CrossMark

Multivariate Gene-Based Association Test on Family Data in MGAS

César-Reyer Vroom[1] · Danielle Posthuma[1,2] · Miao-Xin Li[3,4,5,6] · Conor V. Dolan[7] ·
Sophie van der Sluis[1]

Abstract In analyses of unrelated individuals, the program multivariate gene-based association test by extended Simes (MGAS), which facilitates multivariate gene-based association testing, was shown to have correct Type I error rate and superior statistical power compared to other multivariate gene-based approaches. Here we show, through simulation, that MGAS can also be applied to data including genetically related subjects (e.g., family data), by using p value information obtained in Plink or in generalized estimating equations (with the 'exchangeable' working correlation matrix), both of which account for the family structure on a univariate single nucleotide polymorphism-based level by applying a sandwich correction of standard errors. We show that when applied to family-data, MGAS has correct Type I error rate, and given the details of the simulation setup, adequate power. Application of MGAS to seven eye measurement phenotypes showed statistically significant association with two genes that were not discovered in previous univariate analyses of a composite score. We conclude that MGAS is a useful and convenient tool for multivariate gene-based genome-wide association analysis in both unrelated and related individuals.

Keywords GWAS · Multivariate · Gene-based · Family data · MGAS · GATES · TATES

Introduction

Multivariate genotype–phenotype data are often collected in families, e.g., twins and their siblings, or parent-offspring trios. Yet, genetic association tests that take family relatedness into account generally consider the associations of a single phenotype with many single nucleotide polymorphisms (SNPs), i.e., are univariate and SNP-based in nature. Here we show, through simulation and application to eye measurement data, that the multivariate gene-based tool, multivariate gene-based association test by extended

✉ César-Reyer Vroom
c.vroom@vu.nl

1 Department of Clinical Genetics, Section Complex Traits Genetics, VU Medical Center (VUmc), Neuroscience Campus Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

2 Department of Complex Traits Genetics, Center for Neurogenomics and Cognitive Research (CNCR), VU University, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands

3 Department of Psychiatry, The University of Hong Kong, Pokfulam, Hong Kong

4 State Key Laboratory for Cognitive and Brain Sciences, The University of Hong Kong, Pokfulam, Hong Kong

5 The Centre for Reproduction, Development and Growth, The University of Hong Kong, Pokfulam, Hong Kong

6 The Centre for Genomic Sciences, The University of Hong Kong, Pokfulam, Hong Kong

7 Department of Biological Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands

Simes (MGAS) procedure van der Sluis et al. 2015), can be used on family data.

In genome-wide association studies (GWAS), a univariate phenotype, e.g., a sum score or a case–control dichotomy, is regressed on a large number of common SNPs. Under specific circumstances,[1] a sum score provides a sufficient phenotypic operationalization in the sense that it captures all relevant information in the underlying trait-generating genotype–phenotype model. However, as the specific circumstances do not usually hold, the use of sum scores often implies a loss of information, and consequently a loss in power (van der Sluis et al. 2010; Minică et al. 2010; Medland and Neale 2010). Recently, a new multivariate gene-based test MGAS (van der Sluis et al. 2015) was introduced. MGAS integrates TATES (van der Sluis et al. 2013), a multivariate SNP-based test, and GATES (Li et al. 2011), a univariate gene-based test.

Like these two methods, MGAS combines p-value information obtained in standard univariate SNP-based analyses into multivariate gene-based p-values while correcting for the correlations between the multivariate phenotypes on the one hand, and the correlations between SNPs within genes (i.e., linkage disequilibrium, LD) on the other. MGAS has been shown to perform well under diverse trait-generating genotype–phenotype models (e.g., 1- and 4-factor models, network models, models in which the SNPs in the gene affected either the latent trait or the observed traits directly, and models including opposite effects), and often to have superior power over gene-based analyses of univariate phenotypic sum scores (GATES, multiple regression), MANOVA (all phenotypes as dependent variables and all SNPs as predictors), and GATES on MANOVA p-values (i.e., MANOVA on each SNP separately, and subsequent combining of the multivariate SNP-based p-values using GATES to arrive at a single multivariate gene-based p-value). So far, however, investigations on the MGAS procedure focused on data of unrelated individuals. GWAS may, however, involve related subjects (e.g., family-based data including twins, parents, siblings).

