Hu et al. BMC Proceedings 2014, 8(Suppl 1):S89
http://www.biomedcentral.com/1753-6561/8/S1/S89

BMC Proceedings

PROCEEDINGS Open Access

Association analysis of whole genome sequencing data accounting for longitudinal and family designs

Yijuan Hu[1], Qin Hui[2], Yan V Sun[2,3,4*]

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

Using the whole genome sequencing data and the simulated longitudinal phenotypes for 849 pedigree-based individuals from Genetic Analysis Workshop 18, we investigated various approaches to detecting the association of rare and common variants with blood pressure traits. We compared three strategies for longitudinal data: (a) using the baseline measurement only, (b) using the average from multiple visits, and (c) using all individual measurements. We also compared the power of using all of the pedigree-based data and the unrelated subset. The analyses were performed without knowledge of the underlying simulating model.

Background

Whole genome sequencing (WGS) makes it possible for investigators to extend association studies to rare variants. Rare variants, which have minor allele frequencies (MAFs) of less than 1% to 5%, might play an important role in the etiology of complex traits and account for missing heritability unexplained by common variants [1-3]. However, traditional single-variant tests for common variants have limited power for testing rare variants because of their low frequencies and large numbers. A number of methods [4-7] have been developed to address this challenge by jointly analyzing rare variants within a region. Among these methods, the burden test of Lin and Tang [7] easily fits into the regression framework that can accommodate complex study designs and phenotypes, thus remaining a competitive option.

The longitudinal study design, which collects repeated measurements on the same subject over time, has been routinely used in epidemiologic and clinical research. The repeated measurements can reduce error and thus increase statistical power compared with the single measurement. There have been increasing applications of such a design in genome-wide association studies (GWAS) with a focus on common variants. The analytical strategies include using the measurement at a single time point [8], using the summarized univariate measurement [9] and adopting the linear mixed-effects model to fully exploit information in the repeated measurements [10,11]. However, the implementation of such designs is limited in the context of WGS studies with a focus on rare-variant associations.

In the era of next-generation sequencing studies, the family-based design has the unique advantages of protecting against population stratification, detecting genotyping errors, and facilitating accurate imputation, all of which are challenging issues to cope with in the studies of rare variants using unrelated subjects. However, it is well known that enrolling and sequencing additional family members will not increase statistical power as much as that can be achieved by the same number of unrelated subjects. The degree of power gain from the added family members depends on the extent of the within-family correlation and the size of each family. Thus, it is of interest to assess this power gain in each specific study.

Genetic Analysis Workshop 18 (GAW18) provided WGS data (sequencing plus imputation) in a pedigree-based sample with longitudinal measurements for systolic blood pressure (SBP) and diastolic blood pressure (DBP). In this study, we implement methods to exploit longitudinal and family structures and apply these methods to examine the associations of aggregated rare variants, as well as common single-nucleotide polymorphisms

* Correspondence: yan.v.sun@emory.edu
[2]Department of Epidemiology, Emory University, Atlanta, GA, USA
Full list of author information is available at the end of the article

(SNPs), with SBP and DBP. We compare the power of the methods using the baseline measurement only, using the averaged value over repeated measurements, and using the full information of longitudinal data. We also contrast the power of the methods using all of the pedigree-based data with that using the unrelated subset.

Methods

In this study, we focus on the first replicate of simulated SBP and DBP and the genetic data from sequencing and imputation only on chromosome 3 because of limited computation resources. The analyses were performed without knowledge of the underlying simulating model. We obtained 849 individuals from 20 pedigrees, among which 142 are unrelated. All individuals have SBP and DBP measurements at three time points with no missing data, as well as age, gender, smoking status, and antihypertensive medication status. Note that we preadjust the SBP and DBP measurements by the antihypertensive medication status (i.e., increasing SBP by 10 mm Hg and DBP by 5 mm Hg if the subject is taking medication). We define common variants as those with MAFs 5% or greater and obtain 403,098 SNPs on chromosome 3. We jointly analyze rare variants by mRNA transcripts, which are the functional products of genes. We exclude transcripts whose total rare allele frequency (i.e., sum of MAFs over all inclusive variants) is less than 0.01 and end up with a total of 813 transcripts represented by accession numbers. Given a common single-nucleotide polymorphism (SNP) or a transcript for the phenotype, we consider using (a) the baseline, (b) the time-averaged, and (c) the repeated measurements; for study subjects, we consider using (a) the entire pedigree-based sample and (b) the unrelated subjects only. All statistical analyses were carried out in R (http://www.r-project.org) version 2.15.1.

