Computational and statistical approaches to analyzing variants identified by exome sequencing

Stitziel *et al*. *Genome Biology* 2011, **12**:227
http://genomebiology.com/2011/12/9/227

Genome Biology

REVIEW

Computational and statistical approaches to analyzing variants identified by exome sequencing

Nathan O Stitziel[1,2†], Adam Kiezun[2,3†] and Shamil Sunyaev[2,3*]

Abstract

New sequencing technology has enabled the identification of thousands of single nucleotide polymorphisms in the exome, and many computational and statistical approaches to identify disease-association signals have emerged.

From quantitative trait locus mapping and linkage analysis to genome-wide association studies (GWASs), genetic markers have been used to locate causal genes underlying Mendelian and complex traits with impressive success: the molecular basis for nearly 3,000 Mendelian disorders is known [1] and over 4,500 single nucleotide polymorphisms (SNPs) have been associated with a variety of human traits and complex diseases [2]. These studies rely on linkage with the disease-causing variant and, by their very nature, indirect genetic marker studies have limitations. The causal variant or gene remains unknown for the majority of the 4,500 SNPs associated with complex disease and for over 3,500 Mendelian disorders. New sequencing-based studies have emerged and are poised to change genetic mapping fundamentally by enabling the direct identification of causal sequence variants in a single experiment. We will no longer have to rely on linkage with the disease-causing variant; instead, by obtaining full sequence data for all genes we can now directly test for association with disease. As we have learned in the past few years, however, there is a great deal of human genetic variation [3] and finding the causal variant among thousands of candidates can be difficult.

Here we review the computational and statistical approaches that have emerged for managing these data in this rapidly exploding field. First, we briefly review the process for identifying variants in next-generation sequencing (NGS) studies and then discuss strategies for identifying the causal variant in Mendelian disorders among the total number of variants identified. We also discuss strategies for identifying the causal gene(s) in complex diseases among all genes in the genome, before outlining some challenges facing current exome sequencing studies.

Variant discovery in exome sequencing projects

NGS methods have been developed that harness massively parallel DNA sequencing [4] and enable large-scale sequencing projects that have applications ranging from cataloging genetic diversity on a population level [3] to identifying a disease-causing variant in a single individual, which might lead to directed therapy [5]. Most large-scale medical sequencing projects so far have focused on the protein-coding region of the genome (the 'exome'). This has been driven in part by cost (whole genome sequencing is still relatively expensive for large sample sizes), biology (most known examples of disease-causing variants alter the protein sequence), and practical considerations (there is currently little consensus on interpreting non-coding genetic variation).

Various methods have been developed to select a subset of the genome for sequencing, but only solid-phase hybridization [6] and liquid-phase hybridization [7] have been commercially applied for selecting the entire human exome as the target for sequencing. After target enrichment, sequencing is performed using various NGS technologies, including reversible terminator reactions, sequencing by ligation, pyrosequencing and real-time sequencing [8]. These generate millions of short sequence copies, or reads, tiled across the portions of the reference genome that were targeted. Although numerous algorithms have been developed to align NGS reads to the reference genome (Bowtie, Short Oligonucleotide Analysis Package (SOAP) and Blat-like Fast Accurate Search Tool (BFAST), among others [9]), most sequencing projects use Mapping and Assembly with Qualities (MAQ) [10] or the Burroughs-Wheeler Aligner (BWA) [11] because of computational efficiency and multi-platform compatibility. The resulting

*Correspondence: ssunyaev@rics.bwh.harvard.edu
[2]Program in Medical and Population Genetics, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA
Full list of author information is available at the end of the article
†Contributed equally

aligned sequence is then inspected for positions that vary from the human reference sequence and are identified as SNPs.

As with alignment tools, many algorithms have been developed to identify a high-quality set of variants in NGS projects. Most current SNP discovery tools rely on the calculation of genotype likelihoods at each position [10], defined as the probability of observing the given sequencing data (base calls and base quality scores) at that position given a set of underlying genotypes. Bayesian posterior probabilities can then be calculated for each potential genotype [12]. Two popular tools for SNP discovery in NGS data that are easily incorporated into data-processing pipelines are SAMtools [13] and the Genome Analysis Toolkit UnifiedGenotyper [14,15]. Other tools have been developed to exploit aspects of specific types of NGS technologies (optimizing base quality estimates from pyrosequences, for example) [16-18] or low-coverage sequencing data [18,19].

By applying the appropriate tool one can identify a set of positions in the sequencing data that are different from the reference sequence along with an indication of genotype quality. Typically 15,000 to 20,000 variants are discovered per exome, with the variation in this number occurring from different exome target definitions [20-23] (a target set with fewer genes or exons would be expected to have fewer total variants) and ancestry (individuals of African ancestry have more variants per exome than individuals of European ancestry [3], for example). By contrast, about 3 million SNPs per genome are discovered using whole-genome sequencing [24] because of the larger sequencing target (whole genome sequencing targets about 3 Gb, whereas the typical exome target is about 33 Mb). To facilitate the processing and sharing of these large datasets, the Variant Call Format (VCF) text file format [3] is emerging as the accepted format for reporting sequence variation from NGS projects, and the SAM/BAM file format is routinely being used for storing and sharing raw NGS data [13].

