Liu et al. BMC Proceedings 2014, 8(Suppl 1):S40
http://www.biomedcentral.com/1753-6561/8/S1/S40

BMC Proceedings

PROCEEDINGS Open Access

Extended T^2 tests for longitudinal family data in whole genome sequencing studies

Yiwei Liu, Jing Xuan, Zheyang Wu*

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

Abstract

Family data and rare variants are two key features of whole genome sequencing analysis for hunting the missing heritability of common human diseases. Recently, Zhu and Xiong proposed the generalized T^2 tests that combine rare variant analysis and family data analysis. In similar fashion, we developed the extended T^2 tests for longitudinal whole genome sequencing data for family-based association studies. The new methods simultaneously incorporate three correlation sources: from linkage disequilibrium, from pedigree structure, and from the repeated measures of covariates. We assess and compare these methods using the simulated data from Genetic Analysis Workshop 18. We show that, in general, the extended T^2 tests incorporating longitudinal repeated measures have higher power than the single-time-point T^2 tests in detecting hypertension-associated genome segments.

Background

Compared with traditional genome-wide association studies (GWAS), whole genome sequencing (WGS) is a more efficient way of finding the missing heritability of diseases [1]. While GWAS are mostly based on microarray genotyping, which can discover only common genetic variants, WGS is able to reveal rare and structural variants, which are crucial factors behind disease phenotypes [2]. As the cost of sequencing decreases significantly, we expect that WGS will become increasingly predominant in the hunt for novel disease genes.

Most of the recent discoveries from sequencing studies were based on the Mendelian trait model [3]. Genetic association studies based on the complex trait model are challenging because of limited sample size as well as the new properties of sequencing data. WGS data are distinct from GWAS data in two major aspects. First, WGS provides a huge number of rare variants. Even with large allelic effects, caused by very small minor allele frequencies (MAFs), the association tests between single rare variants and the trait are less powerful and unreliable [4]. Second, family designs play a critical role in WGS. Because of its relatively high cost, WGS tends to exploit families of patients, so that the rare causal variants are likely enriched through cotransmission of the disease [5]. Furthermore, the pedigree structure allows statistical imputation of the genotypes at no experimental cost, which potentially increases the statistical power [6,7].

In a recent celebrated work, Zhu and Xiong proposed a set of generalized T^2 tests for family-based WGS data [8]. These methods simultaneously address the correlations among genetic variants (i.e., linkage disequilibrium [LD]) and the correlations among family members (i.e., kinship). Rare-variant collapsing procedures [9,10] are also integrated into the tests. However, these methods cannot incorporate covariates and do not address the correlation structure for longitudinal repeated measures. In this study, we further extended the methodology of the T^2 tests to address these limits. By applying these methods to an analysis of the Genetic Analysis Workshop 18 (GAW18) simulation data, we showed that the asymptotic null distributions of Zhu and Xiong [8] are problematic in controlling the type I error rate, and that our extended methods are likely more powerful for longitudinal data.

* Correspondence: zheyangwu@wpi.edu
Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, USA

Methods

Generalized T^2 tests for family data

Zhu and Xiong [8] showed that the covariance as a result of both LD and kinship could be explicitly expressed as a Kronecker product of the corresponding covariance matrices. Following the idea of Hotelling's T^2 test [11,12], the authors proposed a generalized T^2 test that incorporates these covariance matrices, which are estimated separately by using the same data. Depending on various strategies of collapsing of rare variants, here we consider three generalized T^2 tests of Zhu and Xiong.

T^2: The genotypes of rare variants between adjacent common variants are summed up, and one covariance matrix is estimated for both common and collapsed rare variants.

CMC.ZXpaper (CMC test): The rare variants are collapsed in the same way as above, but the covariance matrices are estimated separately for common and rare variants (assuming they are uncorrelated).

