Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2019 June 01.

Published in final edited form as:

Genet Epidemiol. 2018 June ; 42(4): 320–332. doi:10.1002/gepi.22121.

Testing Cross-Phenotype Effects of Rare Variants in Longitudinal Studies of Complex Traits

Pratyaydipta Rudra[*,1], **K. Alaine Broadaway**[2], **Erin B. Ware**[3,4], **Min A. Jhun**[3], **Lawrence F. Bielak**[3], **Wei Zhao**[3], **Jennifer A. Smith**[3], **Patricia A. Peyser**[3], **Sharon L.R. Kardia**[3], **Michael P. Epstein**[2], and **Debashis Ghosh**[1]

[1]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO

[2]Department of Human Genetics, Emory University, Atlanta, GA

[3]Department of Epidemiology, University of Michigan, Ann Arbor, MI

[4]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI

Abstract

Many gene mapping studies of complex traits have identified genes or variants that influence multiple phenotypes. With the advent of next-generation sequencing technology, there has been substantial interest in identifying rare variants in genes that possess cross-phenotype effects. In the presence of such effects, modeling both the phenotypes and rare variants collectively using multivariate models can achieve higher statistical power compared to univariate methods that either model each phenotype separately or perform separate tests for each variant. Several studies collect phenotypic data over time and using such longitudinal data can further increase the power to detect genetic associations. While rare-variant approaches exist for testing cross-phenotype effects at a single time point, there is no analogous method for performing such analyses using longitudinal outcomes. In order to fill this important gap, we propose an extension of Gene Association with Multiple Traits (GAMuT) test, a method for cross-phenotype analysis of rare variants using a framework based on the distance covariance. The approach allows for both binary and continuous phenotypes and can also adjust for covariates. Our simple adjustment to the GAMuT test allows it to handle longitudinal data and to gain power by exploiting temporal correlation. The approach is computationally efficient and applicable on a genome-wide scale due to the use of a closed-form test whose significance can be evaluated analytically. We use simulated data to demonstrate that our method has favorable power over competing approaches and also apply our approach to exome chip data from the Genetic Epidemiology Network of Arteriopathy.

Keywords

Complex human traits; gene mapping; longitudinal data; pleiotropy; rare variant

[*]**Correspondence:** 13001 E. 17th Place, Aurora, CO 80045, pratyaydipta.rudra@ucdenver.edu, Phone: 919-699-4043.

Web-resources

GAMuT software: https://epstein-software.github.io/GAMuT/

OMIM catalog: https://www.omim.org/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1 Introduction

Pleiotropy refers to the phenomenon of one genetic variant influencing more than one distinct trait. Several studies in recent years suggest the existence of many genetic variants that influence multiple phenotypes (Lees, Barrett, Parkes, & Satsangi, 2011; Liu et al., 2012). In the presence of such cross-phenotypic effects, joint genetic analysis of multiple phenotypes can be more accurate for phenotype prediction and statistically more powerful for gene mapping than univariate methods that model each phenotype separately (Maier et al., 2015; Galesloot, Van Steen, Kiemeney, Janss, & Vermeulen, 2014; Allison et al., 1998). Genetic pleiotropy induces phenotypic correlation which is more readily detectable through cross-phenotype analyses using the extra information provided by the correlation among the phenotypes. Although several tests of pleiotropy for common variants exist, they are usually less powerful for rare variants (Schaid et al., 2016). A recent method called 'Gene Association with Multiple Traits (GAMuT)' (Broadaway et al., 2016) was proposed to fill this gap for rare variants. The authors also argued that performing cross-phenotypic tests for rare variants might be more important than performing the same types of analyses for common variants since population-genetic models suggest rare variants are likely to be pleiotropic in nature under the model of infintesimal genetic architecture (Broadaway et al., 2016).

In many studies related to genetic epidemiology, such as the Genetic Epidemiology Network of Arteriopathy (GENOA) study (Daniels et al., 2004), observations at multiple time points are available for each subject. More accurate inference can be drawn by exploiting the temporal correlation in these measurements. However, most of the researchers tend to use existing single time point methodologies on such data after collapsing the repeated measurements into a single value (e.g. average across time points). Such a simple approach fails to take advantage of the extra information provided by repeated measurements and can be less powerful as a result.

A rare-variant statistical approach for cross-phenotype analysis of longitudinal outcomes requires a framework that can handle multiple phenotypes observed over multiple time points and furthermore can simultaneously handle information from multiple rare variants within a gene, since gene-based testing can be more powerful than testing of individual variants (Morris & Zeggini, 2010; He et al., 2016; M. C. Wu et al., 2011; Kwee, Liu, Lin, Ghosh, & Epstein, 2008). Such an approach currently does not exist in the statistical or genetics literature. There exist methods for longitudinal analysis of genetic data (Fan et al., 2012; Furlotte, Eskin, & Eyheramendy, 2012) based on random effects models (Fitzmaurice, Laird, & Ware, 2012) or generalized estimating equations (GEE) (Zeger & Liang, 1986). However, such models cannot be applied to test the association of the longitudinal phenotype data with an entire gene or thousands of markers taken together in a flexible manner. A recent method based on the longitudinal genetic random field model allows longitudinal analysis of multiple genetic variants simultaneously (He et al., 2016), but no extension of the method is available for cross-phenotypic effects. Finally, there are other multi-marker approaches such as sequence kernel association test (SKAT) (M. C. Wu et al., 2011) and similarity regression (SIMreg) (Tzeng, Zhang, Chang, Thomas, & Davidian, 2009) that are used for gene mapping, but such approaches do not directly apply to multi-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

phenotype data or longitudinal data. There has been some recent work on multivariate extensions to SKAT (B. Wu & Pankow, 2016; Sun et al., 2016). However, like GAMuT (Broadaway et al., 2016), these methods do not apply to longitudinal data. Another method using multivariate functional linear models (MFLM) (Wang et al., 2015) uses multi-phenotype data, but does not apply to longitudinal data. MFLM has also been shown to have much inferior power compared to GAMuT and SKAT in the single phenotype case (Broadaway et al., 2016). Moreover, the above tests require continuous phenotypes and therefore cannot be applied to important categorical phenotypes like presence or absence of a disease, which is not ideal for a test of pleiotropy.

In this article, we propose an extension of the GAMuT method for longitudinal data using some simple adjustments. The approach utilizes the correlation across time since it does not collapse the repeated measures, and also uses the correlation across phenotypes. Therefore it is appropriate for testing the genetic association for cross-phenotype longitudinal data and is especially powerful for rare variants. We demonstrate its power and control of type-I error through simulations and also apply the method to perform exome-chip analysis of multivariate repeated measures of cardiovascular-related phenotypes from the GENOA study (Daniels et al., 2004).

2 Methods

The GAMuT test (Broadaway et al., 2016) relies on a machine-learning framework called kernel distance-covariance (KDC) (Gretton et al., 2007; Hua & Ghosh, 2015; Székely, Rizzo, Bakirov, et al., 2007) and provides a non-parametric test for independence between a set of phenotypes and a set of genetic variants within a gene of interest. The framework is based on comparing pairwise phenotypic similarity between samples with genotypic similarity between samples. It allows for arbitrary number of genotypes and phenotypes, and therefore it is ideal for testing rare variants. It also allows for covariates and is computationally efficient due to its ability to provide analytic p-values using Davies' (Davies, 1980) method.

