Rare variant association on unrelated individuals in case-control studies using aggregation tests: existing methods and current limitations

Authors:

Simon Boutry[1,2], Raphaël Helaers[1], Tom Lenaerts[2,3,4], Miikka Vikkula[1,5,*]

Affiliations:

[1] Human Molecular Genetics, de Duve Institute, University of Louvain, Avenue Hippocrate 74 (+5) bte B1.74.06, 1200 Brussels, Belgium
[2] Interuniversity Institute of Bioinformatics in Brussels, Université Libre de Bruxelles-Vrije Universiteit Brussels, 1050 Brussels, Belgium
[3] Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium
[4] Artificial Intelligence laboratory, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[5] WELBIO department, WEL Research Institute, avenue Pasteur, 6, 1300 Wavre, Belgium

Corresponding author:

* To whom correspondence should be addressed. Tel: +327647490; fax: +327647460; Email: miikka.vikkula@uclouvain.be

Keywords

Rare variant genetic association studies, GWAS, Case-control samples, Burden test, Variance-components test, Omnibus tests

Key points

- Aggregation tests collapse multiple rare variants within a genetic region (*e.g.* gene, gene set, genomic loci) to test for association. They have successfully identified trait-

associated rare variants and enhanced comprehension of the fundamental disease mechanism.

- We reviewed numerous aggregation tests, presenting their main advantages and drawbacks. We described aggregates methods' specificity given their own underlying assumptions and mathematical models, separating them into 5 classes: Burden, adaptive Burden, Variance-component, Omnibus and others. We also summarize validations and software for such methods.
- We highlight current limitations of aggregation tests: cohort construction, selection of qualifying variants, aggregation unit definition, and test selection. All these having an impact on the success of an aggregation study.
- Aggregation tests capable of managing a broad spectrum of disease mechanisms and can be used to prioritize genetic regions in a complete new way compared to classical bioinformatics tools that only handle single variants.

Author Biographies

Simon Boutry is currently doing a joint PhD at de Duve Institute of the University of Louvain, Brussels, and the Interuniversity Institute of Bioinformatics in Brussels of the Université Libre de Bruxelles, Brussels. His research focuses on the development of a novel genetic regions prioritization to analyze oligogenic diseases.

Raphaël Helaers is a senior bioinformatician at the de Duve Institute of the University of Louvain, Brussels. With a Master of Computer Sciences, he obtained a PhD in bioinformatics in the field of evolutionary biology and is now focused on human genetics and NGS.

Tom Lenaerts is professor of Artificial Intelligence and Computational Biology at the Computer Science Department in the Faculty of Sciences at the Université libre de Bruxelles, Brussels. He obtained his Ph.D. in Computer Science from the Vrije Universiteit Brussel and his research focuses on machine learning and precision medicine, on the one hand, and multi-agent learning and evolution, on the other hand.

Miikka Vikkula is a full professor of Human Genetics at the de Duve Institute of the University of Louvain, Brussels. He obtained his M.D. and Ph.D. in Genetics from University of Helsinki, and his research focuses on Mendelian and multifactorial genetic background of human diseases.

Abstract

Over the past years, progress made in next-generation sequencing technologies and bioinformatics have sparked a surge in association studies. Especially, genome-wide association studies (GWAS) have demonstrated their effectiveness in identifying disease associations with common genetic variants. Yet, rare variants can contribute to additional disease risk or trait heterogeneity. Because GWAS is underpowered for detecting association with such variants, numerous statistical methods have been recently proposed. Aggregation tests collapse multiple rare variants within a genetic region (*e.g.* gene, gene set, genomic loci) to test for association. An increasing number of studies using such methods successfully

identified trait-associated rare variants and led to a better understanding of the underlying disease mechanism. In this review, we compare existing aggregation tests, their statistical features and scope of application, splitting them into the five classical classes: Burden, adaptive Burden, Variance-Component, Omnibus and Other. Finally, we describe some limitations of current aggregation tests, highlighting potential direction for further investigations.

Introduction

The progress in high-throughput next-generation sequencing (NGS) and biostatistics have resulted in a significant increase in the number of associations studies, *e.g.* genome wide association studies (GWAS). To overcome the sparsity and high-dimensional nature of GWAS data, several sophisticated and robust statistical methods have been developed. However, in scenarios where thousands of samples are at our disposal, the efficacy of single-marker association tests can diminish, particularly when confronted with low minor allele frequencies (MAF) and rare variants. Despite the unparalleled potential that sequencing offers for delving into the roles of rare variants in complex diseases, it is important to acknowledge that pinpointing these variants in association studies still remains a significant challenge [1].

The classical strategy to test the connection between genetic variants and complex traits in GWAS is to apply a univariate test (also called single variant test). Multiple testing is taken into account using correction by scaled p-value threshold (e.g. Bonferroni, Benjamini-Hochberg [2], etc.) for declaring significance [3]. Typically, the association between each variant and a trait is assessed through linear regression, particularly for continuous traits (also called quantitative traits). For binary traits (also referred to as case-control data), utilized methods include the Fisher's exact test, χ^2 test, logistic regression and Cochran-Armitage test for trend [4] (CATT) [1, 3]. The latter allowed to identify thousands of trait-associated loci. One of the most popular tests for common variants in GWAS is the univariate minimum p-value (UminP) [5] method that tests separately each SNV and subsequently uses the minimum of their p-values[6]. If some assumptions are met (Sample sizes are sufficiently ample, and effect magnitudes are substantial, MAF not small), these methods can be useful. However, Single-marker examinations exhibit limited ability to identify effects of modest magnitude and collecting large sample cohorts for rare diseases is often unfeasible. Moreover, a single-marker test's power is very sensitive to the effect size, which in most cases is unknown for an individual low-frequency variant [1, 3].

Researchers proposed multiple-marker tests, such as the Hotelling's T^2 test [7] (also referred as TTest test), which represents a multivariate extension of Student's t-test, the ZGlobal statistic

[8] and its variant called Weighted Score Test (WST) [9]. Hotelling's test exhibits a close relationship with the SCORE test [10] Within the framework of logistic regression [6]. The Sum test [11] also performed well under certain situations for common variants[6]. Adaptive combination of p-values for rare variant association testing (ADA) [12] dynamically amalgamates p-values on a per-site basis using MAFs as weights. This approach incorporates a truncation threshold for per-site p-values, strategically employed to mitigate the noise stemming from the inclusion of neutral variants [12]. These methods improve power of multiple moderate single nucleotide polymorphism (SNP) effects [3]. However, the risk allele must be identified at each variant, and the direction of effect affects the power of the test [3]. These methods lose power when There exists solely one robust signal within the genetic region due to their numerous degrees of freedom [3]. To overcome the need of identification of the risk allele, a multivariate distance matrix regression (MDMR) has been proposed [13]. The kernel-based association test (KBAT) [14] does not make any presumptions about the direction of individual SNP impacts and was found to be more powerful than ZGlobal or MDMR methods [3]. A simulation study concerning rare variants demonstrated that multiple marker tests are heavily influenced by the MAF and reduced power when the number of rare causal variants increases in addition to the problem of multiple degrees of freedom [15].

Other techniques that explore the relationship between a trait and numerous variants within a genomic region can enhance statistical power and thereby biological interpretability. In this approach, numerous genetic variants are combined into a single aggregation score. This aggregation score is subsequently used for trait associations instead of assessing each marker separately. This is what aggregation tests do (also referred to as collapsing methods). These approaches apply a solitary univariate test to consolidated data within a cohort, leading to enhanced signals and reduced degrees of freedom, overcoming the multiple correction factors associated with numerous single variant tests or addressing the issue of high degrees of freedom in a multiple-marker test [3, 15]. These tests achieve an increase in power.

