Published in final edited form as:
Genet Epidemiol. 2014 September ; 38(0 1): S37–S43. doi:10.1002/gepi.21823.

Testing Genetic Association with Rare and Common Variants in Family Data

Han Chen[1,‡], **Dörthe Malzahn**[2,‡], **Brunilda Balliu**[3], **Cong Li**[4], and **Julia N. Bailey**[5,6,*]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America [2]Department of Genetic Epidemiology, University Medical Center, Georg-August University Göttingen, Göttingen, Germany [3]Department of Medical Statistics and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands [4]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America [5]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California, United States of America [6]Epilepsy Genetics/ Genomics Laboratory, West LA VA, Los Angeles, California, United States of America

Abstract

With the advance of next-generation sequencing technologies in recent years, rare genetic variant data have now become available for genetic epidemiology studies. For family samples however, only a few statistical methods for association analysis of rare genetic variants have been developed. Rare variant approaches are of great interest particularly for family data because samples enriched for trait-relevant variants can be ascertained and rare variants are putatively enriched through segregation. To facilitate the evaluation of existing and new rare variant testing approaches for analyzing family data, Genetic Analysis Workshop 18 (GAW18) provided genotype and next-generation sequencing data and longitudinal blood pressure traits from extended pedigrees of Mexican-American families from the San Antonio Family Study. Our GAW18 group members analyzed real and simulated phenotype data from GAW18 by using generalized linear mixed-effects models or principal components to adjust for familial correlation or by testing binary traits using a correction factor for familial effects. With one exception, approaches dealt with the extended pedigrees in their original state using information based on the kinship matrix or alternative genetic similarity measures. For simulated data, our group demonstrated that the family-based kernel machine score test is superior in power to family-based single-marker or burden tests, except in a few specific scenarios. For real data, three contributions identified significant associations. They substantially reduced the number of tests before performing the association analysis. We conclude from our real data analyses that further development of strategies for targeted testing or more focused screening of genetic variants is strongly desirable.

[*]Correspondence to: JBailey@mednet.ucla.edu, UCLA Department of Epidemiology, Fielding School of Public Health, Box 9151772, Los Angeles, California, 90095, 310-268-3129.
[‡]These authors contributed equally.

Keywords

extended pedigrees; rare variant analysis; family-based association test; linear mixed effects model; kernel machine score test; principal components

Introduction

Although genome-wide association studies (GWAS) of complex human diseases and quantitative traits have identified large numbers of associated genetic markers, these markers explain only a small proportion of the total heritability for most traits. Rare genetic variants may account for some of the unexplained heritability [Eichler et al., 2010], but single-marker tests, which are widely used in traditional GWAS of common variants, do not have enough power to analyze rare variants. To overcome this difficulty, in recent years investigators have proposed many statistical approaches for rare variant analysis [Han and Pan, 2010; Hoffmann et al., 2010; Lee et al., 2012; Li and Leal, 2008; Lin and Tang, 2011; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Neale et al., 2011; Pan, 2009; Price et al., 2010a; Wu et al., 2011]. However, these approaches were developed for association analysis in unrelated individuals.

Family-based study designs have long been used in association studies of diseases and quantitative traits [Falk and Rubinstein, 1987; Ott, 1989; Spielman et al., 1993; Terwilliger and Ott, 1992]. One way to controll for familial correlation is the family-based association test (FBAT) [Rabinowitz and Laird, 2000]. However, this approach requires known pedigree information, and offspring with homozygous parents (e.g., because of low minor allele frequency) are not useful for the test statistic. Alternatively, linear or generalized linear mixed effects models [Almasy and Blangero, 1998; Amos, 1994; Pankratz et al., 2005; Breslow and Clayton, 1993] and generalized estimating equations [Liang and Zeger, 1986] can be used to account for relatedness.

Family studies have several advantages over studies of independent subjects. First, family structure is useful for imputing missing genotypes [Chen and Abecasis, 2007], which makes them attractive when the genotyping or sequencing budget is limited. Second, higher statistical power can be attained by ascertaining families that are enriched with trait-related variants. Segregation within families can also yield more observations of such rare variants, compared to the general population. Third, family-based designs can immediately determine whether a particular variant is segregating with disease status. For these reasons, family studies have become more popular in recent years.

