STUDY DESIGNS

Statistical power and significance testing in large-scale genetic studies

Pak C. Sham[1] *and Shaun M. Purcell*[2,3]

Abstract | Significance testing was developed as an objective method for summarizing statistical evidence for a hypothesis. It has been widely adopted in genetic studies, including genome-wide association studies and, more recently, exome sequencing studies. However, significance testing in both genome-wide and exome-wide studies must adopt stringent significance thresholds to allow multiple testing, and it is useful only when studies have adequate statistical power, which depends on the characteristics of the phenotype and the putative genetic variant, as well as the study design. Here, we review the principles and applications of significance testing and power calculation, including recently proposed gene-based tests for rare variants.

Likelihoods
Probabilities (or probability densities) of observed data under an assumed statistical model as a function of model parameters.

An important goal of human genetic studies is to detect genetic variations that have an influence on risk of disease or other health-related phenotypes. The typical genetic study involves collecting a sample of subjects with phenotypic information, genotyping these subjects and then analysing the data to determine whether the phenotype is related to the genotypes at various loci. Statistical analysis is therefore a crucial step in genetic studies, and a rigorous framework is required to analyse the data in the most informative way and to present the findings in an interpretable and objective manner. Although there are many frameworks for drawing statistical inferences from data, the most popular framework in genetics is the frequentist significance testing approach, which was proposed by Fisher[1] and further developed by Neyman and Pearson[2] (BOX 1). Most genetic researchers choose to present statistical significance (that is, *P* values) in summarizing the results of their studies. The use of *P* values as a measure of statistical evidence has important limitations[3], and there is little doubt that the Bayesian approach provides a more natural and logically consistent framework for drawing statistical inferences[4,5]. However, Bayesian inference requires prior distributions to be specified for model parameters and intensive computation to integrate likelihoods over the specified parameter space. If different prior distributions are adopted in different studies, then this could complicate the interpretation and synthesis of the findings. Currently, significance testing remains the most widely used, convenient and reproducible method to evaluate the strength of evidence for the presence of genetic effects, although Bayesian analyses may be particularly appealing for fine-mapping a region with multiple significant signals to identify the true causal variants[5].

Inherent in the significance testing framework is the requirement that studies are designed to enable a realistic chance of rejecting the null hypothesis (H_0) when it is false. In the Neyman–Pearson hypothesis testing framework, the probability of rejecting H_0 when the alternative hypothesis (H_1) is true is formalized as the statistical power (BOX 1). Power calculation (BOX 2) is now a required element in study proposals to ensure meaningful results. Although inadequate statistical power clearly casts doubt on negative association findings, what is less obvious is that it also reduces the validity of results that are declared to reach significance. Before the emergence of large-scale association studies and the formation of international consortia in recent years, the study of human genetics has suffered much from the problem of inadequate statistical power, a consequence of which is the frustratingly low rates of successful replication among reported significant associations[6,7].

Power calculations are also important for optimizing study design. Although researchers have no control over the actual genetic architecture that underlies a phenotype, they do have some control of many aspects of study design, such as the selection of subjects, the definition and measurement of the phenotype, the choice of how many and which genetic variants to analyse, the decision of whether to include covariates and other possible confounding factors, and the statistical method to be used. It is always worthwhile to maximize the statistical power of a study, given the constraints imposed by nature or by limitations in resources[8].

[1]Centre for Genomic Sciences, Jockey Club Building for Interdisciplinary Research; State Key Laboratory of Brain and Cognitive Sciences, and Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
[2]Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York 10029–6574, USA.
[3]Center for Human Genetic Research, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA.
Correspondence to P.C.S.
e-mail: pcsham@hku.hk

doi:10.1038/nrg3706

Box 1 | What is statistical power?

The classical approach to hypothesis testing developed by Neyman and Pearson[2] involves setting up a null hypothesis (H_0) and an alternative hypothesis (H_1), calculating a test statistic (T) from the observed data and then deciding on the basis of T whether to reject H_0. In genetic studies, H_0 typically refers to an effect size of zero, whereas H_1 usually refers to a non-zero effect size (for a two-sided test). For example, a convenient measure of effect size in case–control studies is the log odds ratio (log(OR)), where the odds ratio is defined as the odds of disease in individuals with an alternative genotype over the odds of disease in individuals with the reference genotype.