CMC.ZXcode: Rare variants are collapsed by the maximum of their genotypes, and one covariance matrix is estimated for both common and collapsed rare variants. This strategy follows the R function pedCMC of Zhu and Xiong (https://sph.uth.edu/hgc/faculty/xiong/software-D.html).

Extended T^2 test for family data with longitudinal repeated measures

Building on the idea of Zhu and Xiong, we further extend the generalized T^2 tests to account for the longitudinal repeated covariates. Figure 1 shows the data structure and the idea of the extension. Specifically, the extended T^2 tests compare the blocks of common variants, rare variants, and covariates with repeated measures in cases and in controls, while simultaneously accounting for the correlations among genetic factors, among pedigree individuals, and among longitudinal repeated measures. The response is the occurrence of the event at any of the measurement points.

Following the notations in Figure 1, let n_c be the number of the cases, n_d be the number of the controls, and $n = n_c + n_d$. The genotype column vector of the tth common variant is $Z^t = (Z_1^t, \ldots, Z_n^t)'$, the aligned column vector of all T common variants is represented by $Z = (Z^{1'}, \ldots, Z^{T'})'$. Similarly, for the collapsed genotypes of rare variants, the genotype column vector of the sth rare variant is $V^s = (V_1^s, \ldots, V_n^s)'$, and $V = (V^{1'}, \ldots, V^{S'})'$ for totally S rare variants. Considering the covariates with longitudinal repeated measures, the column vector of the cth covariate at the jth repeated measurement point is $A^{cj} = (A_1^{cj}, \ldots, A_n^{cj})'$, and the aligned column vector is $A = (A^{11'}, \ldots, A^{1J'}, A^{21'}, \ldots, A^{2J'}, \ldots, A^{C1'}, \ldots, A^{CJ'})'$ for totally C covariates, each measured for J times. Similarly, the row vectors are denoted as follows. For $i = 1, \ldots, n$, the vectors $Z_i = (Z_i^1, \ldots, Z_i^T)'$ are the rows in the block of common variants, the row vectors $V_i = (V_i^1, \ldots, V_i^S)'$ are for rare variants, and $A_i = (A_i^{11}, \ldots, A_i^{1J}, A_i^{21}, \ldots, A_i^{2J}, \ldots, A_i^{C1}, \ldots, A_i^{CJ})'$ are for longitudinal covariates. The row average in cases is $\bar{Z}_c = \sum_{i=1}^{n_c} Z_i/n_c$, and that in controls is $\bar{Z}_d = \sum_{i=n_c+1}^{n} Z_i/n_d$. The row averages for rare variants and covariates are obtained analogously.

The idea of the extended T^2 test is simply to compare the difference between the row average of the case blocks and the row average of the control blocks. Let $\eta = (Z', V', A')'$. The difference between row averages can