The organizers of Genetic Analysis Workshop 18 (GAW18) provided genotype and whole-genome sequencing (WGS) data and longitudinal blood pressure traits at four time points from 20 extended Mexican-American families with 21 – 76 family members participating in the San Antonio Family Study. Almasy et al. [2014] provide a detailed description of the data. Briefly, there are three sets of genetic data for 959 subjects. The first data set (gwas) contains genotypes for a tag single-nucleotide polymorphism (SNP) panel assayed using different versions of Illumina Infinium Beadchips. The second data set (geno) contains genotypes for a much denser SNP panel (whole-genome sequence for 464 subjects and imputed most probable genotypes for the sequence of all remaining subjects). The third data set (dose) contains dosages for the whole-genome sequence (either original for available genotypes or imputed) and is the only genetic data without missing values. All genetic data contain only odd-numbered chromosomes. Examination year, age, hypertension (HTN), systolic blood pressure (SBP), diastolic blood pressure (DBP), anti-hypertensive medication and smoking status are provided for 939 subjects at four time points, with missing values. GAW18 also provided 200 simulation replicates, in which age, DBP, SBP, HTN, anti-hypertensive medication, and smoking status are reported for 849 subjects at three time points, with no missing values. In addition, a quantitative trait Q1 is simulated under the null hypothesis of no genetic association.

Our GAW18 working group on family-based tests of association for rare variants using real and simulated data explored and evaluated rare and common variant methods for family data. All group contributors analyzed the real phenotype data (Table I), with some also performing additional analyses of the simulated data. Out of nine contributions, seven analyzed at least one quantitative trait and three analyzed the binary trait HTN. Most contributions used the dose genotypes and evaluated multi-marker approaches.

A summary of statistical tests performed in our working group is shown in Table II. Balliu et al. [2014], He and Pitkäniemi [2014], Li et al. [2014], Liu et al. [2014] and Malzahn et al. [2014] developed new methods or extended existing ones; the other contributors applied existing methods for family data. He and Pitkäniemi [2014] constructed their test statistic for nuclear families. All other approaches dealt with the extended pedigrees in their original state using the kinship matrix or another measure of genetic similarity. The applied statistical methods can generally be classified into two categories: testing binary traits with a correction factor for familial effects (as proposed by Zhu and Xiong [2012]), or adjusting for familial correlation in binary and quantitative traits by using the (generalized) linear mixed effects model framework. We describe the two categories in more detail in the Section Methods.

Methods

Testing binary traits with a correction factor for familial effects

Recently, Zhu and Xiong [2012] extended several test statistics for binary traits to accommodate families (among them are the generalized T^2 test [Xiong et al., 2002] and the combined multivariate and collapsing (CMC) test [Li and Leal, 2008]). They showed that familial effects on these test statistics can be written as a correction factor

$$P_{corr}=\frac{n}{n_c(n-n_c)}\left(D_r-\frac{n_c}{n}1\right)^T\Phi\left(D_r-\frac{n_c}{n}1\right), \quad (1)$$

where n is the sample size, D_r is a vector of size n indicating the disease status, $n_c=D_r^T 1$ is the total number of cases, 1 is a vector of size n with all 1′s, and Φ is the kinship matrix. Hainline et al. [2014] compared performances of Zhu and Xiong's [2012] family-based generalized T^2 test and the CMC test on the binary outcome HTN in real data. Originally, these tests did not allow adjusting for covariates.

Liu et al. [2014] extended the T^2 test to adjust for longitudinal covariates, and combined it with various strategies of rare variant collapsing, analyzing the binary outcome HTN in the real data. P-values were obtained by permuting genotypes across all study subjects and across all families.

Generalized linear mixed-effects models

All other approaches used by members of our working group can be formulated in the generalized linear mixed-effects model framework. The basic model is

$$g(E(y)) = X\beta + G\gamma + \delta, \quad (2)$$

where y is the phenotype, X are the covariates, G are the (weighted) genotypes, and δ is the familial random effect. We are interested in testing H_0: $\gamma = 0$.

Three contributions considered extensions of this general framework: Zhang and Pan [2014] compared familial adjustment by random effects $\delta \sim N(0, \sigma_F^2 M)$, in contrast to an adjustment by fixed effects of the most prominent principal components (PC's) of M. M was either the identity-by-state (IBS) matrix or the genetic covariance matrix of the sample. He and Pitkäniemi [2014] decomposed the genotypes into family expected scores and deviations, treated collapsed family expected scores as X and deviations as G, and performed the test using a Bayesian approach. Their model was specifically tailored to rare variants with minor allele frequency (MAF) less than 1% to facilitate computational efficiency. Balliu et al. [2014] used a two-stage approach; in the first stage they used a mixed-effects model with a general random effects structure to capture the correlation between the SNPs in a region, and in the second stage they tested for region specific effects by using the empirical Bayes estimates of the random effects of the first stage as covariates in the model for the longitudinal phenotype.