The aim of the present paper is to demonstrate that the MGAS procedure can be used on family based GWAS data, by applying MGAS to the results of generalized estimating equations (GEE) regression analysis (Dobson 2008), with an appropriate ("sandwich") correction of standard errors to accommodate the effect of family clustering. We demonstrate by means of simulation that the

Type I errors associated with this procedure are correct. We consider the issue of power by applying GEE with two choices of the working correlation matrix (i.e., the provisional model to account for familial clustering): the independence working matrix (as used in Plink; Purcell et al. 2007), and the exchangeable working matrix. We consider both options because Minică et al. (2014b) demonstrated that the choice of working matrix has a bearing on the power of GEE to detect SNP–phenotype associations.

Methods

Simulation

To demonstrate the accuracy of multivariate gene-based analyses using MGAS as performed on p-values obtained in family data, we simulated multivariate phenotypes and genetic data of SNPs in genes in genetically related subjects. All simulations concerned $N_{fam} = 500$ families, including parents and either monozygotic (MZ: 50 % of the families) or dizygotic (DZ) twin pairs, i.e., 2000 individuals in total. For each individual, six normally distributed phenotypes were generated under either a 1- or a 2-latent factor model, with each observed phenotype related to either one of the two factors (i.e., simple structure, see Fig. 1 for a schematic representation of simulation settings). The 1-factor setting was chosen because the use of univariate sum scores or case–control dichotomies as dependent variables in GWAS implies a 1-factor model, and because the 1-factor model is consistent with practical and diagnostic conceptualizations in psychology and medicine. However, many psychological traits are conceptualized and measured in a multidimensional manner. For instance, verbal and non-verbal IQ are distinguished in cognition research, and twin- and family studies have shown that this phenotypic multidimensionality partly reflects genetic multidimensionality (e.g., Rijsdijk and Vernon 2002; Posthuma et al. 2001; Hoekstra et al. 2007; Price et al. 2000). Therefore, we included the 2-factor model in our simulations. Note that for both the 1- and 2-factor simulations, the six observed phenotypes (not the latent factor scores) feature as the dependent variables in the genetic association analyses.

The phenotypes were simulated according to the model

$$\Sigma = \Lambda \times \Psi \times \Lambda^t + \Theta \qquad (1)$$

where, in the 1-factor model, Σ is the 24×24 variance–covariance matrix between the six phenotypes of each of the four family members, Λ is the 24×4 matrix of factor loadings (superscript t denotes matrix transpose), Ψ is a 4×4 variance–covariance matrix of the four common factors within a family, and Θ is the 24×24 diagonal

---

[1] Specifically, the sum score is a sufficient statistic if (a) all correlations between the phenotypes comprising the sum can be explained by 1 latent factor or trait, (b) all phenotypes show identical relations to this latent factor (i.e., identical factor loadings), (c) all phenotypes have identical residual variances (i.e., the phenotypes adhere to a Rasch model (Rasch 1980), and (d) the genetic effect is on this latent factor, see e.g. van der Sluis et al. 2010.