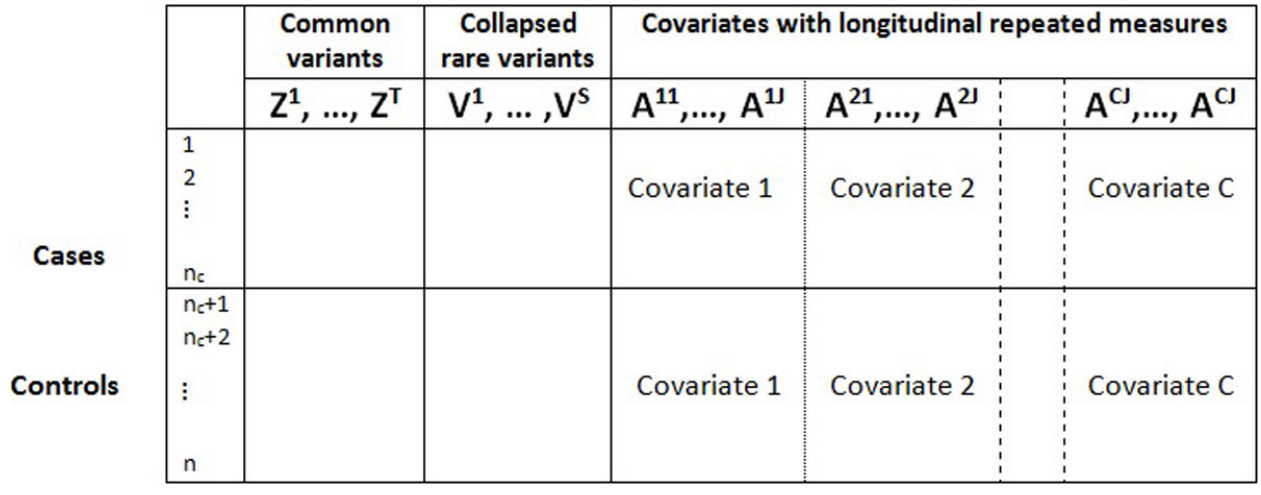

Figure 1 Data structure for composing the extended T^2 tests. Data contain 3 blocks: common variants, rare variants, and longitudinal covariate measures. The statistics integrate the correlations among both rows and columns, and test whether there exists a significant difference between the row vector mean of the cases and that of the controls.

be written in terms of η. That is

$$H\eta = \frac{n_c n_d}{n}\begin{pmatrix} \bar{Z}_c - \bar{Z}_d \\ \bar{V}_c - \bar{V}_d \\ \bar{A}_c - \bar{A}_d \end{pmatrix}, \quad (1)$$

where if we define $D_r = (u_1, \ldots, u_n)'$, with $u_i = 1$ for cases $i = 1, \ldots, n_c$ and $u_i = 0$ for controls $i = n_c + 1, \ldots, n$, and denote 1 as a vector of 1 of length n and $I_{(k)}$ as an identity matrix of dimension k, then the matrix

$$H = \begin{pmatrix} I_{(T)} \otimes (D_r - \frac{n_c}{n}1) & 0 & 0 \\ 0 & I_{(S)} \otimes (D_r - \frac{n_c}{n}1) & 0 \\ 0 & 0 & I_{(CJ)} \otimes (D_r - \frac{n_c}{n}1) \end{pmatrix}. \quad (2)$$

Following the idea of the generalized T^2 test, the extended T^2 test is $T^2 = (H\eta)'\Gamma^{-1}(H\eta)$, where $\Lambda = Var(\eta)$., $\Lambda = Var(\eta)$.

The key problem is to estimate Λ. Following the assumption of Zhu and Xiong [8] that Z and V are independent, we consider two cases. In the first case, assume A is also independent with Z and V. Then $Var(\eta) = \Lambda = diag(\Lambda_Z, \Lambda_V, \Lambda_A)$, where by $\Lambda_Z = Var(Z) = \Sigma_Z \otimes \Phi$ and $\Lambda_V = Var(V) = \Sigma_V \otimes \Phi$.

Σ_Z and Σ_V are the covariance matrix among the elements in Z_i and V_i, respectively (e.g., the LD among the genetic variants), Φ is the kinship matrix, and $\otimes$ denotes the Kronecker product. For the covariate block, we consider $\Lambda_A = Var(A) = \Sigma_A \otimes \Phi^*$, where Σ_A is the covariance matrix among the elements A_i, and Φ^* is a matrix that captures the correlations among individuals in terms of environmental covariates.