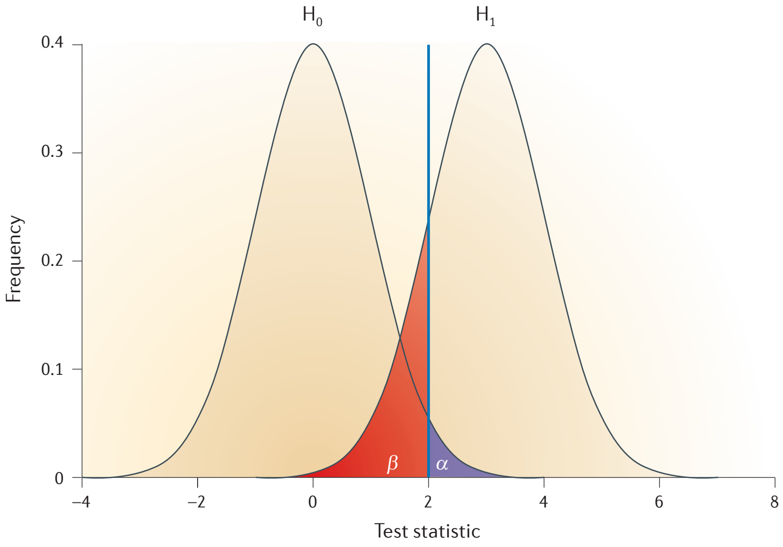
It is important to appreciate that the data obtained from a study and therefore the value of T depend on the particular individuals in the population who happened to be included in the study sample. If the study were to be repeated many times, each drawing a different random sample from the population, then a set of many different values for T would be obtained, which can be summarized as a frequency or probability distribution.

The P value, which was introduced earlier by Fisher[1] in the context of significance testing, is defined as the probability of obtaining — among the values of T generated when H_0 is true — a value that is at least as extreme as that of the actual sample (denoted as t). This can be represented as $P = \mathrm{P}(T \geq t \mid H_0)$.

For a one-sided test (for example, a test for effect size greater than zero), the definition of the P value is slightly more complicated: $P^* = P/2$ if the observed effect is in the pre-specified direction, or $P^* = (1-P)/2$ otherwise, where P is defined as above. In the Neyman–Pearson hypothesis testing framework, if the P value is smaller than a preset threshold α (for example, 5×10^{-8} for genome-wide association studies), then H_0 is rejected and the result is considered to be significant. The range of values of T that would lead to the rejection of H_0 (that is, $T \geq t'$ for which the P value would be less than α) is known as the critical region of the test.

By setting up a hypothesis test in this manner, the probability of making the error of rejecting H_0 when it is true (that is, a type 1 error) is ensured to be α. However, another possible type of error is the failure to reject H_0 when it is false (that is, type 2 error, the probability of which is denoted as β). Statistical power is defined as $1-\beta$ (that is, the probability of correctly rejecting H_0 when a true association is present).

An ideal study should have small probabilities for both types of errors, but there is a subtle asymmetry (see the figure): while the investigator sets the probability of type 1 error (α) to a desired level, the probability of type 2 error (β) and therefore statistical power are subject to factors outside the investigator's control, such as the true effect size, and the accuracy and completeness of the data. Nevertheless, the investigator can try to optimize the study design, within the constraints of available resources, to maximize statistical power and to ensure a realistic chance of obtaining meaningful results.



This schematic representation of the probability distributions of test statistic under H_0 and H_1 shows the critical threshold for significance (blue line), the probability of type 1 error (α; purple) and the probability of type 2 error (β; red). The test statistic is constructed to be standard normal under H_0.

In this Review, we present the basic principles of significance testing and statistical power calculation as applied to genetic studies. We examine how significance testing is applied to large data sets that include millions of genetic variants on a genome-wide scale. We then provide an overview of current tools that can be used to carry out power calculations and discuss possible ways to enhance the statistical power of genetic studies. Finally, we identify some unresolved issues in power calculations for future work.

Multiple testing burdens in genome-wide studies

Genome-wide association studies (GWASs) were made feasible in the late 2000s by the completion of the International HapMap Project[9] and the development of massively parallel single-nucleotide polymorphism (SNP) genotyping arrays, which can now genotype up to 2.5 million SNPs simultaneously[8,10,11]. Partly because of the enormous size of the data sets, GWASs have tended to use simple statistical procedures, for example, logistic regression analysis of either one SNP at a time (with adjustment for potential confounding factors such as ethnic origin) or principal components that are derived from a subset of the SNPs scattered throughout the genome[12,13]. As many SNPs are being tested, keeping the significance threshold at the conventional value of 0.05 would lead to a large number of false-positive significant results. For example, if 1,000,000 tests are carried out, then 5% of them (that is, 50,000 tests) are expected to have $P < 0.05$ by chance when H_0 is in fact true for all the tests. This multiple testing burden has led to the adoption of stringent significance thresholds in GWASs.

In the frequentist framework, the appropriate significance threshold under multiple testing is usually calculated to control the family-wise error rate (FWER) at 0.05. Simulation studies using data on HapMap Encyclopedia of DNA Elements (ENCODE) regions to emulate an infinitely dense map gave a genome-wide significance threshold of 5×10^{-8} (REF. 14). Similarly, by subsampling genotypes at increasing density and extrapolating to infinite density, a genome-wide significance threshold of 7.2×10^{-8} was obtained[15]. Another approach using sequence simulation under various demographic and evolutionary models found a genome-wide significance threshold of 3.1×10^{-8} for a sample of 5,000 cases and 5,000 controls, in which all SNPs were selected with minor allele frequency of at least 5%, for a European population[16]. Subsequently, a genome-wide significance threshold of 5×10^{-8} has been widely adopted for studies on European populations regardless of the actual SNP density of the study. For African populations, which have greater genetic diversity, a more stringent threshold (probably close to 10^{-8}) is necessary[16].