Notation

Assume there are m rare variants in a transcript. Under the population-based design, let Y_{it} denote the phenotype measured for subject i at time t, $\boldsymbol{G}_i = (G_{i1}, G_{i2}, ..., G_{im})^{\mathrm{T}}$ the genotypes of the m variants, and $\boldsymbol{X}_{it} = (X_{it1}, X_{it2}, ..., X_{itq})^{\mathrm{T}}$ the q covariates including the intercept and possibly time-varying ones. Under the family-based design and when subject i belongs to family p, we modify the aforementioned notation to be Y_{pit}, $\boldsymbol{G}_{\mathrm{pi}}$ and $\boldsymbol{X}_{pit}$.

Burden score of rare variants

We focus on variants with MAF less than 5% and that are putatively functional (i.e., nonsense, missense, or splice site mutations). Specifically, using the chromosomal location (NCBI built 37.1) provided by the GAW18 data, we search for functional annotation of 812,234 SNPs with MAFs 5% or less on chromosome 3 using the GVS server (http://snp.gs.washington.edu). We consider the following functional categories: missense, missense-near-splice, splice-3, splice-5, stop-gained, stop-lost, and stop-lost-near-splice. Note that the functional annotation is specific to each transcript. The burden score of the ith subject at a given transcript is defined as the sum of genotypes of the selected variants: $S_i = \sum_{j=1,...,m} G_{ij}$.

In the framework of Lin and Tang [7], the burden score is used in the following regression models as a regular covariate.

Models and assumptions

Unrelated subjects with single measurements of blood pressure

The single measurements of blood pressure can be the baseline or time-averaged SBP or DBP. We denote the baseline and averaged measurements by Y_{i1} and $Y_{i+} = \sum_t Y_{it}$, respectively. Because they are all quantitative, it is natural to relate each of them to S_i and $\boldsymbol{X}_i$ through the linear regression model:

$$Y_{i1} \text{ or } Y_{i+} = bS_i + g^{\mathrm{T}}X_i + e_i, \tag{1}$$

where ϵ_{i} is an error term with mean zero and variance σ^2 and $\boldsymbol{X}_i$ includes the intercept, gender, smoking status, and baseline (averaged) age in the baseline (averaged) phenotype model.

Unrelated subjects with repeated measurements of blood pressure

To account for the correlation of measurements from the same individual, we use the linear mixed-effects model. For subject i at time t, it is written as

$$Y_{it} = bS_i + g^{\mathrm{T}}X_{it} + b_i + e_{it}, \tag{2}$$

where b_i is a random effect that follows $\mathrm{N}(0, \sigma_b^2)$; ϵ_{it} is an error term that follows $\mathrm{N}(0, \sigma^2)$; ϵ_{it} and b_i are mutually independent; and $\boldsymbol{X}_{it}$ consists of the intercept, gender, smoking status, and the time-varying age. By including age into $\boldsymbol{X}_{it}$, we assume that the traits change linearly with time. Both visual inspection of the individual-level trait trajectories and statistical testing of the age coefficient support the linear modeling (results not shown). Because $\mathrm{Cov}(Y_{it}, Y_{it'}) = \sigma_b^2 + \sigma^2$ for $t = t'$ and $\mathrm{Cov}(Y_{it}, Y_{it'}) = \sigma_b^2$ for $t \neq t'$, the random effect b_i induces a squared correlation of $\sigma_b^2/(\sigma_b^2 + \sigma^2)$ between any pair of measurements from subject i. Note that b_i is shared by different measurements of subject i so that the induced correlations are the same.