Let us start with a single time point. Suppose a sample of N subjects has data on L phenotypes. The phenotype vector for the j^{th} subject ($j = 1,2,\ldots,N$) is $P_j = (P_{j1}, P_{j1}, \ldots, P_{jL})$. Each phenotype can be either continuous or categorical. Let $P = \left(P_1^T, P_2^T, \ldots, P_N^T\right)$ be the phenotype matrix for all samples. Similarly, the genotype vector for sample j at V rare-variant sites in the gene of interest, $G_j = (G_{j1}, G_{j2}, \ldots, G_{jV})$, and the genotype matrix $G = \left(G_1^T, G_2^T \ldots, G_N^T\right)$ for all samples can be defined. Here G_{jv} ($j = 1,2,\ldots,N; v = 1,2,\ldots,V$) is the number of copies of the minor allele that the subject possesses at variant v. P is an $N \times L$ matrix and G is an $N \times V$ matrix, and kernel similarity measure between these two matrices is used by GAMuT to test the independence between the set of rare variants and the set of phenotypes.

To apply the GAMuT test, an $N \times N$ phenotypic similarity matrix Y and $N \times N$ genotypic similarity matrix X are first defined. Different kernels can be used to define the similarity matrices. For example, Y can be modeled using a projection matrix (Schork & Zapala,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2012), such that $Y = P(P^T P)^{-1} P^T$. A linear, quadratic or Gaussian kernel can also be used. Let the kernel function $y(P_i, P_j)$ denote the similarity between phenotype vectors of subject i and subject j. For linear kernel, $y(P_i, P_j) = \sum_{l=1}^{L} P_{il} P_{jl}$; for quadratic kernel, $y(P_i, P_j) = \left(1 + \sum_{l=1}^{L} P_{il} P_{jl}\right)^2$; and for Gaussian kernel, $y(P_i, P_j) = \exp\left(-\sum_{l=1}^{L} (P_{il} - P_{jl})^2\right)$ (M. C. Wu et al., 2011; Kwee et al., 2008; Schaid, 2010). Similar kernels can be used for the genotypes.

Once the similarity matrices X and Y are defined, they are centered as $X_c = HXH$ and $Y_c = HYH$, where $H = I - 1_N 1_N^T / N$ is the centering matrix, I being the identity matrix of dimension N, and 1_N being the $N \times 1$ vector with each element equal to 1. The GAMuT test statistic is then defined as

$$T_{GAMuT} = \frac{1}{N} trace(Y_c X_c) \quad (1)$$

Under the null hypothesis that the genotypes and the phenotypes are independent, the test statistic has the same asymptotic distribution as $\frac{1}{N} \sum_{i,j=1}^{N} \lambda_{Xi} \lambda_{Yi} z_{ij}^2$, where λ_{Xi} and λ_{Yi} are the ith ordered nonzero eigenvalue of X_c and Y_c, respectively, and z_{ij}^2 are independent and identically-distributed χ_1^2 random variables. Davies' method can then be used to compute the p-value (Davies, 1980).

When phenotype data are available at multiple time points, the phenotypes for j^{th} subject at time point t can be defined as $P_j^{(t)} = \left(P_{j1}^{(t)}, P_{j2}^{(t)}, \ldots, P_{jL}^{(t)}\right)$. If we combine all samples and time points, it will result in a three-dimensional array, for which it is harder to apply a kernel method. Instead, we concatenate the $P_j^{(t)}$ s for $t = 1,2,\ldots,T$ and define the phenotype vector for jth sample as

$$P_j = \left(P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(T)}\right)$$

which now has dimension $1 \times LT$. Subsequently the $N \times LT$ phenotype matrix P and the genotype matrix G are defined similar to the single time point case. GAMuT (or any other multi-phenotype rare variant model) can be applied using these matrices in the usual manner. For our analysis we applied a weighted linear kernel for modeling the genotypes. A projection matrix or a linear kernel was used for modeling the phenotypes. The weighting scheme for modeling the genotypes, following (M. C. Wu et al., 2011), is based on the minor allele frequency (MAF) of each variant. The observed GAMuT test statistic is then compared with the appropriate mixture of chi-square distribution and p-values are obtained using Davies' method.

We note that the concatenated phenotype vector differs from the normal case since it can impose an unusual correlation structure. The correlation structure among phenotypes at the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

same time point (within time point correlation) may not be similar to the correlation of the same phenotype across time (between time point correlation). Methods that inherently assume exchangeability of the phenotypes in some sense may perform poorly on such concatenated data. However, we have performed extensive simulation studies to verify that in most cases, the performance of the longitudinal version of GAMuT is better than other methods considered here (See 3). We applied GAMuT using both a projection matrix and a linear kernel for measuring phenotypic similarity. These two versions of GAMuT were compared against competing methods.

2.1 Simulations

Various simulations were done to verify that longitudinal GAMuT controls the type-I error at the desired level and to compare its statistical power with that of other methods. The model to generate longitudinal phenotype data is simple yet realistic. For each sample j, first we generated phenotype vectors for each time point t independently. Let us denote the independently generated phenotype vector at time t for subject j as $P_j^{(t)*}$. The model used to generate the $P_j^{(t)*}$ s is the same model used in the original GAMuT paper (Broadaway et al., 2016) which considers different biological factors to simulate realistic datasets. A coalescent model was used to simulate the genotype data, while a multivariate normal distribution was used to simulate the phenotype data. See the supplementary materials for a more detailed description of the simulation procedure.

To obtain the final phenotype vectors $P_j^{(t)*}$ s, a fixed effect of time and an error across time is then added to these values according to the following model:

$$P_j^{(t)} = P_j^{(t)*} + \beta_{time} t + \varepsilon_{tj}, \quad (2)$$

where $\boldsymbol{\varepsilon}_{tj} = (\varepsilon_{tj1}, \varepsilon_{tj2}, \ldots, \varepsilon_{tjL})$. We assume $(\boldsymbol{\varepsilon}_{1ji}, \boldsymbol{\varepsilon}_{2ji}, \ldots, \boldsymbol{\varepsilon}_{Tji})$ are iid from $N(0,\Gamma)$ for all $i = 1,2, \ldots, L$. The variance covariance matrix Γ is assumed to have a first order autoregression (AR(1)) or a compound symmetry (CS) structure. The variances of $P_j^{(t)*}$ and $\boldsymbol{\varepsilon}_{tj}$ are both assumed to be 1. We used several choices of the parameter ρ corresponding to the AR(1) or CS structure. This model ensures that the correlation between $P_{jl1}^{(t)}$ and $P_{jl2}^{(t)}$ is driven by the single time point simulation scheme and the correlation between $P_j^{(t1)}$ and $P_j^{(t2)}$ is driven by Γ. The final correlation between observations across time is half the value of the correlations in the matrix Γ, since the final variance of the observations is 2. We carried out the simulations in such a way (See Supplementary Materials Section 1) that the final correlation between phenotypes within a time point are between 0 and 0.3 (0.3 and 0.4 for high correlation). The coefficients corresponding to effect of time β_{time} were simulated using a $U(0,1)$ distribution.

We also simulated cases where the additive fixed effect of time is not linear. However, at the model fitting stage, the effect of time is regressed away using a linear regression (Hua & Ghosh, 2015). Such simulation frameworks illustrate the performance of our approach under

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

model misspecification. Situations with a quadratic and a sinusoidal time effect are considered. One important advantage of GAMuT is that it is also applicable to categorical data. We simulated a binary phenotype data using a probit model from the already simulated continuous data, we coded the phenotype as 1 when it is greater than some fixed value c and 0 when it is less than c. We chose c to be the arithmetic mean of the phenotype. We used a linear effect of time when simulating binary data. Only GAMuT and univariate SKAT allow for binary data, therefore we compared only these two for the binary data simulations. We used linear regression to adjust for the effect of time even in case of binary phenotypes. Adjustment using a logistic regression was also compared and had very similar results (results not shown).