Zhang and Pan [2014] analyzed both binary and quantitative traits; all other mixed-effects model approaches analyzed only quantitative traits. Family-based single-marker or burden tests and multi-marker tests were performed. In the next subsections we briefly describe these methods.

Single marker tests and burden tests (collapsing)—In the generalized linear mixed-effects model given in Eq. (2), when G is either a single marker, or a variant sum score (collapsed variant burden, which is the weighted sum of multiple genotypes in a gene or a predefined genomic region), the fixed effect test H_0: $\gamma = 0$ becomes a univariate test of fixed effects in a mixed-effects model. Single-marker tests have been widely used in GWAS for common variants. For rare variants single-marker tests are not likely to be powerful because of small genotype subgroup sizes. Burden tests on variant sum scores ameliorate this issue. They often restrict the model space because of an a priori choice of variant weights. Burden tests reduce the number of tests per region and are widely used. Tests on the fixed effects in linear mixed effects models are implemented in various software packages (e.g., kinship in R [Pankratz et al., 2005], SOLAR [Almasy and Blangero, 1998], EMMAX [Kang et al., 2010], and GEMMA [Zhou and Stephens, 2012]).

Kernel machine score tests—Kernel machine score tests provide a joint test over marker sets without collapsing the marker information. These tests were initially proposed to analyze genetic pathway data [Liu et al., 2007; Liu et al., 2008] and to analyze SNP sets of quantitative and binary traits [Kwee et al., 2008; Wu et al. 2010]. In the context of rare variant analyses, Wu et al. [2011] proposed the sequence kernel association test (SKAT) and demonstrated that it is a flexible, computationally easy and powerful approach in various scenarios. Recently, the approach was extended to family samples by several independent contributions [Schifano et al., 2012; Chen et al., 2013; Malzahn et al., 2014; Li et al., 2014], the last two being GAW18 contributions from our working group.

Using the linear kernel, in Eq. (2), we assume genotypic effects $\gamma \sim N(0, \tau W)$ where τ is the variance component parameter and W is either the identity matrix or a diagonal matrix of pre-specified marker weights. The hypothesis H_0: $\gamma = 0$ is equivalent to testing the variance parameter H_0: $\tau = 0$. We first fit the null model

$$y = X\beta + \delta + \varepsilon \quad (3)$$

and get the covariance matrix estimate

$$\widehat{V} = \widehat{\sigma}_E^2 I + \widehat{\sigma}_F^2 \Phi, \quad (4)$$

where $\widehat{\sigma}_E^2$ and $\widehat{\sigma}_F^2$ are variance parameter estimates from the null model corresponding to the random error and the familial correlation, respectively, and Φ is the kinship matrix, either calculated from family structures or estimated from genotype data. The fixed-effect estimates for covariates are

$$\widehat{\beta} = (X^T \widehat{V}^{-1} X)^{-1} X^T \widehat{V}^{-1} y, \quad (5)$$

and the test statistic is

$$Q = (y - X\widehat{\beta})^T \widehat{V}^{-1} G W G^T \widehat{V}^{-1} (y - X\widehat{\beta}). \quad (6)$$

Under the null hypothesis, Q follows a weighted sum of independent chi-square distributions with 1 degree of freedom, with weights equal to the eigenvalues of

$$K = P^{\frac{1}{2}} G W G^T P^{\frac{1}{2}}, \quad (7)$$

where

$$P=\widehat{V}^{-1}-\widehat{V}^{-1}X\left(X^T\widehat{V}^{-1}X\right)^{-1}X^T\widehat{V}^{-1}. \quad (8)$$

In GAW18, Malzahn et al. [2014] estimated type I error, compared power for the WGS in contrast to using only GWAS SNPs and performed candidate gene analysis in the real data set with this method. Moreover, performance of the family-based kernel machine score test was compared with family-based versions of single marker tests by Li et al. [2014], and a burden test (collapsing) by Chen et al. [2014].

Adjustment for covariates and relatedness

Hainline et al. [2014] used methods that do not adjust for any covariates, but all other contributions adjusted for age, and some also adjusted for sex, smoking status or both. Malzahn et al. [2014] also adjusted for the interaction between age and sex.

Except for He and Pitkäniemi [2014], all other contributions used the kinship matrix (or another measure of genetic similarity), either in mixed effects models or in the binary trait tests proposed by Zhu and Xiong [2012]. Six research groups used the exact kinship matrix from the family structures, three contributions used estimated kinship from the genotype data, and Hainline et al. [2014] compared both. Zhang and Pan [2014] compared the use of common variants or rare variants for estimation of genetic correlations expressed by either the genotypic covariance matrix or the IBS matrix.