Challenges for variant discovery in exome sequencing projects

Because even a single base-pair change can be associated with disease, SNP discovery algorithms must robustly distinguish true variation from sequencing errors. This challenge is magnified in exome sequencing projects, in which discovering rare variants is often the goal. NGS has an inherently higher per-base error rate than Sanger sequencing [25] but is generally thought to compensate for these errors with much higher coverage (most NGS experiments for disease-association generate an average of greater than 20- to 30-fold coverage). Despite this degree of coverage, however, the higher error rate of NGS can introduce false-positive associations if cases and controls have differential coverage depths [26]. In large-scale sequencing projects aimed at discovering rare variants associated with complex disease, differential coverage between cases and controls should be one of the quality control metrics (of potentially many); however, a standardized quality control approach to NGS data has not yet emerged.

Applying exome sequencing to Mendelian disorders

Exome sequencing has been successfully used to find the causal variant in several Mendelian disorders, such as Miller syndrome [27] (a rare autosomal recessive disorder characterized by craniofacial abnormalities), Kabuki syndrome [28] (an autosomal dominant form of mental retardation with facial abnormalities), and many others [29]. It is emerging as an attractive method for disease-gene mapping in Mendelian traits when linkage studies have been inconclusive or impossible [23] (often owing to low numbers of affected individuals) or when looking for causal *de novo* mutations [20,28]. Successful studies have typically analyzed fewer than ten individuals and often only affected individuals have been sequenced. These small studies are underpowered for detecting association using currently available association tests and use a different analytic approach for novel gene discovery compared with methods developed for the analysis of complex diseases.

Identifying causal variants: filtering

Various heuristic filtering methods have been used to narrow the search for the causal variant from about 20,000 to often a single variant, or to a single gene (with several independent variants; Figure 1). In general these heuristic filters rely on four main assumptions: (1) the causal variant will alter the protein coding sequence; (2) it will be extremely rare (often assumed to be shared only by cases in one family); (3) every carrier of a putative disease-causing variant will have the phenotype (complete penetrance); and (4) every individual with the disorder will carry the putative disease-causing variant (that is, complete detectance, or 100% probability of observing a genotype given the phenotype). Functional annotation can divide variants into synonymous variants (those that do not change the amino acid sequence), missense variants (those that introduce an amino acid change), and loss-of-function variants (those that prematurely truncate proteins and those disrupting protein splicing). Approximately 50 to 75% of variants can be removed from consideration by focusing only on nonsynonymous (protein-altering) changes [30,31]. Some studies further divide variants into different classes on the basis of the predicted effects of the protein alterations

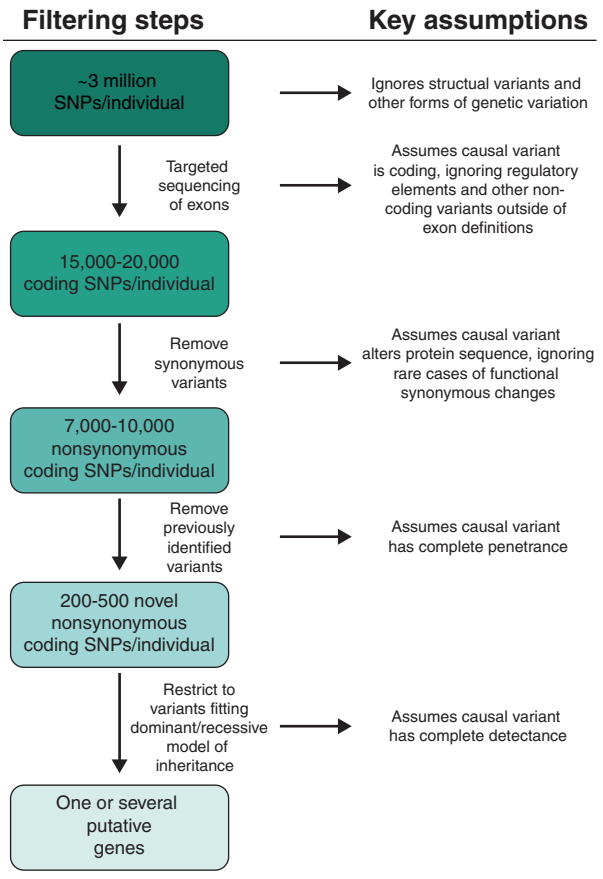


Figure 1. Typical heuristic filtering applied to exome sequencing projects aimed at novel gene discovery for Mendelian disorders, along with key assumptions at each step. Each individual carries approximately 3 million SNPs. Sequential filters shown here can be applied to reduce the number of potential disease-associated variants.

(most commonly using PolyPhen [32], SIFT [33], GERP [34] or PhyloP [35]). Under the assumption that variants responsible for Mendelian disorders will not be present in publicly available databases of human genetic variation, investigators have removed variants for further consideration if they are found in HapMap [36], 1000 Genomes Project [3], dbSNP [37], and privately available variants from other exome sequencing projects (typically shared controls or cases for other phenotypes sequenced locally). Restricting the search to nonsynonymous variants not present in available databases currently reduces the list of putative causal variants to approximately 200 to 500 [23,27,38].