A feature of longitudinal data is potential dropout over time, resulting in missing data. We applied our approach on the case where 20% observation are missing (completely at random) and the effect of time is linear. kNN imputation (Torgo & Torgo, 2011; Kowarik & Templ, 2016) was used to impute the missing data before applying the association tests.

For each simulated dataset, we used 1000 samples, 10 correlated phenotypes and 3 time points (6 time points in one scenario). The correlation between the phenotypes within a time point was either low (between 0 and 0.3) or high (between 0.3 and 0.4). The number of associated phenotypes varied between 0 (null case), 2, 4, 6, and 8. We used 10^6 for type-I error simulations and 10000 replications for power simulations.

For the simulated datasets, we compared the performance of the longitudinal version of GAMuT (using either projection matrix or linear kernel for phenotypes) with variations of other existing approaches. In particular, we considered competitors MFLM and multivariate SKAT (MSKAT) (B. Wu & Pankow, 2016), which can be applied on the concatenated data in a manner similar to GAMuT. The parameters in the MFLM method were used as suggested by the authors (Wang et al., 2015) and the Q-statistic was used for implementing MSKAT. We also comapared univariate SKAT for which the phenotypes are first collapsed into single time point using arithmetic mean and then tested individually. To adjust for the multiple testing due to testing each phenotype individually, we used a 98% principal component approach similar to the original GAMuT paper (Broadaway et al., 2016). The approach finds the effective number of independent tests by computing the number of principal components needed to explain 98% variability. The effective number of tests is then used to adjust the p-values in a way similar to Bonferroni adjustment. Similarly, an approach combining the single time-point GAMuT p-values is also included in the simulation study. The approach computes the GAMuT p-value (Broadaway et al., 2016) for each time point and combines them using a Bonferroni adjustment. This method is referred to as 'GAMuT (Multi)'. A second method GAMuT (Meta) to combine the individual GAMuT p-values is also used in one simulation scenario (see Supplementary Materials Section 2 for details). GAMuT (Meta) performs meta-analysis using the estimate of the correlation structure across the time points, which was only possible in simulations. Therefore, it is unusable in practice and was included only to demonstrate the power gain by the joint analysis compared to the meta-analysis approach.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2.2 Analysis of GENOA data

High body mass index (BMI), low high-density lipoprotein (HDL), and high blood pressure are related phenotypes that are known to be associated with high risk of cardiovascular diseases, stroke and diabetes. These phenotypes are moderately heritable (Vattikuti, Guo, & Chow, 2012; Zarkesh et al., 2012; Hottenga et al., 2005) and understanding their genetic basis is clinically important. The GENOA study (Daniels et al., 2004) seeks to identify genetic variants that influence risk for hypertension and arteriosclerotic complications of hypertension (Lange et al., 2002). It includes a cohort of African American sibships from Jackson, Mississippi, that were genotyped for a large collection of rare variants using the Illumina Human Exome Beadchip. Over two different time points, data were collected on several phenotypes including BMI, HDL, systolic blood pressure (SBP) and diastolic blood pressure (DBP). We selected these four phenotypes for our analysis.

Following Broadaway et al. (2016), we randomly sampled 1 sibling from each sibling pair and performed standard data cleaning on this resulting dataset of independent subjects. After further data cleaning, our sample consisted of 539 subjects. For each subject, we included data from both time points and the samples having missing data at one time point (116 samples) were not dropped. The missing data were subsequently imputed using k nearest neighbors. The data also included covariates such as gender, age, and smoking status (ever smoked at least 100 cigarettes) and use of anti-hypertension or lipid-lowering medication. Following Broadaway et al. (2016), we also calculated the top ten genetic principal components using ancestry informative markers included on the Illumina array. The final imputed phenotype data along with the covariate data were analyzed using longitudinal GAMuT (both projection matrix and linear kernel), univariate SKAT, multivariate SKAT and MFLM. The use of the methods, including the choice of kernels and tuning parameters, was similar to the application to the simulated datasets.

3 Results

3.1 Type-I error simulations

In the main text, we present simulation results for datasets generated under an AR(1) correlation structure across time. Results for simulated datasets generated assuming a compound symmetry structure across time have very similar results and are provided in the Supplementary Materials. Figures 1 and 2 show the quantile-quantile (QQ) plots for the longitudinal version of GAMuT, using linear kernel and projection matrix, respectively, on simulated null datasets. The plots are provided for both linear and binary cases with no missing data. QQ plots for datsets generated with missing data as well as those for high correlation among phenotypes are shown in the Supplementary Materials and they show similar patterns to those in Figures 1 and 2. Table 1 shows the type-I errors for different methods under various simulation set up. It is evident that GAMuT controls the type-I error for almost all the simulation models considered. For tests with small level of significance (10^{-5}), GAMuT (and SKAT) shows a slight inflation in type-I error in a few cases (Supplementary Table 1). However, the inflation is not significant and can be due to unstable estimation of Type-I error for the very small threshold. The only methods that show significant size inflation in some cases are GAMuT (Multi) and MFLM.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3.2 Power simulations

Next we compared the statistical power of the longitudinal GAMuT against univariate SKAT (which uses the mean of each phenotype across phenotypes), multivariate SKAT, MFLM and GAMuT (Multi). All results reported correspond to an AR(1) correlation structure across time and a genomewide p-value threshold of 5×10^{-6}. However, we also performed simulations for CS type correlation structure and less stringent p-value thresholds. They are reported in the Supplementary Materials. Figure 3 shows the power comparison when a linear effect of time was assumed to simulate the data (See Supplementary Materials for quadratic and sinusoidal time effects). Figure 4 shows the case of binary phenotypes and Figure 5 shows the missing data scenario. Simulations to compare the statistical power of the methods for a larger number of time points ($T = 6$) was also done and the results are shown in Supplementary Figure 4. The performance of the methods are similar across all the simulations with low correlation. GAMuT (Linear) has the highest power in all situations except the case when only two phenotypes have genetic association and the correlation across time is 0, which is unlikely in practice. The power of GAMuT (Linear) increases rapidly compared to other methods with the increase in number of associated phenotypes, illustrating its usefulness under pleiotropy. Also, GAMuT (Linear) has much increased power compared to other methods as the correlation of phenotypes across time increases. The power curves for GAMuT (Projection) are not shown since it had very similar performance as MSKAT. Univariate SKAT is the second most powerful method when the number of associated phenotypes is small, but MSKAT (and GAMuT(Projection)) and GAMuT (Multi) have improved power over SKAT as the number of associated phenotypes continues to increase.

However, when the correlation among the phenotypes are high, MSKAT and GAMuT (Projection) performes better than GAMuT (Linear) for low temporal correlation and/or small number of associated phenotypes (Figure 6). Similar results are observed for other simulation scenarios with high correlation (results not shown). MFLM has inferior power compared to the other fours methods in every simulation set up. As discussed in Broadaway et al. (2016), the performance of MFLM is sensitive to departure from very specific type of data they simulated (Wang et al., 2015). If data are simulated using the approach of (Wang et al., 2015), the performance of MFLM is comparatively better, but GAMuT had very similar performance and still outperforms all other approaches. Therefore we have not reported any result using their simulation approach.

