Complex Pedigrees in the Sequencing Era: To Track Transmissions or Decorrelate?

Dalin Li[1,2,†], **Jin Zhou**[3,†], **Duncan C. Thomas**[4], and **David W. Fardo**[5,*]

[1]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California, United States of America

[2]David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America

[3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America

[4]Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America

[5]Department of Biostatistics, University of Kentucky, Lexington, Kentucky, United States of America

Abstract

Next-generation sequencing (NGS) studies are becoming commonplace, and the NGS field is continuing to develop rapidly. Analytic methods aimed at testing for the various roles that genetic susceptibility plays in disease are also rapidly being developed and optimized. Studies that incorporate large, complex pedigrees are of particular importance because they provide detailed information about inheritance patterns and can be analyzed in a variety of complementary ways. The nine contributions from our Genetic Analysis Workshop 18 working group on family-based tests of association for rare variants using simulated data examined analytic methods for testing genetic association using whole-genome sequencing data from 20 large pedigrees with 200 phenotype simulation replicates. What distinguishes the approaches explored is how the complexities of analyzing familial genetic data were handled. Here, we explore the methods that either harness inheritance patterns and transmission information or attempt to adjust for the correlation between family members in order to utilize computationally and conceptually simpler statistical testing procedures. Although directly comparing these two classes of approaches across contributions is difficult, we note that the two classes balance robustness to population stratification and computational complexity (the transmission-based approaches) with simplicity and increased power, assuming no population stratification or proper adjustment for it (decorrelation approaches).

*Correspondence to: David W. Fardo, University of Kentucky, 725 E Rose St, Lexington, KY 40536, USA. david.fardo@uky.edu.
†These authors contributed equally to this work.

Introduction

As DNA sequencing costs continue to decrease, large-scale whole-exome and whole-genome sequencing studies are becoming more feasible. The choice of study design and its implications are of paramount importance in these still expensive studies. In addition, the efficiency of the analytical methods used once the appropriate design has been chosen must be carefully considered.

We examine nine contributions to our Genetic Analysis Workshop 18 (GAW18) working group, which was tasked with studying family-based tests of association for rare variants using simulated data. The GAW18 data comprise whole-genome sequencing data from 20 large pedigrees with 200 phenotype simulation replicates of multiple outcomes observed at three time points [Almasy et al., 2014].We explore the implications of applying various statistical approaches and their resulting operating characteristics. Most of the methods explored by members of our working group fall into two broad categories: transmission-based methods that exploit properties of transmissions from parent to offspring, as in family-based association tests (FBATs); and decorrelation methods that attempt to remove within-family phenotype dependencies by means of some, often regression-based, adjustment. In addition to the presentation of various methods and comparisons within these classes, we discuss a two-stage strategy for sequencing studies of families.

Methods

Transmission-Based Approaches

Linkage Combined with Association—Li et al. [2014b] proposed a method to combine the results of linkage and association analyses. They calculated linkage LOD scores using a variance-components multipoint link-age analysis implemented in SOLAR [Almasy and Blangero, 1998]. They then calculated association test statistics using a multimarker FBAT with an empirical variance estimator to properly incorporate linkage within the extended pedigree [Xu et al., 2006]. Their combining schemes were adapted from the unweighted Liptak method [Liptak, 1958] and used genes as the testing unit. Finally, they calculated the average LOD score of a gene, converted it to a Z-score, and then combined that with an association test Z-score. The variance of the combined test statistic was empirically estimated using permutations [Pesarin, 2001].

Rare Variant Family-Based Association Testing—The FBAT for rare variants (FBAT-RV) is a gene-based burden test developed for rare variant studies [De et al., 2013]. The FBAT software (http://www.biostat.harvard.edu/fbat/default.html) uses both an unweighted rare variant test and a weighted version that uses as weights the inverse of the variance of the allele frequency estimated from the sample. Xu et al. [2014] applied both the

unweighted and weighted FBAT-RV and restricted the analysis to variants with a minor allele frequency (MAF) less than 0.01. Zhou et al. [2014] explored the unweighted FBAT-RV and tested each gene using single-nucleotide polymorphisms (SNPs) filtered on the basis of their predicted functions. They used Polyphen2 [Adzhubei et al., 2010], SnpEff (http://snpEff.sourceforge.net), and lymphoblastoid cell line eQTLs from the HapMap CEU samples to predict SNP function and to highlight SNPs associated with gene transcription [Montgomery et al., 2010]. Their gene-based tests used the combination of rare and common variants. Zhou et al. [2014] also explored several other FBATs, but they reported results only for the unweighted FBAT-RV and the multimarker test because of their superior performances. For the extended pedigree analysis where linkage is present, the empirical variance estimator [De et al., 2013; Rakovski et al., 2007; Xu et al., 2006] is needed to control the type I error.

