Published in final edited form as:

Genet Epidemiol. 2018 July ; 42(5): 447–458. doi:10.1002/gepi.22116.

Powerful and Robust Cross-Phenotype Association Test for Case-Parent Trios

S. Taylor Fischer^{1,*}, Yunxuan Jiang^{2,*}, K. Elaine Broadaway¹, Karen N. Conneely¹, and Michael P. Epstein¹

¹Department of Human Genetics and Center for Computational and Quantitative Genetics, Emory University, Atlanta, GA

²Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA

Abstract

There has been increasing interest in identifying genes within the human genome that influence multiple diverse phenotypes. In the presence of pleiotropy, joint testing of these phenotypes is not only biologically meaningful but also statistically more powerful than univariate analysis of each separate phenotype accounting for multiple testing. While many cross-phenotype association tests exist, the majority of such methods assume samples comprised of unrelated subjects and therefore are not applicable to family-based designs, including the valuable case-parent trio design. In this paper, we describe a robust gene-based association test of multiple phenotypes collected in a case-parent trio study. Our method is based on the kernel distance covariance (KDC) method, where we first construct a similarity matrix for multiple phenotypes and a similarity matrix for genetic variants in a gene; we then test the dependency between the two similarity matrices. The method is applicable to either common variants or rare variants in a gene, and resulting tests from the method are by design robust to confounding due to population stratification. We evaluated our method through simulation studies and observed that the method is substantially more powerful than standard univariate testing of each separate phenotype. We also applied our method to phenotypic and genotypic data collected in case-parent trios as part of the Genetics of Kidneys in Diabetes (GoKinD) study and identified a genome-wide significant gene demonstrating cross-phenotype effects that was not identified using standard univariate approaches.

Keywords

pleiotropy; case-parent trio design; genetic association testing

Address for Correspondence: Michael P. Epstein, Ph.D., Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322, Phone: (404)712-8289, mpepste@emory.edu.

*Contributed equally to this work

Web Resources

dbGaP: <https://www.ncbi.nlm.nih.gov/gap>

Epstein software: <http://genetics.emory.edu/labs/epstein/software/>

Introduction

Pleiotropy has been an increasingly important topic in the genetic association literature since power may be gained by studying the joint influence of a single gene on multiple phenotypes (Kocarnik & Fullerton, 2014; Solovieff, Cotsapas, Lee, Purcell, & Smoller, 2013). A pleiotropic effect occurs when a single molecular function affects multiple biological processes (He & Zhang, 2006). A large number of genes demonstrate pleiotropic effects: a 2011 review article examined NHGRI (National Human Genome Research Institute)'s catalog of common variants and found 233 (16.9%) genes that were significantly associated with multiple traits (Sivakumaran et al., 2011). The importance of studying genes or genetic variants with cross-phenotype effects also extends to secondary phenotypes. For example, diabetes studies typically measure systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL), and body mass index (BMI) as secondary phenotypes. Joint analysis of these phenotypes not only provides biological insight but also increases effective sample size and subsequently improves power (Diggle, 2002; Galesloot, Van Steen, Kiemeney, Janss, & Vermeulen, 2014).

While a variety of statistical methods exist for testing the association between genetic variants and multiple phenotypes (Broadaway et al., 2016; Galesloot et al., 2014; Guo, Liu, Wang, & Zhang, 2013; Maity, Sullivan, & Tzeng, 2012; O'Reilly et al., 2012; Schifano, Li, Christiani, & Lin, 2013; Wang, Meigs, & Dupuis, 2016; Yang & Wang, 2012; Zhao & Thalamuthu, 2011; Zhou & Stephens, 2014), several limitations remain in this area. The majority of existing methods test association between individual genetic variants and multiple phenotypes. However, gene-based tests that jointly consider multiple variants in a region of interest have several advantages over single marker tests. First, gene-based tests combine signals from variants within a gene together, which makes them particularly appealing for rare-variant sequencing studies where individual rare variants may be difficult to detect. Second, gene-based tests can adopt dimension-reduction tools, such as kernel machines, to lower the multiple-comparison burden and accommodate complex relations between markers as well as non-linear effects between genetic variants and phenotypes. Several approaches have used gene-based testing to improve upon traditional single-marker tests (Kwee, Liu, Lin, Ghosh, & Epstein, 2008; Wu et al., 2010) but typically focused on a single phenotype rather than multiple phenotypes.

To address this issue, Maity et al. (2012) proposed a gene-based test of multiple phenotypes for common variants using a multivariate version of kernel machine regression (MV-KMR), which has similar form to a multivariate linear mixed model. Similar to other KMR methods (Kwee et al., 2008; Wu et al., 2010; Wu et al., 2011), they also adopted a variance-component score test to reduce multiple testing burden. However, limitations arise as the MV-KMR method 1) is only applicable to continuous traits, 2) only allows linear correlations between phenotypes, and 3) requires computationally intensive permutation procedures to calculate p-values. To tackle this issue, Broadaway et al. (2016) proposed a method for gene-based testing of multiple rare variants using a kernel distance covariance (KDC) framework. Unlike the method of Maity et al., this method not only defines a similarity matrix among genetic variants in a gene but also defines a similarity matrix for phenotypes. The KDC framework then tests whether the individual elements of the

phenotype similarity matrix are independent of the individual elements of the genotype similarity matrix. This method, called GAMuT, can accommodate binary or continuous traits, and the test statistics asymptotically follow a mixture of chi-square distributions, which makes the calculation of p-values straightforward.