Finding causal variants under a recessive model

To further narrow the search, investigators have imposed a recessive model of disease when the pedigree suggests this mode of inheritance, requiring a putative causal variant to be present in a homozygous state for all individuals (while absent in public databases), or for individuals to be compound heterozygotes in the putative gene (carrying two separate variants in the same gene), which can reduce the list to a single variant or gene [20,22,23]. This has been successfully performed in at least 11 studies of recessive disorders with various numbers of individuals down to as few as one, in which a single individual with Perrault syndrome (ovarian dysgenesis with sensorineural deafness) was found to have two separate non-synonymous variants in *HSD17B4*, a gene that is involved in peroxisomal fatty acid β-oxidation.

These simple filtering techniques may not be sufficient, however, and additional approaches might be needed to further narrow the search. An example of this was the use of an identity by descent analysis in a sequencing study to discover the cause of hyperphosphatemia mental retardation syndrome [39]. After common variants were excluded from the list of shared variants among three affected individuals, 14 candidate genes were left; of these, however, only two were found in regions of the exome that were inferred to be identical by descent. *PIGV* (encoding phosphatidylinositol glycan class V), a gene that is involved in the synthesis of glycosyl-phosphatidylinositol, was identified as the causal gene after the final two candidate genes were sequenced in additional families. Our guess is that after the 'low-hanging fruit' are found, additional novel methods incorporating techniques from population and statistical genetics will be needed to identify causal genes in sequencing projects in which the answer is not immediately apparent.

Finding causal variants under a dominant model

In contrast to the autosomal recessive model of disease, there have been fewer published examples of novel gene association with autosomal dominant disorders (only four have yet been published [29]), perhaps highlighting the relative difficulty in finding such causal genes with exome sequencing. The general approach in the dominant model also relies on filtering a list of nonsynonymous variants to exclude those previously identified in either public databases or shared control exomes, and it requires affected individuals to be heterozygous for the same variant [31] or to be heterozygous for different variants in the same gene [28]. As a proof of principle for exome sequencing in gene discovery for Mendelian disorders, the exomes of four individuals with Freeman-Sheldon syndrome (a rare autosomal dominant disorder previously known to arise from mutations in myosin heavy chain 3, *MYH3*) were sequenced in one of the first publications detailing exome sequencing of multiple individuals [22]. *MYH3* was

identified as the only gene containing non-synonymous variants in all four individuals while being absent from dbSNP and other control exomes.

Challenges for exome sequencing for Mendelian disorders

All exome sequencing studies for gene discovery in Mendelian disorders have relied on the assumption of complete penetrance. Under this assumption, they exclude variants from consideration if present in public catalogs of human genetic variation or unpublished datasets. As these databases expand, however, disease-causing variants might appear in one or more publicly available datasets. The limitation of requiring absence from these datasets is also apparent when one allows for a genetic model of incomplete penetrance (that is, if the phenotype is present in only some fraction of carriers). In the future such a filtering strategy might need to specify a minor allele frequency threshold in such datasets as opposed to requiring complete absence. The converse of penetrance (the probability of observing a phenotype given a genotype) is detectance (the probability of observing a genotype given a phenotype), and almost all exome sequencing studies for Mendelian disorders have relied on a model of complete detectance. The causal gene for Kabuki syndrome**,** however, was found only after allowing for incomplete detectance [28], and might not have been identified as *MLL2* (mixed lineage leukemia 2) if the discovery panel had not been so enriched for carriers (90% of the discovery panel carried a loss-of-function variant in *MLL2* compared with 60% of the replication panel). In the future, better tests will be needed that incorporate incomplete penetrance and detectance. However, it is clear that integration of gene length will be critical, as longer genes will dominate the results given the greater numbers of variants due to their size.