Family-Based Sequence Kernel Association Test—Huang et al. [2014] used a recently proposed family-based sequence kernel association test (SKAT) [Ionita-Laza et al., 2013]. This approach is analogous to SKAT (i.e., a variance-components test) but within the transmission testing frame-work. Huang and colleagues altered the weighted linear kernel by treating the offspring genotypes as the random variables and by conditioning on the phenotypes. Entries in the genotype matrix were offset by the corresponding genotype expectation based on parental genotype transmissions. The family-based SKAT used the same score test statistic as SKAT with the amended kernel.

Decorrelation Approaches

Familial relationships induce a correlation structure between the outcomes of members of the same pedigree conditional on their genotypes at a particular locus. Naively analyzing these types of data without taking into account this correlation can result in inflated type I error and decreased power to detect true associations. One broad alternative approach to handling family data is to adjust for these correlations, essentially treating the relationships as nuisance. In several of the contributions to our working group in this category, including Fardo et al. [2014], Ding et al. [2014], and Li et al. [2014a], rather than using the provided pedigree structures, the investigators estimated kinship matrices from whole-genome SNP data to control for within-family residual phenotype correlations. Estimated kinships can also be used to control for population stratification [Kang et al., 2010; Svishcheva et al., 2012; Zhou and Stephens, 2012], at least for common variants, although estimating kinship for rare variants while adjusting for both population structure and known pedigree relationships simultaneously requires further study.

Mixed-Model Approaches—Mixed models are a natural choice to control for the correlations induced by family structure. Wang et al. [2014] applied a multilevel mixed model in a single-marker analysis of family-based longitudinal data with three levels: an individual level (longitudinal), a within-sibship level, and a between-sibship level. Fardo et al. [2014] performed a similar measured-genotype approach (MGA) analysis [Amin et al., 2007] in which the familial relationships were accounted for by using random effects defined through the kinship matrix, which was estimated using the whole-genome SNP data.

Ding et al. [2014] used a combination of a linear mixed model and a penalized linear regression, called a GRAM-MAR LASSO (genome-wide rapid association using mixed model and regression least absolute shrinkage and selection operator), to detect the genes harboring rare causal variants. In this approach they first regressed out the family structure by performing a polygenic analysis using the kinship matrix estimated from the marker data. This decorrelation step is similar to the method proposed by Svishcheva et al. [2012], with the residuals generated from the first step used as the phenotype in the second step and a penalized regression calculated from a mixture of gene-based group and pure LASSO penalties.

Generalized Least-Squares Approach—Li et al. [2014a] used a generalized least-squares (GLS) framework [Greene, 2012] to decorrelate the family structure. Briefly, they calculated a transformation matrix as the inverse of the decomposition of the kinship matrix estimated from the genetic data and then multiplied both the phenotype and genotype covariance matrices by this transformation matrix. The family-based data were decorrelated after the transformation, and any methods developed for independent data could then be applied. To perform the gene-based rare variant analysis, Li and colleagues [2014a] applied the SKAT-O approach [Lee et al., 2012; Wu et al., 2011] for rare variant detection to the decorrelated data, calling this approach GLS-SKAT.

Prioritizing Variants Based on Cosegregation

Yang and Thomas [2014] used a two-stage strategy to detect rare causal variants in family studies. Their strategy is based on the rationale that exploiting the cosegregation of variants with disease within families can help to distinguish causal variants from noncausal ones and that sequencing a subset of highly informative family members first can be cost-effective. For this first stage, Yang and Thomas proposed a novel score-based statistic (similar to the family-based SKAT tests of Schifano et al. [2012], Chen et al. [2013], and Schaid et al. [2013b]) for SNP prioritization that uses the available phenotype information for all the pedigree members and the genotype information for a subset of the pedigree. Then in the second stage of their analysis, they performed a single-SNP association test in the remaining sample for the top-ranked SNPs obtained in the first stage.

Comparison of Decorrelation and FBAT-Based Approaches

Three of the research groups that used decorrelation approaches made comparisons to the conventional FBAT or its variants. Both Wang et al. [2014] and Fardo et al. [2014] compared the mixed-model approaches to the single-SNP FBAT using quantitative traits. Fardo and colleagues also compared decorrelation to the Van Steen screening approach (FBAT-VS) [Van Steen et al., 2005]. Li et al. [2014a] compared the GLS-SKAT to the weighted FBAT-RV for gene-based inference. Yang and Thomas [2014] compared the performance of their proposed two-stage design to a one-stage $QTDT_M$ [Gauderman, 2003] in which complete phenotype and genotype data were used.

