Published in final edited form as:
Nat Rev Genet. 2010 November ; 11(11): 773–785. doi:10.1038/nrg2867.

Statistical Analysis Strategies for Association Studies Involving Rare Variants

Vikas Bansal[*], **Ondrej Libiger**[*], **Ali Torkamani**[*], and **Nicholas J. Schork**
The Scripps Translational Science Institute (VB, OL, AT, NJS), Scripps Health (VB, AT, NJS), Department of Molecular and Experimental Medicine, The Scripps Research Institute (OL, AT, NJS); Lekarska Fakulta v Hradci Kralove, Charles University, Czech Republic (OL)

Abstract

The limitations of genome-wide association (GWA) studies that focus on the phenotypic influence of common genetic variants have motivated human geneticists to consider the contribution of rare variants to phenotypic expression. The increasing availability of high-throughput sequencing technology has enabled studies of rare variants, but will not be sufficient for their success since appropriate analytical methods are also needed. We consider data analysis approaches to testing associations between a phenotype and collections of rare variants in a defined genomic region or set of regions. Ultimately, although a wide variety of analytical approaches exist, more work is needed to refine them and determine their properties and power in different contexts.

Introduction

Despite the success of genome wide association (GWA) studies in identifying common single nucleotide variants (SNVs) that contribute to complex diseases[1], the vast majority of genetic variants contributing to disease susceptibility are yet to be discovered. In fact, it has been argued that these variants are not likely to be captured in current GWA study paradigms that focus on common SNVs.[2] It is now widely believed that many genetic and epigenetic factors are likely to contribute to common complex diseases, including multiple rare SNVs (defined by convention as those that have frequencies < 1%), copy number variations (CNVs), and other forms of structural variation. [3–12] Irrespective of how one might define 'rare variant' (which, although we have adopted the convention <1% frequency, might range from <0.1% to <0.01% depending on the context[13]) it is essential to recognize that such variants likely contribute to phenotypic expression in conjunction with, or over-and-above, common variants. This consideration has important implications when designing a study or choosing a statistical method for analyzing associations involving rare variants.

There are many reasons to believe that multiple rare variants, both within the same gene and across different genes, collectively influence the expression and prevalence of traits and diseases in the population at large. First, it has been argued that population phenomena, such as the recent expansion of the human population, are likely to have resulted in a large number of segregating, functionally-relevant, rare variants that mediate phenotypic variation.[14, 15] Second, the discovery of rare independent somatic mutations within and across genes contributing to tumorigenesis may parallel the functional effects of inherited

Address correspondence to: Nicholas J. Schork, Ph.D., The Scripps Translational Science Institute, Department of Molecular and Experimental Medicine, The Scripps Research Institute, 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037, nschork@scripps.edu, 858-554-5705, 858-546-9284 (fax).
[*]These authors contributed equally to this review

variants contributing to congenital disease.[11, 16, 17] Third, the identification of multiple rare variants within the same gene contributing to largely monogenic disorders such as Cystic Fibrosis and BRCA1 and BRCA2-associated breast cancer[18, 19] suggests that rare variants might also influence common complex traits and diseases. Fourth, the identification of multiple functional variants within the same gene and the association of these variants with both *in vitro* and clinical phenotypes indicates that multiple rare variants could influence general clinical phenotypic expression[20]. Fifth, importantly, sequencing studies focusing on specific genes have shown that collections of rare variants can indeed associate with particular phenotypes (Table 1).

To comprehensively characterize the contribution of rare variants to phenotypic expression, one could either sequence genomic regions of interest using high-throughput DNA sequencing technologies[21] or genotype common and rare variants identified in previous sequencing studies using custom genotyping chips. There are a number of ways to approach association studies involving rare variants, which are independent of sequencing or genotyping technology. For example, one could: focus on candidate disease genes [22]; focus on genomic regions implicated in linkage or genome-wide association studies, under the assumption that phenotypically-relevant rare variants also exist in those regions; consider multiple functional genomic regions, such as exons [23]; or study entire genomes.[12, 24] The sampling framework for such studies is also extremely important as one could focus on: cases and controls, possibly in DNA pools[22] or with oversampling of controls to achieve greater power in studies of rare diseases; individuals phenotyped for a particular quantitative trait; individuals with ‘extreme’ phenotype values in order to increase efficiency[25, 26]; or families in order to exploit parent-offspring transmission patterns.[12, 24]

In addition to a sequencing technology and an appropriate sampling and study design, bioinformatic methods for analyzing the potentially massive amounts of sequence data likely to be generated in a study are needed, as are algorithms for accurately identifying rare variants and assigning genotypes to individuals from sequence data[12, 27]. Importantly, statistical analysis methods for relating rare variants to phenotypes of interest are needed. Association analyses involving rare variants are not as straightforward as analyses involving common variations since the power to detect an association between a single rare variant is low in even very large samples (Figure 1).[14, 28, 29] Therefore, researchers have begun to develop data analysis strategies that assess the collective effects of multiple rare variants within and across genomic regions [13, 28, 30]. This challenge of statistical analysis is the focus of this Review.

There are many settings in which a collection of rare variants might exhibit an association with a trait. Of the many different methods that could be used for testing associations, not all of them are likely to work well in each of these settings. Here, we consider the rationales behind different data analysis methods, pointing out their limitations and advantages. We also outline areas for further research. As noted, appropriately sophisticated methods for identifying variants, assigning genotypes, and sampling individuals are crucial for rare variant analyses, but we do not discuss them here. There are, however, a few additional issues that researchers need to consider in any association study involving rare variants, as briefly described in Box 1. Finally, although we focus on the analysis of rare SNVs, aspects of the analytical methods discussed can be used with other forms of variation including rare CNVs, although certain caveats apply, which we mention briefly.

Box 1

Issues Impacting the Interpretation of Rare-Variant Association Studies

There are a number of statistical analysis issues that go beyond the choice of an association test statistic in studies of rare variants. These are outlined briefly below.

Sequencing and Genotyping errors

It has been shown that differential genotyping error rate can have substantial impact on common-variant based GWA studies.[89] Given that current sequencing protocols have inherent error rates, more research is needed to understand how false positive variant calls and nucleotide misassignments in sequence-based association studies of rare variants will impact inferences.

Phasing

Rare variant effects can manifest as compound heterozygosity,[90] the 'unmasking' of deleterious variants via deletions on a homologous chromosome[12], and other haplotype context-dependent phenomena. Thus, leveraging phase information in an association study of rare variants may be crucial, but obtaining phase from sequence data alone is not trivial.[24, 91–93]

Stratification

The potential for false positive associations due to population stratification is large in studies involving rare variants since specific rare variants are more likely to be unique to a particular geoethnic group. Thus, even if focus in a rare variant study is on a particular gene or genomic region, it is important to genotype the individuals in the study on enough additional markers to assess and control for stratification using standard strategies.[94, 95]

The Use of In Silico Controls

The practice of identifying and quantifying allele frequencies in a group of individuals and comparing them with historical or publicly available 'control' sets in studies involving rare variants is highly problematic due to the potential for stratification and sampling variation effects.[96] In order to avoid this, either sophisticated genetic background matching strategies or *de novo* sequencing of a case and control group are recommended, but more work in this area is needed.

Genomic Units of Analysis

Different strategies for testing a genomic region for association involving rare variants exist. For example, one could test all the variants in a region (depending on its size) for collective frequency differences between, e.g., cases and controls, define particular regions of interest, such as exons or transcription factor binding sites (Box 2), or pursue a 'moving window' analysis in which variants in contiguous, possibly overlapping, subregions are tested. Each of these strategies impacts the number and nature of multiple testing problems.

Box 2

In Silico Functional Assessment of Sequence Variations

Identifying groups of variants that reside in genomic regions known or likely to be of functional significance, such as exons, promoters, enhancers, etc. can be pursued through the use of genome browsers such as the UCSC genome browser. One can also assess the more specific functional potential of individual sequence variants given

their sequence contexts and incorporate this information into an association analysis (e.g., by weighting them more heavily in test statistics). The table below lists web resources for such assessments. Finally, one could identify variants that participate in common multigene pathway and processes and assess their collective effects on a phenotype.

Functional Element Annotation

Beyond the basic annotations presented in the UCSC genome browser, numerous prediction methods exist for transcription factor binding sites exist (TFsearch, Consite:[100], *TRANSFAC*:[101], enhancers (*VISTA Enhancer Browser*:[102]), microRNAs (miRBase:[103]), microRNA binding sites (Targetscan:[104]), intronic splice sites[105], and exonic splicing enhancers[106, 107], silencers[108, 109], regulatory elements[110–112] (Table B.1). Epigenetic and/or regulatory factors derived from the ENCODE project[113], such as histone binding/methylation/acylation, CpG islands, nuclease accessible sites, transcription start sites, and others are also available through the UCSC Genome Browser[114].

Pathway and Process Assessment

There are numerous resources for pathway information and analysis. Open source databases that include pathway information, but not necessarily analysis of datasets, include Reactome [115], BioCarta and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [116], as well as a biological process resource, The Gene Ontology (GO) database[117]. Publically available pathway analysis tools that link to these databases include, but are not limited to, Cytoscape [118], GenMAPP [119], and the DAVID Bioinformatics Resource [120]. Commercially available tools that build off these databases and include proprietary pathway information include Ingenuity Pathway Analysis and GeneGo by MetaCore. For a more complete review of pathway analysis tools, see Suderman and Hallet.[121]

Functional Impact Prediction Modeling

Functional predictions often leverage various types of information, including but not limited to protein structure information, sequence conservation, motif conservation, etc., in order to build models that generate a probability that a particular variant is functionally important. Some of these methods, and many integrative web servers for this purpose, have been reviewed.[122–124] Functional prediction for non-coding variants are generally limited to scoring the deviation of a polymorphism from known regulatory factor motifs, and examples are limited but include *MaxEntScan* for splicing prediction[105], or *RAVEN* for regulatory regions.[125]

Generality of Annotators

A number of webservers and algorithms attempt to integrate the various functionally-relevant genomic features in order to explicitly weight or prioritize variants investigated in an association study. A subset of the tools attempt to prioritize SNPs based upon scores returned from the various functional impact predictors while many simply present the functional elements and leave it up to the user to draw their own conclusions about ultimate functionality. A few tools, such as *SeattleSeq* and *Sequence Variant Analyzer* integrate various types of biological data in order to annotate novel sequence variants, whereas *Trait-O-Matic* annotates variations with respect to overt phenotypic features that they have been associated with.

