Original Paper

Human Heredity

Hum Hered 2012;73:35–46
DOI: 10.1159/000331219

Received: March 13, 2011
Accepted after revision: July 15, 2011
Published online: January 18, 2012

On the Meta-Analysis of Genome-Wide Association Studies: A Robust and Efficient Approach to Combine Population and Family-Based Studies

Sungho Won[a, b] Qing Lu[c] Lars Bertram[d, e] Rudolph E. Tanzi[d]
Christoph Lange[f–j]

[a]Department of Applied Statistics, and [b]The Research Center for Data Science, Chung-Ang University, Seoul, Republic of Korea; [c]Department of Epidemiology, Michigan State University, East Lansing, Mich., [d]Genetics and Aging Research Unit, Department of Neurology, Massachusetts General Hospital, Charlestown, Mass., USA; [e]Neuropsychiatric Genetics Group, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany; [f]Department of Biostatistics, Harvard School of Public Health, [g]Harvard Medical School, and [h]Center for Genomic Medicine, Brigham and Women's Hospital, Boston, Mass., USA; [i]Institute for Genomic Mathematics, University of Bonn, and [j]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Key Words
Meta-analysis · Genome-wide study · Population stratification

Abstract
For the meta-analysis of genome-wide association studies, we propose a new method to adjust for the population stratification and a linear mixed approach that combines family-based and unrelated samples. The proposed approach achieves similar power levels as a standard meta-analysis which combines the different test statistics or p values across studies. However, by virtue of its design, the proposed approach is robust against population admixture and stratification, and no adjustments for population admixture and stratification, even in unrelated samples, are required. Using simulation studies, we examine the power of the proposed method and compare it to standard approaches in the meta-analysis of genome-wide association studies. The practical features of the approach are illustrated with a meta-analysis of three genome-wide association studies for Alzheimer's disease. We identify three single nucleotide polymorphisms showing significant genome-wide association with affection status. Two single nucleotide polymorphisms are novel and will be verified in other populations in our follow-up study.

Introduction

Over the last several years, genome-wide association studies (GWAS) have become one of the most important and popular tools in the process of gene mapping. While at the beginning their inherent challenges from their costs, data management, cleaning and analysis prohibited their large-scale use, GWAS are now generally feasible for most diseases and phenotypes, and have been extensively used by the scientific community. They have led to the identification of many novel associations between genetic loci and complex diseases/phenotypes, which have been replicated consistently and reliably in other studies.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/12/0731–0035$38.00/0

Accessible online at:
www.karger.com/hhe

Sungho Won
Department of Applied Statistics
Chung-Ang University
Seoul 156-756 (Republic of Korea)
Tel. +82 2 820 5217, E-Mail swon@cau.ac.kr

However, despite their success in terms of the discovery of novel genetic associations, GWAS have not identified the genetic associations that explain the majority of the expected genetic heritabilities for most traits. For example, genes related to height found by GWAS so far only explain 6–7% of the total variability in height [1], even though genetic heritability is believed to be between 70 and 90%. Therefore, there has been a concerted effort to utilize all available GWAS for a particular phenotype of interest by doing a large-scale meta-analysis in order to maximize the statistical power to identify novel disease loci.

Of course, a meta-analysis of several studies at a genome-wide level raises some statistical issues. One is that the same markers should be genotyped in all studies, which is not the case in most scenarios, since different genotyping platforms are likely to have been used in each study. This problem can be addressed by using imputation techniques. Imputation algorithms infer the genotypes at the untyped marker loci using both the linkage disequilibrium (LD) information about the genomic region from a reference population, e.g. HapMap, or other genotypes at the typed marker loci in the same LD region in the same study. Currently, there are a number of efficient imputation approaches [2–4]. For the purpose of this paper, we assume that the same single nucleotide polymorphisms (SNPs) are genotyped in all studies of the meta-analysis.

A second issue is the genetic and phenotypic heterogeneity introduced by the inclusion of different studies into the meta-analysis. Typically, large-scale studies in such meta-analysis projects have been collected at different time points and locations. At the same time, the ascertainment condition defined by phenotype often varies. Since it is very difficult to identify all differences between studies and account for them in the analysis, the meta-analysis results become susceptible to confounding and their control is a major issue in the analysis. The total sample size in a large-scale meta-analysis project can easily exceed more than 10 or 20 thousand individuals, and for such sample sizes, even small imbalances due to population structure can result in a systematic bias in the results of the association tests [5]. While there are several approaches to adjust for population structure in unrelated samples, none of the approaches are able to achieve the complete robustness of family-based design. For example, the differential level of population structure along the genome or its complicated structure can violate the validity of the statistic [6]. For each approach, examples have been reported in the literature in which the approach fails to adjust correctly, and provides either false positive or false negative results [7–10].

Here we propose a new method for the meta-analysis of GWAS that combines both family-based and unrelated samples with rank-based statistics. Other methods such as genomic control and EIGENSTRAT can be sensitive to confounding due to population admixture or stratification for some scenarios. However, the proposed approach for the meta-analysis that consists of both family-based and unrelated samples is completely robust in the sense that the type-1 error does not increase under violations of the model. At the same time, to prevent the proposed method from losing some efficiency, both a principal component analysis (PCA) and linear mixed model are applied. Because PCA and linear mixed model are not necessary for the complete robustness against any confounding, they can be selectively utilized to increase the efficiency for meta-analysis. Using simulation studies, the power and the robustness of the new approach is assessed and compared to standard meta-analysis methods. We illustrate the approach by an application to Alzheimer's disease.

Methods

For simplicity, we assume that all studies that are included in the meta-analysis, regardless of whether samples are family-based or unrelated, have the same phenotypic ascertainment condition. If the phenotypic ascertainment condition is different, the inference by pooled samples can generate substantial false negative results and stratified analysis for each study has to be applied. Depending on the phenotype of interest, the test statistics can be specified so that binary traits, quantitative traits, multivariate traits or time-to-onset can be tested for association with a genetic marker locus [11–14].

Since, as we will see below, the proposed meta-analysis approach inherits the complete robustness of family-based samples against population structure, a pooled meta-analysis where all subjects from the studies are incorporated to a single statistic becomes feasible. The principle that guards the proposed meta-analysis method against genetic confounding also protects it against potential confounding due to study heterogeneity. For the proposed method, first, the population structure between or within studies is adjusted with the eigenvectors from PCA and the linear mixed model with the estimated PC scores as covariates is applied to each study. Second, the distribution-based p values from each study are transformed to the rank-statistic [6, 15] and combined to form an overall statistic. It should be noted that the first step is conducted to maximize the statistical power because the rank-statistic in the second step by itself can lose the efficiency without the first step, even though it always guarantees the robustness against the presence of population structure. The details for each step will be provided in the following subsection.