Anti-hypertensive medication and treatment of repeated measures

Anti-hypertensive medication could affect the quantitative traits SBP and DBP. In our working group five contributions did not adjust for medication and three contributions excluded subjects on medication. Liu et al. [2014] included anti-hypertensive medication in their model as a longitudinal covariate.

Repeated measures were treated differently in our working group. Balliu et al. [2014] and Liu et al. [2014] performed longitudinal data analysis. Five research groups analyzed only the baseline data, defined as either the first examination or the first non-missing examination (see Table 1). Hainline et al. [2014] used an indicator for whether a subject had ever had hypertension at any examination as their outcome of interest, and He and Pitkäniemi [2014] took the maximum log-transformed SBP at any examination.

Selection of genetic variants or genomic regions

Five research groups analyzed both rare and common variants [Balliu et al., 2014; Hainline et al., 2014; Li et al., 2014; Liu et al., 2014; Malzahn et al., 2014]. Chen et al. [2014] and He and Pitkäniemi [2014] focused on rare variants with a MAF less than 5% or 1%. Lacey et al. [2014] and Zhang and Pan [2014] focused on common variants with a MAF greater than 1% or 5% and performed only single-marker tests.

For analyses of the real data, five research groups performed genome-wide analyses [Chen et al., 2014; Hainline et al., 2014; He and Pitkäniemi, 2014; Lacey et al., 2014; Li et al., 2014]. Lacey et al. [2014] compared genome-wide analyses with exome-wide analyses. Three contributions analyzed only chromosome 3 or selected regions on chromosome 3, selecting regions with an excess of identity-by-descent (IBD) sharing of common variants [Browning and Thompson, 2012] in HTN cases compared to control subjects before quantitative trait analysis of these regions [Balliu et al., 2014], or restricting the analysis to candidate genes with previous association reports [Malzahn et al., 2014]. Zhang and Pan [2014] randomly selected 6,228 common variants for analysis.

Evaluation of type I error and power

Three research groups evaluated both type I error based on 200 replicates of the simulated trait Q1 and power using genes, including *MAP4*, the most significant gene in the simulated data [Chen et al., 2014; Li et al., 2014; Malzahn et al., 2014] (see Table 2). Hainline et al. [2014] used HTN only in the first simulation replicate to evaluate type I error and power. Type I error rates were calculated using 6,454 genes with no causal variants, and power was determined using 171 genes with at least 1 causal variant. Liu et al. [2014] used all simulation replicates of HTN but only genotypes on chromosome 3. He and Pitkäniemi [2014] simulated segregation of rare causal variants with MAF less than 1% (genotype and phenotype data) to evaluate type I error and power. Zhang and Pan [2013] investigated only the type I error, using 6,228 randomly selected common variants.

Results

Real data analysis

Results from the real data analysis are summarized in Table 3. Most contributors did not find significant results after correcting for multiple testing. None of the significant findings were obtained consistently by more than a single approach. This is expected because different phenotypes were analyzed, different strategies were used to greatly reduce the number of tests in advance, and, most important, different genetic variants were studied. He and Pitkäniemi [2014] analyzed rare variants (MAF<1%) and tested association of such variants within genes with log-transformed SBP. Balliu et al. [2014] used common variants (MAF>5%) to determine regions with excess IBD sharing on chromosome 3 and tested association of these regions with DBP. Lacey et al. [2014] performed single marker tests on exonic SNPs for SBP. Malzahn et al. [2014] tested association of candidate genes on chromosome 3 with rank-normalized SBP. Interestingly, the candidate genes *SLC4A7* and *ULK4* (examined by Malzahn et al. [2014] as a follow-up to previous association reports) are located close to regions with an excess of IBD sharing in GAW18 HTN cases [Balliu et al., 2014] (see Table 3 for details). *SLC4A7* and *ULK4* were not significantly associated in the GAW18 data. However, using a denser SNP panel (WGS instead of GWAS variants) lowered p-values for rare *SLC4A7* variants with MAF less than or equal to 5% [Malzahn et al., 2014].