In this review, we present existing aggregation methods for rare variant association testing among unrelated individuals, analysed by whole exome sequencing (WES), and using a case-control (binary outcome) design. We present the state-of-the-art methods, their main advantages, drawbacks and some software that implement them. Finally, we highlight some current limitations of aggregation tests.

Existing aggregation methods

Here, we provide a summary of aggregation tests for rare variants (Table 1). The goal of these methods is to overcome the conventional single-variant association techniques, which often lack the necessary power to identify rare variants effectively.

The general statistical model underlying most aggregation tests assumes that *n* individuals have undergone sequencing for a particular genetic region (e.g. genes, WES, WGS, pathways) containing *p* variants. Let us denote by y_i the phenotype of subject *i*, with mean μ_i. For each patient *i*, there are (potentially) *q* covariates $\boldsymbol{X_i} = (x_{i1}, \dots, x_{iq})$, and also $\boldsymbol{G_i} = (g_{i1}, \dots, g_{ip})$ with $g_{ip} = 0, 1\ or\ 2$ denotes the count of the minor allele copies for variant *p* present in the region. Furthermore, let's consider the scenario where y_i conforms to a distribution within the quasi-likelihood family and examine the subsequent generalized linear model [1, 16, 17]:

$$\begin{cases} \mu_i = \alpha_0 + \boldsymbol{X_i}\,\alpha + h(\boldsymbol{G_i})\beta + \varepsilon_i, & continuous\ trait \\ logit(\mu_i) = \alpha_0 + \boldsymbol{X_i}\,\alpha + h(\boldsymbol{G_i})\beta, & binary\ trait \end{cases}$$

Equation 1

Where α_0 is an intercept that aims at modeling the disease prevalence, $\alpha = (\alpha_1, \dots, \alpha_q)^T$ are the regression coefficients for the covariates, $\boldsymbol{X_i}$, $\beta = (\beta_1, \dots, \beta_p)^T$ are the regression coefficients for the allele counts, $\boldsymbol{G_i}$, $h(\cdot)$ is an flexible function characterized solely by a positive semidefinite kernel function $K(\cdot,\cdot)$, and ε_i is an error term. We will show later that equation 1 is a special case (namely a linear kernel). Choices that are more complex can be made. The score statistic of the marginal model for variant j is defined as

$$S_j = \sum_{i=1}^{n} g_{ij}(y_i - \boldsymbol{\mu_i})$$

Equation 2

Where $\boldsymbol{\mu_i}$ is the estimated mean of y_i under the null hypothesis (H_0: $\beta = 0$ and is obtained by application of the null model $\boldsymbol{\mu_i}/logit(\boldsymbol{\mu_i}) = \alpha_0 + \boldsymbol{X_i}\boldsymbol{\alpha}$ [1]. The statistics S_j in equation 2 will be positive if variant *j* increases the risk of disease or trait values, and negative if it decreases it We use the most widely [1, 18-20] used classification of aggregation tests: Burden, adaptive Burden, Variance-component, Omnibus and Other tests, even though some other classifications have been proposed [21].

Burden tests

Burden tests is probably the most famous class of aggregation tests. They amalgamate fata from multiple genetic variants into a singular genetic score, facilitating the assessment of the relationship between these composite scores and a given trait [1, 22]. The classical and basic approach simply count all the numbers of minor alleles present across all variants within the region of interest. Therefore, the summary genetic score is

$$C_i = \sum_{j=1}^{p} w_j g_{ij}$$

Equation 3

Where w_j is the weight for a variant j. In comparison with the previous model, this is identical to putting $\beta_j = w_j\beta$ in the regression model in equation 1 and testing $H_0: \beta = 0$ in the simplified model $\boldsymbol{\mu_i}/logit(\boldsymbol{\mu_i}) = \alpha_0 + \boldsymbol{X_i\alpha} + C_i\,\boldsymbol{\beta}$ [1]. The corresponding score statistic to test $H_0: \beta = 0$ is then

$$Q_{Burden} = \left(\sum_{j=1}^{p} w_j S_j\right)^2$$

Equation 4

A p-value can be acquired by comparing it to a chi-square distribution with a single degree of freedom [1, 20]. The summary genetic score C_i see equation 3 can be modified in order to accommodate various presuppositions regarding the mechanism of the disease. This is the case, for example, of the combined and multivariate collapsing (CMC) [15] or the cohort allelic sums test (CAST) [23]. They both assume that the inheritance mode is collectively dominant [20]. The CAST makes the hypothesis that the existence of any rare genetic variant augments the risk of the disease. Therefore, considering a dominant model, the test sets the genetic score $C_i = 0$ if no minor alleles are present in the region and $C_i = 1$ otherwise. The accumulation of rare variants integrated and extended locus-specific extensions test (ARIEL) [24] broadens burden score to accommodate variability in the quality scores of the variants. The rare variant tests (RVT) [25, 26] (also referred to as MZ tests [1]) is an example of collapsing methods within a regression framework. These approaches characterize the phenotype by employing a summarized representation of variants in one of two manners: RVT1: the fraction of rare

variants that bear at least one copy of the minor allele for every individual; RVT2: whether each individual possesses or lacks at least one minor allele across any rare variant [3].

Classical burden tests make use of the regression method. Yet, some have been constructed outside this framework. This is the case of the CMC test above. CMC aggregates rare variants, akin to CAST, yet with two main differences. It uses distinct MAF categories and appraises the collective impact of both rare and common variants using Hoteling's test [1], which improves the robustness of inclusion of non-causal variants. Another alternative to regression is to obtain p-values by permutation. This is done in the weighted-sum test (WSS) [27] (also referred to as the weighted-sum collapsing approach (WSC) [20], or the w-Sum test [6]). The latter uses the Wilcoxon rank-sum test and posits that the magnitude of the effect is inversely related to the MAF [20]. For CATT, the assumption is that the probability of being disease causing rises in tandem with the count of rare minor alleles, rendering the counts of minor alleles suitable for consideration as ordered categories [3]. A contingency table is used to juxtapose the minor allele count (MAC) between case and control groups. The cumulative minor-allele test (CMAT) [28] takes into consideration the uncertainty introduced by imputation methods on genotypes. This test holds potential for reevaluating established GWAS datasets. Finally, RareCover [29] utilizes a selective approach employing a greedy method to choose a subset of rare variants using forward variable selection that best associates with the phenotype.

While most methods described above are exclusive to case-control data (binary traits), the regression approaches [26] could also handle continuous phenotypes [3]. The strongest assumption behind all Burden tests is that every rare variants within the genetic region of interest are causative and linked to the trait of interest with the same direction and magnitude of effect [22, 30]. If that hypothesis is violated, this could lead to a significant reduction in statistical power.