There have been proponents for an alternative approach to multiple testing adjustments that considers only the SNPs that are actually being tested in the study rather than a SNP set with maximal density. Such an approach may be particularly appropriate for studies adopting custom SNP arrays that are enriched for SNPs in candidate disease-relevant genes or pathways, such as the MetaboChip[17] and ImmunoChip[18]. The

Box 2 | Power calculation: an example

As a simple illustrative example, we consider a case–control study that involves a biallelic locus in Hardy–Weinberg equilibrium with allele frequencies 0.1 (for Allele A) and 0.9 (for Allele B). The risk of disease is 0.01 for the BB genotype and 0.02 for the AA and AB genotypes. The study contains 100 cases of subjects with the disease and 100 normal control subjects, and it aims to test the hypothesis that the AA and AB genotypes increase the risk of disease with a type 1 error rate of 0.05.

The association between disease and the putative high-risk genotypes (that is, AA and AB) can be assessed by the standard test for the difference between two proportions. In this scenario, the two proportions are the total frequencies of the AA and AB genotypes in the cases (p_1) and in the controls (p_2). The null hypothesis (H_0) is that the two proportions are equal in the population, in contrast to the alternative hypothesis (H_1) in which the total frequencies of AA and AB in the cases are greater than those in the controls. For a sample size of n_1 cases and n_2 controls, the test statistic is:

$$Z = \frac{p_1 - p_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{n_1p_1 + n_2p_2}{n_1 + n_2}\right)\left(1 - \frac{n_1p_1 + n_2p_2}{n_1 + n_2}\right)}}$$

In large samples, Z is normally distributed and has a mean of zero and a variance of one under H_0.

The distribution of Z under H_1 depends on the values of the two proportions in the population (see the table). The calculation of these two frequencies proceeds as follows. The population frequencies of the three genotypes under Hardy–Weinberg equilibrium are $0.1^2 = 0.01$ (for AA); $2 \times 0.1 \times 0.9 = 0.18$ (for AB); and $0.9^2 = 0.81$ (for BB). This gives a population disease prevalence (K) of $(0.02 \times 0.01) + (0.02 \times 0.18) + (0.01 \times 0.81) = 0.0119$ according to the law of total probability. The genotype frequencies in the cases are therefore $(0.02 \times 0.01)/0.0119 = 0.0168$ (for AA); $0.02 \times 0.18/0.0119 = 0.3025$ (for AB); and $0.01 \times 0.81/0.0119 = 0.6807$ (for BB). Similarly, the genotype frequencies in the controls are $0.98 \times 0.01/0.9881 = 0.0099$ (for AA); $0.98 \times 0.18/0.9881 = 0.1785$ (for AB); and $0.99 \times 0.81/0.9881 = 0.8116$ (for AA).

	AA	AB	BB
Population frequency	0.01	0.18	0.81
Genotype frequency in cases	0.0168	0.3025	0.6807
Genotype frequency in controls	0.0099	0.1785	0.8116

Thus, the total frequencies of the high-risk genotypes (that is, AA and AB) in the cases and the controls are 0.319328 and 0.188442, respectively.

The distribution of Z under H_1 can now be obtained by simulation. This involves using random numbers to generate a large number of virtual samples. In each sample, each case is assigned a high-risk genotype with probability 0.319328, whereas each control is assigned a high-risk genotype with probability 0.188442, so that the proportions of high-risk genotypes among cases and controls can be counted and used to calculate Z. An empirical distribution of Z is obtained from a large number of simulated samples. The mean and standard deviation of this empirical distribution can be used to characterize the distribution of Z under H_1. When a simulation with 1,000 generated samples was carried out for this example, the mean and the standard deviation of the empirical distribution were 2.126 and 0.969, respectively.

Alternatively, it has been shown analytically that the distribution of Z under H_1 has a mean that is given approximately by substituting the sample proportions p_1 and p_2 in the formula for Z by their corresponding population frequencies, and a variance that remains approximately one[88]. In this example, the population frequencies of 0.319328 and 0.188442, and a sample size of 100 per group, gave a mean value of 2.126.

