Metadata of the chapter that will be visualized online

Chapter Title	Assessing Rare Variation in Complex Traits	
Copyright Year	2018	
Copyright Holder	Springer Science+Business Media, LLC, part of Springer Nature	

Corresponding Author	Family Name	**Kuchenbaecker**
	Particle	
	Given Name	**Karoline**
	Suffix	
	Organization	Wellcome Trust Sanger Institute
	Address	Cambridge, UK
	Email	karoline.kuchenbaecker@sanger.ac.uk
Author	Family Name	**Appel**
	Particle	
	Given Name	**Emil Vincent Rosenbaum**
	Suffix	
	Division	Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Genetics, Faculty of Health Sciences
	Organization	University of Copenhagen
	Address	Copenhagen, Denmark

Abstract

While genome-wide association studies have been very successful in identifying associations of common genetic variants with many different traits, the rarer frequency spectrum of the genome has not yet been comprehensively explored. Technological developments increasingly lift restrictions to access rare genetic variation. Dense reference panels enable improved genotype imputation for rare variants in studies using DNA microarrays. Moreover, the decreasing cost of next generation sequencing makes whole exome and genome sequencing increasingly affordable for large samples. Large-scale efforts based on sequencing, such as ExAC, 100,000 Genomes, and TopMed, are likely to significantly advance this field.

The main challenge in evaluating complex trait associations of rare variants is statistical power. The choice of population should be considered carefully because allele frequencies and linkage disequilibrium structure differ between populations. Genetically isolated populations can have favorable genomic characteristics for the study of rare variants. One strategy to increase power is to assess the combined effect of multiple rare variants within a region, known as aggregate testing. A large number of methods have been developed for this. Model performance depends on the genetic architecture of the region of interest.

Keywords (separated by '-')

Low frequency variants - Rare variants - Sequencing - Association study - Aggregate test - Burden test - Isolated population

Chapter 5

Assessing Rare Variation in Complex Traits

Karoline Kuchenbaecker and Emil Vincent Rosenbaum Appel

AU1

Abstract

While genome-wide association studies have been very successful in identifying associations of common genetic variants with many different traits, the rarer frequency spectrum of the genome has not yet been comprehensively explored. Technological developments increasingly lift restrictions to access rare genetic variation. Dense reference panels enable improved genotype imputation for rare variants in studies using DNA microarrays. Moreover, the decreasing cost of next generation sequencing makes whole exome and genome sequencing increasingly affordable for large samples. Large-scale efforts based on sequencing, such as ExAC, 100,000 Genomes, and TopMed, are likely to significantly advance this field.

The main challenge in evaluating complex trait associations of rare variants is statistical power. The choice of population should be considered carefully because allele frequencies and linkage disequilibrium structure differ between populations. Genetically isolated populations can have favorable genomic characteristics for the study of rare variants.

One strategy to increase power is to assess the combined effect of multiple rare variants within a region, known as aggregate testing. A large number of methods have been developed for this. Model performance depends on the genetic architecture of the region of interest.

Key words Low frequency variants, Rare variants, Sequencing, Association study, Aggregate test, Burden test, Isolated population

1 Background

The discovery of genetic variants contributing to the heritability of complex traits has boomed in recent years. Hundreds of associations, mostly of common variants with small effects, have been identified for outcomes such as anthropometric measures, blood biomarkers, and common diseases. However, rare variants are likely to play an important role in the genetics of many of these traits. Identifying variants with large effects could be particularly useful from a clinical perspective. In the context of disease, the accuracy of predicted risks of carriers of such variants can significantly improve. Furthermore, trait associations with such variants can lead to important biological insights and novel treatments for diseases.

Evangelos Evangelou (ed.), *Genetic Epidemiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1793, https://doi.org/10.1007/978-1-4939-7868-7_5, © Springer Science+Business Media, LLC, part of Springer Nature 2018

A number of empirical findings demonstrate the importance of rare variants and illustrate their clinical potential. Several of these success stories relate to lipid traits. For example, targeted sequencing of data from the Dallas Heart Study revealed an association between low-density lipoprotein (LDL) cholesterol and rare nonsense mutations in *PCSK9* [1]. This gene encodes a protein that is involved in the regulation of LDL cholesterol levels. These LDL-decreasing mutations were also shown to lead to a significant reduction in risk of coronary heart disease (CHD) [2]. Monoclonal antibodies targeting this molecule were developed to reduce CHD risk and these lowered LDL levels beyond what could be achieved by statins alone [3, 4]. As another example, a study using samples from a cosmopolitan UK population [5], as well as studies in isolated populations [6–8], identified several rare variants in the apolipoprotein C-III (APOC3) gene affecting levels of triglycerides in blood with evidence for a cardioprotective effect of these alleles [8–10]. An antisense oligonucleotide was developed to lower APOC3 levels and it also led to decreased triglycerides in patients with high-baseline levels [11].

One of the main technical challenges for the discovery of rare variant associations has been the limited coverage of rare variation by DNA microarrays commonly used in genome-wide association studies (GWAS) (*see* Subheading Technology). However, the decreasing cost of whole exome and whole genome sequencing make these technologies increasingly affordable for larger sample sizes (Table 1). The first large genome sequencing project was the 1000 Genomes Project [13], followed by the UK10K Project [14]. These efforts have significantly advanced the field of genomics. Large numbers of additional variants were discovered and insights into population genetics gained. These projects enabled hundreds of other studies to operate in a very cost-effective way by using DNA microarray genotyping and carrying out genotype imputation with the haplotypes from the sequencing efforts as reference panels. Recognizing the potential of genomics for medicine, governments in the UK and USA seized the opportunity of more affordable sequencing. The precision medicine initiative, launched by US President Barack Obama in 2015, aims to advance personalized medicine through the Trans-Omics for Precision Medicine (TOPMed) programme which involves whole-genome sequencing of 62,000 individuals, possibly up to 100,000 at a later stage [15]. The focus of this programme is on heart, lung, blood, and sleep disorders. There is also a large-scale initiative in the UK, the 100,000 Genomes Project [16]. It involves whole-genome sequencing of germline and tumor DNA of 25,000 cancer patients and also of DNA of 50,000 individuals to study rare diseases. The aim of this programme is to implement genomic medicine in routine clinical practice for rare diseases and cancer [17]. The maximum potential of such initiatives can be realized

Table 1
Overview of essential features of different genotyping technologies

	GWAS Chip	Exome Chip	WES	1× WGS	High depth WGS
Region covered	Genome	Mostly exome	Exome	Genome	Genome
Discovery of novel variants	No	No	Yes	Yes	Yes
Bioinformatics and QC workload	Small	Small[a]	Medium[a]	Large	Large
Cost compared to of a full genome[b]	4%	6%	20%	30%	100%

[a] Exome Chip and WES QC do not have access to genome-wide genotypes, and thus some QC metric are not available when using these technologies

[b] The price of a genome ($1245 in October 2016) was estimated from [12]. The fraction represents an approximate estimation from prices in our laboratory

when data from different sequencing projects are combined. This has recently been done for whole exome sequencing studies. The Exome Aggregation Consortium (ExAC) project, a collection of exome data from more than 60,000 individuals, yielded important findings with implications for the pathogenicity of mutations in coding regions [18]. These large sequencing projects could significantly advance our understanding of the role of rare genetic variation.

