Published in final edited form as:

Genet Epidemiol. 2014 September ; 38(0 1): S13–S20. doi:10.1002/gepi.21820.

Methods for Collapsing Multiple Rare Variants in Whole-Genome Sequence Data

Yun Ju Sung¹, Keegan D. Korthauer², Michael D. Swartz³, and Corinne D. Engelman^{4,*}

¹Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri

²Department of Statistics, University of Wisconsin, Madison, Wisconsin

³Division of Biostatistics, University of Texas School of Public Health, Houston, Texas

⁴Department of Population Health Sciences, University of Wisconsin, Madison, Wisconsin

Abstract

Genetic Analysis Workshop 18 provided whole-genome sequence data in a pedigree-based sample and longitudinal phenotype data for hypertension and related traits, presenting an excellent opportunity for evaluating analysis choices. We summarize the nine contributions to the working group on collapsing methods, which evaluated various approaches for the analysis of multiple rare variants. One contributor defined a variant prioritization scheme, whereas the remaining eight contributors evaluated statistical methods for association analysis. Six contributors chose the gene as the genomic region for collapsing variants, whereas three contributors chose nonoverlapping sliding windows across the entire genome. Statistical methods spanned most of the published methods, including well-established burden tests, variance-components-type tests, and recently developed hybrid approaches. Lesser known methods, such as functional principal components analysis, higher criticism, and homozygosity association, and some newly introduced methods were also used. We found that performance of these methods depended on the characteristics of the genomic region, such as effect size and direction of variants under consideration. Except for *MAP4* and *FLT3*, the performance of all statistical methods to identify rare causal variants was disappointingly poor, providing overall power almost identical to the type I error. This poor performance may have arisen from a combination of (1) small sample size, (2) small effects of most of the causal variants, explaining a small fraction of variance, (3) use of incomplete annotation information, and (4) linkage disequilibrium between causal variants in a gene and noncausal variants in nearby genes. Our findings demonstrate challenges in analyzing rare variants identified from sequence data.

Keywords

Genetic Analysis Workshop 18; rare variants; whole-genome sequence; burden tests; nonburden tests

*Correspondence to: Corinne D. Engelman, Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, 610 Walnut Street, 1007A WARF, Madison, WI 53726-2397. Tel.: (608) 265-5491, Fax: (608) 263-2820, cengelma@wisc.edu.

Introduction

Although genome-wide association studies (GWAS) have successfully identified thousands of new susceptibility loci for complex traits (<http://www.genome.gov/gwastudies>) [Hindorff et al., 2009], these identified loci generally have subtle effects that explain only a small fraction of the heritability of most complex traits [Manolio et al., 2009]. This missing heritability might be partly explained by rare and low-frequency variants, with minor allele frequencies (MAFs) less than 5% [McCarthy and Hirschhorn, 2008]. These variants could have large effect sizes that, when combined, account for a substantial proportion of the missing heritability. GWAS were designed to examine associations with common variants only, as the single-nucleotide polymorphisms (SNPs) included on standard GWAS chips were chosen because they were common variants in the HapMap data. Next-generation sequencing technologies have enabled detection of rare variants, and the increased availability of sequence data has been followed by a surge of statistical methods to analyze rare variants.

The organizers of Genetic Analysis Workshop 18 (GAW18) provided whole-genome sequence data in a pedigree-based sample and longitudinal phenotype data for hypertension and related traits [Almasy et al., 2014]. The approximately 8 million variants available in the GAW18 sequence data exposed challenges but also provided an excellent opportunity for evaluating existing and newly proposed methods to identify associations with rare variants. For sequenced individuals whole-genome sequence data (stored in the variant calling files) and cleaned sequence calls for odd-numbered chromosomes were provided. Also, for nonsequenced family members imputed genotypes based on available GWAS data were provided. Real phenotype data included sex, age, year of examination, systolic and diastolic blood pressure (SBP and DBP, respectively), use of antihypertensive medications, and tobacco smoking status at up to four time points. In addition to real phenotypes, 200 replicates of longitudinal phenotypes were simulated using real genotype data while mimicking the observed distribution of the real phenotypes and maintaining the pedigree structure of the real data. A large number of rare and common variants were selected as causal variants for the simulated SBP and DBP. These causal variants had effects that varied in magnitude and direction (i.e., the simulated data included both risk and protective variants, even within the same gene, so that certain causal variants increased blood pressure and others decreased it).

Here, we summarize the nine contributions to the working group on collapsing methods that evaluated various methods for collapsing and analyzing multiple rare variants using the GAW18 sequence data (see Table 1 for an overview). Because the power to detect individual rare variants is limited by their low frequency, the development of novel statistical approaches for the analysis of rare variants has been an active area of research. Instead of identifying the individual effects of rare variants, these approaches assess the combined effect of multiple rare variants in a genomic region. We first describe various collapsing or binning strategies and then the statistical methods for the analysis of multiple variants.