Applying exome sequencing to complex disease

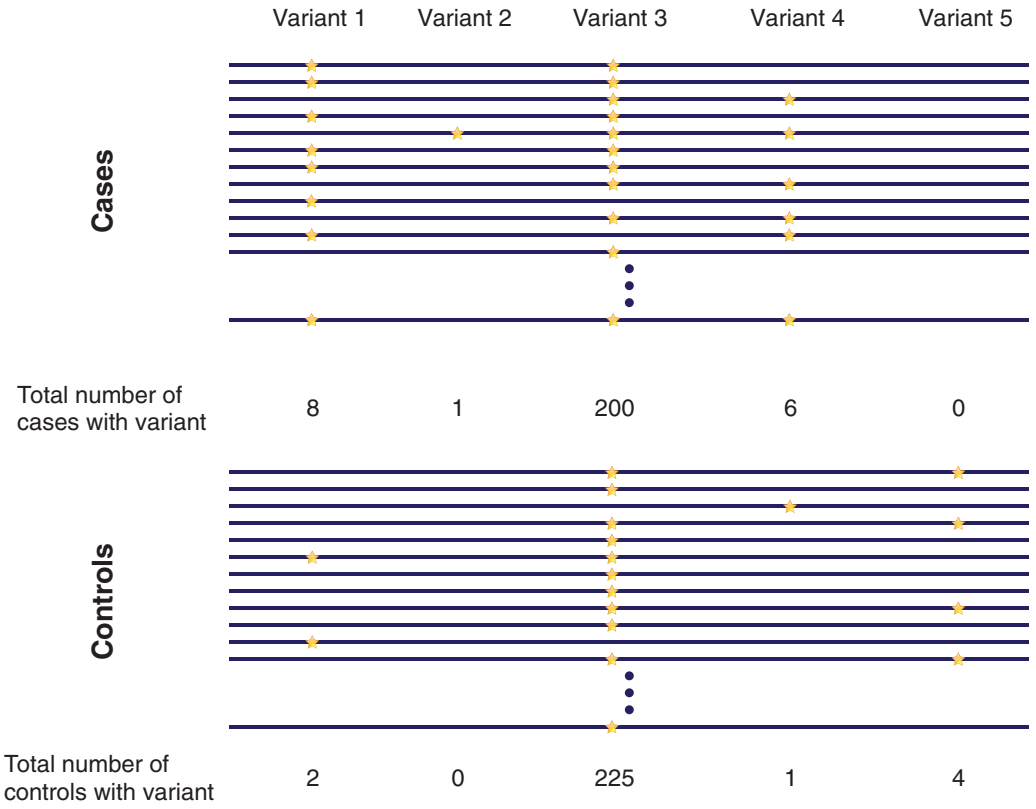
GWASs have been performed for many complex traits and have identified associations with thousands of common variants (minor allele frequency typically over 5%), each conferring a modest increase in risk among carriers (with odds ratios rarely above 1.3 [40]). These 'risk alleles' are typically not causal and are associated with the phenotype of interest because of linkage with the causal variant. Exome sequencing studies fundamentally differ from GWASs because, in theory, they enable unbiased variant discovery and allow for direct association between phenotype and causal variant. The driving hypothesis behind complex disease exome-sequencing studies, motivated by the results of early sequencing studies [41-44], is that multiple rare variants in protein-coding genes contribute to the trait of interest. Focusing on rare genetic variation is also supported by studies predicting that numerous functional and deleterious variants segregate in the population at frequencies (0.5 to 5%) too low to be detected by GWASs [45-47]. These rare variants pose an analytical challenge, however, because they are present in so few individuals that there is low power to detect an association. Although we are still awaiting the results of the first exome sequencing studies for complex diseases, we review (below and in Figure 2 and Additional file 1) the available tests for rare variant association, some of which are likely to be applied in ongoing projects (such as the Exome Sequencing Project from the National Heart Lung and Blood Institute [48]).

Single variant tests

The simplest approach to analyzing variants from exome sequencing data is to examine each one individually for association with the given phenotype. For example, dichotomous traits (myocardial infarction, diabetes, schizophrenia, and so on) can be analyzed using the χ^2 test for contingency tables, Fisher's exact test, Cochran-Armitage test for trend, or logistic regression [49]. These methods test for an enrichment of the 'risk' allele in cases or controls (if seen more frequently in controls, it would be deemed a 'protective' allele). An example would be finding a variant present in 3% of cases but only 1% of controls. Whether this overrepresentation is statistically significant depends on the total number of individuals in the study and the required level of statistical stringency. Quantitative traits (such as blood lipid levels, body mass index or height) can be analyzed by linear regression [49]. By definition, rare variants have low population frequency, and the statistical power to detect association with a phenotype is low for modestly sized studies. For example, assuming 10% disease prevalence, in a study with 1,000 cases and 1,000 controls, there is 2% power to detect an association for a rare variant (minor allele frequency of 0.5%), with a threefold effect at the genome-wide significance level of 5×10^{-8}.

Multiple variant tests

Groups of variants can be analyzed together in an attempt to improve power. In whole genome sequencing, a sliding window can be used to group variants, whereas in exome sequencing the natural unit of grouping is one gene. Alternative splicing can complicate this analysis, however, as a single variant might belong to multiple transcripts of the same gene with different functional effects (a variant might be classified as synonymous for one transcript and missense for another, for example). To extend the single variant tests above, single-SNP *P*-values from multiple variants can be combined by Fisher's [50] or Stouffer's [51] methods. Variants can also be combined in multiple logistic or linear regression models. However,



	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5
Case:Control Summary counts	**8:2**	**1:0**	**200:225**	**6:1**	**0:4**
Collapsing methods	Included	Included	Excluded (above MAF threshold)	Included	Included
Aggregation methods	Included	Included	Included (typically down-weighted owing to high frequency)	Included	Included
Bi-directional methods	Included as risk variant	Included as risk variant	Excluded (above MAF threshold)	Included as risk variant	Included as protective variant
Single SNP methods	Test each variant individually for association with disease				

Figure 2. An illustration of rare variant association tests. Cases and controls from a hypothetical complex disease exome sequencing project are depicted. The horizontal bars indicate aligned exome sequences for individuals; stars indicate the presence of a non-reference allele. Variants 1 and 4 represent low-frequency variants with predominance in cases, Variant 2 represents a singleton, Variant 3 represents a common variant, and Variant 5 represents a low-frequency variant exclusive to controls. For simplicity, these variants are displayed with similar frequency, although very rare variants represent the majority of variation in real sequencing studies. As illustrated, the specific genetic architecture underlying the complex phenotype of interest is expected to have a large role in which test is most powerful for detecting an association. Collapsing methods may be best if a burden of rare variants drives the phenotype, whereas aggregation methods may be more powerful if the full allelic spectrum is contributory. Finally, for genes harboring both risk and protective alleles, bidirectional tests may be most appropriate. See Additional file 1 for examples of methods of each type. MAF, minor allele frequency.

because these simple approaches still essentially test each variant separately and then combine evidence from multiple variants, the results must be adjusted for many degrees of freedom, which will limit the power of these approaches.

