Santorico and Hendricks *BMC Genetics* 2016, **17**(Suppl 2):6
DOI 10.1186/s12863-015-0316-7

BMC Genetics

PROCEEDINGS Open Access

Progress in methods for rare variant association

Stephanie A. Santorico* and Audrey E. Hendricks

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

Empirical studies and evolutionary theory support a role for rare variants in the etiology of complex traits. Given this motivation and increasing affordability of whole-exome and whole-genome sequencing, methods for rare variant association have been an active area of research for the past decade. Here, we provide a survey of the current literature and developments from the Genetics Analysis Workshop 19 (GAW19) Collapsing Rare Variants working group. In particular, we present the generalized linear regression framework and associated score statistic for the 2 major types of methods: burden and variance components methods. We further show that by simply modifying weights within these frameworks we arrive at many of the popular existing methods, for example, the cohort allelic sums test and sequence kernel association test. Meta-analysis techniques are also described. Next, we describe the 6 contributions from the GAW19 Collapsing Rare Variants working group. These included development of new methods, such as a retrospective likelihood for family data, a method using genomic structure to compare cases and controls, a haplotype-based meta-analysis, and a permutation-based method for combining different statistical tests. In addition, one contribution compared a mega-analysis of family-based and population-based data to meta-analysis. Finally, the power of existing family-based methods for binary traits was compared. We conclude with suggestions for open research questions.

Background

Rare variants have increasingly become a focus in studies of complex traits. There are many reasons for this increasing interest. Accessibility, both in cost and technology, of next-generation sequencing has led to the discovery of a plethora of rare variants. Nelson et al. [1] estimated that 95 % of variants were rare with a minor allele frequency (MAF) of less than 0.5 %. This is in stark contrast to previous research suggesting that nearly one-third of variants have a frequency below 5 % [2]. Furthermore, evolutionary theory suggests that deleterious variants are selected against and thus should be rare [3]. Recent research has supported this theory, observing that a large proportion of deleterious variants are indeed rare [4, 5]. Despite the effects of this purifying selection, the 1000 Genomes project estimates that individuals carry 76 to 190 rare nonsynonymous variants predicted to be deleterious [6].

A more contentious argument for focusing genetic research on rare variants pertains to the so-called phenomenon of "missing heritability". Genome-wide association studies (GWAS) have successfully identified numerous common variants associated with complex traits; however, the common variants tend to have relatively small effects and explain only a fraction of the overall heritability [7]. Human height serves as an excellent example with estimates of heritability near 80 %. GWAS variants with genome-wide significant associations explain only approximately 5 % of overall variation in height whereas models that use all high-quality GWAS variants with a MAF of more than 1 % explain approximately 45 % of the variation [8, 9]. Even though the latter is a substantial improvement, it is still well shy of 80 %.

There is emerging evidence that rare variants are involved in complex disease, including Alzheimer disease [10], lipids and coronary artery disease [11], irritable bowel disease [12], prostate cancer [13], and many others [11, 14–16]. Despite these encouraging results, many studies continue to be underpowered to detect

* Correspondence: Stephanie.Santorico@ucdenver.edu
Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO 80217-3364, USA

association to disease-associated rare variants. Continued development of methods is needed to help increase the power to detect these associations. This is particularly true given that tests of individual rare variants are underpowered without exceptionally large sample sizes [14]. Combining variants based on a gene or region is a popular strategy. Other strategies for improving power for detecting rare variants include using family samples or isolated populations to increase the frequency of a variant that is rare in the general or nonisolate population [17, 18]. Ascertaining phenotypic extremes can also increase the likelihood of sampling individuals with disease-associated rare variants, thus increasing the power of rare variants tests [19, 20]. Finally, incorporating biological knowledge and genomic annotation to exclude, or downweight, variants in analyses is also an effective strategy, focusing tests on variants more likely to be deleterious [21].

Here, we provide a summary of the current literature with respect to the association of rare variants and ways to increase the power of these tests. We then provide results from the Collapsing Rare Variants Working Group of Genetic Analysis Workshop 19 (GAW19) and conclude with recommendations and open problems.

Current literature

Although there is no formal definition for a rare variant, variants with a MAF between 5 % and 50 % are generally considered common. Variants with a MAF in the range of 1 % to 5 % [15] or 0.5 % to 5 % [22] are considered low frequency or less common. Rare variants have a MAF falling below these ranges, whereas a private variant is specific to probands and their relatives.

Basic association models for collapsing rare variants

The 2 major types of methods for collapsing rare variants within a meaningful genetic region, such as a gene, consist of burden tests and variance component tests. Burden tests measure the burden of variants within a genetic region and range from a simple indicator of whether a genetic region contains at least 1 rare variant (eg, CAST [cohort allelic sums test] [23]), to a sum of the rare variants within a region (eg, ARIEL [accumulation of rare variants integrated and extended locus-specific test] [24], CMC [combined multivariate and collapsing] [25], and MZ [Morris and Zeggini] [26]), to a weighted sum of the rare variants in a region (eg, WSS [weighted sum statistic] [27] and aSum [data-adaptive sum] [28]).