Adaptive Burden tests

To overcome the natural limitation of the burden tests, several authors have proposed adaptive burden tests. These methods are robust and can handle null variants and both trait-increasing and trait-decreasing variants [20]. For example, Han & Pan [31] proposed five versions of a new adaptive sum (aSum) test made of five steps. It begins with an estimation of the possible direction of effect for each variant. It assigns $w_j = -1$ when β_j is probably negative and $w_j = 1$ otherwise [1]. If w_j is negative, the corresponding SNP coding is reversed. The resulting set of coding SNPs is then employed in fitting the common-effect model [3]. It computes test

statistics, using the estimated directions as weights. The authors recommend to use the aSumC-P test [31]. The latter collapses rare and common variants into two groups, then tests on the two associated regression coefficients within a logistic regression model [31]. In order to compute p-values, this version of the aSum test uses permutation. The aSum+ uses a more constrained form of variable selection, in contrast to the aSum that uses direct variable selection [32]. In order to enhance the performance compared to the aSum+ test, an adaptive Sum test that relies on two directional search approaches (aSum2) [33] (also called aSum2d [34]) proposes utilizing both positively and negatively associated variants. It was found to improve power in detecting gene–gene interactions for common variants [34].

Authors [32] have also proposed an adaptive score test (aScore). The p-value weighted sum test (PWST) [35] weights variant individually, including both direction and significance of individual variant impacts. These are used to calculate a unified weighted sum score derived from rescaled left-tail p-values obtained through single-variant analysis. Subsequently, a permutation test of association is conducted between this score and the given trait [35]. The comprehensive approach to analyzing rare variants (CARV) [36] selects the optimal combination of rare variants and aggregates them into a single group. After testing three approaches (hard, variable and step-up), the agnostic step-up method achieving an optimal grouping of rare variants without relying on preexisting assumptions was recommended [36]. It refines the procedure of the aSum and allows to assign $w_j = 0$ when a variant is unlikely to be associated with a trait [1]. Another data adaptive test is the estimated regression coefficient (EREC), which assigns weights by estimating a regression coefficient of each variant [37]. The idea behind EREC is that true regression coefficient β_j serves as an optimal weight to maximize power [1]. However, the latter is difficult to achieve for rare variants. Indeed, β_j estimates become unstable when the MAC is low [37]. In this case, EREC tries stabilizing the estimates by introducing a slight adjustment to the estimated β_j, which could potentially diminish the optimality of EREC and make it behave more like a burden test (not data adaptive anymore) [37]. The p-values are estimated using parametric bootstrap [1]. aSum, PWST and EREC are computationally intensive because during each permutation, they iteratively compute the marginal estimates for the coefficient β_j [34].

As a third way to improve classical burden tests, the variable threshold (VT) [38] is based on the assumption that the link between allele frequency and effect size might vary greatly with the intensity of selection (i.e. all variants with MAF below the threshold have the same effect size [20]). VT test choses optimal frequency thresholds for conducting burden tests on rare

variants and assesses statistical significance by permutation [38]. The rare variant weighted aggregate statistic (RWAS) [39] computes optimal weights based on the assumption of a constant population-attributable risk for all variants [20]. The kernel-based adaptive cluster (KBAC) method [40] first classifies variants, and then tests for association using kernel-based adaptive weighting. Here, covariates (e.g. sex, population stratification, etc.) can be taken into account. KBAC has a greater power than other rare variant analysis techniques like CMC and WSS, even in scenarios involving variant misclassification and gene interactions, KBAC is particularly advantageous [40]. Finally, the summation of partition approach (SPA), a robust model-free method also exists. It was deliberately tailored for the detection of marginal effects as well as effects arising from gene-gene and gene-environmental interactions within the context of rare variants association studies [41].

This section shows that adaptive burden tests constitute an improved version of the classical burden methods. Indeed, they rely on fewer assumptions concerning the underlying genetic model. However, as limitations, one can observe that estimating the regression coefficients for individual variants is frequently challenging and prone to instability for rare variants, and the permutation needed by most adaptive tests in order to estimate p-values makes them computationally intensive [1]. Many adaptive tests have been compared and they share power similar to that of variance-component and combined tests, which will be discussed next [6].

Variance-Component tests

Variance-Component tests were introduced in order to allow flexibility in regard to two main hypothesis concerning the variants: direction of effect and effect size. These tests are able to handle both protective and deleterious variants, as well as variants with effect sizes of different magnitude [20]. This new class of methods, instead of grouping rare variants of a region of interest, uses random-effect models. These methods ascertain association by examining the distribution of genetic effects across a set of variants[1]. Tests that are part of this class include C-alpha [42], the sum of square score (SSU) test [43], and the sequence kernel association test (SKAT) [17, 30]. They assess the distribution of the aggregated score test statistics (eventually using weights) of each variant [1]. If there is no covariate, SKAT reduces to a C-alpha test. In contrast, SKAT can also handle SNP-SNP interactions [30]. Going back to equation 1, a random effect model states that regression coefficients β_j adhere to a distribution characterized by a mean of 0 and variance $w_j^2\tau$ [20]. In this case, the test hypothesis for association reduces to $H_0: \tau = 0$ by using a variance-component score. As an example, the SKAT test statistic is given by

$$Q_{SKAT} = \sum_{j=1}^{p} w_j^2 S_j^2$$

Equation 5

Equation 5 is a summation of squared single-variant score statistics S_j, with each term weighted accordingly see equation 2 [1]. Contrary to burden tests (see equation 4), SKAT aggregates S_j^2 instead of S_j, which makes it able to handle both protective variants and deleterious variants. Different kernels allow one to model different genetic models. Therefore, the more flexible and adaptable the kernel is, the more complex and different problems (e.g. different diseases, with different underlying genetic models) can be tackled. The element of interest is the function $h(\cdot)$ in equation 1. It is the most important element of the equation as it is the one that dictates how variants are taken into account for disease risk. We detail here the different kernel choices available in the SKAT package. Suppose $h_i = h(G_i) = \sum_{l=1}^{n} K(G_i, G_l)$, and in order to simplify and illustrate the following, the score statistic given by equation 5 will be rewritten, using matrix notation, thanks to equation 2 as

$$Q_{SKAT} = (y - \boldsymbol{\mu_i})^T K (y - \boldsymbol{\mu_i})$$

Equation 6

The kernel function *K*, is an $n \times n$ matrix. The function $K(G_{i_1 j}, G_{i_2 j})$ measures the genetic similarities between individual i_1 and i_2 within the region of interest thanks to the p variants taken into account [1, 17, 30]. By choosing different kernel functions, it is possible to define new bases and association models [44]. A detailed description of kernels can be found in Supplementary Data.

The classic SKAT could lose power if the sample size or the coefficient of variation of the kernel spectrum is small. The power loss is due to ignoring the uncertainty in the error variance estimate. Several versions of adjusted SKAT for different types of outcomes exist to improve the power of SKAT in unfavorable situations. They circumvent the difficulty to estimate error variance under small sample sizes by deriving a scale-free statistic similar to F-statistic. For univariate continuous and binary outcomes, the adjusted SKAT (aSKAT) [45] increased the power in microbiome association studies, although the increase in power for human genetic association studies was constrained. The multivariate continuous outcome (mSKAT) [46] allows one to assess the combined association of rare variant collections with numerous traits. The correlated sequence kernel association (cSKAT) test [47] is more powerful in case of

correlated continuous outcome. The Recalibrated Lightweight SKAT (RL-SKAT) [48] demonstrated that this adjustment could additionally boost statistical power for extensive sample sizes [19], but more work remains to be done to extend the method to multiple kernels and binary phenotypes. A conditional asymptotic distribution has also been introduced for the kernel association test (KAT) [49], which can be used as a screening tool followed by comprehensive permutations (classical SKAT) at genes that display signals. For SKAT, all subjects are used for estimation of the null genetic covariance. The SKAT+ [50] employs identical test statistics as SKAT, but it distinguishes itself through its approach to estimate the null distribution, using only control subjects. For multivariate traits [19], results [51] indicate that no single approach consistently possesses superior power across the methods. The most suitable test hinges on factors such as the extent of phenotype correlation and the nature of effect patterns. However, the multivariate kernel machine regression (MV-KM) [51] seems to be a reasonable approach. When confronted with a substantial quantity of neutral rare variants, SSU and aSSU are more robust [32]. The Bayesian Score Test (BST) [52] is a permutation-based method equivalent to SSU despite its reliance on an empirical Bayesian model featuring an independent prior for genetic variant effects [6].