Fig. 1 Schematic representation of simulation settings. **a** Schematic representation of the 1-factor model. **b** Schematic representation of the 2-factor model. A, C, and E represent the genetic factor (A), shared environmental factor (C), and unique environmental factor (E), influencing the latent factor F of either twin (tw1, tw2) or parents (not shown). Each latent factor (or two in case of the 2-factor model), influences the observed variables, depicted as *squares*. A gene (Genetw1 and Genetw2 are the genes of twin 1 and twin 2; parents not shown) harboring 1 or 0 effective SNPs out of 10, influences the latent factor or an observed variable. Concerning the 1-factor model: in case of 40 % genetic, 40 % shared environmental, and 20 % unique environmental variation; $e1 = \sqrt{0.2}$, $c1 = a1 = \sqrt{0.4}$. In case of 80 % genetic, 0 % shared environmental, and 20 % unique environmental variation; $e1 = \sqrt{0.2}$, $c1 = 0$, $a1 = \sqrt{0.8}$. Concerning the 2-factor model: in case of 40 % genetic, 40 % shared environmental, and 20 % unique environmental variation; $e1 = e3 = \sqrt{0.2}$, $c1 = a1 = \sqrt{0.4}$, $c2 = a2 = \sqrt{0.05625}$, $a3 = c3 = \sqrt{0.34375}$. In case of 80 % genetic, 0 % shared environmental, and 20 % unique environmental variation; $e1 = e3 = \sqrt{0.2}$, $c1 = c2 = c3 = 0$, $a1 = \sqrt{0.8}$, $a2 = \sqrt{0.1125}$, $a3 = \sqrt{0.6375}$
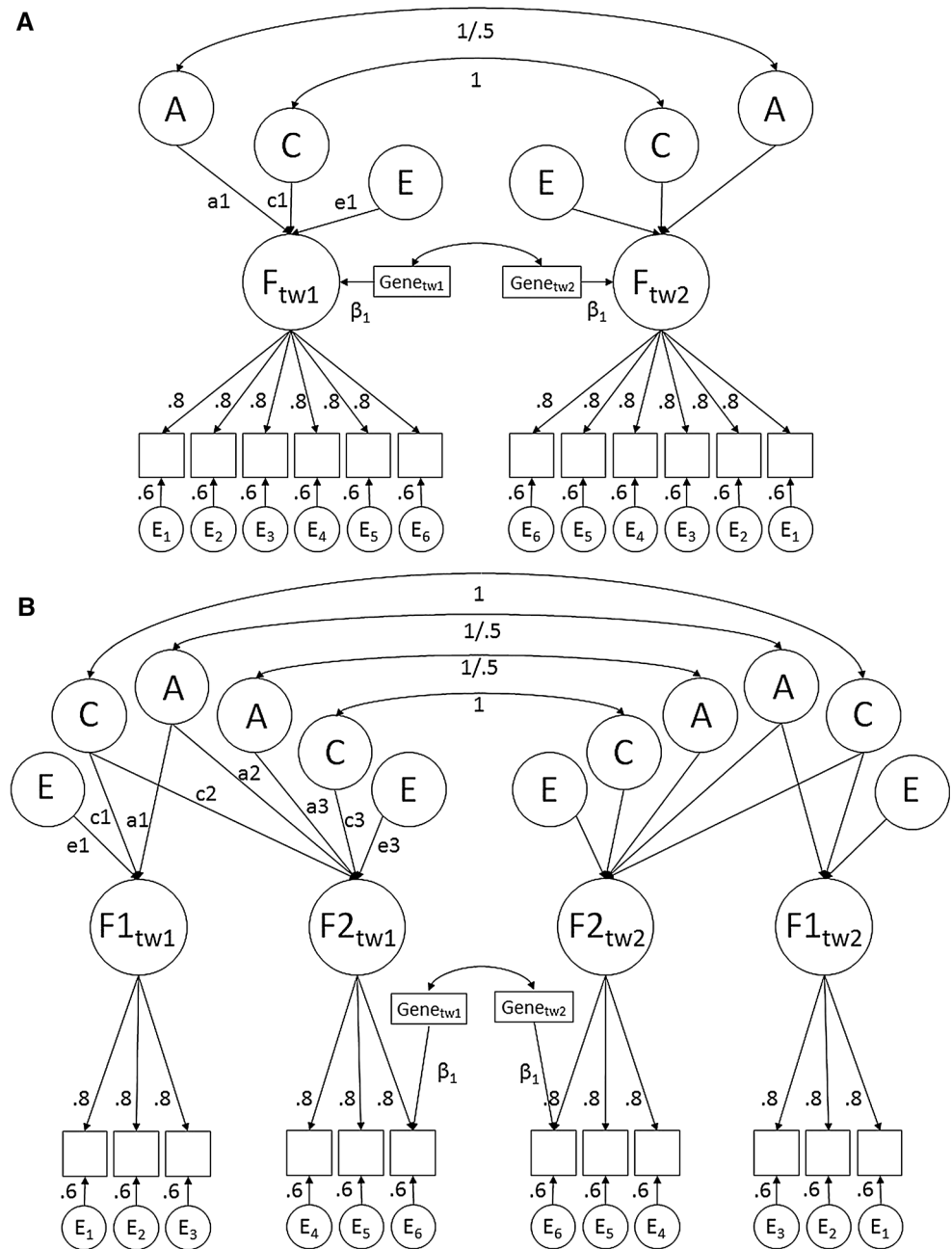
matrix of residual variances. In case of a 2-factor model, the latent factors, each indicated by three phenotypes, were simulated to correlate ($r = 0.3$) within individuals, with the correlation caused by both genetic and shared environmental influences (see Fig. 1b). Λ then is the 24×8 matrix of factor loadings, and Ψ is the 8×8 variance–covariance matrix of the latent factors within a family. In all simulations, factor loadings were set to 0.8, so that 64 % of the phenotypic variation was explained by the latent factor, leaving 36 % unexplained residual variance. All data were generated under the ACE model, including additive genetic (A), shared (C), and unshared environmental (E) effects. The phenotypic variance at the latent level was decomposed as follows: 40 %