As Z has an approximately normal distribution with a mean of zero and a variance of one under H_0, the critical value of Z that corresponds to a type 1 error rate of 0.05 is given by the inverse standard normal distribution function evaluated at 0.95, which is approximately 1.645. Statistical power can be obtained from the empirical distribution obtained by simulation as the proportion of the generated samples for which $Z > 1.645$. In this example, this proportion was 0.701. Alternatively, using the analytic approximation that Z has a mean of 2.126 and a variance of 1, the probability that $Z > 1.645$ is given by the inverse standard normal distribution function evaluated at $1.645 - 2.126 = -0.481$, which is equal to 0.685. The two estimates of statistical power (0.701 and 0.685) are close to each other, considering that the empirical estimate (0.701) was obtained from 1,000 simulated samples with a standard error of 0.014 (that is, the square root of $(0.701 \times 0.299/1{,}000)$).

traditional Bonferroni correction sets the critical significance threshold as 0.05 divided by the number of tests, but this is an overcorrection when the tests are correlated. Modifications of the Bonferroni method have been proposed to allow dependencies between SNPs through the use of an effective number of independent tests (M_e) (BOX 3). Proposed methods for evaluating M_e in a study include simply counting the number of linkage disequilibrium (LD) blocks and the number of 'singleton' SNPs[19], methods based on the eigenvalues of the correlation matrix of the SNP allele counts (which correspond to the variances of the principal components[20–22]) and a method based directly on the dependencies of test outcomes between pairs of SNPs[23]. A recent version of the eigenvalue-based method[24] has been shown to provide good control of the FWER (BOX 3). When applied to the latest Illumina SNP array that contained 2.45 million SNPs, it gave an estimated M_e of 1.37 million and a

Family-wise error rate
(FWER). The probability of at least one false-positive significant finding from a family of multiple tests when the null hypothesis is true for all the tests.

Box 3 | Bonferroni methods and permutation procedures

The Bonferroni method of correcting for multiple testing simply reduces the critical significance level according to the number of independent tests carried out in the study. For *M* independent tests, the critical significance level can be set at 0.05/*M*. The justification for this method is that this controls the family-wise error rate (FWER) — the probability of having at least one false-positive result when the null hypothesis (H_0) is true for all *M* tests — at 0.05. As the *P* values are each distributed as uniform (0, 1) under H_0, the FWER (α*) is related to the test-wise error rate (α) by the formula $\alpha^* = 1 - (1 - \alpha)^M$ (REF. 89). For example, if α* is set to be 0.05, then solving $1 - (1 - \alpha)^M = 0.05$ gives $\alpha = 1 - (1 - 0.05)^{1/M}$. Taking the approximation that $(1 - 0.05)^{1/M} \approx 1 - 0.05/M$ gives $\alpha \approx 0.05/M$, which is the critical *P* value, adjusted for *M* independent tests, to control the FWER at 0.05. Instead of making the critical *P* value (α) more stringent, another way of implementing the Bonferroni correction is to inflate all the calculated *P* values by a factor of *M* before considering against the conventional critical *P* value (for example, 0.05).

The permutation procedure is a robust but computationally intensive alternative to the Bonferroni correction in the face of dependent tests. To calculate permutation-based *P* values, the case–control (or phenotype) labels are randomly shuffled (which assures that H_0 holds, as there can be no relationship between phenotype and genotype), and all *M* tests are recalculated on the reshuffled data set, with the smallest *P* value of these *M* tests being recorded. The procedure is repeated for many times to construct an empirical frequency distribution of the smallest *P* values. The *P* value calculated from the real data is then compared to this distribution to determine an empirical adjusted *P* value. If *n* permutations were carried out and the *P* value from the actual data set is smaller than *r* of the *n* smallest *P* values from the permuted data sets, then an empirical adjusted *P* value (P^*) is given by $P^* = (r + 1)/(n + 1)$ (REFS 25,26,90).

corresponding significance threshold of 3.63×10^{-8} for European populations. This is close to the projected estimates for SNP sets with infinite density. When applied to the 1000 Genomes Project data on Europeans, the same method gave a significance threshold of 3.06×10^{-8}, which again confirmed the validity of the widely adopted genome-wide significance threshold of 5×10^{-8}, at least for studies on subjects of European descent.

An alternative to a modified Bonferroni approach is to use a permutation procedure to obtain an empirical null distribution for the largest test statistic among the multiple ones being tested (BOX 3). This can be computationally intensive because a large number of permutations is required to accurately estimate very small *P* values[25,26]. Some procedures have been proposed to reduce the computational load, for example, by simulation or by fitting analytic forms to empirical distributions[27,28].

The interpretation of association findings

Before GWASs became feasible, association studies were limited to the investigation of candidate genes or genomic regions that have been implicated by linkage analyses. In a review of reported associations for complex diseases, it was found that only 6 of 166 initial association findings were reliably replicated in subsequent studies[6]. This alarming level of inconsistency among association studies may partly reflect inadequate power in some of the replication attempts, but it is also likely that a large proportion of the initial association reports were false positives.