Methods for Collapsing Multiple Rare Variants

Gene-Based Bins, Windows of Fixed Length, Windows of a Fixed Number of Variants

A key component of collapsing methods is defining the genomic region within which variants are jointly considered for association analysis. In principle, collapsing methods can be applied to any arbitrary collection of variants, so it is important to consider the implications of the elected strategy. The most prevalent strategy among our working group contributions was to collapse over genes [Dering et al., 2014; Derkach et al., 2014; Mallaney and Sung, 2014; Sung et al., 2014; Swartz et al., 2014; Xuan et al., 2014]. In these contributions associations between a given gene and the phenotype were assessed using all variants located within that gene. The definition of a gene-based genomic region could also cover a specified number of flanking base pairs to include regulatory regions adjacent to the gene. This approach leads to a natural functional interpretation of each region, especially for well-annotated genes. However, any variants located outside the annotated genes and nearby regulatory regions are ignored. Moreover, the annotation source used for collapsing can result in excluding different sets of variants. For example, Mallaney and Sung [2014] and Sung et al. [2014] mapped variants to genes based on the annotation file v2.20101123.autosomes.SNP_annotation.tgz (available at <http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2012-02-14.html>), which was constructed using the data from the 1000 Genomes Project. Unfortunately, about half the GAW18 simulated causal variants were not in this annotation file, either because they were novel in the GAW18 data or because they were classified as lying outside genes. Thus gene-based windows implicitly assume that all causal variants are located within or nearby annotated genes and can depend on the source of annotation information.

In addition to gene-based bins, several research groups used a competing strategy in which nonoverlapping sliding windows were constructed across the entire genome. The number of variants contained in each window could either be fixed [Agne et al., 2014; Yang and Li, 2014] or depend on the number of variants observed within a set number of base pairs [Xuan et al., 2014; Yang and Li, 2014]. The choice of window sizing was somewhat arbitrary. Agne et al. [2014] considered windows containing 100 variants. Yang and Li [2014] used a window size that contained 5% of variants in a chromosome and therefore contained a larger number of variants for bigger chromosomes. Xuan et al. [2014] explored fixed-length window sizes ranging from 10 kb to 500 kb in addition to gene-based bins. It was not clear from the contributions how the number of variants per window influenced power and type I error. This would likely depend on the statistical method used and the proportion of causal variants in the window. In contrast to gene-based bins, both the fixed number of variants and the fixed-length window strategies are able to incorporate the entire genome sequence, although interpretation of significant regions between genes may be more challenging.

Functional Annotations

An extension to the approach of collapsing all variants in a gene is to consider only variants that are expected to have a functional effect. Accordingly, Nalpathamkalam et al. [2014] chose to focus on defining a simple yet robust variant prioritization scheme using publicly available bioinformatics tools and resources. Nalpathamkalam and colleagues considered all

high-quality variants in the genotype files and divided these into coding (0.9%) and noncoding (99.1%) variants. Coding variants were further divided into protein changing (54.7% of the coding variants) or not, and the protein-changing variants were divided into protein damaging (19.5% of the protein-changing variants) or not. Noncoding variants were prioritized on the basis of cross-species conservation. Just under 1% of noncoding variants met the criteria for “medium conservation” and 39% of those met the criteria for “high conservation.” To bin the variants within each gene for the GAW18 data, Derkach et al. [2014] used the functional classification determined by Nalpathamkalam et al. [2014]. They grouped variants that belonged to the same gene and were of the same annotation type (coding, protein changing, or protein damaging) and found stronger evidence for association with the phenotype for groups of protein-changing variants, with the strongest evidence for groups that were protein damaging.

Statistical Methods for the Analysis of Rare Variants

Figure 1 presents the previously published methods for the analysis of rare variants that were used by members of our working group. We organized these methods into four categories: (1) early-developed burden tests using linear statistics, (2) well-established variance-components-type tests using quadratic statistics, (3) recently developed hybrid methods combining burden and variance-components-type tests, and (4) methods using other dimension-reduction techniques. We first provide a brief summary of the published methods that have received attention for the analysis of rare variants. Next, we provide more details for the novel approaches proposed by our working group.

Early-Developed Burden Tests

Burden tests collapse rare variants in a genomic region (typically a gene) into a single burden statistic and regress the phenotype on the burden statistic to test for the combined effects of all rare variants in the region or gene. For case-control data the cohort allelic sum test (CAST) [Morgenthaler and Thilly, 2007] collapses the genotypes across multiple rare variants (with a MAF threshold set at 1% or 5%) into one indicator variable for the presence or absence of a minor allele at any rare variant within each individual. The combined multivariate and collapsing (CMC) method [Li and Leal, 2008] includes common variants to improve its performance for both rare and common variants. Instead of using a fixed MAF threshold (1%) to define rare variants, as in the CMC method, the group-wise weighted-sum (w-Sum) test [Madsen and Browning, 2009] combines multiple rare variants that are inversely weighted by their frequency in control subjects, thereby overweighting rare alleles.

