Hum Genet (2014) 133:139–150
DOI 10.1007/s00439-013-1361-9

ORIGINAL INVESTIGATION

A unified GMDR method for detecting gene–gene interactions in family and unrelated samples with application to nicotine dependence

Guo-Bo Chen · Nianjun Liu · Yann C. Klimentidis · Xiaofeng Zhu · Degui Zhi · Xujing Wang · Xiang-Yang Lou

Received: 2 July 2013 / Accepted: 5 September 2013 / Published online: 21 September 2013

Abstract Gene–gene and gene–environment interactions govern a substantial portion of the variation in complex traits and diseases. In convention, a set of either unrelated or family samples are used in detection of such interactions; even when both kinds of data are available, the unrelated and the family samples are analyzed separately, potentially leading to loss in statistical power. In this report, to detect gene–gene interactions we propose a generalized multifactor dimensionality reduction method that unifies analyses of nuclear families and unrelated subjects within the same statistical framework. We used principal components as genetic background controls against population stratification, and when sibling data are included, within-family control were used to correct for potential spurious association at the tested loci. Through comprehensive simulations, we demonstrate that the proposed method can remarkably increase power by pooling unrelated and offspring's samples together as compared with individual analysis strategies and the Fisher's combining *p* value method while it retains a controlled type I error rate in the presence of population structure. In application to a real dataset, we detected one significant tetragenic interaction among *CHRNA4*, *CHRNB2*, *BDNF*, and *NTRK2* associated with nicotine dependence in the Study of Addiction: Genetics and Environment sample, suggesting the biological role of these genes in nicotine dependence development.

Introduction

Understanding how genetic mechanisms contribute to the formation of complex traits is one of the major challenges in genetics studies. Although the recent surge of genome-wide association studies (GWASs) has led to the discovery of many new loci that contribute to phenotypic variation, unraveling the so-called "missing heritability" (Manolio et al. 2009) may require more sophisticated strategies not limited to single-marker analysis. The ubiquitous existence of gene–gene (G × G) interaction is well documented, from the molecular interaction to statistical epistasis, and composes pivotal determinants in the formation of phenotypic outcomes. It is consequently anticipated that G × G interaction will help elucidate some of the missing heritability (Zuk et al. 2012).

Conventional single-marker methods that isolate interacting genes from their context likely obfuscate the interconnected networks and plausibly fail to model the complex gene networks that genuinely relate to a phenotypic outcome. Therefore, methods in which association is tested by incorporating multiple genes have been proposed [see a recent review by Cordell (2009)]. Among them, multifactor dimensionality reduction (MDR) method,

Electronic supplementary material The online version of this article (doi:10.1007/s00439-013-1361-9) contains supplementary material, which is available to authorized users.

G.-B. Chen · N. Liu · Y. C. Klimentidis · D. Zhi · X.-Y. Lou (✉)
Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, 1665 University Boulevard, RPHB 327, Birmingham, AL 35294-0022, USA
e-mail: xylou@uab.edu

X. Zhu
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA

X. Wang
The Bioinformatics and Systems Biology Core NHLBI, National Institutes of Health, Bethesda, MD 20892, USA

originally for a case–control study, has sustained its popularity since it was proposed (Ritchie et al. 2001). Rather than modeling the interaction term per se as with regression methods, MDR seeks to capture a combination of loci of interest, a pattern that maximizes the phenotypic variation it explains. It treats the G × G interaction as a whole, coinciding to the very original epistasis described by Bateson and offering a solution that avoids decomposition as concerned in regression methods. As it projects the high-order interaction into one dimension, it theoretically overcomes the issue of high dimensionality, provided that the sample size is sufficient. Further development, such as generalized multifactor dimensionality reduction (GMDR), which integrated generalized linear model into MDR (Lou et al. 2007), and pedigree-based generalized multifactor dimensionality reduction (PGMDR) (Lou et al. 2008), allows MDR to be applied to both binary and continuous traits with adjustment for covariates whenever necessary and to pedigree data.

The family-based design and the population-based design (referred as unrelated-individual design) are among the most commonly used designs in genetic studies. Family-based association tests, such as transmission/disequilibrium test (Spielman et al. 1993), are well known for their robustness against population structure, such as population admixture and stratification. MDR has also been extended to family data (Chen et al. 2011; Lou et al. 2008; Martin et al. 2006). On the other hand, the power of family-based designs may decrease when the parental genotypes are uninformative. Although theoretically attractive, a family design is usually not as economically advantageous as an unrelated-individual design that is less laborious in sample collection. However, the genetic backgrounds of subjects in an unrelated-individual design can be quite different from each other, and if the population structure is not taken into account, false-positive and false-negative associations may arise and thus diminish the advantages of such designs. For unrelated subjects, methods have been proposed to infer genetic ancestry, such as genomic control (Devlin and Roeder 1999), structured association (Pritchard et al. 2000), and the principal components analysis (PCA) method (Price et al. 2006; Zhu et al. 2008) that provides a general solution for more complicated scenarios.

When data from both family-based and population-based studies are available, the ideal strategy is to combine the data, while eliminating the nuisance population structure that may inflate false-positive and false-negative rates. The consequently enlarged sample size will increase the chances of detecting gene–gene interactions. However, several practical issues arise in the application of this strategy. The major issues are how to correct for the population structure in founders of family samples and unrelated samples and how to pool two kinds of samples together. A realized solution in association studies is to correct with a fixed effect model for the structure that can be inferred through a PCA of unrelated individuals (Zhu et al. 2008).

Although the issues related to population structure and sample pooling have been well addressed in single-marker association studies, they remains unexplored in detection of interactions. The purpose of this study is to establish a general framework for detecting gene–gene interactions using unrelated and family samples. We proposed a unified nonparametric method, called unified generalized multifactor dimensionality reduction (UGMDR), which detects gene–gene interactions by incorporating both unrelated individuals and families. Simulations were conducted to demonstrate the benefit of the unified analysis to statistical power. A working example, from the Study of Addiction: Genetics and Environment (SAGE), was used to show the application of this method.

