studies do not report measures of LD and we have to use r^2 as provided by HapMap. First, we have to choose the population that best resembles the one under study, considering both the racial descent (i.e. European, Asian etc.) and the allele frequencies, and second, we have to ensure that the proper sign for r (i.e. $\pm\sqrt{r^2}$) is chosen. In case the opposite sign is chosen for r, the predicted haplotype counts will contain large negative numbers and this could be spotted easily. A straightforward approach that can be easily followed would be to identify the accession numbers of the particular polymorphisms using the Variant Name Mapper (http://www.hugenavigator.net/HuGENavigator/startPageMapper.do) [Yu et al., 2009] and subsequently find the r^2 measures directly using GLIDERS (http://mather.well.ox.ac.uk/GLIDERS/) [Lawrence et al., 2009].

IMPLEMENTATION

The model can be fitted in any statistical package capable of fitting random-effects weighted regression models with an arbitrary covariance matrix, such as SAS (using PROC MIXED or PROC NLMIXED), R (using lme) or Stata (using mvmeta). In this work, we used mvmeta, which performs inferences based on either maximum likelihood (ML) or restricted maximum likelihood (REML), by direct maximization of the approximate likelihood using a Newton-Raphson algorithm [White, 2009]. Alternatively, mvmeta can also implement the multivariate version of the DerSimonian and Laird's method of moments [Jackson et al., 2009]. The last option, being non-iterative, is very attractive in the case of a large number of polymorphisms and/or large number of studies. A Stata program (mpmeta) that performs all the necessary computation will be made available from the authors.

APPLICATION OF THE METHOD

ASSOCIATION OF CX3CR1 GENE POLYMORPHISMS WITH CORONARY ARTERY DISEASE

We have applied the proposed methods in data from the meta-analysis for the association of CX3CR1 gene polymorphisms with coronary artery disease [Apostolakis et al., 2009]. The meta-analysis contains information about two polymorphisms, V249I (rs3732379) and T280M (rs3732378) reported in a total of seven published studies. We chose this particular meta-analysis since in the initial reports the authors have provided detailed information for the combined genotypes as well as the LD between the two polymorphisms. Thus, it is an ideal example in which we can analytically test the appropriateness of our approach.

The reconstruction procedure provided results that are in close agreement with the observed genotypes. For instance, in the six of the included studies that reported both polymorphisms, the correct genotype has been predicted for 96.11% of the controls and 96.63% of the cases, using LD measures (r^2) taken from the individual studies (r^2 ranging from 0.45 to 0.59). When the estimate from HapMap was used ($r^2 = 0.58$), the respective numbers were 95.66 and 95.51%. Thus, the method proposed here is quite successful in determining unobserved combined genotypes. When the two polymorphisms were analyzed separately (single-point analysis) using the general method for meta-analysis, the OR for the TM^{280} genotype was marginally insignificant (OR = 0.769; 95% CI: 0.587, 1.009). The multipoint method for meta-analysis (Fig. 1) did not provide, however, any additional evidence for the association of the TM^{280} genotype compared to TT^{280} genotype (OR = 0.790; 95% CI: 0.601, 1.037). The other three genotypes did not show any significant effect (Fig. 1).

The overall estimates using the HapMap LD correlation coefficient are nearly identical, up to the third decimal place (data not shown). The fact that the OR for MM^{280} genotype compared to TT^{280} was insignificant should be attributed to the small number of persons carrying that genotype. The results obtained here should be compared against those presented in the initial report [Apostolakis et al., 2009], having in mind though, that the authors used the fixed effects estimates which produce narrower confidence intervals (OR for TM^{280} vs. TT^{280} = 0.83, 95% CI: 0.72, 0.97). However, both results point to the same direction concerning the susceptibility of M allele.

ASSOCIATION OF ANGIOTENSINOGEN POLYMORPHISMS WITH HYPERTENSION

We additionally investigated the association of Angiotensinogen (AGT) polymorphisms with hypertension. For the meta-analysis, we included the T174M polymorphism (rs4762) in exon 2, two promoter polymorphisms, G[−6]A (rs5051) and C[−20]A (rs5050), as well as the major polymorphism of AGT, the M235T (rs699). For the complete analysis, we combined the available data from two previous meta-analyses [Pereira et al., 2008; Sethi et al., 2003] yielding in total data for 20,579 cases and 19,784 controls from 79 independent studies. Of the 79 included studies, 40 were performed on Caucasian populations, 28 on Asians, 7 on populations of African-descent whereas 4 studies were performed on mixed populations. Since in this meta-analysis the studies contained overlapping sets of markers, it was possible to investigate the process of borrowing strength from external studies. In Table III, we list the distribution of polymorphism across studies from which it is obvious that only one study contained data for all four polymorphisms.

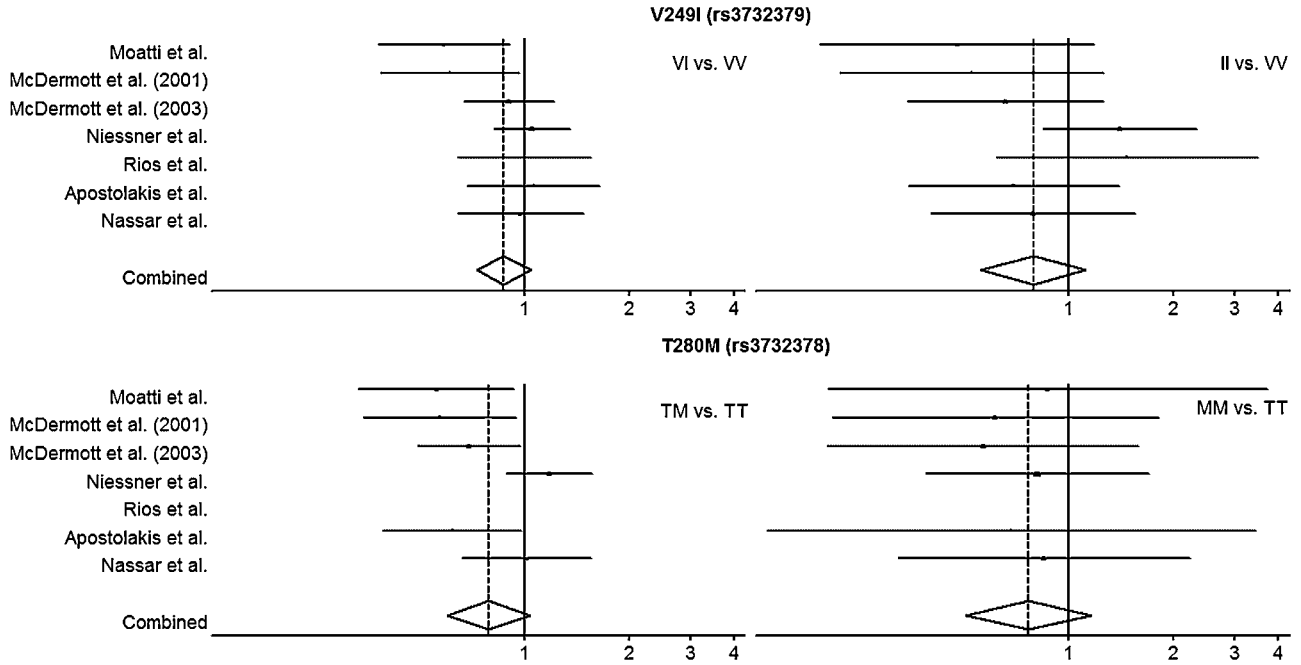

Fig. 1. Forest plots for the meta-analysis for the association of CX3CR1 gene polymorphisms with coronary artery disease. We report two odds ratios for each polymorphism (V249I and T280M) as described in the text. The size of the symbols for the estimates is inversely proportional to their variances.