In the next section, we discuss differences between populations with respect to variant frequency and linkage disequilibrium patterns and how these affect design considerations for studying rare variants. The subsequent part is devoted to the measurement of rare variants and compares DNA microarray genotyping with DNA sequencing technologies. The final part of this describes different statistical analysis techniques to assess trait associations of rare variants. The focus lies on aggregate tests that assess the combined effect of multiple variants in order to improve the power limitations.

While rare structural variants play an important role for some complex traits, this chapter only covers single nucleotide variants. The term “low frequency” is used for variants with minor allele frequencies (MAF) between 1% and 5% and “rare” for variants with MAF less than 1%.

2 Population-Specific Differences in Genetic Variation

Genetic diversity and linkage disequilibrium (LD) structure differ between populations. Some alleles are common in one and rare in another population and some variants are only present in some populations. It is vital to put consideration into the choice of population for a given study, especially for rare variant association studies. Differences in LD structure mean that tagging properties of variants on GWAS arrays can differ greatly between populations resulting in differences in the accuracy with which the signal of a variant can be captured. Variant frequency and imputation accuracy affect statistical power to detect an association. The effect of a variant might also differ between populations due to different environments or epistasis. Several genetic associations with complex traits were found to be population-specific, such as the association of the *MTNR1B* locus with glucose metabolism in European populations but not in East Asian populations [19].

A number of factors shape the genetic make-up of a population including population size, historical bottlenecks, and natural selection [20]. A bottleneck is a period of time stretching across several generations where the population shrinks at the start of the bottleneck and remains stable within the bottleneck. Bottlenecks can be caused, for example, by a famine, pest, or geographical narrow passageway. The effects of a bottleneck are long lasting. After a bottleneck the genetic make-up of the population is composed exclusively of the genetic variation from the lineages that survived the bottleneck while some variants present in the original population are lost. When the population starts expanding again, the variation from the surviving lineages will remain frequent to a much higher extent than variation introduced into the population after the bottleneck. The underlying LD pattern in the surviving linages will be maintained in the expanding population, only broken up by new recombination events [20].

As a consequence, genetic diversity and LD structure are markedly different between Sub-Saharan African and European populations, with higher levels of genetic diversity in the African populations and longer spans of LD in the European populations [21]. This is mainly due to the fact that the European populations share a historic bottleneck, the migration out of Africa, while the populations of Africa consist of several smaller populations, without a common historic event, that continuously admixed, splitting up the LD blocks and allowing for more genetic diversity [21].

2.1 The Special Case of Isolated Populations

An isolated population is a small population that has undergone a bottleneck in its history and remained isolated from other populations after the bottleneck. Due to the genetic drift some variants have risen in frequency and there are higher levels of relatedness and longer LD blocks compared to non-isolated population [22]. Greater environmental and phenotypic homogeneity are often observed as well. Taken together, this gives rise to greater statistical power to detect associations of rare alleles that have drifted to higher frequency, which makes isolated populations particularly attractive for studying rare variants. However, note that in isolated populations only a subset of the rare variants seen in the general population, from which the isolate was derived, will be present, limiting the association testing to those variants.

A number of recent locus discoveries in population isolates have highlighted these properties [7, 23–26]. As an example, the Greenlandic population is a small isolated population with high degrees of relatedness, large LD blocks, fewer rare variants in total, but with higher allele frequency in the average for observed variants [27]. These features were exploited by Moltke et al. who found an association between the nonsense p.Arg684Ter variant in *TBC1D4* and postprandial hyperglycemia, impaired glucose tolerance, and risk of type II diabetes [28]. This variant is extremely rare in the general population (only one allele was found in the 1092 individuals of the 1000 Genome project), but common in the Greenlandic population (MAF = 17%). To observe the same number of alleles seen in the Greenlandic cohort in an outbred population, one would have had to sample over 400,000 individuals. This highlights important considerations regarding the study of rare variants in isolated populations. Rare variants can rise to higher frequencies leading to increased statistical power for discovery, but observed associations may be limited to the isolated population because the variant is not present or extremely rare in other populations. This does not diminish the relevance of the locus discovery, however, as these findings can point to biological pathways involved in complex traits that would otherwise have been overlooked.

3 Genotyping Technologies for Rare Variants

Here, we discuss two types of genotyping technologies, DNA microarrays, and sequencing. We explore the pros and cons of applying these technologies when investigating rare variants. We further differentiate between whole exome sequencing (WES) and whole genome sequencing (WGS).

3.1 DNA Microarray Genotyping

DNA microarray genotyping, also known as chip genotyping, is a comparably cheap and versatile technology, with prices down to $50 per sample for a genome-wide chip. The technology has been widely used and advanced software has been designed to ease the workload of bioinformatics (Table 1).

DNA microarrays are based on known variants and use a calling algorithm based on clustering. Clustering is a method to automatically draw clusters around similar genotype calls, based on the intensity of the colored light used by the high-throughput microarray genotyping machine. Clustering is dependent on the total number of samples in each cluster. This means that the clustering algorithms perform best for common variants where the three clusters, homozygotes wild-type, heterozygote, and homozygote derived, are of similar size. Clustering often performs poorly when only a few samples can be gathered into one cluster which is the case for rare variants.

Genome-wide DNA microarrays are designed on the basis of tagging which exploits the fact that variants are inherited in LD-blocks. Variants are selected for inclusion on the chip in such a way that each LD block is represented. Tagging reduces the number of variants needed to adequately cover the majority of genetic variation down to thousands. Using genotype imputation one can then make use of the information contained in multiple typed variants to infer the genotypes of the variants missing from the array. This requires a reference panel of genomes that contain the variants missing on the chip, so that their relation to typed variants can be inferred. There are general GWAS chips that were designed to capture maximal genetic information with a limited number of variants. There are also custom arrays that were designed to target regions of the genome that are of interest to a specific disease or trait, such as the MetaboChip or OncoArray.

GWAS arrays generally have very good coverage of common genetic variation. However, rare variants are on average in lesser LD with other variants than common variants, resulting in lower coverage. In the context of single SNP association analysis, Yang et al. showed empirically that 81% of common and 25% of rare (MAF $\leq$ 1%) variation can be captured by the best tagging SNP using the CoreExome array in combination with imputation to the 1000 Genomes Project reference panel [29]. Recently, large reference panels from the UK10K study [14] and The Haplotype Reference Consortium [30] have become available and have increased the power to impute rare variants from DNA microarrays [29].