Materials and methods

Correction for population structure in family and unrelated samples

When the dataset consists of both unrelated and family samples, we need to correct for population structure and construct appropriate statistics for combining GMDR analysis. We use unrelated samples including unrelated founders in families to infer ancestral composition of the whole sample and compute the SNP loading for unrelated and children (see the supplementary method for details). Then we can adjust the phenotype of interest for eliminating effects of population structure by fitting a null generalized linear model (i.e., no effects of factors of interest), for example, a linear model $\boldsymbol{Y}=\boldsymbol{\mu}\mathbf{1}+\boldsymbol{PB}+\boldsymbol{Z\gamma}+\boldsymbol{\varepsilon}$ for a continuous phenotype, in which $\boldsymbol{Y}$ the vector for the phenotype, $\boldsymbol{\mu}$ is the grand mean of the model $\mathbf{1}$, is a vector of which all elements are $\mathbf{1}$, $\boldsymbol{P}$ is a $\boldsymbol{N} \times \boldsymbol{L}$ matrix representing the top $\boldsymbol{L}$ principal components for $\boldsymbol{N}$ individuals, $\boldsymbol{B}$ is a vector representing the effects of population structure, $\boldsymbol{Z}$ is the incidence matrix for the covariates such as age and gender, $\boldsymbol{\gamma}$ is the covariate effect vector, and $\boldsymbol{\varepsilon}$ is the vector of residuals. The population structure effects can be corrected by

$$\tilde{\boldsymbol{Y}}=\boldsymbol{Y}-\hat{\boldsymbol{\mu}}\mathbf{1}-\boldsymbol{P}\hat{\boldsymbol{B}}-\boldsymbol{Z}\hat{\boldsymbol{\gamma}} \quad (1)$$

$\tilde{\boldsymbol{Y}}$ is the adjusted phenotypic value for both the principal components and the covariates. In their approach, Zhu et al. (2008) suggested adjusting the genealogical effects both on the phenotypes and on the genetic markers, which is more theoretically attractive. In our approach that treats the markers as categorical variables, differing from the

typical regression methods that treat genetic markers as quantitative or count variables (e.g., the number of alleles of interest in an additive model), we adjust only on the phenotypes, but not on the genetic markers, as incorporating the principal components can substantially eliminate the confounding effects of other covariates. The resulting GMDR is valid in the sense of controlling correct type I error rates. As demonstrated in the simulations, the type I error rates were in good agreement with the given significance levels.

After adjusting for the principal components that account for the potential cryptic population structure, the phenotype and genotype will be independent under the null hypothesis. We use the adjusted phenotype to define an appropriate statistic and integrate with the multifactor data reduction strategy mentioned below. There are two kinds of data involved in the statistic: the siblings in nuclear families are genetically related, and parents of the nuclear families and the singletons are nominally unrelated. For unrelated subjects, the adjusted phenotype is used directly in the data reduction; for convenience of notation, we consider each unrelated individual as a family with only one member and denote statistic $\mathbf{s}_{ij} = \tilde{y}_{ij}$ where $j = 1$. For siblings, to take the genetic dependence among the relatives into account, the within-family association statistic is used in the data reduction—the within-family association statistic can be computed via the conditioning algorithms under the null hypothesis, e.g., the within-family association statistic of the jth individual in the ith nuclear family with respect to a combination of loci L, $\mathbf{s}_{ij}^{L} = \tilde{y}_{ij}\mathbf{g}_{L}(\mathbf{x}_{ij})$, where $\mathbf{g}_{L}(\mathbf{x}_{ij})$ is a function contrasting the transmitted genotype at locus combination L to its reference distribution under the null hypothesis (Chen et al. 2011). For simplicity of notation, we discard the sign of locus combination in $\mathbf{s}_{ij}^{L}$ and $\mathbf{g}_{L}(\mathbf{x}_{ij})$ thereinafter. The principle behind the conditioning algorithm is as follows: given a mating type (parental genotypes) or its minimal sufficient statistic, we have the reference genotypic distribution of offspring under the null hypothesis, denoted by G_{M}; different mating types have their respective genotypic distributions of offspring. Each of these genotypic distributions follows Mendel's law only, and thus is independent of any phenotype. Nevertheless, the observed (or transmitted) genotypic distribution of offspring may differ conditional on the mating type and a trait of interest in the presence of genotype–phenotype association, denoted by $G_{\mathrm{M,T}}$. The discrepancy between them will ascribe to the association of the combination of loci with the trait only, thus eventually eliminating the impact of locus-specific spurious association through comparison between G_{M} and $G_{\mathrm{M,T}}$. Detailed numerical examples for conditional genotype distribution on nuclear families can be found in Rabinowitz and Laird (2000).

Multifactor-reduction algorithm

Our method is devised by integrating the statistic defined in the previous subsection (i.e., $\mathbf{s}_{ij} = \tilde{y}_{ij}$ for unrelated subjects and $\mathbf{s}_{ij} = \tilde{y}_{ij}\mathbf{g}(\mathbf{x}_{ij})$ for siblings) into the GMDR framework, whose implementation of C-fold cross-validation (CV) is summarized as follows.

In step one, regardless of their familial or ethnic origins, individuals are assigned into C even or nearly even subdivisions. One of the subset is used as the testing set and the remaining one(s) as the training set. We set $C = 10$ throughout this report, but it can be other integers, such as $C = 5$ (Motsinger and Ritchie 2006).

In step two, a subset of γ factors are selected from all ω discrete factors of either genetic and/or environmental origin. A total of $\binom{\omega}{\gamma}$ distinct subsets can be chosen in this manner. Each such subset corresponds to a γ-dimensional finite grid, and each subject who is genotyped and assessed for the environmental exposures will fall into exactly one cell in this grid. The values of the statistic defined in the previous subsection are averaged over each cell. Each nonempty cell is labeled either high-risk if its average statistic value is not less than some threshold T, or low-risk otherwise. Without loss of generality, $T = \sum_{i=1}^{N}\sum_{j=1}^{K_i}\hat{s}_{ij}/N_T$, the mean of the sample, is used throughout the paragraphs below.

In step three, a multilocus model is formed by pooling high- and low-risk cells into two groups (i.e., high-risk and low-risk). The classification accuracy can be assessed by the averages of the statistic values in the high-risk group and the low-risk group, respectively.

In step four, the corresponding independent testing set (the set that is left out in steps two and three), is used to evaluate the testing accuracy for the model identified in step three. The testing accuracy is defined in the next subsection.

In step five, as there are C different pairs of training–testing sets, the above procedure is repeated for C rounds on the C training sets. The average testing accuracy over C testing sets can be calculated.

In step six, steps two to five are iterated for all other possible γ factor combinations, and the above procedure is repeated for $\binom{\omega}{\gamma}$ combinations.