TABLE III. The distribution of the polymorphisms across studies in the meta-analysis concerning the association of AGT polymorphisms with hypertension

	Polymorphisms			
Number of studies	T174M (rs4762)	G[-6]A (rs5051)	C[-20]A (rs5050)	M235T (rs699)
29	−	−	−	+
11	+	−	−	+
11	−	+	−	−
10	+	−	−	−
4	−	+	+	−
4	−	−	+	−
3	+	+	−	+
2	−	+	−	+
1	+	+	+	+
79	29	23	13	48

The positive sign (+) denotes that the particular polymorphism has been typed in a given study. The last row contains the total number of studies. It is obvious that the four polymorphisms were differentially typed in different studies (as a matter of fact only one study has typed all four polymorphisms).

The results are presented in Table IV. We initially performed a single-point analysis using the multivariate framework presented earlier [Bagos, 2008]. This analysis yielded results similar to the ones presented in the initial reports [Pereira et al., 2008; Sethi et al., 2003], pointing to a moderate association of T174M, G[−6]A and M235T with hypertension (Table IV). Interestingly, when the multipoint method was applied, the associations of T174M and M235T with hypertension were attenuated, the association of G[−6]A was reversed and more importantly, the effect of C[−20]A became significant. This resembles the well-known situation encountered in ordinary linear regression when including an (intermediate) causal variable. Thus, the results clearly indicate that the effects of T174M and M235T can be explained entirely by the LD patterns with the other two SNPs in the analysis. Moreover, the effect of C[−20]A polymorphism became stronger (i.e. the variance of the estimate was decreased) indicating that the method has borrowed strength from the other studies using the observed LD patterns. The results indicate that carriers of the $AA^{[-6]}$ genotype have approximately 27% higher risk for developing hypertension compared to $GG^{[-6]}$, whereas carriers of the $CC^{[-20]}$ genotype have approximately 50% higher risk for developing hypertension compared to $AA^{[-20]}$.

These findings are very important since they indicate that the postulated association of AGT polymorphisms with hypertension is mediated by the promoter polymorphisms and not by the coding non-synonymous polymorphisms (M235T and T174M). Whereas there is no evidence that the latter polymorphisms have any functional activity, there is plenty of data supporting the idea that the promoter polymorphisms C[−20]A and G[−6]A have direct effects on transcription activity influencing thus the total amount of produced AGT [Dickson et al., 2007; Inoue et al., 1997; Ishigami et al., 1999; Jain et al., 2002]. The effect of another important functional polymorphism, G[−217]A, that also influences the plasma AGT levels could not have been investigated in this meta-analysis, since genotype data as well as data for LD patterns were not available. Interestingly, performing subgroup analyses we found that the effect of M235T persists in Caucasians, whereas that of C[−20]A and G[−6]A, in Asians (data not shown). However, these

TABLE IV. The results obtained from the meta-analysis for the association of AGT polymorphisms with hypertension

	Single-point analysis [Bagos, 2008]		Multipoint analysis [this work]	
SNP	AB vs. AA OR (95% CI)	BB vs. AA OR (95% CI)	AB vs. AA OR (95% CI)	BB vs. AA OR (95% CI)
T174M (rs4762)	1.126 (1.023, 1.239)	1.445 (1.057, 1.975)	0.964 (0.884, 1.050)	0.789 (0.605, 1.027)
G[-6]A (rs5051)	0.821 (0.661, 1.021)	0.660 (0.484, 0.901)	1.182 (0.974, 1.436)	1.273 (1.009, 1.605)
C[-20]A (rs5050)	0.897 (0.617, 1.304)	1.033 (0.719, 1.483)	0.510 (0.417, 0.623)	0.574 (0.476, 0.693)
M235T (rs699)	1.055 (0.992, 1.122)	1.189 (1.103, 1.281)	1.042 (0.958, 1.133)	1.112 (0.975, 1.279)

Single-point analysis consists of using the general multivariate methodology in which each polymorphism is analysed separately, whereas multipoint analysis corresponds to fitting the models described in this work. We list the OR for the two genotype contrasts (AB vs. AA and BB vs. AA) along with their 95% CI.

results may be confounded by the differential reporting of the polymorphisms by the studies in the various ethnic groups and this should be the subject of future research.

Nevertheless, it is clear that the method is very useful in detecting the true causal variants in the presence of LD and can greatly improve the quality of the conclusions drawn from such a meta-analysis. In the initial analysis, Pereira and coworkers tried to investigate the role of T174M and concluded that the polymorphism was independently associated with hypertension [Pereira et al., 2008]. They tried to investigate the effect of LD patterns between T174M and M235T by regressing the logOR derived for the T174M polymorphism against the logOR derived for the M235T polymorphism. However, this approach is flawed since it is limited only in studies reporting both polymorphisms and ignores the within-studies correlation as well as the measurement error in the independent variable, as previously described in a different context [Thompson et al., 1997; van Houwelingen and Senn, 1999].

DISCUSSION

We described here a simple yet powerful method for the meta-analysis of data from published genetic association studies utilizing several correlated markers. We have shown that reliable estimates can be obtained using only the reported marginal genotype counts from each study, along with a measure of LD between the markers. Our approach allows easily estimating the genetic model of inheritance. Compared to simple analyses that use information from only one polymorphism at a time, the method is expected to be more robust since it avoids multiple comparisons. Furthermore, when the studies contain information for different sets of polymorphisms, the method proposed here will be more powerful since it allows borrowing strength from external studies and thus producing narrower confidence intervals [Higgins and Whitehead, 1996].

Application of the method in two already published meta-analyses provided very encouraging results. Applying the method in the meta-analysis concerning the association of CX3CR1 gene polymorphisms with coronary artery disease, we have shown that the method for reconstructing the combined genotypes is very accurate. Furthermore, in the meta-analysis for the association of AGT polymorphisms with hypertension we have shown that the multipoint meta-analysis can yield better results compared to only analyzing one polymorphism at a time, since it provided evidence for the true causal variants in the promoter region. The method was shown to be more powerful when there are several partially typed polymorphisms in the included studies, in which case it would allow borrowing strength from external studies yielding results that could not have been obtained otherwise.

In many types of multivariate meta-analysis, multiple correlated outcomes are encountered; as a result, the within studies correlation needs to be calculated from individual data [Berkey et al., 1998]. The meta-analyses of (a) genetic association studies for a single polymorphism [Bagos, 2008], (b) studies reporting mutually exclusive outcomes [Trikalinos and Olkin, 2008], (c) clinical trials with multiple treatments [Ades, 2003; Higgins and Whitehead, 1996] and (d) studies containing overlapping sets of subjects (cases or controls) [Lin and Sullivan, 2009] are examples where the within studies correlation can be calculated analytically. However, in other situations, such as meta-analysis of Mendelian randomization studies and meta-analysis of surrogate markers, the correlation cannot be calculated. As a result, simulation techniques are used [Daniels and Hughes, 1997; Thompson et al., 2005]. Within studies correlations have been studied in detail in the past; it is generally accepted that by ignoring or approximating those leads to biased estimates of the variance of the overall effect [Riley et al., 2007a,b]. Later on, an alternative model for bivariate meta-analysis was proposed that requires as input only the two correlated outcomes and does not separate the between-studies and the within studies correlation, but instead it computes a single parameter for the ''overall'' correlation [Riley et al., 2008]. The particular method can be used in the meta-analysis of two correlated polymorphisms, but being a bivariate method, this can be accomplished without modification only by assuming a particular genetic model.

The method proposed here is very general and is applicable to both dichotomous and continuous outcomes. Compared to the recently proposed Bayesian methods [Newcombe et al., 2009; Verzilli et al., 2008], it is much simpler and does not require the use of specialized software or the need for extensive computations using MCMC algorithms. We should emphasize here that, besides the computational complexity, the Bayesian methods are not capable of performing an analysis without assuming a particular genetic model, without modification. On the contrary, our approach, being an extension of a previously presented general method [Bagos, 2008], can directly estimate the genetic model of inheritance for each polymorphism without assuming it beforehand. Moreover, Bayesian methods use a complicated parameterization and model specification, whereas the method presented here is a simple extension of the general method for multivariate

meta-analysis [Bagos, 2008]. Thus, a major advantage is that the method relies on simple analytical calculations in order to derive the covariances and not on iterative procedures. Consequently, it can be implemented in almost any statistical package capable of fitting generalized linear mixed models with an arbitrary covariance matrix (Stata, SAS and R/Splus). Bayesian methods are more difficult to be used, since an investigator should have knowledge of Bayesian statistics, as well as programming skills. Moreover, even a simple Bayesian analysis requires a significantly longer amount of time in order to monitor the convergence of MCMC and perform the necessary diagnostics.