We assume that K GWAS with m SNPs are combined for meta-analysis. In order to keep the notation simple, we assume here that all related subjects are trios, i.e. for each study subject, its genotype/phenotype and the parental genotype/phenotype are known. The same statistics simply extend to large pedigrees. There is a large amount of correlations between family members in a large pedigree and the power improvement for the proposed method becomes larger compared to our previous suggestions [6, 15]. If the k-th study subjects are recruited as part of the family-based design, we let the genotype of the i-th marker locus in the offspring of the j-th trio be $x_{j,ik}$ and the phenotype be $y_{j,k}$. The subject's parental genotypes and phenotypes are denoted by $p_{1j,ik}$ ($p_{2j,ik}$) and $q_{1j,ik}$ ($q_{2j,ik}$), respectively. In the families for which the parental information is missing and paired (or more) sibling studies are available without parental information, the parental genotypes can be replaced by the sufficient statistic for pedigrees [16].

Within-Family Component and Between-Family Component

In the subject of the family-based samples, the information about the genetic association can be partitioned at the i-th marker locus into the within-family component and the between-family component [17]. At the within-family level, the evidence for association between the i-th marker and the phenotype of interest is summarized by the statistic $FBAT_i$ that is computed based on the subject's genotype, conditioning on its phenotype and the parental genotypes. The statistics for the within-family level are robust against the population structure. However, the information about the association at the between-family level is sensitive to the population structure as in the statistic for unrelated samples. Therefore, the information about the association at the between-family level is directly included into test statistic T_i for unrelated samples, using offspring's phenotype and its expected genotype which is calculated by Mendelian transmission from the parents to the offspring. This utilization of the proband's phenotype and expected genotype is equivalent to the construction of a VanSteen-type test statistic that is statistically independent of the FBAT statistic under the null hypothesis [12, 18, 19]. We will denote both the unrelated samples and the between-family component by the population-based component in the remainder of this paper.

Principal Component Analysis and Linear Mixed Model

We provide a new method to adjust population stratification in related subjects by extending the EIGENSTRAT approach [9]. However, this approach can also be applied to the population-based component for T_i. Parents in each trio contain the whole information for population structure because the genotypes of nonfounder are transmitted from the founder with Mendelian transmission. We consider only parents, and the computational intensity has been largely improved without any loss of information. Also inclusion of nonfounder for PCA can make spurious PC scores that do not reflect the true underlying population stratification [20]. In the first step, we select/calculate the genotypes/sufficient statistic for parents and they are used to estimate the correlation matrix between individuals as follows:

• For each trio, only parental genotypes from each family are selected to calculate the correlation matrix. If parental genotypes are unknown, the expectations of sufficient statistics (see Rabinowitz and Laird [16]) are utilized.

• PCA is applied to the correlation matrix calculated from the selected genotypes/expectation of the sufficient statistics by the same way as EIGENSTRAT.

• From the PCA, L eigenvectors corresponding to the top large L eigenvalues are considered to adjust the population stratification. L can be decided by considering the relative proportion of each eigenvalue to the total variability or Tracy-Widom statistic [21]. The effect of misspecified L has been shown in the literature [9].

We let $pc_{1jl,k}$ and $pc_{2jl,k}$ be the l-th PC scores for the parents and $pc_{3jl,k}$ for offspring, where $l = 1, ..., L$. $pc_{1jl,k}$ and $pc_{2jl,k}$ are calculated from the correlation matrix, and $pc_{3jl,k}$ is calculated as $0.5(pc_{1jl,k} + pc_{2jl,k})$ because the common diseases are generated by several different genes, and recombination and Mendelian transmission make this relationship preserved. For the j-th trio at the i-th SNP we let

$$X_{j,ik} = (p_{1j,ik}, p_{2j,ik}, x_{j,ik})^t$$
$$Y_{j,k} = (q_{1j,k}, q_{2j,k}, y_{j,k})^t$$
$$PC_{jl,k} = (pc_{1jl,k}, pc_{2jl,k}, pc_{3jl,k})^t$$
$$1_3 = (1, 1, 1)^t.$$

We apply the weighted linear regression with correlation matrix R [22] as follows:

$$X_{j,ik} = \alpha_0 1_3 + \alpha_1 PC_{j1,k} + \cdots + \alpha_L PC_{jL,k} + \gamma_{j,ik},$$
$$\gamma_{j,ik} = \left(\gamma_{1j,ik}, \gamma_{2j,ik}, \gamma_{3j,ik}\right)^t \sim N\left(0, \sigma_i^2 R\right)$$
$$R = \begin{pmatrix} 1 & 2\pi_{12} & 2\pi_{13} \\ 2\pi_{21} & 1 & 2\pi_{23} \\ 2\pi_{31} & 2\pi_{32} & 1 \end{pmatrix},$$

where $\pi_{dd'}$ is the kinship coefficient between the individuals d and d'. If we let the residuals from the above regression for parents and offspring be $p^r_{1j,ik}$ ($p^r_{2j,ik}$) and $x^r_{j,ik}$, respectively, and denote

$$X^r_{j,ik} = (p^r_{1j,ik}, p^r_{2j,ik}, x^r_{j,ik})^t,$$

our final model for trio in the k-th study is

$$Y_{j,k} = \beta_0 1_3 + \beta X^r_{j,ik} + \delta_1 PC_{j1,k} + \cdots + \delta_L PC_{jL,k} + \varepsilon_{j,ik},$$
$$\varepsilon_{j,ik} = \left(\varepsilon_{1j,ik}, \varepsilon_{2j,ik}, \varepsilon_{3j,ik}\right)^t \sim N\left(0, V\right).$$

If there are monozygotic twins, the linear regression of $X_{j,ik}$ on PC scores cannot be conducted. Thus a single individual of each monozygotic twin should be used for the regression but the residual is used for both of each monozygotic twin for the regression of $Y_{j,k}$. For V we need to consider the polygenic effects and common environment effects, and it can be achieved with the proposed linear mixed model (see Appendix). We denote the variance from the polygenic effects and common environment effects by σ_g^2 and σ_c^2 respectively. If the genotypes of each marker are not available, they can be imputed with Hidden Markov model [23] or Bayesian approach [24], and then the proposed method can be applied.

Construction of an Overall Association

The proposed PCA-based methods can be violated if the level of population structure varies along the genomic region as was

shown for EIGENSTRAT in the literature [6]. In this reason we extend the proposed method to the hybrid analysis [6] in order to construct an overall association test for the entire pooled/mixed sample that is robust against confounding due to population structure or study heterogeneity. For the population-based component, the proposed PCA and linear mixed model are applied to the population-based component for T_i across all studies in this step and $FBAT_i$ is calculated by using within-family components. It should be noted that either logistic regression or linear mixed model for T_i can be chosen depending on the ascertainment condition. Then we obtain the Z-statistic that corresponds to the p values of $FBAT_i$ and T_i. pT_i denotes the rank-based p value of the genotype coefficient for T_i that is constructed as follows: if the absolute value of T_i is the o-th ordered among the m SNPs, the rank-based p value, pT_i, for T_i, is defined as

$$(o - \delta)/m,$$

where δ indicates the tuning parameter and $\delta = 0.7$ is recommended for the robustness of the proposed p value in 500K GWAS [6]. We denote $pFBAT_i$ as a distribution-based p value of $FBAT_i$ for the i-th marker. For $pFBAT_i$, we use the one-tailed p value with the direction of T_i to improve the statistical efficiency. If the number of markers whose T_i's are larger than or equal to T_i is smaller than $m/2$, then the one-tailed test for positive direction is applied to $FBAT_i$ and otherwise the one-tailed test for negative direction is applied. We let Z_p and $\Phi(\cdot)$ be the p-th quantile and the cumulative normal distribution of the standard normal distribution, respectively. Then, based on the statistical independence of $FBAT_i$ and T_i [12], we can obtain the following *overall association test for the meta-analysis*, Z_i, at locus i for the pooled samples by combining both Z-statistics in a weighted sum:

$$\Phi(w_{FBAT}Z_{pFBAT_i} + w_T Z_{pT_i})$$

where $w_{FBAT}^2 + w_T^2 = 1$ and $w_{FBAT}Z_{pFBAT_i} + w_T Z_{pT_i} \sim N(0, 1)$ under the null hypothesis. The optimal w_{FBAT} and w_T for continuous and binary trait are discussed in our parallel paper [15], and will be used for the proposed method.