To better account for the heterogeneity of the data in cases and in controls, we applied the method in Hotelling's T^2 test for estimating the covariance matrix (which is different from equation (6) in Ref. [8]). Then equation (3) is simplified to

$$T^2 = \left(\frac{n_c n_d}{n}\right)^2 \left[\frac{(\bar{Z}_c - \bar{Z}_d)'\hat{\Sigma}_Z^{-1}(\bar{Z}_c - \bar{Z}_d) + (\bar{V}_c - \bar{V}_d)'\hat{\Sigma}_V^{-1}(\bar{V}_c - \bar{V}_d)}{\left(D_r - \frac{n_c}{n}1\right)'\Phi\left(D_r - \frac{n_c}{n}1\right)} + \frac{(\bar{A}_c - \bar{A}_d)'\hat{\Sigma}_A^{-1}(\bar{A}_c - \bar{A}_d)}{\left(D_r - \frac{n_c}{n}1\right)'\Phi^*\left(D_r - \frac{n_c}{n}1\right)}\right]. \quad (3)$$

We consider two simplification assumptions: (a) $\Phi^* = I$ indicates that covariate variables among individuals are independent, considering the individual dependence has been captured by the genetics; and (b) $\Phi^* = \Phi$ indicates that covariate variables among individuals have the similar dependence pattern as that according to genetics (e.g., children may be more likely to smoke if parents do, or the age of children is correlated with the age of parents). According to the various rare-variant collapsing strategies in the above generalized T^2 tests by Zhu and Xiong [8], the corresponding extended T^2 tests are denoted T2.longi, CMC.ZXpaper.longi, and CMC.ZXcode.longi, respectively.

Asymptotic and permutation tests

Zhu and Xiong derived asymptotic chi-square distribution for the null. In their paper [8], the degrees of freedom (DF) equal the number of variants; in their R code, the DF equal the rank of data matrix. The latter is better but still gives inflated p values as shown below. Thus, we applied a permutation test for the type I error rate being well controlled. Specifically, let T_g^2 and T_{gl}^2, $l = 1, \ldots, L$, $l = 1, \ldots, L$, denote the test statistics of the gth genome window from the original data and from the lth permutation, respectively. The empirical p value for the gth window is $p_g = \#\left\{T_{gl}^2 \geq T_g^2, l = 1, \ldots, L\right\}/L$, where L =1000. Because the target is to find the associations with genetic variants, not with the covariates, the permutations are applied only to the genotype data to retain the relationship between response and the covariates.

Results

For evaluating the above methods, we used the "dose" genotype data of 1,215,399 single-nucleotide variants (SNVs) on chromosome 3 and the 200 simulation replicates of hypertension outcomes and covariates (age, hypertension medicine use, smoking status). As an arbitrary, yet simple, way to group variants, we split chr3 into 19,080 windows, each 10 kilobase pairs (kbp) long. In each window, rare variants (MAF <0.05) between adjacent common variants were collapsed, leaving 654,415 genetic factors (common or rare variants, or collapsed rare-variant groups) to be analyzed. The average number of genetic factors contained in the windows is 34.3, the median is 32, the minimum is 1, and the maximum is 330. For the simulated phenotypes, the number of individuals is 849 in 20 families, where the family sizes are from 21 to 74, with the mean 42.45 and the median 36.5. There are 188 simulated true SNVs contained in 129 true windows (1, 3, 7, 32, and 86 windows contain 5, 4, 3, 2, and 1 true SNVs, respectively) on chr3. The knowledge of these true SNVs was used only for evaluating the power of these association tests, not for designing data analysis strategy.

To assess the asymptotic null distributions of the tests provided in Zhu and Xiong [8], we obtained the asymptotic p values of these tests for all false windows in chr3. The Q-Q plot of Figure 2 shows that all three methods have inflated p values with large genomic inflation factors λ [13]. For example, when one chooses a p value cutoff of 0.05, the actual (empirical) error rate is approximately 0.1. At the same time, the following results show that permutation test controls the type I error rate well. Thus, the inflated type I error rate is likely caused by the inappropriate asymptotic null distributions, not by possible stratification.