Using a regression framework, Morris and Zeggini [2010] extended these approaches from case-control data to population-based studies with binary and continuous phenotypes. The presence test uses the presence or absence of a minor allele at any rare variant within an individual as the predictor, which is analogous to the CAST. The simple sum test (SST) uses the total count of minor alleles at these variants as the predictor; this is equivalent to the proportion test, which is the SST divided by the total number of variants under consideration. Also using the regression framework, Price et al. [2010] proposed the variable-threshold (VT) approach, which selects variants within each genomic region by choosing the MAF threshold with maximum likelihood across MAF values. Because of this,

each genomic region can have its own MAF threshold, which determines which variants should be collapsed within the region.

Sung et al. [2014] modified the SST of Morris and Zeggini [2010] to account for correlation across related individuals within families and applied it to the GAW18 data. At *MAP4* and *FLT3*, both of which include a common variant with a large effect, they found that collapsing all variants performed much better than collapsing only the low-frequency variants. For several variants in *MAP4* single-marker analysis also provided high power. In contrast, collapsing only the low-frequency variants performed much better when rare variants had larger effects than common variants did. Sung and colleagues also found that collapsing multiple variants did worse than single-marker analysis for several genes that contained causal SNPs with both positive and negative effects.

These burden tests have received much attention [Asimit and Zeggini, 2010; Bansal et al., 2010] and have also been used for the analysis of the mini-whole-exome sequence data provided by GAW17 [Dering et al., 2011]. However, these burden tests implicitly assume that all variants within a gene influence the trait in the same direction (i.e., deleterious or beneficial) and with the same magnitude of effect (homogeneous effects) after incorporating known weights. Therefore their power is reduced if their effects are heterogeneous (e.g., certain rare variants confer increased risk, whereas others are protective) [Cohen et al., 2006].

Well-Established Variance-Components-Type Tests

Variance-components-type tests use quadratic statistics instead of linear burden statistics and, as such, are less sensitive to heterogeneity of effects [Neale et al., 2011; Wu et al., 2011]. Because any rare variant can have no effect, increase risk, or be protective, the C-alpha test uses the distribution of genetic variation observed in case and control subjects to detect the presence of this mixture of effects across a set of rare variants [Neale et al., 2011]. By testing the variance rather than the mean, the C-alpha test maintains consistent power when the target set contains both risk and protective variants. The sequence kernel association test (SKAT) is a score-based variance-components test for the association of variants in a region with continuous and discrete traits that easily adjusts for covariates [Wu et al., 2011]. To efficiently test for the joint effects of the rare variants in a region on the phenotype, SKAT aggregates individual variant-score test statistics with MAF-based weights, assuming a distribution for the genetic effects of the variants, whose variances depend on flexible weights. SKAT with equal weights has been shown to be equivalent to the C-alpha test [Wu et al., 2011].

For the GAW18 data Mallaney and Sung [2014] used three weighting schemes (equal weighting, Madsen-Browning weighting, and SKAT default weighting) and explored how the performance of SKAT depended on the weighting scheme. They found that the SKAT default weighting scheme (which heavily down-weights common variants) performed better for the genes in which causal SNPs are mostly rare. In contrast, the equal-weighting scheme performed the best for *MAP4* and *FLT3*, both of which included a common variant with a large effect.

Recently Developed Hybrid Methods

Hybrid methods have recently been developed that combine burden and variance-components-type tests because they have been shown to be complementary. Burden tests are more powerful for a region in which most rare variants are truly causal and in which their effects are in the same direction and comparable, whereas variance-components-type tests are more powerful if the rare variant effects are heterogeneous [Basu and Pan, 2011; Lee et al., 2012]. Lee et al. [2012] proposed a unified test that combines a burden test and SKAT with an optimally weighted average (SKAT-O). To maximize power, the SKAT-O uses the data to adaptively select the best linear combination of the burden test and SKAT. Derkach et al. [2013] also proposed robust methods involving minimum p -values and Fisher's methods to combine evidence of association from two complementary burden and variance-components-type tests. The minimum p -value approach uses $W_{\min} = \min(p_L, p_Q)$, whereas Fisher's method uses $W_F = -\log(p_L) - \log(p_Q)$, where p_L is the p -value of a linear burden test and p_Q is the p -value from a variance-components-type test.