As outlined in our parallel paper [6, 15] for family-based association analysis, the overall association test for the meta-analysis will maintain the type-1 error in the absence and presence of systematic confounding as long as the assumption of Mendelian transmission in the family-based studies is preserved under the null hypothesis. The rank-based p values from the population components are always uniformly distributed along the whole genomes and $FBAT_i$ are robust to the population structure. Also, because the population-based components and within-family components are independent, our overall statistic provides the valid type-1 error rate at the significance level α under the absence of the genetic effect [6] because

$$\frac{1}{m}\sum_{i=1}^{m} P\left[\Phi\left(w_{FBAT}Z_{pFBAT_i} + w_T Z_{pT_i}\right) \le \alpha\right] \le \alpha.$$

Depending on the heterogeneity of study design, the meta-analysis using stratified analysis for each study can be conducted to minimize the type-2 error. We call this meta-analysis using unpooled data. We denote $pFBAT_{ik}$ as a distribution-based p value of $FBAT$ for the i-th marker when the k-th study is family based. If the k'-th study uses the population-based component from either the family-based or unrelated samples, the p value pT_{ik}' denotes the rank-based p value of the genotype coefficient for T_{ik}'. Then, our overall Z-statistics is

$$\Phi\left(\sum_{k\in u_1} w_{FBAT,k}Z_{pFBAT_{ik}} + \sum_{k'\in u_2} w_{T,k'}Z_{pT_{ik'}}\right),$$

where u_1 indicates the studies using family sample, and u_2 indicates the studies using population-based components. It can be easily shown by mathematical induction that the independence of statistics from each study at the i-th marker provides

$$\frac{1}{m}\sum_{i=1}^{m} P\left[\Phi\left(\sum_{k\in u_1} w_{FBAT,k}Z_{pFBAT_{ik}} + \sum_{k'\in u_2} w_{T,k'}Z_{pT_{ik'}}\right) \le \alpha\right] \le \alpha.$$

It should be noted that this is also preserved even when there is heterogeneity of studies, or non-normality of phenotypes.

Results

Simulation Studies

The type-1 error and power of the proposed family-based association test was assessed in the absence and in the presence of population stratification. For simplicity, in the simulation studies we consider nuclear families for which both parental genotypes and phenotypes are either known or unknown. To alleviate the computational intensity, T_i for $100K$ markers are calculated under null hypothesis, and they are used to get the rank for pT_i in 100,000 replicates for empirical type-1 error and 5,000 replicates for empirical power.

Assuming Hardy-Weinberg equilibrium, the parental genotypes are generated by drawing from Bernoulli distribution defined by the allele frequencies. The offspring genotypes are obtained by simulated Mendelian transmissions from the parents to the offspring. The phenotypes for each individual are decided by summing the phenotypic mean, μ, polygenic effect, common environmental effect, main genetic effect and random error. The polygenic effect is independently generated from $N(0, \sigma_g^2)$ for parents, and their average is combined with the value randomly sampled from $N(0, 1/2\sigma_g^2)$ for offspring. For the main genetic effect, the genetic effect size parameter a is multiplied with the number of disease allele of each subject. For instance, the main genetic effect for the subject with homozygous disease genotype is $2a$. The random error is generated from $N(0, \sigma_e^2 = 1)$. Under the null hypothesis of no association, the coefficient a will be set to 0. For scenarios in which population stratification is present, we assume that the population stratification is created by the presence of 2 subpopulations and each parent is assigned to either of the 2 subpopulations with 50% probability. μ for parents is either 0 or 0.4 depending on

the subpopulation and their average is used for offspring. The other effects for phenotype are sampled in the same way as in the absence of population stratification. The allele frequencies for each marker in the two subpopulations are generated by the Balding-Nichols model [25]. That is, for each marker, the allele frequency, q, in an ancestral population is generated from a uniform distribution between 0.1 and 0.9, $U(0.1, 0.9)$. Then, if we let F_{ST} be the Wright's measure of population subdivision, the marker allele frequencies for the two subpopulations are independently sampled from the beta distributions $(q(1 - F_{ST})/F_{ST}, (1 - q)(1 - F_{ST})/F_{ST})$. F_{ST} was assumed to be 0.001, 0.005, or 0.01.

At first for various scenarios, we verified that the proposed overall family-based association test maintains the α level. We consider three different rank-based p values for conditional mean model [26] (pT^1, pT^2 and pT^3). pT^1 is the proposed rank statistic from the distribution-based p values by the linear mixed model with adjustment of population stratification. For comparison, pT^2 indicates the rank-based p value from the linear mixed model for conditional mean model without adjustment of population stratification, and simple linear model without adjustment of population stratification is applied in pT^3 where the latter is equivalent to our recent works [6, 15]. Each rank-based p value is combined with *pFBAT* with optimal weights. Z^1, Z^2 and Z^3 correspond to the overall p values that combine *pFBAT* with pT^1, pT^2 and pT^3, respectively. The proposed PCA method and linear mixed model are directly applied to whole families without transformation to rank-statistic and they are denoted as *POP*. *POP* is based on the Wald test and it is approximately equal to the variance component model approaches [27, 28]. We assume that parental phenotype/genotypes are known and 500 nuclear families with N_{off} offspring are available for each replicate. We assume that $\sigma_c^2 = \sigma_g^2 = 0.3$. Table 1 shows the empirical type-1 error estimates for the proposed linear mixed model with or without adjustment of population stratification in the absence ($F_{ST} = 0.00$) and presence of the population stratification for 3 different F_{ST} ($F_{ST} = 0.001$, 0.005 and 0.01). The results from *POP* do not provide any evidence for a departure of the empirical type-1 error estimates, both in the absence and presence of population stratification. That is, the proposed PCA method adjusts the population stratification and the linear mixed model fits the correlation of the family members well. Also all of Z^1, Z^2 and Z^3 preserve the α level and we confirm that the rank-based p value is a single requirement to provide the validity under the presence of population stratification and misspecified covariance matrix.