Imputation

There is a great deal of precedent for assigning individuals who have not been sequenced or genotyped at a specific locus common genotypes based on available neighboring locus genotype information and linkage disequilibrium patterns via imputation methods.[97] Although highly problematic in situations involving *de novo* or even moderately rare variants (<1%), imputation methods involving rare variants have begun to receive attention and could be extremely useful in future association studies.[98]

Accommodating Multiple Comparisons

Controlling for false positive findings due to multiple testing is necessary. Pre-specified Bonferroni-like corrections on association p-values are not likely to be appropriate given possible correlations between defined groups of rare variants and/or overlapping windows to be tested. Such correlations will also impact false discovery rate (FDR) procedures for accommodating multiple testing *a posteriori*.[99] Simulation studies and permutation testing that consider the entire set of tests performed (e.g., all windows and groups of variants across all genomic regions considered) to a get a global false positive rate are the most appropriate given their flexibility and sound theoretical bases, but will likely be very computationally intensive.[75] More work in this area is also sorely needed.

Capturing the Effects of Rare Variants

The nature of the effect of rare variants

As noted, rare variants are likely to influence a trait along with common variants.[4, 14] In addition, just as interaction effects involving either genetic or environmental factors must be considered in standard GWA studies[9], they are also likely to be important in association studies involving rare variants. With these facts in mind, there are a number of different settings in which rare variants within a defined genomic region could influence a phenotype. Figure 2 provides a few contrasting examples, including situations in which: a common variant is associated with a phenotype; rare variants influence a phenotype independently of one another; rare variants, along with variants with more moderate or common frequencies, act synergistically to influence a phenotype; or only a subset of the rare variants influences the phenotype due to their locations in a functional element within the region of interest.

Of these possible settings, the one receiving the most attention by statistical geneticists is the ‘extreme allelic heterogeneity (EAH)’ setting in which single or small subgroups of individuals with a particular phenotype or disease possess any one, or some subset, of a larger set of rare variants that all independently perturb a single relevant gene in a similar way.[12, 31] Although conceptually easier to accommodate in statistical analysis models, there is no reason to believe that the EAH setting is the rule rather than the exception with respect to rare variant influences on phenotypic expression. Statistical analysis models and methods for rare variant association studies should therefore be developed and tested in settings that go beyond the EAH model, such as settings implicating synergistic effects of rare (and common) variants within (and across) genomic regions.

Single locus tests vs. ‘collapsing’ sets of rare variants

The simplest approach to testing rare variants for association with a trait is to test them individually using standard contingency table and regression methods of the sort implemented in widely used genetic data analysis packages such as PLINK.[32] This strategy is highly problematic given, for example, the poor power that such statistical tests have to detect small rare variant frequency differences between diagnostic or phenotypic groups (figures 1A and 1B).[14, 28, 29] In order to overcome power issues associated with testing rare variants individually, one could ‘collapse’ sets of rare variants into a single group and test their collective frequency differences between cases and controls.[28, 30] In its simplest form

this strategy could involve counting individuals possessing a rare variant at any position in the genomic region of interest, calculating the frequencies of these individuals, for example in case and control groups, and then testing the two groups for frequency differences. This strategy forms the basis for most of the statistical models described in this review and variations of it have been considered in many studies involving rare variants (Table 1). To make this collapsing strategy more biologically appealing, elaborate ways of leveraging functional elements and annotations in a genomic region to collapse the variants together can be exploited (see below and Box 2). The effect of collapsing variants and testing their collective frequency differences on power can be substantial, as depicted in Figure 1B.

Quantitative Traits and Conditional Analysis

Regression-based collapsed variant and conditional tests can greatly enhance association studies involving rare variants. Consider Figure 1C which plots the power to detect the effect of a variant on a quantitative trait for 1000 individuals as a function of the fraction of variation of the quantitative trait explained by that variant. If a set of rare variants each individually explain only a small fraction of the variation of the trait, they could be combined into a single predictor variable, perhaps by creating a dummy variable which equals 1 if an individual possesses any of the variants and 0 otherwise.[33] This strategy should increase the fraction of variation explained by the variants as a whole and hence increase the power to detect their collective, rather than individual, effects. In addition, if one included other factors in a regression model - such as covariate effects, the effects of previously identified common variants, or other collapsed sets of rare variants - then the power to detect the association involving rare variants could increase substantially (Figure 1C). Not all analysis methods proposed for rare variant studies, however, can accommodate additional factors in their formulations and hence leverage conditioning effects. In addition, not all models can accommodate quantitative trait analysis unless the phenotype is broken into quantiles and stratified analysis is pursued (Table 2).

Defining Collapsing Sets of Rare Variants via Function or Proximity

The collapsing strategy makes important assumptions. First, some formulations of collapsed tests assume that each subject is likely to have only a single rare variant. This may be true given the low frequency of the variants, but could in theory be untrue if the variants interact with one another or large genomic regions are tested.[20, 33] Second, if one collapses variants by counting individuals possessing rare variants, then if either the frequency of those variants is large enough or if there are many of them, the percentage of individuals possessing any one of them could reach 100%. Therefore, ways of circumscribing the variants to be collapsed, such as leveraging functional information (Box 2), or weighting the variants in some way,[34, 35] are important. Alternatively, one could employ statistics that do not rely on simple counting. For example, one could tally the number of variants within a collapsed set possessed by each individual.[33]

Although there are a number of ways to leverage functional annotations to guide the collapsing of rare variants in association studies, their use will only be as good as the science behind those annotations. It is also possible that different functional ‘levels’ of annotation can be used to define collapsed sets of rare variants. For example, one could define a set of variants as ‘genic’ if they reside in the open reading frame associated with a gene; as ‘exonic’ if they reside in coding regions within that frame; as ‘non-synonymous coding variants’ if they perturb an encoded amino acid; and as ‘non-synonymous coding within an active site of the encoded protein’ if a variant impacts a residue within the active site of the encoded protein. With this in mind, one could perhaps test hierarchies of hypotheses about collections of variants and their biological impact on a phenotype.

It is important to note the distinction between leveraging functional annotations to collapse a set of rare variants based on their location versus predictions that the variants themselves have a functional effect (Box 2).[35] In fact, two recent papers[23, 36] suggest that leveraging functional annotations and computational methods for predicting the consequences of specific rare variants can be used to great advantage in the identification of disease-predisposing variants, at least for rare monogenic conditions. Functional annotations for rare CNVs and other forms of structural variation can also be leveraged in collapsed or group-wise analyses. However, many of these forms of variation are thought to exert or manifest their effects throughout the genome and not necessarily as a group of variants in a singular region of the genome. Thus, pathway-based (Box 2) and other higher-order approaches to collapsing or summarizing rare CNV effects have been proposed, especially in the context of neuropsychiatric disease.[3, 37]

Specific analysis models

There are a number of statistical analysis strategies that can be used to test the hypothesis that specific collections of rare variants are associated with a particular trait or disease. Some of these methods have been developed in contexts beyond human association studies, such as assessing genetic differentiation between human geoethnic groups or pathogen sequences. In addition, some methods are more or less agnostic to variant frequencies. In order to facilitate their descriptions, we have grouped various methods together in three broad and somewhat arbitrary categories: tests based on the use of group summary information on variant frequencies compared between, for example, case and control groups; tests based on the similarity or diversity of unique DNA sequences possessed by different individuals; and regression models that consider collapsed sets of variants and other factors as predictors of a phenotype. We consider each of these three categories separately below, although Table 2 provides brief summaries of representative methods from each category. Each of the methods discussed can leverage functional annotations to define collapsed variant sets or can be used in a moving window setting (Box 2).

Box 3

Measures of Diversity and Genomic Similarity

Exploiting sequence similarity or diversity in genetic association studies can be problematic due to the fact that the choice of a similarity or diversity measure can impact the interpretation of the results. This issue is well-documented in the cluster analysis literature[59, 126] but has been shown to influence the interpretation of genomic studies as well. For example, the determination of phylogenetic patterns among different species based on DNA sequences requires the choice of a DNA sequence alignment method in order to identify patterns of orthology, and it has been shown that, depending on how DNA similarities are defined and the alignments are determined, different conclusions can be drawn about the phylogenetic, and hence evolutionary, relationships between species.[127]

For within-species studies assessing the ancestral relationships between populations based on DNA sequence, it has been shown that the choice of a distance measure can impact the interpretation of the results[50, 128]. Measures of nucleotide similarity for the comparison of DNA sequences between pairs of individuals within a species are also problematic for this reason. This issue is no less problematic for the assessment of the difference in the diversity of DNA sequences obtained from two or more groups of individuals when summary allele frequency measures are used.[50] For example, consider the classical general formula for diversity measures[129, 130] for a single population:

$$\Delta=\left(\sum_{i=1}^{k} p_i^q\right)^{\frac{1}{1-q}}$$

Where p_i is the frequency of the ith allele out of a total of k (i=1,…, k) and the exponent q determines the Δ measure's sensitivity to the frequency of the alleles. Thus, the use of q values less than 1.0 produces a measure that emphasizes rare variants and the use of λ values greater than 1.0 produces a measure that emphasizes common variants.[50, 129] The use of different q values in the construction of Δ measures for the comparison of the genetic diversities of two (or more) populations will have the same effect[50, 130, 131]: small q values will impact differences in rare variants and large q values emphasize differences in common variants[132]. Since a genomic region may harbor common, moderately common, and rare variants, some of which may influence phenotypic expression, the choice of a q value for association studies based on diversity indices may be problematic.