Evaluation of *p* value

In each round of cross-validation, testing accuracy (TA) is defined as

Table 1 Design of the simulation experiments

	Design I	Design II	Design III	Design IV
Samples	200 families each with a discordant sibling pair and 200 cases and 200 controls	200 families each with three siblings and 200 cases and 200 controls	200 families each with three siblings and 500 cases and 500 controls	320 families each with three siblings and 200 cases and 200 controls
Case–control	400	400	1,000	400
Unrelated	800	800	1,400	1,040
Siblings	400	600	600	960
Total individuals	1,200	1,400	2,000	2,000

In design I, neither parent was affected, whereas in design II–IV at least one parent was affected

$$\mathrm{TA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \quad (2)$$

where TP is True Positive, defined as having a high-risk value in the high-risk group, TN is True Negative, defined as having a low-risk value in the low-risk group, FP is False Positive defined as having a low-risk value in the high-risk group, and FN is False Negative defined as having a high-risk value in the low-risk group. For a training set, the rule of classification guarantees that classification accuracy is not less than 0.5, whereas TA may be lower than 0.5 due to statistical fluctuation. TA has an expected value of about 0.5 under the null hypothesis. Over C-fold CV, the mean of TA, i.e., $\overline{\mathrm{TA}} = \sum_{i=1}^{C} \mathrm{TA}_i$, is calculated and employed as the test statistic for evaluating G × G interaction.

In general, we use a permutation method to determine empirical p value from the distribution of the permuted TAs under the null hypothesis. When the sample size is sufficiently large, as the result of the central limit theory, the p value can be approximately assessed by the normal distribution of the C-fold mean of TA under the null hypothesis. An approximate Z score is $Z = \frac{\overline{\mathrm{TA}} - E(\overline{\mathrm{TA}})}{\sqrt{\mathrm{var}(\overline{\mathrm{TA}})}}$. The mean and standard deviation of TA could also be computed through permutations. It should be noticed that there are two kinds of data involved in the test statistic. The siblings in nuclear families are genetically related, and parents of the nuclear families and the singletons are nominally unrelated. Although the genealogical effects of unrelated individuals can be adjusted through regression on the principal components, the family structure should be fully accounted for in building the test statistic. We use a hybrid strategy to evaluate the mean and empirical variance of the test statistic in permutations. As the genealogical effects of the unrelated individuals have already been adjusted, these singletons are exchangeable with each other in permutations, but the siblings are randomly shuffled only within the family because of the family structure effects. Permutations can be run for either the phenotype or the genotype at loci under consideration; both permutation schemes often yield nearly identical results. In this report, we permute phenotypes only.

Monte Carlo simulations

Systematic simulations were performed to investigate the power in various scenarios. A recent admixed population with a similar ancestry to African–Americans was simulated for the scenarios considered. Four study designs with different sample size and ratio of families to singletons were adopted in the simulation study as tabulated in Table 1. Various disease models, relative risks, and allele frequencies were considered in simulations (refer to the supplementary materials for details). To compare the proposed unified strategy with the separate analysis strategies, we computed the power of four methods: FAM for family-based method conditional on parental genotypes in which only sibs were used, CC for case–control method in which only case–control samples but no family samples were used, UN for method of unrelated individuals in which cases, controls and founders of families were used, and UI for the proposed unified method in which all cases, controls, founders of families, and siblings are used. These first three methods are used as the reference methods for power comparison.

UI was also compared with a benchmark method, the meta-analysis implemented with the Fisher's combining p value method for individual UN and FAM analyses (Fisher 1954). A Chi square test statistic with four degrees of freedom was computed from the p values of UN and FAM to determine the overall p value and statistical power.

A case study

In this study, we managed to detect interactions among genes in the cohort for SAGE. Majority of SAGE samples are unrelated, in addition to a few families, including, after

quality control, a total of 3,897 individuals from three subsamples: the Collaborative Study on the Genetics of Alcoholism (COGA) (1,178 individuals), the Collaborative Study on the Genetics of Nicotine Dependence (COGEND) (1,427 individuals) and the Family Study of Cocaine Dependence (FSCD) (1,292 individuals). Although many phenotypes were recorded, we were primarily interested in the genetic mechanism of nicotine dependence. SNPs in the nicotinic acetylcholine receptor (nAChR) $\alpha 4$ subunit (*CHRNA4*), the nAChR $\beta 2$ subunit (*CHRNB2*), the neurotrophic tyrosine kinase receptor 2 (*NTRK2*, also known as the tyrosine kinase receptor gene, *TrkB*), and the brain-derived neurotrophic factor (*BDNF*) were selected to detect the G $\times$ G interaction among these genes.

The PCA was run for the SAGE data to investigate the population mixture. The score statistics for nicotine dependence were computed in a logistic regression with adjustment for age, sex, and the top five principal components. The unified method proposed in our study was used for three sub-samples individually and the whole sample. As a contrast, the meta-analysis was also conducted with the Fisher's combining *p* value method.

Results

Simulation study

As the principal component method can precisely identify the ancestry of each individual (see the supplementary result section and supplementary Figs. 1, 2), we could use principal components to control population structure and get well-controlled type I error rates (supplementary Table 1). Simulations suggested UI in general outperformed the three reference methods in terms of power under various settings in the simulations (see supplementary Table 2 for the impact of the power due to the simulated factors simulated). Figure 1 presents the power comparison of UI to the three reference methods. As shown in the first vertical panels (on the left side in Fig. 1), the means of power over the 1,200 scenarios, denoted by the black circles whose coordinates in the horizontal and the vertical axes were the mean of UI and that of a method compared in each panel, respectively, were about 0.55 for UI, 0.22 for FAM, 0.21 for CC, and 0.41 for UN. In other words, UI had a higher, at least 0.14, average power than the other methods (Also see Table 2 for details). The dots below the green lines indicate the power values of the other three methods that were less than 80 % of UI, and most of those scenarios seemed to be of moderate statistical power values. And for those over the green lines, most were of powers close to 1 (few were close to zero), when relative risk was not less than 2.5 as indicated in panel B. In terms of power, the second best method was UN (Fig. 1a3), since around 35 % scenarios reached 80 % power of UI when relative risk not less than 2.5 under designs III and IV. In very few scenarios, the dots highlighted in brown in the first vertical panels, the powers of UI appeared to be lower than those of the other three methods, but their values were extremely low. It seemed to be more likely attributed to sampling errors. Compared with other two reference methods, UN had the closest power to UI (supplementary Fig. 3) probably because UN can use more individuals than CC and FAM (Table 1). As UI can use all individuals in the simulated samples, whereas the other three methods could only use a part of them, it seemed quite reasonable that UI outperformed other methods.