A, 40 % C, and 20 % E (the ACE model), or 80 % A, 0 % C, and 20 % E (i.e., the AE model). The residual variances were all specified to be due to E only. For each individual, 10 SNPs were simulated (MAF = 0.5), that together formed a gene (LD ~0.9). In simulating the genotypes, first the haplotypes of the parents were simulated, under the assumption of Hardy–Weinberg Equilibrium, with known phase, and then these haplotypes were used to simulate the offspring haplotypes with recombination rate based on 5 cM distance between SNPs, which is roughly equivalent to 5 % recombination probability. We note that the resulting region is unrealistically large for a gene, but as MGAS can also be used to test regions rather than genes, we wanted to study the

Type I error rate under circumstances including recombination. We also note that the power of MGAS is known to depend on the number of SNPs covering the gene or region, and the LD structure between these SNPs, but not on the physical distance between the SNPs per se (van der Sluis et al. 2015). Only one of the 10 SNPs in the region was actually associated with the phenotype(s). We considered the power to detect the signal attributable to this disease-susceptibility locus (DSL) given that the DSL explained either 1 % variation in the (first) latent factor, or 0.5 % variation in the last observed phenotype. We also evaluated the Type I error rate by simulating data in which none of the SNPs were related to any of the phenotypes. These settings resulted in a total of 16 simulation scenarios: one or two factor models, two different ACE decompositions, DSL-effect either on the observed or latent phenotype, and DSL either with an effect (on the factor or on the last observed phenotype) or without an effect (to ascertain the Type I error). All simulations were repeated $N_{sim} = 2000$ times. Note that the standard error of the p-value equals $\sqrt{((p(1 - p))/N_{sim})}$. Therefore, p-values between 0.04 and 0.06 lie within the 95 % confidence interval for an unbiased nominal p-value when there is no effect (given $\alpha = 0.05$, $p \approx 0.05$ is expected), and are thus considered correct.

Analyses

All analyses were performed in R, using the packages *MASS*, *psych*, and *corpcor* for simulation, and the packages *OpenMx* (Boker et al. 2011) and *gee* for data analysis. We denote the model used to simulate the data the true model, and used the package *OpenMx* to fit it. In the true model, all parameters were estimated (i.e., variable means, factor loadings, residual variances, regression coefficients, and a, c, and e paths; with variable means, factor loadings, and residuals error variances constrained to be equal between family members), and the family relations were represented correctly. In practice, the actual data generating genotype–phenotype system is generally unknown. However, fitting the true model is useful in simulations because it provides information on power and Type I error rate under the optimal circumstances of no misfit or misspecification of the actual data generating genotype–phenotype model. In the true model, the p-value for the test of association between the 10 SNPs on the one hand and the latent trait or last phenotype on the other hand, was obtained using a 10 df likelihood ratio test, i.e., comparing the fit of the true model with all 10 regression parameters estimated freely, to the fit of the model with these parameters fixed to 0 (i.e., because the gene is our functional unit of interest, we tested the effects of all 10 SNPs in the gene simultaneously, instead of only the effective SNP).