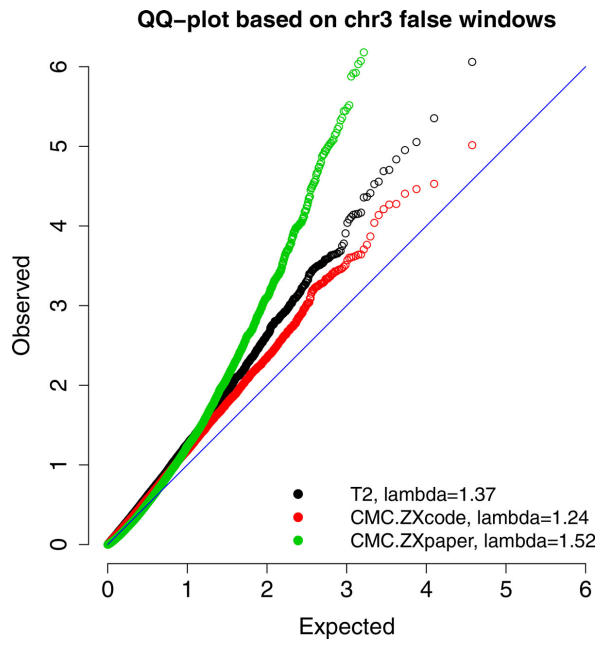

Figure 2 Q-Q plot for asymptotic *p* values of Zhu and Xiong's generalized T^2 tests. Results are based on the 18,951 false 10 kbp-long windows on chr3.

We studied the power of these tests in detecting true windows over chr3. Based on the phenotype data in the simulation replicate 1, the right panel of Figure 3 shows the receiver operating characteristic (ROC) curve for power (estimated by the true positive rate) over a variety of *p* value cutoffs. In general, the power is low at small or moderate *p* values. This phenomenon indicates that the sample size is still relatively too small for detecting many weak genetic effects simulated in the data. At the same time, it is clear that the 3 extended T^2 tests that incorporate longitudinal information are significantly better than the generalized T^2 tests that only consider the measures at the first time point. Because the two setups: $\Phi^* = I$ and $\Phi^* = \Phi$ in (3) led to similar results, we only report that by $\Phi^* = I$ for simplicity. The left panel of Figure 3 shows that the permutation test controls the type I error rates well for all methods.

To compare the overall capabilities of these tests, we studied their power (i.e., true positive rates) in detecting each of all windows over 200 simulation replicates. As illustrated in Figure 4, there are 4 representative patterns of the comparisons for the 129 true windows on chr3. In particular, 93 windows have longitudinal extended T^2 tests more powerful than generalized T^2 tests (illustrated in Figure 4, left panel), 5 windows have similar results for both (Figure 4, middle panel), 15 windows have generalized T^2 tests more powerful (Figure 4, right panel), and the remaining 16 windows have almost no power for any tests. So, the longitudinal extended T^2 tests are significantly superior to the single-time-point generalized T^2 tests (93 vs. 15, *p* value = 3.8e-15 based on binomial model). For all windows, the type I error rates of all methods were well controlled (results are available upon request).