We also examined the influences of different relative risks on power. Under each factor, simulations under each relative risk level were plotted as scattered points according to the means of power of UI (*x*-axis) and a reference method (*y*-axis) in each panel, providing a straight comparison. Then the distributions of the points filled with different color elucidated the pattern of power value under each method. As anticipated, the power increased with the relative risk. UI appeared to increase power substantially by 0.46 when the relative risk was increased in the interval of 2.0 (mean power = 0.39) to 2.5 (mean power = 0.85), but increased by only 0.09 in the interval of 2.5–3.0. Similar trends were observed in the other three methods. When relative risk was as low as 1.5, neither UI nor a reference method demonstrated a practically appreciated statistical power regardless of the change of other factors.

The mean powers of UI were 0.45 and 0.53 for designs I and II, respectively, showing an improvement of $\sim$0.08 caused by the addition of one offspring in each of the 200 nuclear families. After adding other 600 individuals to design II, by either recruiting more unrelated samples in design III or family samples in design IV, the power increased to 0.606 and 0.607, respectively. This indicates that an increase in unrelated samples can give a power gain similar to an increase in nuclear families and, in practical application, we can adopt either of the two alternative recruitment schemes according to how easily the sample can be recruited.

The powers always increased corresponding to the magnitude of relative risk. The checkerboard models tended to have higher powers compared with the other two models. The general patterns are summarized in Fig. 1. The corresponding patterns could be connected to their causes. For example, in Fig. 1b2, red points (RR = 2.0) were clustered into two groups, and Table 1 indicates that the upper group was arisen from an increase in the case–control sample size by 600 individuals. When the alpha was decreased to 0.01, powers dropped off (supplementary Table 3). But the averaged powers (in bold font) dropped less with UI

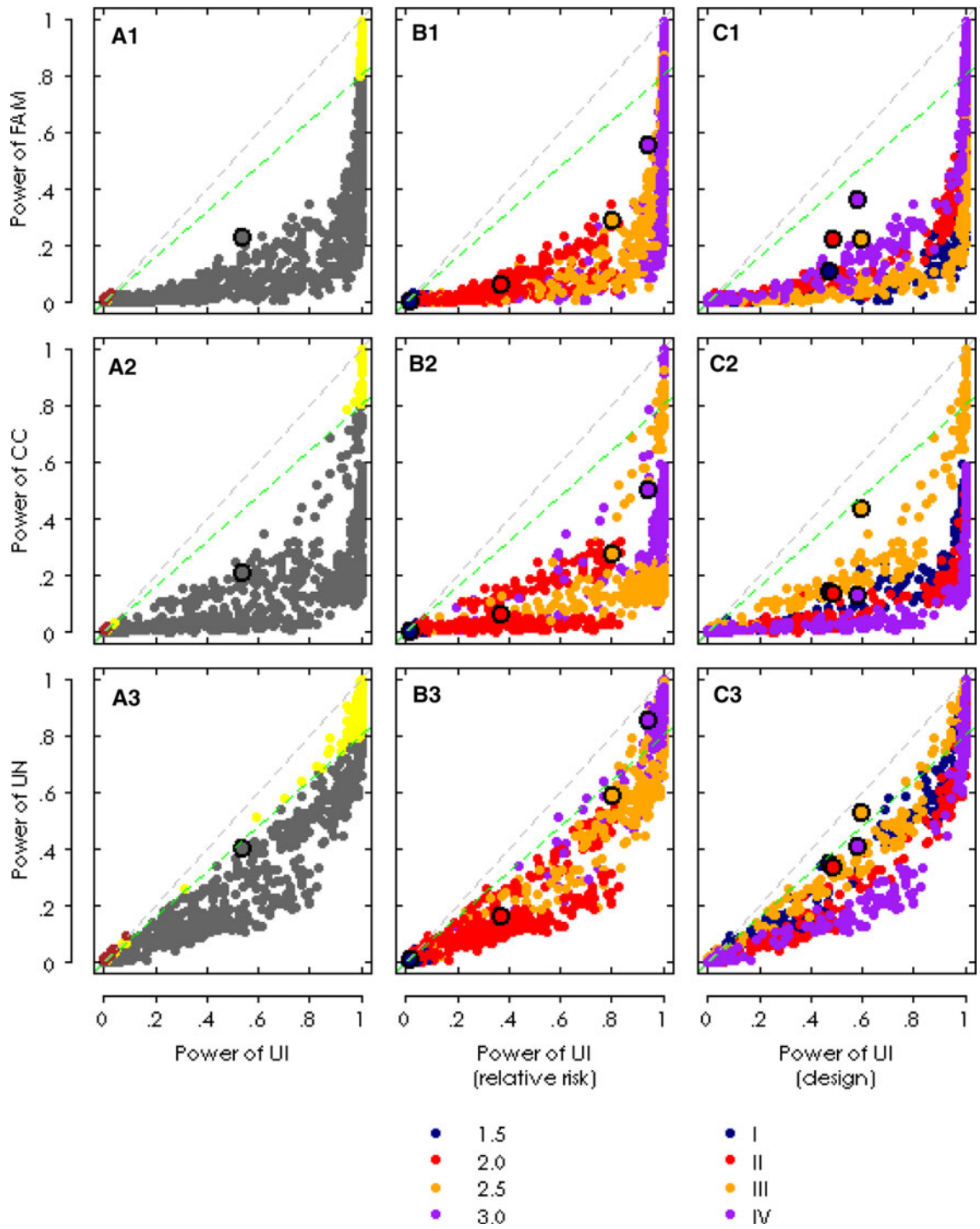

Fig. 1 Power comparison between the unified method and three reference methods given $\alpha = 0.05$. **a** The overall comparison of the power values. The *gray points* below the *blue lines* indicate the power values in their respective reference methods were less than 80 % as much as the UI method, *yellow points* (*yellow*) larger than 80 % but less than UI, the *brown points* greater than UI. The means of power are indicated with *black circle*, *filled with gray*, the horizontal value. **b** Comparison of the power with regard to the levels of relative risk. The means of power are indicated by the *black circles*, *filled with their respective color* of the level which they referred to. **c** Comparison of the power with regard to the study design. The means were represented the same as the *panel* **b** (color figure online)

compared with the reference methods. In design IV, the mean powers of UI decreased from 0.607 down to 0.519 (by 14 %), but for CC, from 0.138 down to 0.07 (by 49 %).