Given the large amount of human genetic variation, it would not be surprising to find neutral variants in a causal gene. Therefore, selecting a subset of variants for regression can improve the power to detect an association. For example, synonymous variants are typically discarded because they are less likely to be causal. Shrinkage and regularization regression methods such as LASSO [52], ridge regression [53], and stepwise regression have been proposed for association studies. In these methods, the regression model is fitted while accounting for the cost of adding each additional variable to the model. Other approaches, such as logic regression [54] and the method proposed by Han and Pan [55], use data-driven combinations of variants to select variables for regression.

Collapsing methods

Another approach to increasing power is to collapse multiple rare variants together for analysis. The framework of these tests involves collapsing all variants across a unit (each gene being a unit, for example) together so that even if variants are individually rare, they might be jointly present in sufficient frequency to be used in a univariate test. When used for dichotomous traits, collapsing methods test whether the overall burden of rare variants is higher in cases than controls. For example, CAST [56] examines the differences in the number of individuals with one or more rare variants between cases and controls, and the CMC test [57] is based on comparison of non-synonymous rare variants between cases and controls. These tests rely on designating a set of variants as 'rare' for inclusion, and it is not surprising that altering this definition can greatly influence the association results. Unfortunately there is little guidance in this area and allele frequency thresholds of 1% or 5% are commonly (and arbitrarily) chosen. An alternative approach has been developed that uses the data to select the best variants. The variable-threshold test [58] finds the frequency threshold that best discriminates cases from controls. Similarly, RareCover [59] aims to find the optimal set of variants to collapse together. Although there have been no published complex-disease exome sequencing studies, these tests have been applied to candidate gene sequencing results [58,60].

Aggregation methods

An alternative to the collapsing methods involves aggregation, which aims to summarize the information from many variants while appropriately weighing the contribution of each variant. Although collapsing methods discard variants that are considered unlikely to be causal, aggregation methods aim to include the full frequency spectrum of alleles (rare and common) into the association test. The weighted-sum statistic [61] weighs variants according to allele frequency (rare variants are given stronger weighting) because of an assumption that functional variants of large effect are kept at a low population frequency by purifying selection. Weighing variants by apparent effect size is also effective and is implemented in KBAC [62] and the test described by Ionita-Laza *et al.* [63]. These tests have been applied to candidate gene sequencing results [58].

Extensions to these methods

Accounting for covariates

The association of genotype with phenotype can be confounded by various factors such as ancestry, age and sex. Methods that can directly account for such covariates can be advantageous in discerning the causal effect of genetic variants. When a test does not directly accommodate covariates, regressing the genotype and phenotype on the covariate and using the residuals for the association analysis can remove the effect of the covariate on the phenotype.

Accounting for risk and protective alleles together

The effects of genetic variants can be neutral, protective or detrimental for a given disease trait. Many existing methods test for a frequency differential of variants between cases and controls and a mixture of positive and negative effects will adversely affect these tests. For example, *PCSK9* (encoding proprotein convertase subtilisin/kexin type 9), a gene associated with cholesterol levels and coronary artery disease, contains both risk-lowering loss-of-function variants and gain-of-function variants that increase risk [64]. Testing for a difference in the aggregate of these alleles in either cases or controls would not be expected to yield significant results as cases will be enriched for risk variants and controls will be enriched for protective variants, effectively canceling each other out in the sum total. Methods that account for a mixture of directions of effects can be more powerful in such scenarios, and several tests explicitly account for bidirectionality of effects (Additional file 1). The prevalence of genes with variants having bidirectional effects is currently unknown but loss-of-function variants are expected to be more abundant in the general population and this bidirectional effect may be less apparent for sequencing studies not focusing on phenotypic extremes. Regardless, it is likely that multiple genes in a common pathway would have alleles with bidirectional effects, and if a collapsing method is used to

group variants across a pathway, these tests can be increasingly used.

Incorporating functional annotations

Several studies have shown that using functional information improves the power to detect association [58,65-67]. For protein-coding variants this can include the predicted effect on protein function, using programs such as SIFT [33,68], PolyPhen [32,69], Panther [70,71], MutationAssessor [72], SNAP [73] and PupaSuite [74]. For non-coding variants, evolutionary conservation and functional effects can be assessed using programs such as PhyloP [75], PhastCons [76], SCONE [77] and SiPhy [78].

Statistical power

The statistical power of the methods to test for association with rare variants has not been systematically analyzed. Although articles that describe novel association tests usually provide power comparisons to previous methods, these calculations are prone to being performed under specific assumptions about the genetic architecture of the trait that often favors the test being implemented and might not be representative for human traits in general [79,80]. Extending the results from theoretical studies [81] and early sequencing studies of candidate genes [41,42,82] would suggest that approximately 10,000 exomes are needed to achieve genome-wide significance for complex traits (in which a Bonferroni-corrected *P*-value for 20,000 genes would require $P < 2.5 \times 10^{-6}$). Even the most powerful of the methods available for analyzing sequencing data will not lower these requirements substantially. It would not be surprising, then, that the first exome sequencing association studies will be underpowered and exome sequencing will need to be replicated with additional sequencing or genotyping (or both) [83].