There are several strategies to improve access to rare variation through chip genotyping. The Exome Chip [31] was designed to capture rare coding variants based on exome sequencing and has since been used to genotype millions of samples in different association studies which successfully identified rare variant associations with various traits and diseases [26, 32, 33]. The Exome Chip offers a very cost-effective solution for large-scale genotyping of rare variants in exons (Table 1). However, the focus on rare exonic variants also represents an important limitation because the majority of complex trait associations identified so far were with noncoding variants. Furthermore, the array is targeted to European populations and is not suited to discover de novo mutations. For this, exome sequencing is a better option.

3.2 Next Generation Sequencing

Generally, genotype sequencing is more expensive than DNA microarrays, but has several advantages, especially in the context of low frequency and rare variants. Prices for whole exome sequencing are around three times cheaper than for whole genome sequencing which cost ~1200$. However, the costs have been AU5 decreasing continuously as the technology matures (Table 1). There are options to make sequencing more cost-effective. One common approach is to lower the depth, the average number of overlapping sequence fragments, called reads, mapped to the same position. Alternatively, one can opt to cover only specific regions, such as candidate genes.

Advanced software is available for researchers working with sequencing data. However, the bioinformatics workload involved in the quality control (QC) and analysis of sequencing data is more taxing than for DNA microarrays (Table 1).

3.2.1 Whole Exome Sequencing

Whole exome sequencing (WES) is a common strategy to investigate rare variants while keeping the cost down. This is done by limiting the regions that are sequenced to only the exome, without compromising the sequencing depth. This allows for accurate calling of rare variants in regions where they are likely to have an effect. WES also enables the detection of novel variants, which is not possible with DNA microarrays.

Focusing on the exome is motivated by the fact that missense variants found in an exon of a gene can be disruptive to the protein sequence and can therefore have an effect on the function of the protein. Mendelian diseases represent an extreme case of this where the disease can be caused by a single missense variant. Evolutionary conservation has therefore restricted the frequency of exonic variants. WES is recommended when investigating the effects of rare variants on monogenetic diseases. However, the majority of previously identified associations identified for complex traits were for noncoding variants [34].

3.2.2 Whole Genome Sequencing

Whole-genome sequencing (WGS) offers the potential to access the entire genetic information of an individual. It enables the discovery of novel variants, and makes it possible to access rare variants outside as well as within coding regions. As WGS covers the whole genome, it also enables the mapping of the underlying genetic architecture of complex polygenetic traits and the study of large structural variations, such as copy-number variations (CNV).

The amount of data generated per individual is considerably larger than for WES or chip genotyping. For example, in comparison with a chip-based GWAS, WGS requires about 1000 times more space to store the post-QC genotype information for chromosome 1 (~2 million WGS variants and ~ 58,000 GWAS tag-SNPs) for 1200 individuals (~13 GB, in a compressed VCF-file [12], for the WGS genotypes versus ~20 MB, in binary plink-files, for Omni Exome Chip genotypes of the same individuals). Processing of these files requires more computational resources, is more time consuming, and requires technical expertise. Furthermore, control of type I error requires consideration as a larger number of statistical tests are carried out (*see* Subheading Significance Thresholds).

The biggest drawback of WGS is its cost. It is considerably more expensive than WES and DNA Microarray genotyping. While cost can be lowered by using a low read depth, this is at the expense of quality of the genotype calls. Using a low depth, e.g., an average depth of one read per position, known as 1 × WGS, will lead to more errors in calling variants. This can affect the discovery of novel variants in particular. One strategy to improve on this is by using imputation with large reference panels. A strict QC pipeline, especially when investigating novel rare variants, is needed to avoid type I errors. Overall, low depth WGS offers a cost effective method for studying rare variants.

While it is a significant advantage of WGS over WES to be able to access noncoding variation, the interpretation of the findings can be much less straightforward in comparison with associations of mutations affecting protein sequence. Understanding regulatory effects is considerably more complex and represents a very active area of research. One approach to ease interpretation of association findings is to use annotation scores that represent the likelihood of a given variant to affect protein expression. This has been done using different sources of information for coding as well as noncoding variants, e.g., for the Eigen [35], GWAVA [36], or CADD scores [37].

As the technology develops and genome annotations improve, the challenges involved in sequencing will become easier to meet, and WGS will become more feasible for increasingly large sample sets.

4 Association Analyses Methods for Rare Variants

4.1 Single Variant Association Tests

Fast and efficient estimation procedures have been developed to carry out association tests for large numbers of variants. Most genetic association studies assume an additive genetic model where the SNP effect is estimated per copy of the effect allele. Usually, either linear or logistic regressions are used to estimate and test SNP associations for continuous or dichotomous outcomes, respectively. Increasingly, linear mixed models are applied which allow for the inclusion of relatives and account for possible population stratification by adjusting for genetic similarity between individuals. Details are described elsewhere is this book (*see* Chapters 3 and 4). These methods are also applicable to low frequency and rare variants. However, in case-control studies for variants with small numbers of carriers of the rare allele, the p-values of asymptotic logistic regression tests can be inaccurate [38, 39]. In this context, the minor allele count (MAC) has been established as a more useful metric than the minor allele frequency (MAF) because it is the absolute number of alleles that affects the performance of the test [38]. It has been shown that logistic regression tests can perform poorly for variants with MAC of less than 400 which can be used as a guidance for choosing an appropriate test.

One solution to the problem is to use Fisher's exact test instead which represents the gold standard to assess an association between categorical variables with small counts [40]. However, it is rarely used in this context because it cannot adjust for covariates. Alternative methods include Firth regression which is a penalized likelihood-based method that has been shown to perform well for rare variants [38, 41]. Permutation approaches have also been proposed [39]. Finally, a computationally efficient resampling approach for score tests has been developed [42].

4.1.1 Effect of Population Stratification, Non-normality and Outliers

Association testing for rare variants can be less robust to violations of assumptions. Rare variant association analyses (both single variant and aggregate tests) can be more strongly affected by non-normality, outliers and population stratification than association analyses for common variants [43]. With respect to outliers, it should be taken into account that extreme values of a trait could also be observed as the result of a rare high penetrance mutation, as seen in Mendelian diseases. Therefore, exclusions of outliers and variable transformations need to be considered carefully. Furthermore, association tests are particularly sensitive to population stratification because even small levels of stratification can lead to different frequencies of rare variants [44–49]. Therefore, quality control has to be particularly thorough. However, adjusting for fine-scale patterns of population stratification can be difficult with traditional methods when stratification for rare variants differs from that of common variants. For more details *see* Chapter 3.

4.1.2 Significance Threshold

For single variant association testing, multiple testing is an important consideration. In GWAS and sequencing studies, the associations of hundreds of thousands or even millions of genetic variants are evaluated, leading to a high multiple testing burden. Most of these variants are unlikely to causally affect the trait of interest so that the prior probability of association is small for each variant. The majority of previously published genetic association studies used an adjusted *p*-value threshold to account for the number of independent tests. Because many variants are in LD with each other and therefore not independent, a Bonferroni adjustment for the total number of variants tested would be too conservative. For chip-based genome-wide association studies, a *p*-value threshold of 5×10^{-8} has been established and is used routinely as it has been demonstrated to be valid for many GWAS arrays [50–52]. However, this threshold is not valid for whole exome or whole genome sequencing. The addition of many rare variants that tend to be in less strong LD with other variants leads to an increased number of independent tests.

