PERSPECTIVE

nature genetics

Exome sequencing and the genetic basis of complex traits

Adam Kiezun[1,2,16], Kiran Garimella[2,16], Ron Do[2,3,16], Nathan O Stitziel[2,4,16], Benjamin M Neale[2,3,5], Paul J McLaren[1,2], Namrata Gupta[2], Pamela Sklar[6,7], Patrick F Sullivan[8], Jennifer L Moran[2], Christina M Hultman[9], Paul Lichtenstein[9], Patrik Magnusson[9], Thomas Lehner[10], Yin Yao Shugart[11], Alkes L Price[2,12,13,17], Paul I W de Bakker[1,2,14,15,17], Shaun M Purcell[5,17] & Shamil R Sunyaev[1,2,17]

Exome sequencing–based studies are emerging as a popular approach to test for association of rare coding variants with complex phenotypes. The promise of exome sequencing is grounded in theoretical population genetics and in the empirical successes of candidate gene sequencing studies. We discuss here several aspects of exome sequencing studies that we view as particularly important. We analyze exome sequencing data from 438 individuals and use this as a basis to review processing and quality control of raw sequence data, evaluate the statistical properties of exome sequencing studies, discuss rare variant burden tests to detect association to phenotypes and show the importance of accounting for population stratification in the analysis of rare variants. We conclude that enthusiasm for exome sequencing studies to identify the genetic basis of complex traits should be combined with caution stemming from the observation that on the order of over 10,000 samples may be required to reach sufficient statistical power.

Next-generation sequencing[1–5] coupled with efficient DNA capture[6–8] enables the use of exome sequencing as a new approach to study the genetic basis of human phenotypes. A number of genes underlying Mendelian diseases have been mapped using this approach[6,9–15]. Exome sequencing has also been applied to tumors[16–20], where sample purity, read mapping and chromosomal rearrangements are critical and form a very distinctive set of issues. Here, we restrict our focus to complex traits.

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [2]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]The Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]Division of Cardiovascular Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [6]Department of Psychiatry, Friedman Brain Institute, Mount Sinai School of Medicine, New York, New York, USA. [7]Institute for Genomics and Multi-scale Biology, Mount Sinai School of Medicine, New York, New York, USA. [8]Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA. [9]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [10]Division of Neuroscience and Basic Behavioral Science, National Institute of Mental Health, Bethesda, Maryland, USA. [11]Division of Intramural Research Program, National Institute of Mental Health, Bethesda, Maryland, USA. [12]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [13]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [14]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [15]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. [16]These authors contributed equally to this work. [17]These authors jointly directed this work. Correspondence should be addressed to S.R.S. (ssunyaev@rics.bwh.harvard.edu).

Published online 29 May 2012; doi:10.1038/ng.2303

In complex trait genetics, exome sequencing studies can be used to identify rare coding variants that are not detected by microarray-based genome-wide association studies (GWAS). The promise of exome sequencing studies of complex traits has its basis in the success of candidate gene studies[21–26] and has firm roots in population genetic theory[27–35].

Large-scale GWAS of complex traits have consistently shown, with few exceptions, that common variants have modest effects, often requiring over 10,000 samples for their detection. Exome sequencing provides a complementary approach by comprehensively assessing the role of all coding variation, both common and rare. With mutations continually occurring in each protein-coding gene (at a rate of $\sim 1 \times 10^{-5}$ per gene per generation for nonsynonymous variants)[36–39] and fitness losses of less than 1% for most novel nonsynonymous mutations[29–31,34], almost every gene is expected to harbor functionally important variants that can be tested through sequencing, even if these variants are rare. Therefore, the strong interest in exome sequencing stems from three factors: the potential to identify many genes underlying complex traits, straightforward functional annotation of coding variation and a substantially lower cost (approximately five times lower) than that of whole-genome sequencing.

We evaluate the extent of rare coding variation in empirical data, discuss data processing and quality control of raw sequence data and review analytical methods for detecting genotype-phenotype associations, their expected statistical power and the potential for confounding due to population stratification. To illustrate our analyses, we used empirical whole-exome sequence data from 184 individuals from the International HIV Controllers Study (HIV)[40] and 254 control individuals from the Schizophrenia (SCZ) exome sequencing study (**Supplementary Note**).

Assessment of rare coding variation in empirical data

Exome sequencing data contain an abundance of rare coding variation and indicate that a large fraction of this variation is functional. Not only are there many more rare variants than common ones, but sequencing additional samples continues to uncover additional rare variants. In fact, as sample size increases, the number of observed variants grows much faster than is predicted by the neutral model with constant population size[41,42] (**Fig. 1**). This relative excess of rare variants can, in part, be attributed to recent population expansion[43–45] but is also likely to be due to purifying selection. As a consequence, rare variation is enriched for evolutionarily deleterious, and thus functional, variants. Additionally, the proportion of nonsynonymous variants is higher among rare than among common variants[45]. Finally, among rare variants, missense variants predicted[46] to be damaging are more prevalent than variants predicted to be benign (**Fig. 1**). These findings are consistent with studies that showed that rare

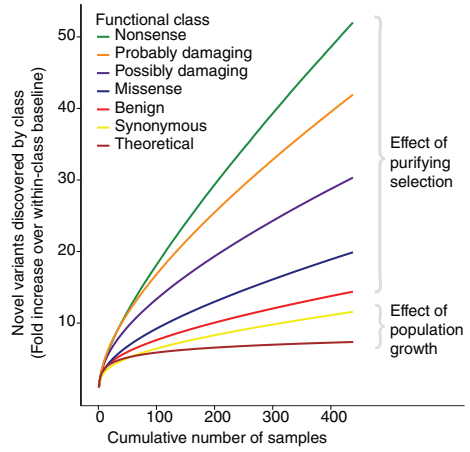

Figure 1 Discovery of novel variants for increasing numbers of samples. For each functional class, the fold increase over the number of variants in one sample for that class is plotted as a function of the number of samples in a sequencing experiment. For example, the number of nonsense variants discovered in 300 samples is 40 times greater than the average number discovered in a single sample, whereas the number of synonymous variants is only 10 times greater (although the absolute number of nonsense variants is a relatively minor proportion of the total variation discovered); this effect is due to purifying selection. All classes of variants are discovered at rates exceeding what would be predicted under a neutral model of evolution in a population of constant size, an effect of population growth. The crossing between curves for synonymous variants and the theoretical prediction most likely is a signature of the out-of-Africa bottleneck. Additional details are provided in Methods.

variants in protein-coding regions are under purifying selection[35,47–51]. Because sequencing larger samples continuously uncovers functionally relevant variants, exome sequencing studies enable the direct identification of causal variants (in contrast to GWAS that use linkage disequilibrium patterns between common markers).