While the methods of Maity et al. (2012) and Broadaway et al. (2016) are valuable, they are also limited to studies of unrelated subjects. It would be useful to create a gene-based test of multiple phenotypes applicable to family designs, particularly the valuable case-parent trio design that has been utilized in both GWAS and next-generation sequencing studies. An attractive feature of case-parent trio designs is that, by leveraging the Quantitative Transmission Disequilibrium Testing (QTDT) framework (Abecasis, Cardon, & Cookson, 2000), one can create tests that by design are robust to population stratification. Population stratification occurs when samples originate from multiple populations with different disease prevalence and different distributions of minor allele frequencies. It can lead to power loss and substantially inflated type I error rates when left unaddressed (Epstein et al., 2012; Jiang, Epstein, & Conneely, 2013).

In this manuscript, we propose a novel gene-based test for cross-phenotype association testing in case-parent trio studies that is robust to population stratification. We base our approach on the kernel distance-covariance (KDC) framework utilized in the GAMuT test (Broadaway et al., 2016) but replace the observed genotype information in that test with robust within-family genotypic information derived from the QTDT framework (Abecasis et al., 2000). In the following sections, we first introduce how we construct our test statistics using the KDC framework and how we make these statistics robust to population stratification via the QTDT framework. If phenotype data are available on parents within each trio, we further describe how to leverage this information as a screening tool to further improve power of the robust approach (using similar ideas as discussed in Ionita-Laza et al. (2007) and Jiang et al. (2014)). We evaluate our method using simulations and further compare results of our method with robust gene-based testing of univariate phenotypes using the robust framework kernel-machine regression (RF-KMR) test of Jiang et al. (2014). We also apply our method to a real GWAS case-parent trio study of type 1 diabetes-related phenotypes collected by the Genetics of Kidneys in Diabetes (GoKinD) study. Finally, we conclude with a summary of our findings and discussion of future extensions.

Materials and Methods

Notation

We assume a sample of N case-parent trios (parents and offspring) that are genotyped in a gene or region of interest and are measured for multiple phenotypes (continuous or binary, in nature). Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iL})$ denote the L phenotypes for the offspring in family i , where $i=1, 2, \dots, N$. Let $\mathbf{G}_i = (G_{i1}, G_{i2}, \dots, G_{iS})$, a $1 \times S$ vector, denote the S genotyped variants for the same offspring, where G_{is} is coded as number of minor alleles that the offspring possesses at site s . The variants in the gene can consist of either common variants or rare variants. We further define $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iC})$ as a $1 \times C$ vector of covariates for the offspring. Let $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \dots, \mathbf{Y}_N^T)^T$ denote the $N \times L$ matrix of offspring phenotypes in the dataset and let $\mathbf{G} = (\mathbf{G}_1^T, \mathbf{G}_2^T, \dots, \mathbf{G}_N^T)^T$ denote the $N \times S$ matrix of offspring genotypes.

Kernel Distance Covariance Test of Independence

We wish to create a robust association test between phenotypes $\mathbf{Y}$ and gene-based genotypes $\mathbf{G}$ for case-parent trios using the kernel distance-covariance (KDC) framework (Gretton et al., 2007). To do this, we leverage the work of Broadaway et al. (2016), who showed how to create such a KDC-based test (Gene Association with Multiple Traits, named “GAMuT”) for gene-based testing of multiple phenotypes in population-based studies. We develop their approach first and then discuss how we leverage their work to develop the robust test of cross-phenotype effects for case-parent trio studies.

The GAMuT test of Broadaway et al. (2016) is based on the independence test between kernels on reproducing kernel Hilbert spaces (RKHS) first introduced by Bach et al. (2002). For Hilbert spaces, Bach et al. showed that the canonical correlation of two kernels equals zero if and only if the two variables are independent. Based on this finding, Gretton et al. (2007) extended the test to use the Hilbert-Schmidt norm as a measure to test the independence between two kernels. The advantages of their method are: 1) the calculation is straightforward and computational complexity is proportional to the square of the sample size, which is appealing for high-dimensional genomics data, and 2) the test statistics asymptotically follow a mixture of chi-square distributions, which makes the calculation of p-value efficient to derive. These characteristics make KDC ideal for testing independence between a kernel similarity matrix based on multiple phenotypes, $\mathbf{Y}$, and a kernel similarity matrix based on multiple variants in $\mathbf{G}$, the gene/region of interest.