A general formula for the burden of rare variants within a region is shown in Eq. (1).

$$B_i = \sum_{m=1}^{M} G_{i,m} w_m \tag{1}$$

where $G_{i,m}$ is the genotype coding based on a genetic model (eg, recessive, dominant, or additive) for individual i and variant m, w_m is the weight for variant m, and M is the total number of variants in the region. This model is applicable to nearly all burden tests mentioned above. For example, the CAST is often written in terms of an indicator function such as $B_i = I\left(\sum_{m=1}^{M} G_{i,m} w_m\right)$ where $G_{i,m}$ is either 1 or 0, depending on whether the subject has or does not have variant m, respectively, and $w_m = 1$. However, we can use Equation (1) directly by making $w_m = \frac{1}{\sum_{m=1}^{M} G_{i,m}}$ to ensure that B_i is 1 for subjects who have at least 1 variant in the gene and 0 otherwise. For a simple count of the number of variants within a region, a weight of 1 for each variant is used. Others have proposed more informative weights. Madsen and Browning [27] proposed a weight that increases as MAF decreases while Asimit et al. [24] weighted genotypes by their quality.

Even though several tests were first developed outside of the regression framework [23, 25, 27], nearly all can be easily implemented in a generalized regression framework (Eq. 2) by incorporating B_i as a covariate in the regression model. This greatly generalizes the statistical framework allowing for many types of outcome variables (eg, continuous, binary, survival, etc.), and the incorporation of additional possible confounders and covariates:

$$f(\mu) = \gamma_0 + \boldsymbol{\gamma}'\boldsymbol{X} + \beta B \tag{2}$$

where $f(\mu)$ is a function that links a linear combination of the predictors and the mean, μ, of the outcome (eg, disease or trait); γ_0 is the intercept; $\boldsymbol{\gamma}'$ is a vector of parameters for the covariates, $\boldsymbol{X}$; β is the regression parameter for the burden of rare variants within a region, B; and bolded symbols denote a vector. For a quantitative trait $f(\mu) = \mu$ is used within a linear regression framework, and for a qualitative trait $f(\mu) = \text{logit}(\mu)$ is typically used within a logistic regression framework. Although several test statistics can be implemented within the generalized regression format, we focus on the score statistic, U, testing whether $\beta = 0$. The burden score statistic is shown in Equation (3) and under the null hypothesis of no association has a chi-square distribution with 1 degree of freedom (df). Note that the burden score statistic can be written as a weighted sum of the marginal score statistics, S_m, for each genetic variant where, $S_m = \sum_{i=1}^{n} G_{i,m}(y_i - \hat{\mu}_i)$ for n individuals, with $\hat{\mu}_i$ being the estimated mean for individual i, which includes the effects of covariates as estimated through generalized regression.

$$U_{burden} = \left[\sum_{i=1}^{n} B_i(y_i - \hat{\mu}_i)\right]^2 = \left[\sum_{m=1}^{M} w_m \sum_{i=1}^{n} G_{i,m}(y_i - \hat{\mu}_i)\right]^2 \tag{3}$$
$$= \left[\sum_{m=1}^{M} w_m S_m\right]^2 \sim \chi_1^2$$

As marginal score statistics can first be calculated on each variant, this alternative form lends itself nicely to extensions such as meta-analysis as described later.

Instead of calculating the burden of variants within a genetic region, variance component tests (eg, sequence kernel association test [SKAT] [29], C-alpha [30] and SumSqU [31]) evaluate the similarity of the variants within the region. Simply, we expect the distribution of variants to be more similar for subjects with similar trait values than for subjects with different trait values. Like with the burden test, a general equation for the score statistic of the variance component test can be written and is shown in Eq. (4).

$$U_{VC} = \sum_{m=1}^{M} (w_m S_m)^2 \quad (4)$$

where S_m is the previously defined marginal score statistic. U_{VC} follows a mixture chi-square distribution. Because the marginal score statistic is squared, both negative and positive effects can be included in the statistic. This is a notable advantage of variance component tests over burden tests, for which effects of different directions can cancel each other out. For both C-alpha [30] and SumSqU [31], the weights equal 1. C-alpha is further restricted to scenarios where the phenotype is dichotomous, and there are no covariates. The SKAT statistic [29] is identical to U_{VC}, accommodating a variety of weights; as such, C-alpha and SumSqU are special cases of SKAT.

Burden tests tend to be most powerful when the majority of variants have an effect in the same direction [25, 29, 32]. Variance component tests are more powerful when the variants have different effects (ie, many variants with no effect or effects in opposite directions) [29, 32]. To combine the different strengths of the burden and variance component tests, Lee et al. [32] developed an optimal unified approach called the SKAT-O, where the burden and SKAT tests are combined with a weighting parameter, ρ (Eq. 5). Note that the optimal test is equivalent to the burden test and SKAT (ie, variance component test) when ρ is 1 or 0, respectively.

$$U_{optimal} = (1-\rho)U_{SKAT} + \rho U_{burden},\ 0 \le \rho \le 1 \quad (5)$$

Others have explored combining the burden and variance components tests as well [33, 34]. Finally, more recently the EC test [35] was developed under a Bayesian framework with an alternative hypothesis prior that gives a higher probability to only 1 causal variant per genetic region.