Results obtained by fitting the true model in *OpenMx*, were compared with MGAS results. The MGAS procedure is explained in van der Sluis et al. (2015). Briefly, MGAS uses the p-values obtained in regressing each of the six observed phenotypes individually on each of the 10 SNPs while correcting for familial relatedness between participants by using GEE with the working correlation matrix set to independent or exchangeable. The working correlation matrix represents the background correlations between the family members conditional on the SNP under study. As our simulations concerned four members per family, the working correlation matrix is a 4×4 matrix. Under the independent setting, this matrix is diagonal, with the variances equal (i.e., one parameter), and all off-diagonal elements fixed to 0. Under the exchangeable setting, the matrix contains equal variances and equal covariances (i.e., two parameters). Given four family members, i.e., MZ or DZ twins and their parents, both GEE settings entail misspecification, as residual covariances between family members conditional on the tested SNP are unlikely to be either zero or equal. Practically, this misspecification results in underestimation of the standard error of the parameter of interest, i.e., the regression weight relating the SNP to the phenotypic data. The sandwich corrected standard error has been shown to yield correct Type I error rates in the context of family based association analysis (Minică et al. 2014b).

MGAS then uses all the GEE based univariate p-values (i.e., nvar $\times$ nsnp = $6 \times 10 = 60$) to obtain one multivariate gene-based p-value P_{MGAS} as follows:

$$P_{MGAS} = \min\left(\frac{q_e p_j}{q_{ej}}\right) \qquad (2)$$

Here, q_e denotes the effective number of p-values within a gene, q_{ej} the effective number of p-values among the top j p-values where j runs from 1 to nvar $\times$ nsnp, and p_j denotes the jth p-value in the list of ordered p-values. P_{MGAS} is thus the smallest weighted p-value within a gene associated with the null hypothesis that none of the nvar phenotypes are related to the nsnp SNPs within the gene under study, and the alternative hypothesis that at least one of the nvar phenotypes is related to at least one of the nsnp SNPs.

Since the SNPs are correlated and the phenotypes are correlated, the p-values from the univariate regressions are also correlated. These p-value correlations are not observed but can be accurately approximated from the p-value-sorted Kronecker product between the SNP- and phenotype correlation matrices using a sixth order polynomial (Li et al. 2011; van der Sluis et al. 2015). The effective number of p-values q_{ej} among the top j p-values is then calculated from this p-value-sorted Kronecker product using

$$q_{ej} = j - \sum_{i=1}^{j} I(\lambda_i) \times (\lambda_i - 1) \tag{3}$$

where λ_i is the ith eigenvalue of the sorted Kronecker product, and $I(\lambda_i)$ is an indicator function taking on value 1 if $\lambda_i > 1$ and value 0 if $\lambda_i \leq 1$. q_{ej} is thus calculated as j minus the sum of the difference between λ_i and 1 for those eigenvalues >1, and $q_e = q_{ej}$ for the special case that $j = nvar \times nsnp$ (i.e., when the selection of j p-values covers all nvar $\times$ nsnp p-values). MGAS thus accommodates correlations between phenotypes and SNPs within individuals, while GEE corrects for the relatedness between subjects but only in a univariate SNP-based setting. Whether the combination of these two corrections suffices to account for the correlations that are expected within families between phenotypes and between SNPs, is the topic of our investigation.

Results

Table 1 shows the Type I error rates and power of the true model, and the two different MGAS models, based on GEE-independent (i.e., Plink) or GEE-exchangeable. None of the Type I error rates deviated significantly from 0.05, irrespective of the used analysis method or simulation settings. The fact that both MGAS models have correct

Type I error rates indicates that the sandwich correction of the standard error following the GEE regression adequately handles the relatedness between participants in the sample. MGAS can thus be used on family data as long as the relatedness between participants has been sufficiently accommodated in the univariate SNP-based association analyses.

Furthermore, as shown by Minică et al. (2014b), the test of association based on the GEE-independent option is less powerful than that based on the GEE-exchangeable option. The difference in power is a function of the phenotypic covariance matrix: the greater the phenotype covariance between family members, the larger the power advantage of GEE-exchangeable over GEE-independent.