Families with single measurements of blood pressure

To account for the phenotype correlation among subjects from the same family, we also adopt a linear mixed-effects model. For the ith subject in the pth pedigree, we formulate that

$$Y_{pi1} \text{ or } Y_{pi+} = \beta S_{pi} + \gamma^{\mathrm{T}} X_{pi} + g_{pi} + \epsilon_{pi}, \quad (3)$$

where g_{pi} is a random effect representing genetic similarity among family members, ϵ_{pi} is an error term that follows $\mathrm{N}(0, \sigma^2)$, g_{pi} is independent of ϵ_{pi} ϵ_{pi}, and $\boldsymbol{X}_{pi}$ is the same as $\boldsymbol{X}_i$ in (1). In addition, we assume that $\mathrm{Corr}(g_{pi}, g_{pi'}) = 2\psi_{ii'}$, $\mathrm{Corr}(g_{pi}, g_{pi'}) = 2\psi_{ii'}$, where $\psi_{ii'}$ is the kinship coefficient between family member i and i'. Unlike b_i in (2), which is shared among correlated units (time), there is a random effect g_{pi} for each unit (subject) here, and their covariance matrix is specified so that correlations among different pairs of family members are different.

Families with repeated measurements of blood pressure

Model (3) can be readily extended to accommodate repeated measurements. We include the repeated measurements in (3) as follows:

$$Y_{pit} = \beta S_{pi} + \gamma^{\mathrm{T}} X_{pit} + g_{pi} + b_{pi} + \epsilon_{pit}, \quad (4)$$

where $\boldsymbol{X}_{pit}$ contains the time-varying age, g_{pi} is introduced previously, b_{pi} follows the same distribution as b_i in model (2), and g_{pi} and b_{pi} are independent of each other. Unlike in (2), b_{pi} here characterizes the additional correlation between repeated measurements after adjusting for the genetically induced portion. To see this, we consider the reduced model without b_i:

$$Y_{pit} = \beta S_{pi} + \gamma^{\mathrm{T}} X_{pit} + g_{pi} + \epsilon_{pit} \quad (5)$$

For both (4) and (5), the covariance between different family members is $\mathrm{Cov}\left(Y_{pit}, Y_{pi't'}\right) = 2\psi_{ii'}\sigma_g^2$ for any t and t'. However, the covariance between measurements $\left(t \neq t'\right)$ from the same subject is $\mathrm{Cov}\left(Y_{pit}, Y_{pit'}\right) = \sigma_g^2 + \sigma_b^2$ based on (4) and $\mathrm{Cov}\left(Y_{pit}, Y_{pit'}\right) = \sigma_g^2$ based on (5).

Although model (4) is more flexible than (5), the chromosome-wide scan based on (4) is not feasible within the given timeframe and available computational resources. We thus adopt a two-stage strategy that first scans chromosome 3 using (5) and then refines the p-values of top SNPs using (4).

Population stratification

GAW18 data consist of Mexican Americans from San Antonio, a population that may have an admixed ancestry of whites and Native Americans. To account for possible population stratification, we include top principal components (PCs) of SNP genotypes as covariates in the above regression models. We first obtain independent SNPs (linkage disequilibrium R^2 <0.2) restricted to those with MAFs 5% or greater using the unrelated subjects. Then we project the SNP loadings of unrelated subjects to their relatives to calculate the eigenvectors of the entire sample of families.