Large-sample-based p-value computation might produce inaccurate type I error rates if sample sizes or total MAC are small [1]. In this situation, if the number of cases and controls are equal, the false positive rate can be underestimated. On the other hand, if they are not equal, the latter can be overestimated [44]. To address the latter problem, a moment-based method capable of adjusting the asymptotic null distribution using estimates of the exact small-sample variance and kurtosis of the test statistic was developed [1, 44]. Despite this, if the MAC is very low, obtaining accurate p-value estimates might require bootstrap or a permutation approach.

A comprehensive examination of kernel methods is conducted in an in-depth review [19], delving further into aspects such as hypothesis testing, extensions for various traits, and the intricate design of underlying kernels. This review [19] establishes linkages between kernel tests and other statistical methodologies.

Omnibus tests

Also called combination tests, omnibus tests combine burden and variance-component tests. The main idea behind them is to aggregate evidence from multiple complementary sources [20]. On one hand, variance-component tests exhibit greater statistical power when compared to burden tests if the region of interest has many non-causal variants or if the causal variants are

both protective and deleterious. On the other hand, burden tests tend to exhibit greater statistical power compared to variance-component tests when the region of interest contains a significant proportion of causative variants that share the same association direction [1]. Because both previously described situations can happen, omnibus tests can be more robust. For example, a method was designed to combine p-values of the two tests (burden and variance-component) with Fisher's method, and a permutation in order to compute the significance of the test [20, 53]. The Fisher statistic takes the form $Fisher = -2\log(p_{Burden}) - 2\log(p_{SKAT})$. Where p_{Burden} and p_{SKAT} are p-values obtained from Burden and SKAT tests, respectively. Another approach is to use convex combination of burden test and the SKAT statistics with a predetermined weight coefficient or adaptive data. The optimal SKAT (SKAT-O) is a linear combination in the following form $Q_\rho = \rho Q_{Burden} + (1-\rho)Q_{SKAT}, 0 \leq \rho \leq 1$ [22, 44]. Where parameter ρ can be viewed as a pairwise correlation among the genetic effect coefficients β_j in equation 1 [1]. The parameter ρ is likely to be unknown. Therefore, SKAT-O computes an approximate optimal weight coefficient, $0 \leq \rho \leq 1$ heuristically. It will evaluate p-values over a range of ρ values and selects the value that results in stronger evidence of an association (i.e. the minimum p-value computed over the grid) [20]. One of the strength of the SKAT-O relies in the fact that its asymptotic p-value can be computed efficiently with one dimensional numerical integration tools [1, 20].

The evolutionary mixed model for pooled association testing (EMMPAT) [54] is similar to the SKAT-O. EMMPAT uses a hierarchical model for rare variant associations [48]. An important advantage is that it allows the incorporation of known characteristics of variants [20]. This method loses power if the variant annotations are not correlated with the variant effect size.

Another example of hierarchal model such as EMMPAT is implemented in the mixed effects score test (MiST) [55] that has the The ability to discern which aspects of variant properties and heterogeneity contribute to the observed association. The sum of powered score (SPU) tests [34] generalize the sum test [43], a representative burden test and the SSU, a variance-component test. SPU may be useful in the presence of many non-associated rare variants [34].

The variant-set mixed model association tests (SMMAT) [56] are computationally efficient and scalable for WGS analysis. SMMAT comes in four forms, the burden test (SMMAT-B), SKAT (SMMAT-S), SKAT-O (SMMAT-O), as well as the recommended version (SMMAT-E). This test combines the burden test and an adjusted mixed model SKAT statistic maintains asymptotic independence from the mixed model burden test statistics [56], such as the MiST in

non-mixed model setting [55]. DoEstRare [21] combines a burden test along with a position test comparing rare variant position distribution and average allele frequencies of a region. The position test computes variant position densities within the genetic region for both cases and controls using a kernel method, where the weights are determined by a function of allele frequencies [21].

Omnibus tests reach robust power and outperform Burden or variance-component tests in a broad spectrum of situations. However, if the assumptions behind them (burden or variance-component) are largely true, omnibus tests can be less powerful. Yet, as investigators seldom possess prior knowledge regarding the underlying genetic structure, omnibus tests constitute an attractive choice. As a global remark on combination tests, the naive procedure of picking the minimal p-value of different methods will most of the time yield to an inflated type I error rate [1].

Other methods

There are still other methods that do not fit the above categories. As seen previously, burden and variance-component tests use linear and quadratic sums of S_j respectively, whereas, for example, the exponential-combination (EC) test [57], uses an exponential sum of S^2_j. The EC test is based on the assumption that only one or very few variants within a genetic region are causal [57]. The EC test statistic is as follow $Q_{EC} = \sum_{j=1}^{p} \exp\left(\frac{S_j^2}{2\,var(S_j)}\right)$ where S_j is defined by equation 2. If the assumption fits well the unknown underlying genetic model, the test will outperform the others. If only an exceedingly minor fraction of variations are causing the disease, EC achieve higher power than the above tests, thanks to the exponential function increasing very quickly as S_j^2 increases [1]. Obviously, if too many variants are associated with the trait, the EC test will perform poorly. Moreover, in order to estimate p-values, the permutations method is required as the null distribution of Q_{EC} is not known.

Another method that depends on sparsity of the signal is the least absolute shrinkage and selection operator (LASSO) and penalized regression [58]. The replication-based test (RBT) [59] (also referred to as RB [41]) uses a replication-based strategy, which is based on a weighted-sum statistic to measure: the enrichment of rare alleles in cases (trait-increasing alleles); and controls (trait-decreasing alleles) [20]. This method holds the benefit of being less influenced by the existence of both protective and risk variants within a genetic region [59]. In the past, RBT has been classified as an adaptive burden test [1]. The Bayesian risk index (BRi)

[60] incorporates model uncertainty when deciding which variants to include in the index, along with the direction of their associated effects.

Nengjun Yi and Degui Zhi introduced a novel Bayesian hierarchical generalized linear model (we refer to it as the YZ test) that uses prior distributions to specify aspects of the distribution of effect sizes [61]. This method allows dealing with disparate effects and non-functional variants. However, Bayesian approaches [60, 61] have not been as popular as frequentist approaches for rare variants because of computational issues [20].

Methodology for aggregation tests benchmarking

We give an overview of the methods see Table 2, their scope and data used to validate them by extending the table proposed in [6]. There are three steps to establish the validity of a new aggregation method. First of all, one should simulate data to mimic the population genetic samples under a given demographic model (*e.g.* using a home-made design, available software/resources [62-69] or reuse design from existing methods [6, 9, 15, 27, 37, 43, 44, 70, 71]) and build a simulation pipeline to analyze such data. The second step is to select state-of-the-art method to benchmark the new aggregation test. One can see that most methods have never been compared see Table 3. Last step is to apply the new method on real data when available [72-89]. There is no consensus on how to perform these three steps.