The significance threshold for rare variant studies depends on the genotyping technology used, MAF threshold for variants considered (related to sequencing depth) and the population as that affects the genomic LD structure (*see* Subheading Populations). For samples from cosmopolitan populations of European ancestry it has been demonstrated that a threshold of 1×10^{-8} for whole-genome and 3×10^{-7} for whole exome sequencing provide a level of adjustment for variants with MAF > 0.001 that is equivalent to the adjustment of the 5×10^{-8} threshold for common variants [53]. There is a higher burden of multiple testing for samples of African ancestry due to greater genetic diversity. Isolated populations on the other hand have longer shared haplotypes and therefore require adjustment for a smaller number of independent tests which renders them particularly suitable for the analysis of rare variation (*see* Subheading Populations).

4.1.3 Statistical Power to Identify Novel Associations

Due to the high multiple testing burden, one of the main challenges for genetic association studies is to provide sufficient statistical power to detect novel associations with a trait of interest. For the identification of associations of low frequency and rare variants, statistical power is an even greater challenge. Factors impacting the power to detect a trait association include frequency and effect size of a variant and how well it can be imputed in case it was not genotyped or sequenced directly [54]. As discussed in Subheading Genotyping Technologies, in GWAS the average imputation accuracy for rare variants is lower than for common variants due to their reduced linkage disequilibrium. Therefore, the power to detect associations of rare variants can be low in GWAS.

Low frequency of variants can severely limit statistical power to find trait associations. For example, given a disease prevalence of 10%, a sample size of 10,000 cases and 10,000 controls, an OR of 1.2 (additive effect), the power to detect an association at $p < 5 \times 10^{-8}$ for a common variant with MAF = 0.4 is 98% whereas the power for a low frequency variant with MAF = 0.05 is 16%. As Fig. 1 demonstrates, given a moderate effect size (e.g., OR = 1.5) variants with MAF = 0.01 require more than 30,000 samples while variants with MAF = 0.001 require more than 300,000 samples to achieve sufficient discovery power (>80%). This demonstrates that in this setting, associations of rare variants are realistically discoverable only if the variants have moderate to large effect sizes.

Therefore, an important question concerns the effect size distribution of rare variants. If effect sizes are consistently small, then even large studies have limited power to detect rare variant associations. For many health-related complex traits it is now firmly established that almost all associated common variants have relatively

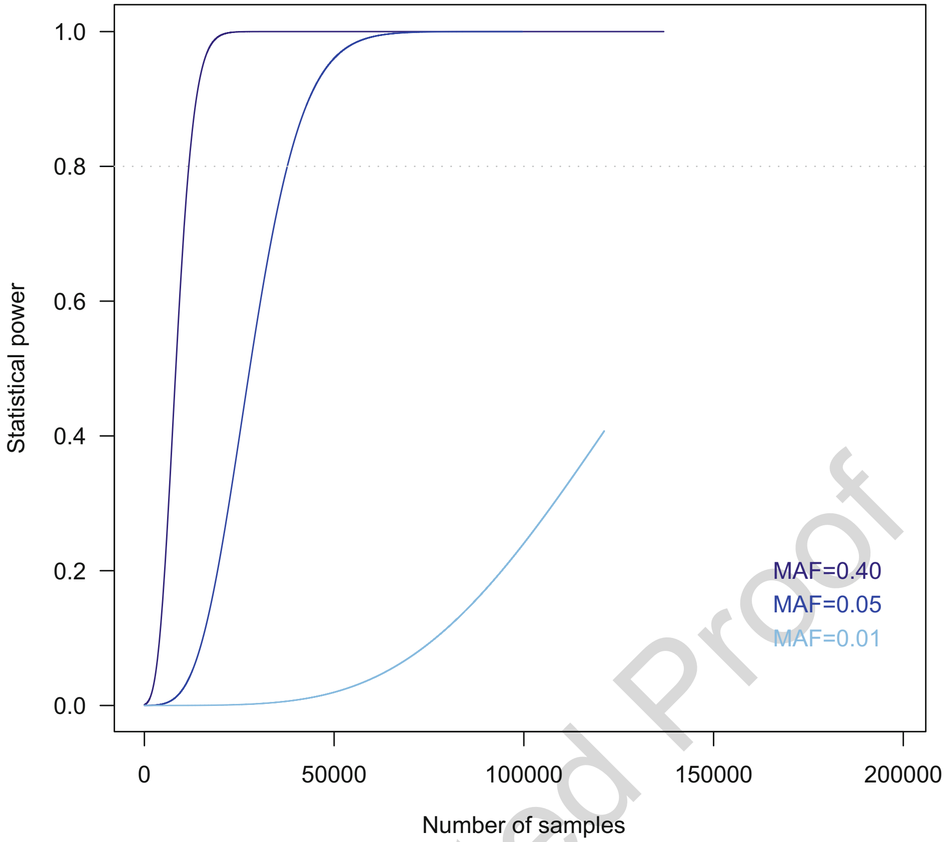

Fig. 1 Power to detect a variant association with OR= 1.5 (additive) at $p < 5 \times 10^{-8}$ in a case-control study with a 50:50 ratio of cases to controls and a disease prevalence of 10%

small effects (i.e., OR < 1.5). Despite very high statistical power, common variants with large effects have not been discovered. Similar conclusions cannot be drawn with respect to the rare variants. As the power calculations demonstrate, much larger samples are needed to identify associations of rare variants given the same effect size as common variants. Moreover, all genetic association studies with more than 100,000 samples that have been published to-date used GWAS genotyping and had therefore limited coverage of rare variants (*see* Subheading Genotyping Technologies). For traits under selection it is likely that variants of moderate to large effect are rare. In line with this, rare and low-frequency variants are strongly enriched for functional and deleterious variants [55–57]. However, genetic architecture differs between traits and is an ongoing field of research.

4.2 Aggregate Testing

In order to increase statistical power to detect rare variant associations, analysis methods have been developed to test the combined effect of several variants. These tests are known as aggregate or gene-based tests. There are several arguments supporting the use of aggregate methods. These include the observations that recent population expansion may have led to high numbers of functional variants, that a combination of variants can be necessary to create a phenotype, and that an increasing number of genes have been discovered with multiple common and/or rare associated variants. Finally, a number of previous successful discoveries from gene-based tests provide proof of principle [58]. Variants are usually combined within genes. An alternative unit can be sliding windows across the genome to assess the combined effect of variants located close to each other. Combining variants from genes in a common pathway has also been suggested [59].