Results

Figure 1 displays the quantile-quantile (QQ) plots of p-values for testing the association between common SNPs and SBP. All 6 tests produced proper type I error because their genomic control parameter λ's are close to 1. This suggests that the data are well described by our models and population stratification is appropriately adjusted by the PCs. Clearly, using all pedigree-based samples is substantially more powerful than using the unrelated subjects only. In addition, using the averaged SBP yielded smaller p-values for top SNPs than using the baseline or repeated measurements, and using the repeated measurements is slightly more powerful than using the baseline. This pattern can also be seen in Figure 2. The top five SNPs based on the method using the averaged SBP and all subjects are listed in Table 1, whose last column provides the refined p-values from model (4). Note that the use of model (4) does not alter the aforementioned order based on power, although it tends to slightly improve on the use of model (5). Using the Bonferroni correction, the genome-wide significance threshold is 1.3×10^{-7}, at which the top five SNPs can be declared as genome-wide significant by any method. Note that we only focused on chromosome 3, so what we are assessing is in fact chromosome-wide significance. For testing the association between rare variants and SBP, the 6 tests also have controlled type I error (see Figure 3 for QQ plots of p-values). Again, compared with the unrelated subset, the relatives added considerable information on the associations of the top three transcripts. All three types of SBP generated comparable power with all individuals, and the three consensus top transcripts are described in Table 2. Using the Bonferroni correction, the genome-wide significance threshold is 6.2×10^{-5}, at which the three top transcripts can be declared as genome-wide significant by any method. All of the identified common and rare variants map to the gene *MAP4*, which spans from 47,892,180 to 48,130,769 on chromosome 3.

The results of testing the genetic association with DBP show similar patterns as with SBP (data not shown). In particular, using the averaged DBP yielded better power than using the repeated measurements. Tables 1 and 2 provide the top common SNPs and transcripts, respectively.

Discussion and conclusions

We investigated three approaches to exploiting longitudinal phenotype data and assessed the power gain of adding family members in the context of WGS studies. Most GWAS have focused on the population-based design, which maximizes the power per genotyped subject. Our results demonstrated that including family members can also significantly boost the power. Most