Variant calling and quality control filtering

An exome sequencing study starts with exome capture and sequencing of DNA samples followed by the identification of sequence variants. Exome capture may be realized on many platforms (for example, Illumina HiSeq, Roche 454 and ABI SOLiD) and through a variety of probe definitions (for example, Agilent SureSelect and Nimblegen SeqCap EZ). Recent advances have made it possible to sequence an entire exome or even several exomes at deep coverage in a single run of the sequencing instrument. However, exome capture technologies differ in what they target, how much they capture and how consistently they do so[8]. Moreover, only 80–90% of the targeted regions are covered by greater than 10×, which may leave 4–8 Mb (or 1,000–2,000 genes) without sufficient coverage for variant detection.

Exome sequencing coverage shows tremendous regional variation[8]. Some regions may be over-covered, representing true structural variation (for example, segmental duplications for which only one copy of the region exists in the reference genome) or technical artifacts (for example, greater abundance of capture probes or overlapping probe definitions resulting in 'double capturing'). Similarly, some areas may be under-covered for biological reasons (for example, segmental duplications where more than one copy exists in the reference sequence, preventing the aligner from placing the read uniquely) or for technical reasons (for example, high GC content or density of variation, which impair hybridization of probes). Furthermore, some near-target regions within 50 bp of the target boundary can have sufficient coverage to warrant inclusion in variant calling. Critically, whichever capture technology is used, either all samples should be processed using the same technology or the variability should be accounted for, for example, by stratifying the study by technology.

We generated whole-exome data (targeting 28 Mb) from 438 samples (Methods) using a two-stage approach[52,53]. First, we applied a data processing and variant calling protocol described previously[54]. Second, we applied post–SNP calling quality control filters.

For quality control of the resulting SNPs, we used population genetic statistics and properties of human genetic variation. Using those statistics helps to identify true variants, because the properties of the mutational process[37,55] are different from errors of the sequencing technology.

We compared statistics computed in the 438-sample data set and in 37 whole-genome data sets released by Complete Genomics, Inc. (CGI; see URLs), focusing only on the same genomic regions as in the exome data. The CGI whole-genome data set is good for comparison, because whole-genome sequencing is not dependent on exome capture technology. We further stratified these per-sample statistics into classes that are biologically interesting (by functional class and CpG status) but may also show different rates of technical artifacts. We show that filtering is critical for achieving high-quality calls (**Table 1**). Before filtering, the metrics showed significant deviation from their expected values, which may indicate a high false positive rate. After filtering, the statistics converged to those in the CGI data set. The effectiveness of the filters is evident also in comparison of the mutation pattern with human-chimpanzee divergence[55].

The number of novel variant sites (defined here as those not present in dbSNP 129) is another metric of SNP call quality (**Supplementary Table 1**). Most novel variants have low frequency and are especially enriched in the singletons and doubletons. Singletons and doubletons are particularly important to distinguish from false positives, because technical artifacts or errors in data processing can easily manifest as novel variation.

Statistics such as the transition/transversion ratio (Ti/Tv) and the number of novel variants are useful as broad indicators of the quality of the data set and enable comparison of two sets of calls from the same data. However, precise expectations of these statistics are unknown, because they depend on many factors, including uneven coverage, variability in DNA quality and other sources of technical bias, such as machine error. Therefore, interpreting small differences from expectation in these statistics is nontrivial. Genotyping validation provides an additional measure of call set quality, independent of the population genetics statistics. Comparing genotyping data to sequencing data makes it possible to directly measure call set quality by calculating non-reference sensitivity (NRS—the rate at which non-reference sites in the genotyping data are recovered in the sequencing data) and non-reference discrepancy rate (NRD—the rate at which genotypes from sequencing and genotyping data differ). A genotyping assay should include sites at various allele frequencies, especially at low frequencies (~1%). When available, data from families, particularly from parent-child trios, can also be useful in assessing call set quality.

Comparing our call set with GWAS data in the same samples at overlapping sites suggested high sensitivity for common variants (98.6% NRS). To assess the quality of low-frequency variant calls in a comparable sequencing data set, we compared CGI data to that generated with the HumanOmni2.5-8 (Omni2.5) BeadChip for the 1000 Genomes Project (**Supplementary Table 2**). This comparison resulted in 95.65% NRS, 1.79% NRD and 1.12% NRD for novel variants.

Despite stringent quality control, genotyping and sequencing errors are still present. Unfortunately, when stratifying variants on the basis of putative functional consequences, the class of variation that is annotated to be most deleterious is also more heavily enriched for errors[56]. This underscores the importance of rigorous quality control.

It is critical that great care is taken to prevent technical biases and confounding in sequencing to avoid distorting association results. For

instance, differences in how (rare and precious) case samples are handled compared to control samples may lead to systematic false positives that masquerade as interesting associations. Likewise, simultaneous multi-sample variant calling on only cases or only controls may lead to differential detection of variants across batches, negatively impacting the accuracy of allele frequency estimates and association analyses. Many other, often poorly understood or hidden, technical confounders (for example, in DNA preparation, exome capture technology, machine type, read length, depth of coverage, SNP calling algorithm or quality control filters) may influence the properties of exome sequencing data. Therefore, although the use of shared controls (for example, from the 1000 Genomes Project) has been helpful in filtering approaches applied to Mendelian disorders[6,9], it is not likely to be applicable to association analysis of complex diseases.