Malzahn et al. [2014] compared the performance of a multimarker test (family extension of the kernel machine score test) when applied to WGS and GWAS SNP panels. This altered the number of jointly tested SNPs within gene regions, whereas always the same subjects were analyzed and the same types of tests were performed. Sets of common and rare SNPs should be tested separately or with unequal weights. Compared to GWAS, WGS data were found to improve the power especially for joint tests of rare variants (MAF<1%) in the simulated GAW18 data (see also the real data result for *SLC4A7*). When more SNPs than subjects are available in a region of interest, Malzahn et al. [2014] suggested splitting the large SNP sets and jointly testing only fewer SNPs than subjects.

Lacey et al. [2014] compared the WGS and GWAS SNP panels with three exome sequencing panels (mimicked by restricting the analysis to targeted regions designed by the exome sequencing platforms). At a false discovery rate of 5%, results from both the GWAS and the WGS data did not pass multiple-testing adjusted significance thresholds, whereas exome sequencing data had a much smaller multiple-testing burden and found that variants at positions 11022564 and 11022230 on chromosome 7 were significantly associated with SBP consistently in all three exome platforms. For blood pressure, Lacey et al. [2014] concluded that exome sequencing was a more cost-effective way to capture trait-associated variants.

Zhang and Pan [2014] found that the top PCs from the genetic covariance matrix controlled the type I error rates as effectively as the top PCs of the IBS matrix, provided that both matrices were estimated from common variants. However, using the covariance matrix was less effective than using the IBS matrix when the matrices were constructed from rare variants. They also demonstrated that in the GAW18 data PC adjustment gave equally good type I error control compared to adjustment for random familial effects. This finding is different from previous studies, which favored the random effects adjustment over PCs [Price et al., 2010b].

Simulation results

Members of our working group found correct type I error for the kernel machine score tests with familial adjustment in simulated data. Power analyses on association tests with SBP in simulated data are summarized in Table 4. The performance of family-based versions of the kernel machine score tests was compared to the performance of burden tests [Chen et al., 2014] and single-marker tests [Li et al., 2014]. When the proportion of causal variants was large and all variants had the same direction of effects (as for all nonsynonymous coding variants of *MAP4*), burden tests were likely to outperform kernel machine score tests. This is consistent with the conclusions by Liu et al. [2013], who classified burden tests as length tests and SKAT as a joint test of lengths and angles from a geometric point of view. On the other hand, single-marker tests can be more powerful than kernel machine score tests when the association in a region is driven by one dominating causal variant (as in *LEPR*). For most other scenarios however, kernel machine score tests have much higher power.

Hainline et al. [2014] found correct type I error for the T^2 test and CMC approach with the familial correction factor proposed by Zhu and Xiong [2012], however the power was even lower than the nominal level. In contrast, Liu et al. [2014] found inflated type I errors of these tests on unassociated SNPs. Correct type I error, however, was seen for this set of methods with permutation testing. When Liu et al. extended these methods to adjust for longitudinal covariates, they found improved power in 93 out of 129 windows on chromosome 3. Their test extension was less powerful than Zhu and Xiong's tests in only 15 windows.

By comparing results using either estimated kinship from genotypes or exact kinship from pedigrees, Hainline et al. [2014] found high correlation of resulting p-values but tests were more conservative on average when using exact kinship from pedigrees.

Discussion

Adjustment for familial correlation can be accomplished by using family-based tests, principal components analysis, or mixed-effects models. Existing methods for association analysis in unrelated individuals, such as single-marker tests, burden tests, and kernel machine score tests, can be extended to related individuals based on the linear mixed models framework.

It has long been debated whether the exact kinship from pedigree information or the estimated kinship from genotype data should be used to account for correlation [Astle and Balding, 2009]. The exact kinship matrix is often easier to compute but does not capture cryptic relatedness, a problem often encountered when samples from isolated populations or with inbreeding are analyzed. In contrast, estimated kinship captures more information and can also be used when pedigree information is not available. The exact kinship matrix might be preferred, however, when pedigree information is available but only a few genetic markers are genotyped (e.g., in candidate gene studies, targeted sequencing studies, or fine mapping studies). Hainline et al. [2014] showed that use of the exact kinship matrix yields slightly more conservative results than using the estimated kinship matrix in the GAW18 data.

The longitudinal observations in GAW18 were used in only two contributions in our working group. In the mixed effects model context, it is straightforward to incorporate repeated measurements and familial correlation as separate random effects and these models can also accommodate missing data resulting from loss to follow up. However, several methods in our contributions depend on complete data, and in the real GAW18 data about 33% of the original sample was lost to follow-up at the third examination (about 71% were lost to follow-up at the fourth examination). This necessitates future method development to accommodate missing data.