What is often not appreciated is the fact that both inadequate statistical power and an insufficiently stringent significance threshold can contribute to an increased rate of false-positive findings among significant results (which is known as the false-positive report probability (FPRP)[29]). Although significance (that is, the *P* value) is widely used as a summary of the evidence against H_0, it cannot be directly interpreted as the probability that H_0 is true given the observed data. To estimate this probability, it is also necessary to consider the evidence with regard to competing hypotheses (as encapsulated in H_1), as well as the prior probabilities of H_0 and H_1. This can be done using Bayes' theorem as follows:

$$P(H_0 | P \leq \alpha) = \frac{P(P \leq \alpha | H_0)P(H_0)}{P(P \leq \alpha | H_0)P(H_0) + P(P \leq \alpha | H_1)P(H_1)}$$

$$= \frac{\alpha\pi_0}{\alpha\pi_0 + (1 - \beta)(1 - \pi_0)}$$

In this formula, $P(H_0 | P \leq \alpha)$ is the FPRP given that a test is declared significant, and π_0 is the prior probability that H_0 is true. Although the term $P(P \leq \alpha | H_1)$ is often interpreted as the statistical power $(1 - \beta)$ under a single H_1, for complex traits and in the context of GWASs, it is likely that multiple SNPs have a true association with the trait, so that it would be more accurate to consider $P(P \leq \alpha | H_1)$ as the average statistical power of all SNPs for which H_1 is true. This formula indicates that, when a study is inadequately powered, there is an increase in the proportion of false-positive findings among significant results (FIG. 1). Thus, even among association results that reach the genome-wide significance threshold, those obtained from more powerful studies are more likely to represent true findings than those obtained from less powerful studies.

The above formula can be used to set α to control the FPRP as follows:

$$\alpha = \frac{P(H_0 | P \leq \alpha)}{1 - P(H_0 | P \leq \alpha)} \, \frac{1 - \pi_0}{\pi_0} (1 - \beta)$$

When the power $(1 - \beta)$ is low, α has to be set proportionately lower to maintain a fixed FPRP; that is, the critical *P* value has to be smaller to produce the same FPRP for a study with weaker power than one with greater power. Similarly, when the prior probability that H_0 is true (that is, π_0) is high, $(1 - \pi_0)/\pi_0$ is low, then α again has to be set proportionately lower to keep the FPRP fixed at the desired level.

The fact that multiple hypotheses are tested in a single study usually reflects a lack of strong prior hypotheses and is therefore associated with a high π_0. The Bonferroni adjustment sets α to be inversely proportional to the number of tests (*M*), which is equivalent to assuming a fixed π_0 of $M/(M + 1)$; this means that one among the *M* tests is expected to follow H_1. This is likely to be too optimistic for studies on weak candidate genes but too pessimistic for GWASs on complex diseases. As genomic coverage increases, hundreds (if not thousands) of SNPs are expected to follow H_1. As studies become larger by combining data from multiple centres, the critical significance level that is necessary for controlling the FPRP is expected to increase so that many results that are close to the conventional genome-wide significance level