Meta-analysis is used as a method to strengthen the signal from independent studies. Although FAM method using siblings only and the UN method using unrelated

Table 2 Power comparison for admixture population at $\alpha = 0.05$

Design	I				II				III				IV			
Methods	FAM	CC	UN	UI	FAM	CC	UN	UI	FAM	CC	UN	UI	FAM	CC	UN	UI
RR = 1.5																
Checkerboard	0.005	0.003	0.005	0.017	0.008	0.003	0.007	0.026	0.006	0.012	0.023	0.047	0.012	0.003	0.009	0.037
Diagonal	0.002	0.003	0.006	0.013	0.004	0.001	0.003	0.013	0.003	0.008	0.012	0.027	0.004	0.001	0.005	0.016
Upper corner	0.000	0.001	0.004	0.008	0.001	0.001	0.003	0.008	0.001	0.006	0.009	0.015	0.003	0.002	0.005	0.016
RR = 2.0																
Checkerboard	0.036	0.023	0.119	0.324	0.079	0.025	0.122	0.405	0.080	0.249	0.457	0.719	0.221	0.028	0.217	0.693
Diagonal	0.019	0.024	0.092	0.210	0.034	0.019	0.074	0.261	0.037	0.161	0.286	0.477	0.127	0.026	0.145	0.486
Upper corner	0.006	0.016	0.065	0.120	0.016	0.015	0.052	0.152	0.022	0.133	0.248	0.376	0.049	0.013	0.083	0.274
RR = 2.5																
Checkerboard[a]	0.203	0.174	0.613	0.899	0.389	0.182	0.602	0.949	0.410	0.826	0.956	0.994	0.787	0.183	0.783	0.994
Diagonal	0.082	0.121	0.431	0.702	0.227	0.124	0.404	0.736	0.268	0.724	0.899	0.974	0.505	0.119	0.569	0.816
Upper corner	0.043	0.101	0.357	0.553	0.084	0.091	0.259	0.500	0.125	0.528	0.693	0.797	0.315	0.093	0.453	0.779
RR = 3.0																
Checkerboard	0.460	0.489	0.932	0.995	0.764	0.484	0.933	0.999	0.790	0.986	1.00	1.00	0.973	0.491	0.987	1.00
Diagonal	0.256	0.375	0.837	0.970	0.653	0.398	0.862	0.991	0.573	0.917	0.960	0.980	0.739	0.308	0.820	0.962
Upper corner	0.140	0.323	0.742	0.863	0.360	0.275	0.655	0.849	0.312	0.672	0.739	0.801	0.624	0.262	0.770	0.909
Mean power	0.104	0.138	0.350	0.473	0.218	0.135	0.331	0.491	0.219	0.435	0.524	0.601	0.363	0.127	0.404	0.582

FAM family-based method, *CC* case–control method, *UN* unrelated individuals (unbalanced case–control), *UI* unified method

[a] Power, the proportion of true models significant at the given significance level in all simulations. Each power in this table is the mean of 25 scenarios, each of which was sampled on allele frequency between 0.05 and 0.95 and then replicated 500 times

individuals can add up together to use the whole sample, their combined p values, determined by the Fisher's method, were not as powerful as our proposed unified method (supplementary Table 4), consistent with the results from single-marker association studies (Macgregor 2008; Skol et al. 2007).

Real data analysis

As illustrated in Fig. 2a, there were black, white, and mixed individuals in the SAGE cohort, and the admixed genetic background in fact was across each of the three subsamples in SAGE (Fig. 2b). In this sense, SAGE made itself a suitable sample for demonstrating the unified GMDR methods.

Recent studies revealed genetic associations with nicotine dependence of *CHRNA4* (Feng et al. 2004; Li et al. 2005), *NTRK2* and *BDNF* (Beuten et al. 2005). As indicated by biochemical studies, in the brain the $\alpha 4\beta 2$-containing nAChR subtype makes up the majority of the high-affinity nicotine-binding sites and that under chronic nicotine exposure the genes for both subunits are upregulated. In our previous study, we also discovered the interaction among *CHRNA4*, *CHRNB2*, *BDNF,* and *NTRK2* underlying nicotine dependence (Li et al. 2008; Lou et al. 2007). Given the SNP information (dbSNP, Build 135), SAGE sample was mapped to eight SNP markers in *CHRNA4*, four in *CHRNB2*, 25 in *BNDF*, and 130 in *NTRK2*, respectively, and in total it generated 104,000 ($8 \times 4 \times 25 \times 130$) tetragenic interactions, one SNP from each of the four genes. The phenotype of interest was nicotine dependence, of which SAGE had 1,765 nicotine-dependent individuals and 2,036 -nondependent individuals. The numbers of individuals that survived after quality control and also had the nicotine dependence phenotype are shown in Table 3 but the exact individuals used varied, due to missing genotypes or availability of other covariates, with each interaction model tested.

Table 3 Interaction SNPs detected among *CHRNA2*, *CHRNA4*, *BDNF,* and *NTRK2*

Model[a]	Effective individuals[b]	Variance contributed	Testing accuracy	p value
rs1013402–rs1044394–rs2072660–rs6559840				
SAGE	3,786 (134)	0.0176	0.5468	6.46e−06
FSCD	1,275 (121)	0.036	0.5428	5.81e−03
COGA	1,089 (5)	0.0352	0.5156	1.35e−01
COGEND	1,422 (6)	0.0125	0.4691	9.20e−01
META				2.48e−02

[a] In each model, from left to right, the SNPs are located in *BDNF*, *CHRNA4*, *CHRNB2*, and *NTRK2*, respectively

[b] The used individual for detecting each tetragenic interaction model, and in the parenthesis were the number of siblings

Using the unified method, we tested 104,000 tetragenic interaction models, which include one SNP marker from each gene. As expected, the distribution of the testing accuracy is a normal distribution (supplementary Fig. 4). Figure 3 shows Manhattan plots of the p values from the analyses of the whole sample and three sub-samples, and