Which test(s) should be used?

The decision regarding the use of specific tests will depend on many factors, including study design (if the trait is quantitative or dichotomous), the assumption of the underlying genetics (whether only rare variants or both rare and common variants are expected to contribute to disease, whether protective and risk variants are expected), and pragmatic considerations (which test is available for use). Most importantly, different tests are powered to detect associations for different aspects of genetic architectures (number of affected loci, associated population frequencies, or associated effect sizes and directions) [79,84,85]. Currently, no software suite contains more than a small number of tests and input formats vary between available software packages, which complicates applying multiple tests to the same study. In the future we expect multiple tests to be implemented in available software suites.

Challenges for exome sequencing applied to complex disease

Numerous tests have been developed for analyzing sequencing data (Additional file 1). Running a large battery of these tests comes at the cost, however, of having to penalize multiple hypothesis testing, as well as potential confusion over inconsistent results (a gene can be highly ranked in one test and not significant in another, for instance). Regardless of the test, unless rare variants have a surprisingly large phenotypic effect on complex diseases, achieving sufficient statistical power will require large studies. DNA sequencing costs will continue to decrease, however, and adequately sized studies might soon be performed (simulations suggest that 5,000 cases and 5,000 controls would provide adequate power to detect association for rare variants with modest effect [81]). Combining results from different studies on the same phenotype is an attractive intermediate option (as has been seen with increasingly larger GWAS meta-analyses). This will probably prove more challenging than GWAS meta-analysis, however, as differences in results from multiple sequencing centers (perhaps with different sequencing technologies or different exome target definitions, for example) can introduce significant technical artifacts. Once putative variants have been discovered, the replication strategy for exome studies will depend on the genetic architecture discovered in the analysis. Disease-associated low-frequency polymorphisms can be verified with follow-up genotyping. If the phenotype is caused by a collection of singleton variants, however, further sequencing in additional individuals will be needed and might prove expensive (especially if multiple genes are being considered or if genes are large or have many exons).

Prospects for the future

The growing number of exome sequencing studies demonstrates the power of this approach in mapping genes involved in Mendelian phenotypes. The success of this approach is uncertain, however, as publication bias makes it unclear how many studies fail to identify a causal locus by exome sequencing. Non-allelic heterogeneity, regulatory variation and structural variation underlying phenotypes all pose challenges for sequencing-based discovery of Mendelian genes. It is possible that new statistical and computational methods will increase the already impressive success rate of exome sequencing studies for Mendelian disorders.

Although we are still awaiting the completion of the first exome sequencing studies focusing on complex phenotypes, the early studies will probably be underpowered because current sequencing costs prohibit the adequately sized samples discussed above (10,000 samples). Owing to this lack of power, the first studies

may not result in the discovery of numerous novel loci involved in traits of medical relevance. We believe that the enthusiasm for sequencing studies should not be diminished, however, because this technology has already shown great promise in the field of Mendelian disorders and sequencing costs will continue to decline, leading to adequately powered studies for complex traits. Technology already allows for the complete characterization of genetic diversity. The success of complex trait genetic research will now be determined by our ability to interpret the data and assemble sufficiently large well-phenotyped clinical populations.

Additional materials

Additional file 1: A table of available statistical methods for analyzing variants discovered in sequencing studies.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported in part by National Institutes of Health grants R01-MH084676 and R01-GM078598. NS was supported in part by National Institutes of Health Training Grant T32-HL07604-25, Brigham and Women's Hospital, Division of Cardiovascular Medicine.

Author details
[1]Division of Cardiovascular Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. [2]Program in Medical and Population Genetics, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA. [3]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Published: 14 September 2011