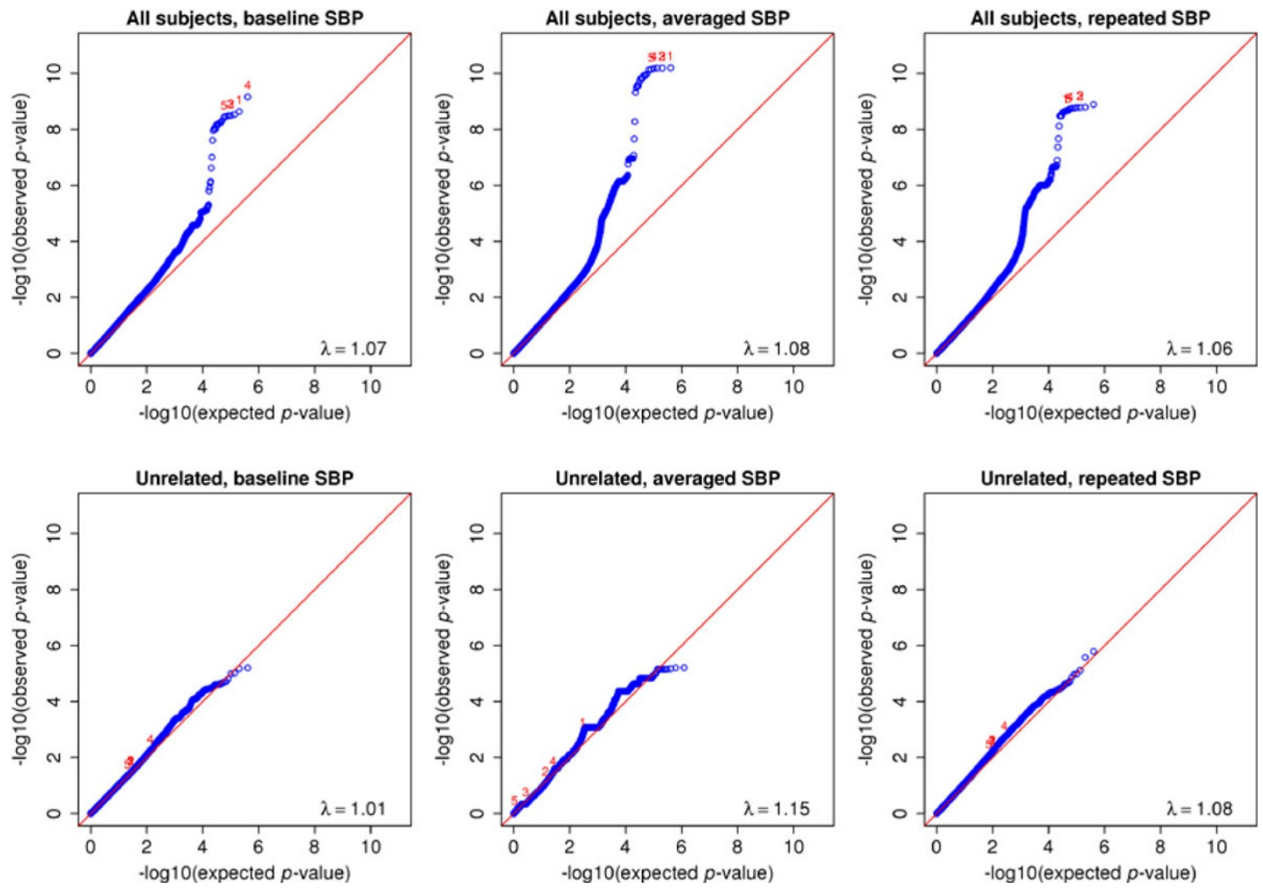

Figure 1 Quantile-quantile plots of *p*-values for tests between common single-nucleotide polymorphisms and systolic blood pressure (SBP)

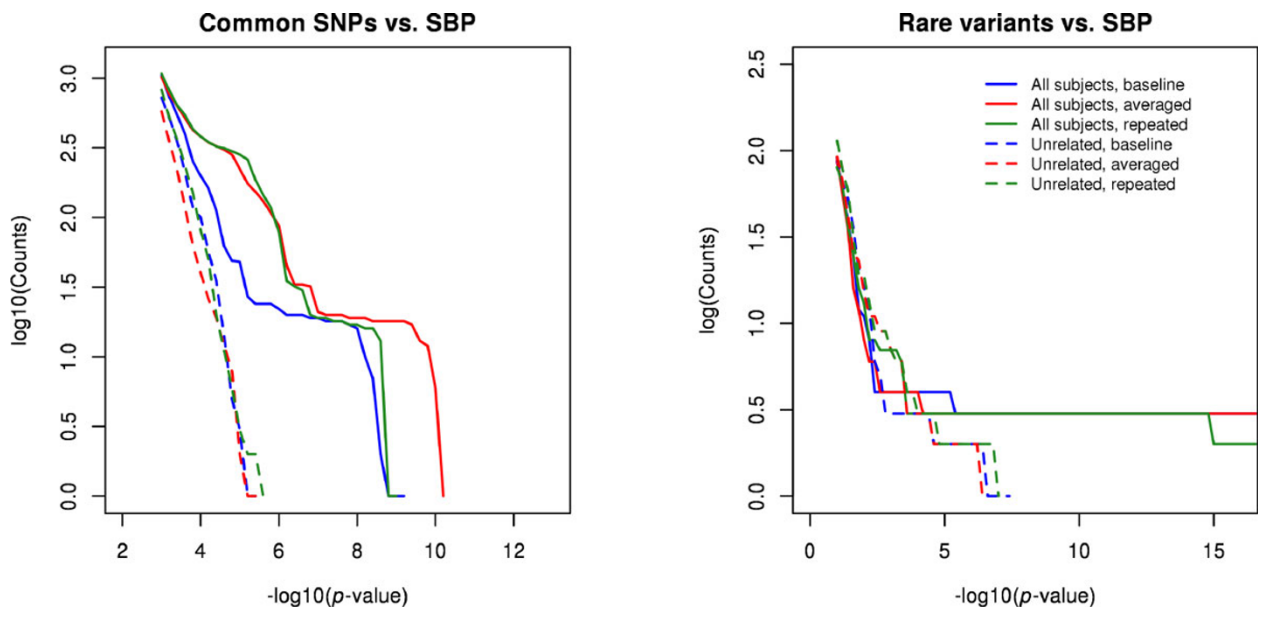

Figure 2 Count of tests with *p*-values lower than the *p*-value threshold across a range of *p*-value cutoffs