Here, we provide a basic overview of general methods; others have done this as well in more detail [22]. In addition, Derkach et al. [36] have provided an excellent review and comparison (both empirical and theoretical) of existing methods. Important conclusions and results include: weighting variants inversely to the MAF does not always increase power even under scenarios where rare variants where simulated to have a larger effect; as the sample size increases the variance component statistic tends to have a higher power than the burden statistic; uniformly optimal tests are difficult to achieve in practice.

Incorporating additional information

There have been many extensions to the basic frameworks and models to include and account for additional information. Various weights can be defined based on the MAF [27], quality of genotype calls [24], previous evidence for association, direction of effect (eg, aSum [28]), evolutionary conservation (eg, phastCons [phylogenetic analysis with space/time models conservation] [37], and GERP [genomic evolutionary rate profiling] [38]), probability of being functional, and likelihood of being deleterious. There exists several algorithms/software that predict whether a variant is likely to be deleterious, including CONDEL (consensus deleteriousness) [39], SIFT (sorting intolerant from tolerant) [40], PolyPhen (polymorphism phenotyping) [41], CADD (combined annotation-dependent depletion) [42], and several others (see Castellana and Mazza [43]). Although the predictions of these programs can differ greatly [39, 43, 44], variants that have consistent predictions of either being benign or deleterious across all programs may be more likely to be truly benign or deleterious. Variants can be removed entirely from the model by using a weight of 1 and a weight of 0 for variants fulfilling or not fulfilling a requirement or threshold, respectively. It is often difficult to know the true or best threshold to use when determining which variants to include in the model. Adaptive methods implement the region-based methods over a variety of thresholds (such as various MAF thresholds) and then adjust for multiple comparisons using permutation [28, 45].

It is worth emphasizing that the proportion of variants in the collapsing test with association to the outcome is directly related to the power [46, 47]. As such, choosing which variants to include is extremely important. When choosing variants, various factors should be considered such as the likely penetrance of the variants, the prevalence of the disease or trait, and the predicted deleteriousness of the variants. As discussed in the previous paragraph, instead of weighting variants, only a subset of variants can be kept, such as those predicted to be deleterious or to result in loss of function.

Once a gene or region has been identified as being associated to a disease or trait, an important next step is to identify the causal variants within the region. Experimental studies to determine the functional effects are often costly both in effort and money. In a recent paper, Ionita-Laza et al. [48] proposed and compared 2 methods to identify likely causal variants within gene regions.

Unlike rare variants, the parameter estimation for common variants is generally stable. Including disease-

associated common variants within a gene region could help to identify genetic regions associated with a trait as well as to help determine if a collapsed set of rare variants produces an independent signal above that from the common variants. Determining which common variants to include in the model is not always straightforward as too many variants will dilute the signal and decrease the power by using up valuable degrees of freedom, while including too few variants may miss a signal all together. Penalized regression methods, such as LASSO (least absolute shrinkage and selection operator), have been proposed [49] as well as an extension to the SKAT framework that incorporates common variants [50].

More recently, methods have been developed to compare the observed number of filtered variants within a genetic region to that expected genome- or exome-wide [51] or expected by an estimated mutation rate [52]. These methods are most often implemented in a case-only framework, and are thus sensitive to the estimates of comparison (eg, genome-/exome-wide averages, mutation rates, etc.). These methods are discussed further below.

Study design considerations

Although sequencing costs continue to decline, the cost of sequencing continues to impose a limit on the number of samples that can be sequenced. There is increasing evidence that the power to detect an association to rare variants is low regardless of the type of test or statistic used [46, 47]. As such, study design is of utmost importance and includes, among others, family-based, trio, case-only, case–control, and population cohort designs.

Study design affects the power and generalizability of the study. Certain study designs may increase the power to detect an association in certain situations while decreasing the ability to detect other genetic associations. For instance, sampling families with a particular rare disease increases the likelihood of observing multiple copies of the causal rare variant, thus increasing power to detect an association to that particular variant [53]. However, this sampling framework may reduce the number of detected genetic variants, making it more difficult to discover the variety of genetic variants that would be seen when sampling the general population. Recently, several methods have been developed or extended to accommodate related samples [54, 55]. Probably the most widely used of these is famSKAT (family-based sequence kernel association test) developed by Chen et al. [56], which extends SKAT by using a linear mixed effects model to account for the family structure in tests of quantitative traits. For GAW19, Wang et al. [57] studied the type 1 error and power of current family-based methods for rare variant association tests for dichotomous phenotypes. It is also important to note that valid permutation to assess significance in the context of dependent samples (such as with related samples or population stratification) is not straightforward. Others have explored permutation in this setting and have proposed modified permutation procedures [58].

For extremely rare, highly penetrant disorders, researchers have had success sequencing a set of cases [59, 60] or trios where the offspring has an extremely rare disorder and the parents are unaffected [61–63]. Specific software exists for detecting de novo mutations within trio designs [64]. For more complex and common diseases or traits, study designs such as a case–control or population-cohort are often used [21, 65, 66]. Although many case–control studies are retrospective, few incorporate the retrospective ascertainment of the sampling design into the statistical framework. Such methods were included in GAW19 contributions [57, 67]. Unfortunately, detecting rare genetic associations in complex diseases has continued to prove difficult and much larger sample sizes are needed to achieve adequate power. Some study designs use extreme sampling either of cases [68] or of quantitative phenotypes [21] to increase power. For complex traits, extreme sampling can lead to an increase in the number of rare variants detected and subsequently an increase in power [69]. However, not accounting for the trait-dependent sampling when analyzing quantitative traits can lead to biased estimates, inflated type 1 error, and even a decrease in power [70]. In 2013, Barnett et al. [69] and Lin et al. [70] each developed novel statistical methods to appropriately analyze quantitative traits with extreme sampling study designs.