Statistical methods for the analysis of rare variants

Analysis of rare variants requires statistical methods that are fundamentally different from association statistics used for testing common variants. There are two reasons for this. First, rare variants have to be combined in a gene (or pathway) for an association test to reach sufficient power[57]. For example, a causal SNP at a frequency of 1 in 500 and genotype relative risk of 10 in a sample of 200 cases and 200 controls has 0.2% power to be detected at a conventional significance threshold for GWAS ($P < 5 \times 10^{-8}$). Second, functional and population genetics information can be added to the testing approach, because exome sequencing comprehensively captures variation that can be annotated with such information.

Early candidate gene sequencing studies for complex traits were based on comparison of the numbers of nonsynonymous alleles exclusive to cases or controls (or samples at the extremes of the trait distribution)[21,26]. This approach has limited power, because it ignores common and low-frequency polymorphisms, as most such variants would be present both in cases and controls. Recently, a number of statistical tests have been designed for rare variant analysis. The Combined Multivariate and Collapsing (CMC) test[58] jointly assesses the role of rare and common variation. For common variants, traditional regression-based association is applied. For rare variation, an individual's predictor in a regression model is defined as 1 if the individual possesses at least one rare variant in the region (for example, a gene) and 0 otherwise. The weighted-sum statistic (WSS) test[59] creates a composite genotype score for all individuals. This score is the sum of alternate alleles weighted by the inverse of the binomial variance. A rank-sum test is then performed on the genotype scores between phenotypic groups. The kernel-based adaptive cluster (KBAC) test[60] also uses a weighting scheme that reflects apparent effect sizes of individual variants. An alternative approach to combine rare variants into a single test selects an allele frequency threshold on the basis of the observed data. The development of this variable threshold (VT) approach[61] was motivated by population genetics simulations that showed that there is no single optimal weighting scheme or allele frequency threshold. There are numerous other statistical tests for rare variants in complex traits (reviewed in refs. 62–65).

In simulation studies[64], most tests behave similarly in many situations. However, the results may depend on assumptions used in simulated data. The relative power to detect association depends on factors such as the number and proportion of causal variants, their population frequencies and their effect sizes, as well as the directionality of effects, the number of genes contributing to the trait and the fraction of causal genetic variation located in the exome. Statistical tests were developed with various combinations of these factors in mind and therefore are likely to be sensitive to different disease architectures. For example, the simulation framework used in the development of the WSS test assumes effect size proportional to $1/x(1-x)$ (where x is the population frequency of the causal allele), the sequence kernel association test (SKAT)[66] simulation framework uses effect size proportional to $-\log(x)$, and the VT test simulations use a demographic history model with a range of possible values of strength of selection leading to different relationships between effect size and x. These simulations were designed to show the strengths of each method under different effect size distributions: the WSS test is designed for effect sizes proportional to $1/x(1-x)$, SKAT is designed for effect sizes proportional to β density $\beta(x; a_1, a_2)$ for prespecified a_1 and a_2, the C-alpha test[67] is designed for

Table 1 SNP counts per Illumina-sequenced sample

	Filter	Counts (% filtered)	Number of heterozygotes (% filtered)	Number of homozygotes (% filtered)	Ti/Tv
Total					
	Unfiltered	18,626	11,761	3,007	2.92
	Filtered	16,776 (10%)	10,242 (13%)	2,785 (7%)	3.21
	Comparison	16,914	10,464	2,492	3.31
By functional class					
Silent	Unfiltered	9,536	5,933	1,601	4.80
	Filtered	8,845 (7%)	5,372 (9%)	1,514 (5%)	5.10
	Comparison	8,987	5,514	1,352	5.22
Missense	Unfiltered	8,698	5,557	1,350	1.92
	Filtered	7,644 (12%)	4,685 (16%)	1,220 (10%)	2.11
	Comparison	7,723	4,772	1,095	2.17
Nonsense	Unfiltered	70	60	9	1.31
	Filtered	48 (31%)	39 (35%)	8 (11%)	1.65
	Comparison	46	38	6	2.00
By CpG status					
CpG	Unfiltered	2,213	1,539	422	4.82
	Filtered	2,030 (8%)	1,390 (10%)	397 (6%)	5.12
	Comparison	2,098	1,448	350	5.44
Non-CpG	Unfiltered	16,415	10,218	2,585	2.75
	Filtered	14,752 (10%)	8,852 (13%)	2,338 (10%)	3.03
	Comparison	14,822	9,901	2,145	3.11

SNP counts (computed as the median of the metrics value in each sample over 438 samples) were localized to the exome, stratified by functional and biological criteria and compared to SNP counts per the Complete Genomics–sequenced sample (computed as the median of the metrics value in each sample over 37 samples). Before the application of filters, counts significantly differed from the expectations derived from the independently obtained comparison set, indicating the presence of many false positives. For example, the SNPs identified as nonsense mutations initially appeared to be enriched by 1.5-fold for false positives. As nonsense variants are rare, an unfiltered call set may contain many artifactual events that masquerade as nonsense variants. Quality control filters helped to align the metrics with the comparison set and the data from human-chimpanzee divergence. Number of homozygotes includes only homozygotes of the derived allele.

Table 2 Summary of gene burden test results for rare variant studies

Trait	Gene	Test	AC[a] low	AC[a] high	*n*	*P*	Ref.
Triglycerides	*ANGPTL4*	Fisher's exact	13	2	1,775	0.016[b]	26
Triglycerides	*ANGPTL5*	Fisher's exact	9	1	1,775	0.022[b]	
HDL	*ABCA1*	RVE	28	4	519	<0.0001[b]	21
	APOA1		1	0	519		
	LCAT		6	1	519		
Blood pressure	*SLC12A1, SLC12A3, KCNJ1*	Fisher's exact	9	1	626	0.02	22
Obesity	Obesity[c]	Fisher's exact	73	97	757	0.061	25
Type 1 diabetes	*IFIH1*	Fisher's exact	21	39	960	0.025	24
Triglycerides	*APOA5*	Fisher's exact	1	5	765	0.25	23
	GCKR	Fisher's exact	5	20	765	0.024	
	LPL	Fisher's exact	8	44	765	2.47×10^{-5}	
	APOB	Fisher's exact	39	85	765	0.008	