Table 1 *p*-Values for the top five common single-nucleotide polymorphisms based on the analysis of averaged systolic blood pressure

					SBP *p*-values			
SNP ID	Gene	Chr	Position	MAF	Baseline	Averaged	Repeated-1	Repeated-2
3_47903424	*MAP4*	3	47903424	0.124	2.3×10^{-9}	6.3×10^{-11}	2.2×10^{-9}	1.1×10^{-10}
3_47903305	*MAP4*	3	47903305	0.123	3.2×10^{-9}	6.5×10^{-11}	1.7×10^{-9}	1.0×10^{-10}
3_47905079	*MAP4*	3	47905079	0.124	3.2×10^{-9}	6.5×10^{-11}	1.7×10^{-9}	1.0×10^{-10}
3_47588649	*MAP4*	3	47588649	0.122	3.5×10^{-9}	6.7×10^{-11}	1.8×10^{-9}	1.1×10^{-10}
3_47990500	*MAP4*	3	47990500	0.123	3.6×10^{-9}	7.2×10^{-11}	2.1×10^{-9}	1.1×10^{-10}
					DBP *p*-values			
SNP ID	Gene	Chr	Position	MAF	Baseline	Averaged	Repeated-1	Repeated-2
3_48064367	*MAP4*	3	48064367	0.128	1.4×10^{-11}	3.6×10^{-13}	3.5×10^{-13}	3.6×10^{-13}
3_47711490	*MAP4*	3	47711490	0.120	1.8×10^{-11}	3.9×10^{-13}	5.4×10^{-12}	3.9×10^{-13}
3_48092335	*MAP4*	3	48092335	0.127	2.8×10^{-11}	4.8×10^{-13}	2.2×10^{-12}	5.2×10^{-13}
3_48105528	*MAP4*	3	48105528	0.127	2.8×10^{-11}	4.8×10^{-13}	2.2×10^{-12}	5.2×10^{-13}
3_47990500	*MAP4*	3	47990500	0.123	4.9×10^{-11}	5.0×10^{-13}	5.8×10^{-12}	5.4×10^{-13}

Note: Repeated-1 and repeated-2 correspond to model (5) and model (4), respectively.
DBP, diastolic blood pressure; *MAF*, minor allele frequency; *SBP*, systolic blood pressure