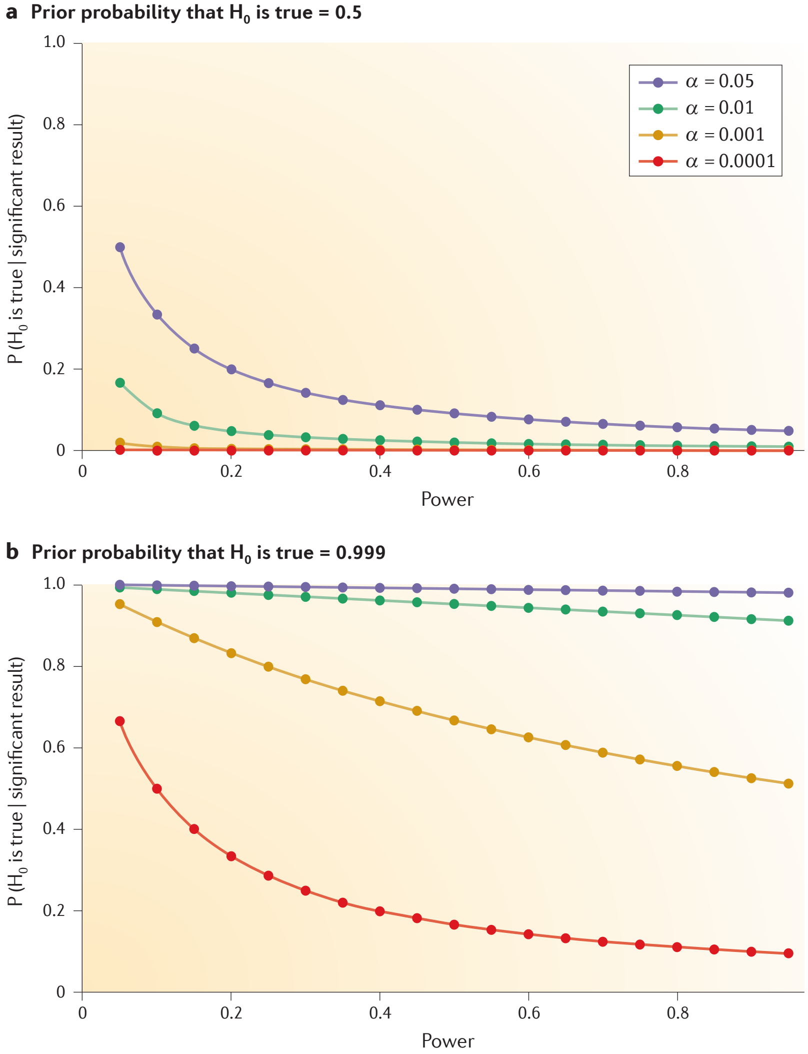


Figure 1 | **Posterior probability of H_0 given the critical significance level and the statistical power of a study, for different prior probabilities of H_0.** The probability of false-positive association decreases with increasing power, decreasing significance level and decreasing prior probability of the null hypothesis (H_0).

of 5×10^{-8} will turn out to be true associations. Indeed, it has been suggested that the genome-wide threshold of significance for GWASs should be set at the less stringent value of 10^{-7} (REF. 30).

Although setting less stringent significance thresholds for well-powered studies has a strong theoretical basis, it is complicated in practice because of the need to evaluate the power of a study, which requires making assumptions about the underlying disease model. An alternative way to control the FPRP directly without setting a significance threshold is the false discovery rate (FDR) method[31], which finds the largest *P* value that is substantially smaller (by a factor of at least 1/φ, where φ is the desired FDR level) than its expected value given that all the tests follow H_0, and declares this and all smaller *P* values as being significant. Although FDR statistics are rarely presented in GWAS publications, it is common to present a quantile–quantile plot of the *P* values, which captures the same information as the FDR method by displaying negatively ranked log *P* values against their null expectations (the expectation that the r^{th} smallest *P* value of *n* tests is $r/(n+1)$, when H_0 is true for all tests). The quantile–quantile plot has the added advantage that very early departure of negatively ranked log *P* values from their expected values is a strong indication of the presence of population stratification[8].

Another approach to control the FPRP is to abandon the frequentist approach (and therefore *P* values) completely and to adopt Bayesian inference using Bayes factors as a measure of evidence for association[4]. A Bayes factor can be conveniently calculated from the maximum likelihood estimate (MLE) of the log odds ratio and its sampling variance, by assuming a normal prior distribution with a mean of zero and variance *W* (REF. 32). By specifying *W* as a function of an assumed effect size distribution, which may be dependent on allele frequency, one obtains a Bayes factor that can be interpreted independently of sample size. It is interesting that, if *W* is inappropriately defined to be proportional to the sampling variance of the MLE, then the Bayes factor will give identical rankings as the *P* value, which offers a link between these divergent approaches[32]. A greater understanding of Bayesian methods among researchers, and the accumulation of empirical data on effect sizes and allele frequencies to inform specification of prior distributions, should promote the future use of Bayes factors.

Determinants of statistical power

Many factors influence the statistical power of genetic studies, only some of which are under the investigator's control. On the one hand, factors outside the investigator's control include the level of complexity of genetic architecture of the phenotype, the effect sizes and allele frequencies of the underlying genetic variants, the inherent level of temporal stability or fluctuation of the phenotype, and the history and genetic characteristics of the study population. On the other hand, the investigator can manipulate factors such as the selection of study subjects, sample size, methods of phenotypic and genotypic measurements, and methods for data quality control and statistical analyses to increase statistical power within the constraints of available resources.

Mendelian diseases are caused by single-gene mutations, although there may be locus heterogeneity with different genes being involved in different families; the genetic background or the environment has little or no effect on disease risk under natural conditions. The causal mutations therefore have an enormous impact on disease risk (increasing it from almost zero to nearly one), and such effects can be easily detected even with modest sample sizes. An efficient study design would be to genotype all informative family members using SNP chips for linkage analysis to narrow down the genome to a few candidate regions, and to capture and sequence these regions (or carry out exome sequencing followed by *in silico* capture of these regions, if this is more convenient and cost effective) in one or two affected family members to screen for rare, nonsynonymous mutations. Nevertheless, statistical power can be reduced both when there is misdiagnosis of some individuals owing to phenotypic heterogeneity and phenocopies, and when there is locus heterogeneity in which mutations from multiple loci all cause a similar phenotype.

Some diseases have rare Mendelian forms and common complex forms that are phenotypically similar. Cases caused by dominant mutations (for example, familial Alzheimer's disease and familial breast cancer) will usually cluster in multiplex families and are therefore easily distinguishable from complex forms. Such families can be investigated using the same methodology as for Mendelian diseases. However, for a common disease in which a small proportion of cases are caused by recessive mutations, these cases will tend to be sporadic and difficult to distinguish from other cases. However, as few genes would contain two rare nonsynonymous (homozygous or compound heterozygous) mutations in the same subject by chance, when this occurs in a patient there is a strong possibility that the mutations are causal to the disease.