A number of different approaches have been developed for aggregate testing. In general, decisions involved in aggregate testing include the unit of aggregation (e.g., gene, region of a certain size), the coding scheme for the genotypes (e.g., score, carrying any vs no rare alleles, recoding of variants with effects in the opposite direction), variant filtering (e.g., frequency, functional annotation), weighting scheme (e.g., frequency, predicted functional effect, imputation accuracy), and whether to include covariates (e.g., principle components). The following sections describe different aggregate testing methods. Please note that meta-analysis methods for aggregate tests are described elsewhere in this book.

4.2.1 Collapsing Tests

In collapsing tests the numbers of rare alleles carried are summed up for all variants within a specified region (e.g., gene). Each variant can be weighted. The association between this aggregate and the trait of interest is then tested through regression:

$$f(y_i) = \alpha + \beta \sum_j w_j g_{ij}$$

where y_i is the phenotype of individual i, g_{ij} is the genotype of individual i for variant j, w_j is an optional weight for variant j, $f()$ represents the link function and is the logit for dichotomous traits and linear for continuous traits. Note that there is just one regression coefficient β for the aggregate effect rather than separate ones for individual variants. AU6

Several different implementations of the collapsing approach have been developed. RVT can be used for continuous as well as dichotomous outcomes [60]. It can estimate the effect per additional minor allele carried or compare individuals who carry at least one minor allele with those who do not. The Cohort Allelic Sums Test (CAST) [61], Combined Multivariate and Collapsing (CMC) [62], and Weighted Sum Statistic (WSS) [63] were designed specifically for dichotomous outcomes and differ in terms of their coding of the genotypes, variant filtering, and weighting. For regions that contain a mix of causal and non-causal variants, the CMC test had highest statistical power among these methods [62, 64].

It has been demonstrated that for studies based on GWAS chip genotyping, imputation of variants improves power to detect gene-based associations [65]. There are several modified versions of the collapsing tests that can use imputed variants and account for variant quality. The cumulative minor allele test (CMAT) [66] and GRANVIL [67], an implementation of RVT, can use dosages for imputed variants. The Accumulation of Rare variants Integrated and Extended Locus-specific test (ARIEL) is another adaptation of RVT that can also use weights to adjust for variant quality scores [68].

In order to overcome some of the limitations of collapsing tests, modifications have been developed that adapt to properties of the data. The data adaptive test (aSum) [64] involves two stages. Results from a marginal model evaluating single SNP associations are used to recode variants. An extension, the step-up test [69], can be used to filter variants if their marginal test provides little evidence for association. The estimated regression coefficient test (EREC) [70] is another two-stage procedure that uses the regression coefficients from the marginal test as weights for the collapsing test. It adds a small constant to each weight because regression coefficients from single variants tests tend to be unstable for rare variants. The Kernel-based adaptive cluster method (KBAC) [71] uses Kernel-based adaptive weighting in order to select likely causal variants. The variable threshold (VT) approach [72] changes the MAF thresholds for each region in order to identify the optimal variant selection.

Most of the original collapsing methods are less powerful when the associations of the rare alleles of different variants are in opposite directions [73–75]. In the presence of different directions of effect, the data-adaptive approach performed well while the VT method performed well in the case of consistent direction of effect but existence of non-causal variants [74, 76]. However, adaptive methods tend to be computationally intensive because most of them require permutation tests in order to obtain p-value estimates.

4.2.2 Variance-Component Methods

The most widely used variance-component method is SKAT [77]. It assumes a multiple regression model with variants as predictors and variant-specific regression coefficients so that the direction and magnitude of the association of each variant can vary. A mixed model is fitted assuming a random effect for genotype with $\beta_j \sim N(0, w_j\tau)$ where τ is the variance component. The overall effect of the variants can then be assessed by testing whether $\tau = 0$ via a variance-component score test. Covariates are incorporated as fixed effects. It is also possible to include interaction effects. For a dichotomous outcome without covariates SKAT and the C-alpha test [73] are equivalent. Without weights, SKAT reduces to the sum of squares of the marginal score statistics, SSU test [78].

There are a number of modified versions of SKAT. For example, C-SKAT was designed to estimate aggregate effects for both common and rare variants [79]. AP-SKAT is an implementation that avoids deriving p-values from an asymptotic distribution which can lead to bias while reducing the computational load from permutation [80].

4.2.3 Combined Tests

SKAT is a popular choice because it accounts for differences in direction and magnitude of effect between variants. Moreover, it outperforms most adaptive testing methods in terms of computational efficiency because it does not require permutation testing. However, which one of the models has the highest statistical power depends on the underlying genetic architecture of the region and trait under consideration. Collapsing methods have higher power when the majority of variants are causal and have the same direction of effect [74, 77]. In practice, there usually is little prior knowledge about the genetic architecture. Therefore, SKAT-O [81] has been developed. It combines variance component and collapsing approaches in order to maximize power for different scenarios. Alternative unified approaches include MiST [82] and CCS for case control studies [83]. CCS models the variant distributions in cases and controls and can account for ascertainment by using a retrospective likelihood approach. It has been shown to perform favorably when samples sizes are small, variants are rare, and when there is a high proportion of non-causal variants [83]. In a recent simulation study, unified approaches had higher power than collapsing and variance component tests given a range of genetic architectures [84].

A general framework has been developed that enables combining any gene-based tests of choice into a unified approach [85]. This strategy provided higher statistical power than running tests separately and using Bonferroni correction.

One potential problem with both collapsing and variance component methods is that these tests can yield inflated type I error levels [86]. Therefore, inflation should be assessed.

4.2.4 Bayesian Approaches

Several Bayesian approaches have been developed. One advantage is that they can make use of prior information regarding variants [87, 88]. The exponential combination (EC) approach [89] uses a quadratic score term for the aggregate effect of variants and is particularly powerful when the proportion of causal variants is low. However, it requires permutation in order to estimate *p*-values and is therefore computationally demanding. The Variational Bayes discrete mixture test (VBDM) [90] on the other hand is very computationally efficient because it is based on Bayes approximate inference. VBDM explicitly models non-causal variant and therefore performs particularly well in a scenario with many non-causal variants.

4.2.5 Functional Data Analysis

In the framework of functional data analysis, the genomic region of interest is conceptualized as a sequence of variants which was the result of a stochastic process that depends on linkage and linkage disequilibrium and the genetic effects are therefore a function of variant location [91]. While variance component methods only account for LD between pairs of variants, this approach makes optimal use of the LD structure between multiple genetic variants in the region. Moreover, it is possible to include rare as well as common variants. Aggregate tests have been developed within this framework for continuous [91, 92] and dichotomous traits [93–95]. Using the same simulation setup as the original studies for variance component methods, these functional linear model approaches were shown to have higher statistical power than variance component methods in most of the tested scenarios [91, 92, 96, 97].