3.3 Analysis of GENOA data

We applied longitudinal GAMuT and the competing methods to the four phenotypes BMI, HDL, SBP and DBP from the GENOA exome-chip study. Prior to the analysis, covariate adjustment were made for gender, age, smoking status, use of anti-hypertension medication, use of lipid lowering medication, and ancestry on the 539 unrelated subjects. We applied longitudinal GAMuT using both projection matrix and linear kernel to the information collected at two time points. We also applied MSKAT and MFLM to this same dataset. We also applied univariate SKAT on the collapsed data using the average value of phenotypes across time. For GAMuT and SKAT, we applied a weighted linear kernel to quantify

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

pairwise genotypic similarity, where weights were a function of minor-allele frequency and was of the same form to that typically utilized in SKAT (M. C. Wu et al., 2011).

We used a study-wise significance threshold of 1.5×10^{-5}, which corresponds to a Bonferroni correction based on the number of genes tested (3278). P-values less than 10^{-3} were considered as suggestive.

Figure 7 shows the QQ-plot for different methods. MFLM shows sizeable inflation of p-values which could not be resolved by transformations. The MFLM method was subsequently dropped from the analysis of GENOA data. The other four approaches did not show such large p-value inflation.

Figure 8 shows the genome-wide results using different methods on the GENOA data. All the genes passing the suggestive or genomewide threshold by at least one method are reported in Table 2. Consistent with our simulation findings, we observed that GAMuT with a projection phenotype matrix yielded quite similar findings to MSKAT in the GENOA analyses. The top two genes identified by these two methods were *EFCAB7* and *COL9A3*, with the results from MSKAT being genomewide significant while the result from GAMuT (projection) being borderline significant. No variants in EFCAB7 or COL9A3 have been reported to be associated with any traits in genome wide studies. EFCAB7 (EF-hand calcium binding domain 7) may play a physiological role in cardiac development based on a study of patients with Ellis-van Creveld syndrome (Nguyen et al., 2016). COL9A3 (collagen type 9 alpha 3) is associated with intervertebral disc disease, a disease frequent among older individuals (Martirosyan et al., 2016).

ZNF667 (zinc finger protein 667) and ENPP3 (ectonucleotide pyrophosphatasep/phosphodiesterase 3) were genomewide suggestive with GAMuT with a projection phenotype matrix. No ZNF667 variants have been reported to be associated with any traits in genome wide studies. The ZNF667 protein, however, may play a role in neuroprotection by acting as a transcriptional repressor (Yuan, Huang, Yuan, Zhao, & Jiang, 2013). A variant in ENPP3 was associated with response of circulating adiponection in response to fenofibrate treatment (Aslibekyan et al., 2013). Adiponectin, a protein secreted by adipose tissue, is associated with improved insulin sensitivity, suppressed development of atherosclerosis, and altered inflammation. Fenofibrate targets circulating adiponectin levels to prevent the onset and progression of cardiovascular disease.

The top gene identified by longitudinal GAMuT with a linear kernel was *CD33*, which also was borderline significant. CD33 (sialic acid-binding immunoglobulin-like lectin 3) is an established gene for Alzheimers disease (Lambert et al., 2013). Recently this gene was associated with cognitive decline (Nettiksimmons, Tranah, Evans, Yokoyama, & Yaffe, 2016). Mid-life hypertension is associated with cognitive decline (Gottesman et al., 2014).

Univariate SKAT analysis identified a different gene, *FNDC1* (fibronectin type III domain containing 1), which was genomewide significant. Interestingly, none of the methods that directly model cross-phenotype effects showed a strong signal for *FNDC1*, suggesting perhaps that this gene is associated only with a single phenotype rather than the multiple phenotypes that were considered. In a case-control study, several variants in FNDC1 showed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

suggestive association with coronary artery disease (Wild et al., 2011). Several suggestive genes were identified by each method. However, only MSKAT found two genes just passing the genomewide significance threshold. All the genes passing the suggestive or genomewide threshold by at least one method are reported in Table 2. We performed a non-parametric bootstrap analysis (with 1000 bootstraps) for each gene to confirm that the results are stable and not extra-sensitive to sampling fluctuations. The bootstrap confidence interval of the p-values are narrow in most cases indicating the stability of the methods. (Supplementary Table 2).

4 Discussion

There is increasing evidence that genetic variants can exhibit cross-phenotypic effects and the statistical power to find such cross-phenotypic effects can be enhanced by utilizing longitudinal phenotype data collected over time. We propose a simple extension of GAMuT for longitudinal data which can be applied on both continuous and binary phenotypes and can adjust for covariates. It is computationally efficient due to the use of Davies' exact method for p-value computation and becomes usable at the genome-wide scale (Software will be made available for public use and will be made available on GitHub pages).

Instead of defining a kernel on the complicated three dimensional data (Phenotypes × Subjects × Time), we considered a concatenated data and have shown in simulation studies that such adjustment results in high statistical power and achieves proper control of the type-I error. The longitudinal extension of GAMuT was also shown to have higher power compared to other competing methods, similarly adjusted for longitudinal data.

The concatenated data may result in a completely different correlation structure among the columns of the phenotype matrix compared to what is usually observed for a single time point phenotype matrix. We performed extensive simulation studies to verify that GAMuT preserves its advantages of being highly powerful and being able to control type-I error at the desired level for such data. The construction of a projection matrix may be harder for longitudinal data and our results indicate that the linear kernel for measuring the phenotypic similarity has favorable statistical power over projection matrix. For low correlation among phenotypes, GAMuT with linear kernel dominated every other competing methods in almost every situation that were considered. However, GAMuT (Projection) had higher power for scenarios with low temporal correlation and/or small number of associated phenotypes.

Additionally, the longitudinal extension of GAMuT was also applied on the GENOA data collected at two time points. Every gene passing the suggestive threshold using any of the competing methods were also detected by at least one of the two GAMuT methods. GAMuT (Projection) was able to identify more genes passing the suggestive threshold. It is not surprising since the phenotypes in this study had high correlations and only 4 phenotypes were used. In general, we suggest to use GAMuT (Projection) for longitudinal data when the correlation among the phenotypes is high and/or the correlation across time is very small. However, GAMuT (linear) may be the better choice if many phenotypes are expected to be associated.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In the age of information explosion, more and more data are collected and often data are available on many phenotypes and over several time points. Our extension of GAMuT for longitudinal data can scale efficiently to handle an arbitrary number of phenotypes and time points, and enables the statistical analysis of cross-phenotypic effect of rare variants by taking advantage of such data to enhance statistical power and helps to develop a better understanding of the genetic background of important complex traits.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants GM117946, HG007508, HL054457, HL086694, HL119443, MH071537, and AR060893.

References

Allison DB, Thiel B, Jean PS, Elston RC, Infante MC, Schork NJ. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. The American Journal of Human Genetics. 1998; 63(4):1190–1201. [PubMed: 9758596]

Aslibekyan S, An P, Frazier-Wood A, Kabagambe E, Irvin M, Straka R, et al. Preliminary evidence of genetic determinants of adiponectin response to fenofibrate in the genetics of lipid lowering drugs and diet network. Nutrition, Metabolism and Cardiovascular Diseases. 2013; 23(10):987–994.

Broadaway KA, Cutler DJ, Duncan R, Moore JL, Ware EB, Jhun MA, et al. A statistical approach for testing cross-phenotype effects of rare variants. The American Journal of Human Genetics. 2016; 98(3):525–540. [PubMed: 26942286]

Daniels PR, Kardia SL, Hanis CL, Brown CA, Hutchinson R, Boerwinkle E, Turner ST. Familial aggregation of hypertension treatment and control in the genetic epidemiology network of arteriopathy (genoa) study. The American journal of medicine. 2004; 116(10):676–681. [PubMed: 15121494]

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.