For the GAW18 data Derkach et al. [2014] used all three hybrid methods and evaluated their performance along with other well-established burden and nonburden methods (CAST, w-SUM, C-alpha, and Hotelling statistic). They found that all seven tests that they applied provided good power for *MAP4* in the analysis of simulated DBP and SBP phenotypes. They also found that performance of the hybrid methods was indeed robust, providing intermediate power between the burden and nonburden tests, and was often comparable to the method with the maximum power at each gene. Furthermore, they observed that Fisher's method was the best or the second best option in many cases, providing higher power than the other two robust statistics.

Replication-Based Weighted-Sum Statistic

To quantify enrichment in risk variants and protective variants, Ionita-Laza et al. [2011] proposed the replication-based weighted-sum statistic (RBS) using two one-sided statistics. These two statistics are constructed as a weighted sum of rare variants, where the rare variants are classified as either at risk or protective according to their frequencies in the case subjects and the control subjects. The overall statistic is the sum of the at-risk and protective variant statistics. However, this implicitly assumes that at-risk and protective variants play an equal role in disease etiology.

Swartz et al. [2014] developed a modified RBS that uses a data-based weight for combining the at-risk variant statistic and the protective variant statistic. For the GAW18 data Swartz et al. [2014] used their modified RBS and compared its performance to other published methods (VT, C-alpha, SKAT, and RBS). For all causal genes on chromosome 3 they found that all six methods that they applied performed similarly (within 5%) and that the modified RBS performed competitively and similarly to the other replication-based statistics. They also found that performance of the VT method varied more than that of the other methods, providing higher power for some genes and lower power for other genes.

Functional Principal Components Analysis

Functional principal components analysis (FPCA) has recently been applied to association studies for sequence data. Luo et al. [2011] used the genome continuum model [Bickeboller and Thompson, 1996], which treats a genomic location as a real continuous value. By treating the genotype data of each individual along the genome as a genetic variant function, Luo et al. [2011] applied FPCA to transform a genetic variant function into a linear combination of principal component functions to capture the genetic variations [Ramsay and Silverman, 2005]. Because FPCA is applicable to sparse and irregularly spaced genomic variant data, FPCA-based statistics have higher power to detect association of rare variants and better ability to filter sequence errors than other methods [Luo et al., 2011].

For the GAW18 data Dering et al. [2014] compared FPCA to the CMC method by evaluating type I error and power using the simulated hypertension phenotype. They found that CMC and FPCA failed to detect most of the truly associated genes. CMC identified one gene (*KRT23*) with a power of 0.66 at the cost of a very high false-positive rate (75%), whereas FPCA roughly maintained the type I error level but at the cost of low power.

Higher Criticism

Higher criticism (HC), or second-level significance testing, was originally suggested by Tukey [1976] for the case in which there are many independent tests of significance and one is interested in rejecting the joint null hypothesis. The HC statistic compares the fraction of observed significances at a given level to the expected fraction under the joint null hypothesis. Donoho and Jin [2004] considered a generalized HC statistic and showed that the resulting HC statistic is effective for detecting weak and sparse effects.

Because whole-genome sequence data share similar characteristics, in that the effects of individual rare variants are sparse and weak, Xuan et al. [2014] applied this HC approach to detect multiple rare variants in the GAW18 sequence data by using *p*-values obtained from an association analysis of individual variants in a genomic region. They compared the HC approach with SKAT and the minimum *p*-value approach. To accommodate the different variant numbers and correlation structures within different genome segment windows, they used permutation for all three methods. Xuan and colleagues found that the minimum *p*-value approach was more powerful for genomic regions that contained a true variant with strong effect, whereas SKAT was more powerful for regions that contained a large number of variants with weak effects. They also found that the HC approach was more powerful for regions that contained a smaller number of variants with weak effects.

Homozygosity Intensity Across Multiple Variants

Regions of extended homozygosity across large numbers of consecutive SNPs are a function of linkage disequilibrium within a chromosomal region, and regions of extensive long-range linkage disequilibrium can be used to identify functional significance. Homozygosity haplotyping has successfully identified the shared segments for Marfan syndrome in a Japanese population [Miyazawa et al., 2007]. Moreover, runs of homozygosity were found in European populations [McQuillan et al., 2008] and were used to identify genes for schizophrenia [Lencz et al., 2007] and autism spectrum disorder [Casey et al., 2012]. The

Loss-of-Heterozygosity Analysis Suite (LOHAS) software [Yang et al., 2011] was developed to detect loss of heterozygosity in cancer research and to identify long contiguous stretches of homozygosity in population genetic studies using SNP genotype data.

For the GAW18 data Yang and Li [2013] extended LOHAS to handle large numbers of SNPs and rare variants in the whole-genome sequence data. LOHAS provides a two-step procedure for homozygosity association studies. First, LOHAS constructs sliding windows on a chromosome using a nearest neighbor method and estimates homozygosity intensity in each window for each individual using a local polynomial model. Second, the software performs a linear rank association test to identify runs of homozygosity with differential homozygosity intensities between the study groups. For the analysis of family data Yang and Li [2013] used generalized estimating equations (GEEs) to account for correlation among related individuals within families. They identified five and three homozygosity regions associated with the real DBP and hypertension phenotypes, respectively. They further showed through their own simulation studies that homozygosity association tests controlled type I error and had promising power.