Based on these findings, Broadaway et al. (2016) first constructed a similarity matrix between phenotypes ($\mathbf{Y}$) and a similarity matrix between genotypes ($\mathbf{G}$), and then tested for dependency between the two similarity matrices. Let $\mathbf{P}$ denote the phenotype similarity matrix, where commonly used similarity methods to construct $\mathbf{P}$ include the construction of a projection matrix $\mathbf{P}=\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$ (Wessel & Schork, 2006) or linear kernels $\mathbf{P}=\mathbf{Y}\mathbf{Y}^T$. Let $\mathbf{K}$ denote the genotype similarity matrix, where commonly used methods to construct $\mathbf{K}$ include construction of an identity by state (IBS) kernel where the $(i,j)^{th}$ element of the matrix is $K(\mathbf{G}_i, \mathbf{G}_j) = \sum_{s=1}^S IBS(G_{is}, G_{js})/2S$, where $IBS(G_{is}, G_{js})$ calculates the number of alleles shared in common by subjects i and j at the s^{th} variant or a weighted linear kernel $\mathbf{K}=\mathbf{G}\mathbf{T}\mathbf{G}^T$, where $\mathbf{T}=\text{diag}(\text{weight}_1, \text{weight}_2, \dots, \text{weight}_s)^T$ is a diagonal matrix with relative weight for each variant. Typical weights include minor-allele frequency (such that rarer alleles are upweighted over more common alleles) or functional information (Wu et al., 2010). Choice of kernels depends on prior assumptions of the relationships between phenotypes or genotypes; appropriate choice of the kernel can increase power (Kwee et al., 2008; Schaid, 2010; Wu et al., 2010).

The GAMuT test of Broadaway et al. (2016) is based on the test of Bach et al. (2002) which relies on centered kernels for inference. Therefore we further define a centering matrix $\mathbf{H} = (\mathbf{I} - \mathbf{1}_N \mathbf{1}_N^T / N)$ (Schölkopf, Smola, & Müller, 1998), where $\mathbf{I}$ is a N dimensional identity matrix and $\mathbf{1}_N$ is a $N \times 1$ vector of 1, such that $\mathbf{K}_c = \mathbf{H}\mathbf{K}\mathbf{H}$ and $\mathbf{P}_c = \mathbf{H}\mathbf{P}\mathbf{H}$ are centered matrices. Following the above notation, GAMuT is constructed as

$$Q = \frac{1}{N} \text{trace}(\mathbf{K} \mathbf{C} \mathbf{P} \mathbf{C}). \quad (1)$$

Under the null hypothesis of no association, the test statistic follows the asymptotic mixed chi-square distribution defined as $\frac{1}{N^2} \sum_{m,n} \lambda_{P_{C,m}} \lambda_{K_{C,n}} z_{mn}^2$, where $\lambda_{K_{C,n}}$ is the n th non-zero eigenvalue of $\mathbf{K} \mathbf{C}$, $\lambda_{P_{C,m}}$ is the m th non-zero eigenvalue of $\mathbf{P} \mathbf{C}$, and z_{mn} are independent and identically distributed standard normal variables. GAMuT uses Davies' method (Davies, 1980) to analytically calculate the p-value of Q in (1), thereby avoiding the need for computationally expensive permutations for inference.

Kernel Distance Covariance Test for Case-Parent Trios

As we discussed in the introduction, the limitation of the original GAMuT test based on KDC is it cannot be used to construct robust cross-phenotype association tests for use in case-parent trio studies. To address this issue, we propose a robust modification of the GAMuT test statistic for trio data by integrating the QTDT (Abecasis et al., 2000) framework within the KDC framework. The QTDT framework decomposes the observed genotype of a trio offspring G_{is} into a between-family component, B_{is} , and a within-family component W_{is} . B_{is} is calculated as the average genotype of the offspring's parents, which can be viewed as the mean genotype of the founder's sub-population and is thus sensitive to population stratification. The within-family component W_{is} is calculated as $G_{is} - B_{is}$. Because W_{is} can be viewed as the deviate of the observed genotype from the sub-population mean, it is thus robust to population stratification. Thus, to create a robust gene-based test of multiple phenotypes for use in case-parent trios, we first use parental genotypes to calculate B_{is} and W_{is} for each variant in the gene. We then create a $N \times S$ matrix $\mathbf{W}$ with $(i,s)^{\text{th}}$ element corresponding to W_{is} and then construct the genotype similarity matrix $\mathbf{K}$ using $\mathbf{W}$ instead of $\mathbf{G}$, such that (for example) $\mathbf{K} = \mathbf{W} \mathbf{T} \mathbf{W}^T$ under a weighted linear kernel function. Using this modified version of $\mathbf{K}$, we then construct the score statistic Q in (1) as previously shown. By modifying the GAMuT test to use within-family information, our resulting test is robust to population stratification.

While the use of within-family information makes our approach robust to population stratification, the discarding of between-family information can lead to a nontrivial loss in power (Ionita-Laza, Lee, Makarov, Buxbaum, & Lin, 2013a). Thus, our robust method will lose power compared to a GAMuT-based test of the observed offspring genotypes when the latter test is valid (i.e. no confounding exists). To recover some of this lost power for the within-family test, we can in certain situations follow a similar strategy to that of Jiang et al. (2014) and leverage between-family information as a screening tool. Specifically, if parental phenotype and genotype information are available, we perform a first-stage GAMuT association test of each gene using the phenotypes and genotypes of the parents. We then identify those genes with the smallest first-stage p-values and follow these genes in the second stage using the robust test that compares offspring phenotypes to within-family genotypes. Since the first- and second-stage analyses are independent (since they use independent pieces of genotype and phenotype information), the two-stage procedure

preserves size. On the other hand, by reducing the number of within-family tests needing to be performed due to the first-stage screening procedure, we can improve the overall power of our robust test.