Gene burden test results are summarized from published candidate gene resequencing studies. Only one signal for the *LPL* gene is strongly associated ($P = 2.47 \times 10^{-5}$) but does not attain a genome-wide significance of $P < 2.5 \times 10^{-6}$ ($P < 0.05$ after applying a Bonferroni correction for 20,000 genes tested). This highlights the importance of sequencing large numbers of samples. Counts for refs. 21 and 26 are as reported in the published studies. Counts for ref. 22 are based on functional mutation carriers, as described in the published study. Counts for refs. 24 and 25 are based on SNPs with minor allele frequency (MAF) < 0.01. Counts for ref. 23 are based on SNPs with MAF < 0.01 in controls only. All *P* values are from two-sided tests unless reported otherwise in the published study. RVE, rare variant exclusive test; HDL, high-density lipoprotein.

[a]Allele count for non-reference allele. [b]As reported in the published study. [c]Obesity refers to a combination of 21 candidate genes.

effects in opposite directions in the same region, and the VT test makes no assumptions about effect size distributions.

When combining rare variants, all functional variants may either be assumed to influence the trait in the same direction, or some may be allowed to have opposite directions of effect. A biochemical argument can be made that most nonsynonymous variants are loss-of-function hypomorphs, whereas gain-of-function variants are infrequent. However, some genes (for example, *PCSK9*)[68] have variants of both kinds. Several tests allow for rare variants to have opposite effects on the trait (for example, Step-up[69], C-alpha, the replication-based test[70] and SKAT). These tests are based either on the analysis of over-dispersion or on explicit linear models that determine the contribution of a variant to a score on the basis of the direction of effect observed in the data.

Rare variant tests can benefit from stratifying or weighting rare alleles by functional significance, as evidenced by simulations and sequencing studies of candidate genes[61,64,71–73]. The power of rare variant tests is strongly influenced by the fraction of causal variants among all variants analyzed, and using functional information is an effective way to give greater weight to likely causal variants. For example, nonsense variants should be prioritized above non-conserved missense variants. Similarly, missense variants should be prioritized above synonymous variants. Functional consequences of variants can be predicted by examining the effects of amino-acid changes using comparative sequence and protein structure analyses. Many computational prediction and conservation methods[74,75] are available (reviewed in refs. 76–79). The accuracy of those methods is approximately 80% (ref. 80), and it is likely highest for rare variants. (Truly functional variants are most likely deleterious and are kept at low frequencies by purifying selection, and, so, common variants are most likely neutral and nonfunctional.) Therefore, using prediction methods enriches for functional variants and thereby boosts the power of association tests. Because such predictions are not perfect, however, they should be used quantitatively by weighing variants rather than qualitatively by filtering out variants. A number of tests allow the inclusion of prediction scores in test statistics, including the VT test, KBAC, SKAT, the rare variant weighted aggregate statistic (RWAS)[72] and the likelihood ratio test (LRT)[73]. The PLINK/SEQ suite includes previously computed PolyPhen-2 (ref. 46) prediction scores for all possible missense changes in humans, making these scores readily applicable.

An important consideration for exome sequencing studies is selecting the significance threshold that accounts for multiple testing. A simple way of doing this is to adopt a Bonferroni correction for 20,000 independent tests (1 test for each gene), which, for an experiment-wide significance of 0.05 gives a *P*-value threshold of 2.5×10^{-6} per gene. However, such a threshold may be overly conservative, because it assumes that each tested gene has sufficient variation to achieve the asymptotic properties for the test statistic. For example, if only two individuals carry nonsynonymous variants in a given gene, the difference between cases and controls never exceeds two total observations—thus, the most significant *P* value that can be achieved is approximately 0.25, assuming that these two variants are independent. Therefore, unless the study is large, association *P* values will generally be less significant than expected under the null hypothesis of no association. This deficiency of significant *P* values is apparent in the 438 whole exomes (**Fig. 2a**). The PLINK/SEQ suite computes from data the i-stat, which is an estimate of the minimal achievable *P* value for a gene. The i-stat can be used by setting a threshold (for example, 1×10^{-3}) and only correcting for the number of genes that have an i-stat value below the threshold, in agreement with the idea that, for the genes with an i-stat value above the threshold, there is no power to detect an association. Another way to correct for multiple testing is to compute an experiment-wide significance threshold by permutations of phenotype labels, create the empirical distribution of minimal *P* values for all genes across permutations and compare the minimal *P* value from the real data to that distribution (**Fig. 2b**). This approach efficiently controls type 1 error (the probability of rejecting the null hypothesis when it is true) and is less conservative than the Bonferroni correction. Notably, the *P*-value threshold computed by permutations is dependent on both the study and the statistical test. However, the experiment-wide correction via permutation is not robust to confounding, and it is essential to assess the quality of the distribution of test statistics for those genes that have i-stat values less than the threshold to ensure appropriate calibration of the distribution. Nevertheless, with increasing sample sizes, the dimensionality of the tests will also increase, and studies will be assessing close to 20,000 tests. Therefore, for large studies, we consider the Bonferroni threshold to be preferable.

Statistical power of exome sequencing studies

The power of an exome sequencing study is limited by the amount of variation in a gene. Therefore, power is higher for genes with more variants, for example, larger genes or genes in regions with elevated mutation rate. Additionally, genes in which most variants are causal are easier to identify than those in which few variants are causal. In individual candidate gene sequencing studies, estimates of this proportion ranged from 30–70% (refs. 21,22,26). Consequently, the effect size is not only a property of an individual variant but rather a reflection of the distribution of effects coupled with how those effects are interpreted via the test. Some statistical tests explicitly account for differences in power when evaluating evidence of association[81].