Methods based on summary statistics

Morgenthaler and Thilly[30] were the first to describe a version of the collapsing approach in which the frequency of individuals carrying any one of a number of rare variants is contrasted between case and control groups. They termed this approach the 'cohort allelic sums test' or 'CAST' method and suggested the use of standard contingency table-based Chi-square or Fisher's exact tests for obtaining p-values. The method as first proposed does not easily accommodate covariates, cannot be used with quantitative phenotypes, and does not consider weighting of the variants using, for example, variant frequency or functional annotations. Li and Leal considered an extension of the CAST method, which they termed the 'Combined Multivariate and Collapsing (CMC)' method.[28] Here, rare variants are collapsed, as in the CAST method, and treated as a single set of variants whose frequency differences are then tested between groups. This testing could potentially be done simultaneously with frequency differences at other individual loci or among other collapsed sets using a summary distance-based Hotelling's T-Squared statistic.[28, 38] The CMC statistic has desirable properties in that it appropriately controls type I error rates even when non-functional variants are included in the set of variants to be tested, and has better power than the standard CAST method. In addition, the CMC statistic can be implemented in a regression modeling framework as discussed later.

Madsen and Browning proposed a statistic for testing a prespecified collapsed set of variants that leverages weighting of each variant by its frequency, thus allowing one to include variants of any frequency into the collapsed set.[34] A score is calculated for each individual using that individual's genotypes and the frequency-determined weights. The sum of ranks of the scores among the cases is then used as a summary statistic to be compared to the same statistic computed among the controls using permutation methods, in a manner analogous to the Wilcoxon rank test.[39] Madsen and Browning showed that their proposed statistic is more powerful than either the CAST or CMC methods in a number of settings, but more work in this area is needed to clarify the advantages, if any, of each.[34] Other strategies for testing groupwise frequency differences of genetic variations between cases and controls in an analogous manner to the CAST method have been proposed, although many have only been implemented in settings involving common variants. [34, 40, 4142]

Recently, Price et al.[35] implemented a method for testing rare coding variants that considers optimal or variable weighting of the variants in a procedure resembling Madsen and

Browning's.[34] Price et al.[35] showed that their method is more powerful than approaches that consider fixed weights. In addition, they argued that the use of the predicted functional impact of each individual non-synonymous coding variant could be leveraged in their model. Finally, Han and Pan[40] recently devised a method that cleverly considers the direction of the effect of the implicated variants (e.g., protective or deleterious) which can be implemented in a regression model framework (see below). Other summary statistic methods essentially ignore direction of effect and hence may be problematic in settings in which rare variants are not necessarily more frequent in disease or certain *a priori* defined phenotypic states.

Another way of exploiting summary statistics for rare variant analysis involves comparing haplotype frequencies between, for example, case and control groups, as opposed to genotype or single variant carrier status frequencies.[43–45] Haplotype analyses require phase information, which is not trivial to obtain for genotyped rare variants or variants derived from sequence data (Box 1). In addition, if enough rare variants are studied, each individual in a sample of cases and controls may have their own unique haplotypes, making summary statistic approaches impossible. A recently proposed two-stage approach to haplotype analysis of rare variants could alleviate this problem since it collapses haplotypes into groups and eliminates variants not likely to be relevant prior to contrasting haplotype frequencies.[46]

Other potential methods that leverage summary statistics to test multiple variant frequency differences across groups include classical DNA sequence diversity measures such as nucleotide polymorphism, θ, and nucleotide diversity, π[47], as well as traditional measures of population differentiation such as that statistics referred to as Fst and Gst.[48, 49] These methods are more or less agnostic to allele frequencies, but can provide insight into differences between groups over many rare variants. However, their utility and power have not been assessed in association analysis settings. In addition, flaws with measures such as Fst and Gst have been pointed out that may not allow them to reliably capture diversity, differences in diversity, or population differentiation in general in some of the most trivial settings, given their focus on heterozygosity.[50] Jost[50] discusses alternatives to traditional Fst, Gst and related DNA sequence population differentiation measures, but these measures still require assumptions about the best way to apply them in any one particular setting. Interestingly, the methods described by Jost can be easily adjusted to assess group differences attributable to many rare variants (see Box 3).[50]

Approaches based on similarities among individual sequences

Instead of constructing statistics based on the frequencies of individual or collapsed variants, statistics that reflect the similarity of the unique DNA sequences possessed by individuals can be constructed. Such statistics have their roots in the assessment of cross-species orthology, protein family determination, phylogeny construction and a number of other molecular genetic analyses based on DNA sequence similarity and are more or less agnostic to the frequencies of the variants being considered.[20, 51] The main motivation for similarity-based approaches to assessing rare variant associations is that the general nucleotide background or context within which a rare variant can influence a phenotype may be important. Thus, such approaches assume some form of interaction among variants or at least a simple shaping of gene function by the balance of variations an individual possesses.

Many recent papers have described flexible strategies for testing genetic associations that leverage individual sequence similarity information,[20, 52–57] and it has been shown that such strategies can be as powerful, if not more so, than some traditional tests of association in many settings involving common variations.[58] However, the performance of these methods when many rare variants and no common variants are considered is an open question. In

addition, a limitation of these methods is that a specific DNA similarity or distance measure or metric must be chosen and this can be problematic (Box 3).[59] For example, a number of approaches have described DNA sequence similarity metrics that consider the origins or phylogenetic relationships between sequences.[60–62] In addition, other approaches, some of which have their roots in comparing pathogen sequences, consider weighting individual nucleotides by their frequency or putative functional effects.[54, 63, 64]

The problem of choosing a DNA sequence similarity measure based purely on nucleotide content matching or genealogical or cladistic distance is rooted in the fact that, ultimately, functional nucleotide content (i.e., what nucleotides and nucleotide combinations an individual possesses that impact function) determines gene activity, rather than the phylogenetic origins of those nucleotides. Thus, in theory, similarity measures that build off the functional features and functional capacities of impacted genes associated with DNA sequence (Box 2) – as shaped by particular nucleotides and nucleotide combinations – are likely to be more appropriate for association studies than measures based on either phylogenetic relationships between sequences or the mere equality of aligned nucleotides.

Alternatively, statistics that exploit pairwise sequence similarity can be used[65] as alternatives to classical summary statistic measures of sequence diversity differences between groups. Such statistics would be highly appropriate in situations, such as the EAH situation, in which a group of individuals (e.g., cases) are hypothesized to simply possess more unique variants or more unique combinations of variants than another group of individuals (e.g., controls) in a defined genomic region.

In the absence of knowledge of which rare variants to collapse or consider as a set, one could potentially search for a subset of variants that maximally discriminates between, for example cases and controls, based on the distances between the sequences in the two groups.[66] Permutation methods could be used to derive p-values for discriminative ability. Searches for optimal sets of variations in this manner have parallels to the approach underlying logic regression[67] and the method of Han and Pan[40], which are discussed later in the section on regression methods. Although intuitively appealing, such methods are problematic in that the determination of an optimal subset of variants based on group differences can be computationally-intensive. In addition, if a large enough genomic region is considered, then one could merely ‘collapse’ all variants unique to each case and then unique to each control, resulting in a set of variants that completely and perfectly discriminate cases from controls. The possibility of this phenomenon emphasizes a need for considering functional annotations in relevant data analyses or other ways of circumscribing rare variants to be considered as a collapsed set.

Finally, traditional family-based linkage analyses consider the consistency of within-family sharing of specific transmitted chromosomal segments among affected family members rather than the consistency or similarity of the nucleotide content of those segments across different families. As a result, such methods are fairly robust to allelic heterogeneity.[68] However, not all approaches to linkage analysis are very powerful, and this is especially true for non-parametric approaches involving small families[69, 70], although transmission/disequilibrium tests may have merit in the analysis of rare variants.[71] In addition, linkage analysis approaches not only come with the often difficult and expensive need to sample family members, but many phenotypes may not exhibit familial aggregation, undermining the motivation to consider family-based studies[10].

Multiple regression and data mining methods

Regression models treat the phenotype as a dependent variable and collapsed sets of variants as independent or predictor variables. Such methods provide a flexible framework for

assessing the contribution of collections of rare variants to a phenotype.[28, 33] Such models can accommodate a number of additional predictor variables, including common variants, covariates such as gender and age, and interaction terms. Recently, Morris and Zeggini[33] assessed the power of simple regression methods for testing collapsed sets of rare variants for association with a quantitative trait and found that such approaches are indeed intuitive, flexible and powerful. The authors compared the use of a simple tally of the number of rare variants possed by an individual across a large region as a predictor of a phenotype against the use of a simple indicator of the possession of any rare variant. They found that the use of a tally may be more powerful.[33] However, they did not consider conditioning effects (Figure 1C) or problems associated with analyses involving many correlated predictor variables.[33]

Multiple regression models have been applied in many standard GWA studies in an effort to identify the most likely causal variants in a particular genomic region harboring many associated variants[72, 73]. However, their direct application via simple extensions of the methods described by Morris and Zeggini[33] to the analysis of multiple individual rare variants or collapsed sets of variants may be problematic. For example, collapsed sets of variants might be correlated due to LD with an additional common variant included in the model or due to the manner in which different subsets of variants are collapsed based on functional annotations, as discussed previously in the context of the hierarchical nature of collapsing sets of variants based on functional annotations. Furthermore, strong multicollinearity is known to cause numerical and interpretation issues in traditional linear regression analysis. In addition, there will likely be many potential predictor variables to choose from if many individual common and rare variants, as well as collapsed sets of variants, are considered. Having many independent variables, or more independent variables than subjects, creates enormous potential for numerical instabilities and overfitting in standard linear regression models.