Application to the GAW18 Data

All investigators performed analyses with the knowledge of the true simulation model. A summary of the GAW18 data that each individual researcher used can be found in Table 1.

All transmission-based tests were gene-based and region-based and were focused on continuous phenotypes, with systolic blood pressure (SBP) and diastolic blood pressure (DBP) for power comparison and Q1 for evaluating type I error.

All the decorrelation studies used the simulated SBP or DBP data at baseline, with the exception of Wang et al. [2014], who used all three time points. Fardo et al. [2014] also used the simulated Q1 data, which had no genetic contribution, for type I error comparison. Three research groups [Fardo et al. 2014; Wang et al. 2014; Yang and Thomas 2014] performed single-SNP analyses, and two groups [Ding et al. 2014; Li et al. 2014a] performed gene-based inference, although the definition of gene regions varied.

Results

Transmission-Based Approaches

Linkage Combined with Association—Li et al. [2014b] adopted the method of Levy et al. [2000] to adjust for the effects of age, sex, and medication and then calculated mean SBP over the three time points for the analysis. They showed that chromosome 3 had LOD scores greater than 1.5 in three of the first 10 replicates for the null trait of Q1 compared with nine of 10 replicates for mean SBP. The FBAT was applied to the 8,047 genes with more than one nonsynonymous SNP. On average, over replicates 1–10, there were 49 genes of 8,047 with combined P-values for mean SBP less than 0.001. Only two causal genes, *MAP4* and *FLNB* on chromosome 3, were ever among the top 49 genes. For Q1, on average, there were 9.5 and 9.1 genes of 8,047 with FBAT P-values and combined P-values less than 0.001, corresponding to an empirical false-positive rate of 0.0012 and 0.0011, respectively. After combining linkage and FBAT P-values, the detection power for *MAP4* improved from 50% to 100%. For *FLNB*, which explains a much lower percentage of SBP variance (0.29%), FBAT had no detection power. Combined P-values improved the power to 40%. The type I error rate was well controlled in the combined P-values. When the correlations between linkage and association P-values were corrected, the ranks of *MAP4* and *FLNB* (out of 8,047) based on the combined P-values did not change.

Rare Variant Family-Based Association Testing—Both Xu et al. [2014] and Zhou et al. [2014] focused their studies on chromosome 3 and applied various FBAT methods to all 200 simulations. Xu and colleagues used both SBP alone and SBP adjusted for the effects of sex, age, and medication, although simple covariate adjustment for medication can be inappropriate when it serves as both a confounder and an intermediate variable on a causal pathway [Tobin et al., 2005]. Causal genes from chromosome 3 were used for the assessment of power, and the rate of type I errors was calculated using noncausal genes. *MAP4* gave the strongest signal when collapsing all rare variants (MAF< 0.01) within the true causal genes. Both the unweighted and weighted FBAT-RV methods detected the *MAP4* association signal with P-values less than 10^{-4}. In*SCAP*, another causal gene for SBP, only the weighted FBAT-RV detected associations at this level of significance, and neither the weighted or the unweighted method detected associations in other causal genes. Xu and colleagues found that the weighted FBAT-RV consistently outperformed the unweighted FBAT-RV. At a significance level of 10^{-3}, the unweighted FBAT-RV identified

less than 6% of significant associations in the 1,209 noncausal genes. Although the number of false positives decreased as the threshold became more stringent, the unweighted FBAT-RV still identified 2% of the associations among the noncausal genes as significant ($\alpha = 10^{-6}$), indicating inflated type I error.

Xu et al. [2014] also used a sliding window approach for association testing. The entire chromosome was divided into a series of disjoint 100 kb windows. Four such windows covered *MAP4*. After Bonferroni correction, the power (at $\alpha = 10^{-7}$) of the best performing window out of these four decreased from 0.745 to 0.455, which was better than the power for *MAP4* evaluated as a whole (0.005). However, the sliding window method resulted in additional type I error inflation, making power comparisons difficult, if not impossible.