As sequencing costs become lower, more related individuals in family-based studies will be sequenced and more rare variants will likely be discovered. How to best analyze the data remains a topic of active research in this field, and there is no simple answer yet. A key to the success of real data analyses in our working group was strategies to reduce the number of association tests and tested regions. In this respect, screening for IBD sharing [Browning and Thompson, 2012] was an attractive strategy successfully explored in our group. Interestingly it highlighted regions close to previous genome-wide association reports and also novel regions, some of which formed regional clusters in close proximity to each other. However, whether trait-related variants are located exclusively in genomic areas with certain putative functions may well depend on the considered trait (see, e.g., [Freedman et al., 2011]).

Our GAW18 group contributed one novel rare variants approach [He and Pitkäniemi, 2014]. This method is restricted to variants with MAF less than 1% (for computational efficiency of Bayesian estimates), is family-based in an FBAT-like manner (considering nuclear families), and allows for covariate adjustment. However, statistical significance is quantified in terms of Bayes factors instead of p-values.

The binary trait analysis with a familial correction factor according to the method of Zhu and Xiong [2012] appears to have problems on the extended GAW18 pedigrees in terms of type I error and power. Members of our working group adjusted for covariates, but that was computationally tedious. The family-based version of the kernel machine score test can be straightforwardly extended to binary traits and allows for covariate adjustment and for partial collapsing of rare variants in a simple manner.

Acknowledgments

The Genetic Analysis Workshops are supported by the National Institutes of Health (NIH) through grant R01 GM031575. HC was supported by NIH grant U01 DK85526. DM was supported by the Deutsche Forschungsgemeinschaft (through grant Klinische Forschergruppe [KFO] 241: TP5, BI 576/5-1) and the German Federal Ministry of Education and Research BMBF (through German National Genome Research Net NGFN grant 01GS0837). CL was supported by a graduate student fellowship from the China Scholarship Council. JNB was supported by NIH grant R01 NS055057. We would like to thank Liang He, Nathan Tintle, Yiwei Zhang and Zheyang Wu for their comments and suggestions.

References

Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998; 62:1198–1211. [PubMed: 9545414]

Almasy L, Dyer TD, Peralta JM, Jun G, Fuchsberger C, Almeida MA, Kent JW Jr, Fowler S, Duggirala R, Blangero J. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. BMC Proc. 2014; 8(Suppl 1):S2. [PubMed: 25519314]

Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet. 1994; 54:535–543. [PubMed: 8116623]

Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Statistical Science. 2009; 24:451–471.

Balliu B, Uh HW, Tsonaka R, Boehringer S, Helmer Q, Houwing-Duistermaat JJ. Combining information from linkage and association mapping for next generation sequencing longitudinal family data. BMC Proc. 2014; 8(Suppl 1):S34. [PubMed: 25519382]

Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993; 88:9–25.

Browning SR, Thompson EA. Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics. 2012; 190:1521–1531. [PubMed: 22267498]

Chen H, Choi SH, Hong J, Lu C, Milton JN, Allard C, Lacey SM, Lin H, Dupuis J. Rare genetic variant analysis on blood pressure in related samples. BMC Proc. 2014; 8(Suppl 1):S35. [PubMed: 25519320]

Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013; 37:196–204. [PubMed: 23280576]

Chen WM, Abecasis GR. Family-based association tests for genomewide association scans. Am J Hum Genet. 2007; 81:913–926. [PubMed: 17924335]

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450. [PubMed: 20479774]

Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet. 1987; 51:227–233. [PubMed: 3500674]

Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, de Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Vikis HG, You M, Mills IG. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011; 43:513–518. [PubMed: 21614091]

Hainline A, Alvarez C, Luedtke A, Greco B, Beck A, Tintle NL. Evaluation of the power and type I error of recently proposed family-based tests of association for rare variants. BMC Proc. 2014; 8(Suppl 1):S36. [PubMed: 25519321]

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70:42–54. [PubMed: 20413981]

He L, Pitkäniemi JM. Family-based Bayesian collapsing method for rare variant association study. BMC Proc. 2014; 8(Suppl 1):S37. [PubMed: 25519322]

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One. 2010; 5:e13584. [PubMed: 21072163]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354. [PubMed: 20208533]

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 2008; 82:386–397. [PubMed: 18252219]

Lacey S, Chung JY, Lin H. A comparison of whole genome sequencing with exome sequencing for family-based association studies. BMC Proc. 2014; 8(Suppl 1):S38. [PubMed: 25519383]

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13:762–775. [PubMed: 22699862]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–321. [PubMed: 18691683]