Similar to the Bayesian methods [Newcombe et al., 2009; Verzilli et al., 2008], the method proposed in this work makes the assumption of HWE that can be questionable especially among cases [Salanti et al., 2005; Trikalinos et al., 2006]. However, this assumption is only used to calculate the covariances, so it will only slightly influence the overall estimates even when departures from HWE are observed. Another assumption that the method makes similarly to the Bayesian methods is that the LD patterns in the populations under study are similar to the ones reported in the literature or in HapMap. If this assumption fails, the calculated covariances will be questionable, irrespective of the method used (Bayesian of frequentist). Nevertheless, up to some degree, this assumption can be tested by comparing the allele frequencies in the studied populations against the ones reported in HapMap, as discussed previously [Newcombe et al., 2009; Verzilli et al., 2008].

Several methods have been proposed previously for multilocus association in which information from untyped markers has been shown to be useful in inferring or imputing the missing genotypes [Browning, 2008; Marchini et al., 2007; Nicolae, 2006; Wen and Nicolae, 2008]. However, these methods require the complete genotype for each individual in order to infer the genotype of the missing marker. The method proposed here is in some respects the meta-analytic analogue of such methods, using though only published summary data. However, it differs in that it does not directly impute the missing marginal genotypes, but instead allows to borrow strength from studies reporting both polymorphisms by using the within studies covariance of the estimates, in order to improve the overall estimate. The method that we proposed also shares some analogies with the recently proposed method of ''synthesis analysis'' [Samsa et al., 2005]. Whereas traditional meta-analysis search for a single summary measure for the relation of two variables (Y, X, i.e. gene and disease) across k studies, synthesis analysis seeks to integrate estimates for two or more predictors (X_1, X_2), in a multivariate model for predicting Y, using only information from the pairwise comparisons, for instance using estimates from the relations (Y, X_1), (Y, X_2) and (X_1, X_2) [Samsa et al., 2005; Zhou et al., 2009]. In principle, a crude synthesis method can be used for the meta-analysis if we use the estimates derived from individual meta-analyses (i.e. for each polymorphism). However, the method has been presented only for continuous variables and calculation of the covariances between logORs has to be done with a method similar to the one proposed here. Nevertheless, the method proposed here integrates elegantly in a unified framework meta-analysis and synthesis analysis and we anticipate that it will be useful and widely used.

ACKNOWLEDGMENTS

The authors thank Dr Vasilios Plagianakos and Dr Maria Adam for their valuable discussions and help in certain algebraic derivations used in this work.

REFERENCES

Ades AE. 2003. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. Stat Med 22:2995–3016.

Agresti A. 2002. Categorical Data Analysis. New York: Wiley.

Apostolakis S, Amanatidou V, Papadakis EG, Spandidos DA. 2009. Genetic diversity of CX3CR1 gene and coronary artery disease: new insights through a meta-analysis. Atherosclerosis 207:8–15.

Bagos PG. 2008. A unification of multivariate methods for meta-analysis of genetic association studies. Stat Appl Genet Mol Biol 7:Article31.

Bagos PG, Nikolopoulos GK. 2007. A method for meta-analysis of case-control genetic association studies using logistic regression. Stat Appl Genet Mol Biol 6:Article17.

Becker KG, Barnes KC, Bright TJ, Wang SA. 2004. The genetic association database. Nat Genet 36:431–432.

Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. 1998. Meta-analysis of multiple outcomes by regression with random effects. Stat Med 17:2537–2550.

Browning SR. 2006. Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78:903–913.

Browning SR. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124:439–450.

Chalmers TC, Berrier J, Sacks HS, Levin H, Reitman D, Nagalingam R. 1987. Meta-analysis of clinical trials as a scientific discipline. II: Replicate variability and comparison of studies that agree and disagree. Stat Med 6:733–744.

Daniels MJ, Hughes MD. 1997. Meta-analysis for the evaluation of potential surrogate markers. Stat Med 16:1965–1982.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Devlin B, Roeder K, Wasserman L. 2003. Analysis of multilocus models of association. Genet Epidemiol 25:36–47.

Dickson ME, Zimmerman MB, Rahmouni K, Sigmund CD. 2007. The −20 and −217 promoter variants dominate differential angiotensinogen haplotype regulation in angiotensinogen-expressing cells. Hypertension 49:631–639.

Glass G. 1976. Primary, secondary and meta-analysis of research. Educ Res 5:3–8.

Greenland S. 1998. Meta-analysis. In: Rothman KJ, Greenland S, editors. Modern Epidemiology. Philadelphia: Lippincott Williams & Wilkins. p 643–673.

Gu CC, Yu K, Rao DC. 2008. Characterization of LD structures and the utility of HapMap in genetic association studies. Adv Genet 60:407–435.

Guo SW. 1997. Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum Hered 47:301–314.

HapMap. 2003. The International HapMap Project. Nature 426: 789–796.

Higgins JP, Whitehead A. 1996. Borrowing strength from external trials in a meta-analysis. Stat Med 15:2733–2749.

Higham NJ. 2002. Computing the nearest correlation matrix—a problem from finance IMA. J Num Anal 22:329–343.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. Genet Med 4:45–61.

Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, Witte JS. 2004. Using hierarchical modeling in genetic association