Zhou et al. [2014] tested both SBP and DBP over the three time points adjusted for age, sex, age $\times$ sex, and medication at each exam. They also analyzed average residuals over three exams. The Q1 phenotype was used to evaluate type I error and was adjusted for age and sex only. Of the 894 variants in *MAP4*, Zhou and colleagues identified 28 SNPs that met the functional criteria of Polyphen scores above 0.5, splice, stop variants, and an eQTL cutoff of 3.4 (–log *P*-value from eQTL analysis). Of these, eight were true causal variants. More than half (57%) of the 28 SNPs were rare (MAF < 0.05). The same set of functional variants was used for the comparison of both family- and population-based designs. The type I errors of both the unweighted FBAT-RV and the multimarker FBAT were well controlled, and this analysis highlighted the association of *MAP4* across all simulation replicates. The highest power and the strongest association signals were identified using the multimarker FBAT with power around 0.9.

Family-Based SKAT—Huang et al. [2014] used the baseline SBP and DBP from all 200 simulations and the 31 causal genes on chromosome 3 to evaluate power for the family-based SKAT and the burden test (unweighted FBAT-RV). Q1 was used to evaluate type I error rate. Ninety-three trios were extracted for the gene-based association test. The empirical type I error rates were close to the nominal level of 0.05, with a range of 0.043–0.059. Using a nominal significance level of 0.05, both the family-based SKAT and the family-based burden test were able to detect *MAP4*. The proportions of causal SNPs in the analysis (i.e., 10%, 25%, 50%) did not substantially affect the power of either test, nor did the inclusion or not of rare variants. The power of both the family-based SKAT and the family-based burden test were comparable across most genes, but the performance of the two tests depended on the true simulation model; for example, when the causal variants' β coefficients were in different directions, the family-based SKAT had better power than the burden test (e.g., *MAP4* gene). The combination of both common and rare variants provided the best performance.

Decorrelation Approaches

Both Wang et al. [2014] and Fardo et al. [2014] reported nominal type I error rates for the mixed model based strategy. Li et al. [2014a] also observed a nominal type I error rate for the GLS-SKAT. Type I error rate was not available in the GRAMMAR LASSO method [Ding et al., 2014].

Several research groups compared the performance of different variants of the decorrelation strategy to that of FBAT. Wang et al. [2014] compared the power of the mixed model applied to the family-based longitudinal model to the single time point FBAT analysis. In general, higher power was observed in the mixed-model approach, although when applied to the rare variants only, both the mixed model and FBAT-based approaches had low power (less than 20% across all the causal variants at a significance level of 0.05). Fardo et al. [2014] found that the MGA approach also tended to have higher power compared to the conventional FBAT. The FBAT-VS approach [Van Steen et al., 2005], which is much less computationally intensive than the MGA method, had better performance in terms of power than the conventional FBAT but was less powerful than MGA. For genes that accounted for only a small proportion of the variance, power of all three approaches was low, although the MGA approach was still somewhat better. For *MAP4*, Li et al. [2014a] saw a clear advantage for the GLS-SKAT (with a power of 0.34 at an α level of 4.0×10^{-5}) over the weighted FBAT-RV (with a power of 0.08). Neither approach detected any other causal gene region on chromosome 3 after a stringent multiple-testing correction.

In the GRAMMAR LASSO approach, Ding et al. [2014] showed that the *MAP4* gene, which contributes more than 6% of the heritability of both DBP and SBP, could be consistently discovered. Detection probability increased as more weight was placed on the pure LASSO penalty. False discovery rates were often above 90%, although the gene-based false-positive rates were reasonably maintained (all less than 0.03%).

Performance of the Two-Stage Approach

In the two-stage approach of Yang and Thomas [2014], the first-stage mean score statistics showed a clear gradient across the negative, null, and positive variants, indicating the potential of the score statistics for prioritizing variants. Compared to the one-stage procedure (power of 13.4% for the top 100 ranked variants), lower power was observed for the two-stage design (power of 4.0% for the top 100 ranked variants).

Discussion

Members of our working group applied two major strategies to the family-based sequence data. The first category follows the traditional family-based association testing approach and utilizes transmission patterns for inference. In contrast, the second strategy class treats family structure as a nuisance and tries to remove the correlation resulting from family structure to ensure valid inference. Those two categories can be viewed as extensions of similar strategies in the common variants era, that is, the traditional transmission-based tests such as FBAT and the transmission disequilibrium test (TDT) [Abecasis et al., 2000; Allison, 1997; Cleves et al., 1997; Laird et al., 2000; Spielman et al., 1993] and the recently developed variance-components or mixed model based tests in which kinship is incorporated to control for (known or unknown) family structure [Kang et al., 2010; Svishcheva et al., 2012; Zhou and Stephens, 2012].