Simulations

We first evaluate the type I error rate and power of our method using simulated data. For each simulation, we use the coalescent simulator *cosi* (Schaffner et al., 2005) to simulate a pool of 5,000 European haplotypes and 5,000 African haplotypes, each 30kb long. *Cosi* uses a coalescent model to simulate haplotypes based on empirical patterns of genetic variation observed in different ancestral populations. To simulate sequencing data for trios, we first randomly select haplotypes within each population and pair them for the father and mother of the trio. We then randomly select one haplotype from each parent to form the offspring's haplotypes. In order to examine whether our method is robust to population stratification, we assume the sample consists of 75% trios of African origin and 25% trios of European origin. We further assume the mean trait difference between European and African subjects is 0.3 (R^2 : 0.69). We consider tests both of common variation as well as rare variation in a gene. Rare variants are defined as variants with minor allele frequency less than 3%. Common variants are defined as variants with minor allele frequency greater than 5%.

For type I error rate simulations, we assume six phenotypes are recorded, and the residual correlation among them follows a multivariate normal distribution with pairwise-trait correlation sampled from a uniform distribution (0,0.3). To examine the robustness of our model to confounding due to population stratification, we assume two, four, or six phenotypes are affected by population stratification. We constructed the test both using the observed genotype information (corresponding to the GAMuT test) and just using the robust within-family component. For the analysis of binary phenotypes, continuous phenotype measurements in the top quartile were considered affected ($Y_{ij}=1$) and measurements falling below the 75th percentile were considered unaffected ($Y_{ij}=0$).

For power simulations, we also assume six phenotypes are recorded. As discussed in Kocarnik et al. (2014), pleiotropy involves both highly correlated phenotypes and phenotypes that are very diverse. To simulate this, we again assume that continuous phenotypes follow a multivariate normal distribution but that the pairwise trait correlations are drawn from a uniform distribution. We separately consider scenarios reflecting low correlation (pairwise correlation of phenotypes drawn from a uniform distribution with bounds (0,0.3)), medium correlation (0.3,0.5), or high correlation (0.5,0.7). We also varied the number of phenotypes that are truly associated: we assume either two or four of the six phenotypes are associated with the causal variants in the region. For the rare variants test, we assume that 5% of rare variants in the region are causal. For each causal variant, we simulated the effect as $\beta = (0.4 + N(0,0.1)) \times |\log_{10} MAF|$ such that less frequent alleles have larger effects on the outcome (Wu et al., 2010). For the common variants test, we assume that 1% of common variants in the region are causal. For each causal common variant, we assume a fixed effect where $\beta = \log_e 1.5$, which is based on the work of Ionita-Laza et al. (2013b). For tests of continuous traits, we compare our method with RF-KMR, a robust univariate gene-based test of a continuous phenotype using within-family information

(Jiang et al., 2014) that does not consider multiple phenotypes simultaneously. As RF-KMR tests each phenotype individually, it is necessary to adjust for the six tests performed. As the phenotypes are correlated, direct application of the Bonferroni correction will be conservative. We instead calculate the number of effective tests, $M_{effective}$, as the number of principal components able to explain 90% of the variance of phenotypes. We then calculate the adjusted threshold as $\alpha_{effective} = \alpha / M_{effective}$. Compared to traditional Bonferroni correction, this threshold will achieve appropriate type I error rate and increased power.

We performed additional simulations to address how much of a power loss our robust method suffered compared to using observed genotype information when confounding due to population stratification does not exist and furthermore investigated whether our proposed two-stage strategy for testing can help restore the performance of the robust method in such instances. Here we simulate only European haplotypes. Given N trios, we separately consider population-based samples of $\frac{1}{2}N$, $\frac{3}{4}N$, and N unrelated individuals. We again consider six continuous phenotypes with medium pairwise trait correlation, of which either two or four are assumed to be truly associated with the causal variants in the region. For the trios, we implemented the two-stage screening procedure described by Jiang et al. (2014) where observed parental genotypes (from which the between-family component is derived) across several gene regions are first tested for cross-phenotype associations using the GAMuT method. We then use the resulting p-values to select a subset of top gene regions of interest to test using the robust within-family information in offspring. The tests of the first stage are independent of the tests of the second stage due to orthogonality of between- and within-family components.