Results

In this section we present results from the eight research groups who evaluated statistical methods for association analysis, excluding Nalpathamkalam et al. [2014], who defined a variant prioritization scheme. As shown in Table 1, Yang and Li [2014] used real phenotypes only, Derkach et al [2014] used both real and simulated phenotypes, and the remaining six contributors used simulated phenotypes only. Genomic data in these eight contributions varied. Mallaney and Sung [2014] and Yang and Li [2014] used all odd-numbered chromosomes that were provided; Dering et al. [2014] excluded chromosome 5; Xuan et al. [2014] used chromosomes 1, 3, and 5; Agne et al. [2014], Derkach et al. [2014], and Sung et al. [2014] used chromosome 3 only; and Swartz et al. [2014] further restricted their analysis to the causal genes on chromosome 3.

All these statistical methods for multiple rare variants were originally developed for unrelated individuals. Six research groups applied these methods to a smaller sample of unrelated individuals in the GAW18 data; Agne et al. [2014], Dering et al. [2014], Mallaney and Sung [2014], and Xuan et al. [2014] used 142 unrelated individuals with simulated phenotype data, and Derkach et al. [2014] and Swartz et al. [2014] used 103 unrelated individuals with real phenotype data. Sung et al. [2014] used all related individuals by extending the SST to account for correlation across 847 related individuals within families. Yang and Li [2014] used both data sets (a subset of unrelated individuals and all related individuals using the GEE approach, as described in the Methods for Collapsing Multiple Rare Variants section).

Correct Type I Error

Several contributors observed that the type I error was controlled at the specified level of 0.05. For the analysis of both continuous and dichotomized values of the simulated Q1 phenotype, for which no SNPs were simulated to be causal, Derkach et al. [2014] observed the correct type I error for all seven methods that they compared, as shown in Table 1. For

the analysis of the simulated hypertension phenotype, Swartz et al. [2014] found that the type I error for all six statistical methods that they compared was controlled at the 0.05 level. Dering et al. [2014] observed that the FPCA also provided the correct type I error for the analysis of the simulated hypertension phenotype. Similarly, for the analysis of the simulated SBP and DBP phenotypes, Mallaney and Sung, [2014] found that SKAT controlled type I error at the 0.05 level using all three weighting schemes. For the analysis of the simulated SBP phenotypes, Sung et al. [2014] found that SSTs that were modified for the analysis of family data also controlled type I error at the 0.05 level whether all variants or only rare variants were considered. Similarly, Xuan et al. [2014] found that type I error for HC, SKAT, and minimum p -value approaches was well controlled under various grouping and collapsing strategies.

Poor Performance

For *MAP4* several contributors observed that most statistical methods performed relatively well for the simulated DBP and SBP phenotypes [Derkach et al., 2014; Mallaney and Sung, 2014; Sung et al., 2014]. When using related individuals, Sung et al. [2014] found that a collapsing method was able to identify *FLNB*, which contained one common causal variant with a relatively large effect (MAF = 0.49, explaining 0.3% of the variance in SBP).

However, other contributors observed that most statistical methods performed poorly in terms of overall power. For the analysis of the simulated hypertension phenotype, Dering et al. [2014] found that CMC and FPCA failed to detect most of the truly associated genes. Derkach et al. [2014] found low power for most of the causal genes on chromosome 3. Swartz et al. [2014] found that power from each of six methods was close to 0.05, after averaging each method's power across all causal genes on chromosome 3. Mallaney and Sung [2014] found that power from SKAT with all three weighting schemes was about 0.05 on average across all causal genes on all odd-numbered chromosomes.

These four research groups used a subset of unrelated individuals, resulting in low power. Using all 847 related individuals, Sung et al. [2014] found that power for the modified SST was about 0.14 on average across all causal genes on chromosome 3. Moreover, using a subset of unrelated individuals affected power beyond a simple reduction in sample size. Out of 8.3 million variants provided by GAW18, about 90,000 variants were monomorphic when individuals were restricted to those with the simulated phenotype data (847 related individuals). When individuals were further restricted to those 142 unrelated individuals with phenotype data, about 1.2 million variants were monomorphic. Our working group found that several causal variants (in particular rare variants) were absent in the unrelated-individuals data set.