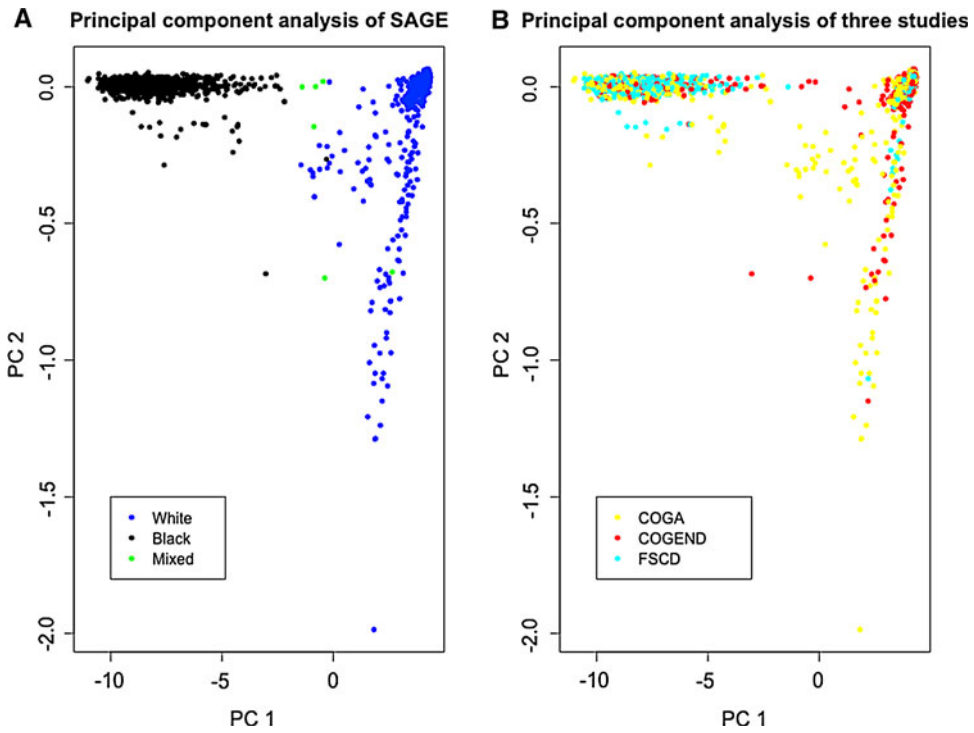

Fig. 2 The principal components analysis for SAGE. As there were relatives in the SAGE sample, two-stage method, as described in our method, was used for building principal components. **a** The genetic background of the SAGE sample were plotted into the first and the second PC axes. **b** The genetic background for the three sub samples in SAGE in the first and the second PC spaces

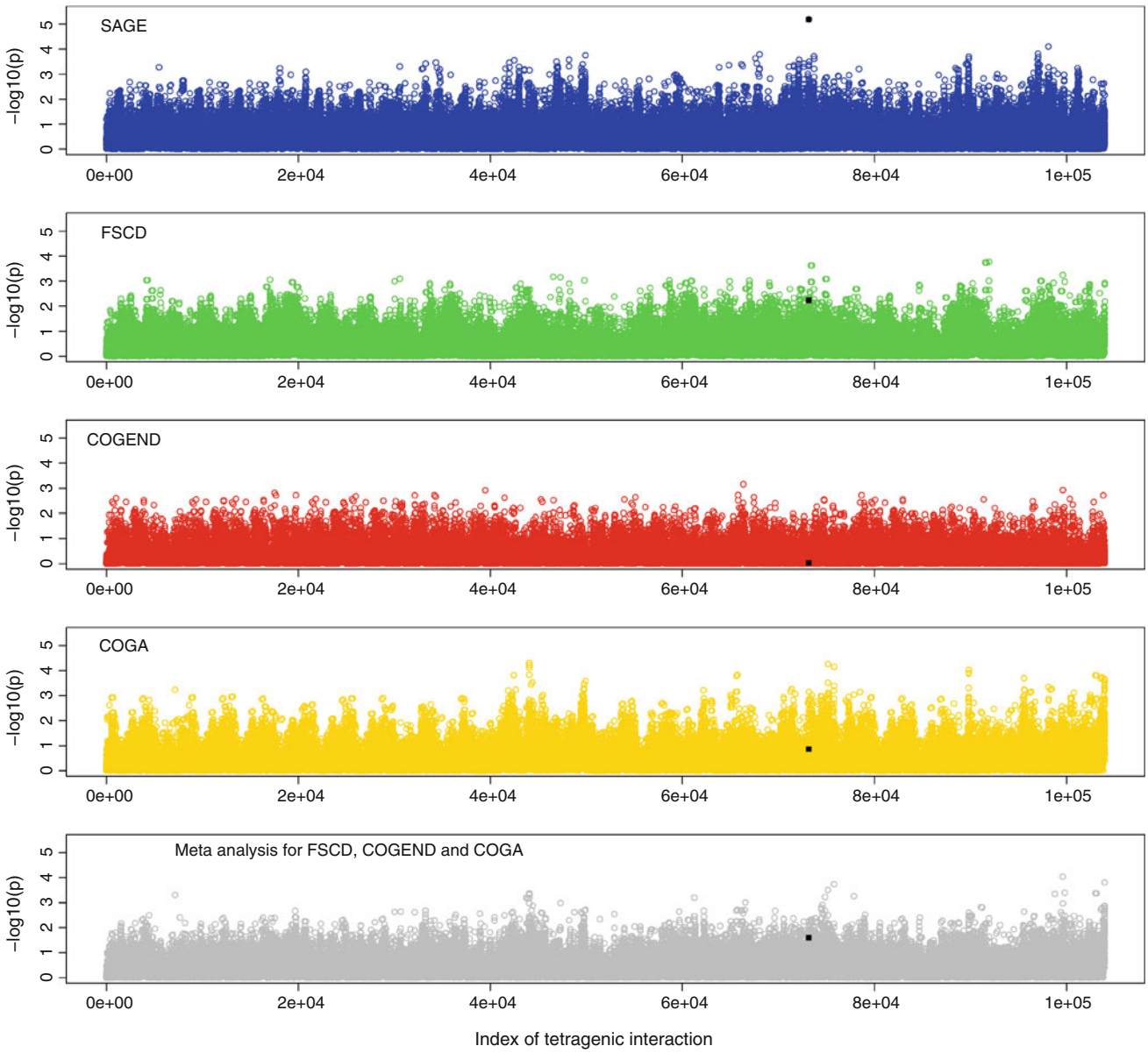

Fig. 3 Manhatton plots for SAGE, FSCD, COGEND, COGA, and the meta-analysis. The *black square* represents the interaction, rs1013402–rs1044394–rs2072660–rs6559840, which had the highest *p* value in SAGE and in the three sub samples

the meta-analysis, respectively. The most significant tetragenic interaction model was rs1013402–rs1044394–rs2072660–rs6559840, having a *p* value of 6.46e−06, which was detected in SAGE, whereas its *p* values in the each of the subsamples and the meta-analysis for the three subsamples were less significant. It should also be noticed that because of the high linkage disequilibrium between SNPs within genes, the practical threshold of *p* value would not be as conservative as the one given by Bonferroni correction. The *p* value to declare significance remains an open question for the detected interactions. However, accounting for our previous discovery (Li et al. 2008; Lou et al. 2007), this *p* value indicated that there was potential interaction of these four genes underlying nicotine dependence.