Table 1. Empirical type-1 error estimates at the 0.001 α level

N_{off}	F_{ST}	*POP*	*FBAT*	Z^1	Z^2	Z^3
1	0	1.10	0.99	0.95	1.07	1.06
	0.001	1.15	0.93	1.03	1.03	0.95
	0.005	1.09	0.80	0.99	1.10	1.03
	0.01	1.03	1.04	1.14	1.02	1.09
2	0	1.05	1.03	1.03	1.08	0.98
	0.001	0.89	0.85	1.01	0.84	0.99
	0.005	1.00	1.12	1.05	0.91	1.01
	0.01	1.11	0.93	0.98	1.03	0.99
3	0	1.03	1.05	0.89	1.05	1.11
	0.001	0.93	0.92	0.95	0.99	1.08
	0.005	0.86	1.12	0.85	0.90	0.89
	0.01	1.25	1.01	1.25	1.16	1.33

Empirical type-1 error at the scale $\times 10^{-3}$ is calculated for different levels of population stratification, F_{ST}, when parental phenotypes and genotypes are known. We assume that $\sigma_g^2 = \sigma_c^2 = 0.3$ and $\sigma_e^2 = 1$, and the disease allele frequency is 0.1. N_{off} means the number of offspring.

In the next set of simulation studies, we assess the empirical power over the different nuclear family structures. Under the assumption of an additive disease model for a quantitative trait, the genetic effect, a, is given as a function of the heritability attributable to main genetic effect, h^2, the disease allele frequency, p_D, and the phenotypic variance, $\sigma^2 = \sigma_c^2 + \sigma_g^2 + \sigma_e^2$, by:

$$a = \sigma \sqrt{\frac{1}{2p_D(1-p_D)} \frac{h^2}{(1-h^2)}}$$

where h^2 indicates the heritability attributable to the main genetic effect. We assume that $p_D = 0.1$ and $\sigma_e^2 = 1$. We assume 500 nuclear families available in our simulations. Table 2 shows the empirical power estimates from 5,000 replicates when parental genotypes/phenotypes are known. The results show that *POP* is the most powerful under both absence and presence of population stratification, followed by Z^1. However, their difference is usually very small. *FBAT* does not use the between-family components and it is the least informative. The results indicate that the hybrid analysis [6, 15] can be improved when the linear model with appropriate correlation structure is applied to conditional mean model and the population stratification is adjusted. For the application of the linear mixed model to the between-family component of the subjects from the family-based designs, the expected ge-

Table 2. Empirical power estimates at the 0.0001 α level

N_{off}	h^2	F_{ST}	*POP*	*FBAT*	Z^1	Z^2	Z^3
1	0.005	0.001	0.1483	0.0030	0.1447	0.1206	0.1328
		0.005	0.1461	0.0016	0.1423	0.1094	0.1354
		0.01	0.1421	0.0044	1.1452	0.0678	0.1390
	0.010	0.001	0.5451	0.0134	0.5326	0.4694	0.5092
		0.005	0.5364	0.0134	0.5255	0.4106	0.5056
		0.01	0.5421	0.0114	0.5389	0.3352	0.5304
2	0.005	0.001	0.2440	0.0136	0.2314	0.1856	0.2124
		0.005	0.2398	0.0116	0.2202	0.1700	0.2080
		0.01	0.2288	0.0154	0.2034	0.1218	0.1890
	0.010	0.001	0.7252	0.0606	0.7106	0.6402	0.6794
		0.005	0.7131	0.0634	0.6855	0.5820	0.6642
		0.01	0.7216	0.0586	0.6893	0.4992	0.6626
3	0.005	0.001	0.3552	0.0322	0.3310	0.2692	0.2948
		0.005	0.3509	0.0344	0.3335	0.2318	0.3042
		0.01	0.3431	0.0342	0.3174	0.1826	0.2760
	0.010	0.001	0.8565	0.1694	0.8257	0.7642	0.7942
		0.005	0.8423	0.1560	0.8239	0.7072	0.7870
		0.01	0.8485	0.1598	0.8205	0.6390	0.7834

Empirical power estimates at the 0.0001 α level are calculated for different level of population stratification, F_{ST}, when parental phenotypes and genotypes are known. We assume that $\sigma_g^2 = \sigma_c^2 = 0.3$ and $\sigma_e^2 = 1$, and the disease allele frequency is 0.1. N_{off} means the number of offspring.

notypes instead of the observed genotypes for offspring are used for the computation of the test statistic T_i. For the between-family component the phenotypic variance of parents, $\mathrm{Var}(q_{1j,ik} \mid p_{1j,ik})$, is different from the phenotypic variance of the subjects from the family-based design, $\mathrm{Var}(y_{j,k} \mid p_{1j,ik}, p_{2j,ik})$, but our linear mixed model to the between-family component assumes that both are same. The test statistic is generally the most efficient when the variance/covariance structure is correctly specified and it may explain the reason for the power loss of Z^1 compared to *POP*.

Power Analysis in an Unpooled Meta-Analysis

We calculate the effect of rank-based p-value transformation on the power. The empirical power at 0.05 genome-wide significance level is calculated from 10,000 replicates for 500K GWAS under Bonferroni correction. We consider the meta-analysis in which one family-based and one unrelated sample are combined, and one family-based and two unrelated samples are combined. When two unrelated samples are used, their sample sizes are assumed to be equal. 1,000 trios for family-based samples are generated with parental phenotypes unknown for both cases. The results from each study are combined with Liptak method using the optimal weights. *POP* is the empirical power estimates when the linear regression is applied to both family-based and unrelated samples. For other statistics, family-based samples are split to between-family and within-family components, and then *FBAT* for within-family component is calculated. For between-family component, the linear regression is used and it is combined with *FBAT*. While the distribution-based p value from between-family component and unrelated samples is combined with *FBAT* in *NONRANK*, the rank-based p value for between-family component and unrelated samples is combined in *RANK*.

For our empirical power estimates, we assume that $h^2 = 0.005$ and $\sigma_c^2 + \sigma_g^2 + \sigma_e^2 = 1$ under the absence of population structure. Because we assume that the parental phenotype is unknown, the identification of σ_c^2, σ_g^2 and σ_e^2 is not required. For disease model, we assume that $p_D = 0.1$ and the genetic effect a is calculated in the same way as in the simulation study for table 2. Table 3 shows the empirical power estimates for each meta-analysis. The results show that *POP* is always the most efficient, followed by *NONRANK* while both are virtually equal. *RANK* is always least efficient. However the power loss is negligible if three or more studies are combined, even when unrelated samples have much larger sample size, compared to family-based design. If the sample sizes for family-based and unrelated samples are not highly different, the power loss of *RANK* is also very small.