Software for aggregation tests

Most articles introducing a new method for aggregation provide an implementation freely available. The complexity of rare variant analyses makes it difficult to establish a global framework able to meet all analysis needs for each study. In such situations, investigators might find it desirable to implement custom-designed approaches to carry out specific analyses [90]. While reviewing available software [17, 21, 22, 30, 37, 41, 44-51, 55, 91-98] see Table 4, we came across some outdated packages. These previously described programs [19], are no longer available: the SPA3G CRAN package for gene-gene interaction analysis for continuous phenotypes using kernels [99]; the iSKAT, implementing the GESAT for gene-environment interaction kernel [19] (note that the GESAT method can be available by contacting the authors [100]); FAmily-based rare variant association test for gene-based association test with related samples (FARVAT) [101]; and the KMgene CRAN package [102].

General comments

We reviewed numerous aggregation tests, illustrating the fast evolution of the field. Table 2 aggregates methods' specificity given their own underlying assumptions and mathematical

models. It shows that the robustness and power of most methods tend to be validated based on simulated data given very specific numerical simulations that tend to model the best scenario for the method under study [103]. It is clearly necessary to establish a general simulation framework to enhance our knowledge on the behavior of the methods, as proposed by several authors [6, 11, 104-106]. It is necessary to rigorously compare the performance of different methods using well-established benchmark datasets, to evaluate the respective strengths and to conduct solid benchmarking studies [107, 108]. We focused on methods for binary traits, but a vast amount of aggregation tests have been developed outside this framework (*e.g.* Family-based association test [101, 109], pedigree data [19, 95], quantitative traits [51], interaction testing [19], and haplotype-based methods [105]). The latter might be more useful in frequent disorders where common variants are looked for, compared to rare disease research in which private variants are usually searched. Other methods reuse the statistical methods described above, adapting them for specific sequencing technologies. For example, WGScan test [110] has been specifically designed for whole genome sequencing (WGS) data. The three version of WGSscan are based on classical CMC, SKAT and SKAT-O methods, respectively. Some methods try to improve existing methods by repeating the same idea but on different cluster of variants. For example, the subregion-based burden test (REBET) [111] performs a classical burden test and then splits the significant regions (*e.g.* genes) based on variable criteria (*e.g.* the same biological function, similar functional impact) into all possible combinations of subregions to find the combination that shows the strongest evidence of association.

Current limitations of aggregation tests

The outcome of an aggregation study relies on several factors that influence the range of detectable effect sizes [112]. Most methods have been developed in order to fit a particular set of parameters (*i.e.* set of answers to the questions below). We focus on methods for binary trait (cases versus controls), using WES data and rare variants. An evaluation of the methods dedicated to related samples is available elsewhere [113]. There are no clear-cut answers to the questions below and further investigations regarding the behavior of aggregation tests is needed, as it has been shown that they have a non-negligible impact on the association results [1, 16, 90, 104, 106, 113, 114].

How to construct the optimal cohort?

Two main points have been raised regarding cohort construction for association studies, namely the proportion of cases versus controls and the cohort size.

Across most aggregation methods (see Table 2), simulation is based on the same number of cases and controls, while some real data applications are based on unbalanced cohorts. In the presence of imbalance between the numbers of cases and controls, methods that rely on asymptotic properties cannot be used [114]. Further investigations are required to evaluate the impact of unbalanced cohort on type I error and power of these methods.

The second limitation comes from the cohort size. It requires a large amount of individuals to observe enough copies of rare variants [90]. As an illustration, to have a 99% probability of sampling alleles with a frequency of 0.5% or 0.05%, sequencing at least 460 or 4,600 individuals, respectively, is necessary [1]. Comparative studies showed that the power of detection ranges from 5 to 20% for a cohort size of 3K, while using a 10K cohort, the power was around 60% [106]. Background variations, which vary across the genome, directly affect sample sizes needed to detect associations of a genetic region [104]. Locus heterogeneity and mode of inheritance are primary drivers determining the needed sample size [104].

How to include qualifying variants?

This general question should be investigated in regard to four dimensions on the qualifying variants: How to weight them, handle protective effects, handle non-causal variants and set an optimal MAF threshold?

The first dimension only concerns methods able to include a weighting scheme within their mathematical model (see Table 2). Setting an optimal weighting procedure for a particular association study is not feasible in practice, because it requires to know the effect of each qualifying variant within the disease model. It is recommended to check the ratio of potential weights before applying them to variants to ensure that the range is compatible [90]. For example, it is rare that the weight of one variant is hundreds of times higher than another one, and that can lead to false positives (*e.g.* if the up weight variant is non-causal). The classical weighting scheme [27] for aggregation tests, using only the MAF, has already been proven to be powerful. As an extension of that, possible weighting of variants based on prior knowledge such as functional prediction scores [90] can be used.

The second issue, the presence of protective variants, gave birth to a new class of aggregation tesst (see Variance-component tests) able to handle both risk-increasing and risk-decreasing variants [1]. Any association study willing to address this issue must therefore focus on methods able to take the direction of effect into account in their mathematical model (see Table 2).

Regarding the third dimension, increasing sample sizes (from 3000 to 10 000 individuals) also increases the number of neutral variants, which may limit gains of power for some methods [106]. As a consequence, non-causal variants may dilute the association signal. It is recommended to use methods that are robust for the inclusion of such variants (*e.g.* CMC, SKAT-O) [90]. While studying the correlation between statistical power and the filters employed to prioritize variants that have a higher likelihood of being pathogenic, less stringent filters may be advantageous [104]. This merits further study.

How to set an optimal MAF threshold is closely linked to establishing a weighting procedure, in the sense that it will require knowing in advance which variants are involved in a particular disease mechanism. Classical GWAS focuses on the identification of common variants. The idea of focusing on variants below a MAF threshold (*e.g.* $MAF < 1\%$), is that these might elucidate supplementary disease susceptibility or variability in trait expression [1]. Although MAF resolution depends on the database used, the strongest signals from this type of association analyses are concentrated among the rarest variants [114]. Future work should characterize aggregation methods given different MAF threshold [106].

Which genetic region for aggregation unit?

One of the first decision in order to perform association studies is selecting a particular unit, or genetic region in which all qualifying variants will be aggregated (*e.g.* genes, pathways, GO terms). These units are established based on genomic coordinates, gene annotations or functional characterization [16], and therefore depend on the annotator tool or database used. To reduce the impact of multiple testing correction (assuming 20 000 genes in the human genome, a Bonferoni correction would lead to an $\alpha = 2.5 \times 10^{-6}$ as a significance threshold in an exome-wide search), it might be of interest to restrict an analysis to a gene list of interest. Constructing the latter raise other issues and considerations, such as the need to construct a curated gene list for the disease being studied, which are well beyond the scope of this review [115]. However, gene-sets have been used as testing unit [116]. These also rely on databases such as MSigDB [117], Reactome [118] and KEGG for pathways [119], or the Gene Ontology Resource for GO terms [120]. However, such genetic regions might include an excessive total number of variants leading to both computational time explosion and association signal dilution. It is expected that only a few genes within these gene-sets truly harbor causal variants [90].

Which aggregation tests to use?

All the above questions can be summarized in one: which aggregation tests to use? The comparative efficacy of aggregation techniques is contingent upon the underlying disease architecture, which is typically unknown [1]. In cases where it is anticipated that a genetic region contains a substantial proportion of causal rare variants, with the majority of them promoting disease risk, Burden tests are expected to exhibit higher statistical power [1]. However, for other association studies, it may be challenging to properly answer the above question. A current recommendation is to try multiple methods and subsequently adjust p-values while considering the utilization of diverse methods. This strategy is employed to prevent the inflation of type I errors. Alternatively, one can consider employing an Omnibus test, which is anticipated to possess robust statistical power across a broader spectrum of disease models [1]. No single approach outperforms all others in every situation [90], and results may vary based on the specific tests that are used [106]. All this indicates the difficulty for researchers, aiming to apply aggregation tests, to select the optimal statistical test for a particular problem at hand.