Almost all common diseases have a complex genetic architecture that involves multiple risk loci and environmental factors, but the level of complexity seems to differ for different classes of disease. Generally, as the sample size of GWASs increases, the number of SNPs with evidence of association that reaches genome-wide significance will tend to increase, but this rate of increase seems to differ for different diseases. Thus, the minimum sample size that is required to detect any significant SNPs differs among diseases, but beyond this threshold a doubling of the sample size generally results in approximately doubling the number of significant SNPs[33]. For some diseases, such as Crohn's disease and ulcerative colitis, the numbers of significant SNPs reached double figures when the sample sizes were increased beyond ~5,000 cases, whereas for type 2 diabetes and breast cancer sample sizes of more than 30,000 cases were required to reach the same numbers of significant SNPs[33]. These differences in the ease of SNP detection among diseases clearly reflect differences in the numbers of susceptibility SNPs that underlie the diseases, which in turn are determined by the level of complexity of the mechanisms and processes leading to disease occurrence.

The sample size required for detecting a particular genetic variant is determined by both the frequency and the effect size of the variant. For population cohort studies, power calculations can be simplified by assuming a liability threshold model for disease. This model postulates that individual predisposition to disease is characterized by an unobserved variable called liability, which is continuous and normally distributed in the population. The liability is determined by both genetic and environmental factors, and disease occurs in an individual whose liability exceeds a certain threshold. Under this liability threshold model, the non-centrality parameter (NCP) (BOX 4) of the association test of the SNP in a random sample from the population is directly related to the proportion of variance in liability explained by the locus (V_G), which is in turn approximately determined by the allele frequency (p) and the effect size (ln(OR)) as follows[34]:

$$V_G = \frac{2p(1-p)(\ln(\mathrm{OR}))^2}{\frac{\pi^2}{3} + 2p(1-p)(\ln(\mathrm{OR}))^2}$$

This implies that low-frequency alleles of relatively small effect sizes will be especially hard to detect because they will individually account for very little variance in liability. For case–control samples, power is still strongly influenced by V_G, but the relationship is more complex and it is simpler to calculate power directly from the assumed odds ratio and the allele frequency of the putative risk variant using software such as the Genetic Power Calculator (GPC)[35].

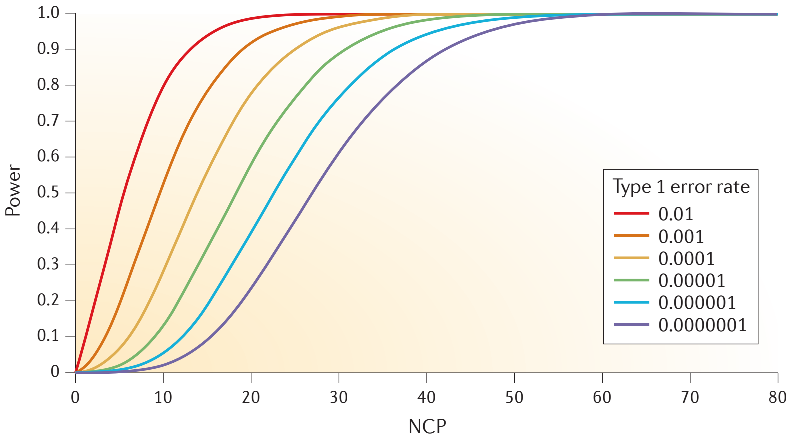
There is a caveat when specifying the assumed odds ratio in a replication study to be the estimate obtained from the original discovery sample: the odds ratio estimate is likely to be upwardly biased by the "winner's curse" phenomenon, particularly if the original report was a screen of a large number of variants and if the original P value was close to the significance threshold[36]. Methods for correcting this bias have been proposed[37–39], and the corrected effect sizes should be used in power calculations of replication studies.

As most current GWASs genotype only a subset of all variants in the genome and rely on LD between typed markers and untyped variants to increase the coverage of the genome, an important determinant of power in an association study is the level of LD between the typed marker and the true causal variant. If a direct association analysis of a causal SNP would provide an NCP of λ, then an indirect association analysis of a SNP that has correlation R with the causal SNP will have an NCP of $R^2\lambda$ (REF. 40). In other words, when a proxy SNP that has correlation R with the causal SNP is analysed, (instead of the causal SNP itself) the sample size required to obtain the same level of statistical power is increased by a factor of $1/R^2$.

The frequency of an allele is an important determinant of statistical power for two reasons. First, the phenotypic variance explained by a genetic locus is directly proportional to heterozygosity, which is in turn determined by the frequencies of the alleles at the locus, under random mating. Second, a rare variant is less likely than a common allele to have a high R^2 with SNPs that are included in commercial SNP arrays because these SNPs have been selected to represent common variation, and the R^2 between two loci that have very different allele frequencies is necessarily small. These considerations remain relevant in modern studies that use imputation to analyse untyped SNPs. The statistical power for such SNPs depends on the quality of imputation, which is in turn determined by the level of LD between the untyped and typed SNPs (or between the untyped SNP and haplotypes formed by the typed SNPs). Generally, rare variants are more difficult to impute accurately than common alleles.