Also we consider the meta-analysis with only unrelated samples as an extreme case. In figure 1, we assume the meta-analysis without family-based samples and it shows the relative power of rank-based p value compared to the distribution-based p value. We assume the meta-analysis with 2, 4, 10 and 20 studies, and their Z-statistics for each study are simply generated from $N(0.05, 1)$. Then their p values using both distribution-based and rank-based approaches are calculated respectively, and the Liptak method is applied to combine the p values. We assume GWAS with 1,000 SNPs or 10,000 SNPs, and the relative efficiencies of empirical powers are calculated from 2,000,000 replicates at the genome-wide $\alpha = 0.05$ level. Figure 1 shows that the relative power of the rank-based p values is proportional to the number of studies and SNPs. Our results confirm that the proposed meta-analysis can have more than 90% relative efficiency with more than two GWAS for large-scale genome-wide meta-association analysis. As a result, because *RANK* is a unique choice that preserves the complete robustness to

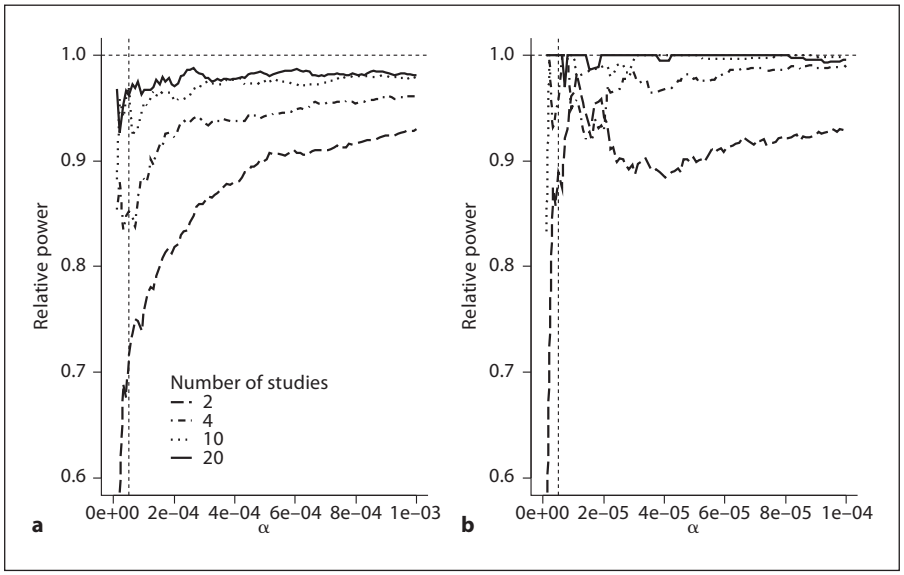

Fig. 1. Relative power of rank-statistics in a meta-analysis. The relative power of the rank-based p values in meta-analysis is compared with the distribution-based p values when different numbers of studies are combined. X-axis indicates the nominal significance level and y-axis indicates the empirical relative power for the given nominal significance level. If the relative efficiency is 1, it indicates that the rank-based p values are equally efficient as the distribution-based p values. The dashed vertical line indicates the 0.05 genome-wide significance level under Bonferroni correction and rank-based p values are calculated for genome-wide studies with 1,000 (**a**) and 10,000 markers (**b**).

any type of violations, such as non-normality, population stratification, etc., we recommend using the rank-based p value for unrelated samples.

Application to Alzheimer Studies

Late-onset Alzheimer's disease is a progressive and fatal brain disease. During the last 30 years, people have tried to localize the Alzheimer's disease susceptibility genes, but except for APOE those efforts have mostly led to inconsistent findings, even though Alzheimer's disease is highly inheritable [29]. To overcome this problem, several 500K GWAS [30–33] have been conducted to identify the causal gene. However, each single GWAS may be under insufficient efficiency and results for each marker need to be confirmed with a robust approach.

In our analysis, we applied the proposed method to one family-based design and the meta-analysis is performed with two additional case-control designs. The family sample was collected as part of the National Institute of Mental Health Genetics Initiative Study and it is denoted as NIMH. For meta-analysis one case-control sample was collected by the Translational Genomics Research Institute and the other case-control sample was collected by GlaxoSmithKline. We denote the former and the latter by TGEN [33] and GSK [32], respectively. Each of these three studies performed whole genome association analysis using 500,668 SNPs on the GeneChip Human Mapping 500K Array Set (Affymetrix, Santa Clara, Calif., USA). For our analysis, if the minor allele frequency is less than 0.05, or the p value from Hardy-Weinberg

Table 3. Empirical power estimates at the 0.05 genome-wide significance level

	1 family based + 1 unrelated			1 family based + 2 unrelated		
N	*POP*	*NORANK*	*RANK*	*POP*	*NORANK*	*RANK*
1,000	0.0118	0.0110	0.0104	0.0421	0.0394	0.0385
1,500	0.0297	0.0282	0.0246	0.1440	0.1381	0.1353
2,000	0.0536	0.0522	0.0400	0.2793	0.2706	0.2667
2,500	0.0973	0.0941	0.0575	0.4540	0.4440	0.4384
3,000	0.1585	0.1532	0.0828	0.6244	0.6135	0.6050
3,500	0.2212	0.2136	0.0942	0.7605	0.7515	0.7396
4,000	0.3106	0.3025	0.1248	0.8699	0.8632	0.8544
4,500	0.3898	0.3812	0.1413	0.9247	0.9217	0.9136
5,000	0.4826	0.4722	0.1555	0.9595	0.9583	0.9538

Empirical power is estimated with 10,000 replicates at the 0.05 genome-wide significance level. The multiple testing problem is addressed with Bonferroni correction. We assume 1,000 trios available for family-based samples with parental phenotypes unknown. *N* denotes the sample size for each unrelated sample. *POP* indicates the empirical power estimates when a simple linear regression is applied to trio and unrelated samples, and the results from each study are combined with Liptak's method with the optimal weights. For *NORANK* and *RANK*, trio data are split to between-family and within-family components, and then *T* and *FBAT* are calculated respectively. While the distribution-based p values for between-family component and unrelated samples are combined with *FBAT* with Liptak's method in *NONRANK*, the rank-based p values for between-family component and unrelated samples are combined in *RANK*.

Table 4. Genome-wide association analysis with NIMH (500K)

SNP	*pFBAT*	*pT*	Overall
SNP1	5.54×10^{-14}	8.32×10^{-7}	2.83×10^{-18}
SNP2	5.63×10^{-5}	2.06×10^{-4}	8.55×10^{-8}
SNP3	2.78×10^{-4}	8.95×10^{-5}	1.79×10^{-7}
SNP4	1.39×10^{-5}	1.90×10^{-3}	2.73×10^{-7}
SNP5	1.58×10^{-5}	1.96×10^{-3}	3.14×10^{-7}
SNP6	3.13×10^{-5}	1.26×10^{-3}	3.39×10^{-7}
SNP7	5.34×10^{-4}	1.01×10^{-4}	3.86×10^{-7}
SNP8	6.41×10^{-3}	3.60×10^{-6}	4.05×10^{-7}
SNP9	3.62×10^{-3}	9.15×10^{-6}	4.13×10^{-7}
SNP10	1.50×10^{-4}	4.11×10^{-4}	4.29×10^{-7}
SNP11	2.41×10^{-5}	2.02×10^{-3}	4.61×10^{-7}
SNP12	1.31×10^{-3}	4.52×10^{-5}	4.89×10^{-7}
SNP13	2.29×10^{-3}	4.24×10^{-5}	8.60×10^{-7}
SNP14	4.04×10^{-3}	2.58×10^{-5}	1.09×10^{-6}
SNP15	3.13×10^{-5}	3.67×10^{-3}	1.14×10^{-6}
SNP16	1.00×10^{-4}	1.69×10^{-3}	1.29×10^{-6}
SNP17	2.49×10^{-6}	1.88×10^{-2}	1.31×10^{-6}
SNP18	2.15×10^{-3}	7.57×10^{-5}	1.31×10^{-6}
SNP19	6.60×10^{-5}	2.93×10^{-3}	1.65×10^{-6}
SNP20	7.32×10^{-6}	1.31×10^{-2}	1.77×10^{-6}

After the quality control, 360,742 SNPs are analyzed and the genome-wide significance level at 0.05 is 1.386×10^{-7}. The p values of the top 20 SNPs are shown.