Another cause of the poor performance may have been that most of the causal variants explained a small fraction of phenotypic variance [Sung et al., 2014]. The simulated DBP had 1,243 causal variants in 245 genes, and the simulated SBP had 1,040 causal variants in 205 genes. Chromosome 3, which was analyzed by most research groups, contained 134 causal variants in 22 genes; the detailed description for these 22 genes is available in Table 3 of Sung et al. [2014]. The gene with the largest effect on both DBP and SBP was *MAP4* (on chromosome 3), which included 15 causal variants (3 common and 12 rare variants). *MAP4*

explained 6.5% of DBP and 7.8% of SBP. Except for the *MAP4*, *LEPR*, and *FLT3* genes, most causal variants explained a small fraction of phenotypic variance; they ranged from 0.0% (although not identical to 0) to 2.7% (in *MAP4*) with a median of 0.001% for simulated SBP. Causal variants of the simulated DBP were similar.

Annotation Information and Binning Strategies

Using both real and simulated GAW18 data, Derkach et al. [2014] assessed the effect of different gene annotation methods and binning strategies, determined by Nalpathamkalam et al. [2014], on the analyses of rare variants. Using the real data, Derkach and colleagues found that quantile-quantile (Q-Q) plots of their results showed more deviation from the null hypothesis (with an excess of smaller *p*-values) when only protein-changing variants were included and still further deviation (again with smaller *p*-values) when a subset of protein-damaging variants were included. By focusing their analysis on protein-damaging variants, they found some promising signals, evidenced by a marked departure from the expected values in the Q-Q plot. Using the simulated data, they observed that top-ranked genes differed considerably across the different binning strategies, further confirming the practical importance of annotation in the analysis of rare variants.

Mallaney and Sung [2014] and Sung et al. [2014] used the annotation information based on the 1000 Genomes Project. They found that this annotation information contained about half the GAW18 causal variants and therefore that power was reduced by the use of incomplete annotation information. This further illustrates the importance of annotation in these rare variant analyses.

Discussion

The approximately 8 million variants available in the GAW18 sequence data exposed challenges but also provided an excellent opportunity for evaluating analysis choices. One contributor investigated annotation of variants observed in the GAW18 sequence data, whereas the remaining eight contributors evaluated several statistical methods for the association analysis of the provided phenotypes. Six contributors chose the gene, a natural unit of the human genome, as the genomic region for collapsing multiple variants. In contrast, three contributors chose nonoverlapping sliding windows across the entire genome. Gene-based bins led to a natural functional interpretation of each region but implicitly assumed that all causal variants were located within or nearby annotated genes and depended on the source of annotation information. Restricting variants based on bioinformatically predicted functional annotations (such as coding, protein changing, or protein damaging) may improve the power to identify causal variants but may also overlook causal variants that were not annotated. In contrast, sliding windows can incorporate the entire genome sequence, although interpretation of significant regions outside genes may be more challenging. However, one of the remarkable findings from the ENCODE Project is that biochemical functions were assigned for 80% of the genome [ENCODE Project Consortium, 2012], so there is hope that this challenge will diminish in the future.

The statistical methods used by our working group span most of the methods for the analysis of rare variants, including burden and variance-components-type tests, as well as hybrid

approaches. They also include lesser known methods, such as FPCA, higher criticism, and homozygosity association. Some contributors also introduced novel methods. We observed that the performance of these various methods depended on the characteristics of the genomic region, such as effect size and direction of variants under consideration. Performance of all these statistical methods was disappointingly poor, except for *MAP4* and *FLT3*. For most of these statistical methods the overall power was almost identical to the type I error. All these methods were originally developed for unrelated individuals. Therefore most of these methods were applied to a smaller sample of just over 100 unrelated individuals in the GAW18 data, resulting in low power. With such low power it was difficult to identify any advantages or disadvantages of newly developed methods [Derkach et al., 2014; Swartz et al., 2014]. By extending one of these methods to account for correlation across 847 related individuals within families, Sung et al. [2014] were able to improve overall power somewhat, although not satisfactorily.

The GAW18 data provided an excellent opportunity to evaluate existing and novel methods for identifying associations with rare variants. However, we observed that most of these methods performed poorly, demonstrating challenges in the analysis of rare variants. This poor performance may have arisen from a combination of reasons. First, as discussed in the previous paragraph, it may have been due to small sample size. Second, poor performance may have occurred because most of the causal variants explained only a small fraction of phenotypic variance (with a median of 0.001% for simulated DBP and SBP) [Sung et al., 2014]. Because the effect sizes used in the GAW18 simulations are likely realistic, this further demonstrates challenges in the analysis of rare variants. Third, the observed low power of some methods may have resulted from the use of incomplete annotation information [Mallaney and Sung, 2014; Sung et al., 2014]. Finally, linkage disequilibrium between causal variants in a gene and noncausal variants in nearby genes could have resulted in potential confounding [Derkach et al., 2014].