studies with multiple markers: application to a case-control study of bladder cancer. Cancer Epidemiol Biomarkers Prev 13:1013–1021.
Inoue I, Nakajima T, Williams CS, Quackenbush J, Puryear R, Powers M, Cheng T, Ludwig EH, Sharma AM, Hata A, Jeunemaitre X, Lalouel JM. 1997. A nucleotide substitution in the promoter of human angiotensinogen is associated with essential hypertension and affects basal transcription in vitro. J Clin Invest 99:1786–1797.
Ishigami T, Tamura K, Fujita T, Kobayashi I, Hibi K, Kihara M, Toya Y, Ochiai H, Umemura S. 1999. Angiotensinogen gene polymorphism near transcription start site and blood pressure: role of a T-to-C transition at intron I. Hypertension 34:430–434.
Jackson D, White IR, Thompson SG. 2010. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. Stat Med 29:1282–1297.
Jain S, Tang X, Narayanan CS, Agarwal Y, Peterson SM, Brown CD, Ott J, Kumar A. 2002. Angiotensinogen gene polymorphism at -217 affects basal promoter activity and is associated with hypertension in African-Americans. J Biol Chem 277:36889–36896.
Lawrence R, Day-Williams AG, Mott R, Broxholme J, Cardon LR, Zeggini E. 2009. GLIDERS—a web-based search engine for genome-wide linkage disequilibrium between HapMap SNPs. BMC Bioinformatics 10:367.
Lin DY, Sullivan PF. 2009. Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet 85:862–872.
Liu N, Zhang K, Zhao H. 2008. Haplotype-association analysis. Adv Genet 60:335–405.
Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417.
Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P. 2006. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78:437–450.
Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.
Minelli C, Thompson JR, Tobin MD, Abrams KR. 2004. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. Am J Epidemiol 160:445–452.
Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. 2005. The choice of a genetic model in the meta-analysis of molecular association studies. Int J Epidemiol 34:1319–1328.
Newcombe PJ, Verzilli C, Casas JP, Hingorani AD, Smeeth L, Whittaker JC. 2009. Multilocus Bayesian meta-analysis of gene-disease associations. Am J Hum Genet 84:567–580.
Nicolae DL. 2006. Testing untyped alleles (TUNA)-applications to genome-wide association studies. Genet Epidemiol 30: 718–727.
Niu T. 2004. Algorithms for inferring haplotypes. Genet Epidemiol 27:334–347.
Normand SL. 1999. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 18:321–359.
Pereira TV, Nunes AC, Rudnicki M, Yamada Y, Pereira AC, Krieger JE. 2008. Meta-analysis of the association of 4 angiotensinogen polymorphisms with essential hypertension: a role beyond M235T? Hypertension 51:778–783.
Petiti DB. 1994. Meta-Analysis Decision Analysis and Cost-Effectiveness Analysis. Oxford: Oxford University Press.
Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14.
Rebonato R, Jäckel P. 1999. The most general methodology to create a valid correlation matrix for risk management and option pricing purposes. J Risk 2:17–28.
Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. 2007a. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat Med 26:78–97.
Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. 2007b. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Med Res Methodol 7:3.
Riley RD, Thompson JR, Abrams KR. 2008. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics 9:172–186.
Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. 1987. Meta-analyses of randomized controlled trials. N Engl J Med 316:450–455.
Salanti G, Amountza G, Ntzani EE, Ioannidis JP. 2005. Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. Eur J Hum Genet 13:840–848.
Samsa G, Hu G, Root M. 2005. Combining information from multiple data sources to create multivariable risk models: illustration and preliminary assessment of a new method. J Biomed Biotechnol 2005:113–123.
Schaid DJ. 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348–364.
Schwertman NC, Allen DM. 1979. Smoothing an indefinite variance-covariance matrix. J Stat Comput Simulation 9:183–194.
Sethi AA, Nordestgaard BG, Tybjaerg-Hansen A. 2003. Angiotensinogen gene polymorphism, plasma angiotensinogen, and risk of hypertension and ischemic heart disease: a meta-analysis. Arterioscler Thromb Vasc Biol 23:1269–1275.
Shoemaker CA, Pungliya M, Sao Pedro MA, Ruiz C, Alvarez SA, Ward M, Ryder EF, Krushkal J. 2001. Computational methods for single-point and multipoint analysis of genetic variants associated with a simulated complex disorder in a general population. Genet Epidemiol 21:S738–S745.
Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. 2000. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 283:2008–2012.
Thakkinstian A, McElduff P, D'Este C, Duffy D, Attia J. 2005. A method for meta-analysis of molecular association studies. Stat Med 24:1291–1306.
Thompson SG, Smith TC, Sharp SJ. 1997. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 16: 2741–2758.
Thompson JR, Minelli C, Abrams KR, Tobin MD, Riley RD. 2005. Meta-analysis of genetic studies using Mendelian randomization—a multivariate approach. Stat Med 24:2241–2254.
Trikalinos TA, Olkin I. 2008. A method for the meta-analysis of mutually exclusive binary outcomes. Stat Med 27:4279–4300.
Trikalinos TA, Salanti G, Khoury MJ, Ioannidis JP. 2006. Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations. Am J Epidemiol 163:300–309.
Trikalinos TA, Salanti G, Zintzaras E, Ioannidis JP. 2008. Meta-analysis methods. Adv Genet 60:311–334.
Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. 2000. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 19:3417–3432.
van Houwelingen H, Senn S. 1999. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 18:110–115.
van Houwelingen HC, Arends LR, Stijnen T. 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 21:589–624.
Verzilli CJ, Stallard N, Whittaker JC. 2005. Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. Genet Epidemiol 28:313–325.
Verzilli C, Shah T, Casas JP, Chapman J, Sandhu M, Debenham SL, Boekholdt MS, Khaw KT, Wareham NJ, Judson R, Benjamin EJ, Kathiresan S, Larson MG, Rong J, Sofat R, Humphries SE, Smeeth L, Cavalleri G, Whittaker JC, Hingorani AD. 2008. Bayesian meta-analysis of genetic association studies with different sets of markers. Am J Hum Genet 82:859–872.

Wang K, Abbott D. 2008. A principal components regression approach to multilocus genetic association studies. Genet Epidemiol 32:108–118.

Wen X, Nicolae DL. 2008. Association studies for untyped markers with TUNA. Bioinformatics 24:435–437.

White IR. 2009. Multivariate random-effects meta-analysis. Stata J 9:40–56.

Xu H, Wu X, Spitz MR, Shete S. 2004. Comparison of haplotype inference methods using genotypic data from unrelated individuals. Hum Hered 58:63–68.

Xu J, Yang Y, Ying Z, Ott J. 2008. Testing linkage disequilibrium from pooled DNA: a contingency table perspective. Stat Med 27: 5801–5815.

Yu W, Ned R, Wulf A, Liu T, Khoury MJ, Gwinn M. 2009. The need for genetic variant naming standards in published abstracts of human genetic association studies. BMC Res Notes 2:56.

Zaitlen N, Kang HM, Eskin E, Halperin E. 2007. Leveraging the HapMap correlation structure in association studies. Am J Hum Genet 80:683–691.

Zhou XH, Hu N, Hu G, Root M. 2009. Synthesis analysis of regression models with a continuous outcome. Stat Med 28:1620–1635.

APPENDIX A

The covariance between two logORs can be derived by treating the observed counts in each 2×3 table representing a single study, as independent Poisson variables with $E[Y_i] = \text{var}[Y_i] = Y_i$ and the logORs as contrasts among the log counts, then using the delta-method we can compute the variance by:

$$\text{var}[f(Y)] \approx \text{var}\,(Y)\left(\frac{\partial f(E[Y])}{\partial (E[Y])}\right)^2 \tag{A1}$$

Thus:

$$\begin{aligned}\text{var}[\log(Y_i)] &\approx \text{var}(Y)\left(\frac{\partial \log(E[Y_i])}{\partial (E[Y_i])}\right)^2 = Y_i\left(\frac{\partial \log(Y_i)}{\partial Y_i}\right)^2 \\ &= Y_i\left(\frac{1}{Y_i}\frac{\partial Y_i}{\partial Y_i}\right)^2 = Y_i\left(\frac{1}{Y_i}\right)^2 = \frac{1}{Y_i}\end{aligned} \tag{A2}$$

For calculating the covariances, we also make use of the following properties of the covariance function:

$$\begin{aligned}\text{cov}(aX + bY,\ cW + dV) = ac\text{cov}(X, W) + ad\text{cov}(X, V) \\ + bc\text{cov}(Y, W) + bd\text{cov}(Y, V)\end{aligned} \tag{A3}$$

$$\text{cov}(Y, Y) = \text{var}\,(Y) \text{ and} \tag{A4}$$

$$\text{cov}(Y_i, Y_j) = 0, \forall i \neq j \text{ for each } 2 \times 3 \text{ table} \tag{A5}$$

Thus, concerning the covariance between y_{1i} and y_{2i} we will have:

$$\begin{aligned}\text{cov}(y_{1i}, y_{3i}) &= \text{cov}\left(\log\left(\frac{A^1B^1_{1i}A^1A^1_{0i}}{A^1A^1_{1i}A^1B^1_{0i}}\right),\ \log\left(\frac{A^2B^2_{1i}A^2A^2_{0i}}{A^2A^2_{1i}A^2B^2_{0i}}\right)\right) \\ &= \text{cov}(\log(A^1B^1_{1i}A^1A^1_{0i}) - \log(A^1A^1_{1i}A^1B^1_{0i}), \\ &\quad \log(A^2B^2_{1i}A^2A^2_{0i}) - \log(A^2A^2_{1i}A^2B^2_{0i}))\end{aligned} \tag{A6}$$

which reduces to:

$$\begin{aligned}&= \text{cov}(\log(A^1B^1_{1i}A^1A^1_{0i}), \log(A^2B^2_{1i}A^2A^2_{0i})) \\ &\quad - \text{cov}(\log(A^1B^1_{1i}A^1A^1_{0i}), \log(A^2A^2_{1i}A^2B^2_{0i})) \\ &\quad - \text{cov}(\log(A^1A^1_{1i}A^1B^1_{0i}), \log(A^2B^2_{1i}A^2A^2_{0i})) \\ &\quad + \text{cov}(\log(A^1A^1_{1i}A^1B^1_{0i}), \log(A^2A^2_{1i}A^2B^2_{0i})) \\ &= \text{cov}(\log(A^1B^1_{1i}) + \log(A^1A^1_{0i}), \log(A^2B^2_{1i}) + \log(A^2A^2_{0i})) \\ &\quad - \text{cov}(\log(A^1B^1_{1i}) + \log(A^1A^1_{0i}), \log(A^2A^2_{1i}) + \log(A^2B^2_{0i})) \\ &\quad - \text{cov}(\log(A^1A^1_{1i}) + \log(A^1B^1_{0i}), \log(A^2B^2_{1i}) + \log(A^2A^2_{0i})) \\ &\quad + \text{cov}(\log(A^1A^1_{1i}) + \log(A^1B^1_{0i}), \log(A^2A^2_{1i}) + \log(A^2B^2_{0i})) \\ &= \text{cov}(\log(A^1B^1_{1i}), \log(A^2B^2_{1i})) + \text{cov}(\log(A^1A^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad - \text{cov}(\log(A^1B^1_{1i}), \log(A^2A^2_{1i})) - \text{cov}(\log(A^1A^1_{0i}), \log(A^2B^2_{0i})) \\ &\quad - \text{cov}(\log(A^1A^1_{1i}), \log(A^2B^2_{1i})) - \text{cov}(\log(A^1B^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad + \text{cov}(\log(A^1A^1_{1i}), \log(A^2A^2_{1i})) + \text{cov}(\log(A^1B^1_{0i}), \log(A^2B^2_{0i}))\end{aligned} \tag{A7}$$

Similarly, we will have for the covariance between y_{2i} and y_{4i}:

$$\text{cov}(y_{2i}, y_{4i}) = \text{cov}\left(\log\left(\frac{B^1B^1_{1i}A^1A^1_{0i}}{A^1A^1_{1i}B^1B^1_{0i}}\right),\ \log\left(\frac{B^2B^2_{1i}A^2A^2_{0i}}{A^2A^2_{1i}B^2B^2_{0i}}\right)\right) \tag{A8}$$

$$\begin{aligned}&= \text{cov}(\log(B^1B^1_{1i}), \log(B^2B^2_{1i})) + \text{cov}(\log(A^1A^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad - \text{cov}(\log(B^1B^1_{1i}), \log(A^2A^2_{1i})) - \text{cov}(\log(A^1A^1_{0i}), \log(B^2B^2_{0i})) \\ &\quad - \text{cov}(\log(A^1A^1_{1i}), \log(B^2B^2_{1i})) - \text{cov}(\log(B^1B^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad + \text{cov}(\log(A^1A^1_{1i}), \log(A^2A^2_{1i})) + \text{cov}(\log(B^1B^1_{0i}), \log(B^2B^2_{0i}))\end{aligned} \tag{A9}$$

For the covariance between y_{2i} and y_{3i} we will have:

$$\text{cov}(y_{2i}, y_{3i}) = \text{cov}\left(\log\left(\frac{B^1B^1_{1i}A^1A^1_{0i}}{A^1A^1_{1i}B^1B^1_{0i}}\right),\ \log\left(\frac{A^2B^2_{1i}A^2A^2_{0i}}{A^2A^2_{1i}A^2B^2_{0i}}\right)\right) \tag{A10}$$

$$\begin{aligned}&= \text{cov}(\log(B^1B^1_{1i}), \log(A^2B^2_{1i})) + \text{cov}(\log(A^1A^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad - \text{cov}(\log(B^1B^1_{1i}), \log(A^2A^2_{1i})) - \text{cov}(\log(A^1A^1_{0i}), \log(A^2B^2_{0i})) \\ &\quad - \text{cov}(\log(A^1A^1_{1i}), \log(A^2B^2_{1i})) - \text{cov}(\log(B^1B^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad + \text{cov}(\log(A^1A^1_{1i}), \log(A^2A^2_{1i})) + \text{cov}(\log(B^1B^1_{0i}), \log(A^2B^2_{0i}))\end{aligned} \tag{A11}$$

And, finally for the covariance between y_{1i} and y_{4i} we will have:

$$\text{cov}(y_{1i}, y_{4i}) = \text{cov}\left(\log\left(\frac{A^1B^1_{1i}A^1A^1_{0i}}{A^1A^1_{1i}A^1B^1_{0i}}\right),\ \log\left(\frac{B^2B^2_{1i}A^2A^2_{0i}}{A^2A^2_{1i}B^2B^2_{0i}}\right)\right) \tag{A12}$$

$$\begin{aligned}&= \text{cov}(\log(A^1B^1_{1i}), \log(B^2B^2_{1i})) + \text{cov}(\log(A^1A^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad - \text{cov}(\log(A^1B^1_{1i}), \log(A^2A^2_{1i})) - \text{cov}(\log(A^1A^1_{0i}), \log(B^2B^2_{0i})) \\ &\quad - \text{cov}(\log(A^1A^1_{1i}), \log(B^2B^2_{1i})) - \text{cov}(\log(A^1B^1_{0i}), \log(A^2A^2_{0i})) \\ &\quad + \text{cov}(\log(A^1A^1_{1i}), \log(A^2A^2_{1i})) + \text{cov}(\log(A^1B^1_{0i}), \log(B^2B^2_{0i}))\end{aligned} \tag{A13}$$

The covariances between log-counts from polymorphism 1 and 2 will be non-zero and at this point we can make use of the delta function to calculate it:

$$\text{cov}[\log(Z_i),\ \log(X_i)] = \text{cov}(Z_i, X_i)\frac{\partial \log Z_i}{\partial Z_i}\frac{\partial \log X_i}{\partial X_i} \quad \text{(A14)}$$

For instance, using the above result we will have:

$$\text{cov}[\log(A^1B^1_{1i}),\ \log(A^2B^2_{1i})] = \frac{\text{cov}(A^1B^1_{1i},\ A^2B^2_{1i})}{(A^1B^1_{1i})\,(A^2B^2_{1i})} \quad \text{(A15)}$$

and similarly for the other genotypes, for cases and controls. The covariance $\text{cov}(A^1B^1_{1i}, A^2B^2_{1i})$ can be calculated by:

$$\begin{aligned}
&\text{cov}(A^1B^1_{1i}, A^2B^2_{1i}) \\
&= \text{cov}\begin{pmatrix} A^1B^1_{1i}A^2B^2_{1i} + A^1B^1_{1i}A^2A^2_{1i} + A^1B^1_{1i}B^2B^2_{1i}, \\ \\ A^1A^1_{1i}A^2B^2_{1i} + A^1B^1_{1i}A^2B^2_{1i} + B^1B^1_{1i}A^2B^2_{1i} \end{pmatrix} \\
&= \text{cov}(A^1B^1A^2B^2_{1i},\ A^1B^1A^2B^2_{1i}) \\
&= \text{var}(A^1B^1A^2B^2_{1i}) \\
&= E(A^1B^1A^2B^2_{1i}) \\
&= A^1B^1A^2B^2_{1i} \quad \text{(A16)}
\end{aligned}$$

Intuitively, the covariance between two marginal genotype counts for polymorphism 1 and 2 is equal to the count of persons carrying both genotypes. Similar calculations can be made when we assume a particular genetic model of inheritance. For instance, if we assume a dominant genetic model for both genes we will have the logarithm of OR[1] (the OR of heterozygous vs. homozygous for the wild type concerning gene 1) to be given by $y_{1i} = \log\left(\frac{(A^1B^1_{1i}+B^1B^1_{1i})\,A^1A^1_{0i}}{A^1A^1_{1i}(A^1B^1_{0i}+B^1B^1_{0i})}\right)$ with an approximate variance calculated by:

$$\begin{aligned}
s^2_{1i} &= 1/A^1A^1_{0i} + 1/A^1A^1_{1i} + 1/(A^1B^1_{0i} + B^1B^1_{0i}) \\
&\quad + 1/(A^1B^1_{1i} + B^1B^1_{1i})
\end{aligned}$$