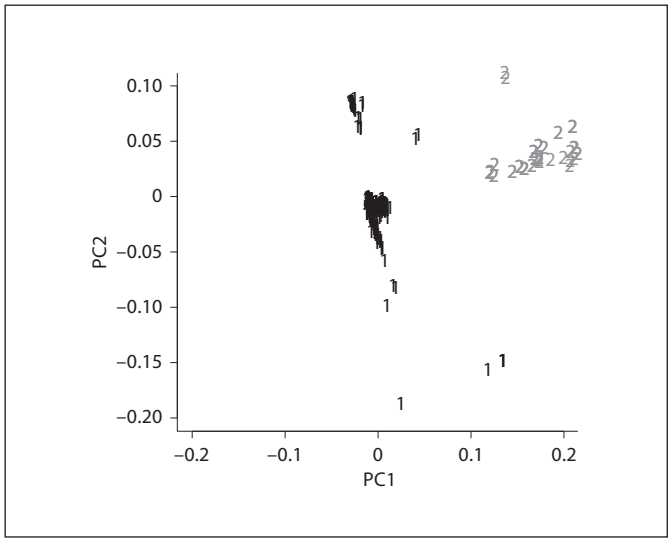

Fig. 2. Population stratification in NIMH. The proposed PCA approach is applied to NIMH. The first PC scores are plotted against the second PC scores. There are two populations, European American and African American. 1 indicates European American and 2 indicates African American.

proportion test is less than 0.0001, these SNPs are excluded from the analysis. The analysis was conducted in the Linux system with 2,412.309 cpu MHz, and the computation took around one week with a single node. The programs for PCA were developed with C++ and will be included in PBAT software.

NIMH consists of two populations, European American and African American, and we can expect that there may be population stratification. *FBAT* is applied to the within-family component of NIMH. The proposed PCA method is conducted to adjust population stratification and the linear regression using PC scores as covariates is applied to the between-family component. It was empirically shown that through the analysis of Phase II HapMap the differential level of linkage disequilibrium does not significantly affect the results [9] and we use all SNPs that passed the quality control. Then the rank-based p value for between-family component and the exact p values of *FBAT* are combined with Liptak's method.

Figure 2 shows the PC scores for NIMH. The first PC score is plotted against the second PC score. We found that two PC scores explained more than 90% of the total variability and chose two PC scores for the analysis. 1,376 individuals from 410 families are self-reported European ancestry and they are labelled with 1 in figure 2. Fifty-eight individuals from 24 families were of African descent and they are labelled with 2. The results show that the proposed method completely identifies African descent and European ancestry, and it seems there may be two subgroups in European ancestry.

Table 4 shows the results from NIMH. 360,742 SNPs in NIMH after quality control are analyzed and the 0.05 genome-wide significance level is 1.386×10^{-7}. We have two genome-wide significant SNPs. The most significant and the second genome-wide significant SNPs are SNP1 and SNP2, respectively. In our paper, their rs numbers are not shown but they will be updated with additional results in our follow-up studies.

Table 5 shows the results from our meta-analysis. The SNPs from SNP1 to SNP20 in table 5 correspond to the SNPs in table 4. For TGEN the PC score is also obtained with EIGENSTRAT and the logistic regression is applied. For GSK Cochran-Armitage test is calculated without application of EIGENSTRAT because only limited data are available. Their distribution-based p values from between-family component for NIMH and two case-control designs are transformed to rank-based p values. The p values from between-family and within-family components in NIMH are combined with Liptak

Table 5. Meta-analysis with rank-based p value (500K)

SNP	NIMH		TGEN	GSK	Overall
	pFBAT	*pT*	*pT*	*pT*	
SNP1	5.54×10^{-14}	2.57×10^{-6}	9.63×10^{-1}	2.57×10^{-6}	$<1.00 \times 10^{-20}$
SNP21	1.13×10^{-2}	1.53×10^{-3}	2.78×10^{-3}	1.54×10^{-3}	1.14×10^{-7}
SNP2	5.63×10^{-5}	2.23×10^{-4}	1.38×10^{-1}	3.90×10^{-2}	1.26×10^{-7}
SNP12	1.31×10^{-3}	4.66×10^{-5}	1.55×10^{-2}	1.02×10^{-1}	1.90×10^{-7}
SNP22	3.47×10^{-4}	1.17×10^{-3}	1.37×10^{-2}	2.67×10^{-2}	2.31×10^{-7}
SNP3	2.78×10^{-4}	9.79×10^{-5}	7.90×10^{-1}	7.42×10^{-3}	2.66×10^{-7}
SNP9	3.62×10^{-3}	1.36×10^{-5}	3.45×10^{-1}	2.01×10^{-2}	8.18×10^{-7}
SNP23	6.41×10^{-3}	3.60×10^{-6}	4.05×10^{-7}	7.96×10^{-5}	1.44×10^{-6}
SNP24	3.62×10^{-3}	9.15×10^{-6}	4.13×10^{-7}	8.40×10^{-1}	1.63×10^{-6}
SNP10	1.50×10^{-4}	4.11×10^{-4}	4.29×10^{-7}	5.56×10^{-2}	2.78×10^{-6}
SNP25	2.41×10^{-5}	2.02×10^{-3}	4.61×10^{-7}	9.90×10^{-6}	3.47×10^{-6}
SNP7	1.31×10^{-3}	4.52×10^{-5}	4.89×10^{-7}	9.39×10^{-2}	3.93×10^{-6}
SNP26	2.29×10^{-3}	4.24×10^{-5}	8.60×10^{-7}	1.79×10^{-1}	4.37×10^{-6}
SNP27	4.04×10^{-3}	2.58×10^{-5}	1.09×10^{-6}	1.11×10^{-2}	4.90×10^{-6}
SNP28	3.13×10^{-5}	3.67×10^{-3}	1.14×10^{-6}	1.46×10^{-3}	5.00×10^{-6}
SNP19	1.00×10^{-4}	1.69×10^{-3}	1.29×10^{-6}	2.74×10^{-1}	6.27×10^{-6}
SNP29	2.49×10^{-6}	1.88×10^{-2}	1.31×10^{-6}	1.18×10^{-3}	6.39×10^{-6}
SNP30	2.15×10^{-3}	7.57×10^{-5}	1.31×10^{-6}	2.19×10^{-1}	6.45×10^{-6}
SNP31	6.60×10^{-5}	2.93×10^{-3}	1.65×10^{-6}	7.28×10^{-4}	7.76×10^{-6}
SNP32	7.32×10^{-6}	1.31×10^{-2}	1.77×10^{-6}	1.74×10^{-1}	8.04×10^{-6}

After the quality control, 272,769 SNPs are available for all three studies and the 0.05 genome-wide significance level is 1.833×10^{-7}. The directions of each statistic are not considered for overall p values. The p values of the top 20 SNPs are shown.

method, and those are combined with the other rank-based p values from TGEN and GSK with Fisher's method. Imputation is not performed for the analysis and 272,769 SNPs after quality control are available for all studies. The 0.05 genome-wide significance level is 1.833×10^{-7}. Our results show that there are three genome-wide significant SNPs, SNP1, SNP21, and SNP2. SNP1 and SNP2 were also significant in table 4. *pT* in NIMH in table 5 are different from those in table 4 and it happens that the number of available SNPs in both tables are different.