References

1. Online Mendelian Inheritance in Man [http://www.ncbi.nlm.nih.gov/omim/]
2. A Catalog of Published Genome-wide Association Studies [http://www.genome.gov/gwastudies/]
3. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.
4. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135-1145.
5. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL, Basehore MJ, Broeckel U, Tomita-Mitchell A, Arca MJ, Casper JT, Margolis DA, Bick DP, Hessner MJ, Routes JM, Verbsky JW, Jacob HJ, Dimmock DP: **Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease.** *Genet Med* 2011, **13:**255-262.
6. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: **Microarray-based genomic selection for high-throughput resequencing.** *Nat Methods* 2007, **4:**907-909.
7. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.** *Nat Biotechnol* 2009, **27:**182-189.
8. Metzker ML: **Sequencing technologies – the next generation.** *Nat Rev Genet* 2010, **11:**31-46.
9. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing.** *Brief Bioinform* 2010, **11:**473-483.
10. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18:**1851-1858.
11. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26:**589-595.
12. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: **A general approach to single-nucleotide polymorphism discovery.** *Nat Genet* 1999, **23:**452-456.
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
14. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
15. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-498.
16. Quinlan AR, Stewart DA, Stromberg MP, Marth GT: **Pyrobayes: an improved base caller for SNP discovery in pyrosequences.** *Nat Methods* 2008, **5:**179-181.
17. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K: **SNP detection for massively parallel whole-genome resequencing.** *Genome Res* 2009, **19:**1124-1132.
18. Malhis N, Jones SJ: **High quality SNP calling using Illumina data at shallow coverage.** *Bioinformatics* 2010, **26:**1029-1035.
19. Le SQ, Durbin R: **SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples.** *Genome Res* 2011, **21:**952-960.
20. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, Veltman JA: **De novo mutations of SETBP1 cause Schinzel-Giedion syndrome.** *Nat Genet* 2010, **42:**483-485.
21. Bolze A, Byun M, McDonald D, Morgan NV, Abhyankar A, Premkumar L, Puel A, Bacon CM, Rieux-Laucat F, Pang K, Britland A, Abel L, Cant A, Maher ER, Riedl SJ, Hambleton S, Casanova JL: **Whole-exome-sequencing-based discovery of human FADD deficiency.** *Am J Hum Genet* 2010, **87:**873-881.
22. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461:**272-276.
23. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, Garimella KV, Fisher S, Abreu J, Barry AJ, Fennell T, Banks E, Ambrogio L, Cibulskis K, Kernytsky A, Gonzalez E, Rudzicz N, Engert JC, DePristo MA, Daly MJ, Cohen JC, Hobbs HH, Altshuler D, Schonfeld G, Gabriel SB, Yue P, Kathiresan S: **Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia.** *N Engl J Med* 2010, **363:**2220-2227.
24. Koboldt DC, Ding L, Mardis ER, Wilson RK: **Challenges of sequencing human genomes.** *Brief Bioinform* 2010, **11:**484-498.
25. Chan EY: **Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery.** *Methods Mol Biol* 2009, **578:**95-111.
26. Garner C: **Confounded by sequencing depth in association studies of rare alleles.** *Genet Epidemiol* 2011. doi: 10.1002/gepi.20574.
27. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet* 2010, **42:**30-35.
28. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J: **Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome.** *Nat Genet* 2010, **42:**790-793.
29. Ng SB, Nickerson DA, Bamshad MJ, Shendure J: **Massively parallel sequencing and rare disease.** *Hum Mol Genet* 2010, **19:**R119-124.
30. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG, Roepman R, Knoers NV, Veltman JA, Brunner HG: **Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome.** *Am J Hum Genet* 2010, **87:**418-423.
31. Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, Jiang H, Zhang P, Shen L, Guo JF, Li N, Li YR, Lei LF, Zhou J, Du J, Zhou YF, Pan Q, Wang J, Wang J, Li RQ, Tang BS: **TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing.** *Brain* 2010, **133:**3510-3518.
32. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7:**248-249.

33. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31**:3812-3814.
34. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA: **Single-nucleotide evolutionary constraint scores highlight disease-causing mutations.** *Nat Methods* 2010, **7**:250-251.
35. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence.** *Genome Res* 2005, **15**:901-913.
36. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, *et al.*: **Integrating common and rare genetic variation in diverse human populations.** *Nature* 2010, **467**:52-58.
37. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**:308-311.
38. Pierce SB, Walsh T, Chisholm KM, Lee MK, Thornton AM, Fiumara A, Opitz JM, Levy-Lahad E, Klevit RE, King MC: **Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome.** *Am J Hum Genet* 2010, **87**:282-288.
39. Krawitz PM, Schweiger MR, Rödelsperger C, Marcelis C, Kölsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Köhler S, Jäger M, Grünhagen J, de Condor BJ, Doelken S, Brunner HG, Meinecke P, Passarge E, Thompson MD, Cole DE, Horn D, Roscioli T, Mundlos S, Robinson PN: **Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome.** *Nat Genet* 2010, **42**:827-829.
40. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106**:9362-9367.
41. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science* 2004, **305**:869-872.
42. Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA: **Medical sequencing at the extremes of human body mass.** *Am J Hum Genet* 2007, **80**:779-791.
43. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP: **Rare independent mutations in renal salt handling genes contribute to blood pressure variation.** *Nat Genet* 2008, **40**:592-599.
44. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC: **Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans.** *J Clin Invest* 2009, **119**:70-79.
45. Kryukov GV, Pennacchio LA, Sunyaev SR: **Most rare missense alleles are deleterious in humans: implications for complex disease and association studies.** *Am J Hum Genet* 2007, **80**:727-739.
46. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD: **Assessing the evolutionary impact of amino acid mutations in the human genome.** *PLoS Genet* 2008, **4**:e1000083.
47. Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome.** *Genetics* 2001, **158**:1227-1234.
48. **NHLBI Exome Sequencing Project** [http://www.nhlbi.nih.gov/resources/exome.htm]
49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559-575.
50. Fisher RA: *Statistical Methods for Research Workers.* Edinburgh, London: Oliver and Boyd; 1925.
51. Stouffer SA: *The American Soldier.* Princeton: Princeton University Press; 1949.
52. Wu TT, Chen YF, Hastie T, Sobel E, Lange K: **Genome-wide association analysis by lasso penalized logistic regression.** *Bioinformatics* 2009, **25**:714-721.
53. Malo N, Libiger O, Schork NJ: **Accommodating linkage disequilibrium in genetic-association analyses via ridge regression.** *Am J Hum Genet* 2008, **82**:375-385.
54. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L: **Sequence analysis using logic regression.** *Genet Epidemiol* 2001, **21 Suppl 1**:S626-S631.
55. Han F, Pan W: **A data-adaptive sum test for disease association with multiple common or rare variants.** *Hum Hered* 2010, **70**:42-54.
56. 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.
57. 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**:311-321.
58. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR: **Pooled association tests for rare variants in exon-resequencing studies.** *Am J Hum Genet* 2010, **86**:832-838.
59. Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V: **A covering method for detecting genetic associations between rare variants and common phenotypes.** *PLoS Comput Biol* 2010, **6**:e1000954.
60. Luo L, Boerwinkle E, Xiong M: **Association studies for next-generation sequencing.** *Genome Res* 2011, **21**:1099-1108.
61. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5**:e1000384.
62. Liu DJ, Leal SM: **A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions.** *PLoS Genet* 2010, **6**:e1001156.
63. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C: **A new testing strategy to identify rare variants with either risk or protective effect on disease.** *PLoS Genet* 2011, **7**:e1001289.
64. Kotowski IK, Pertsemlidis A, Luke A, Cooper RS, Vega GL, Cohen JC, Hobbs HH: **A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol.** *Am J Hum Genet* 2006, **78**:410-422.
65. Sul JH, Han B, Eskin E: **Increasing power of groupwise association test with likelihood ratio test.** In *Proceedings of the Fifteenth Annual Conference on Research in Computational Biology (RECOMB-2011)*. Heidelberg: Springer; 2011 [http://portal.acm.org/citation.cfm?id=1987628]
66. Sul JH, Han B, He D, Eskin E: **An optimal weighted aggregated association test for identification of rare variants involved in common diseases.** *Genetics* 2011, **188**:181-188.
67. Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallée MP, Voegele C, Webb PM, Whiteman DC, Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G: **Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer.** *Am J Hum Genet* 2009, **85**:427-446.
68. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Res* 2001, **11**:863-874.
69. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P: **Prediction of deleterious human alleles.** *Hum Mol Genet* 2001, **10**:591-597.
70. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Res* 2003, **13**:2129-2141.
71. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B: **Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools.** *Nucleic Acids Res* 2006, **34**:W645-W650.
72. Reva B, Antipin Y, Sander C: **Determinants of protein function revealed by combinatorial entropy optimization.** *Genome Biol* 2007, **8**:R232.
73. Bromberg Y, Rost B: **SNAP: predict effect of non-synonymous polymorphisms on function.** *Nucleic Acids Res* 2007, **35**:3823-3835.
74. Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J: **PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes.** *Nucleic Acids Res* 2006, **34**:W621-625.
75. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies.** *Genome Res* 2010, **20**:110-121.
76. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15**:1034-1050.
77. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S: **Analysis of sequence conservation at nucleotide resolution.** *PLoS Comput Biol* 2007, **3**:e254.

78. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X: **Identifying novel constrained elements by exploiting biased substitution patterns.** *Bioinformatics* 2009, **25:**i54-i62.
79. Bansal V, Libiger O, Torkamani A, Schork NJ: **Statistical analysis strategies for association studies involving rare variants.** *Nat Rev Genet* 2010, **11:**773-785.
80. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ: **Testing for an unusual distribution of rare variants.** *PLoS Genet* 2011, **7:**e1001322.
81. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: **Power of deep, all-exon resequencing for discovery of human trait genes.** *Proc Natl Acad Sci U S A* 2009, **106:**3871-3876.
82. 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:**387-389.
83. Liu DJ, Leal SM: **Replication strategies for rare variant complex trait association studies via next-generation sequencing.** *Am J Hum Genet* 2010, **87:**790-801.
84. Asimit J, Zeggini E: **Rare variant association analysis methods for complex traits.** *Annu Rev Genet* 2010, **44:**293-308.
85. 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.** *Pac Symp Biocomput* 2011:76-87.
86. Lawrence R, Day-Williams AG, Elliott KS, Morris AP, Zeggini E: **CCRaVAT and QuTie-enabling analysis of rare variants in large-scale case control and quantitative trait association studies.** *BMC Bioinformatics* 2010, **11:**527.
87. Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S: **Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes.** *Am J Hum Genet* 2010, **87:**604-617.
88. 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:**171-187.
89. Morris AP, Zeggini E: **An evaluation of statistical approaches to rare variant analysis in genetic association studies.** *Genet Epidemiol* 2010, **34:**188-193.
90. Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel EM, Lange K: **Penalized regression for genome-wide association screening of sequence data.** *Pac Symp Biocomput* 2011:106-117.
91. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: **Powerful SNP-set analysis for case-control genome-wide association studies.** *Am J Hum Genet* 2010, **86:**929-942.
92. Hoffmann TJ, Marini NJ, Witte JS: **Comprehensive approach to analyzing rare genetic variants.** *PLoS One* 2010, **5:**e13584.
93. Li Y, Byrnes AE, Li M: **To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests.** *Am J Hum Genet* 2010, **87:**728-735.
94. King CR, Rathouz PJ, Nicolae DL: **An evolutionary framework for association testing in resequencing studies.** *PLoS Genet* 2010, **6:**e1001202.

doi:10.1186/gb-2011-12-9-227
Cite this article as: Stitziel NO, *et al*.: Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biology* 2011, **12:**227.