Conclusions and future directions

In this review, we have focused on rare-variant aggregation tests for testing association among unrelated individuals, sequenced by WES, and using a case-control (binary outcome) design. An extensive literature for other settings exists (*e.g.* family-based [113], quantitative traits [121]). Most of the observations and conclusions drawn here may directly apply to other sequencing technologies (*e.g.* WGS [16]), although, each design comes with its own challenges.

In the past few years, the decreasing cost of sequencing has facilitated data procurement. These databases, along with modern statistical methods, have allowed to overcome the classical limitations of GWAS imposed by low allelic frequency. Therefore, the literature on association studies has mainly focused on such variants, which has led to the explosion in the number of methods available to perform such analysis. These aggregation tests have increased sensitivity and constitute a new powerful way to explore genetic data in order to improve our understanding of a wide range of disease mechanisms (*e.g.* Mendelian, oligogenic and complex). They already have proven to be useful to uncover genetic regions associated with many traits (*e.g.* see Table 2, real data), although the identification of causal variants must yet to be documented for the majority of the studies [105]. Indeed, a challenging task will be to establish biological methodologies to confirm (*e.g.* in vitro, in vivo) such discoveries.

Nonetheless, aggregation tools can be used to prioritize genetic regions in a complete new way compared to classical bioinformatics tools that only handle single variants.

Our growing understanding of the benefits of each class of methods (*e.g.* underlying assumptions behind the models, type I error, power, computational efficiency), have facilitated the informed choices about which aggregation test to use for a particular scenario. This is imperative to ensure minimizing false positive discoveries [18]. Developing new methods for the analysis of rare variants should not prevent future work to focus on comparing already existing ones. Indeed, the robustness and power of the statistical models still needs to be assessed and compared in a wide variety of contexts in order to enhance our ability to choose the right statistical test for a problem at hand [103]. An important field of investigation to overcome current limitations in aggregation tests consists of follow-up studies and meta-analysis [18, 122] that were not discussed here.

Consent for publication

All authors acknowledge the content of this manuscript and grant permission for its publication.

Author's contributions

S.B. performed literature review and designed the review. S.B. wrote the draft an revised the manuscript. All authors reviewed the manuscript.

Acknowledgements

The authors thank all members of the laboratory of Human Molecular Genetics and members of the oligogenic team at the Interuniversity Institute of Bioinformatics in Brussels for their supports and feedbacks. The studies in the laboratory received financial support from the Fund Generet managed by the King Baudouin Foundation (Grant 2018-J1810250-211305), the Fonds de la Recherche Scientifique (FNRS Grants T.0026.14 & T.0247.19), and by la Région wallonne dans le cadre du financement de l'axe stratégique FRFS-WELBIO (WELBIO-CR-2019C-06) (all to MV). We also thank the Foundation against Cancer (2010-101), Belgium and the National Lottery, Belgium for their support to the Genomics Platform of University of Louvain and de Duve Institute, as well as the Fonds de la Recherche Scientifique - FNRS Eguipment (Grant U.N035.17) for the «Big data analysis cluster for NGS at UCLouvain». S.B.

was supported by fellowships from F.R.I.A. (Fonds pour la formation à la recherche dans l'industrie et dans l'agriculture), and Patrimoine UCL.

Funding

This work received financial support from the Fund Generet managed by the King Baudouin Foundation [Grant 2018-J1810250-211305]; the Fonds de la Recherche Scientifique -FNRS Grants [T.0026.14, T.0247.19]; and by la Région wallonne dans le cadre du financement de l'axe stratégique FRFS-WELBIO [WELBIO-CR-2019C-06] (all to MV).

References

1. Lee, S., et al., *Rare-variant association analysis: study designs and statistical tests.* Am J Hum Genet, 2014. **95**(1): p. 5-23.
2. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society, 1995. **Serie B (Methodological)**(No. 1): p. 289-300.
3. Asimit, J. and E. Zeggini, *Rare variant association analysis methods for complex traits.* Annu Rev Genet, 2010. **44**: p. 293-308.
4. Li, B. and S.M. Leal, *Discovery of rare variants via sequencing: implications for the design of complex trait association studies.* PLoS Genet, 2009. **5**(5): p. e1000481.
5. Conneely, K.N. and M. Boehnke, *So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests.* Am J Hum Genet, 2007. **81**(6): p. 1158-68.
6. Basu, S. and W. Pan, *Comparison of statistical tests for disease association with rare variants.* Genet Epidemiol, 2011. **35**(7): p. 606-19.
7. Xiong, M., J. Zhao, and E. Boerwinkle, *Generalized T2 Test for Genome Association Studies.* Am. J. Hum. Genet., 2002. **70**: p. 1257–1268.
8. Daniel J. Schaid, S.K.M., Scott J. Hebbring, Julie M. Cunningham, and a.S.N. Thibodeau, *Nonparametric Tests of Association of Multiple Genes with Human Disease.* Am. J. Hum. Genet., 2005. **76**: p. 780–793.
9. Wang, T. and R.C. Elston, *Improved Power by Use of a Weighted Score Test for Linkage Disequilibrium Mapping.* Am J Hum Genet, 2007: p. 353-360.
10. Clayton, D., J. Chapman, and J. Cooper, *Use of unphased multilocus genotype data in indirect association studies.* Genet Epidemiol, 2004. **27**(4): p. 415-28.
11. Chapman, J. and J. Whittaker, *Analysis of multiple SNPs in a candidate gene or region.* Genet Epidemiol, 2008. **32**(6): p. 560-6.
12. Lin, W.Y., et al., *Rare variant association testing by adaptive combination of P-values.* PLoS One, 2014. **9**(1): p. e85728.
13. Wessel, J.a.S., N. J., *Generalized Genomic Distance–Based Regression Methodology for Multilocus Association Analysis.* Am J Hum Genet, 2006. **79(5)**: p. 792-806.
14. Mukhopadhyay, I., et al., *Association tests using kernel-based measures of multi-locus genotype similarity between individuals.* Genet Epidemiol, 2010. **34**(3): p. 213-21.
15. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet, 2008. **83**(3): p. 311-21.
16. Auer, P.L. and G. Lettre, *Rare variant association studies: considerations, challenges and opportunities.* Genome Med, 2015. **7**(1): p. 16.
17. Wu, M.C., et al., *Powerful SNP-set analysis for case-control genome-wide association studies.* Am J Hum Genet, 2010. **86**(6): p. 929-42.