The high-risk and low-risk distribution of the identified multilocus models could be further illustrated (Fig. 4). The patterns of high-risk and low-risk cells varied across each of the different multilocus dimensions, presenting evidence of epistasis. With increasing interaction loci, it is possible, given limited sample size, that merging empty genotypic cells might decrease robustness of a model. The biological mechanism, partially as revealed (Li et al. 2008), underlying the tetragenic model requires further investigation both through in silico analysis and laboratory verification in the future. However, it should be noticed that genetic

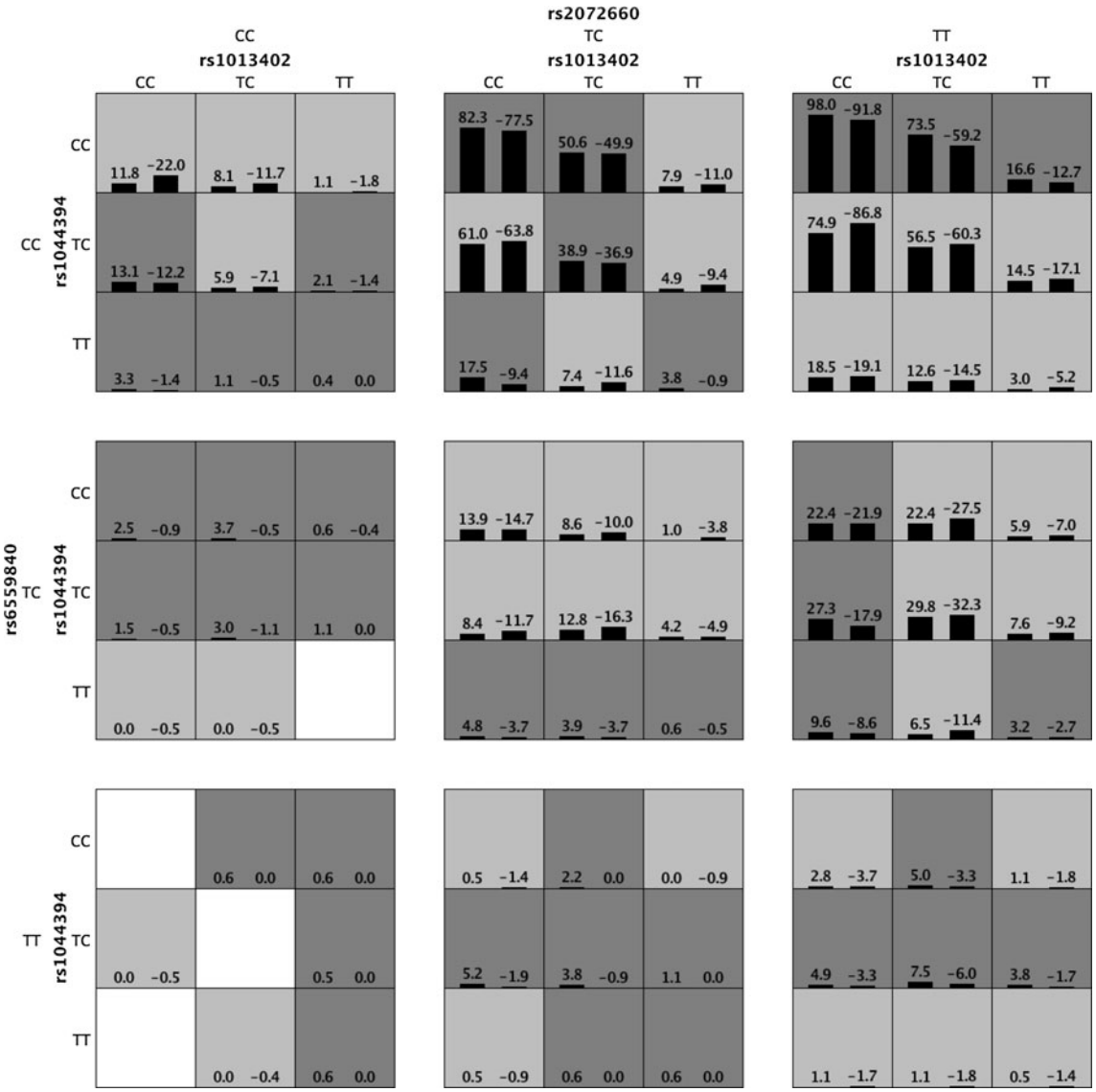

Fig. 4 The interaction pattern among rs1013402–rs1044394–rs2072660–rs6559840. In each cell, the *left bar* represents a positive score, and the *right bar* a negative score. High-risk cells are indicated by *dark shading*, low-risk cells by *light shading*, and empty cells by *no shading*. Note that the patterns of high-risk and low-risk cells differ across each of the different multilocus dimensions, presenting evidence of epistasis

heterogeneity on etiology has not been considered especially when across multiple cohorts of differential genetic admixture. It remains to be further investigated whether the variation of the strength of the illustrated tetragenic signals in three cohorts reflects power issue or various genetic etiologies.

Discussion

Detecting G × G interaction underlying complex traits is getting increasing attention in genetic studies. Many theoretical and application studies have revealed the importance of interactions in the formation of phenotypic outcomes (Zuk et al. 2012). There is little doubt that interactions among genes play an important role in the genetic architecture of complex traits. In order to foster drug development and establish proper medical interventions, identifying G × G can be crucial. There are a few terms, such as statistical interaction, epistatic interaction, and additive interaction, commonly used in describing gene–gene interactions, as summarized in the literature (Wang et al. 2010). In our study, the interpretation of the interaction defined in this report is close to multilocus model or joint action of genes. Once interaction models of interest are identified using the method proposed, follow-

up analysis might be applied depending on the purpose of the study. If statistical interaction is of interest, for example, main effects and interaction effects can be further estimated for a detected multilocus model. Given the method introduced in this report, after correction for population stratification, the unified GMDR can maximize the number of individuals available in the sample. As demonstrated in our simulations, the unified method had higher power in many of the scenarios simulated.