Newer regression techniques that make use of regularization and shrinkage parameters to control for collinearity and overfitting can be used to overcome these problems. Two such techniques, ridge regression[74] and the LASSO[75,76] have been considered in genetic association analysis contexts, and other methods have also been proposed as well.[77–80, 81] Tibshirani[82] compared the relative merits of standard stepwise regression, ridge regression, and the LASSO in different non-genetic contexts and concluded that each method seems is best suited for different specific settings, depending on the number and effect sizes of the predictors. This is problematic in the context of genetic association analyses since one will not necessarily know *a priori* how many common, rare, or collapsed sets of variants might influence a phenotype, nor what kind of effects those variants have. One possible solution to this problem is to devise methods that combine elements of many different regression procedures, such as the ‘bridge (GPS)’ regression procedure of Friedman[83] that exploits constructs forming the basis for both ridge and LASSO-based regression. Alternatively, ‘ensemble’ methods or ‘super learners’ that combine the results of different regression and prediction methods[84] could be used. However, it is not clear that such methods will pick out functional or causal variants in an association study involving a large number of variants or collapsed sets of variants over those that may, due to LD, merely act as strong predictors of the phenotype.

Logic regression[67] may be a particularly attractive regression-based approach, at least in theory, for the analysis of rare variants. Logic regression, which is similar in ways to the method proposed by Han and Pan,[40] was initially proposed for analyzing sequence data and does not assume that variants have been collapsed *a priori*. Instead, it constructs, and then tests for association, combinations of variants held together through the creation of dummy independent variables. These variables are constructed from logical operators such as ‘AND’ and ‘OR’ that connect and combine sets of variants into potential predictors of the

phenotype. There are many issues with logic regression and related approaches that are similar to the issues discussed previously in the context of selecting an optimal subset of rare variants[40, 66]. These include: computational burden; difficulty in obtaining p-values for each potential independent variable (or individual rare variant, as opposed to a collapsed group of rare variants); and the identification of the optimal, and hence the biologically most-plausible, set of genetic predictors. The development of regression analysis methods for rare variant association analyses is an important area of research, however, as the flexibility, conditioning strategies, and ability to accommodate many effects make them particularly appealing.

Power Studies

Most studies assessing associations between rare variants and a phenotype have relied on rather simple collapsing strategies (Table 1). The advantages of more sophisticated data analysis methods are therefore unclear from a practical and implementation standpoint. However, power studies comparing newer methods with more simplistic methods for rare variant analysis have been pursued (Table 3). The studies we list in Table 3 are in no way exhaustive, but their consideration can provide insight into the limitations of the different strategies and, therefore, motivation for further studies. For example, almost all such studies consider comparisons between a proposed novel method and simple single locus analyses, which is an obvious comparison at some level, but does not reflect the sophistication and utility of the proposed method. In addition, almost all of the studies considered simulations under some version of the EAH model of rare variant effects and do not consider other scenarios (Figure 2) or the influence of LD structure among multiple common and rare variants (of the type that might create ‘synthetic associations’[85]). In addition, studies so far have not considered tests within a hierarchical collapsing framework that leverages functional annotations of genomic regions to separate truly causal variants from collections of rare variants that merely contain causal variants.

Other obvious issues with the current assessments of the power and other properties of rare variant analysis methods concern the simple fact that not enough time has elapsed since their introduction for someone to compare them all in a large study. In addition, some methods are clearly nuanced and are unlikely to work in situations other than those for which they were designed. For example, some methods do not take into account the possible direction of a rare variant effect, such as the methods described by Li and Leal[28] and Madsen and Browning[34] whereas other methods are designed to handle these situations[40]. Finally, although many such published power studies simulate data assuming a population genetics model for the propagation of rare variants, the appropriateness of the assumptions of these models is unclear. We believe that the best approach will be to take real sequence data obtained from many individuals (e.g., the 1000 Genomes Project data) and simulate phenotypes based on variants in those sequences, making assumptions only about phenotypic effect sizes and interactions between variants.

In this light, Bansal et al.[86] recently considered the analysis of sequencing data obtained on two genes, FAAH and MGLL, thought be associated with morbid obesity among 142 morbidly obese and 147 control subjects discussed in a previous study[66]. They applied 11 of the methods described in this review plus 9 high-dimensional regression procedures, and showed that the methods do not consistently agree on the most strongly associated regions of the genes or the most likely causal variants. Their results emphasize the need for simulation and theoretical studies of different methodologies.

Conclusions and future directions

The identification and characterization of the effects of collections of rare variants on common complex disease susceptibility and general phenotypic expression will play prominent roles in future genetic studies. Appropriate data analysis methods for associating rare variants to a phenotype are therefore needed. A number of rare variant association analysis methods have been proposed that build off the notion of collapsing variants into groups based on either functional annotations of the genomic regions they reside in or on their location in a defined genomic region or 'window.' The power and robustness of these models need to be assessed in a wide variety of contexts. In addition, future studies of rare variants will likely be pursued in the context of a broader understanding of the genetic and environmental factors contributing to a particular common complex disease, making it unlikely that an exclusive focus on the influence of rare variants would be appropriate. Furthermore, as DNA sequencing and other genomic technology costs decrease, the frequency and functional impact of different forms of variation beyond SNPs will also be better understood. In this context merely finding that a set of rare variants appears to be collectively associated with a phenotype in no way suggests that all those variants are indeed functional or causally related to the phenotype. Thus, the problem of assigning causality to rare variants in a set may be more pronounced than it is in assigning causality to a single common variant.

A better understanding of the genetic architecture of disease, as well as a better appreciation of the forms and functions of DNA sequence variation, will undoubtedly impact the choice of a statistical method for rare variant association studies. Thus, for example, methods which can accommodate covariates, previously identified genetic factors, allelic heterogeneity, and different sets of collapsed variants simultaneously, such as regression-based methods, are clearly advantageous. However, methods which can account for subtle synergistic effects of many loci within a defined region and/or different forms of variation that might contribute to gene function, such as those rooted in sequence or functional similarity[56, 57, 87, 88] are also likely to be appropriate. It is arguable that, in general, variants or groups of variants should always be studied in a more comprehensive regression model that includes covariates and other confounding variables no matter how the collapsed set was initially identified. Such an approach might mitigate a range of concerns, for example about accommodating confounding variables and the functional assessment of variants.

Acknowledgments

This work was supported in part by the following research grants: U19 AG023122-05; R01 MH078151-03; N01 MH22005; U01 DA024417-01; P50 MH081755-01; R01 AG030474-02; N01 MH022005; R01 HL089655-02; R01 MH080134-03; U54 CA143906-01; UL1 RR025774-03 as well as the Price Foundation and Scripps Genomic Medicine. Ondrej Libiger is also supported by a grant from Charles University: GAUK #134609. The authors would like to thank the reviewers for their comments on previous versions of the review as well as Drs. Eric Topol, Sarah Murray, Sam Levy and the entire team at the STSI for support.

GLOSSARY

Contingency table	a way of representing categorical data in a matrix that is often used to record and analyze the relation between two or more categorical variables. Also referred to as cross-tabulation or a cross-tab table
Regression methods	Statistical methods for predicting or relating a variable (or set of variables) known as the 'dependent' variable to another variable (or set of variables) known as the 'independent' or 'predictor' variable. The resulting relationship defines a 'regression function'

Conditional tests	In regression analysis, the importance of additional variables (or 'covariates') one be included in the model - that is, the model can be 'conditioned' on the additional variables. A 'conditional test' of the relationship between the primary independent variable and the dependent variable can therefore be performed
Covariate effects	The influence of non-primary independent variables on the relationship between a primary independent variable and a dependent variable in a regression analysis setting
Quantiles	Points taken at regular intervals in the cumulative distribution function (CDF) of a random variable that are used to define discrete categories of that variable
Stratified analysis	Data analysis that proceeds by breaking up the units of observation into groups and analyzing those groups independently
Group summary information	Statistics that capture frequencies, counts, and other measures that reflect information at the population or sample level, in contrast to measures reflecting information that is unique to each individual
Moving window	A method for testing genetic associations in which a subregion of a larger region is defined. Variants with the defined region are test for association, the region is shifted to an adjacent region, and the process repeated until all the subregions have been assessed
Type 1 error rates	The probability of a false positive result from a statistical hypothesis test
Permutation methods	Strategies for assessing the probability of observing the value of a particular statistic. The probability is computed from a data set in which the data are randomly shuffled and the statistic is recomputed from the shuffled data many times and ultimately compared to the value of the statistic obtained with the non-shuffled data
Phase information	The determination of the nucleotide content of each of the homologous chromosomes in a diploid individual
Fst/Gst	Two classical measures of population differentiation at the nucleotide level. Essentially, Fst and Gst capture and quantify the allele frequency differences between populations
Logic regression	A regression analysis procedure in which sets of independent variables are groups together using logical operators such as 'AND' and 'OR.' These sets of independent variables, rather than the individual variables themselves, are tested for association with a dependent variable
Non-parametric approaches	Statistical analysis methods that do not rely on specific distributional assumptions (e.g., normality) for the variables being analyzed
Multicollinearity	The situation in which two or more predictors (or subsets of predictors) are strongly (but not perfectly) correlated to one other, making it difficult to interpret the strength of the effect of each predictor (or predictor subset). For example, it would be hard to detect a gene if its effect is 'absorbed' (or masked) by

	combinations of genetic background action/interaction parameters in the model
Overfitting	A phenomenon in which predictions of a dependent variable, based on a set independent variables in a regression setting, are complicated by the fact that there are many more independent variables used in the prediction than there are individuals who have been measured on these independent variables
Regularization/ Shrinkage	A method for combating overfitting in regression models. Most of the independent variables are assumed to make only a small or non-existent contribution to the prediction of a dependent variable. Hence their impact is 'shrunk' or 'regulated' to be close to zero when estimating relevant parameters governing the regression model
Compound heterozygosity	A situation in medical genetics in which two normally recessive alleles of a gene cause disease when they are located on different chromosome homologues in the same individual
Population stratification	The phenomenon of an apparently homogeneous population that is actually composed of subgroups of individuals with distinct ancestral origins and differing allele frequencies at many loci. This leads to bias in the assessment of the significance of associations of a trait with particular loci
Multiple testing	In statistics, multiple testing occurs when one considers a more than one statistical inference from a single data set. Errors in inference are more likely to occur when one considers all the inferences as a whole
Imputation	Based on the known linkage disequilibrium structure in fully genotyped individuals, the genotype of untyped variants can be inferred or imputed in individuals who are genotyped for a smaller number of variants