Given the sample sizes, the likely effect sizes and frequencies of causal variants and the proportion of causal variants in a gene, do current exome sequencing studies have sufficient power to detect genes underlying

complex phenotypes? Enthusiasm for exome sequencing studies stems, in part, from successful candidate gene sequencing studies, and, so, we sought to test whether exome sequencing would be expected to have sufficient power to detect genes discovered by the candidate gene approach. To date, no published candidate gene study reported *P* values that would be significant in the context of the complete exome (**Table 2**). This is particularly notable, because some candidate gene studies used much larger sample sizes (thousands of individuals) than ongoing exome sequencing studies (hundreds of individuals). This shows that current exome sequencing studies are underpowered to detect genes with the allelic distribution and effect sizes similar to those in the published examples. Indeed, extrapolation of effect sizes and frequencies from published studies shows that thousands of individuals are required to reach acceptable statistical power (**Fig. 3**). This analysis is consistent with an earlier study based on population genetics simulations that concluded that as many as 10,000 individuals at phenotypic extremes would be needed to achieve satisfactory power[30]. The very first GWAS[82–84] were also highly underpowered, but falling costs and the ability to combine studies in meta-analyses have enabled the rapid creation of well-powered studies and have resulted in many discoveries. Similarly, with the falling cost of sequencing and targeted enrichment[85], exome sequencing will soon be affordable to many research groups, and we expect that consortia will form to facilitate the pooling of exome sequencing data, thus enabling better-powered studies and a new wave of discoveries.

Replication to confirm association

To discover robust associations, replication in exome sequencing studies will be critical. Because small early studies will inevitably be underpowered, no gene may achieve exome-wide statistical significance. In such cases, unless strict correction for multiple tests is performed, researchers should resist the temptation to apply a battery of statistical tests, each with various weighing schemes and variant selection. We strongly argue that an association can only be considered real if it has been replicated. A reasonable replication strategy is to select a few genes (for example, ten) from the discovery stage on the basis of the strength of association[86] and biological plausibility. Sequencing and rare variant associations must then be performed on new samples using a multiple test correction threshold applied only to the (smaller) set of candidate genes.

Population stratification

Population stratification—systematic ancestry differences between cases and controls—is a well-studied confounder in genetic association studies[87]. In GWAS, commonly used approaches to correct for stratification include stratifying by population cluster (structured association), principal-components analysis (PCA) and mixed models[87–90]. Genomic control may also be applied, but it is generally more useful for assessing stratification than correcting for it[87,91].

An important question is whether population stratification can confound exome sequencing studies, and, if so, how should stratification be corrected for in this context? Although excess of rare variant tests are fundamentally different from single-variant tests, the possibility of stratification still exists, because different ancestries within a structured population sample (for example, African and European ancestry in African-Americans or northern and southern European ancestry in European-Americans) may have different allele frequency spectra due to their different demographic histories. For example, in an exome sequencing study in African-Americans in which disease cases have more African ancestry than controls, one expects to see an excess of rare variants in cases, because African chromosomes carry more rare variants[92].

We created a hypothetical case-control exome sequencing study involving real sequencing data and simulated phenotype data using 438 individuals divided into two populations (Methods). To induce population stratification, we assigned case or control status to each sample randomly, with a bias whereby more cases were taken from one population and more controls were taken from the other population. Association tests indicated inflated rates of spurious statistically significant *P* values. We corrected for population stratification by modifying the permutation scheme to account for subpopulations. This correction was effective at controlling type 1 errors in all association tests.

Our simulations show that exome sequencing studies for complex traits can be affected by population stratification, which may produce spurious associations. We have shown that a simple permutation scheme is sufficient to correct for population stratification when discrete clusters corresponding to genome-wide ancestry are known or can be inferred by applying standard methods to GWAS chip data[88,89,93]. The permutation scheme is appealing in that it generalizes most burden of multiple rare variants tests; however, some tests may also be amenable to the use of PCA covariates in instances in which population structure is best described by continuous clines rather than discrete clusters[89].

Discussion

Exome sequencing studies enable the unbiased discovery of coding variations for subsequent association testing for complex traits. However, we expect that initial studies will be underpowered, and we have highlighted a number of technical issues that could bias the interpretation and analysis of rare variant data, especially for novel variants. We expect that over 10,000 exomes will be required to achieve sufficient statistical power to robustly detect associations of rare variation with complex traits. Issues affecting the design of exome sequencing studies that we have discussed here are also relevant to whole-genome sequencing studies, in which the analysis of protein-coding variation will remain the same as in exome sequencing studies.

Focusing exclusively on the exome sequence is a limitation in complex trait genetics, where noncoding genetic variation is believed to

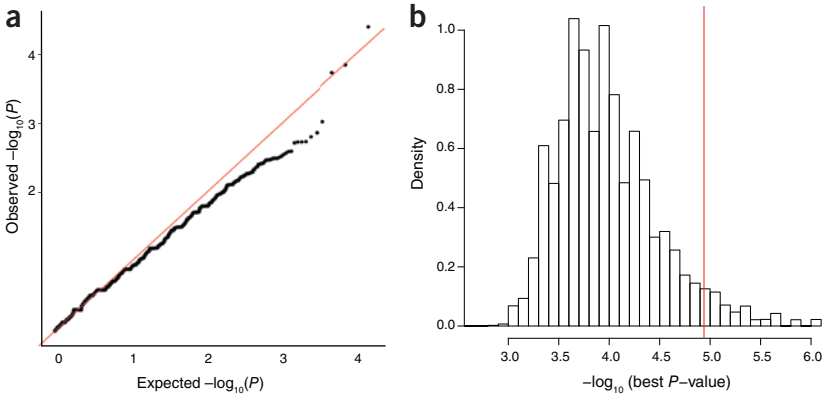

Figure 2 Association analysis. (**a**) Quantile-quantile plot of association *P* values under the null hypothesis. (**b**) Distributions of lowest *P* values under whole-exome permutations. Histograms show the distributions of the lowest *P* values across permutations for the T5 test. The red vertical line indicates the exome-wide significance level of 0.05 for the most significant gene (the most significant gene reaches exome-wide significance if its *P* value is lower that the level indicated by the red line).