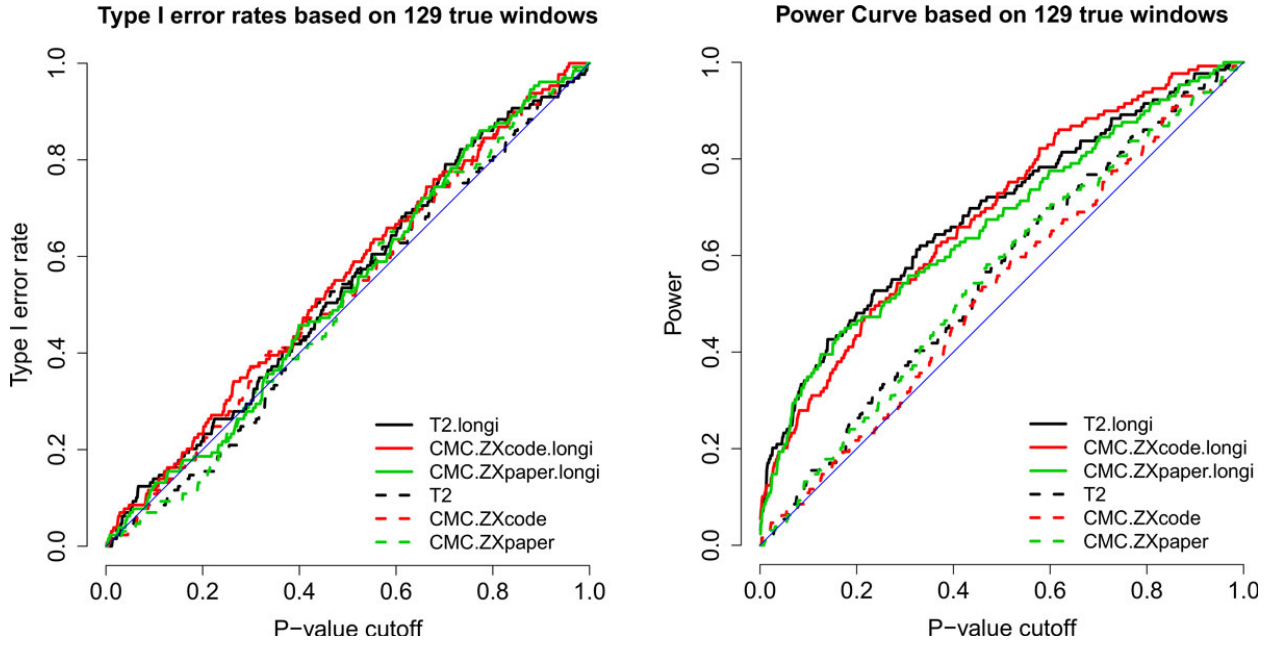

Figure 3 Type I error rate and power of detecting true windows on chr3. The power is the percentage of all 129 true windows that are detected at various *p* value cutoffs. The type I error rate is the same percentage of these windows, except the genotypes were permuted to destroy the genetic associations.

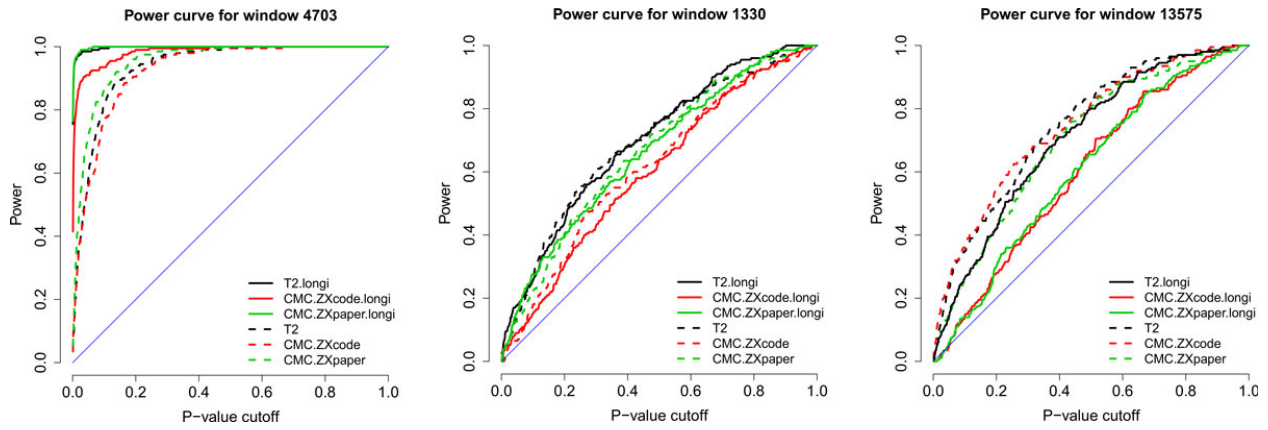

Figure 4 Three patterns of comparisons between the longitudinal tests and one-time-point tests. *Left:* longitudinal tests are better (illustrated by window 4703, 79 windows in total); *middle:* they are similar (illustrated by window 1330, 19 windows in total); *right:* one-time-point tests are better (illustrated by window 13575, 5 windows in total).