In particular, table 5 shows that the results from TGEN are inconsistent with the other studies. For example, SNP1 is very significant in NIMH and GSK while it is not in TGEN. At the same time, the allelic direction in TGEN does not correspond to the others. Because SNP1 is associated with APOE that is generally known to be associated with Alzheimer's disease, the systematic heterogeneity for ascertainment condition may exist in TGEN. The insignificance of SNP1 in TGEN can also be explained with this systematic heterogeneity. If we exclude TGEN from our meta-analysis and consider their directions, the p values for SNP2 and SNP21 are 3.82×10^{-8} and 1.40×10^{-6}, respectively. Thus, SNP2 seems the more promising SNP associated with Alzheimer's disease. This SNP2 is novel and will be verified in the other populations in our follow-up study.

Discussion

In genetic association analysis, it has been known that the presence of population stratification can deteriorate the validity of genetic association and many different methods have been proposed. Under population stratification we can adjust them either by modeling them or by using the robust statistics without any statistical model. We will call the former a model-based approach and the latter a model-free approach. For instance, STRUCTURE and EIGENSTRAT are model-based, and TDT and rank-based p values are model-free approaches. The former can increase the efficiency when the assumed models are correct but otherwise they can lose both efficiency and validity. The model-free approaches are beneficial in

terms of validity, but they are less efficient under the general population stratification model. Our recommendation is to use both approaches, such as PCA-based approach and rank-based p value, to improve the efficiency and guarantee the validity.

The rank-based p value also provides the statistical inference robust against the other types of violations, such as non-normality and the existence of the confounders. This is because the rank-based p values over the whole genome are always uniformly distributed. One useful application is for computations in a family-based association analysis. The covariance structure in family-based samples makes the calculation of the test statistics intensive and in this reason the statistical analysis in general pedigree structures has been limited. However, if the distribution-based p values are transformed to the rank-based p value, we can use the simplified covariance structure. For instance, for a large pedigree we can split each large pedigree into several nuclear families and the proposed linear mixed model for nuclear family can be applied. The slight power loss is expected but it is worthwhile to try because we can accelerate the computation.

However, even though the proposed method provides flexibility of a genetic analysis, there are still some limitations. First, the pooled meta-analysis under the presence of heterogeneity in the study can lead to reduced statistical power of our meta-analysis approach even though the rank-based p values enable the meta-analysis using the pooled samples. In situations in which substantial study heterogeneity is present and known prior to the analysis, test statistics should be calculated for each study separately and combined into overall statistics, using Fisher's method of combining p values or the Liptak approaches of weighted Z-statistics [34, 35]. Second, the proposed method requires the independence of each statistic for the complete validity. For instance, some SNPs can be the first ranked for all studies because of population stratification. In this scenario, the final p value becomes 8.9×10^{-12} for a meta-analysis using two 500K GWAS. Even though it seems uncommon, it is still possible for ancestry informative markers [36]. To prevent this effect, some modification for the proposed method is necessary for the meta-analysis using unpooled samples while the meta-analysis using pooled samples is always fine. Statistics for case-control designs and between-family component from family-based designs are calculated separately and then combined to calculate the summary statistic, such as weighted Z-score. Then the rank-based p values for the summary statistics are combined with the p values from the within-family components in family-based design. This 'double stratification' can lead to the loss of efficiency, but no alternative has been suggested yet in this scenario. Third, if the number of markers is small compared to the sample size or the amount of heritability, our transformation can deteriorate the efficiency. For instance, when 1,000 SNPs are analyzed as was shown in figure 1, the power of rank-statistic can decrease to around 80% compared to the distribution-based p value. For such a number of SNPs, some modification is necessary but the power loss may be negligible in most GWAS. In future work, we will investigate these problems and some extensions will be provided.

Acknowledgements

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government 2010-0023594, and by the Industrial Strategic Technology Development Program (10040176) funded by the Ministry of Knowledge Economy (MKE, Korea). Also this work was supported by the NIMH 1R37MH60009 and Cure Alzheimer's Fund. The authors would like to thank the anonymous referees for their helpful suggestions and comments.

Appendix

Linear Mixed Model for Related Subjects

We provide the linear mixed model to consider the correlation between family members for our simulations. Even though the proposed linear mixed model is provided for our simulation studies, it can be extended to the related subjects in real analysis with popular statistical software such as R and SAS if family structure is not highly unbalanced. Also, when the rank-based p value is applied to T_{ik}, overall statistics, Z_{ik}, is always valid and only efficiency depends on the chosen covariance structure. Thus, if the correlations between family members are expected to be small, it is better to assume the simplified covariance to reduce the computational intensity.

For the correlations between family members we assume that there is no polygenic dominant effect. If the k-th study consists of related subjects, we let polygenic additive effects for two parents and offspring $a_{1j,k}$, $a_{2j,k}$ and $a_{3j,k}$, and common environmental effects be $c_{1j,k}$, $c_{2j,k}$ and $c_{3j,k}$. If we let σ_g^2 and σ_c^2 be their variances for the polygenic additive effects and common environmental effects, it is generally assumed that $c_{1j,k}$, $c_{2j,k}$ and $c_{3j,k}$ are identical and follow $N(0, \sigma_c^2)$. Also $a_{1j,k}, a_{2j,k} \sim N(0, \sigma_g^2)$ and $a_{3j,k} \mid a_{1j,k}, a_{2j,k} \sim N(0.5\,(a_{1j,k} + a_{2j,k}), 0.5\sigma_g^2)$. Thus, if we let $A_{j,k} = (a_{1j,k}, a_{2j,k}, a_{3j,k})^t$ and $C_{j,k} = (c_{1j,k}, c_{2j,k}, c_{3j,k})^t$, then $Y_{j,k}$ can be expressed with random effects as

$$Y_{j,k} = \beta_0 + \beta X^r_{j,ik} + \delta_1 PC_{j1,k} + \ldots + \delta_L PC_{jL,k} + C_{j,k} + A_{j,k} + \varepsilon_{j,ik}$$

where $\varepsilon_{j,ik}$ is a random error vector for the j-th family member in the k-th study. While $C_{j,k}$ can be easily incorporated into the linear

mixed model, $A_{j,k}$ requires some transformations because the probability density function of $A_{j,k}$ is