Because linkage and association metrics capture distinct and almost independent information from phenotype-genotype correlations, various efforts have been made to model linkage and association jointly [Biernacka and Cordell, 2007; Chen et al., 2005; Dupuis and

Van Eerdewegh, 2003; Goring and Terwilliger, 2000; Li et al., 2004, 2005; Roeder et al., 2006; Sun et al., 2002; Thornton and McPeek, 2010] to improve power for detecting causal variants. Joint modeling methods are usually computationally intensive; hence they cannot currently accommodate large pedigrees with dense markers. The method proposed by Li et al. [2014b] combines *P*-values from the linkage LOD score and *P*-values from the multimarker FBAT. The performance of the proposed methods, which can be viewed as an average of the linkage and association signals, depends largely on the strength of both linkage and association signals. Moderate signals in both linkage and association will generate a more significant combined *P*-value than a significant signal in one test but a null signal in the other.

The unweighted FBAT-RV (and the weighted FBAT-RV) is a recently developed FBAT designed specifically for rare variant studies [De et al., 2013]. Several research groups applied this method to the GAW18 simulated data [Huang et al., 2014; Xu et al., 2014; Zhou et al., 2014]. Xu and colleagues examined both the weighted and unweighted FBATRV methods, focusing on rare variants (MAF < 1%); they found that the weighted FBAT-RV had better power. Zhou and colleagues focused on both the unweighted and weighted FBAT-RV and found that the multimarker FBAT outperformed the unweighted FBAT-RV when testing *MAP4*; this could be due to some causal variants of *MAP4* within the variants for analysis being common. Other multimarker tests of FBAT did not perform well in the GAW 18 data set. Huang and colleagues compared the unweighted FBAT-RV (collapsing method) with the family-based SKAT from Ionita-Laza et al. [2013]. The family-based SKAT outperformed the unweighted FBAT-RV when the signs of effects from causal SNPs were in different directions.

The family-based SKAT and the various versions of FBAT, which are TDT-type tests conditioning on founder genotypes and comparing allelic transmissions, had low power except when common variants were included in the analysis. Although smaller sample size can partly explain the lower power, the simulated data of GAW18, which was generated by fixing the genotypes across the simulations, is also a concern when examining approaches that treat genotypes as the random variable, because the transmissions are exactly the same in all replicates. Zhou et al.'s [2014] population-based comparison suggests that in the absence of population substructure or with proper adjustment [Price et al., 2010], the population-based association tests using the whole families are more powerful. However, in the presence of population substructure that is not properly taken into account, the population-based association tests may lead to inflated type I errors, whereas the transmission-based methods are fully robust to population substructure [Laird and Lange, 2006].

Similar in spirit to a few recent studies [Kang et al., 2010; Svishcheva et al., 2012; Zhou and Stephens, 2012], a direct extension of the mixed model in the family-based data is attractive because it allows investigators to use most of the phenotype and genotype information for individuals within a family. Wang et al. [2014] extended the mixed model to accommodate both longitudinal observations and family structure, and the MGA approach of Fardo et al. [2014] used an estimated kinship matrix based random effect for polygenic components to account for family structure. Both of these approaches showed higher power than

transmission-based tests. Although both approaches are for single marker based analysis, direct extension to gene or gene region based inference is possible, because a variance component or random effect for local genetic structure can be incorporated into the mixed model, similar to that applied in the SKAT approach for unrelated individuals [Lee et al., 2012; Wu et al., 2011]. However, the model will be more cumbersome and the computational burden can be formidable for genome-wide data, so more efficient algorithms might be necessary for developments in this direction.

Another way of accounting for family structure is to decorrelate the family data before applying methods developed for unrelated subjects. Those methods include the GRAMMAR LASSO approach of Ding et al. [2014] and the GLS-SKAT of Li et al. [2014a]. Similar to the original GRAMMAR approach [Svishcheva et al., 2012], the GRAMMAR LASSO method uses the kinship matrix to calculate the residuals of the phenotype after controlling for covariates and family structure and then uses penalized regression to detect the gene regions harboring rare causal variants. In the GLS-SKAT, the family-based data were decorrelated by using a GLS transformation of phenotype, genotype, and covariates. An advantage of these approaches is that the data only need to be decorrelated once, and established methods for unrelated subjects can then be applied. This makes a seamless interface for many recent methodological developments, most of which require independence. A caveat of this type of approach is that decorrelation using the kinship matrix may not completely remove dependence resulting from family structure, because there could be residual correlations within families as a result of shared environmental factors. In addition, these approaches may actually remove the genotype and phenotype correlations that are of primary interest. More exploration to test this is needed.