Li C, Yang C, Chen M, Chen X, Hou L, Zhao H. Adjustment of familial relatedness in association test for rare variants. BMC Proc. 2014; 8(Suppl 1):S39. [PubMed: 25519384]

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

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. [PubMed: 21885029]

Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008; 9:292. [PubMed: 18577223]

Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007; 63:1079–1088. [PubMed: 18078480]

Liu K, Fast S, Zawistowski M, Tintle NL. A geometric framework for evaluating rare variant tests of association. Genet Epidemiol. 2013; 37:345–357. [PubMed: 23526307]

Liu Y, Xuan J, Wu Z. Extended T2 tests for longitudinal family data in whole genome sequencing studies. BMC Proc. 2014; 8(Suppl 1):S40. [PubMed: 25519385]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]

Malzahn D, Friedrichs S, Rosenberger A, Bickeböller H. Kernel score statistic for dependent data. BMC Proc. 2014; 8(Suppl 1):S41. [PubMed: 25519324]

Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007; 615:28–56. [PubMed: 17101154]

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

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 genetics. 2011; 7:e1001322. [PubMed: 21408211]

Ott J. Statistical properties of the haplotype relative risk. Genet Epidemiol. 1989; 6:127–130. [PubMed: 2731704]

Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009; 33:497–507. [PubMed: 19170135]

Pankratz VS, de Andrade M, Therneau TM. Random effects Cox proportional hazard model: general variance components methods for time-to-event data. Genet Epidemiol. 2005; 28:97–109. [PubMed: 15532036]

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010a; 86:832–838. [PubMed: 20471002]

Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010b; 11:459–463. [PubMed: 20548291]

Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered. 2000; 50:211–223. [PubMed: 10782012]

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, Lin X. SNP set association analysis for familial data. Genet Epidemiol. 2012; 36:797–810. [PubMed: 22968922]

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993; 52:506–516. [PubMed: 8447318]

Terwilliger JD, Ott J. A haplotype-based ‘haplotype relative risk’ approach to detecting allelic associations. Hum Hered. 1992; 42:337–346. [PubMed: 1493912]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Poweful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010; 86:929–942. [PubMed: 20560208]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]

Xiong M, Zhao J, Boerwinkle E. Generalized T2 test for genome association studies. Am J Hum Genet. 2002; 70:1257–1268. [PubMed: 11923914]

Zhang Y, Pan W. Adjusting for population stratification and relatedness with sequencing data. BMC Proc. 2014; 8(Suppl 1):S42. [PubMed: 25519386]

Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44:821–824. [PubMed: 22706312]

Zhu Y, Xiong M. Family-based association studies for next-generation sequencing. Am J Hum Genet. 2012; 90:1028–1045. [PubMed: 22682329]

Table 1

Summary of real data analyzed.

Contribution	Phenotypes	Genotypes	Covariates	Medication	Longitudinal	Variants	Scope
Balliu et al. [2014]	DBP	dose, geno	Age, smoke	unadjusted	longitudinal	all	Chromosome 3 with excess IBD
Chen et al. [2014]	rSBP, DBP	dose	Age, sex, smoke	excluded	baseline: first non-missing	MAF < 5%	genome
Hailine et al. [2014]	HTN	dose	None	unadjusted	any(HTN)	all	genome
He and Pitkäniemi [2014]	log(SBP)	dose	Age	unadjusted	max(SBP)	MAF < 1%	genome
Lacey et al. [2014]	SBP	dose, gwas	Age, sex, smoke	excluded	baseline: time1	MAF > 1%	genome
Li et al. [2014]	SBP	dose	Age, sex	unadjusted	baseline: first non-missing	all	genome
Liu et al. [2014]	HTN	dose	Age, smoke	adjusted	longitudinal	all	Chromosome 3
Malzahn et al. [2014]	rSBP	dose, gwas	Age, sex, age×sex	excluded	baseline: time1	all	Five genes on Chromosome 3
Zhang and Pan [2014]	HTN, SBP	geno	Age, sex, smoke	unadjusted	baseline: time1	MAF > 5%	6,228 common variants

DBP, diastolic blood pressure; HTN, hypertension; IBD, identity-by-descent; MAF, minor allele frequency; rSBP, rank-transformed SBP; SBP, systolic blood pressure.

Table 2

Summary of statistical tests performed.