Conclusions

We have summarized the work of nine contributions to our working group that evaluated various approaches to the analysis of multiple variants. The approaches evaluated were sophisticated state-of-the-art statistical methods, aimed at identifying rare variants associated with human complex diseases. Most existing methods were able to identify *MAP4*, the gene with the largest effect on both DBP and SBP. Using all related individuals, a collapsing method was able to identify *FLNB*, another gene containing a common causal variant with a relatively large effect. However, most existing methods do not appear to be adequate for identifying rare variants with smaller effect sizes. This highlights the need to improve existing methods or to develop novel approaches for this task. Furthermore, we found that proper use of predicted biological information is as important as sophisticated use of statistical methods. Information about the biological and functional properties of genetic variants is essential for applying these statistical methods. With both biological and statistical insights, we should eventually be able to understand the genetic causes of complex diseases.

Acknowledgments

We thank Andrew Paterson, Senior Editor, and two anonymous reviewers for their constructive and insightful comments, which substantially improved the manuscript. The Genetic Analysis Workshop 18 was supported by the National Institutes of Health (NIH) through grant R01 GM031575. YJS was supported by NIH grants HL086694, HL107552, and HL111249.

References

- Agne M, Huang C, Lo S. Considering interactive effects in the identification of influential regions in extremely rare variants via fixed bin approach. *BMC Proc.* 2014; 8 (suppl 2):S7. [PubMed: 25519400]
- 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 2):S2. [PubMed: 25519314]
- Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet.* 2010; 44:293–308. [PubMed: 21047260]
- Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet.* 2010; 11(11):773–785. [PubMed: 20940738]
- Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol.* 2011; 35(7):606–619. [PubMed: 21769936]
- Bickeboller H, Thompson EA. The probability distribution of the amount of an individual's genome surviving to the following generation. *Genetics.* 1996; 143:1043–1049. [PubMed: 8725249]
- Casey JP, Magalhaes T, Conroy JM, Regan R, Shah N, Anney R, Shields DC, Abrahams BS, Almeida J, Bacchelli E, et al. A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Hum Genet.* 2012; 131(4):565–579. [PubMed: 21996756]
- Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH. Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci USA.* 2006; 103(6):1810–1815. [PubMed: 16449388]
- Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: An overview of collapsing methods. *Genet Epidemiol.* 2011; 35(suppl 1):S12–S17. [PubMed: 22128052]
- Dering C, Schillert A, Konig I, Ziegler A. A comparison of two collapsing methods in different approaches. *BMC Proc.* 2014; 8 (suppl 2):S8. [PubMed: 25519408]
- Derkach A, Lawless J, Merico D, Paterson A, Sun L. Evaluation of gene-based association tests for analyzing rare variants using Genetic Analysis Workshop 18 data. *BMC Proc.* 2014; 8 (suppl 2):S9. [PubMed: 25519417]
- Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet Epidemiol.* 2013; 37(1):110–121. [PubMed: 23032573]
- Donoho D, Jin JS. Higher criticism for detecting sparse heterogeneous mixtures. *Ann Stat.* 2004; 32(3):962–994.
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. [PubMed: 22955616]
- Hindorf LA, Junkins HA, Mehta JP, Manolio TA. [accessed September 7, 2013] A catalog of published genome-wide association studies. 2009. www.genome.gov/gwastudies
- Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* 2011; 7(2):e1001289. [PubMed: 21304886]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. Team NGENS-ELP. 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(2):224–237. [PubMed: 22863193]

- Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci USA*. 2007; 104(50):19,942–19,947.
- 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(3):311–321. [PubMed: 18691683]
- Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. *Genome Res*. 2011; 21(7):1099–1108. [PubMed: 21521787]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009; 5(2):e1000384. [PubMed: 19214210]
- Mallaney C, Sung YJ. Rare variant analysis of blood pressure phenotypes in the Genetic Analysis Workshop 18 whole-genome sequencing data using SKAT. *BMC Proc*. 2014; 8 (suppl 2):S10. [PubMed: 25519353]
- 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(7265):747–753. [PubMed: 19812666]
- McCarthy MI, Hirschhorn JN. Genome-wide association studies: Potential next steps on a genetic journey. *Hum Mol Genet*. 2008; 17(R2):R156–R165. [PubMed: 18852205]
- McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, et al. Runs of homozygosity in European populations. *Am J Hum Genet*. 2008; 83(3):359–372. [PubMed: 18760389]
- Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Huqun, Kyo S, Okazaki Y, et al. Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *Am J Hum Genet*. 2007; 80(6):1090–1102. [PubMed: 17503327]
- 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(1–2):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(2):188–193. [PubMed: 19810025]
- Nalpathamkalam T, Ziman R, Derkach A, Scherer S, Paterson A, Merico D. Genetic Analysis Workshop 18 single-nucleotide variant prioritization based on protein impact, sequence conservation, and gene annotation. *BMC Proc*. 2014; 8 (suppl 2):S11. [PubMed: 25519362]
- 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(3):e1001322. [PubMed: 21408211]
- 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*. 2010; 86(6):832–838. [PubMed: 20471002]
- Ramsay, JJO.; Silverman, BW. *Functional Data Analysis*. New York: Springer Science + Business Media; 2005.
- Sung YJ, Basson J, Rao DC. Whole genome sequence analysis of the simulated SBP in Genetic Analysis Workshop 18 family data: long-term average and collapsing methods. *BMC Proc*. 2014; 8 (suppl 2):S12. [PubMed: 25519365]
- Swartz M, Kim T, Niu J, Yu R, Shete S, Ionita-Laza I. Small sample properties for rare variant analysis methods. *BMC Proc*. 2014; 8 (suppl 2):S13. [PubMed: 25519366]
- Tukey, JW. *Course Notes, Statistics 411*. Princeton University; 1976. The higher criticism.
- 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(1):82–93. [PubMed: 21737059]
- Xuan J, Yang L, Wu Z. Higher criticism approach to detect rare variants using whole-genome sequencing data. *BMC Proc*. 2014; 8 (suppl 2):S13. [PubMed: 25519366]
- Yang HC, Chang LC, Huggins RM, Chen CH, Mullighan CG. LOHAS: Loss-of-heterozygosity analysis suite. *Genet Epidemiol*. 2011; 35(4):247–260. [PubMed: 21312262]