References

1. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008; 118:1590–605. [PubMed: 18451988]
2. Manolio TA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–53. This paper describes the motivation for considering alternative approaches to discovering genes that influence common complex diseases. It is essentially argues that current GWAS paradigms focusing on common variants have simple failed to identify the majority of genetic variants that influence particular phenotypes. [PubMed: 19812666]
3. Pinto D, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010; 466:368–72. [PubMed: 20531469]
4. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009; 10:241–51. [PubMed: 19293820]
5. Tycko B. Mapping allele-specific DNA methylation: a new tool for maximizing information from GWAS. Am J Hum Genet. 2010; 86:109–12. [PubMed: 20159108]
6. Kong A, et al. Parental origin of sequence variants associated with complex diseases. Nature. 2009; 462:868–74. [PubMed: 20016592]
7. Eichler EE, et al. Completing the map of human genetic variation. Nature. 2007; 447:161–5. [PubMed: 17495918]

8. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet. 2005; 6:287–98. [PubMed: 15803198]
9. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009; 10:392–404. [PubMed: 19434077]
10. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008; 40:695–701. [PubMed: 18509313]
11. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009; 19:212–9. [PubMed: 19481926]
12. Cirulli ET, et al. Common genetic variation and performance on standardized cognitive tests. Eur J Hum Genet. 2010
13. Asimit J, Zeggini E. Rare Variant Association Analysis Methods for Complex Traits. Annual Review of Genetics. 2010; 44:293–308.
14. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet. 2008; 82:100–12. [PubMed: 18179889]
15. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001; 69:124–37. [PubMed: 11404818]
16. Wood LD, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007; 318:1108–13. This study suggests that many different mutations within key genes are likely to drive tumorigenesis, so that although patients might have unique mutations, these mutations are likely to be in genes that harbor mutations across many patients. This rare variant heterogeneity may also contribute to the inherited basis of many common chronic diseases. [PubMed: 17932254]
17. Lahiry P, Torkamani A, Schork NJ, Hegele RA. Kinase mutations in human disease: interpreting genotype-phenotype relationships. Nat Rev Genet. 2010; 11:60–74. [PubMed: 20019687]
18. Bobadilla JL, Macek M Jr, Fine JP, Farrell PM. Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening. Hum Mutat. 2002; 19:575–606. [PubMed: 12007216]
19. Easton DF, et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am J Hum Genet. 2007; 81:873–83. [PubMed: 17924331]
20. Schork NJ, Wessel J, Malo N. DNA sequence-based phenotypic association analysis. Adv Genet. 2008; 60:195–217. [PubMed: 18358322]
21. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010; 11:31–46. [PubMed: 19997069]
22. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324:387–9. [PubMed: 19264985]
23. Ng SB, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42:30–5. [PubMed: 19915526]
24. Roach JC, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010; 328:636–9. [PubMed: 20220176]
25. Schork NJ, Nath SK, Fallin D, Chakravarti A. Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. Am J Hum Genet. 2000; 67:1208–18. [PubMed: 11032785]
26. Lanktree MB, Hegele RA, Schork NJ, Spence JD. Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease. Circ Cardiovasc Genet. 2010; 3:215–21. [PubMed: 20407100]
27. Gilad Y, Pritchard JK, Thornton K. Characterizing natural variation using next-generation sequencing technologies. Trends Genet. 2009; 25:463–71. [PubMed: 19801172]
28. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–21. One of the first papers to comprehensively evaluate statistical methods for testing ‘collapsed’ sets of rare variants to a trait. The paper discussed both distance-based and regression approaches. [PubMed: 18691683]

29. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322:881–8. [PubMed: 18988837]
30. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007; 615:28–56. This paper introduced the notion of 'collapsing' sets of variants into a single group whose collective frequency could be contrasted between groups. [PubMed: 17101154]
31. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010; 141:210–7. [PubMed: 20403315]
32. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–75. [PubMed: 17701901]
33. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34:188–93. [PubMed: 19810025]
34. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]
35. Price AL, et al. Pooled association tests for rare variants in exon-resequencing studies. American Journal of Human Genetics. 2010; 86:832–838. This paper describes a method for explicitly incorporating information about the likely functional effect of specific rare variants into the formulation of an association statistic. The proposed method only considers coding variations however. [PubMed: 20471002]
36. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–6. [PubMed: 19684571]
37. Sebat J, Levy D, McCarthy SE. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. Trends in Genetics. 2009; 25:528–535. [PubMed: 19883952]
38. Xiong M, Zhao J, Boerwinkle E. Generalized T2 test for genome association studies. American Journal of Human Genetics. 2002; 70:1257–1268. [PubMed: 11923914]
39. Lehmann, EL. Nonparametric Statistical Methods Based on Ranks. McGraw-Hill; New York, NY: 1975.
40. Han F, Pan W. A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. Hum Hered. 2010; 70:42–54. [PubMed: 20413981]
41. Hoh J, Ott J. Scan statistics to scan markers for susceptibility genes. Proc Natl Acad Sci U S A. 2000; 97:9615–7. [PubMed: 10931953]
42. Pan W, Han F, Shen X. Test selection with application to detecting disease association with multiple SNPs. Hum Hered. 2010; 69:120–30. [PubMed: 19996609]
43. Fallin D, et al. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res. 2001; 11:143–51. [PubMed: 11156623]
44. Zhao JH, Curtis D, Sham PC. Model-free analysis and permutation tests for allelic associations. Hum Hered. 2000; 50:133–9. [PubMed: 10799972]
45. Zhu X, Fejerman L, Luke A, Adeyemo A, Cooper RS. Haplotypes produced from rare variants in the promoter and coding regions of angiotensinogen contribute to variation in angiotensinogen levels. Human Molecular Genetics. 2005; 14:639–643. [PubMed: 15649942]
46. Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol. 2010; 34:171–87. [PubMed: 19847924]
47. Hartl, DL.; Clark, AG. Principles of population genetics. Sinauer Associates; Sunderland, Mass: 2007.
48. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 2009; 10:639–50. [PubMed: 19687804]
49. Nei, M. Molecular Evolutionary Genetics. Columbia University Press; New York, NY: 1987.
50. Jost L. G(ST) and its relatives do not measure differentiation. Mol Ecol. 2008; 17:4015–26. [PubMed: 19238703]
51. Mount, DW. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press; New York: 2001.

52. Qian D, Thomas DC. Genome scan of complex traits by haplotype sharing correlation. Genet Epidemiol. 2001; 21 (Suppl 1):S582–7. [PubMed: 11793742]
53. Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet. 2003; 72:891–902. [PubMed: 12610778]
54. Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet. 2006; 79:792–806. [PubMed: 17033957]
55. Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol. 2009; 34:213–221. [PubMed: 19697357]
56. Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol. 2004; 27:415–28. [PubMed: 15481099]
57. Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009; 65:822–32. [PubMed: 19210740]
58. Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. Genet Epidemiol. 2009; 33:183–97. [PubMed: 18814307]
59. Ickstadt, K.; Selinski, S.; Muller, TD. SFB 475 Komplexitatsreduktion in Multivariaten Datenstrukturen. Universitat Dortmund; 2005.
60. Templeton AR, et al. Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. Genetics. 2005; 169:441–53. [PubMed: 15371364]
61. Nair RP, et al. Localization of psoriasis-susceptibility locus PSORS1 to a 60-kb interval telomeric to HLA-C. Am J Hum Genet. 2000; 66:1833–44. [PubMed: 10801386]
62. Tachmazidou I, Verzilli CJ, De Iorio M. Genetic association mapping via evolution-based clustering of haplotypes. PLos Genetics. 2007; 3:111.
63. Kowalski J, Pagano M, DeGruttola V. A nonparametric test of gene region heterogeneity associated with phenotype. Journal of the American Statistical Association. 2002; 97:398–408.
64. Gilbert PB, Novitsky VA, Montano MA, Essex M. An efficient test for comparing sequence diversity between two populations. J Comput Biol. 2001; 8:123–39. [PubMed: 11454301]
65. Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. Biometrics. 2006; 62:245–53. [PubMed: 16542252]
66. Bhatia G, et al. A covering method for detecting genetic associations between rare variants and common phenotypes. PLos Genetics. 2010 in press.
67. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. Genet Epidemiol. 2001; 21(Suppl 1):S626–31. One of the first papers to consider statistical methods for identifying optimal sets of predictors of a phenotype from sequence data based purely on the strength of statistical association. The paper proposed a novel regression method for this task. [PubMed: 11793751]
68. Ott, J. Analysis of Human Genetic Linkage. Johns Hopkins University Press; Baltimore: 1991.
69. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996; 58:1347–63. [PubMed: 8651312]
70. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273:1516–7. [PubMed: 8801636]
71. Oexle K. A remark on rare variants. J Hum Genet. 2010; 55:219–26. [PubMed: 20203695]
72. Haiman CA, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet. 2007; 39:638–44. [PubMed: 17401364]
73. Clarke R, et al. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. N Engl J Med. 2009; 361:2518–28. [PubMed: 20032323]
74. Malo N, Libiger O, Schork NJ. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. American Journal of Human Genetics. 2008; 82:375–385. [PubMed: 18252218]
75. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. PLos Genetics. 2008; 4 This paper, along

with the paper by Malo, Libiger, and Schork (2008) introduced regularized regression techniques for accommodating a large number of predictors in a genetic association study as well as to separate causally-associated from non-causally asociated variants.