Similarly, the corresponding ORs with their respective variances, for gene 2, will be:

$$\begin{aligned}
y_{2i} &= \log\left(\frac{(A^2B^2_{1i} + B^2B^2_{1i})\,A^2A^2_{0i}}{A^2A^2_{1i}(A^2B^2_{0i} + B^2B^2_{0i})}\right) \quad \text{with} \\
s^2_{2i} &= 1/A^2A^2_{0i} + 1/A^2A^2_{1i} + 1/(A^2B^2_{0i} + B^2B^2_{0i}) \\
&\quad + 1/(A^2B^2_{1i} + B^2B^2_{1i})
\end{aligned}$$

In this case, we will have a bivariate model of y_{1i} and y_{2i}, with covariance given by:

$$\begin{aligned}
\text{cov}(y_{1i}, y_{2i}) = \text{cov}\Bigg[&\log\left(\frac{(A^1B^1_{1i} + B^1B^1_{1i})A^1A^1_{0i}}{A^1A^1_{1i}(A^1B^1_{0i} + B^1B^1_{0i})}\right), \\
&\log\left(\frac{(A^2B^2_{1i} + B^2B^2_{1i})A^2A^2_{0i}}{A^2A^2_{1i}(A^2B^2_{0i} + B^2B^2_{0i})}\right)\Bigg] \quad \text{(A17)}
\end{aligned}$$

$$\begin{aligned}
&= \text{cov}\begin{bmatrix} \log((A^1B^1_{1i}+B^1B^1_{1i})A^1A^1_{0i}) - \log(A^1A^1_{1i}(A^1B^1_{0i}+B^1B^1_{0i})), \\ \\ \log((A^2B^2_{1i}+B^2B^2_{1i})A^2A^2_{0i}) - \log(A^2A^2_{1i}(A^2B^2_{0i}+B^2B^2_{0i})) \end{bmatrix} \\
&= \text{cov}[\log((A^1B^1_{1i}+B^1B^1_{1i})A^1A^1_{0i}), \log((A^2B^2_{1i}+B^2B^2_{1i})A^2A^2_{0i})] \\
&\quad - \text{cov}[\log(A^1A^1_{1i}(A^1B^1_{0i}+B^1B^1_{0i})), \log((A^2B^2_{1i}+B^2B^2_{1i})A^2A^2_{0i})] \\
&\quad - \text{cov}[\log((A^1B^1_{1i}+B^1B^1_{1i})A^1A^1_{0i}), \log(A^2A^2_{1i}(A^2B^2_{0i}+B^2B^2_{0i}))] \\
&\quad + \text{cov}[\log(A^1A^1_{1i}(A^1B^1_{0i}+B^1B^1_{0i})), \log(A^2A^2_{1i}(A^2B^2_{0i}+B^2B^2_{0i}))] \\
&= \text{cov}[\log(A^1B^1_{1i}+B^1B^1_{1i}), \log(A^2B^2_{1i}+B^2B^2_{1i})] \\
&\quad + \text{cov}[\log(A^2A^2_{0i}), \log(A^1A^1_{0i})] \\
&\quad - \text{cov}[\log(A^1A^1_{1i}), \log(A^2B^2_{1i}+B^2B^2_{1i})] \\
&\quad - \text{cov}[\log(A^2A^2_{0i}), \log(A^1B^1_{0i}+B^1B^1_{0i})] \\
&\quad - \text{cov}[\log(A^1A^1_{0i}), \log(A^2B^2_{0i}+B^2B^2_{0i})] \\
&\quad - \text{cov}[\log(A^1B^1_{1i}+B^1B^1_{1i}), \log(A^2A^2_{1i})] \\
&\quad + \text{cov}[\log(A^1B^1_{0i}+B^1B^1_{0i}), \log(A^2B^2_{0i}+B^2B^2_{0i})] \\
&\quad + \text{cov}[\log(A^1A^1_{1i}), \log(A^2A^2_{1i})] \quad \text{(A18)}
\end{aligned}$$

Here, we will have:

$$\begin{aligned}
&\text{cov}(A^1B^1_{1i} + B^1B^1_{1i}, A^2B^2_{1i} + B^2B^2_{1i}) \\
&\quad = \text{cov}(A^1B^1_{1i}, A^2B^2_{1i}) + \text{cov}(A^1B^1_{1i}, B^2B^2_{1i}) \quad \text{(A19)} \\
&\qquad + \text{cov}(B^1B^1_{1i}, A^2B^2_{1i}) + \text{cov}(B^1B^1_{1i}, B^2B^2_{1i})
\end{aligned}$$

whereas,

$$\begin{aligned}
\text{cov}(A^1A^1_{1i},\ A^2B^2_{1i} + B^2B^2_{1i}) &= \text{cov}(A^1A^1_{1i},\ A^2B^2_{1i}) \\
&\quad + \text{cov}(A^1A^1_{1i},\ B^2B^2_{1i}) \quad \text{(A20)}
\end{aligned}$$

Similarly, using Equation (A16) we can calculate any covariance between two genotype counts. In a similar fashion, we can also calculate the covariance of the logORs when assuming a recessive model of inheritance (not shown). If we assume a co-dominant model of inheritance (allele-based), we will also have:

$$\begin{aligned}
\text{cov}(y_{1i}, y_{2i}) = \text{cov}\Bigg[&\log\left(\frac{(2B^1B^1_{1i} + A^1B^1_{1i})\,(2A^1A^1_{0i} + A^1B^1_{0i})}{(2B^1B^1_{0i} + A^1B^1_{0i})\,(2A^1A^1_{1i} + A^1B^1_{1i})}\right), \\
&\log\left(\frac{(2B^2B^2_{1i} + A^2B^2_{1i})\,(2A^2A^2_{0i} + A^2B^2_{0i})}{(2B^2B^2_{0i} + A^2B^2_{0i})\,(2A^2A^2_{1i} + A^2B^2_{1i})}\right)\Bigg] \\
&= \text{cov}[\log(2B^1B^1_{1i} + A^1B^1_{1i}), \log(2B^2B^2_{1i} + A^2B^2_{1i})] \\
&\quad + \text{cov}[\log(2A^1A^1_{0i} + A^1B^1_{0i}), \log(2A^2A^2_{0i} + A^2B^2_{0i})] \\
&\quad - \text{cov}[\log(2B^1B^1_{1i} + A^1B^1_{1i}), \log(2A^2A^2_{1i} + A^2B^2_{1i})] \\
&\quad - \text{cov}[\log(2A^1A^1_{0i} + A^1B^1_{0i}), \log(2B^2B^2_{0i} + A^2B^2_{0i})] \\
&\quad - \text{cov}[\log(2B^1B^1_{0i} + A^1B^1_{0i}), \log(2A^2A^2_{0i} + A^2B^2_{0i})] \\
&\quad - \text{cov}[\log(2A^1A^1_{1i} + A^1B^1_{1i}), \log(2B^2B^2_{1i} + A^2B^2_{1i})] \\
&\quad + \text{cov}[\log(2B^1B^1_{0i} + A^1B^1_{0i}), \log(2B^2B^2_{0i} + A^2B^2_{0i})] \\
&\quad + \text{cov}[\log(2A^1A^1_{1i} + A^1B^1_{1i}), \log(2A^2A^2_{1i} + A^2B^2_{1i})] \quad \text{(A21)}
\end{aligned}$$

The terms in the right-hand side of Equation (A21) can be calculated by using Equation (A14), for instance by:

$$\begin{aligned}&\mathrm{cov}(2B^1B^1_{1i}+A^1B^1_{1i},2B^2B^2_{1i}+A^2B^2_{1i})\\&\quad=\mathrm{cov}(2B^1B^1_{1i},\ 2B^2B^2_{1i})+\mathrm{cov}(A^2B^2_{1i},\ A^1B^1_{1i})\\&\qquad+\mathrm{cov}(A^1B^1_{1i},\ 2B^2B^2_{1i})+\mathrm{cov}(2B^1B^1_{1i},A^2B^2_{1i})\\&\quad=4\mathrm{cov}(B^1B^1_{1i},\ B^2B^2_{1i})+\mathrm{cov}(A^2B^2_{1i},\ A^1B^1_{1i})\\&\qquad+2\mathrm{cov}(A^1B^1_{1i},\ B^2B^2_{1i})+2\mathrm{cov}(B^1B^1_{1i},\ A^2B^2_{1i})\end{aligned}\tag{A22}$$

and similar for the other terms.