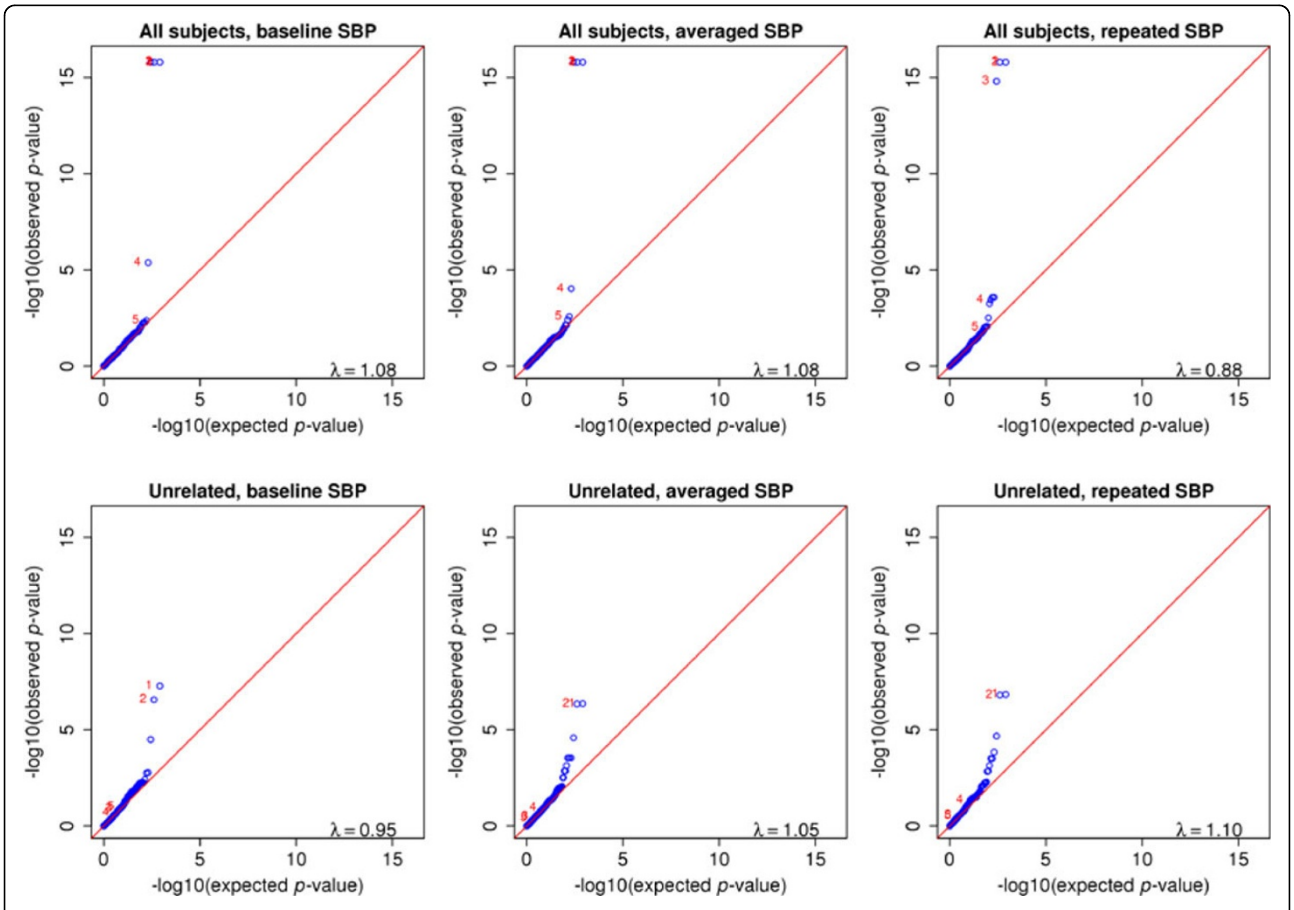

Figure 3 Quantile-quantile plots of *p*-values for tests between common single-nucleotide polymorphisms and systolic blood pressure (SBP)

Table 2 *p*-Values for the top three transcripts based on the analysis of averaged systolic blood pressure (diastolic blood pressure

				SBP *p*-values			
Accession ID	**Gene**	**Chr**	**SMAF**	**Baseline**	**Averaged**	**Repeated-1**	**Repeated-2**
NM_001134364.1	*MAP4*	3	0.080	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
NM_002375.4	*MAP4*	3	0.074	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
NM_030885.3	*MAP4*	3	0.036	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	1.6×10^{-15}	$<2.2 \times 10^{-16}$
				DBP *p*-values			
Accession ID	**Gene**	**Chr**	**SMAF**	**Baseline**	**Averaged**	**Repeated-1**	**Repeated-2**
NM_001134364.1	*MAP4*	3	0.080	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
NM_002375.4	*MAP4*	3	0.074	4.4×10^{-16}	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$
NM_030885.3	*MAP4*	3	0.036	6.0×10^{-15}	2.2×10^{-16}	5.3×10^{-15}	$<2.2 \times 10^{-16}$

Note: SMAF is the sum of minor allele frequencies (MAFs) of single-nucleotide polymorphisms included in the burden score of the gene.
DBP, diastolic blood pressure; SBP, systolic blood pressure.