For the unified method proposed, the first step is to capture the genetic background of the sample. Currently, a couple of methods have been proposed (Price et al. 2010). PC coordinates can be inferred from the unrelated set of the sample (Zhu et al. 2008), such as used in our method, or inferred from another independent dataset. As demonstrated in our study, this method, either applied to an admixed population or a discrete population, can extract population structure in terms of ancestral origin very well and consequently control the type I error rate. The genetic interpretation of the first principal component was well connected to the averaged coalescent times between populations and Wright's F_{st} statistics (McVean 2009), whereas the interpretation of admixed population, such as African Americans, requires careful modeling of the historical gene flow (Gravel 2012). Alternatively, mixed model approaches are also applied to control for population structure which may inflate the type I error rate (Wu et al. 2011). From the viewpoint of genetics, both methods used nearly the same genetic information. Using PCA tends to consider the effects due to genetic origin as fixed, whereas mixed model approaches treat them as random. So far, no conclusion is reached on which method is more appropriate in application. Although we demonstrated the advantage of improving statistical power after adjusting population structure by PCA, which are served as covariates in building the score statistic, the impact of including covariates may depend on other factors, such as prevalence of the disease. It should be noticed that decreasing power may occur under some scenarios (Pirinen et al. 2012). This topic deserves further investigation in relation to our method.

We assume there is no relatedness between the case–control individuals and the founders, who are used in estimating eigenvectors. In some cases, particularly in the context of samples from isolated populations, cryptic relatedness may be problematic or there exists known relatedness. The kinship coefficients or estimated kinship coefficients need to be incorporated into the statistical model for eliminating relatedness effects (Bourgain et al. 2003; Choi et al. 2009). Furthermore, stringent quality control should be applied for excluding subjects with cryptic relatedness from GMDR analysis when it is an issue.

In the real data analysis, although our previous analysis (Lou et al. 2007) also detected tetragenic interaction analysis for nicotine dependence, a case–control sample, with 382 subjects, was used at that time and as a result the p value was not as significant as demonstrated in the present study. Classifying nicotine dependence into cases and controls also results in loss of information and consequently in underestimating the genetic variance it explains. The real data analysis in this report showed more advantageous in power than our previous study. With our proposed method, it used a much bigger sample size, superior in statistical power after its population stratification had been corrected. As a typical complex trait, the genetic variants underlying the formation of nicotine dependence may be tiny in effect size but large in their total number. To reveal their genetic architecture, using single-marker based methods seems insufficient. As demonstrated in this report, as well as previous success cases, methods such as UGMDR may empower the discovery of the genetic determinants.

Limitations should be noticed for the proposed method. As it is currently designed for pooling unrelated sample with siblings from the nuclear families, complex pedigrees require more sophisticated calculation, such as attempt of controlling population structures with linear models or using flexible permutation strategies. In theory, the unified method can be extended to complex pedigrees but will increase the computational time exponentially with the generations included. The computational challenge and multiple testing problem also pose another hurdle in practice, especially for detecting high-order interactions for the whole genome data. More theoretical and computational work is required to address these challenges.

Acknowledgments This work was funded in part by the National Institutes of Health Grants DA025095, GM081488, GM077490, HG003054, and DK080100. Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative (GEI) (U01 HG004422). The datasets used for the analyses described in this manuscript was obtained from the database of Genotypes and Phenotypes (dbGaP) found at http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000092.v1.p1 through dbGaP accession number phs000092.v1.p.

Conflict of interest The authors declare no conflict of interest.

References

Beuten J, Ma JZ, Payne TJ, Dupont RT, Quezada P, Huang W, Crews KM, Li MD (2005) Significant association of BDNF haplotypes in European–American male smokers but not in European–American female or African–American smokers. Am J Med Genet B Neuropsychiatr Genet 139:73–80

Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS (2003) Novel case–control test

in a founder population identifies P-selectin as an atopy-susceptibility locus. Am J Hum Genet 73:612–626

Chen GB, Zhu J, Lou XY (2011) A faster pedigree-based generalized multifactor dimensionality reduction method for detecting gene–gene interactions. Stat Interface 4:295–304

Choi Y, Wijsman EM, Weir BS (2009) Case–control association testing in the presence of unknown relationships. Genet Epidemiol 33:668–678

Cordell HJ (2009) Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet 10:392–404

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Feng Y, Niu T, Xing H, Xu X, Chen C, Peng S, Wang L, Laird N (2004) A common haplotype of the nicotine acetylcholine receptor alpha 4 subunit gene is associated with vulnerability to nicotine addiction in men. Am J Hum Genet 75:112–121

Fisher AR (1954) Statistical methods for research workers, 12th edn. Hafner, New York

Gravel S (2012) Population genetics models of local ancestry. Genetics 191:607–619

Li MD, Beuten J, Ma JZ, Payne TJ, Lou XY, Garcia V, Duenes AS, Crews KM, Elston RC (2005) Ethnic- and gender-specific association of the nicotinic acetylcholine receptor alpha4 subunit gene (CHRNA4) with nicotine dependence. Hum Mol Genet 14:1211–1219

Li MD, Lou XY, Chen G, Ma JZ, Elston RC (2008) Gene-gene interactions among CHRNA4, CHRNB2, BDNF, and NTRK2 in nicotine dependence. Biol Psychiatry 64:951–957

Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet 80:1125–1137

Lou XY, Chen GB, Yan L, Ma JZ, Mangold JE, Zhu J, Elston RC, Li MD (2008) A combinatorial approach to detecting gene–gene and gene–environment interactions in family studies. Am J Hum Genet 83:457–467

Macgregor S (2008) Optimal two-stage testing for family-based genome-wide association studies. Am J Hum Genet 82:797–799 (author reply 799–800)

Manolio T, Collins F, Cox N, Goldstein D, Hindorff L, Hunter D, McCarthy M, Ramos E, Cardon L, Chakravarti A, Cho J, Guttmacher A, Kong A, Kruglyak L, Mardis E, Rotimi C, Slatkin M, Valle D, Whittemore A, Boehnke M, Clark A, Eichler E, Gibson G, Haines J, Mackay T, McCarroll S, Visscher P (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH (2006) A novel method to identify gene–gene effects in nuclear families: the MDR-PDT. Genet Epidemiol 30:111–123

McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5:e1000686

Motsinger AA, Ritchie MD (2006) The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genet Epidemiol 30:546–555

Pirinen M, Donnelly P, Spencer CC (2012) Including known covariates can reduce power to detect genetic effects in case–control studies. Nat Genet 44:848–851

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

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463

Pritchard J, Stephens M, Rosenberg N, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

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

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2007) Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31:776–788

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Wang X, Elston RC, Zhu X (2010) The meaning of interaction. Hum Hered 70:269–277

Wu C, DeWan A, Hoh J, Wang Z (2011) A comparison of association methods correcting for population stratification in case–control studies. Ann Hum Genet 75:418–427

Zhu X, Li S, Cooper RS, Elston RC (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 82:352–365

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109:1193–1198