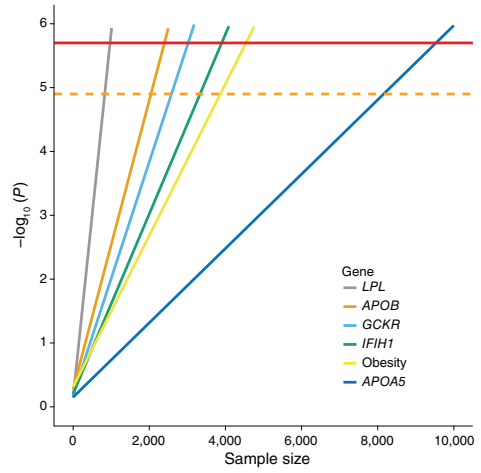

Figure 3 Extrapolation of gene burden results. Horizontal solid red line shows Bonferroni genome-wide significance threshold of $P = 2.5 \times 10^{-6}$. Horizontal dashed line shows the threshold derived from whole-exome permutations (**Fig. 2b**). For larger sample sizes, the permutation threshold would be closer to the Bonferroni threshold, asymptotically approaching it as the sample sizes increase. Obesity refers to a combination of 21 candidate genes.

have a larger role than in Mendelian genetics or in somatic cancer genetics. However, there are several reasons to perform exome sequencing. First, statistical approaches combining multiple rare variants are problematic in noncoding regions, because there is no easily identifiable set of sites harboring variants with unidirectional phenotypic effects. Second, variants in regulatory regions are likely to have smaller effect sizes. In contrast, protein-coding genes provide a well-defined and interpretable target for mutations in the locus. These mutations create variants at this locus that, in a well-powered study, can be identified as being associated with the trait. Thus, although the associated variants identified by focusing on the exome sequence alone are unlikely to explain all of the heritability for a complex trait, this approach does have the potential to highlight the genes involved.

Despite the challenges in the design of exome sequencing studies of complex traits discussed here, the observation that a large number of functionally significant coding variants exist in the human population brings hope that the exome sequencing approach will be useful in identifying loci important for complex traits and diseases.

Methods

Simulating discovery of novel variants

To calculate the discovery rate of novel variants for increasing numbers of samples, first, all exome samples were arranged in a random order. Then, samples were analyzed sequentially, starting with the first sample, and the cumulative set of identified variants was computed. For every subsequent sample, a variant site was considered novel if that site had not been identified as a variant in the cumulative set of the preceding samples. The fold increase over baseline (where the baseline for each class was the number of variants discovered in the first sample) was plotted (**Fig. 1**). To avoid sampling bias, random resampling was performed, and the overall mean was calculated. Nonsense, missense and synonymous classes were defined on the basis of RefSeq annotations. The missense class was further divided into probably damaging, possibly damaging and benign subclasses according to PolyPhen-2 predictions[46]. The theoretical line plotted the expected number of segregating sites under a neutral model of evolution in a population of constant size[41].

Data generation

Reads were aligned to the reference genome using the Burrows-Wheeler Aligner (BWA)[94], duplicate PCR reads were removed using Picard (see URLs), base quality scores were recalibrated using the Genome Analysis Toolkit (GATK), and alignments near putative indels were refined using GATK. The resulting data were run through GATK to discover and genotype SNP candidates.

Quality control filters

We used the following quality control filters: (i) a quality score versus depth filter that excluded variants whose depth-normalized discovery confidence did not exceed 2.0; (ii) a homopolymer run filter that excluded variants that had an alternate allele that matched the allele in an immediately adjacent homopolymer run of greater than five nucleotides; (iii) a strand bias filter that excluded variants whose alternate allele was preferentially found on one of the two available read orientations at the site, and (iv) an indel mask filter that excluded variants discovered at sites that overlap with indels.

Association analysis

Case-control status was assigned randomly, and a T5 test for burden of rare variants was executed on all genes (T5 is a variant of the CMC test[58] that considers only nonsynonymous variants with MAFs less than 5%, uses the total count of alternative minor alleles in cases as the test statistic and assigns significance by permuting phenotype labels). The overall deflation in significant *P* values (fewer genes associated at any significance level than expected by chance) was due to low counts of variants in genes. Results were similar for the T1 version of CMC, as well as for WSS[59] and the VT test[61]. This pattern is expected in studies with small sample sizes (with fewer than approximately 1,000 individuals). Whole-exome permutations can be used to establish exome-wide significance in such cases.

Whole-exome permutations

Phenotype labels of full exomes were permuted 1 million times, that is, permuted phenotype affected all genes in an individual. In each permutation, the lowest exome-wide *P* value was computed. It took fewer than 1,000 computing hours to run 8 statistical tests on the 1 million whole-exome permutations of 15,122 genes in 438 individuals. The computation was very easy to parallelize and was thus quite affordable using cluster or cloud computing.

Power calculations

Data were extrapolated from results from five candidate genes and one obesity gene set from published studies (**Table 2**). Fisher 's exact test was used to calculate *P* values after sample size extrapolations.

Population stratification

We induced population stratification in a hypothetical exome sequencing study involving real sequencing data and simulated phenotype data using 184 individuals from the HIV and 254 individuals from the SCZ exome sequencing studies. We observed that there were exome-wide differences in allele frequencies between the populations, which we quantified by estimating the F_{ST} between HIV and SCZ samples using exome sequencing data[95]. F_{ST} was estimated using EIGENSOFT software. Using variants with MAFs of at least 5%, we observed an F_{ST} value of 0.003, which is consistent with the different European ancestries of the HIV (European-American) and SCZ (Swedish) samples and with previous estimates of genetic distances between European populations[96]. We considered the possibility that the observed differences between HIV and SCZ samples could be due to differential bias resulting from differences in sample collection, sequencing or data processing[97], but we view this as unlikely because we

applied identical data processing and quality control procedures to both sample sets and because quality control metrics revealed no systematic differences between the sample sets.