Fan R, Zhang Y, Albert PS, Liu A, Wang Y, Xiong M. Longitudinal association analysis of quantitative traits. Genetic epidemiology. 2012; 36(8):856–869. [PubMed: 22965819]

Fitzmaurice, GM., Laird, NM., Ware, JH. Applied longitudinal analysis. Vol. 998. John Wiley & Sons; 2012.

Furlotte NA, Eskin E, Eyheramendy S. Genome-wide association mapping with longitudinal data. Genetic epidemiology. 2012; 36(5):463–471. [PubMed: 22581622]

Galesloot TE, Van Steen K, Kiemeney LA, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. PloS one. 2014; 9(4):e95923. [PubMed: 24763738]

Gottesman RF, Schneider AL, Albert M, Alonso A, Bandeen-Roche K, Coker L, et al. Midlife hypertension and 20-year cognitive change: the atherosclerosis risk in communities neurocognitive study. JAMA neurology. 2014; 71(10):1218–1227. [PubMed: 25090106]

Gretton A, Fukumizu K, Teo CH, Song L, Schölkopf B, Smola AJ, et al. A kernel statistical test of independence. Nips. 2007; 20:585–592.

He, Z., Zhang, M., Lee, S., Smith, JA., Kardia, SL., Diez Roux, AV., Mukherjee, B. Journal of the American Statistical Association. 2016. Set-based tests for gene-environment interaction in longitudinal studies. (just-accepted)

Hottenga JJ, Boomsma DI, Kupper N, Posthuma D, Snieder H, Willemsen G, de Geus EJ. Heritability and stability of resting blood pressure. Twin Research and Human Genetics. 2005; 8(05):499–508. [PubMed: 16212839]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Hua WY, Ghosh D. Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. Biometrics. 2015; 71(3):812–820. [PubMed: 25939365]

Kowarik A, Templ M. Imputation with r package vim. Journal of Statistical Software. 2016; 74(7):1–16.

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]

Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. Nature genetics. 2013; 45(12):1452–1458. [PubMed: 24162737]

Lange LA, Lange EM, Bielak LF, Langefeld CD, Kardia SL, Royston P, Peyser PA. Autosomal genome-wide scan for coronary artery calcification loci in sibships at high risk for hypertension. Arteriosclerosis, thrombosis, and vascular biology. 2002; 22(3):418–423.

Lees C, Barrett J, Parkes M, Satsangi J. New ibd genetics: common pathways with other diseases. Gut. 2011; 60(12):1739–1753. [PubMed: 21300624]

Liu F, Van Der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, et al. A genome-wide association study identifies five loci influencing facial morphology in europeans. PLoS Genet. 2012; 8(9):e1002932. [PubMed: 23028347]

Maier R, Moser G, Chen GB, Ripke S, Coryell W, Potash JB, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. The American Journal of Human Genetics. 2015; 96(2):283–294. [PubMed: 25640677]

Martirosyan NL, Patel AA, Carotenuto A, Kalani MYS, Belykh E, Walker CT, Theodore N. Genetic alterations in intervertebral disc disease. Frontiers in Surgery. 2016; 3

Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology. 2010; 34(2):188–193. [PubMed: 19810025]

Nettiksimmons J, Tranah G, Evans DS, Yokoyama JS, Yaffe K. Gene-based aggregate snp associations between candidate ad genes and cognitive decline. Age. 2016; 38(2):1–10. [PubMed: 26695510]

Nguyen TQN, Saitoh M, Trinh HT, Doan NMT, Mizuno Y, Seki M, Mizuguchi M. Truncation and microdeletion of evc/evc2 with missense mutation of efcab7 in ellis-van creveld syndrome. Congenital anomalies. 2016; 56(5):209–216. [PubMed: 26748586]

Schaid DJ. Genomic similarity and kernel methods ii: methods for genomic information. Human heredity. 2010; 70(2):132–140. [PubMed: 20606458]

Schaid DJ, Tong X, Larrabee B, Kennedy RB, Poland GA, Sinnwell JP. Statistical methods for testing genetic pleiotropy. Genetics. 2016; 204(2):483–497. [PubMed: 27527515]

Schork NJ, Zapala MA. Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. Frontiers in genetics. 2012; 3:190. [PubMed: 23060897]

Sun J, Oualkacha K, Forgetta V, Zheng HF, Richards JB, Ciampi A, Greenwood CM. A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. European Journal of Human Genetics. 2016; 24(9):1344–1351. [PubMed: 26860061]

Székely GJ, Rizzo ML, Bakirov NK, et al. Measuring and testing dependence by correlation of distances. The Annals of Statistics. 2007; 35(6):2769–2794.

Torgo, L., Torgo, L. Data mining with r: learning with case studies. Chapman & Hall/CRC; Boca Raton, FL: 2011.

Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009; 65(3):822–832. [PubMed: 19210740]

Vattikuti S, Guo J, Chow CC. Heritability and genetic correlations explained by common snps for metabolic syndrome traits. PLoS Genet. 2012; 8(3):e1002637. [PubMed: 22479213]

Wang Y, Liu A, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, Fan R. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genetic epidemiology. 2015; 39(4):259–275. [PubMed: 25809955]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Wild PS, Zeller T, Schillert A, Szymczak S, Sinning CR, Deiseroth A, et al. A genome-wide association study identifies lipa as a susceptibility gene for coronary artery diseaseclinical perspective. Circulation: Cardiovascular Genetics. 2011; 4(4):403–412. [PubMed: 21606135]

Wu B, Pankow JS. Sequence kernel association test of multiple continuous phenotypes. Genetic epidemiology. 2016; 40(2):91–100. [PubMed: 26782911]

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]

Yuan D, Huang J, Yuan X, Zhao J, Jiang W. Zinc finger protein 667 expression is upregulated by cerebral ischemic preconditioning and protects cells from oxidative stress. Biomed Rep. 2013; 1:534–538. [PubMed: 24648981]

Zarkesh M, Daneshpour MS, Faam B, Fallah MS, Hosseinzadeh N, Guity K, Azizi F. Heritability of the metabolic syndrome and its components in the tehran lipid and glucose study (tlgs). Genetics research. 2012; 94(06):331–337. [PubMed: 23374242]

Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986:121–130. [PubMed: 3719049]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

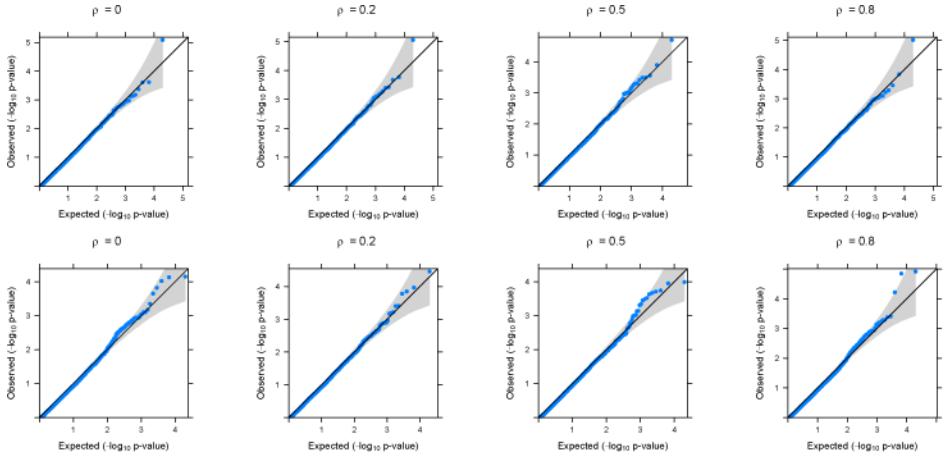

Figure 1.