When the DSL-effect is limited to one latent variable, the power of the MGAS model is adequate (~ 0.8) compared to the true model (~ 0.8). However, when the DSL-effect is specific to one observed variable, a notable drop in power (from ~ 0.9 to 0.5) is observed for MGAS compared to the true model. This is understandable since in the true model the DSL-effect is part of the residual of the observed variable (i.e., the variance in the observed variable that is not explained by the latent factor), while it is part of the total variance in the GEE models used in MGAS. That is, the true model tests how much of the residual variance of the specific phenotype (i.e., conditional on the latent factor) is due to the DSL, while the GEE models test how much of

Table 1 Type I error rates and power for the true model and MGAS

	Simulation settings					True model	MGAS	
	Effect size	Location effect	A^2	C^2	E^2		GEE-indep.	GEE-exch.
							Type I error	
1fac	0	Lat[1]	0.4	0.4	0.2	0.060	0.051	0.052
			0.8	0	0.2	0.050	0.037	0.041
	0	Obs[2]	0.4	0.4	0.2	0.049	0.042	0.044
			0.8	0	0.2	0.059	0.039	0.045
2fac	0	Lat	0.4	0.4	0.2	0.055	0.039	0.047
			0.8	0	0.2	0.050	0.052	0.050
	0	Obs	0.4	0.4	0.2	0.055	0.045	0.049
			0.8	0	0.2	0.050	0.055	0.059
							Power	
1fac	0.01	Lat	0.4	0.4	0.2	0.848	0.757	0.927
			0.8	0	0.2	0.848	0.757	0.927
	0.005	Obs	0.4	0.4	0.2	0.946	0.405	0.539
			0.8	0	0.2	0.942	0.442	0.490
2fac	0.01	Lat	0.4	0.4	0.2	0.802	0.685	0.855
			0.8	0	0.2	0.749	0.764	0.827
	0.005	Obs	0.4	0.4	0.2	0.892	0.404	0.539
			0.8	0	0.2	0.899	0.462	0.507

[1] DSL-effect is on the level of the latent variable

[2] DSL-effect is on the level of the observed variable

the total variance in the observed phenotype is due to the DSL.

Occasionally, the power of MGAS is higher than the power of the true model. Although this may seem counterintuitive since the true model is the model used for simulation, this can be explained by the difference in the number of estimated parameters between the true model and the MGAS model; the true model estimates all parameters describing the relations between the phenotypes and SNPs within and between subjects (31 in case of a 1-factor model, 37 in case of a 2-factor model) while MGAS is based on univariate regression, in which only one parameter is estimated. When constraining the association parameters (β_1 in Fig. 1a, b) to zero, the resulting misfit can, in the *OpenMx* model, partly be accommodated by other parameters, while such "compensation" is impossible in the univariate regression context.

Finally, because MGAS is based on TATES (i.e., multivariate SNP-based test) and GATES (i.e., univariate gene-based test), we also considered the Type I error rates of these procedures using the same simulated family-data (see Online Resource 1). Since TATES is based on a single SNP, only the effective SNP was used for analysis. Since GATES is based on a single phenotype, three different phenotypic approaches were taken. First, only one (directly or indirectly) affected observed variable was regressed on the gene. Second, in case of the 1-factor model, the overall sum score calculated across all six observed variables was regressed on the gene. Third, in case of the 2-factor model, the sum score of the observed variables under the latent variable including the affected observed variable, or the sum score of the observed variables under the affected latent variable, was regressed on the gene. Type I error rates proved correct for TATES and all variants of GATES, and again analyses based on GEE-exchangeable proved more powerful than analyses based on GEE-independent. Note that occasionally, TATES and GATES outperform MGAS with respect to power. This is due to only using the effective SNP in case of TATES, and only using the affected observed variable, or factor sum score including the affected observed variable, in case of GATES. Note also that when the DSL is on the observed variable, the sum score approach of GATES has very low power (~ 0.1, see also van der Sluis et al. 2015).

Implementation: Myopia data

To illustrate the application of MGAS to family data, we analyzed data consisting of seven eye measurements obtained in families of the Twins Eye Study of Refractive Error and Glaucoma Endophenotypes (TES) Database of the TwinsUK Cohort. We gained acces to these data through dbGaP. The TES cohort initially consisted of 2928 subjects, of whom 84 were removed during quality control (Quality control was performed according to the Anderson et al. (2010) protocol; see Online Resource 2 for more information), leaving 2844 subjects, clustered in 2115 families. Of the 2115 families, 1416 included a single member, 674 included two members, 20 included three members, and five included four members. The seven phenotypes were: sphere right eye (SPHR), cylinder right eye (CYLR), axis right eye (AXISR), sphere left eye (SPHL), cylinder left eye (CYLL), axis left eye (AXISL), and cataract (CATARACT). The genotypic data initially included 589,296 SNPs, of which 51,462 were removed during quality control, leaving 537,834 SNPs, to be included in analyses. MGAS was run in KGG v3.5 (van der Sluis et al. 2015) (http://statgenpro.psychiatry.hku.hk/limx/kgg/). KGG requires as input a phenotypic correlation file, and a file containing p-values from the seven univariate analyses of all SNPs. The univariate analyses were run in GEE (option: 'exchangeable') through PLINK (see Minică et al. 2014b) with sex and age as covariates. The genomic inflation factor was 1.03. Principal components were calculated using multidimensional scaling (MDS) based on raw Hamming Distances in Plink. As adding Principal Components as covariates did not reduce the inflation factor, all reported analyses were run without Principal Components as covariates. The phenotypic correlations entered in MGAS (Online Resource 3) were corrected for sex and age.