Yang HC, Li H. Analysis of homozygosity disequilibrium using whole-genome sequence data. BMC Proc. 2014; 8 (suppl 2):S15. [PubMed: 25519368]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Summary of statistical methods used by the working group on collapsing methods. The methods are organized into four broad categories: (1) early-developed burden tests based on linear statistics, (2) well-established variance-components-type tests based on quadratic statistics, (3) recently developed hybrid methods that combine burden and nonburden tests, and (4) other methods based on dimension-reduction techniques. The burden tests use linear weights to compute a weighted-sum statistic representing all rare variants in a region. These early tests are unidirectional by construction, assuming that all variants affect the phenotype in the same direction. The well-established variance-components-type tests use quadratic statistics for rare variants. These methods make no assumption on the direction of the effect on phenotype and therefore are able to handle both at-risk and protective variants. Both the recently developed hybrid methods and the other methods based on dimension-reduction techniques are also bidirectional approaches. CAST, cohort allelic sum test; CMC, combined multivariate and collapsing; PCA, principal components analysis; RBS, replication-based weighted-sum statistic; SKAT, sequence kernel association test; SKAT-O, SKAT with the use of the optimal weighted average; SST, simple sum test; VT, variable-threshold; w-SUM, group-wise weighted sum.

Table 1

Overview of contributions to the working group on collapsing methods

Contribution	Phenotype	Chromosomes	Unit to combine variants	MAF threshold	Statistical methods for association analysis of multiple variants
Agne et al. [2014]	Simulated HTN	3	Bins (with 100 SNPs)	Private SNPs	Proportion of private SNPs in cases for low- and high-environment groups
Dering et al. [2014]	Simulated HTN	All but chromosome 5	Gene	0.05	FPCA, CMC
Derkach et al. [2014]	Real HTN; simulated DBP, SBP	3	Gene, coding, protein changing, protein damaging	0.05	CAST, w-Sum, C-alpha, Hotelling, SKAT-O, minimum <i>p</i> -value, Fisher's
Mallaney and Sung [2014]	Simulated DBP, SBP	All	Gene	All, 0.05	SKAT with three weighting schemes (equal, Madsen-Browning, and default)
Nalpathamkalam et al. [2014]	Not applicable	All	Not applicable	All	Variant prioritization (e.g., coding, protein-changing, and protein-damaging variants)
Sung et al. [2014]	Simulated SBP	3	Gene	All, 0.05	SST modified for family data
Swartz et al. [2014]	Simulated HTN	Causal genes on chromosome 3	Gene	0.05	VT, RBS, C-alpha, SKAT, modified RBS
Xuan et al. [2014]	Simulated SBP	1, 3, 5	Bins (with 10 kbp, 100 kbp, 500 kbp), gene	0.01, 0.05	HC, minimum <i>p</i> -value, SKAT
Yang and Li [2014]	Real DBP, SBP, HTN	All	Sliding windows (5% SNPs on chromosome)	0.05	Runs of homozygosity

CAST, cohort allelic sum test; CMC, combined multivariate and collapsing; DBP, diastolic blood pressure; FPCA, functional principal components analysis; HC, higher criticism; HTN, hypertension; MAF, minor allele frequency; RBS, replication-based weighted-sum statistic; SBP, systolic blood pressure; SKAT, sequence kernel association test; SKAT-O, SKAT with the use of the optimal weighted average; SNP, single-nucleotide polymorphism; SST, simple sum test; VT, variable-threshold; w-Sum, group-wise weighted sum.