QQ-plots based on application of GAMuT (Linear) on simulated null datasets. Top row shows the plots for continuous phenotypes and the bottom row shows the plots for binary phenotypes (linear time effect). The correlation structure across time is considered to be AR(1) with parameter ρ.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

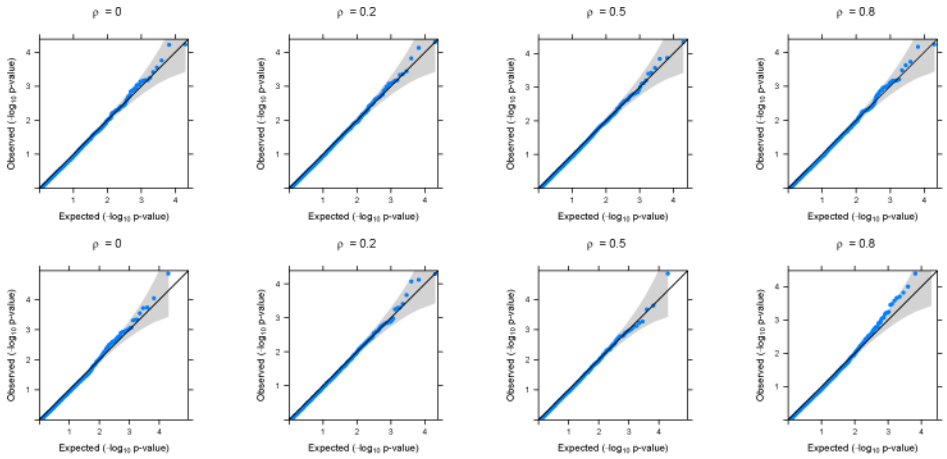

Figure 2.

QQ-plots based on application of GAMuT (Projection) on simulated null datasets. Top row shows the plots for continuous phenotypes and the bottom row shows the plots for binary phenotypes (linear time effect). The correlation structure across time is considered to be AR(1) with parameter ρ.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

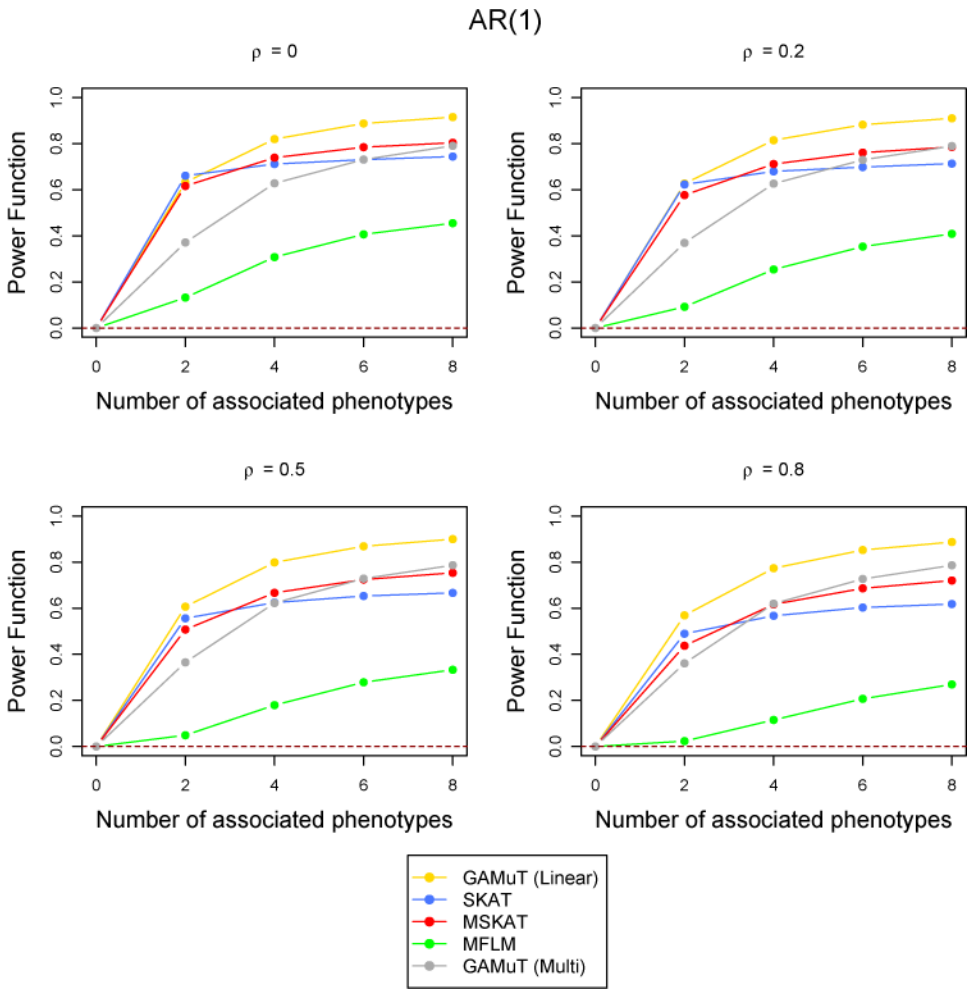

Figure 3.

Comparison of power curves for different methods in the linear time effect set up. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

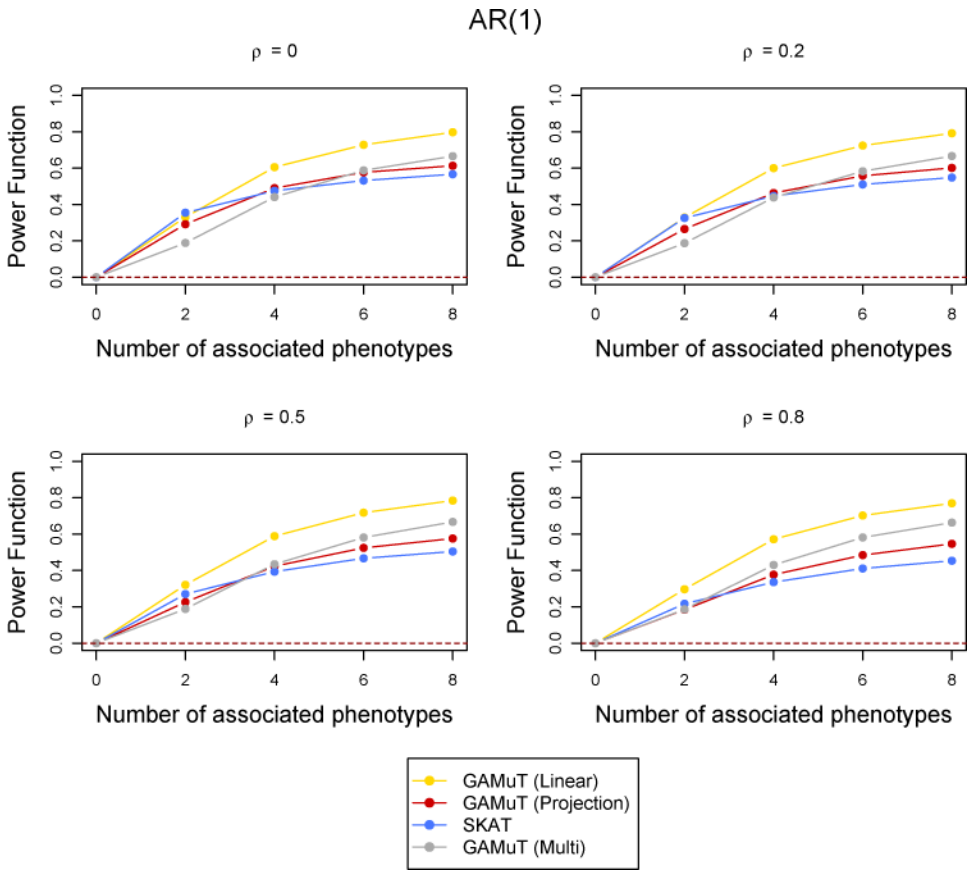

Figure 4.