Of the total 537,834 SNPs, 314,900 were located within genes, covering 22,739 genes in total. In KGG v3.5, SNPs were mapped onto genes or genic regions defined by the RefGene database with 5 kb boundary extensions on both sides. Using a Bonferroni Family Wise Error Rate threshold of $\alpha = 2.20e-6$, MGAS identified two significant genic regions that were not previously reported for these phenotypes: *LINC00583* ($P_{MGAS} = 1.03e-8$, mainly contributing to CYLR), and *OGDH* ($P_{MGAS} < 8.95e-7$, mainly contributing to AXISR; see Online Resource 4 for univariate p-values of all SNPs within these two genes).

Although GWAS has not been performed on this specific dataset, this dataset has been included in a larger GWAS (Hysi et al. 2010) and was part of a GWAS meta-analysis (Verhoeven et al. 2013). In both studies, the dependent variable was "spherical equivalent", which is a composite score created from an individual's sphere and cylinder measures. Using MERLIN (Abecasis et al. 2002) to analyze that data, the SNP rs8027411, located in the *RASGRF1* gene, had the strongest (yet not genome-wide significant) association with spherical equivalent in (Hysi et al. 2010) (p = 7.91e−8), and this SNP did reach genome-wide significance (p = 2.07e−9) in a replication sample of six cohorts combined. The *RASGRF1* gene was also significantly associated with spherical equivalent in Verhoeven et al. (2013).

SNP rs8027411 was only included in our analyses when larger boundary extensions than 5 kb were used. However, we did not replicate this result in our analyses. Even when we used the composite score of spherical equivalent as a dependent variable, i.e., using the GATES method, we did not replicate this finding. In this GATES analysis, the strongest associated gene, *PXDNL*, did not reach genome-wide significance (p = 3.42e−5). Our non-replication could be due to the fact that both the studies by Hysi et al. (2010) and Verhoeven et al. (2013) included larger samples. However, it is interesting to note that most SNPs that are strongly associated with mean spherical equivalent in our analyses, are most strongly (but not significantly) associated to both left and right eye sphere, and not to cylinder of either eye. We refer the reader to Online Resource 5 for the details of these results. Furthermore, while sphere of left and right eye are correlated ($r = 0.87$) and both cylinder measures are correlated ($r = 0.60$), neither sphere measure is correlated with cylinder (r between -0.03 and 0.06). These correlations do not support the choice to use a composite of sphere and cylinder, such as spherical equivalent, as a composite of uncorrelated phenotypes may decrease the power to detect associated SNPs or genes (van der Sluis et al. 2013). This present application shows that multivariate gene-based methods like MGAS can reveal associations which may be missed when using univariate SNP-based, or univariate gene-based analyses on composite scores.

Discussion

Here we have shown that MGAS, which combines p-value information obtained in regressing univariate phenotypes on common SNPs into one multivariate gene-based p-value, has a correct Type I error rate when the data include (genetically) related participants. This implies that no additional corrections are required after using a sandwich correction in GEE to account for the residual relatedness in families conditional on the SNP under study. Therefore, MGAS is a convenient tool for multivariate gene-based genome-wide association analysis that can be used on data of unrelated as well as related individuals. We found that GEE-exchangeable generally has greater power than GEE-independent (see Minică et al. 2014b). The difference in power (between 4.5 and 17 % for the current simulation settings) depends on the magnitude of the covariance between the family members conditional on the SNP under study. As GEE is a freely available R-package (http://cran.r-project.org/web/packages/gee/gee.pdf) that can be called from the Plink environment, we advise to use MGAS, TATES, and GATES on p-values obtained in GEE-exchangeable when data include related individuals.