4.2.6 Relatedness

Most of the methods described so far assume that samples are independent. However, including relatives can increase statistical power to detect a genetic association [98]. For family-based studies with known pedigrees there are transmission-based tests [99, 100]. There is also a pedigree-based option for SKAT for continuous traits, famSKAT [101]. Other models use a genetic relatedness matrix rather than pedigree structures. This provides more flexibility for incorporating complex or unknown family structures. These methods are also applicable when there is a mix of related and unrelated individuals. Pedgene [102] offers rapid collapsing as well as variance-component tests for dichotomous and continuous traits and so do famrvtests for continuous traits [103]. There are other family-based modifications of SKAT, including FFBSKAT [104] and ASKAT [105]. MONSTER is a generalization of SKAT-O that accounts for relatedness [106]. Finally, there is also a modification of the functional linear model approach to use data from related individuals [107].

4.2.7 Survival Analysis

Some studies assess associations of genetic variants with time to an event within a survival analysis framework. A modified version of collapsing tests and SKAT, the CoxBT and CoxSKAT likelihood ratio tests were developed for this setting [108]. Other variance component implementations exist [109, 110]. There is also an extension of the functional linear model approach to assess region-based associations using Cox regression [111].

5 Conclusion

Method development for aggregate testing of rare variants is a dynamic area of research. One of the advantages is that tests have been developed for a variety of different study designs. On the other hand, it can be difficult to navigate this field and identify the optimal test for a given study. The statistical power of each method is dependent on the genetic architecture of the trait (and region) of interest and the ranking of tests changes for different scenarios. In situations with little prior knowledge regarding the genetic architecture of the trait of interest, unified approaches incorporating methods that perform well given high as well as low proportions of causal variants can be a good choice.

As in single variant association testing, hits from aggregate tests also require confirmation using an independent replication sample. However, the locus needs to be validated rather than a single variant. There are different strategies to do this that may need to involve targeted sequencing of the locus [112].

References

1. Cohen J, Pertsemlidis A, Kotowski IK et al (2005) Low LDL cholesterol in African Americans resulting from frequent nonsense mutations in PCSK9. Nat Genet 37(3):328–328. https://doi.org/10.1038/ng0305-328c
2. Cohen JC, Boerwinkle E, Mosley TH (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. New Engl J Med 354(12):1264–1272. https://doi.org/10.1056/NEJMoa054013
3. Roth EM, McKenney JM, Hanotin C, Asset G, Stein EA (2012) Atorvastatin with or without an antibody to PCSK9 in primary hypercholesterolemia. New Engl J Med 367(20):1891–1900. https://doi.org/10.1056/NEJMoa1201832
4. Koren MJ, Scott R, Kim JB et al (2012) Efficacy, safety, and tolerability of a monoclonal antibody to proprotein convertase subtilisin/kexin type 9 as monotherapy in patients with hypercholesterolaemia (MENDEL): a randomised, double-blind, placebo-controlled, phase 2 study. Lancet 380(9858):1995–2006. https://doi.org/10.1016/S0140-6736(12)61771-1
5. Timpson NJ, Walter K, Min JL et al (2014) A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. Nat Commun 5:4871. https://doi.org/10.1038/ncomms5871
6. Gilly A, Ritchie GR, Southam L (2016) Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. Hum Mol Genet 25(11):2360–2365. https://doi.org/10.1093/hmg/ddw088
7. Tachmazidou I, Dedoussis G, Southam L et al (2013) A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. Nat Commun 4:2872. https://doi.org/10.1038/ncomms3872
8. Pollin TI, Damcott CM, Shen HQ et al (2008) A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent Cardioprotection. Science 322(5908):1702–1705. https://doi.org/10.1126/science.1161524
9. Jorgensen A, Frikke-Schmidt R, Nordestgaard BG, Tybjaerg-Hansen A (2014) Loss-of-function mutations in Apoc3 and reduced risk of ischemic vascular disease. Atherosclerosis 235(2):E18–E18
10. Crosby J, Peloso GM, Auer PL et al (2014) Loss-of-function mutations in APOC3, triglycerides, and coronary disease. New Engl J Med 371(1):22–31. https://doi.org/10.1056/NEJMoa1307095
11. Gaudet D, Alexander VJ, Baker BF et al (2015) Antisense inhibition of apolipoprotein C-III in patients with hypertriglyceridemia.

New Engl J Med 373(5):438–447. https://doi.org/10.1056/NEJMoa1400283

12. Wetterstrand K (2016) DNA Sequencing Costs: Data from NHGRI Genome Sequencing Program (GSP). http://www.genome.gov/sequencingcostsdata. Accessed 28 Oct 2016
13. Altshuler DM, Durbin RM, Abecasis GR et al (2015) A global reference for human genetic variation. Nature 526(7571):68–74. https://doi.org/10.1038/nature15393
14. UK10K Consortium, Walter K, Min JL et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526 (7571):82–90. https://doi.org/10.1038/nature14962
15. National Heart Lung and Blood Institute (2016) Trans-Omics for Precision Medicine (TOPMed) Program. https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed. Accessed 8 Nov 2016
16. Genomics England (2016) The 100,000 Genomes Project. https://www.genomicsengland.co.uk/the-100000-genomes-project/. Accessed 8 Nov 2016
17. Peplow M (2016) The 100 000 genomes project. BMJ 353. ARTN i1757. https://doi.org/10.1136/bmj.i1757
18. ExAC project pins down rare gene variants (2016). Nature 536(7616):249. https://doi.org/10.1038/536249a
19. Wang H, Liu L, Zhao J et al (2013) Large scale meta-analyses of fasting plasma glucose raising variants in GCK, GCKR, MTNR1B and G6PC2 and their impacts on type 2 diabetes mellitus risk. PLoS One 8(6):e67665. https://doi.org/10.1371/journal.pone.0067665
20. Gillespie JH (2010) Population genetics: a concise guide. Johns Hopkins University Press, Baltimore
21. Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258
22. Varilo T, Peltonen L (2004) Isolates and their potential use in complex gene mapping efforts - commentary. Curr Opin Genet Dev 14 (3):316–323. https://doi.org/10.1016/j.gde.2004.04.008
23. Minster RL, Hawley NL, Su CT et al (2016) A thrifty variant in CREBRF strongly influences body mass index in Samoans. Nat Genet 48 (9):1049–1054. https://doi.org/10.1038/ng.3620
24. Steinthorsdottir V, Thorleifsson G, Reynisdottir I et al (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39(6):770–775. https://doi.org/10.1038/ng2043
25. Holm H, Gudbjartsson DF, Sulem P et al (2011) A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 43(4):316–320. https://doi.org/10.1038/ng.781
26. Huyghe JR, Jackson AU, Fogarty MP et al (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45(2):197–201. https://doi.org/10.1038/ng.2507
27. Moltke I, Fumagalli M, Korneliussen TS et al (2015) Uncovering the genetic history of the present-day greenlandic population. Am J Hum Genet 96(1):54–69. https://doi.org/10.1016/j.ajhg.2014.11.012
28. Moltke I, Grarup N, Jorgensen ME et al (2014) A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature 512(7513):190–193. https://doi.org/10.1038/nature13425
29. Yang J, Bakshi A, Zhu Z et al (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet 47(10):1114–1120. https://doi.org/10.1038/ng.3390
30. McCarthy S, Das S, Kretzschmar W et al (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48(10):1279–1283. https://doi.org/10.1038/ng.3643
31. Abecasis G, Altshuler D, Boehnke M, et al (2016) Exome Chip. http://genome.sph.umich.edu/wiki/Exome_Chip_Design. Accessed 31 Oct 2016
32. Wessel J, Chu AY, Willems SM et al (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun 6:5897. https://doi.org/10.1038/ncomms6897
33. Peloso GM, Auer PL, Bis JC et al (2014) 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 94(2):223–232. https://doi.org/10.1016/j.ajhg.2014.01.009
34. Visscher PM, Brown MA, McCarthy MI et al (2012) Five years of GWAS discovery. Am J