Comparison of power curves for different methods in the binary phenotypes case. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

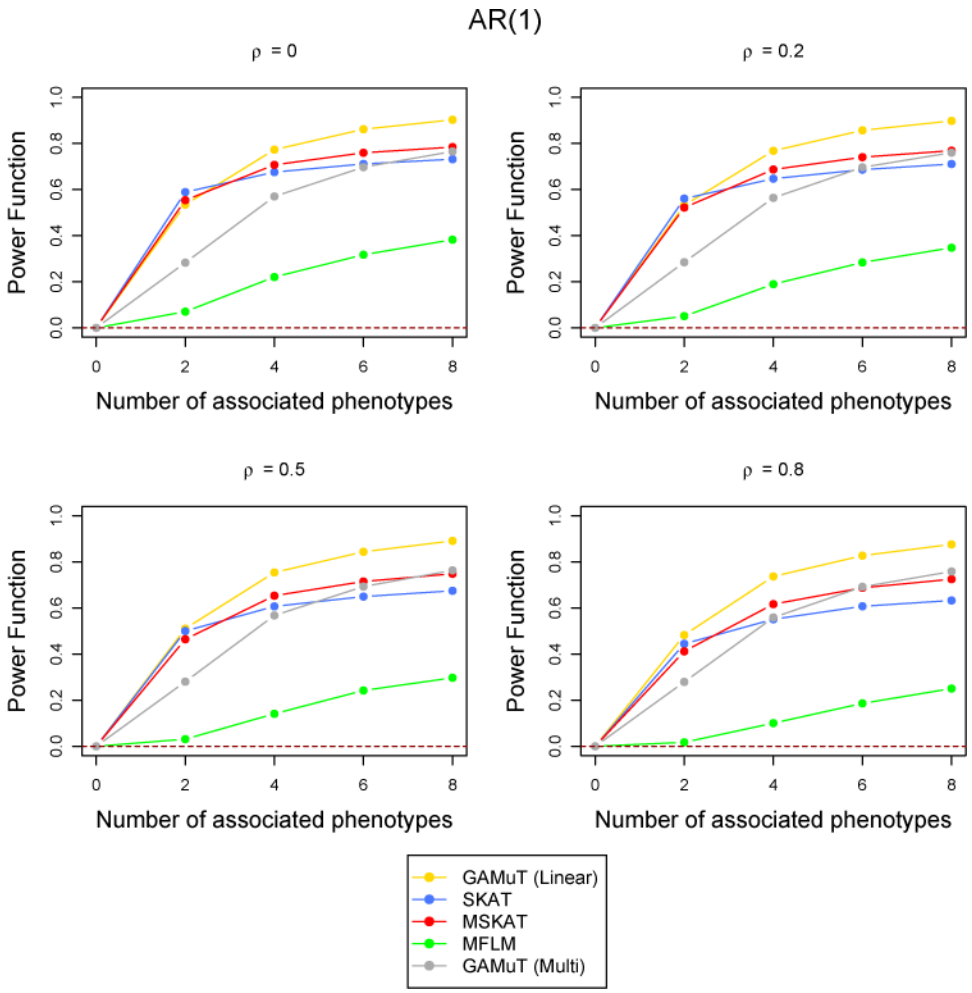

Figure 5.

Comparison of power curves for different methods in the missing data scenario. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

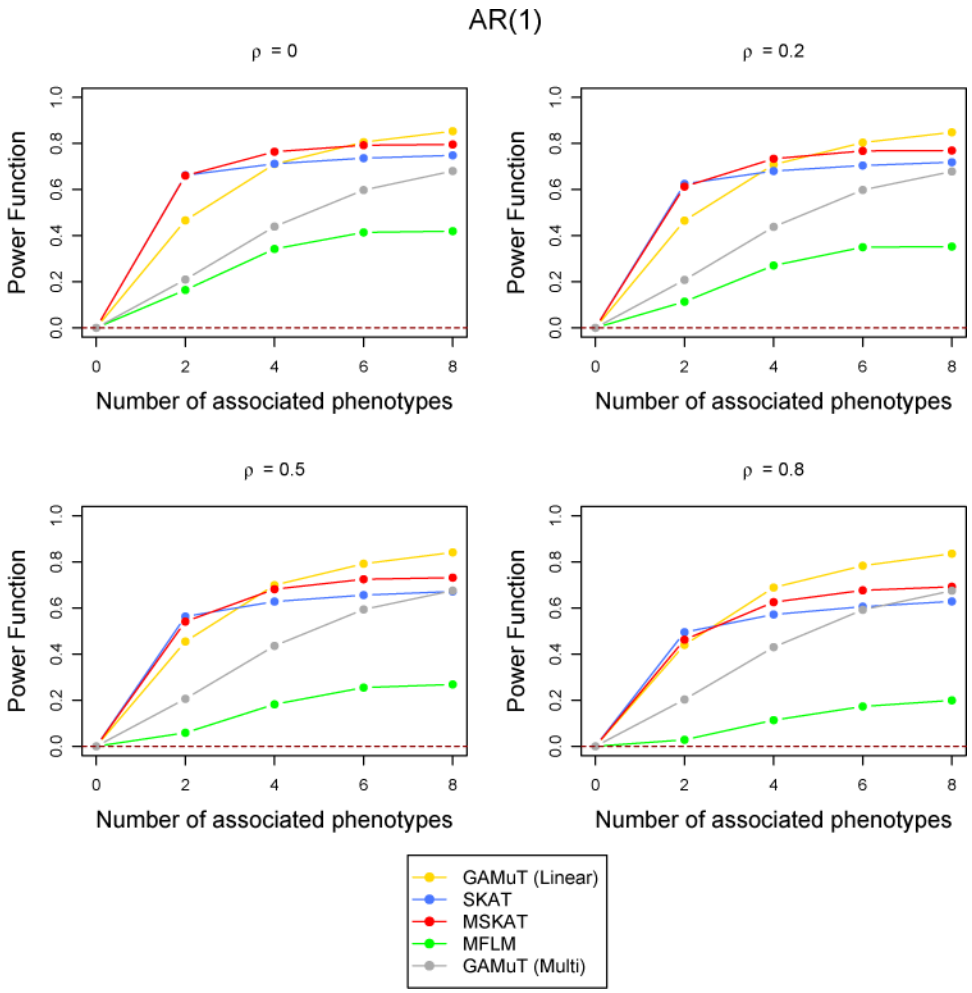

Figure 6.