Within the study design of sequencing a unique and homogeneous set of cases, case-only statistical frameworks exist for detecting exceedingly rare or de novo and highly penetrant variants [51, 52]. Statistical frameworks also exist to incorporate external population controls with the unique set of cases in a case–control analysis [71], although more research in this area is needed.

As previously discussed, most methods can be expressed within a regression framework. Many of the burden methods are within a generalized linear regression framework while the variance component methods, such as SKAT, are implemented within a mixed effects regression model. The original regression frameworks of these methods required large enough sample sizes to reach an asymptotic distribution of the test statistics and independent observations. Few methods have been developed specifically for small samples, although Lee et al. [72] extended SKAT for use with small sample sizes.

Meta-analysis

Meta-analysis of test statistics across multiple studies is widely used in GWAS and other genetic studies of common variants to replicate, confirm, and find new associations. Meta-analysis is arguably even more important for

studies of rare variants where extremely large sample sizes are important for achieving adequate power. Many simple meta-analysis frameworks that combine information about the test statistic or p value (such as Fisher's and Z-score methods [73]) can be applied to test statistics from current region-based methods. (Although it should be noted that, as there is no direction of effect for variance component tests, only weights based on sample size and not direction of effect can be incorporated into Z-score meta-analysis for variance component tests.) Although simple and easy to implement, these meta-analysis methods do not account for different variants that may be included in the region-based statistics for each study.

Lee et al. [74] developed a meta-analysis framework for rare variants that achieves nearly identical empirical power as analyses based on combined individual level data (sometimes called mega-analysis). This framework uses single-variant score statistics and the corresponding between-variant covariance matrix. Importantly, the framework allows for variants to be monomorphic (ie, the alternate allele is not seen) in some of the individual studies. To be included in the meta-analysis statistic, a variant only has to be polymorphic in at least 1 study. Further, meta-analysis has other advantages such as easier sharing of data (given consent or computational barriers to sharing raw data) and controlling for potential confounders or population stratification specific to each study. For instance, one study may adjust for 5 principal components whereas another study may adjust for 3 principal components and recruitment center. In addition to being able to include different study-specific covariates, one can also further account for possible heterogeneity in study statistics in the meta-analysis statistic itself as described below.

Here, we briefly outline Lee et al's [74] meta-analysis framework. If we define the single-variant (ie, marginal) score statistics as $S_{k,\,m}$ for study k and variant m, we can then rewrite the burden score statistic as a combination of the single-variant statistics over all studies:

$$U_{burden\_meta} = \left[\Sigma_{m=1}^{M}\Sigma_{k=1}^{K} w_{k,m}S_{k,m}\right]^2.$$

We can also square the single-variant score statistics summed over the studies and then summed over variants to produce a meta-analysis score statistic for the variance component region test:

$$U_{VC\_meta\_hom} = \Sigma_{m=1}^{M}\left(\Sigma_{k=1}^{K} w_{k,m}S_{k,m}\right)^2.$$

The above variance component statistic requires the additional assumption of homogeneous genetic effects across all studies. If we believe that the genetic effects are instead heterogeneous, the meta-analysis score statistic for the variance component region test can be written as follows:

$$U_{VC\_meta\_het} = \Sigma_{m=1}^{M}\Sigma_{k=1}^{K}\left(w_{k,m}S_{k,m}\right)^2.$$

If we believe the heterogeneity can be isolated to clusters of studies, such as ethnicity, the statistics can be combined, first over the studies in each cluster and then over each cluster and marker. Note that the burden and variance component meta-analysis test statistics can be combined in an optimal way similar to that shown for single studies in Equation (5). More details are in Lee et al. [74]. Others have explored meta-analysis for rare variants as well [75].

Contributions from the collapsing rare variants working group

GAW19 provided real human sequence and phenotype data for data sets of Mexican American families and unrelated individuals. In addition, 200 simulated data sets were provided based off the real sequence data for phenotypes with true underlying genetic associations as well as a null polygenic trait. Family data (for 959 individuals in 20 pedigrees) consisted of whole genome sequencing calls and GWAS single-nucleotide polymorphisms (SNPs) for odd-numbered chromosomes, as well as longitudinal real phenotype data for systolic and diastolic blood pressure, age, sex and indicators of hypertension, antihypertensive medication use, and cigarette smoking, collected at up to 4 time points. Family data also included genome-wide measures of gene expression for a smaller set of individuals; however, no contribution in our group utilized this data nor did any contribution utilize the longitudinal nature of the data. The data set of 1943 unrelated individuals contained exome sequence calls and the same phenotypes as the family data, at a single time point. More detailed information on the GAW19 data sets is available in Blangero et al. [76].

The 6 contributions from the Collapsing Rare Variants Working Group of GAW19 extend upon the current literature and reflect varied goals, including the creation of new statistical tests, developments of meta-analytic techniques and a comparison of existing statistical tests. Table 1 provides overall characteristics of each contribution.