76. Zhou H, Sehl ME, Sinsheimer JS, Lange K. Association Screening of Common and Rare Genetic Variants by Penalized Regression. Bioinformatics. 2010
77. Clark TG, De Iorio M, Griffiths RC, Farrall M. Finding associations in dense genetic maps: a genetic algorithm approach. Hum Hered. 2005; 60:97–108. [PubMed: 16220001]
78. Guo W, Lin S. Generalized linear modeling with regularization for detecting common disease rare haplotype association. Genet Epidemiol. 2009; 33:308–16. [PubMed: 19025789]
79. Luan YH, Li HZ. Group additive regression models for genomic data analysis. Biostatistics. 2008; 9:100–113. [PubMed: 17513311]
80. Kwee LC, Liu DW, Lin XH, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. American Journal of Human Genetics. 2008; 82:386–397. [PubMed: 18252219]
81. Capanu M, Begg CB. Hierarchical Modeling for Estimating Relative Risks of Rare Genetic Variants: Properties of the Pseudo-Likelihood Method. Biometrics. 2010
82. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological. 1996; 58:267–288.
83. Friedman, JH. Technical Report. Stanford University; 2008. Fast sparse regression and classification.
84. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical Applications in Genetics and Molecular Biology. 2007; 6:1.
85. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. 2010; 8:e1000294. [PubMed: 20126254]
86. Bansal V, Libiger O, Torkamani A, Schork NJ. An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. Pacific Symposium on Biocomputing Proceedings. 2011; 16 to appear.
87. Wessel J, Schork AJ, Tiwari HK, Schork NJ. Powerful designs for genetic association studies that consider twins and sibling pairs with discordant genotypes. Genet Epidemiol. 2007; 31:789–96. [PubMed: 17549743]
88. Nievergelt CM, Libiger O, Schork NJ. Generalized analysis of molecular variance. PLoS Genet. 2007; 3:e51. [PubMed: 17411342]
89. Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC. Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered. 2006; 61:55–64. [PubMed: 16612103]
90. Zschocke J. Dominant versus recessive: molecular mechanisms in metabolic disease. J Inherit Metab Dis. 2008; 31:599–618. [PubMed: 18932014]
91. Andres AM, et al. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. Genet Epidemiol. 2007; 31:659–71. [PubMed: 17922479]
92. Kim JH, Waterman MS, Li LM. Accuracy assessment of diploid consensus sequences. Ieee-Acm Transactions on Computational Biology and Bioinformatics. 2007; 4:88–97.
93. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. [PubMed: 17803354]
94. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–9. [PubMed: 16862161]
95. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–54. [PubMed: 20208533]
96. Li B, Leal SM. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet. 2009; 5:e1000481. [PubMed: 19436704]
97. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009; 10:387–406. [PubMed: 19715440]

98. Wang K, et al. Interpretation of association signals and identification of causal variants from genome-wide association studies. American Journal of Human Genetics. 2010; 86:730–742. [PubMed: 20434130]
99. Efron B. Correlation and large-sclae simultaneous significance testing. Journal of American Statistical Association. 2007; 102:92–103.
100. Sandelin A, Wasserman WW, Lenhard B. ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Research. 2004; 32:W249–W252. [PubMed: 15215389]
101. Matys V, et al. TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes. Nucleic Acids Research. 2006; 34:D108–D110. [PubMed: 16381825]
102. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser - a database of tissue-specific human enhancers. Nucleic Acids Research. 2007; 35:D88–D92. [PubMed: 17130149]
103. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008; 36:D154–8. [PubMed: 17991681]
104. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005; 120:15–20. [PubMed: 15652477]
105. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004; 11:377–94. [PubMed: 15285897]
106. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. 2003; 31:3568–71. [PubMed: 12824367]
107. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 2002; 297:1007–13. [PubMed: 12114529]
108. Sironi M, et al. Silencer elements as possible inhibitors of pseudoexon splicing. Nucleic Acids Res. 2004; 32:1783–91. [PubMed: 15034146]
109. Wang Z, et al. Systematic identification and analysis of exonic splicing silencers. Cell. 2004; 119:831–45. [PubMed: 15607979]
110. Goren A, et al. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. Mol Cell. 2006; 22:769–81. [PubMed: 16793546]
111. Zhang L, et al. Functional allelic heterogeneity and pleiotropy of a repeat polymorphism in tyrosine hydroxylase: prediction of catecholamines and response to stress in twins. Physiol Genomics. 2004; 19:277–91. [PubMed: 15367723]
112. Zhang C, Li WH, Krainer AR, Zhang MQ. RNA landscape of evolution for optimal exon and intron discrimination. Proc Natl Acad Sci U S A. 2008; 105:5797–802. [PubMed: 18391195]
113. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
114. Kuhn RM, et al. The UCSC Genome Browser Database: update 2009. Nucleic Acids Research. 2009; 37:D755–D761. [PubMed: 18996895]
115. Matthews L, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009; 316:D16–D22.
116. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010; 38:D35–D60.
117. Ashburner M, et al. Gene ontology: tool for the unification of biology. Nature Genetics. 2000; 25:25–29. [PubMed: 10802651]
118. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–504. [PubMed: 14597658]
119. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet. 2002; 31:19–20. [PubMed: 11984561]
120. Dennis G Jr, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003; 4:P3. [PubMed: 12734009]

121. Suderman M, Hallett M. Tools for visually exploring biological networks. Bioinformatics. 2007; 23:2651–9. [PubMed: 17720984]
122. Karchin R. Next generation tools for the annotation of human SNPs. Briefings in Bioinformatics. 2009; 10:35–52. [PubMed: 19181721]
123. Plumpton, M.; Barnes, MR. Bioinformatics for Geneticists. Barnes, M., editor. John Wiley and Sons; New York: 2007. An excellent review of the methods available for computationally assessing the functional impact of DNA sequence variants. The review also provides lists of available tools
124. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annual Review of Genomics and Human Genetics. 2006; 7:61–80.
125. Andersen MC, et al. In silico detection of sequence variations modifying transcriptional regulation. PLoS Comput Biol. 2008; 4:e5. [PubMed: 18208319]
126. Everitt, BS. Cluster Analysis. John Wiley and Sons; New York, NY: 2009.
127. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008; 319:473–6. [PubMed: 18218900]
128. Libiger O, Nievergelt CM, Schork NJ. Comparison of genetic distance measures using human SNP genotype data. Hum Biol. 2009; 81:389–406. [PubMed: 20067366]
129. Hill MO. Diversity and Evenness - Unifying Notation and Its Consequences. Ecology. 1973; 54:427–432.
130. Keylock CJ. Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. Oikos. 2005; 109:203–207.
131. Lande R. Statistics and partitioning of species diversity, and similarity among multiple communities. Oikos. 1996; 76:5–13.
132. Jost L, et al. Partitioning diversity for conservation analyses. Diversity and Distributions. 2010; 16:65–76.
133. Johansen CT, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet. 2010; 42:684–7. [PubMed: 20657596]
134. Romeo S, et al. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest. 2009; 119:70–9. [PubMed: 19075393]
135. Slatter TL, Jones GT, Williams MJ, van Rij AM, McCormick SP. Novel rare mutations and promoter haplotypes in ABCA1 contribute to low-HDL-C levels. Clin Genet. 2008; 73:179–84. [PubMed: 18199144]
136. Marini NJ, et al. The prevalence of folate-remedial MTHFR enzyme variants in humans. Proc Natl Acad Sci U S A. 2008; 105:8055–60. [PubMed: 18523009]
137. Ji W, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet. 2008; 40:592–9. [PubMed: 18391953]
138. Frikke-Schmidt R, Sing CF, Nordestgaard BG, Steffensen R, Tybjaerg-Hansen A. Subsets of SNPs define rare genotype classes that predict ischemic heart disease. Hum Genet. 2007; 120:865–77. [PubMed: 17006673]
139. Azzopardi D, et al. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res. 2008; 68:358–63. [PubMed: 18199528]
140. Masson E, Chen JM, Scotet V, Le Marechal C, Ferec C. Association of rare chymotrypsinogen C (CTRC) gene variations in patients with idiopathic chronic pancreatitis. Hum Genet. 2008; 123:83–91. [PubMed: 18172691]
141. Ma X, et al. Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease. PLoS One. 2007; 2:e1318. [PubMed: 18091991]
142. Ahituv N, et al. Medical sequencing at the extremes of human body mass. Am J Hum Genet. 2007; 80:779–91. [PubMed: 17357083]
143. Wang J, et al. Resequencing genomic DNA of patients with severe hypertriglyceridemia (MIM 144650). Arterioscler Thromb Vasc Biol. 2007; 27:2450–5. [PubMed: 17717288]
144. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006; 354:1264–72. [PubMed: 16554528]

145. Kotowski IK, et al. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. Am J Hum Genet. 2006; 78:410–22. [PubMed: 16465619]

146. Cohen JC, et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A. 2006; 103:1810–5. [PubMed: 16449388]

147. Cohen J, et al. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet. 2005; 37:161–5. [PubMed: 15654334]

148. Cohen JC, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305:869–72. One of the first papers to explicitly consider the association and impact of a collection of rare variants on a complex phenotype. [PubMed: 15297675]

149. Fearnhead NS, et al. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci U S A. 2004; 101:15992–7. [PubMed: 15520370]

150. Calvo SE, et al. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet. 2010

Biographies

Vikas Bansal, Ph.D. Vikas Bansal received his Ph.D. in Computer Science from the University of California, San Diego, USA. He is currently a Research Scientist at the Scripps Translational Science Institute in La Jolla, California. His current research interests include developing computational methods for the detection of human genetic variation using high-throughput sequencing technologies, reconstructing diploid human genomes, and statistical methods for enabling sequencing-based disease association studies.

Ondrej Libiger, Ph.Dc. Ondrej Libiger holds a Master of Science degree in Computer Science from Charles University in the Czech Republic, and is currently finishing his doctoral degree in biomedical sciences there. Ondrej worked in laboratory of Dr. Nicholas Schork as a research programmer at the University of California, San Diego between 2003 and 2007. Since 2007, he has been a research programmer at The Scripps Translational Institute and The Scripps Research Institute. His interests include applying multivariate statistics and data mining techniques to data generated by various types of genomic technology with the aim of improving health care.