Contribution	**Methods**	**Mixed models**	**Kinship**	**Type I error**	**Power**
Balliu et al. [2014]	Two-stage	Yes	Exact	No	No
Chen et al. [2014]	BT, SKAT	Yes	Exact	Q1	*MAP4*
Hailine et al. [2014]	SMT, T^2, CMC	No	Exact, estimated	HTN (sim1)	HTN (sim1)
He and Pitkäniemi [2014]	FBCM	Yes	Family-level indicator	own simulation	own simulation
Lacey et al. [2014]	SMT	Yes	Exact	No	No
Li et al. [2014]	SMT, SKAT	Yes	Estimated	Q1	7 genes
Liu et al. [2014]	T^2, CMC	No	Exact	HTN (chromosome 3)	HTN (chromosome 3)
Malzahn et al. [2014]	SKAT	Yes	Exact	Q1	*MAP4*
Zhang and Pan [2014]	SMT, PCA	Yes	Estimated	SBP, HTN	No

BT, family-based burden test; CMC, family-based combined multivariate and collapsing method; FBCM, family-based Bayesian collapsing method; PCA, principal components analysis; SBP, systolic blood pressure; SKAT, family-based sequence kernel association test; SMT, family-based single marker test.

Table 3

Summary of real data analysis results.

Contribution	Phenotype	Chromosome	Region or gene	Significance Level	P-value	Bayes Factor
Balliu et al. [2014]	DBP	3	Excess of IBD sharing:	7×10^{-3}		NA
			29239664 – 29531222		n.s.	
			34834899 – 35282759		n.s.	
			35718847 – 36018767		n.s.	
			36815704 – 37526013		n.s.	
			40249244 - 41025167		2.3×10^{-3}	
			167635899 - 168125439		n.s.	
			168621773 - 168859006		n.s.	
He, Pitkäniemi [2014]	log(SBP)	5	*HNRNPA1P13*	NA	NA	8.52
	log(SBP)	19	*OR7C2*	NA	NA	4.82
	log(SBP)	15	*RP11-138E16.2*	NA	NA	4.43
Lacey et al. [2014]	SBP	7	11022564	0.05 (FDR)	8.8×10^{-7}	NA
	SBP	7	11022230	0.05 (FDR)	1.0×10^{-6}	NA
Malzahn et al. [2014]	rSBP	3	*MECOM*: 168800190 – 169397295 *ULK4*: 41301508 - 41979954 *PPARG*: 12340020 – 124797518 *AGTR1*: 148400139 – 148499684 *SLC4A7*: 27410561 - 27489996	0.01 per gene	n.s., joint test of rare variants: p-values reduce with WGS for *AGTR1, SLC4A7*	NA

DBP, diastolic blood pressure; FDR, false discovery rate; IBD, identity by descent; log(SBP), log-transformed SBP; NA, not applicable; n.s., not significant after multiple testing adjustment; rSBP, rank-transformed SBP; SBP, systolic blood pressure; WGS, whole-genome sequencing.

Table 4

Summary of power analyses on association tests with SBP in simulated data.

Contribution	Variants	Chromosome	Region	Total variance explained (%)	Significance level	SMT power[a]	BT power	SKAT power
Chen et al. [2014]	MAF<5%, nonsynonymous	3	*MAP4*	7.79	2.5×10^{-6}	NA	1.000	1.000
	MAF<5%, sliding window[b]	3	*MAP4*	7.79	3.3×10^{-8}	NA	0.005	0.945
Li et al. [2014]	Nonsynonymous	3	*MAP4*	7.79	0.05	1.000	NA	1.000
	Nonsynonymous	1	*TNN*	3.87	0.05	0.740	NA	0.825
	Nonsynonymous	1	*LEPR*	2.23	0.05	0.895	NA	0.730
	Nonsynonymous	13	*FLT3*	0.97	0.05	0.300	NA	0.305
	Nonsynonymous	3	*FLNB*	0.29	0.05	0.165	NA	0.515
	Nonsynonymous	19	*ZNF443*	0.22	0.05	0.035	NA	0.050
	Nonsynonymous	9	*GSN*	0.21	0.05	0.130	NA	0.385
Malzahn et al. [2014]	All	3	*MAP4*	7.79	0.05	NA	NA	1.000
	All	3	*MAP4*	7.79	1.0×10^{-4}	NA	NA	0.625
	All	3	*MAP4*	7.79	1.0×10^{-8}	NA	NA	0.215

[a] Bonferroni correction on the minimum p-values.

[b] Sliding windows of length 4 kb, with 2-kb overlaps. P-values were calculated by multiplying the minimum p-value by 60 (*MAP4* gene length: 239 kb).

BT, family-based burden test; MAF, minor allele frequency; NA, not applicable; SKAT, family-based sequence kernel association test; SMT, family-based single-marker test.