New statistics

Green et al. [77] developed a general framework for combining different statistical tests of association of rare variants with a continuous trait in family-based studies. A linear mixed model was used to derive residuals by adjusting for covariates as well as a random effect for familial correlation. These residuals were then permuted to create data sets reflective of the null hypothesis of no association, allowing for the derivation of empirical

Table 1 Contributions from the GAW19 Collapsing Rare Variants working group

Goal	Reference	Trait	Data type	Statistic type
New statistic	Green et al. [77]	Quantitative	Family simulated	Combined burden and variance component
	Jadhav et al. [78]	Dichotomous	Unrelated, simulated	Burden and Variance component
	Zhu et al. [67]	Quantitative	Family simulated	Variance component
Meta-analysis	Katsumata and Fardo [81]	Quantitative	Family and unrelated simulated	Variance component
	Wang et al. [82]	Quantitative	Family and unrelated real data	Variance component, haplotype model
Method comparison	Wang et al. [57]	Dichotomous	Family simulated	Burden, variance component and an optimal combination

p values that combine information over a set of rare variant tests yielding a single overall test of association. In the Green et al. formulation [77], evidence was combined over 4 burden tests and 4 variance–component tests representing different powers of the marginal score statistics (U, U^2, U^3 and U^4) as well as over 2 weight functions, one based on the Beta distribution [29], and the other based on the inverse standard deviation of the allele count [45]. With increasing powers of the marginal score statistics, the contribution of noncausal variants to the overall statistic is lessened, and the use of the Beta distribution more severely down weights common variants compared to those based on the inverse standard deviation of the allele count. By utilizing all combinations of weight function with powers of the score statistic, a variety of models are included within the test. However, given the permutation framework, their method can be generalized to any set of statistical tests. In evaluating their method, Green et al. [77] focused on the GAW19 simulated data set of 30 genes on chromosome 3 that have at least 1 casual variant; type 1 error and power were estimated for the combined approach, as well as for each of the 4 burden tests and 4 variance-component tests. Type 1 error was controlled at the 0.05 level based on the null trait, Q1, provided with the simulated data. The combined approach consistently yielded intermediate power relative to the power of the four burden and four variance-component tests. Given there is no single best test and that the optimal statistic is unknown a priori, the combined approach allows for proper control of type 1 error and is an approach that is robust to differing genetic architectures. Further research is needed to determine an optimal combination of tests, ones that are uncorrelated, reflecting different patterns of association, and that maximize power.

Zhu et al. [67] derived a score test, OW-score, based on the retrospective likelihood for a continuous trait formed by conditioning on observed phenotypes. The resulting test is a function of a weighted combination of genotypes over the variants included in the test, where the weighting is derived to maximize the score statistic. Power for the OW-score method was compared to that of famSKAT [56] for the 14 genes with the largest simulated signal for diastolic blood pressure in the GAW19 data at a significance level of 0.05. Only 4 genes, yielded power that was greater than 40 % for either method; of these, the OW-score test was more powerful for 3 genes and the famSKAT was more powerful for 1 gene. It is important to note that the distribution of the weights by MAF differs between the OW-score test and famSKAT, with famSKAT more highly weighting variants with a MAF within (0.01, 0.05). This aligned with the simulation results and performance of the OW-score test compared to famSKAT: when causal variants fell within this MAF range, the famSKAT was more powerful than the OW-score method. Thus, more research is needed to determine if the retrospective nature of the OW-score test or the varied weighting structure is leading to increased power in certain scenarios.

Jadhav et al. [78] used a method, from the branch of statistics called *functional data analysis*, which is based on analysis of curves, surfaces, or functions [79]. Specifically, a functional analysis of variance (ANOVA) model compared the difference in the genetic structure of a genomic region between cases and controls. To do so, a continuous function was fit to each individual's genotype using cubic B-splines over a 30-kb region, and the resulting mean function was compared between cases and controls using an ANOVA test. Results were compared with a burden test that weighted minor allele counts by the inverse standard deviation for the minor allele count in controls [27], as well as to a burden test that incorporates linkage disequilibrium through a genetic covariance matrix [80]. Simulations were conducted for a 1.4-Mb region of chromosome 3 where causal variants were randomly selected to be 1 % to 50 % of the region, and phenotypes were simulated using both unidirectional and bidirectional effects. The functional ANOVA test had greater power, up to 0.135 higher, compared to the burden test (with or without incorporation of linkage disequilibrium) over all but 1 scenario, in which 50 % of the variants in the region were causal and with unidirectional effect. In this scenario, power was comparable.

Developments in meta-analysis

Katsumata and Fardo [81] applied the famSKAT statistic to each of the GAW19 family- and population-based data sets, as well as to the combined set of data, resulting in a mega-analysis. These 3 analyses were compared against a meta-analysis of the family- and population-based data sets for the 15 most causal genes influencing each of diastolic and systolic blood pressure in the GAW19 simulation model (23 genes in total, given overlapping causal genes). They found that mega-analysis could be substantially more powerful than meta-analysis (*NRF1*, *LEPR*, *LRP8*, *GAB2* with systolic blood pressure [SBP]) with meta-analysis resulting in discernibly higher power compared to mega-analysis for only one of the top genes (*TNN* with both SBP and diastolic blood pressure [DBP]). However, when the power to detect association to a gene-region was considerably lower within the family-based sample versus the population-based sample, the power of the mega-analysis was much lower than the analysis based on the population-based sample alone, while the meta-analysis had a less-severe power loss. This suggests that the mega-analysis may be better when there is sufficient power to detect an association in both samples, but a meta-analysis might be more suited to situations where one study is underpowered and/or there is heterogeneity in the genetic associations between study samples. Both meta-analysis and mega-analysis indicated elevated type 1 error with estimates based on the 200 simulated data sets of the null trait Q1 ranging from 0.055 to 0.130 and 0.050 to 0.135, respectively, for the 23 genes.