Ali Torkamani, Ph.D. Ali Torkamani received his Ph.D. training in the laboratory of Dr. Nicholas Schork in the School of Medicine at the University of California at San Diego. He then joined the Scripps Translational Sciences Institute as a Research Scientist before being appointed as an Assistant Professor of Molecular Medicine at the The Scripps Translational Science Institute and The Scripps Research Institute. His research interests are in developing computational and analytical methods for understanding the functional impact of inherited and somatically-acquired DNA sequence variation from multiple perspectives, including sequence and structure based analyses as well as systems biology approaches.

Nicholas J. Schork, Ph.D. Nicholas J. Schork is currently Professor, Molecular and Experimental Medicine, The Scripps Research Institute (TSRI) and Director of Biostatistics and Bioinformatics at the Scripps Translational Science Institute (STSI). Prior to joining TSRI and the STSI, Dr. Schork held faculty positions at the University of California, San Diego and Case Western Reserve University. His professional interests are in statistical genetics and integrated biomedical research. He received an M.A. in Philosophy, an M.A. in Statistics, and a Ph.D. in Epidemiology under Drs. Michael Boehnke and Patricia Peyser from the University of Michigan in Ann Arbor.

‘AT A GLANCE’ BULLET POINTS

- We review the motivation for exploring the role of rare variants in phenotypic expression
- There are several problems with capturing the effects of rare variants in association studies using current statistical analysis methods
- We discuss the concept and use of ‘collapsing’ sets of rare variants into predictors of phenotypic expression, to aid statistical analyses of rare variant associations.
- Functional annotations of specific variants and genomic regions can be used to define collapsed sets of rare variants
- A range of statistical analysis models and inference-making procedures could be exploited to assess the association between rare variants and phenotypic expression. We discuss the relative merits of these approaches.
- We compare ‘Moving window’ and ‘defined region’ approaches to the analysis of rare variant effects
- We discuss the importance for rare variant analysis of the flexibility of statistical analysis models and methods in accommodating factors including common variants, interactions between variants, beneficial and deleterious effects of variants and environmental factors.

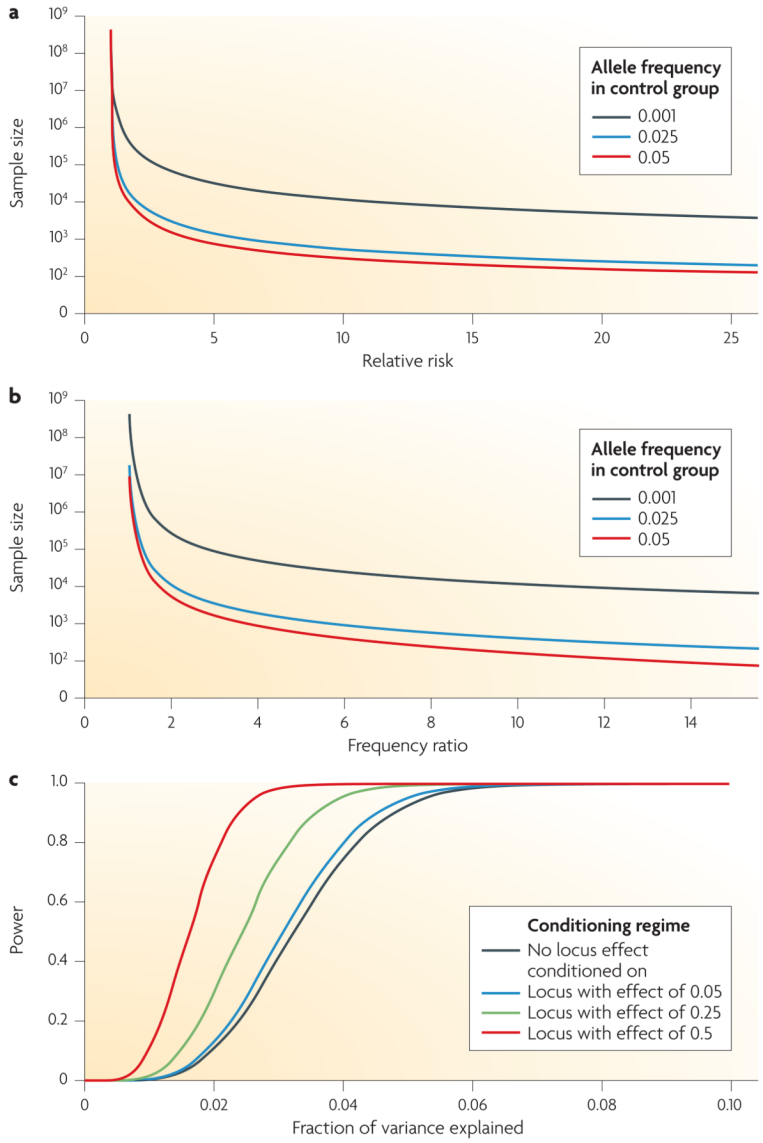

Figure 1. Sample size requirements and statistical power for variants of different frequencies

(**A**). Sample sizes necessary to detect an association between an allele with a specific effect size and a binary trait. The plots assume a standard z-test for the difference in the frequency of the allele between the two phenotypic categories. A genome-wide type I error rate of 10^{-9} was assumed, under the assumption that one may perform 2 orders of magnitude more tests in a complete sequence-based GWAS than a standard GWAS. (**B**). Similar setting to that provided in Figure 1A except the effect size depicted on the x axis gives the ratio of the frequency of the allele in the case vs. control groups. These curves give insight into the power gains associated with the collapsing strategy. Consider the black line in Figure 1B and testing a single rare variant with a frequency of 0.01 in the controls and 0.02 in the cases. This difference would require approximately 250,000 cases and controls to detect with 80% power at a super genome-wide level of significance. However, if one were to test 5 such variants with the same frequencies after collapsing them (assuming they are independent and no individual has more than one such variant), then one would effectively be testing a 0.05 frequency among the controls and a 0.10 frequency among the cases. From

the red line in Figure 1B this difference would require only 3000 cases and controls. (**C**). Power to detect a quantitative trait locus with a sample of 1000 individuals as a function of fraction of phenotypic variation explained by the locus via standard linear regression analysis. A genome-wide type I error rate of 10^{-9} was assumed.

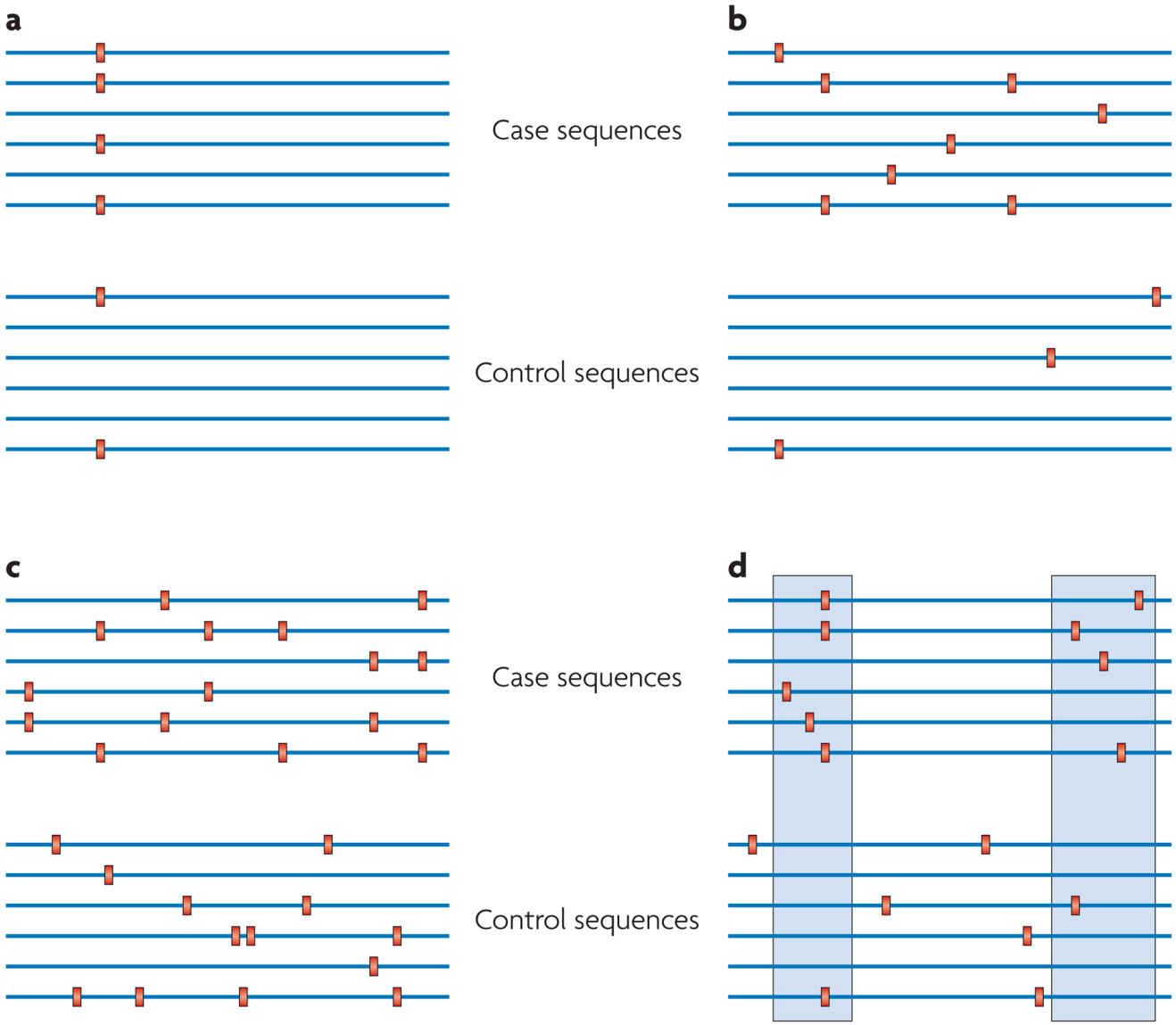

Figure 2. Scenarios in which DNA sequence variants distinguish cases and controls

Blue lines indicate genomic regions; red boxes indicate variants. **A**. Variants at a single locus with common alleles are more frequent in cases then controls. **B**. Multiple rare variations contribute to the phenotype such that the collective frequency of these variations is greater in cases. This would create a greater diversity of haplotypes or DNA sequences among the cases. **C**. Multiple rare variations contribute to the phenotype, but act in a synergistic fashion such that cases are likely to have more similar DNA sequences compared to controls. **D**. Multiple rare variations contribute to a phenotype, but the variations contributing to the phenotype reside in specific genomic regions. This situation would create greater sequence diversity among the cases, as in setting B, but only within the genomic regions of relevance.