We note that while our simulations are limited (i.e., only two trait-generating models with two choices for ACE-decomposition), the main objective of this study was to verify that MGAS has correct Type I error rate when used on family data. To test this, we choose realistic settings for A and C that created considerable covariance between phenotypes of different family members, ranging from 0 to 0.26 between parents, from 0.26 to 0.38 between DZ twins, and equaling 0.51 for MZ twins. As the Type I error rates of MGAS were correct for all these settings, we believe that further testing using different settings (e.g., higher or lower factor loadings, different choices for the ACE-decomposition) will add little, especially since in other papers (Minică et al. 2014a, b), Type I error rates were found to be correct in both Plink and GEE-exchangeable for various genetic settings, suggesting sufficient correction for relatedness. For this reason we also did not include scenarios with mixes of families of different sizes and single member families, as this would not pose additional problems to MGAS. We refer to the original MGAS paper (van der Sluis et al. 2015) for a comprehensive simulation study on the power of MGAS under many different circumstances (e.g., 1- and 4-factor models, network models, models in which the SNPs in the gene affected either the latent trait or the observed traits directly, and models including opposite effects). In conclusion, MGAS, GATES, and TATES can be applied to results obtained in samples characterized by family clustering (e.g., samples from twin- and family registers).

Compliance with ethical standards

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Anderson CA, Pettersson FH, Clarke GM, Cardon GM, Morris AP, Zondervan KT (2010) Data quality control in genetic case–control association studies. Nat Protoc 5:1564–1573

Boker S, Neale M, Maes H et al (2011) OpenMx: an open source extended structural equation modeling framework. Psychometrika 76:306–317

Dobson AJ (2008) In: Chatfield C, Zidek J (eds) An introduction to generalized linear models, 3rd edn. Chapman & Hall/CRC, London

Hoekstra RA, Bartels M, Boomsma DI (2007) Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. Learn Individ Differ 17:97–114

Hysi PG, Young TL, Mackey DA et al (2010) A genome-wide association study for myopia and refractive error identifies a susceptibility locus at 15q25. Nat Genet 42:902–905

Li MX, Gui HS, Kwan JSH, Sham PC (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet 88:283–293

Medland S, Neale MC (2010) An integrated phenomic approach to multivariate allelic association. Eur J Hum Genet 18:233–239

Minică CC, Boomsma DI, van der Sluis S, Dolan CV (2010) Genetic association in multivariate phenotypic data: power in five models. Twin Res Hum Genet 13:525–543

Minică CC, Boomsma DI, Vink JM, Dolan CV (2014a) MZ twin pairs or MZ singletons in population family-based GWAS? More power in pairs. Mol Psychiatry 19:1154–1155

Minică CC, Dolan CV, Kampert MM, Boomsma DI, Vink JM (2014b) Sandwich corrected standard errors in family-based genome-wide association studies. Eur J Hum Genet 23(3):388–394

Posthuma D, de Geus EJC, Boomsma DI (2001) Perceptual speed and IQ are associated through genetic factors. Behav Genet 31:593–602

Price TS, Eley TC, Dale PS, Stevenson J, Saudino K, Plomin R (2000) Genetic and environmental covariation between verbal and nonverbal cognitive development in infancy. Child Dev 71:948–959

Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Rasch G (1980) Probabilistic models for some intelligence and attainment tests. MESA Press, Chicago

Rijsdijk FV, Vernon PA, Boomsma DI (2002) Application of hierarchical genetic models to Raven and WAIS subtests: a Dutch twin study. Behav Genet 32:199–210

van der Sluis S, Verhage M, Posthuma D, Dolan CV (2010) Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. PLoS One 5:e13929

van der Sluis S, Posthuma D, Dolan CV (2013) TATES: efficient multivariate genotype–phenotype analysis for genome-wide association studies. PLoS Genet 9:e1003235

van der Sluis S, Dolan CV, Li J, Song Y, Sham P, Posthuma D, Li MX (2015) MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. Bioinformatics 31:1007–1015

Verhoeven VJM, Hysi PG, Wojciechowski R et al (2013) Genome-wide meta analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. Nat Genet 45:313–318