18. Weissenkampen, J.D., et al., *Methods for the Analysis and Interpretation for Rare Variants Associated with Complex Traits.* Curr Protoc Hum Genet, 2019. **101**(1): p. e83.
19. Larson, N.B., J. Chen, and D.J. Schaid, *A review of kernel methods for genetic association studies.* Genet Epidemiol, 2019. **43**(2): p. 122-136.
20. Nicolae, D.L., *Association Tests for Rare Variants.* Annu Rev Genomics Hum Genet, 2016. **17**: p. 117-30.
21. Persyn, E., et al., *DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease.* PLoS One, 2017. **12**(7): p. e0179364.
22. Lee, S., M.C. Wu, and X. Lin, *Optimal tests for rare variant effects in sequencing association studies.* Biostatistics, 2012. **13**(4): p. 762-75.
23. Morgenthaler, S. and W.G. Thilly, *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): p. 28-56.
24. Asimit, J.L., et al., *ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data.* Hum Hered, 2012. **73**(2): p. 84-94.
25. Morris, A.P. and E. Zeggini, *An evaluation of statistical approaches to rare variant analysis in genetic association studies.* Genet Epidemiol, 2010. **34**(2): p. 188-93.
26. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.
27. Madsen, B.E. and S.R. Browning, *A groupwise association test for rare mutations using a weighted sum statistic.* PLoS Genet, 2009. **5**(2): p. e1000384.
28. Zawistowski, M., et al., *Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes.* Am J Hum Genet, 2010. **87**(5): p. 604-17.
29. Bhatia, G., et al., *A covering method for detecting genetic associations between rare variants and common phenotypes.* PLoS Comput Biol, 2010. **6**(10): p. e1000954.
30. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet, 2011. **89**(1): p. 82-93.
31. Han, F. and W. Pan, *A data-adaptive sum test for disease association with multiple common or rare variants.* Hum Hered, 2010. **70**(1): p. 42-54.
32. Pan, W. and X. Shen, *Adaptive tests for association analysis of rare variants.* Genet Epidemiol, 2011. **35**(5): p. 381-8.
33. Pan, W., S. Basu, and X. Shen, *Adaptive tests for detecting gene-gene and gene-environment interactions.* Hum Hered, 2011. **72**(2): p. 98-109.
34. Pan, W., et al., *A powerful and adaptive association test for rare variants.* Genetics, 2014. **197**(4): p. 1081-95.
35. Zhang, Q., et al., *A data-driven method for identifying rare variants with heterogeneous trait effects.* Genet Epidemiol, 2011. **35**(7): p. 679-85.
36. Hoffmann, T.J., N.J. Marini, and J.S. Witte, *Comprehensive approach to analyzing rare genetic variants.* PLoS One, 2010. **5**(11): p. e13584.
37. Lin, D.Y. and Z.Z. Tang, *A general framework for detecting disease associations with rare variants in sequencing studies.* Am J Hum Genet, 2011. **89**(3): p. 354-67.
38. Price, A.L., et al., *Pooled association tests for rare variants in exon-resequencing studies.* Am J Hum Genet, 2010. **86**(6): p. 832-8.
39. Sul, J.H., et al., *An optimal weighted aggregated association test for identification of rare variants involved in common diseases.* Genetics, 2011. **188**(1): p. 181-8.
40. Liu, D.J. and S.M. Leal, *A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions.* PLoS Genet, 2010. **6**(10): p. e1001156.
41. Fan, R. and S. Lo, *A Robust Model-free Approach for Rare Variants Association Studies incorporating Gene-Gene and Gene-Environmental interactions.* PLoS One, 2013. **8(12)**: p. e83057.

42. Neale, B.M., et al., *Testing for an unusual distribution of rare variants.* PLoS Genet, 2011. **7**(3): p. e1001322.
43. Pan, W., *Asymptotic tests of association with multiple SNPs in linkage disequilibrium.* Genet Epidemiol, 2009. **33**(6): p. 497-507.
44. Lee, S., 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): p. 224-37.
45. Chen, J., et al., *Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies.* Genet Epidemiol, 2016. **40**(1): p. 5-19.
46. Wu, B. and J.S. Pankow, *Sequence Kernel Association Test of Multiple Continuous Phenotypes.* Genet Epidemiol, 2016. **40**(2): p. 91-100.
47. Zhan, X., et al., *A small-sample kernel association test for correlated data with application to microbiome association studies.* Genet Epidemiol, 2018. **42**(8): p. 772-782.
48. Schweiger, R., et al., *RL-SKAT: An Exact and Efficient Score Test for Heritability and Set Tests.* Genetics, 2017. **207**(4): p. 1275-1283.
49. Wang, K., *Conditional asymptotic inference for the kernel association test.* Bioinformatics, 2017. **33**(23): p. 3733-3739.
50. Wang, K., *Boosting the Power of the Sequence Kernel Association Test by Properly Estimating Its Null Distribution.* Am J Hum Genet, 2016. **99**(1): p. 104-14.
51. Maity, A., P.F. Sullivan, and J.Y. Tzeng, *Multivariate phenotype association analysis by marker-set kernel machine regression.* Genet Epidemiol, 2012. **36**(7): p. 686-95.
52. Goeman, J.J., S.A.v.d. Geer, and H.C.v. Houwelingen, *Testing against a high dimensional alternative.* J. R. Statist. Soc, 2006. **68**: p. 477–493.
53. Derkach, A., J.F. Lawless, and L. Sun, *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): p. 110-21.
54. King, C.R., P.J. Rathouz, and D.L. Nicolae, *An evolutionary framework for association testing in resequencing studies.* PLoS Genet, 2010. **6**(11): p. e1001202.
55. Sun, J., Y. Zheng, and L. Hsu, *A unified mixed-effects model for rare-variant association in sequencing studies.* Genet Epidemiol, 2013. **37**(4): p. 334-44.
56. Chen, H., et al., *Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies.* Am J Hum Genet, 2019. **104**(2): p. 260-274.
57. Chen, L.S., et al., *An exponential combination procedure for set-based association tests in sequencing studies.* Am J Hum Genet, 2012. **91**(6): p. 977-86.
58. Zhou, H., et al., *Association screening of common and rare genetic variants by penalized regression.* Bioinformatics, 2010. **26**(19): p. 2375-82.
59. Ionita-Laza, I., et al., *A new testing strategy to identify rare variants with either risk or protective effect on disease.* PLoS Genet, 2011. **7**(2): p. e1001289.
60. Quintana, M.A., et al., *Incorporating model uncertainty in detecting rare variants: the Bayesian risk index.* Genet Epidemiol, 2011. **35**(7): p. 638-49.
61. Yi, N. and D. Zhi, *Bayesian analysis of rare variants in genetic association studies.* Genet Epidemiol, 2011. **35**(1): p. 57-69.
62. Almasy, L., et al., *Genetic Analysis Workshop 17 mini-exome simulation.* BMC Proc, 2011. **5 Suppl 9**: p. S2.
63. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
64. Spencer, C.C., et al., *Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.* PLoS Genet, 2009. **5**(5): p. e1000477.
65. Kryukov, G.V., et al., *Power of deep, all-exon resequencing for discovery of human trait genes.* Proc Natl Acad Sci U S A, 2009: p. 3871–3876.