GoKinD Data Analysis

As a real-data application example, we applied our method to common variants from a case-parent trio GWAS study of type 1 diabetes from the GoKinD study (Mueller et al., 2006; Pezzolesi et al., 2009). While GoKinD was initially designed to identify genes associated with diabetic nephropathy in type 1-diabetes patients, the study collected additional phenotypes that can potentially provide more insights in this line of research, such as systolic blood pressure (SBP), diastolic blood pressure (DBP), high-density lipoprotein (HDL), and body mass index (BMI). Via dbGaP (see Web Resources), the study made available phenotype and genotype data on 584 parent-offspring trios on dbGaP (dbGaP accession numbers phs000018.v2.p1 and phs000088.v1.p1). All subjects were genotyped using the Affymetrix Mapping 500K array. We used the annotation file from the 1000 Genomes Project to identify common SNPs that fell within known genes. After excluding genes with less than two common variants, 9,647 genes and 131,366 SNPs were included in our analysis. We used our novel cross-phenotype test to test the association between the 9,647 genes and SBP, DBP, HDL, BMI. We also adjusted for important covariates in our model: gender, age, renal function status (proteinuric, dialysis, renal transplant or other), smoking status, insulin intake (yes or no), anti-hypertension drug intake (yes or no), lipid lowering medication intake (yes or no). We applied both our method and univariate RF-KMR testing to the dataset.

Results

Type I Error Rate

We first applied our method to 5,000 simulated datasets that were subjected to confounding. For each simulation, we sampled 500 trios (125 European and 375 African) from the pool of 10,000 simulated haplotypes. We constructed the test statistics using both the original GAMuT test based on observed genotype (sensitive to population stratification) as well as our modified approach based only on the within-family component (which is robust to population stratification). We chose the weighted linear kernel to form the genotype similarity matrix, where weight was generated through the beta distribution density function evaluated at the minor allele frequency: $\text{weight} \sim \text{beta}(\text{MAF}, 1, 25)$ (Wu et al., 2011). We further chose the projection matrix to form the phenotype similarity matrix.

We summarize type I error rates using Quantile-Quantile (QQ) plots in Figure 1 (rare variants) and Figure 2 (common variants). In the presence of population stratification, the distribution of p-values for the GAMuT test of observed genotypes significantly deviated from the expected uniform distribution (top panels of Figures 1 and 2). As more phenotypes are affected by stratification, the deviation becomes more extreme. However, our method (bottom panels of Figure 1 and 2) yields the expected distribution of p-values under the null under all circumstances, suggesting that the correct type I error rate is achieved at all significance levels. These observations also hold for rare variant and common variant tests of binary phenotypes (Figures S1 and S2, respectively).

Power

Our type I error rate simulations showed that our method using the within-family component is robust to population stratification. In this section, we evaluate the power of our method and, where possible, compare the results with the robust univariate test, RF-KMR, which is the within-family KMR method of Jiang et al. (2014). We applied our method to 1,000 simulated datasets and evaluated power at type I error rate of 0.05. Similar to above, each simulation sampled 500 trios (125 European and 375 African) from the pool of haplotypes.

As described in Methods, we simulated such that 5% of rare variants in the region were causal, and we varied the number of phenotypes associated with the causal variants. Simulation results for continuous traits are summarized in Figure 3, where we see our method (dark gray bars) outperforms univariate kernel machine regression (light gray bars). Our method can capture the correlation between phenotypes: power increases as correlation between phenotypes increases (Figure 3, left to right), while the univariate test cannot exploit this information. Power also increases as the number of phenotypes associated with the causal variants increases for all levels of phenotypic correlation (Figure 3). For the common variants test, we simulated such that 1% of common variants in the region were causal, but with an effect size smaller than that assumed for rare variants. For common variant tests of continuous phenotypes, we also observed that our method achieves higher power compared to the univariate test under all simulation settings (Figure 4). We show power results for rare variant and common variant tests of binary traits in Figures S3 and S4,

respectively. As expected, the power of our test for binary phenotypes generally increases with an increase in the number of binary phenotypes associated with causal variants.

Two-Stage Screening Procedure

While our results have shown that the use of within-family information provides necessary protection against inflated type I error in the presence of population stratification, discarding the between-family genotypic information can incur a noticeable loss in power compared to tests on observed genotypes. Unsurprisingly, in the absence of population stratification, the power of tests on observed genotypes from a population-based sample exceeds the power of tests on within-family information of trio offspring when sample sizes are equivalent ($N=500$; Figure S5). We therefore assessed the extent to which power could be recovered using the two-stage screening approach of Jiang et al. (2014). Here, we again simulated haplotypes for 500 parent-offspring trios (125 European, 375 African). In stage one, we analyzed the observed parental genotypes by applying the GAMuT rare-variant association test separately to ten 30kb regions. We then selected the top one, two, three, or four regions with lowest p-values for testing with robust within-family information in offspring. We see that appropriate type I error is maintained under the null across all screening scenarios (Figure S6) and that, assuming four phenotypes are associated with the causal variants, the power of the robust test can indeed be recovered by initial screening gene regions using observed parental phenotype and genotype data (Figure S7).