Discussion

In the simulation data of GAW18, true SNVs are always allocated on genes. Using genes as windows to group SNVs may concentrate the true SNVs and has the potential to improve the detection power. However, the idea of WGS, instead of exome sequencing, is that the disease-related genetic factors might allocate outside of genes. So we did not use the knowledge that true SNVs are in genes; instead, we evaluated the methods based on fixed-genome segment windows.

There are several limitations and future research topics based on the current work. First, matrix estimation is a key issue in this methodology development. Good estimation of matrices and their inverses can better incorporate correlation structures' potential to improve the performance. Here we simply adopted the same variance matrix estimate in Hotelling's T^2 test. This is a maximal likelihood estimate if observations are independent. Unfortunately, independency is not true for family data in the first place. Besides the correlation issue, for a high-dimensional matrix with a potentially sparse structure, there are better estimates of the covariance matrix and its inverse [14]. Second, the permutation test is relatively slow, especially for handling large amounts of data in WGS. It would be desirable to derive more accurate asymptotic distributions for fast and precise p value calculation. Third, necessary modifications of these tests are needed to handle missing data and unequal numbers of repeated measures, which are common problems.

Conclusions

We have extended Zhu and Xiong's [8] generalized T^2 tests to incorporate the covariates with longitudinal repeated measures. These methods account for 3 sources of correlation structures among genetic variants, family members, and time series observations. Compared with the T^2 test methods for snapshot observations, the new methods have higher power to detect hypertension-related genome segments according to the GAW18 simulation data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZW designed the overall study. ZW, YL, and JX conducted statistical analyses and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to the National Institutes of Health funding support for GAW18 and for the student travel awards to YL and JX. We are grateful to WPI Computing and Communications Center for computational support. 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.

Published: 17 June 2014

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, *et al*: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747-753.
2. Cirulli ET, Goldstein DB: **Uncovering the roles of rare variants in common disease through whole-genome sequencing.** *Nat Rev Genet* 2010, **11**:415-425.
3. Stitziel NO, Kiezun A, Sunyaev S: **Computational and statistical approaches to analyzing variants identified by exome sequencing.** *Genome Biol* 2011, **12**:227.

4. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: **Power of deep, all-exon resequencing for discovery of human trait genes.** *Proc Natl Acad Sci U S A* 2009, **106**:3871-3876.
5. Ott J, Kamatani Y, Lathrop M: **Family-based designs for genome-wide association studies.** *Nat Rev Genet* 2011, **12**:465-474.
6. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: **Merlin–rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2001, **30**:97-101.
7. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: **Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.** *Nat Genet* 2012, **44**:955-999.
8. Zhu Y, Xiong M: **Family-based association studies for next-generation sequencing.** *Am J Hum Genet* 2012, **90**:1028-1045.
9. 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.
10. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ: **Testing for an unusual distribution of rare variants.** *PLoS Genet* 2011, **7**:e1001322.
11. Hotelling H: **The generalization of Student's ratio.** *Ann Math Statist* 1931, **2**:360-378.
12. Xiong M, Zhao J, Boerwinkle E: **Generalized T^2 test for genome association studies.** *Am J Hum Genet* 2002, **70**:1257-1268.
13. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M, *et al*: **Genomic inflation factors under polygenic inheritance.** *Eur J Hum Genet* 2011, **19**:807-812.
14. Cai TT, Zhang CH, Zhou HH: **Optimal rates of convergence for covariance matrix estimation.** *Ann Statist* 2010, **38**:2118-2144.

doi:10.1186/1753-6561-8-S1-S40
Cite this article as: Liu *et al.*: Extended T^2 tests for longitudinal family data in whole genome sequencing studies. *BMC Proceedings* 2014 **8**(Suppl 1):S40.