APPENDIX B

The pairwise covariances in the case of continuous outcomes can be calculated by expressing the mean value of the continuous outcome Y for persons carrying a particular marginal genotype (i.e. A^1A^1), as the weighted average of the mean values of the persons carrying all composite genotypes that include this marginal (i.e. $A^1A^1A^2A^2$, $A^1A^1A^2B^2$ and $A^1A^1B^2B^2$) with weights given by the combined genotype counts that we have already computed. For instance, in order to calculate:

$$\begin{aligned}\mathrm{cov}(y_{2i},y_{4i})=\mathrm{cov}(Y_{B^1B^1_i},Y_{B^2B^2_i})-\mathrm{cov}(Y_{B^1B^1_i},Y_{A^2A^2_i})\\-\mathrm{cov}(Y_{A^1A^1_i},Y_{B^2B^2_i})+\mathrm{cov}(Y_{A^1A^1_i},Y_{A^2A^2_i})\end{aligned}\tag{B1}$$

we need to calculate four covariances in the right-hand side. In case of $\mathrm{cov}(Y_{B^1B^1_i},Y_{B^2B^2_i})$, we will have:

$$Y_{B^1B^1_i}=\frac{(B^1B^1A^2A^2_i)Y_{B^1B^1A^2A^2_i}+(B^1B^1A^2B^2_i)Y_{B^1B^1_iA^2B^2_i}+(B^1B^1B^2B^2_i)Y_{B^1B^1B^2B^2_i}}{(B^1B^1A^2A^2_i+B^1B^1A^2B^2_i+B^1B^1B^2B^2_i)}\tag{B2}$$

$$Y_{B^2B^2_i}=\frac{(B^1B^1B^2B^2_i)Y_{B^1B^1B^2B^2_i}+(A^1B^1B^2B^2_i)Y_{A^1B^1_iB^2B^2_i}+(A^1A^1B^2B^2_i)Y_{A^1A^1B^2B^2_i}}{(B^1B^1B^2B^2_i+A^1B^1B^2B^2_i+A^1A^1B^2B^2_i)}\tag{B3}$$

Then:

$$\begin{aligned}\mathrm{cov}(Y_{B^1B^1_i},Y_{B^2B^2_i})&=\mathrm{cov}\left[\begin{array}{l}\dfrac{(B^1B^1A^2A^2_i)Y_{B^1B^1A^2A^2_i}+(B^1B^1A^2B^2_i)Y_{B^1B^1_iA^2B^2_i}+(B^1B^1B^2B^2_i)Y_{B^1B^1B^2B^2_i}}{(B^1B^1A^2A^2_i+B^1B^1A^2B^2_i+B^1B^1B^2B^2_i)},\\ \dfrac{(B^1B^1B^2B^2_i)Y_{B^1B^1B^2B^2_i}+(A^1B^1B^2B^2_i)Y_{A^1B^1_iB^2B^2_i}+(A^1A^1B^2B^2_i)Y_{A^1A^1B^2B^2_i}}{(B^1B^1B^2B^2_i+A^1B^1B^2B^2_i+A^1A^1B^2B^2_i)}\end{array}\right]\\&=\mathrm{cov}\left[\frac{(B^1B^1B^2B^2_i)Y_{B^1B^1B^2B^2_i}}{(B^1B^1)},\ \frac{(B^1B^1B^2B^2_i)Y_{B^1B^1B^2B^2_i}}{(B^2B^2_i)}\right]\\&=\frac{(B^1B^1B^2B^2_i)}{(B^2B^2_i)(B^2B^2_i)}sd^2_{B^1B^1B^2B^2_i}\end{aligned}\tag{B4}$$

Repeating the above procedure for the other terms in the right-hand side of Equation (B1), we will have:

$$\begin{aligned}\mathrm{cov}(y_{2i},y_{4i})=&\frac{(B^1B^1B^2B^2_i)}{(B^1B^1_i)(B^2B^2_i)}sd^2_{B^1B^1B^2B^2_i}+\frac{(A^1A^1A^2A^2_i)}{(A^1A^1_i)\,(A^2A^2_i)}sd^2_{A^1A^1A^2A^2_i}\\&-\frac{(B^1B^1A^2A^2_i)}{(B^1B^1_i)(A^2A^2_i)}sd^2_{B^1B^1A^2A^2_i}-\frac{(A^1A^1B^2B^2_i)}{(A^1A^1_i)\,(B^2B^2_i)}sd^2_{A^1A^1B^2B^2_i}\end{aligned}\tag{B5}$$

Similarly we will have:

$$\begin{aligned}\mathrm{cov}(y_{1i},y_{3i})&=\mathrm{cov}(Y_{A^1B^1_i}-Y_{A^1A^1_i},Y_{A^2B^2_i}-Y_{A^2A^2_i})\\&=\frac{(A^1B^1A^2B^2_i)}{(A^1B^1_i)(A^2B^2_i)}sd^2_{A^1B^1A^2B^2_i}+\frac{(A^1A^1A^2A^2_i)}{(A^1A^1_i)\,(A^2A^2_i)}sd^2_{A^1A^1A^2A^2_i}\\&\quad-\frac{(A^1B^1A^2A^2_i)}{(A^1B^1_i)(A^2A^2_i)}sd^2_{A^1B^1A^2A^2_i}-\frac{(A^1A^1A^2B^2_i)}{(A^1A^1_i)(A^2B^2_i)}sd^2_{A^1A^1A^2B^2_i}\end{aligned}\tag{B6}$$

$$\begin{aligned}\mathrm{cov}(y_{1i},y_{4i})&=\mathrm{cov}(Y_{A^1B^1_i}-Y_{A^1A^1_i},Y_{B^2B^2_i}-Y_{A^2A^2_i})\\&=\frac{(A^1B^1B^2B^2_i)}{(A^1B^1_i)(B^2B^2_i)}sd^2_{A^1B^1B^2B^2_i}+\frac{(A^1A^1A^2A^2_i)}{(A^1A^1_i)\,(A^2A^2_i)}sd^2_{A^1A^1A^2A^2_i}\\&\quad-\frac{(A^1B^1A^2A^2_i)}{(A^1B^1_i)\,(A^2A^2_i)}sd^2_{A^1B^1A^2A^2_i}-\frac{(A^1A^1B^2B^2_i)}{(A^1A^1_i)\,(B^2B^2_i)}sd^2_{A^1A^1B^2B^2_i}\end{aligned}\tag{B7}$$

and

$$\begin{aligned}\mathrm{cov}(y_{2i},y_{3i})&=\mathrm{cov}(Y_{B^1B^1_i}-Y_{A^1A^1_i},Y_{A^2B^2_i}-Y_{A^2A^2_i})\\&=\frac{(B^1B^1A^2B^2_i)}{(B^1B^1_i)\,(A^2B^2_i)}sd^2_{B^1B^1A^2B^2_i}+\frac{(A^1A^1A^2A^2_i)}{(A^1A^1_i)\,(A^2A^2_i)}sd^2_{A^1A^1A^2A^2_i}\\&\quad-\frac{(B^1B^1A^2A^2_i)}{(B^1B^1_i)\,(A^2A^2_i)}sd^2_{B^1B^1A^2A^2_i}-\frac{(A^1A^1A^2B^2_i)}{(A^1A^1_i)\,(A^2B^2_i)}sd^2_{A^1A^1A^2B^2_i}\end{aligned}\tag{B8}$$