Hum Genet 90(1):7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

35. Ionita-Laza I, McCallum K, Xu B et al (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48(2):214–220. https://doi.org/10.1038/ng.3477

36. Ritchie GRS, Dunham I, Zeggini E et al (2014) Functional annotation of noncoding sequence variants. Nat Methods 11 (3):294–U351. https://doi.org/10.1038/nmeth.2832

37. Kircher M, Witten DM, Jain P et al (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315. https://doi.org/10.1038/ng.2892

38. Ma C, Blackwell T, Boehnke M et al (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol 37 (6):539–550. https://doi.org/10.1002/gepi.21742

39. Bigdeli TB, Neale BM, Neale MC (2014) Statistical properties of single-marker tests for rare variants. Twin Res Hum Genet 17 (3):143–150. https://doi.org/10.1017/thg.2014.17

40. Fisher RA (1922) On the interpretation of chi-squared from contingency tables, and the calculation of P. J R Stat Soc 85(1):87–94. https://doi.org/10.2307/2340521

41. Wang X (2014) Firth logistic regression for rare variant association tests. Front Genet 5:187. https://doi.org/10.3389/fgene.2014.00187

42. Lee S, Fuchsberger C, Kim S et al (2016) An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. Biostatistics 17(1):1–15. https://doi.org/10.1093/biostatistics/kxv033

43. Auer PL, Reiner AP, Leal SM (2016) The effect of phenotypic outliers and non-normality on rare-variant association testing. Eur J Hum Genet 24 (8):1188–1194. https://doi.org/10.1038/ejhg.2015.270

44. O'Connor TD, Kiezun A, Bamshad M et al (2013) Fine-scale patterns of population stratification confound rare variant association tests. PLoS One 8(7):e65834. https://doi.org/10.1371/journal.pone.0065834

45. Zhang Y, Shen X, Pan W (2013) Adjusting for population stratification in a fine scale with principal components and sequencing data. Genet Epidemiol 37(8):787–801. https://doi.org/10.1002/gepi.21764

46. Babron MC, de Tayrac M, Rutledge DN et al (2012) Rare and low frequency variant stratification in the UK population: description and impact on association tests. PLoS One 7(10): e46519. https://doi.org/10.1371/journal.pone.0046519

47. Liu Q, Nicolae DL, Chen LS (2013) Marbled inflation from population structure in gene-based association studies with rare variants. Genet Epidemiol 37(3):286–292. https://doi.org/10.1002/gepi.21714

48. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44(3):243–246. https://doi.org/10.1038/ng.1074

49. Tintle N, Aschard H, Hu I et al (2011) Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 genomes project exon sequencing data in unrelated individuals: summary results from group 7 at genetic analysis workshop 17. Genet Epidemiol 35 Suppl 1:S56–S60. https://doi.org/10.1002/gepi.20650

50. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063):1299–1320. https://doi.org/10.1038/nature04226

51. Pe'er I, Yelensky R, Altshuler D et al (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32 (4):381–385. https://doi.org/10.1002/gepi.20303

52. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32 (3):227–234. https://doi.org/10.1002/gepi.20297

53. Fadista J, Manning AK, Florez JC et al (2016) The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet 24(8):1202–1205. https://doi.org/10.1038/ejhg.2015.269

54. Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 15(5):335–346. https://doi.org/10.1038/nrg3706

55. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80(4):727–739. https://doi.org/10.1086/513473

56. Nelson MR, Wegmann D, Ehm MG et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337(6090):100–104. https://doi.org/10.1126/science.1217876

57. Fu W, O'Connor TD, Jun G et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493(7431):216–220. https://doi.org/10.1038/nature11690

58. Bansal V, Libiger O, Torkamani A et al (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11(11):773–785. https://doi.org/10.1038/nrg2867

59. Wu G, Zhi D (2013) Pathway-based approaches for sequencing-based genome-wide association studies. Genet Epidemiol 37(5):478–494. https://doi.org/10.1002/gepi.21728

60. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34(2):188–193. https://doi.org/10.1002/gepi.20450

61. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615(1–2):28–56. https://doi.org/10.1016/j.mrfmmm.2006.09.003

62. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311–321. https://doi.org/10.1016/j.ajhg.2008.06.024

63. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384. https://doi.org/10.1371/journal.pgen.1000384

64. Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70(1):42–54. https://doi.org/10.1159/000288704

65. Magi R, Asimit JL, Day-Williams AG et al (2012) Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. Genet Epidemiol 36(8):785–796. https://doi.org/10.1002/gepi.21675

66. Zawistowski M, Gopalakrishnan S, Ding J et al (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87(5):604–617. https://doi.org/10.1016/j.ajhg.2010.10.012

67. Magi R, Kumar A, Morris AP (2011) Assessing the impact of missing genotype data in rare variant association analysis. BMC Proc 5 (Suppl 9):S107. https://doi.org/10.1186/1753-6561-5-S9-S107

68. Asimit JL, Day-Williams AG, Morris AP et al (2012) ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered 73 (2):84–94. https://doi.org/10.1159/000336982

69. Hoffmann TJ, Marini NJ, Witte JS (2010) Comprehensive approach to analyzing rare genetic variants. PLoS One 5(11):e13584. https://doi.org/10.1371/journal.pone.0013584

70. Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89(3):354–367. https://doi.org/10.1016/j.ajhg.2011.07.015

71. Liu DJ, Leal SM (2010) 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 6(10): e1001156. https://doi.org/10.1371/journal.pgen.1001156

72. Price AL, Kryukov GV, de Bakker PI et al (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86(6):832–838. https://doi.org/10.1016/j.ajhg.2010.04.005

73. Neale BM, Rivas MA, Voight BF et al (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7(3):e1001322. https://doi.org/10.1371/journal.pgen.1001322

74. Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35(7):606–619. https://doi.org/10.1002/gepi.20609

75. Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762–775. https://doi.org/10.1093/biostatistics/kxs014

76. Ladouceur M, Dastani Z, Aulchenko YS et al (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet 8(2):e1002496. https://doi.org/10.1371/journal.pgen.1002496

77. Wu MC, Lee S, Cai T et al (2011) Rare-variant association testing for sequencing

data with the sequence kernel association test. Am J Hum Genet 89(1):82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

78. Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33(6):497–507. https://doi.org/10.1002/gepi.20402

79. Ionita-Laza I, Lee S, Makarov V et al (2013) Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet 92(6):841–853. https://doi.org/10.1016/j.ajhg.2013.04.015

80. Hasegawa T, Kojima K, Kawai Y et al (2016) AP-SKAT: highly-efficient genome-wide rare variant association test. BMC Genomics 17(1):745. https://doi.org/10.1186/s12864-016-3094-3

81. Lee S, Emond MJ, Bamshad MJ et al (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91(2):224–237. https://doi.org/10.1016/j.ajhg.2012.06.007

82. Sun J, Zheng Y, Hsu L (2013) A unified mixed-effects model for rare-variant association in sequencing studies. Genet Epidemiol 37(4):334–344. https://doi.org/10.1002/gepi.21717

83. Li H, Chen J (2016) Efficient unified rare variant association test by modeling the population genetic distribution in case-control studies. Genet Epidemiol 40(7):579–590. https://doi.org/10.1002/gepi.21995

84. Moutsianas L, Agarwala V, Fuchsberger C et al (2015) The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet 11(4):e1005165. https://doi.org/10.1371/journal.pgen.1005165

85. Greco B, Hainline A, Arbet J et al (2016) A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architectures. Eur J Hum Genet 24(5):767–773. https://doi.org/10.1038/ejhg.2015.194

86. Dering C, Konig IR, Ramsey LB et al (2014) A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required. Front Genet 5:323. https://doi.org/10.3389/fgene.2014.00323

87. Yi N, Zhi D (2011) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35(1):57–69. https://doi.org/10.1002/gepi.20554

88. Quintana MA, Berstein JL, Thomas DC et al (2011) Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. Genet Epidemiol 35(7):638–649. https://doi.org/10.1002/gepi.20613

89. Chen LS, Hsu L, Gamazon ER et al (2012) An exponential combination procedure for set-based association tests in sequencing studies. Am J Hum Genet 91(6):977–986. https://doi.org/10.1016/j.ajhg.2012.09.017

90. Logsdon BA, Dai JY, Auer PL et al (2014) A variational Bayes discrete mixture test for rare variant association. Genet Epidemiol 38(1):21–30

91. Fan R, Wang Y, Mills JL et al (2013) Functional linear models for association analysis of quantitative traits. Genet Epidemiol 37(7):726–742. https://doi.org/10.1002/gepi.21757

92. Luo L, Zhu Y, Xiong M (2012) Quantitative trait locus analysis for next-generation sequencing with the functional linear models. J Med Genet 49(8):513–524. https://doi.org/10.1136/jmedgenet-2012-100798

93. Luo L, Boerwinkle E, Xiong M (2011) Association studies for next-generation sequencing. Genome Res 21(7):1099–1108. https://doi.org/10.1101/gr.115998.110

94. Fan R, Wang Y, Mills JL et al (2014) Generalized functional linear models for gene-based case-control association studies. Genet Epidemiol 38(7):622–637. https://doi.org/10.1002/gepi.21840

95. Vsevolozhskaya OA, Zaykin DV, Greenwood MC et al (2014) Functional analysis of variance for association studies. PLoS One 9(9):e105074. https://doi.org/10.1371/journal.pone.0105074

96. Fan R, Wang Y, Boehnke M et al (2015) Gene level meta-analysis of quantitative traits by functional linear models. Genetics 200(4):1089–1104. https://doi.org/10.1534/genetics.115.178343

97. Wang Y, Liu A, Mills JL et al (2015) Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genet Epidemiol 39(4):259–275. https://doi.org/10.1002/gepi.21895

98. Wijsman EM (2012) The role of large pedigrees in an era of high-throughput sequencing. Hum Genet 131(10):1555–1563. https://doi.org/10.1007/s00439-012-1190-2

99. De G, Yip WK, Ionita-Laza I et al (2013) Rare variant analysis for family-based design. PLoS One 8(1):e48495. https://doi.org/10.1371/journal.pone.0048495

100. Ionita-Laza I, Lee S, Makarov V et al (2013) Family-based association tests for sequence data, and comparisons with population-based association tests. Eur J Hum Genet 21 (10):1158–1162. https://doi.org/10.1038/ejhg.2012.308

101. Chen H, Meigs JB, Dupuis J (2013) Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37(2):196–204. https://doi.org/10.1002/gepi.21703

102. Schaid DJ, McDonnell SK, Sinnwell JP et al (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 37(5):409–418. https://doi.org/10.1002/gepi.21727

103. Feng S, Pistis G, Zhang H et al (2015) Methods for association analysis and meta-analysis of rare variants in families. Genet Epidemiol 39(4):227–238. https://doi.org/10.1002/gepi.21892

104. Svishcheva GR, Belonogova NM, Axenovich TI (2014) FFBSKAT: fast family-based sequence kernel association test. PLoS One 9(6):e99407. https://doi.org/10.1371/journal.pone.0099407

105. Oualkacha K, Dastani Z, Li R et al (2013) Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. Genet Epidemiol 37 (4):366–376. https://doi.org/10.1002/gepi.21725

106. Jiang D, McPeek MS (2014) Robust rare variant association testing for quantitative traits in samples with related individuals. Genet Epidemiol 38(1):10–20. https://doi.org/10.1002/gepi.21775

107. Svishcheva GR, Belonogova NM, Axenovich TI (2015) Region-based association test for familial data under functional linear models. PLoS One 10(6):e0128999. https://doi.org/10.1371/journal.pone.0128999

108. Chen H, Lumley T, Brody J et al (2014) Sequence kernel association test for survival traits. Genet Epidemiol 38(3):191–197. https://doi.org/10.1002/gepi.21791

109. Cai T, Tonini G, Lin X (2011) Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67(3):975–986. https://doi.org/10.1111/j.1541-0420.2010.01544.x

110. Lin X, Cai T, Wu MC et al (2011) Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet Epidemiol 35(7):620–631. https://doi.org/10.1002/gepi.20610

111. Fan R, Wang Y, Yan Q et al (2016) Gene-based association analysis for censored traits via fixed effect functional regressions. Genet Epidemiol 40(2):133–143. https://doi.org/10.1002/gepi.21947

112. Liu DJ, Leal SM (2015) Replicating sequencing-based association studies of rare variants. In: Zeggini E, Morris A (eds) Assessing rare variation in complex traits: design and analysis of genetic studies. Springer, New York, NY, pp 201–213. https://doi.org/10.1007/978-1-4939-2824-8_14

Author Queries

Chapter No.: 5	394545_1_En

Query Refs.	Details Required	Author's response
AU1	Please check whether the affiliation and correspondence details are presented correctly.	
AU2	References were renumbered for sequential purpose. Please check.	
AU3	Please check and provide appropriate Section cross-link throughout the chapter.	
AU4	Please check the edit made to the footnote.	
AU5	Please check we have deleted "‡" in the text since its significance is provided in Table 1.	
AU6	Please check this term "f()"	