GWAS have ignored the longitudinal nature of the phenotype data, which are available from many prospective cohorts. The use of the longitudinal data can provide a more accurate measurement of the phenotype and thus serves as a powerful tool in genetic association studies. With more clinic data being available through the electronic medical record (EMR) system and more clinic populations with genotypic data, the search for disease-associated common and rare variants can be more fruitful by improving the phenotyping via longitudinal information.

It appears somewhat counterintuitive that using the time-averaged measurement is more powerful than using the repeated measurement in the analysis of the GAW18 data. This is possible because, in the presence of linear time effect, the averaged measurement does not lose any information compared with the repeated measurements but simply reduces error. When more complex longitudinal structures exist, the repeated measurements retain full information and are expected to outperform the averaged measurement.

A family-based design can allow us to test association and linkage simultaneously. In this paper, we focused on the association analysis only. We modeled the association in the fixed-effect parameters and accounted for family relatedness using the random-effect parameters, whose covariances among family members are formulated through the kinship coefficient. Our models can be readily extended to linkage analysis by including another set of random-effect parameters whose covariances depend on the proportion of alleles shared identical by descent at the marker locus between a relative pair [12].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
YJH and YVS designed the overall study and drafted the manuscript. QH conducted statistical analyses. All authors read and approved the final manuscript.

Acknowledgements
YVS and QH were supported in part by National Institutes of Health (NIH) grant RC1 HL100245 from the National Heart, Lung, and Blood Institute. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575. This article has been published as part of *BMC Proceedings* Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

Authors' details
[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA. [2]Department of Epidemiology, Emory University, Atlanta, GA, USA. [3]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA. [4]Center for Health Research, Kaiser Permanente Georgia, Atlanta, GA, USA.

Published: 17 June 2014

References
1. Schork NJ, Murray SS, Frazer KA, Topol EJ: **Common vs. rare allele hypotheses for complex diseases.** *Curr Opin Genet Dev* 2009, **19**:212-219.
2. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305**:869-872.
3. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: **Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.** *Science* 2009, **324**:387-389.
4. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
5. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR: **Pooled association tests for rare variants in exon-resequencing studies.** *Am J Hum Genet* 2010, **86**:832-838.
6. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare variant association testing for sequencing data using the sequence kernel association test (SKAT).** *Am J Hum Genet* 2011, **89**:82-93.
7. Lin DY, Tang ZZ: **A general framework for detecting disease associations with rare variants in sequencing studies.** *Am J Hum Genet* 2011, **89**:354-367.
8. Sabatti C, Service SK, Hartikainen A, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, *et al*: **Genome-wide association analysis of metabolic traits in a birth cohort from a founder population.** *Nat Genet* 2008, **41**:35-46.
9. Ionita-Laza I, McQueen M, Laird N, Lange C: **Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan.** *Am J Hum Genet* 2007, **81**:607-614.

10. Smith EN, Chen W, Kähönen M, Kettunen J, Lehtimäki T, *et al*: **Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa Heart Study.** *PLoS Genet* 2010, **6**:e1001094.
11. Luan J, Kerner B, Zhao JH, Loos RJF, Sharp SJ, Muthén BO, Wareham NJ: **A multilevel linear mixed model of the association between candidate genes and weight and body mass index using the Framingham longitudinal family data.** *BMC Proc* 2009, **3(suppl 7)**:S115.
12. Abecasis GR, Cookson WOC, and Cardon LR: **Pedigree tests of transmission disequilibrium.** *Eur J Hum Genet* 2000, , **8:** 545-551.

doi:10.1186/1753-6561-8-S1-S89
Cite this article as: Hu *et al.*: Association analysis of whole genome sequencing data accounting for longitudinal and family designs. *BMC Proceedings* 2014 **8**(Suppl 1):S89.