GoKinD Data Analysis

Using genotype and phenotype data from the GoKinD study available from dbGaP (see Web Resources and Acknowledgements), we tested the phenotypes SBP, DBP, HDL, and BMI for association with common variants. We removed variants whose missing rate is larger than 5%. For each phenotype, we replaced missing values with the median value of the phenotype. The final sample consisted of 544 trios with genotypes on 131,366 common variants from 9,647 genes, with a median of 13.6 variants in each gene. We analyzed the data using our method, which tests all four phenotypes simultaneously, and the univariate RF-KMR method, which tests each phenotype individually. For both tests, we tried using both linear kernels and weighted linear kernels to form the genetic similarity matrix. We also adjusted for age, gender, renal function status, smoking status, insulin intake, anti-hypertension drug intake, and lipid lowering medication intake. To adjust for covariates in our method, we first regress each phenotype separately on the covariates, and then use the residuals from each regression to form the phenotype similarity matrix (linear kernel). We assume a Bonferroni-adjusted genome-wide significance level of $0.05/9647 \approx 5 \times 10^{-6}$ and a suggestive level of 1×10^{-4} . As the RF-KMR method tests each phenotype individually, we adjusted for multiple testing through the following procedure: we first find the minimum p-value among the four tests, and then multiply it by an estimate of the effective number of tests (the number of principal components that can explain 90% variation of the four phenotypes). This procedure is less conservative than Bonferroni correction, allowing for a fairer comparison between methods.

We first formed the genetic similarity matrix using the linear kernel and summarized the results in Manhattan plots and QQ plots shown in Figure 5 (for the cross-phenotype

approach) and Figure 6 (for univariate analysis using RF-KMR). By utilizing information from the correlation between phenotypes, Figure 5 shows our novel method identified a gene Vacuolar Protein Sorting 41 (*VPS41*, containing 47 SNPs in our data) on chromosome 7 that passes the genome-wide significance threshold. We also formed the genetic similarity matrix using a weighted linear kernel, where weight is calculated as $-\log_{10}(MAF)$ and, for these weighted analyses, *VPS41* remained the top hit and nearly approached genome-wide significance as well (Figure S8). We also repeated analyses imputing missing phenotypes to the mean rather than median and the Manhattan and QQ plots using such imputation were qualitatively similar to those that imputed to the median (Figure S9). On the other hand, univariate analyses of these same phenotypes (Figure 6 and Figure S10 for unweighted and weighted linear kernel, respectively) failed to identify any genome-wide significant genes across the genome. Moreover, the minimum univariate p-value of *VPS41* was > 100 fold larger than the corresponding cross-phenotype p-value, thereby showing the potential value of jointly analyzing multiple correlated phenotypes together in association analysis.

Discussion

In this paper, we introduced a robust multivariate method for identifying genes associated with one or more phenotypes in case-parent trio studies. Our method is a gene-based method that can incorporate prior information about the gene and is suitable for testing sets of either rare or common variants. The test is a non-parametric test based on the KDC framework. The framework is applicable to high-dimensional data, and the resulting test statistic follows an asymptotic distribution enabling efficient calculation of p-values. Our method further incorporated the QTD framework with the KDC framework to make the model robust to population stratification. We performed simulation studies on both sets of rare variants and sets of common variants; both studies showed that our method is more powerful compared to robust univariate testing of phenotypes using techniques we previously developed (Jiang et al., 2014). We make our code available through the web resource (<http://genetics.emory.edu/labs/epstein/software/>).

While we restricted our robust approach to case-parent trios in this manuscript, extensions of our strategy to enable robust gene-based association testing of multiple phenotypes in general pedigrees is certainly feasible. If we possess continuous phenotypes and assume a linear kernel for such phenotypes, we can represent our test in multivariate kernel-machine regression frameworks previously used by Jiang et al. (2014) and Jiang et al. (2017) and add an additional random effect within the model to handle within-subject correlation of the multivariate phenotype data. For a more flexible test, another possibility to attempt is to implement the GAMuT test directly on offspring genotypes conditional on sufficient statistics (using the approach of Rabinowitz and Laird (2000)) and then account for within-family correlation of outcomes and conditional offspring genotypes using a statistical-whitening procedure (Kessy, Lewin, & Strimmer, 2017). Another possible technique would be to modify the robust family-based U-statistic approach of Li et al. (2015), which currently handles univariate phenotype and genotype data, to handle multivariate phenotype and genotype data. As the authors' approach uses a kernel function to model phenotype similarity between relatives, extension of the framework to handle multivariate phenotypes is straightforward. Extending the approach of Li et al. (2015) to multiple genetic variants is

more challenging but, by constructing the authors' test statistic as a sum over the Euclidean distances across each variant, it may be possible to create a closed-form test for inference.

A reviewer pointed out a recent method by Won et al. (2015) for gene-based association testing of common variants with multiple phenotypes in family-based studies. The approach constructs an omnibus statistic based on a quasi-likelihood score function where phenotypes are adjusted with an offset based on incorporating the best linear unbiased predictor from a linear mixed model. To correct for population stratification, the method does not use parental genotype information to ensure robustness and validity like our proposed method does. Rather, the method requires estimating the genetic relationship matrix using genomewide genetic data. As such, the method of Won et al. is not applicable in settings without genomewide data, such as targeted resequencing projects, whereas our method can be applied in this setting. Furthermore, the work of Won et al. produces omnibus tests of association that, under the null hypothesis, follow a chi-square distribution with degrees of freedom equal to MQ , where M equals the number of variants and Q equals the number of phenotypes. By not accounting for correlation among phenotypes and correlation among genotypes (as our method does), the test needlessly expends degrees of freedom and will likely lose power compared to kernel-based approaches particularly as the number of phenotypes and genotypes considered becomes increasingly large.