Overall, the comparison between accounting for family structure and transmission-based inference indicates a higher power when mixed-model and decorrelation approaches are used to adjust for family structure. This is expected because sample sizes are larger when founders can be incorporated directly, as in the adjustment approaches. However, as in the population studies, adjustment for population stratification is still needed for nontransmission-based approaches, which can be especially challenging with complex pedigree structures and, in particular, rare variants. Nevertheless, a direct comparison across different contributions is impossible with the differences in the data used, the manner of handling multiple testing, the significance thresholds used, the definition of genes or gene regions, and the comparability of the type I error rate for some approaches. These difficulties are present even when differences appear minor, and this is evidenced by the two contributions that found contradicting type I error rates with similar approaches.

Yang and Thomas [2014] proposed a two-stage approach in which a subset of subjects was selected for SNP prioritization and the remaining "maximally unrelated" subjects were used for the second-stage validation. Compared to a one-stage approach ($QTDT_M$), Yang and Thomas observed lower power in the second stage. This is not at all surprising, because in the proposed two-stage approach only a small subset of the data was used. However, from a design point of view, with the success of similar two-stage ideas in genome-wide association studies [Schaid et al., 2013a; Skol et al., 2006], this approach could have potential, given the current cost of whole-genome sequencing.

From the analysis point of view this two-stage idea can also be useful. Several two-step analysis strategies for the same data have been proposed[Millstein et al., 2006;Murcray et al., 2009, 2011; Van Steen et al., 2005; Wason and Dudbridge, 2012; Zheng et al., 2007]. The advantage of the two-stage analysis is that, with an independent filtering stage, which is usually based on a certain hypothesis, the signal-to-noise ratio can be increased and/or the multiple testing penalty can be decreased [Wason and Dudbridge, 2012]. Built on a similar idea of increasing the signal-to-noise ratio, Zhou et al. [2014] filtered the SNPs and gene regions based on predicted biological functions of SNPs. They showed that the filtering based on external functional and evolutionary data can prioritize the causal variants and improve success for later analysis, although one might expect that the performance also depends strongly on the relevance and accuracy of prior information. These filtering-based strategies can be useful because the multiple-testing penalty can be tremendous for whole-genome data, and previous investigations have indicated that gene-based inference of rare variants can be sensitive to the signal-to-noise ratio in the region. The GAW18 simulated data were based on eQTL analysis, so using this information to prioritize variants would likely be the best scenario for this application.

Conceptually, the two classes of approaches examined by members of our working group vary greatly (e.g., the random variables are different across approaches and the association metrics are on different scales), and the implications of their differences are illuminated somewhat by comparisons among our working group's results. However, broader comparisons will require more in-depth exploration using varied data structures (e.g., samples subject to population stratification). In addition, most contributions examined methods within only one of these classes, making direct comparisons between classes difficult. This is particularly challenging with different numbers of tests, distinct multiple-testing corrections, and/or significance thresholds used and sometimes inconsistent type I error rates across the contributions. In general, the two classes balance the robustness to population stratification and computational complexity of the transmission-based approaches with the simplicity and increased power (assuming no stratification or proper adjustment) of the decorrelation approaches.

Acknowledgments

References

Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. Am J Hum Genet. 2000; 66:279–292. [PubMed: 10631157]

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]

Allison DB. Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet. 1997; 60:676–690. [PubMed: 9042929]

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, Wood AR, Fuchsberger C, Almeida MA, Kent JW Jr, Fowler S, Blackwell TW, et al. 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]

Amin N, van Duijn CM, Aulchenko YS. A genomic background based method for association analysis in related individuals. PLoS One. 2007; 2:e1274. [PubMed: 18060068]

Biernacka JM, Cordell HJ. Exploring causality via identification of SNPs or haplotypes responsible for a linkage signal. Genet Epidemiol. 2007; 31:727–740. [PubMed: 17508343]

Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013; 37(2):196–204. [PubMed: 23280576]

Chen MH, Van Eerdewegh P, Dupuis J. Identification of polymorphisms explaining a linkage signal: application to the GAW14 simulated data. BMC Genet. 2005; 6(Suppl 1):S88. [PubMed: 16451703]

Cleves MA, Olson JM, Jacobs KB. Exact transmission-disequilibrium tests with multiallelic markers. Genet Epidemiol. 1997; 14:337–347. [PubMed: 9271708]

De G, Yip WK, Ionita-Laza I, Laird N. Rare variant analysis for family-based design. PLoS One. 2013; 8:e48495. [PubMed: 23341868]

Ding X, Su S, Nandakumar K, Wang X, Fardo DW. A two-step penalized regression method for family-based next-generation sequencing association studies. BMC Proc. 2014; 8(Suppl 1):S25. [PubMed: 25519315]

Dupuis J, Van Eerdewegh P. Identification of polymorphisms that explain a linkage peak: conditioning on parental genotypes. Genet Epidemiol. 2003; 25:247.