To induce population stratification, we randomly assigned 80% of the HIV samples and 20% of the SCZ samples as cases and assigned the remaining samples as controls. We then used case-control labels to run four association tests: two fixed-threshold approaches (T1 and T5 versions of the CMC test[58]), WSS[59] and the VT test[61]. We quantified the evidence of population stratification by considering the most significant *P* value (of 15,122 genes) and the proportion of *P* values < 0.05 and < 0.01. As seen in the null distribution (**Fig. 2**), it is expected that, due to low counts, *P* values will have a deficiency of statistically significant signals. Before correction for population stratification, however, our metrics indicated an excess of statistically significant signals. For example, for T5, the most significant *P* value was < 0.000001, and the proportions of *P* values were 0.0595 at the 0.05 level and 0.0136 at the 0.01 level. Results were similar for the other statistical tests and for other proportions of HIV samples assigned as cases (we experimented with 90%, 80% and 70%, as well as 30%, 20% and 10%). We note that, when the proportion of HIV individuals assigned as cases was above 50%, the induced inflation was higher than when the proportion was below 50%, which could be due to a population genetic excess of rare variants in Swiss and European-American samples relative to Swedish samples.

To correct for stratification, we modified the script that implements association tests (see URLs) to employ a permutation scheme in which case-control status was permuted within each population (HIV and SCZ), assuming known population labels. This permutation scheme does not change the computational cost of the study. The results show that the permutation procedure adequately controlled for population stratification, removing the excess of significant signals. For example, for T5, the most significant *P* value after correction was 0.0001, and the proportions of *P* values were 0.0340 at the 0.05 level and 0.0060 at the 0.01 level. The deficiency of statistically significant signals is due to low counts and is consistent with the null distribution (**Fig. 2**). Results were similar for the other statistical tests and for other proportions of HIV samples assigned as cases. These results show that the permutation-based correction was effective at controlling type 1 errors.

Data access

SCZ control data can be accessed via the database of Genotypes and Phenotypes (dbGAP; phs000473.v1.p1). To access the HIV data, investigators can submit a brief concept sheet detailing their study design, research questions and other needs. The concept sheet with detailed instructions can be downloaded (see URLs). Please send completed forms to P. Richtmyer (prichtmyer@partners.org). Requests will be reviewed on the basis of scientific merit, feasibility and overlap with ongoing concept sheets and investigations.

URLs. Complete Genomics data set, http://www.completegenomics.com/sequence-data/download-data/; R script used for all association analyses (contains T1, T5, WSS and VT tests and optionally weighted PolyPhen-2 predictions), http://genetics.bwh.harvard.edu/rare_variants/; EIGENSOFT, http://www.hsph.harvard.edu/faculty/alkes-price/software/; Picard utilities for manipulation of Sequence Alignment/Map (SAM) files, http://picard.sourceforge.net/index.shtml; Burrows-Wheeler Aligner (BWA), http://bio-bwa.sourceforge.net/; Genome Analysis Toolkit (GATK) suite, http://www.broadinstitute.org/gsa/wiki/index.php/Home_Page; PLINK/SEQ library for management, quality control and analysis of exome sequencing data, including several statistical tests, http://atgu.mgh.harvard.edu/plinkseq/; sample repository research concept sheet, http://cfar.globalhealth.harvard.edu/fs/docs/icb.topic938249.files/Harvard%20CFAR%20Concept%20Sheet%20Template%20.docx.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors are grateful to S. Pollack for assistance with EIGENSOFT. This work was made possible, in part, by the US National Institutes of Health (NIH; grant 5R01 MH084676) and, in part, by the International HIV Controllers Study, supported by the Collaboration for AIDS Vaccine Discovery of the Bill and Melinda Gates Foundation (to P.I.W.d.B.), and the AIDS Clinical Trials Group, supported by the NIH (grants AI069513, AI34835, AI069432, AI069423, AI069477, AI069501, AI069474, AI069428, AI69467, AI069415, Al32782, AI27661, AI25859, AI28568, AI30914, AI069495, AI069471, AI069532, AI069452, AI069450, AI069556, AI069484, AI069472, AI34853, AI069465, AI069511, AI38844, AI069424, AI069434, AI46370, AI68634, AI069502, AI069419, AI068636, RR024975 and AI077505). Sequencing of the SCZ control individuals was funded by the NIH (grant RC2MH089905), the Herman Foundation and the Stanley Medical Research Institute. N.O.S. was supported, in part, by an NIH Training Grant (T32-HL07604-25; Division of Cardiovascular Medicine, Brigham and Women's Hospital). B.M.N. was supported by a National Institute of Mental Health (NIMH) grant (1R01MH089208-01). R.D. is supported by a Canadian Institutes of Health Research Banting Postdoctoral Fellowship. The views expressed in this paper do not necessarily represent the views of the NIMH, NIH, Department of Health and Human Services (HHS) or the US government.

AUTHOR CONTRIBUTIONS

A.K. developed the computational analysis pipeline and analyzed data, K.G. performed upstream quality control and analysis of sequencing data, R.D. performed the power analysis, N.O.S. performed the assessment of rare variants in empirical data, B.M.N. contributed to statistical analyses, and P.J.M. assisted with data analysis. T.L. participated in designing the study. P.S., P.F.S., J.L.M., C.M.H., P.L., P.M., P.I.W.d.B., N.G. and S.M.P. contributed data. A.L.P., P.I.W.d.B. and S.R.S. conceived and designed the study. A.L.P., P.I.W.d.B., S.M.P. and S.R.S. supervised the work. A.K., K.G., R.D., N.O.S., B.M.N., Y.Y.S., A.L.P., P.I.W.d.B. and S.R.S. wrote the manuscript. All authors approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2303.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Fuller, C.W. *et al.* The challenges of sequencing by synthesis. *Nat. Biotechnol.* **27**, 1013–1023 (2009).
2. Rusk, N. & Kiermer, V. Primer: Sequencing—the next generation. *Nat. Methods* **5**, 15 (2008).
3. Metzker, M.L. Sequencing technologies the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
4. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
5. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
6. Ng, S.B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
7. Teer, J.K. & Mullikin, J.C. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145–R151 (2010).
8. Hedges, D.J. *et al.* Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS ONE* **6**, e18595 (2011).
9. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
10. Pierce, S.B. *et al.* Am. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. *J. Hum. Genet.* **87**, 282–288 (2010).
11. Krawitz, P.M. *et al.* Identity-by-descent filtering of exome sequence data identifies *PIGV* mutations in hyperphosphatasia mental retardation syndrome. *Nat. Genet.* **42**, 827–829 (2010).
12. Wang, J.L. *et al. TGM6* identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain.* **133**, 3510–3518 (2010).
13. Ng, S.B., Nickerson, D.A., Bamshad, M.J. & Shendure, J. Massively parallel sequencing and rare disease. *Hum. Mol. Genet.* **19**, R119–R124 (2010).
14. Musunuru, K. *et al.* Exome sequencing, *ANGPTL3* mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* **363**, 2220–2227 (2010).
15. Hoischen, A. *et al. De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).
16. Zhao, Q. *et al.* Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing. *Genome Biol.* **11**, R114 (2010).
17. Wei, X. *et al.* Exome sequencing identifies *GRIN2A* as frequently mutated in melanoma. *Nat. Genet.* **43**, 442–446 (2011).
18. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene *PBRM1* in renal carcinoma. *Nature* **469**, 539–542 (2011).
19. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in *NOTCH1*. *Science* **333**, 1154–1157 (2011).
20. Chang, H. *et al.* Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PLoS ONE* **6**, e21097 (2011).
21. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).

22. Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.* **40**, 592–599 (2008).
23. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684–687 (2010).
24. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
25. Ahituv, N. *et al.* Medical sequencing at the extremes of human body mass. *Am. J. Hum. Genet.* **80**, 779–791 (2007).
26. Romeo, S. *et al.* Rare loss-of-function mutations in *ANGPTL* family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.* **119**, 70–79 (2009).
27. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
28. Pritchard, J.K. & Cox, N. J. The allelic architecture of human disease genes: common disease–common variant…or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
29. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
30. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci. USA* **106**, 3871–3876 (2009).
31. Boyko, A.R. *et al.* Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083 (2008).
32. Williamson, S.H. *et al.* Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* **102**, 7882–7887 (2005).
33. Eyre-Walker, A., Woolfit, M. & Phelps, T. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**, 891–900 (2006).
34. Yampolsky, L.Y., Kondrashov, F.A. & Kondrashov, A.S. Distribution of the strength of selection against amino acid replacements in human proteins. *Hum. Mol. Genet.* **14**, 3191–3201 (2005).
35. Fay, J.C., Wyckoff, G.J. & Wu, C.-I. Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234 (2001).
36. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
37. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
38. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
39. Xue, Y. *et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.* **19**, 1453–1457 (2009).
40. The HIV Controllers Study. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551–1557 (2010).
41. Ewens, W.J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112 (1972).
42. Kimura, M. Molecular evolutionary clock and the neutral theory. *J. Mol. Evol.* **26**, 24–33 (1987).
43. Marth, G.T., Czabarka, E., Murvai, J. & Sherry, S.T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).
44. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1**, 131 (2010).
45. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
46. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
47. Halushka, M.K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**, 239–247 (1999).
48. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
49. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
50. Sunyaev, S., Ramensky, V. & Bork, P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* **16**, 198–200 (2000).
51. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597 (2001).
52. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
55. Hellmann, I. *et al.* Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13**, 831–837 (2003).
56. MacArthur, D.G. & Tyler-Smith, C. Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* **19**, R125–R130 (2010).
57. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149–150 (2003).
58. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
59. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
60. Liu, D.J. & Leal, S.M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.* **6**, e1001156 (2010).
61. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
62. Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11**, 773–785 (2010).
63. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44**, 293–308 (2010).
64. Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* **35**, 606–619 (2011).
65. Stitziel, N.O., Kiezun, A. & Sunyaev, S.R. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* **12**, 227 (2011).
66. Wu, M.C. *et al.* Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am. J. Hum. Genet.* **89**, 82–93 (2011).
67. Neale, B.M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322 (2011).
68. Kotowski, I.K. *et al.* A spectrum of *PCSK9* alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* **78**, 410–422 (2006).
69. Hoffmann, T.J., Marini, N.J. & Witte, J.S. Comprehensive approach to analyzing rare genetic variants. *PLoS ONE* **5**, e13584 (2010).
70. Ionita-Laza, I., Buxbaum, J.D., Laird, N.M. & Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* **7**, e1001289 (2011).
71. Tavtigian, S.V. *et al.* Rare, evolutionarily unlikely missense substitutions in *ATM* confer increased risk of breast cancer. *Am. J. Hum. Genet.* **85**, 427–446 (2009).
72. Sul, J.H., Han, B., He, D. & Eskin, E. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics* **188**, 181–188 (2011).
73. Sul, J.H., Han, B. & Eskin, E. Increasing power of groupwise association test with likelihood ratio test. in *Research in Computational Molecular Biology, Lecture Notes in Computer Science* Vol. 6577/2011 452–467 (Springer, Berlin/Heidelberg, 2011).
74. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
75. Cooper, G.M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* **7**, 250–251 (2010).
76. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
77. Jordan, D.M., Ramensky, V.E. & Sunyaev, S.R. Human allelic variation: perspective from protein function, structure, and evolution. *Curr. Opin. Struct. Biol.* **20**, 342–350 (2010).
78. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32**, 358–368 (2011).
79. Cooper, G.M. & Shendure, J. Needles in stacks of needles: finding disease-causing variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
80. Hicks, S., Wheeler, D.A., Plon, S.E. & Kimmel, M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum. Mutat.* **32**, 661–668 (2011).
81. Stephens, M. & Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
82. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
83. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
84. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
85. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
86. Lipman, P.J. *et al.* On the follow-up of genome-wide association studies: an overall test for the most promising SNPs. *Genet. Epidemiol.* **35**, 303–309 (2011).
87. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
88. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
89. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
90. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
91. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
92. Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.* **39**, 1251–1255 (2007).
93. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
94. Li, H. & Durbin, R. ast and accurate short read alignment with Burrows Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
95. Holsinger, K.E. & Weir, B.S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.* **10**, 639–650 (2009).
96. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
97. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).