Our method tests on either rare or common variants in a gene. There is a growing interest in examining the combined effect of rare and common variants. Ionita-Laza et al. (2013b) proposed such a framework for univariate tests of phenotypes. In their paper, they constructed the combined test as the weighted sum of the test statistics from rare and common variant tests, where the weights can be assigned using prior knowledge. It should be easy to incorporate their method into our framework to test the combined effect of rare and common variants on multiple phenotypes.

Applying our method to publicly available data from the GoKinD Study, we identified a gene not previously reported to associate with diabetes-related phenotypes (DBP, SBP, HDL, and BMI). *VPS41* is a member of Vesicle mediated protein sorting family, which plays an important role in segregation of intracellular molecules into distinct organelles. Previous work has shown that *VPS41* associates with class C VPS proteins to form the complete homotypic fusion and protein sorting (HOPS) complex (Plemel et al., 2011). Expression studies have shown that *VPS41* is potentially involved in the formation and fusion of transport vesicles from the Golgi.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants GM117946 and HG007508. We thank the reviewers for their helpful comments. The dataset used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype (dbGaP) found at <http://www.ncbi.nlm.nih.gov/gap> through dbGaP accession numbers phs000018.v2.p1 and phs000088.v1.p1. The Genetics of Kidneys in Diabetes (GoKinD) Study was conducted by the GoKinD Investigators and supported by the Juvenile Diabetes Research Foundation, the CDC, and the Special

Statutory Funding Program for Type 1 Diabetes Research administered by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data [and samples] from the GoKinD study were supplied by the NIDDK Central Repositories. This manuscript was not prepared in collaboration with investigators of the GoKinD study and does not necessarily reflect the opinions or views of the GoKinD study, the NIDDK Central Repositories, or the NIDDK.

References

- Abecasis G, Cardon L, Cookson W. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*. 2000; 66(1):279–292. [PubMed: 10631157]
- Bach FR, Jordan MI. Kernel independent component analysis. *Journal of machine learning research*. 2002; 3(Jul):1–48.
- Broadaway KA, Cutler DJ, Duncan R, Moore JL, Ware EB, Jhun MA, ... Epstein MP. A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. *Am J Hum Genet*. 2016; 98(3): 525–540. DOI: 10.1016/j.ajhg.2016.01.017 [PubMed: 26942286]
- Davies RB. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1980; 29(3):323–333.
- Diggle, PJ., Heagerty, P., Liang, K-Y., Zeger, SL. *Analysis of longitudinal data*. Oxford University Press; 2002.
- Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *The American Journal of Human Genetics*. 2012; 91(2):215–223. [PubMed: 22818855]
- Galesloot TE, Van Steen K, Kiemeny LA, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. *PLoS One*. 2014; 9(4):e95923. [PubMed: 24763738]
- Gretton, A., Fukumizu, K., Teo, CH., Song, L., Schölkopf, B., Smola, AJ. A kernel statistical test of independence. Paper presented at the Advances in neural information processing systems; 2007.
- Guo X, Liu Z, Wang X, Zhang H. Genetic association test for multiple traits at gene level. *Genetic epidemiology*. 2013; 37(1):122–129. [PubMed: 23032486]
- He X, Zhang J. Toward a molecular understanding of pleiotropy. *Genetics*. 2006; 173(4):1885–1891. [PubMed: 16702416]
- Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur J Hum Genet*. 2013a; 21(10): 1158–1162. DOI: 10.1038/ejhg.2012.308 [PubMed: 23386037]
- Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*. 2013b; 92(6):841–853. [PubMed: 23684009]
- 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(3):607–614. DOI: 10.1086/519748 [PubMed: 17701906]
- Jiang Y, Conneely KN, Epstein MP. Flexible and robust methods for rare-variant testing of quantitative traits in trios and nuclear families. *Genet Epidemiol*. 2014; 38(6):542–551. DOI: 10.1002/gepi. 21839 [PubMed: 25044337]
- Jiang Y, Conneely KN, Epstein MP. Robust Rare-Variant Association Tests for Quantitative Traits in General Pedigrees. *Statistics in Biosciences*. 2017; doi: 10.1007/s12561-017-9197-9
- Jiang Y, Epstein MP, Conneely KN. Assessing the impact of population stratification on association studies of rare variation. *Human heredity*. 2013; 76(1):28–35. [PubMed: 23921847]
- Kessy A, Lewin A, Strimmer K. Optimal Whitening and Decorrelation. *The American Statistician*. 2017; doi: 10.1080/00031305.2016.1277159
- Kocarnik JM, Fullerton SM. Returning pleiotropic results from genetic testing to patients and research participants. *JAMA*. 2014; 311(8):795–796. [PubMed: 24481117]
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*. 2008; 82(2):386–397. [PubMed: 18252219]