Table 1

Recent Studies Pursuing Rare Variant Association Analyses

Phenotype	Method	Sample	Genes	Variants	Associated	Comments	Ref
HTG levels	CAST	438/327	4	187	154	Associated variants across 4 genes	133
Type 1 Diabetes	CS/FET	480/480	10	212	4	Four rare variants in one gene	22
Plasma HDL levels	FET	3551	4	93	NP	Rare NS cSNPs more frequent in low TG subjects	134
Plasma HDL levels	Observe	154/102	1	NP	3	5 carriers of rare variants with low HDL	135
Folate response	FET	564	1	14	5	Functional evaluation of NS mutations	136
Blood pressure	FET	3125	3	138	30	Rare mutations affect blood pressure	137
Plasma HDL levels	FET	95/95	1	51	3	Variants in ABCA1 influence HDL-C	138
Colorectal cancer	FET	691/969	1	61	NP	Rare NS variants in patients	139
Pancreatitis	CS	216/350	1	20	18	Rare variants common in patients	140
Tuberculosis	FET	1312	5	179	NP	Rare NS variants in tuberculosis cases	141
BMI	CS	379/378	58	1074	NP	Rare NS variants in obese vs. lean	142
HTG levels	CS	110/472	3	NP	10	Single common variant combined with rare variants = HTG	143
Heart Disease	CS	3363	1	2	2	Rare variants associated with lower plasma LDL	144
Plasma LDL levels	FET	3543	4	17	1	PCSK9 variants associated with low LDL	145
Plasma LDL levels	NP	512	1	26	NP	Variants in NPC1L1 associated with low cholesterol	146
Plasma LDL levels	NP	128	1	2	2	2 missense mutations associated with low LDL	147
Plasma AGT levels	FET	29/28	1	93	11	Rare haplotypes associated with high AGT levels	45
Plasma HDL levels	FET	519	3	NP	NP	Used collapsing of rare variants	148
Colorectal Adenoma	NP	124/483	4	NP	NP	25% vs. 12% rare variants in cases vs controls	149
Complex I	Observe	Pooled	103	898	151	More likely deleterious variants in Complex I Deficiency	150

Key: ABCA1: ATP-binding cassette transporter 1; HTG: Hypertryglycerides; HDL: High density lipoproteins; LDL: Low density lipoproteins; BMI: Body mass index; AGT: angiotensinogen; CAST: Cohort allelic sums test[30]; CS: Contingency table Chi-square test; FET: Fisher's Exact Test; NP: Not Provided in the text in an obvious way; NS, non-synonymous; cSNP, SNPs that occur in cDNAs; TG, triglycerides; PCSK9: Proprotein convertase subtilisin/kexin type 9; NPC1L1: Niemann-Pick C1 Like 1; Genes = number of genes/genomic regions sequenced; Variants = total number of variants found; Associated = number of variants associated with the phenotype.

Table 2

Statistical Analysis Approaches that Accommodate Rare Variants

Approach	Category	Description	QTL?	Cov?	Comp?	Reference
Simple CAST *	Sum	Collapse variants and test for overall frequency differences	Strat	Strat	Triv	28, 30
Differentiation	Sum	Assess the overall genetic distance between groups over multiple loci	Strat	Strat	Triv	50
Nucleotide Divers.	Sum	Compare nucleotide diversity in a genomic region between groups	Strat	Strat	Triv	47
Combine SL tests	Sum	Combine test statistics at each locus via, e.g., Fisher's p-value method	Yes	Strat	Triv	42
T-Square distance *	Sum	Compute the distance between allele frequency profiles	Strat	Strat	Mod	28
Freq. Weighting *	Sum	Compute individual carrier status scores weighted by allele frequency	Strat	Strat	Triv	34
Variable Weight *	Sum	Find optimal weights of variants and leverage functional impact	Yes	Strat	Mod	35
Haplotype Freq. *	Sum	Omnibus test of haplotype frequency differences between groups	Strat	Strat	Mod	43, 44
Seq. Diversity	Dis	Compare individual sequence differences across groups	Strat	Strat	Triv	65
MDMR	Dis	Directly relate a sequence dissimilarity matrix to phenotypic variation	Yes	Direct	Intens	20, 54
Sim. regression	Dis	Non-matrix-based regression of phenotype on sequence similarity	Yes	Direct	Mod	56, 57
IBD sharing *	Dis	Evaluate identity-by-descent sharing within families	Yes	Strat	Mod	69, 70
Subset Selection	Dis	Identify minimal set of variants that maximally discriminate groups	Strat	Strat	Intens	66
Linear regression *	Reg	Regress phenotype on collapsed sets of variants	Yes	Direct	Triv	33
Adaptive Sums *	Reg	Identify optimal subset of variants as predictors considering effect sign	Yes	Direct	Intens	40
Logic regression *	Reg	Optimize collapsed set of predictors in regression framework	Yes	Direct	Intens	67
Ridge regression	Reg	L2-regularized regression to accommodate variant correlations	Yes	Direct	Mod	74
Lasso *	Reg	L1-regularized regression to accommodate large number of variants	Yes	Direct	Mod	75
Lasso/Ridge *	Reg	Grouped parameter L1 and L2 regularized regression	Yes	Direct	Mod	76

Key: CAST=Cohort Allelic Sums Test; Comp?=Computational burden (Triv=Trivial; Mod=Moderate; Intens=Intensive); Cov?=the ability of the statistics directly accommodate covariates in their formulation ('Direct') or can they be accommodate only through stratified analyses ('Strat'); Dis=Dissimilarity in individual sequences-based test; Freq.=Frequency; Divers.=Diversity; IBD=Identity-By-Descent; L1=linear penalty; L2=quadratic penalty; MDMR=Multivariate distance matrix regression; QTL?=the ability of statistics to deal with quantitative phenotypes either directly ('Yes') or only by stratifying the phenotype into categories that can be compared ('Strat'); Reg=Regression model-based test; Seq.=Sequence; Sim.=Similarity; SL=Single locus; Sum=Summary statistic based test; T-Square=Hotelling's T-square statistic for comparing;

*Denotes a method explicitly proposed within the context of a genetic association study.

Table 3

Power Studies Comparing Statistical Methods that Explicitly Consider Rare Variants in Association Analysis Settings

Primary	Methods	Sample	Region Size	Variants	Quantitative	Population?	Comments	Ref
VW	MB	10000	9kb	-	Yes	Yes	VW>MB; only simulated missense mutations	35
HC	SL, Link	1000	149	8	No	Yes	HC>Link>SL; family and 2 stage designs	46
LReg	SL	5000	50kb	-	Yes	Yes	LReg>SL; # variants>presence/absence	33
Asum	CMC, MB	500	9–18	9	No	No	Asum>CMC and MB for directional effects	40
MB	SL, CAST,	1000	50	50	No	Yes	MB>LL>CAST>SL	34
CMC	SL, Hotel	1000	10–40	5–20	No	No	CMC>Hotel>SL; analytical power studies	28
Las	SL	1000	WG	3000	No	Yes	Las>SL; Las gives fewer FP; common and rare	75

Key: Region Size and Variants reflect the size of the genomic region and the number of variants considered in the power comparison. Population?: was a specific population genetic model assumed in the studies? VW: Variable Weighting[35]; MB: Madsen and Browning[34]; HC: Haplotype collapsing[45]; SL: Single Locus; Link: Linkage analysis; LReg: Linear regression assuming the number of rare variants or the presence of absence of rare variants as predictors[33]; Asum: Adaptive sums test[40]; CAST: Cohort allelic sums test[30]. Hotel: Hotelling's T-square[38]; Las: Lasso[75]; FP: False positives. '>' indicates the relative power of two methods. Note that for the 'LReg vs SL' comparison (row 3) Morris and Zeggini found that using the number of variants in a region ('# variants') as a predictor was more powerful than simply using the presence or absence of a variant in a region as a predictor ('presence or absence') in certain regression contexts.

Table B.1

Integrative Web-Servers for Variant Annotation

Server Name	URL	Types of variant annotated
FASTSNP	http://fastsnp.ibms.sinica.edu.tw/	Precalculated SNPs
F-SNP	http://compbio.cs.queensu.ca/F-SNP/	Precalculated SNPs
Human Splicing Finder	http://www.umd.be/HSF/	Any Sequence / Splicing Only
MutDB	http://mutdb.org/	Precalculated SNPs
PharmGKB	http://www.pharmgkb.org/index.jsp	Pharmacogenetic SNPs
PolyDoms	http://polydoms.cchmc.org/polydoms/	Precalculated SNPs
PupaSuite	http://pupasuite.bioinfo.cipf.es/	Precalculated SNPs
SeattleSeq	http://gvs.gs.washington.edu/SeattleSeqAnnotation/	Any Sequence
Sequence Variant Analyzer	http://www.svaproject.org/	Any Sequence
SNP@Domain	http://bioportal.net/.	Precalculated SNPs
SNPeffect	http://snpeffect.vib.be/	Precalculated SNPs
SNP Functional Portal	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx	Precalculated SNPs
Trait-o-matic	http://snp.med.harvard.edu/	SNPs Associated with Traits