Wang et al. [82] also considered a meta-analysis of the famSKAT statistic applied to the family- and population-based data sets for DBP. These results were compared to a meta-analysis of results from a haplotype-based association model. For haplotype analysis, a mixed linear model was fit, allowing for covariates, fixed effects of haplotypes (with haplotypes with frequency of less than 0.5 % collapsed into 1 group) and random components for family structure and error. Haplotypes were coded using dosages estimated from genotypes using the expectation–maximization algorithm. Models were fit separately for the family-based and population-based samples, and the weighted-least squares method of meta-analysis was followed by a Wald test of equal haplotype effects. Type 1 error for the haplotype model was found to be elevated for genes with more than 14 haplotypes; hence, results on the real data set were given for only genes with fewer than 14 haplotypes. None of the genes were significant for famSKAT after correcting for multiple testing; however, multiple genes did achieve statistical significance using the haplotype model indicating a potentially more powerful method for association testing. As these results are from real data, further study is needed to understand relative performance of the 2 methods over a range of models.

Method comparison

Finally, Wang et al. [57] compared existing family-based methods for binary traits including the rare variant transmission disequilibrium test (RV-TDT) [55], the generalized estimating equations–based–kernel association (GEE-KM) test [83], an extended CMC test for pedigree data known as PedCMC [84], a gene-level kernel and burden association tests for pedigree data (PedGene) [80], and the family-based rare variant association test (FARVAT) [85]. Through simulation based on the 6 genes with the largest effects on both simulated SBP and DBP, they found that the FARVAT method based on optimal weights (that adaptively use the data to combine burden and variance component tests) was more powerful than the PedCMC, GEE-KM, or any of the RV-TDT tests. The power for the PedGene method was comparable with that of FARVAT; however, FARVAT required substantially less computing time. Based on dichotomization of the simulated null trait Q1 to correspond to a prevalence of 22.6 %, type 1 error was demonstrated to be deflated for the RV-TDT and inflated for the GEE-KM test, while the PedCMC, PedGene, and FARVAT had reasonable control of type 1 error across a range of significance levels.

Discussion and conclusions

Over the last 10 years, there has been considerable methods development for association tests of rare variants. Tests have been proposed that are ideal for unidirectional and bidirectional effects, as well as an optimized combination of the 2 types of effects. Methods have been proposed for binary as well as normally distributed traits, for population-based and family studies. Most tests allow for the use of different weighting schemes (eg, based on MAF or genomic annotation), and meta-analysis procedures have also been developed.

Contributions to the GAW19 Collapsing Rare Variants group expanded upon the literature in several ways. Green et al. [77] provided a method that could be used to combine any collection of statistics for rare variant association. This is particularly important given there are numerous types of annotation that could be used as weights and these weights could be implemented in a burden model, variance component model, or combination of the 2 models. While an oft-used strategy is to conduct all tests separately, the method proposed by Green et al. would allow for an empirical combination in a statistically rigorous framework while controlling for total type 1 error.

New statistical tests were developed to allow for a retrospective likelihood based on optimized variant weights [67] and to incorporate genomic structure into the test of rare variants [78].

Katsumata and Fardo [81] provided guidance regarding design and meta-analysis. Based on GAW19 simulated

data, they found that mega-analysis generally led to higher power than a meta-analysis; however, if there were large differences in power between the family-based and population-based studies, a mega-analysis could have power less than that of the studies being combined, meta-analysis was less affected by this scenario. Wang et al. [82] compared meta-analysis based on haplotypes to meta-analysis based on famSKAT statistics, demonstrating the 2 approaches to be complementary by detecting associations to different genes.

Finally, Wang et al. [57] compared existing family-based methods for rare variant association to binary traits and demonstrated PedGene and FARVAT to be powerful methods for rare variant association, with FARVAT being more computationally efficient.

Although much work has been done, there are still many open research areas pertaining to the analysis of rare variants. We mention a number of these areas; however, this list is by no means exhaustive. For example, great care has been taken in studies of common variants to control for population substructure. These have included the use of genetic principal components, genomic matching and linear mixed models; see, eg, the review by Price et al. [86]. Given rare variants are confounded with population ancestry, it is not clear how to best control for this substructure. Although there has been some work in this area in showing that population substructure is indeed different for common and rare variants [87, 88], more work, especially in method development, is needed.