APPENDIX C

In order to compute the combined genotypes, we will initially need to estimate the haplotype counts. Table II denotes the cross-tabulation of alleles for polymorphism 1 and 2 and obviously is a 2×2 contingency table with given marginals. If a measure of association is known, the interior cells are uniquely identified [Agresti, 2002]. The unique solution can be found non-iteratively by solving a nonlinear system of five equations with four unknowns. For instance, in study i, the equations

are ($j = 0,1$):

$$A^1A^2_{ji} + B^1A^2_{ji} = 2A^2A^2_{ji} + A^2B^2_{ji} \tag{C1}$$

$$A^1B^2_{ji} + B^1B^2_{ji} = 2A^2A^2_{ji} + A^2B^2_{ji} \tag{C2}$$

$$A^1A^2_{ji} + A^1B^2_{ji} = 2A^1A^1_{ji} + A^1B^1_{ji} \tag{C3}$$

$$B^1A^2_{ji} + B^1B^2_{ji} = 2B^1B^1_{ji} + A^1B^1_{ji} \tag{C4}$$

$$\frac{A^1A^2_{ji}B^1B^2_{ji} - A^1B^2_{ji}B^1A^2_{ji}}{\sqrt{(A^1A^2_{ji}+B^1A^2_{ji})(A^1B^2_{ji}+B^1B^2_{ji})(A^1A^2_{ji}+A^1B^2_{ji})(B^1A^2_{ji}+B^1B^2_{ji})}} = r \tag{C5}$$

Equation (C5) is derived by expressing the correlation coefficient (r) in terms of the haplotype counts (the interior cells of the table). This system of equations can be solved analytically to give:

$$B^1B^2_{ji} = r\frac{\sqrt{(2A^1A^1_{ji} + A^1B^1_{ji})\,(2B^1B^1_{ji} + A^1B^1_{ji})\,(2A^2A^2_{ji} + A^2B^2_{ji})\,(2B^2B^2_{ji} + A^2B^2_{ji})}}{2\,(A^1A^1_{ji} + B^1B^1_{ji} + A^1B^1_{ji})} + \frac{(2A^1A^1_{ji} + A^1B^1_{ji})\,(2B^1B^1_{ji} + A^1B^1_{ji})}{2\,(A^1A^1_{ji} + B^1B^1_{ji} + A^1B^1_{ji})} + \frac{(2B^1B^1_{ji} + A^1B^1_{ji})^2}{2(A^1A^1_{ji} + B^1B^1_{ji} + A^1B^1_{ji})} - \frac{(2B^1B^1_{ji} + A^1B^1_{ji})\,(2A^2A^2_{ji} + A^2B^2_{ji})}{2(A^1A^1_{ji} + B^1B^1_{ji} + A^1B^1_{ji})} \tag{C6}$$

$$A^1A^2_{ji} = 2A^1A^1_{ji} + A^1B^1_{ji} - 2B^2B^2_{ji} - A^2B^2_{ji} + B^1B^2_{ji} \tag{C7}$$

$$A^1B^2_{ji} = 2B^2B^2_{ji} + A^2B^2_{ji} - B^1B^2_{ji} \tag{C8}$$

$$B^1A^2_{ji} = 2B^1B^1_{ji} + A^1B^1_{ji} - B^1B^2_{ji} \tag{C9}$$

Accordingly, the probabilities of observing the four haplotypes are given by:

$$P(A^1A^2_{ji}) = \frac{A^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}} \tag{C10}$$

$$P(A^1B^2_{ji}) = \frac{A^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}} \tag{C11}$$

$$P(B^1A^2_{ji}) = \frac{B^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}} \tag{C12}$$

$$P(B^1B^2_{ji}) = \frac{B^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}} \tag{C13}$$

Given the haplotype probabilities, the probabilities of the combined genotypes can be calculated assuming HWE. Under this assumption, the mating of chromosomes is considered random and thus the probability of a combined genotype is the product of respective haplotype probabilities:

$$P(A^1A^1A^2A^2_{ji}) = P(A^1A^2_{ji}, A^1A^2_{ji}) = P(A^1A^2_{ji})^2 = \left(\frac{A^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)^2 \tag{C14}$$

$$P(A^1A^1B^2B^2_{ji}) = P(A^1B^2_{ji}, A^1B^2_{ji}) = P(A^1B^2_{ji})^2 = \left(\frac{A^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)^2 \tag{C15}$$

$$P(B^1B^1A^2A^2_{ji}) = P(B^1A^2_{ji}, B^1A^2_{ji}) = P(B^1A^2_{ji})^2 = \left(\frac{B^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)^2 \tag{C16}$$

$$P(B^1B^1B^2B^2_{ji}) = P(B^1B^2_{ji}, B^1B^2_{ji}) = P(B^1B^2_{ji})^2 = \left(\frac{B^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)^2 \tag{C17}$$

$$\begin{aligned} P(A^1B^1B^2B^2_{ji}) &= P(A^1B^2_{ji}, B^1B^2_{ji}) + P(B^1B^2_{ji}, A^1B^2_{ji}) \\ &= P(A^1B^2_{ji})P(B^1B^2_{ji}) + P(A^1B^2_{ji})P(B^1B^2_{ji}) \\ &= 2\left(\frac{A^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) \\ &\quad \times \left(\frac{B^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) \end{aligned} \tag{C18}$$

$$\begin{aligned} P(B^1B^1A^2B^2_{ji}) &= P(B^1A^2_{ji}, B^1B^2_{ji}) + P(B^1B^2_{ji}, B^1A^2_{ji}) \\ &= P(B^1A^2_{ji})P(B^1B^2_{ji}) + P(B^1B^2_{ji})P(B^1A^2_{ji}) \\ &= 2\left(\frac{B^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) \\ &\quad \times \left(\frac{B^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) \end{aligned} \tag{C19}$$

$$\begin{aligned} P(A^1B^1A^2A^2_{ji}) &= P(A^2A^2_{ji}, B^1A^2_{ji}) + P(B^1A^2_{ji}, A^2A^2_{ji}) \\ &= P(A^2A^2_{ji})P(B^1A^2_{ji}) + P(A^2A^2_{ji})P(B^1A^2_{ji}) \\ &= 2\left(\frac{A^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) \\ &\quad \times \left(\frac{B^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) \end{aligned} \tag{C20}$$

$$
\begin{aligned}
P(A^1A^1A^1B^2_{ji}) &= P(A^1A^2_{ji}, A^1B^2_{ji}) + P(A^1B^2_{ji}, A^1A^2_{ji}) \\
&= P(A^1A^2_{ji})\, P(A^1B^2_{ji}) + P(A^1B^2_{ji})\, P(A^1A^2_{ji}) \\
&= 2\left(\frac{A^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)\left(\frac{A^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)
\end{aligned}
\tag{C21}
$$

$$
\begin{aligned}
P(A^1B^1A^2B^2_{ji}) &= P(A^1A^2_{ji}, B^1B^2_{ji}) + P(A^1B^2_{ji}, B^1A^2_{ji}) + P(B^1A^2_{ji}, A^1B^2_{ji}) + P(B^1B^2_{ji}, A^1A^2_{ji}) \\
&= P(A^1A^2_{ji})P(B^1B^2_{ji}) + P(A^1B^2_{ji})P(B^1A^2_{ji}) + P(B^1A^2_{ji})P(A^1B^2_{ji}) + P(B^1B^2_{ji})P(A^1A^2_{ji}) \\
&= 2\left(\frac{A^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)\left(\frac{B^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) + 2\left(\frac{A^1B^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right) \\
&\quad \times \left(\frac{B^1A^2_{ji}}{2A^1A^1_{ji} + 2A^1B^1_{ji} + 2B^1B^1_{ji}}\right)
\end{aligned}
\tag{C22}
$$