Fardo DW, Zhang X, Ding L, He H, Kurowski B, Alexander E, Baye T, Pilipenko V, Kottyan L, Nandakumar K, Martin LJ. On family-based genome-wide association studies with large pedigrees: observations and recommendations. BMC Proc. 2014; 8(Suppl 1):S26. [PubMed: 25519377]

Gauderman WJ. Candidate gene association analysis for a quantitative trait, using parent-offspring trios. Genet Epidemiol. 2003; 25:327–338. [PubMed: 14639702]

Goring HHH, Terwilliger JD. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. Am J Hum Genet. 2000; 66:1310–1327. [PubMed: 10731466]

Greene, WH. Econometric Analysis. Boston: Prentice Hall; 2012.

Huang J, Chen Y, Swartz MD, Ionita-Laza I. Comparing the power of family-based association tests for sequence data with applications in the GAW18 simulated data. BMC Proc. 2014; 8(Suppl 1):S27. [PubMed: 25519316]

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Family-based association tests for sequence data, and comparisons with population-based association tests. Eur J Hum Genet. 2013; 21:1158–1162. [PubMed: 23386037]

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]

Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet. 2006; 7:385–394. [PubMed: 16619052]

Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. Genet Epidemiol. 2000; 19(Suppl 1):S36–S42. [PubMed: 11055368]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012; 91:224–237. [PubMed: 22863193]

Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. Evidence for a gene influencing blood pressure on chromosome 17: genome scan linkage results

for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. Hypertension. 2000; 36:477–483. [PubMed: 11040222]

Li C, Scott LJ, Boehnke M. Assessing whether an allele can account in part for a linkage signal: the genotype-IBD sharing test (GIST). Am J Hum Genet. 2004; 74:418–431. [PubMed: 14872409]

Li D, Rotter J, Guo X. A generalized least-squares framework for rare variant analysis in family data. BMC Proc. 2014a; 8(Suppl 1):S28. [PubMed: 25519378]

Li M, Boehnke M, Abecasis GR. Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am J Hum Genet. 2005; 76:934–949. [PubMed: 15877278]

Li Y, Foo JN, Liany H, Low HQ, Liu J. Combined linkage and family-based association analysis improves candidate gene detection in Genetic Analysis Workshop 18 simulation data. BMC Proc. 2014b; 8(Suppl 1):S29. [PubMed: 25519379]

Liptak T. On the combination of independent tests. Magyar Tud Akad Mat Kutato Int Kozl. 1958; 3:171–196.

Millstein J, Conti DV, Gilliland FD, Gauderman WJ. A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet. 2006; 78:15–27. [PubMed: 16385446]

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]

Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. Am J Epidemiol. 2009; 169:219–226. [PubMed: 19022827]

Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genet Epidemiol. 2011; 35:201–210. [PubMed: 21308767]

Pesarin, F. Multivariate Permutation Tests: With Applications in Biostatistics. New York: Wiley; 2001.

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

Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM. A new multimarker test for family-based association studies. Genet Epidemiol. 2007; 31:9–17. [PubMed: 17086514]

Roeder K, Bacanu S, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet. 2006; 78:243–252. [PubMed: 16400608]

Schaid DJ, Jenkins GD, Ingle JN, Weinshilboum RM. Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies. Genet Epidemiol. 2013a; 37:229–238. [PubMed: 23348637]

Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol. 2013b; 37(5):409–418. [PubMed: 23650101]

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]

Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006; 38:209–213. [PubMed: 16415888]

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]

Sun L, Cox NJ, McPeek MS. A statistical method for identification of polymorphisms that explain a linkage result. Am J Hum Genet. 2002; 70:399–411. [PubMed: 11791210]

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. Nat Genet. 2012; 44:1166–1170. [PubMed: 22983301]

Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet. 2010; 86:172–184. [PubMed: 20137780]

Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Stat Med. 2005; 24:2911–2935. [PubMed: 16152135]

Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, DeMeo DL, Murphy A, Su J, Datta S, Rosenow C, et al. Genomic screening and replication using the same data set in family-based association testing. Nat Genet. 2005; 37:683–691. [PubMed: 15937480]

Wang J, Yu R, Shete S. Comparison of multilevel modeling and the family-based association test for identifying genetic variants associated with systolic and diastolic blood pressure using Genetic Analysis Workshop 18 simulated data. BMC Proc. 2014; 8(Suppl 1):S30. [PubMed: 25519380]