It often makes sense to focus on the gene region as the unit for collapsing methods, especially given analyses within the coding regions of the genome. However, GWAS associations are often in intergenic regions, and there is building evidence that much of the noncoding region of the genome is indeed functional [89, 90]. Thus, there is an interest in testing noncoding regions of the genome for association with rare variants and how best to define the regions is an open question. A sliding-window based approach is often used to group regions of the genome for testing. There are many additional questions when using sliding windows such as number, size of window, and size of overlap between windows. Genomic windows will need to be large enough to capture the causal region without being too large so as to include too much noise. There is likely to be a tradeoff between multiple testing adjustments necessary to account for many small windows versus the potential power loss from using fewer windows that are too large. In addition, as the functionality of the noncoding regions continues to be discovered and defined, it is likely that there will be useful information to use when building or defining the windows or meaningful genetic units within the non-coding regions.

As we have detailed, methods exist for incorporating genomic annotation as weights in region based methods. The choice of the best weight and, in fact, which information to consider at all, remain somewhat open questions. Currently, MAF, functionality, consequence, evolutionary conservation and many other metrics can be used as weights and the list continues to grow, especially in non-coding regions as functional research continues at a rapid pace. Thus, there is a need to further develop efficient methods of deriving the most appropriate weight. This can be done to some extent through the adaptive methods discussed previously. However, the adaptive methods, which often rely on permutation, may become computationally infeasible given the increasing amount of information on which to weight, increasing sample size, and analysis on the entire genome. Thus, there will continue to be a need for computationally efficient methods of determining the weights while retaining the appropriate type 1 error.

To date, most collapsing of rare variants is done on a contiguous region of the genome, whether it is a gene or a genomic window. Alternative approaches include the use of pathways or gene sets developed, for example, from expression studies or protein-protein interaction studies. Recent studies have found some success with this approach [91], but more research is needed.

Finally, given the continued struggle to adequately power studies of rare variants, more work is needed on ways to improve power. One approach is to continually increase the sample size of the studies, perhaps through including publically available population controls. Another, perhaps more feasible approach may require refocusing the phenotype through use of multidimensional phenotypes or homogeneous subphenotypes.

Given this relatively brief discussion of remaining areas of research for the association of rare variants, there is little doubt that this will continue to be an active area of research for several more years.

Competing interests
The authors declare they have no competing interests.

Authors' contributions
SAS served as group leader during for the GAW19 Collapsing Rare Variants working group, editor for the corresponding contributions and summarizes the group work herein. AEH summarized existing literature. Both authors read and approved the final manuscript.

Declarations
This article has been published as part of *BMC Genetics* Volume 17 Supplement 2, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at www.biomedcentral.com/bmcgenet/supplements/17/S2. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Acknowledgements
We would like to express our thanks to members of the Collapsing Rare Variants working group for their participation, many active discussions and for their subsequent comments and assistance. In addition, we are particularly grateful for the in-depth critique and suggestions made by two anonymous reviewers. GAW19 was supported by NIH grant R01 GM031575.

Published: 3 February 2016

References

1. Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean PS, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337(6090):100–4.
2. Hartl D, Clark A: Principles of Population Genetics. Sunderland, Sinauer Associates 1998
3. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2011;13(2):135–45.
4. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007;80(4):727–39.
5. Zhu Q, Ge D, Maia JM, Zhu M, Petrovski S, Dickson SP, et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. Am J Hum Genet. 2011;88(4):458–68.
6. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.
8. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24.
9. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011;43(6):519–25.
10. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature. 2014;505(7484):550–4.
11. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet. 2014;94(2):223–32.
12. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066–73.
13. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet. 2012;44(12):1326–9.
14. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010;11(11):773–85.
15. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415–25.
16. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009;19(3):212–9.
17. Hatzikotoulas K, Gilly A, Zeggini E. Using population isolates in genetic association studies. Brief Funct Genomics. 2014;13(5):371–7.
18. Wang SR, Agarwala V, Flannick J, Chiang CWK, Altshuler D, Hirschhorn JN, et al. Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. Am J Hum Genet. 2014;94(5):710–20.
19. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet. 2005;37(2):161–5.
20. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A. 2006;103(6):1810–5.
21. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216–20.
22. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23.
23. 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.
24. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered. 2012;73(2):84–94.
25. 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):311–21.
26. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34(2):188–93.
27. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):e1000384.
28. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010;70(1):42–54.
29. 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.
30. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7(3):e1001322.
31. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009;33(6):497–507.
32. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13(4):762–75.
33. 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):110–21.
34. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. Genet Epidemiol. 2013;37(4):334–44.
35. Chen LS, Hsu L, Gamazon ER, Cox NJ, Nicolae DL. An exponential combination procedure for set-based association tests in sequencing studies. Am J Hum Genet. 2012;91(6):977–86.
36. Derkach A, Lawless JF, Sun L. Pooled association tests for rare genetic variants: a review and some new results. Stat Sci. 2014;29(2):302–21.
37. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 2011;12(1):41–51.
38. Cooper GM, Stone EA. Asimenos G; NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A: Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901–13.
39. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score. Condel Am J Hum Genet. 2011;88(4):440–9.
40. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–82.
41. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
42. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
43. Castellana S, Mazza T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. Brief Bioinform. 2013;14(4):448–59.
44. Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, et al. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet. 2012;91(6):1022–32.
45. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86(6):832–8.
46. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB. The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet. 2012;8(2):e1002496.
47. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 2015;11(4):e1005165.