Comparison of power curves for different methods in the linear time effect set up when the correlations between phenotypes were higher. The correlation structure across time is considered to be AR(1) with parameter ρ and the tests are done at the p-value threshold 5×10^{-6}. The value of the power function corresponding to 0 associated phenotypes shows the type-I error and the horizontal dotted line indicates the level of the test.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

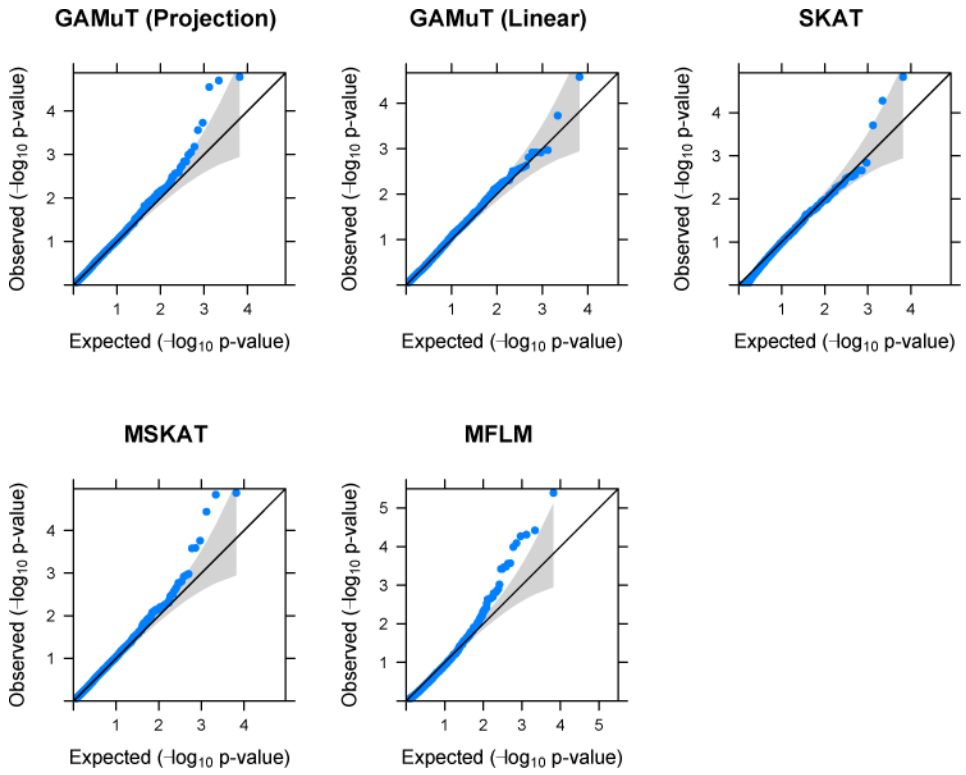

Figure 7.
QQ-plots based on application of different methods on GENOA data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

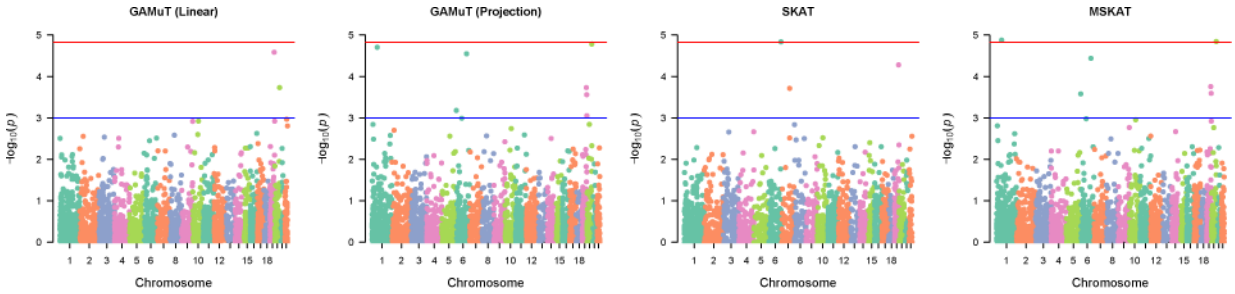

Figure 8.
Manhattan plots based on application of different methods on GENOA data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Type-I error of different methods under various simulations, target $\alpha = 0.01$. The correlation structure across time is considered to he AR(1) with parameter ρ.

	GAMuT (Projection)				**GAMuT (Linear)**				**GAMuT (Multi)**			
	ρ = 0	**ρ = 0.2**	**ρ =0.5**	**ρ = 0.8**	**ρ = 0**	**ρ = 0.2**	**ρ =0.5**	**ρ = 0.8**	**ρ = 0**	**ρ = 0.2**	**ρ =0.5**	**ρ = 0.8**
Linear	0.010	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.008	0.008
Linear(high correlation)	0.009	0.009	0.008	0.009	0.010	0.010	0.010	0.011	0.009	0.010	0.009	0.009
Linear (6 time points)	0.006	0.006	0.006	0.007	0.007	0.007	0.006	0.008	0.008	0.008	0.007	0.007
Quadratic	0.010	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.008	0.009
Sine	0.010	0.008	0.010	0.009	0.010	0.009	0.009	0.009	0.009	0.008	0.008	0.009
Binary	0.010	0.010	0.009	0.010	0.010	0.009	0.009	0.010	0.012	0.012	0.011	0.013
Missing	0.008	0.008	0.007	0.007	0.008	0.008	0.007	0.008	0.008	0.008	0.007	0.009
	SKAT				MSKAT				MFLM			
	ρ = 0	ρ = 0.2	ρ =0.5	ρ = 0.8	ρ = 0	ρ = 0.2	ρ =0.5	ρ = 0.8	ρ = 0	ρ = 0.2	ρ = 0.5	ρ = 0.8
Linear	0.008	0.008	0.008	0.008	0.010	0.009	0.009	0.009	0.007	0.008	0.009	0.007
Linear (high correlation)	0.007	0.007	0.007	0.007	0.009	0.009	0.008	0.009	0.007	0.007	0.008	0.009
Linear (6 time points)	0.007	0.008	0.007	0.007	0.006	0.006	0.006	0.007	0.007	0.008	0.007	0.007
Quadratic	0.008	0.008	0.008	0.007	0.010	0.009	0.009	0.009	0.007	0.008	0.009	0.007
Sine	0.008	0.008	0.008	0.007	0.010	0.008	0.010	0.009	0.008	0.008	0.009	0.007
Binary	0.007	0.007	0.008	0.006	0.010	0.010	0.009	0.010	0.010	0.009	0.008	0.009
Missing	0.008	0.009	0.008	0.009	0.008	0.008	0.007	0.007	0.008	0.009	0.008	0.009

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Genomewide results based on the application of different methods on GENOA data.

Chromosome	Gene Name	OMIM Number	Number of Variants	GAMuT (Projection)	GAMuT (Linear)	SKAT	MSKAT
1	EFCAB7	–	6	1.98E-05	0.010	0.325	1.33E-05
6	ENPP3	602182	6	2.84E-05	0.009	0.431	3.66E-05
6	NQO2	160998	5	0.001	0.007	0.061	2.63E-04
6	FNDC1	609991	10	0.006	0.003	1.46E-05	0.005
7	ZNF655	–	5	0.009	0.006	1.94E-04	0.009
19	CD33	159590	5	1.85E-04	2.61E-05	5.25E-05	1.75E-04
19	ZNF551	–	5	0.001	0.001	0.033	0.001
19	ZNF667	611024	5	2.75E-04	0.008	0.005	2.55E-04
20	COL9A3	120270	5	1.66E-05	1.85E-04	0.031	1.43E-05