Wason JMS, Dudbridge F. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. Am J Hum Genet. 2012; 90:760–773. [PubMed: 22560088]

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]

Xu M, Wang HZ, Guo W, Qin H, Shugart YY. Family-based tests applied to extended pedigrees identify rare variants related to hypertension. BMC Proc. 2014; 8(Suppl 1):S31. [PubMed: 25519318]

Xu X, Rakovski C, Laird N. An efficient family-based association test using multiple markers. Genet Epidemiol. 2006; 30:620–626. [PubMed: 16868964]

Yang Z, Thomas DC. Two-stage family-based designs for sequencing studies. BMC Proc. 2014; 8(Suppl 1):S32. [PubMed: 25519319]

Zheng G, Song K, Elston RC. Adaptive two-stage analysis of genetic association in case-control designs. Hum Hered. 2007; 63:175–186. [PubMed: 17310127]

Zhou JJ, Yip W, Cho MH, Qiao D, McDonald MN, Laird NM. GAW18: a comparative analysis of family- and population-based association tests using whole genome sequence data. BMC Proc. 2014; 8(Suppl 1):S33. [PubMed: 25519381]

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

Table 1

Summary of contributions to the working group on family-based tests of association for rare variants using simulated data

Contributor	Methods	Phenotypes	Genotypes	Highlighted conclusions/results
Li et al. [2014b]	Linkage and multimarker FBAT	SBP (first 10 simulations); Q1	Genes on all chromosomes; focused on *MAP4* and *FLNB*	Type I error rate was well controlled in the combined P-values; power for *MAP4* improved from 50% to 100%; power for *FLNB* improved from nothing to 40%.
Xu et al. [2014]	Unweighted and weighted FBAT-RV	SBP baseline (200 simulations)	Genes on chromosome 3; sliding window of 100 kb regions over chromosome 3	Weighted FBAT-RV outperformed unweighted FBAT-RV consistently; both methods have inflated type I error.
Zhou et al. [2014]	Unweighted FBAT-RV; multimarker FBAT; SKAT; GCTA	SBP,DBP at all three time points (200 simulations); Q1	Genes on chromosome 3; focused on *MAP4*	Performances were evaluated at a nominal significant level of 0.05; type I error was controlled; using filtered SNPs of *MAP4* (common and rare), multimarker FBAT has the best power (over 90%); using 142 unrelated individuals, SKAT and adapted GCTA have better than 80% power to detect *MAP4*.
Huang et al. [2014]	Family-based SKAT	SBP baseline (200 simulations); Q1	31 causal genes on chromosome 3	Performances were evaluated at a nominal significant level of 0.05; both the family-based SKAT and the family-based burden test were able to detect *MAP4*, and performances were comparable for other genes.
Ding et al. [2014]	GRAMMAR LASSO	SBP and DBP (first time point, 200 simulations)	Genes on all chromosomes	*MAP4* could be consistently discovered; detection probability increased as more weight was placed toward the pure LASSO penalty; false discovery rates were often greater than 90%, although the gene-based false-positive rates were reasonably maintained.
Wang et al. [2014]	Mixed-model FBAT	SBP and DBP (all three time points for mixed model; first time point for FBAT; 200 simulations)	All the causal variants and randomly selected noncausal variants	Power of the mixed model tended to be higher than the single time point FBAT.
Fardo et al. [2014]	MGA, FBAT, FBAT-VS	DBP and Q1 (first time point, 200 simulations)	Causal variants (DBP); a subset of uncorrelated SNPs on chromosome 3	MGA tended to have higher power than the conventional FBAT; FBAT-VS approach, which is much less computationally intensive than the MGA, had better performance in terms of power compared to the conventional FBAT but was less powerful than MGA.
Li et al. [2014a]	GLS-SKAT; weighted FBAT-RV	SBP (first time point, 100 simulations)	All the genes on chromosome 3	GLS-SKAT is more powerful than weighted FBAT-RV.
Yang and Thomas [2014]	Two-stage	SBP and DBP (first time point, five simulations)	*MAP4* gene region	Two-stage approach is less powerful than the one-stage approach, which uses more data.

DBP, diastolic blood pressure; FBAT, family-based association test; FBAT-RV, FBAT for rare variants; FBAT-VS, Van Steen's FBAT [Van Steen et al., 2005]; GCTA, genome-wide complex trait analysis; GLS-SKAT, generalized least-squares SKAT; GRAMMAR LASSO, genome-wide rapid association using mixed model and regression least absolute shrinkage and selection operator; MGA, measured-genotype approach; SBP, systolic blood pressure; SKAT, sequence kernel association test.