48. Ionita-Laza I, Capanu M, De Rubeis S, McCallum K, Buxbaum JD. Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. PLoS Genet. 2014;10(12):e1004729.
49. Chen H, Hendricks AE, Cheng Y, Cupples AL, Dupuis J, Liu CT: Comparison of statistical approaches to rare variant analysis for quantitative traits. BMC Proc 2011, 5 Suppl 9: S113.
50. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92(6):841–53.
51. Zhi D, Chen R. Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing. PLoS One. 2012;7(2):e31358.
52. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944–50.
53. Wijsman EM. The role of large pedigrees in an era of high-throughput sequencing. Hum Genet. 2012;131(10):1555–63.
54. Epstein MP, Duncan R, Ware EB, Jhun MA, Bielak LF, Zhao W, et al. A statistical approach for rare-variant association testing in affected sibships. Am J Hum Genet. 2015;96(4):543–54.
55. He Z, O'Roak BJ, Smith JD, Wang G, Hooker S, Santos-Cortez RLP, et al. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. Am J Hum Genet. 2014;94(1):33–46.
56. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013;37(2):196–204.
57. Wang L, Choi S, Lee S, Park T, Won S. Comparing family-based rare variant association tests for dichotomous phenotypes. BMC Proc. 2015;9 Suppl 8:S21.
58. Abney M. Permutation testing in the presence of polygenic variation. Genet Epidemiol. 2015;39(4):249–58.
59. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010;42(9):790–3.
60. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010;42(1):30–5.
61. Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, et al. De novo mutations in epileptic encephalopathies. Nature. 2013;501(7466):217–21.
62. Rivière JB, van Bon BW, Hoischen A, Kholmanskikh SS, O'Roak BJ, Gilissen C, et al. De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. Nat Genet. 2012;44(4):440–4. S1-S2.
63. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013;498(7453):220–3.
64. Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, et al. DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods. 2013;10(10):985–7.
65. Taylor PN, Porcu E, Chew S, Campbell PJ, Traglia M, Brown SJ, et al. Whole-genome sequence-based analysis of thyroid function. Nat Commun. 2015;6:5681.
66. Timpson NJ, Walter K, Min JL, Tachmazidou I, Malerba G, Shin SY, et al. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. Nat Commun. 2014;5:4871.
67. Zhu H, Wang Z, Wang X, Sha Q: A novel statistical method for rare variant association studies in general pedigrees. BMC Proc 2015, 9 Suppl 8: S23.
68. Futema M, Plagnol V, Li K, Whittall RA, Neil HA, Seed M, et al. Whole exome sequencing of familial hypercholesterolaemia patients negative for LDLR/APOB/PCSK9 mutations. J Med Genet. 2014;51(8):537–44.
69. Barnett IJ, Lee S, Lin X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genet Epidemiol. 2013;37(2):142–51.
70. Lin DY, Zeng D, Tang ZZ. Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proc Natl Acad Sci U S A. 2013;110(30):12247–52.
71. Derkach A, Chiang T, Gong J, Addis L, Dobbins S, Tomlinson I, et al. Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. Bioinformatics. 2014;30(15):2179–88.
72. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. 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):224–37.
73. Hedges LV, Olkin I. Statistical Method for Meta-Analysis. 1st ed. Orlando: Academic Press Inc; 1985.
74. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet. 2013;93(1):42–53.
75. Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. Nat Genet. 2014;46(2):200–4.
76. Blangero J, Teslovich TM, Sim X, Almeida MA, Jun G, Dyer TD, et al. Omics-squared: human genomic, transcriptomic and phenotypic data for Genetic Analysis Workshop 19. BMC Proc. 2015;9 Suppl 8:S2.
77. Green A, Cook K, Grinde K, Valcarcel A, Tintle N. A general method for combining different family-based, rare variant tests of association to improve power and robustness to a wide range of genetic architectures. BMC Proc. 2015;9 Suppl 8:S18.
78. Jadhav S, Vsevolozhskaya OA, Tong X, Lu Q. The impact of genetic structure on sequencing analysis. BMC Proc. 2015;9 Suppl 8:S19.
79. Ramsay J, Silverman B. Functional Data Analysis. 2nd ed. New York: Springer; 2005.
80. 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. 2013;37(5):409–18.
81. Katsumata Y, Fardo DW. On combining family- and population-based sequencing data. BMC Proc. 2015;9 Suppl 8:S20.
82. Wang S, Fisher V, Chen Y, Dupuis J. Comparison of multi-SNV association tests in a meta-analysis of GAW19 family and unrelated data. BMC Proc. 2015;9 Suppl 8:S22.
83. Wang X, Lee S, Zhu X, Redline S, Lin X. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. Genet Epidemiol. 2013;37(8):778–86.
84. Zhu Y, Xiong M. Family-based association studies for next-generation sequencing. Am J Hum Genet. 2012;90(6):1028–45.
85. Choi S, Lee S, Cichon S, Noethen MM, Lange C, Park T, et al. FARVAT: a family-based rare variant association test. Bioinformatics. 2014;30(22):3197–205.
86. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459–63.
87. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012;44(3):243–6.
88. Zawistowski M, Reppell M, Wegmann D, St Jean PL, Ehm MG, Nelson MR, et al. Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. Eur J Hum Genet. 2014;22(9):1137–44.
89. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799–816.
90. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30.
91. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216–21.