66. Hernandez, R.D., *A flexible forward simulator for populations subject to selection and demography.* Bioinformatics, 2008. **24**(23): p. 2786-7.
67. Schaffner, S.F., et al., *Calibrating a coalescent simulation of human genome sequence variation.* Genome Res, 2005. **15**(11): p. 1576-83.
68. Montana, G., *HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients.* Bioinformatics, 2005. **21**(23): p. 4309-11.
69. Hudson, R.R., *Generating samples under a Wright–Fisher neutral model of genetic variation.* Bioinformatics, 2002. **18**(2): p. 337-338.
70. Zhao, N., et al., *Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test.* Am J Hum Genet, 2015. **96**(5): p. 797-807.
71. Wang, K., *Genetic association tests in the presence of epistasis or gene-environment interaction.* Genetic Epidemiology, 2008. **32**: p. 606–614.
72. Nicolas, G., et al., *SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease.* Mol Psychiatry, 2016. **21**(6): p. 831-6.
73. Le Scouarnec, S., et al., *Testing the burden of rare variation in arrhythmia-susceptibility genes provides new insights into molecular diagnosis for Brugada syndrome.* Hum Mol Genet, 2015. **24**(10): p. 2757-63.
74. Goodrich, J.K., et al., *Human genetics shape the gut microbiome.* Cell, 2014. **159**(4): p. 789-99.
75. Grove, M.L., et al., *Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium.* PLoS One, 2013. **8**(7): p. e68095.
76. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
77. Tavtigian, S.V., et al., *Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer.* Am J Hum Genet, 2009. **85**(4): p. 427-46.
78. Romeo, S., et al., *Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans.* J Clin Invest, 2009. **119**(1): p. 70-9.
79. Nejentsev, S., et al., *Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.* Science, 2009. **324**(5925): p. 387-9.
80. Nair, R.P., et al., *Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways.* Nat Genet, 2009. **41**(2): p. 199-204.
81. Smith, J.A., et al., *The genetic architecture of fasting plasma triglyceride response to fenofibrate treatment.* Eur J Hum Genet, 2008. **16**(5): p. 603-13.
82. Firmann, M., et al., *The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome.* BMC Cardiovasc Disord, 2008. **8**: p. 6.
83. Schymick, J.C., et al., *Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data.* The Lancet Neurology, 2007. **6**(4): p. 322-328.
84. Hunter, D.J., et al., *A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.* Nat Genet, 2007. **39**(7): p. 870-4.
85. Ahituv, N., et al., *Medical sequencing at the extremes of human body mass.* Am J Hum Genet, 2007. **80**(4): p. 779-91.
86. Victor, R.G., et al., *The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health.* Am J Cardiol, 2004. **93**(12): p. 1473-80.
87. Bernstein, J.L., et al., *Study design: evaluating gene-environment interactions in the etiology of breast cancer - the WECARE study.* Breast Cancer Res, 2004. **6**(3): p. R199-214.
88. Vijver, M.J.v.d., et al., *A gene-expression signature as a predictor of survival in breast cancer.* The New England Journal of Medicine, 2002. **347**(25).

89. Gordon, M.O., M.A. Kass, and O.H.T.S. Group, *The ocular hypertension treatment study: Design and Baseline Description of the Participants.* ARCH OPHTHALMOL, 1999. **117**: p. 573-583.
90. Li, B., D.J. Liu, and S.M. Leal, *Identifying rare variants associated with complex traits via sequencing.* Curr Protoc Hum Genet, 2013. **Chapter 1**: p. Unit 1 26.
91. Zhan, X., K. Banerjee, and J. Chen, *Variant-set association test for generalized linear mixed model.* Genet Epidemiol, 2021. **45**(4): p. 402-412.
92. Gogarten, S.M., et al., *Genetic association testing using the GENESIS R/Bioconductor package.* Bioinformatics, 2019. **35**(24): p. 5346-5348.
93. Lumley, T., et al., *FastSKAT: Sequence kernel association tests for very large sets of markers.* Genet Epidemiol, 2018. **42**(6): p. 516-527.
94. Wang, G.T., B. Peng, and S.M. Leal, *Variant association tools for quality control and analysis of large-scale sequence and genotyping array data.* Am J Hum Genet, 2014. **94**(5): p. 770-83.
95. Schaid, D.J., et al., *Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data.* Genet Epidemiol, 2013. **37**(5): p. 409-18.
96. Ionita-Laza, I., et al., *Sequence kernel association tests for the combined effect of rare and common variants.* Am J Hum Genet, 2013. **92**(6): p. 841-53.
97. Li, B., G. Wang, and S.M. Leal, *SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits.* Bioinformatics, 2012. **28**(20): p. 2703-4.
98. Liu, H., Y. Tang, and H.H. Zhang, *A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables.* Computational Statistics & Data Analysis, 2009. **53**(4): p. 853-856.
99. Li, S. and Y. Cui, *Gene-centric gene–gene interaction: A model-based kernel machine method.* The Annals of Applied Statistics, 2012. **6**(3): p. 1134-1161.
100. Lin, X., et al., *Test for interactions between a genetic marker set and environment in generalized linear models.* Biostatistics, 2013. **14**(4): p. 667-81.
101. Choi, S., et al., *FARVAT: a family-based rare variant association test.* Bioinformatics, 2014. **30**(22): p. 3197-205.
102. Yan, Q., Z. Fang, and W. Chen, *KMgene: a unified R package for gene-based association analysis for complex traits.* Bioinformatics, 2018. **34**(12): p. 2144-2146.
103. Bansal, V., et al., *Statistical analysis strategies for association studies involving rare variants.* Nat Rev Genet, 2010. **11**(11): p. 773-85.
104. Guo, M.H., et al., *Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders.* Am J Hum Genet, 2016. **99**(3): p. 527-539.
105. Wang, M. and S. Lin, *Detecting associations of rare variants with common diseases: collapsing or haplotyping?* Brief Bioinform, 2015. **16**(5): p. 759-68.
106. Moutsianas, L., 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): p. e1005165.
107. Weber, L.M., et al., *Essential guidelines for computational method benchmarking.* Genome Biol, 2019. **20**(1): p. 125.
108. Boutry, S., et al., *Excalibur: a new ensemble method based on an optimal combination of aggregation tests for rare-variant association testing for sequencing data.* PLoS Comput Biol, 2023. **Manuscript under revision**.
109. Saad, M. and E.M. Wijsman, *Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees.* Genet Epidemiol, 2014. **38**(7): p. 579-90.
110. He, Z., et al., *A genome-wide scan statistic framework for whole-genome sequence data analysis.* Nat Commun, 2019. **10**(1): p. 3018.

111. Zhu, B., L. Mirabello, and N. Chatterjee, *A subregion-based burden test for simultaneous identification of susceptibility loci and subregions within.* Genet Epidemiol, 2018. **42**(7): p. 673-683.
112. Agarwala, V., et al., *Evaluating empirical bounds on complex disease genetic architecture.* Nature Genetics, 2013. **45**(12): p. 1418-1427.
113. Chen, M.H., A. Pitsillides, and Q. Yang, *An evaluation of approaches for rare variant association analyses of binary traits in related samples.* Sci Rep, 2021. **11**(1): p. 3145.
114. Povysil, G., et al., *Rare-variant collapsing analyses for complex traits: guidelines and applications.* Nat Rev Genet, 2019. **20**(12): p. 747-759.
115. Reimand, J., et al., *Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap.* Nat Protoc, 2019. **14**(2): p. 482-517.
116. Wightman, D.P., et al., *Rare variant aggregation in 148,508 exomes identifies genes associated with proxy dementia.* Sci Rep, 2023. **13**(1): p. 2179.
117. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.
118. Gillespie, M., et al., *The reactome pathway knowledgebase 2022.* Nucleic Acids Res, 2022. **50**(D1): p. D687-D692.
119. Kanehisa;, M. and S. Goto, *KEGG kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000(28): p. 27-30.
120. The Gene Ontology, C., *The Gene Ontology Resource: 20 years and still GOing strong.* Nucleic Acids Res, 2019. **47**(D1): p. D330-D338.
121. Wei, P., et al., *On Robust Association Testing for Quantitative Traits and Rare Variants.* G3 (Bethesda), 2016. **6**(12): p. 3941-3950.
122. Wang, L., et al., *metaFARVAT: An Efficient Tool for Meta-Analysis of Family-Based, Case-Control, and Population-Based Rare Variant Association Studies.* Front Genet, 2019. **10**: p. 572.