- Li M, He Z, Schaid DJ, Cleves MA, Nick TG, Lu Q. A Powerful Nonparametric Statistical Framework for Family-Based Association Analyses. *Genetics*. 2015; 200(1):69. [PubMed: 25745024]
- Maity A, Sullivan PF, Tzeng Ji. Multivariate Phenotype Association Analysis by Marker-Set Kernel Machine Regression. *Genetic epidemiology*. 2012; 36(7):686–695. [PubMed: 22899176]
- Mueller PW, Rogus JJ, Cleary PA, Zhao Y, Smiles AM, Steffes MW, ... Krolewski AS. Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. *Journal of the American Society of Nephrology*. 2006; 17(7):1782–1790. [PubMed: 16775037]
- O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, Coin LJ. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*. 2012; 7(5):e34861. [PubMed: 22567092]
- Pezzolesi MG, Poznik GD, Mychaleckyj JC, Paterson AD, Barati MT, Klein JB, ... Bochenski J. Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. *Diabetes*. 2009; 58(6):1403–1410. [PubMed: 19252134]
- Plemel RL, Lobingier BT, Brett CL, Angers CG, Nickerson DP, Paulsel A, ... Merz AJ. Subunit organization and Rab interactions of Vps-C protein complexes that control endolysosomal membrane traffic. *Mol Biol Cell*. 2011; 22(8):1353–1363. DOI: 10.1091/mbc.E10-03-0260 [PubMed: 21325627]
- 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(4):211–223. DOI: 10.1159/000022918 [PubMed: 10782012]
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome research*. 2005; 15(11):1576–1583. [PubMed: 16251467]
- Schaid DJ. Genomic similarity and kernel methods II: methods for genomic information. *Human heredity*. 2010; 70(2):132–140. [PubMed: 20606458]
- Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics*. 2013; 92(5):744–759. [PubMed: 23643383]
- Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*. 1998; 10(5):1299–1319.
- Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, ... Campbell H. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*. 2011; 89(5):607–618. [PubMed: 22077970]
- Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*. 2013; 14(7):483–495.
- Wang S, Meigs JB, Dupuis J. Joint association analysis of a binary and a quantitative trait in family samples. *Eur J Hum Genet*. 2016; 25(1):130–136. DOI: 10.1038/ejhg.2016.134 [PubMed: 27782109]
- Wessel J, Schork NJ. Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis. *Am J Hum Genet*. 2006; 79(5):792–806. [PubMed: 17033957]
- Won S, Kim W, Lee S, Lee Y, Sung J, Park T. Family-based association analysis: a fast and efficient method of multivariate association analysis with multiple variants. *BMC Bioinformatics*. 2015; 16:46.doi: 10.1186/s12859-015-0484-5 [PubMed: 25887481]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*. 2010; 86(6):929–942. [PubMed: 20560208]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*. 2011; 89(1): 82–93. [PubMed: 21737059]
- Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of probability and statistics*. 2012. 2012
- Zhao, J., Thalamuthu, A. Gene-based multiple trait analysis for exome sequencing data. Paper presented at the BMC proceedings; 2011.

Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods*. 2014; 11(4):407–409. DOI: 10.1038/nmeth.2848 [PubMed: 24531419]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Q-Q plots of p-values for gene-based tests of rare variants with six null continuous phenotypes using 5,000 simulations. Top panel: tests on observed genotype. Bottom panel: tests on within-family component. Left panel: two phenotypes affected by population stratification. Middle panel: four phenotypes affected by population stratification. Right panel: six phenotypes affected by population stratification.

Figure 2.

Q-Q plots of p-values for gene-based testing of common variants with six null continuous phenotypes using 5,000 simulations. Top panel: tests on observed genotype. Bottom panel: tests on within-family component. Left panel: two phenotypes affected by population stratification. Middle panel: four phenotypes affected by population stratification. Right panel: six phenotypes affected by population stratification.

Figure 3.

Power for gene-based testing of rare variants with six continuous phenotypes. Dark gray bar: power of cross-phenotype test using KDC. Light gray bar: power of test using univariate RF-KMR with adjusted Bonferroni to correct for multiple comparisons. Left panel: two phenotypes associated with the causal rare variants. Right panel: four phenotypes associated with the causal rare variants. All tests are constructed using robust within-family component. The results are based on 1,000 simulations.

Figure 4.

Power for gene-based testing of common variants with six continuous phenotypes. Dark gray bar: power of cross-phenotype test using KDC. Light gray bar: power of test using univariate RF-KMR with adjusted Bonferroni to correct for multiple comparisons. Left panel: two phenotypes associated with the causal common variants. Right panel: four phenotypes associated with the causal common variants. All tests are constructed using robust within-family component. The results are based on 1,000 simulations.

Figure 5. Cross-phenotype test on GoKinD data. Top: Manhattan plot for cross-phenotype test with linear kernel. Red line: genome-wide significance level. Blue line: suggestive level. Bottom: Quantile-Quantile plot.

Figure 6. RF-KMR test on GoKinD data. Top: Manhattan plot for RF-KMR test with linear kernel. Red line: genome-wide significance level. Blue line: suggestive level. Bottom: Quantile-Quantile plot.