$$f\left(A_{j,k}\right)=\phi\left(\frac{a_{1j,k}}{\sigma_g}\right)\phi\left(\frac{a_{2j,k}}{\sigma_g}\right)\phi\left(\frac{a_{3j,k}-0.5\left(a_{1j,k}+a_{2j,k}\right)}{\sigma_g/\sqrt{2}}\right),$$

where $\Phi(\cdot)$ is a probability density function for standard normal distribution. If we let I_d be the identity matrix with $d \times d$ dimension, with simple algebra we can show that it is equivalent to the following linear mixed effect model:

$$Y_{j,k} = \beta_0 + \beta X^r_{j,ik} + \delta_1 PC_{j1,k} + \ldots + \delta_L PC_{jL,k} + C_{j,k} + Z^t_{j,k} A'_{j,k} + \varepsilon_{j,ik}$$
$$C_{j,k} \sim N(0, \sigma_c^2 I_3),\ A'_{j,k} \sim N(0, \sigma_g^2 I_3),\ \varepsilon_{j,ik} \sim N(0, \sigma_e^2 I_3)$$

where

$$A'_{j,k} \sim N\left(0, \sigma_g^2 I_3\right),\ \text{and}\ Z_{j,k} = \begin{pmatrix} I_2 & \underline{0}_2 \\ 1/2 \cdot \underline{1}_2^t & 1/\sqrt{2} \end{pmatrix}$$
$$\underline{0}_2 = (0,0)^t,\ \underline{1}_2 = (1,1)^t.$$

As a result, via the proposed transformation we can use the linear mixed model to test family-based association with standard statistical software such as SAS (MIXED, NLMIXED) or R (nlme, lme4). Also, the proposed model can be further extended to binary trait with the generalized linear mixed model.

References

1 Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CN, Morris AD, Ouwehand WH, Zhao JH, Li S, Loos RJ, Barroso I, Deloukas P, Sandhu MS, Wheeler E, Soranzo N, Inouye M, Wareham NJ, Caulfield M, Munroe PB, Hattersley AT, McCarthy MI, Frayling TM: Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 2008;40:575–583.

2 Dai JY, Ruczinski I, LeBlanc M, Kooperberg C: Imputation methods to improve inference in SNP association studies. Genet Epidemiol 2006;30:690–702.

3 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–913.

4 Servin B, Stephens M: Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 2007;3:e114.

5 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512–517.

6 Won S, Wilk JB, Mathias RA, O'Donnell CJ, Silverman EK, Barnes K, O'Connor GT, Weiss ST, Lange C: On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. PLoS Genet 2009;5:e1000741.

7 Epstein MP, Allen AS, Satten GA: A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 2007;80:921–930.

8 Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA: Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet 2005;76:592–608.

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

10 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959.

11 Laird NM, Horvath S, Xu X: Implementing a unified approach to family-based tests of association. Genet Epidemiol 2000;19 Suppl 1:S36–S42.

12 Laird NM, Lange C: Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 2006;7:385–394.

13 Lange C, Laird NM: On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. Genet Epidemiol 2002;23:165–180.

14 Lange C, Silverman EK, Xu X, Weiss ST, Laird NM: A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics 2003;4:195–206.

15 Won S, Bertram L, Becker D, Tanzi RE, Lange C: Maximizing the power of genome-wide association studies: a novel class of powerful family-based association tests. Stat Biosci 2009;1:125–143.

16 Rabinowitz D, Laird N: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 2000;50:211–223.

17 Fulker DW, Cherny SS, Sham PC, Hewitt JK: Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 1999;64:259–267.

18 Ionita-Laza I, McQueen MB, Laird NM, Lange C: Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. Am J Hum Genet 2007;81:607–614.

19 Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C: Genomic screening and replication using the same data set in family-based association testing. Nat Genet 2005;37:683–691.

20 Lee S, Zou F, Wright FA: Convergence and prediction of principal component scores in high-dimensional settings. Ann Stat 2010;38:3605–3629.

21 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. PLoS Genet 2006;2:e190.

22 Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS: Novel case-control test in a founder population identifies p-selectin as an atopy-susceptibility locus. Am J Hum Genet 2003;73:612–626.

23 Browning BL, Browning SR: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 2009;84:210–223.

24 Croiseau P, Genin E, Cordell HJ: Dealing with missing data in family-based association studies: a multiple imputation approach. Hum Hered 2007;63:229–238.

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

26 Lange C, Lyon H, DeMeo D, Raby B, Silverman EK, Weiss ST: A new powerful nonparametric two-stage approach for testing multiple phenotypes in family-based association studies. Hum Hered 2003;56:10–17.

27 Chen WM, Abecasis GR: Family-based association tests for genomewide association scans. Am J Hum Genet 2007;81:913–926.

28 Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E: Variance component model to account for sample structure in genome-wide association studies. Nat Genet 2010;42:348–354.

29 Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL: Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 2006;63:168–174.

30 Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, Schjeide BM, Hooli B, Divito J, Ionita I, Jiang H, Laird N, Moscarillo T, Ohlsen KL, Elliott K, Wang X, Hu-Lince D, Ryder M, Murphy A, Wagner SL, Blacker D, Becker KD, Tanzi RE: Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. Am J Hum Genet 2008;83:623–632.

31 Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, Jehu L, Segurado R, Stone D, Schadt E, Karnoub M, Nowotny P, Tacey K, Catanese J, Sninsky J, Brayne C, Rubinsztein D, Gill M, Lawlor B, Lovestone S, Holmans P, O'Donovan M, Morris JC, Thal L, Goate A, Owen MJ, Williams J: Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet 2007;16:865–873.

32 Li H, Wetten S, Li L, St Jean PL, Upmanyu R, Surh L, Hosford D, Barnes MR, Briley JD, Borrie M, Coletta N, Delisle R, Dhalla D, Ehm MG, Feldman HH, Fornazzari L, Gauthier S, Goodgame N, Guzman D, Hammond S, Hollingworth P, Hsiung GY, Johnson J, Kelly DD, Keren R, Kertesz A, King KS, Lovestone S, Loy-English I, Matthews PM, Owen MJ, Plumpton M, Pryse-Phillips W, Prinjha RK, Richardson JC, Saunders A, Slater AJ, St George-Hyslop PH, Stinnett SW, Swartz JE, Taylor RL, Wherrett J, Williams J, Yarnall DP, Gibson RA, Irizarry MC, Middleton LT, Roses AD: Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. Arch Neurol 2008;65:45–53.

33 Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, Joshipura KD, Pearson JV, Hu-Lince D, Huentelman MJ, Craig DW, Coon KD, Liang WS, Herbert RH, Beach T, Rohrer KC, Zhao AS, Leung D, Bryden L, Marlowe L, Kaleem M, Mastroeni D, Grover A, Heward CB, Ravid R, Rogers J, Hutton ML, Melquist S, Petersen RC, Alexander GE, Caselli RJ, Kukull W, Papassotiropoulos A, Stephan DA: GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron 2007;54:713–720.

34 Fisher RA: Statistical Methods for Research Workers, ed 11. Edinburgh, Oliver & Boyd, 1950.

35 Liptak T: On the combination of independent tests. Magyar Tud Akad Mat Kutato Int Kozl 1958;3:171.

36 Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM: Control of confounding of genetic associations in stratified populations. Am J Hum Genet 2003;72:1492–1504.