Another important determinant of statistical power is phenotype definition. The ideal phenotype, in terms of ease of finding susceptibility variants, should be rare rather than common (so that it represents more extreme

Box 4 | Non-central chi-squared distribution



The central limit theorem states that a test statistic that consists of additive contributions from multiple observations would tend to have a normal distribution in large samples. The variance of the statistic is usually standardized to one by appropriate scaling. Furthermore, the statistic is usually constructed such that its mean is zero when the null hypothesis (H_0) is true. Remarkably, when the alternative hypothesis (H_1) is true, the distribution is often simply 'shifted' so that the mean becomes non-zero, whereas the variance remains approximately one (BOX 3).

The square of a standard normal variable has a chi-squared distribution with one degree of freedom. Thus, many statistical tests in genetic studies can be expressed in either normal or chi-squared forms. Furthermore, when considering H_1, it is convenient to define the square of a normal variable (which has mean μ and a variance of one) to have a non-central chi-squared distribution with a non-centrality parameter (NCP, which is denoted as λ) which has a mean of μ^2 and one degree of freedom. The mean of a variable with a non-central chi-squared distribution is equal to the sum of its NCP and degrees of freedom, compared with a central chi-squared distribution in which the mean is simply equal to its degrees of freedom.

The non-central chi-squared distribution is useful in analytic power calculations[91]. This is because calculating the NCP is often a convenient intermediate step in power calculations, as it has a fairly simple relationship to sample size and various parameters assumed under H_1. Moreover, the NCP fully specifies the statistical power for any chosen type 1 error rate (see the figure). In particular, for any given set of parameters under H_1, the NCP is directly proportional to sample size, so that it is simple to extrapolate the NCP to any sample size from the NCP value that has been calculated for a particular sample size.

For the example in BOX 2, the test can be formulated as a chi-squared test by taking the square of Z as the test statistic. The NCP of the chi-squared test statistic would be $2.126^2 = 4.522$ under H_1, whereas the critical value for significance is $1.645^2 = 2.706$. From the distribution function of the non-central chi-squared distribution (for example, that obtained using the Genetic Power Calculator (GPC)), the statistical power of the test is found to be 0.685. If the sample size is doubled to 200 cases and 200 controls, then the statistical power can be found by simply doubling the NCP to 9.044 which, when looked up in the non-central chi-squared distribution function at 2.706, gives a statistical power of 0.913.

selection under a liability threshold model), distinctive rather than obscure, stable rather than changeable, and highly familial. Phenotypes without these favourable characteristics (for example, depression and hypertension) may require much larger sample sizes for susceptibility loci to be detected. Despite the large amount of resources required, genetic studies of such traits may nevertheless be warranted because of their public health importance and the lack of alternative strategies for elucidating their aetiology and mechanisms. In such cases, there may be ways of efficiently improving statistical power, for example, by taking either longitudinal approaches (that is, multiple measurements of the same trait at different time points) or multivariate approaches (that is, measurement of multiple correlated but different traits) for phenotype definition to reduce the influence of temporal fluctuations and measurement errors. Other methods to improve statistical power include sampling patients who are most likely to have a high genetic loading (for example, those with familial, early-onset, severe and chronic disease), and sampling controls who are 'super normal' in terms of a suitable endophenotype (for example, controls whose glucose tolerance is above the population average when studying diabetes, for which impaired glucose tolerance is an important diagnostic criterion) (BOX 5).

Statistical power is also influenced by the procedures used in data analyses. Appropriate data quality control to filter out problematic SNPs and subjects helps to reduce false-positive associations and to improve power. Statistical analyses that take appropriate account of covariates (for example, age and sex) and potential confounding factors (for example, hidden population stratification) also tend to improve statistical power. An interesting exception is the use of logistic regression in case–control studies, in which the inclusion of a covariate can result in a reduction in power[41]. Statistical power is influenced by the hypothesis for which the test is designed. For example, a test that assumes additive effects would have greater power than a test that also allows dominance, if the true effects at the locus are indeed additive and do not show dominance. Conversely, if the underlying causal variant is recessive, then power would be lost by carrying out an analysis that assumes additivity, and the analysis should consider the overall risk genotype rather than a single risk allele. If there is uncertainty regarding the true pattern of effects at a locus, then it might be appropriate to use several statistical tests to ensure adequate statistical power for all possible scenarios. Although this raises the issue of multiple testing, there are proposed procedures for dealing with this[42–44].

Power for tests of rare variants

Recent advances in large-scale sequencing are driving a wave of studies that aim to uncover rare variants that may underlie disease risk. Such studies often target the coding portion of the genome: exome sequencing can be now routinely applied to both Mendelian[45] and common[46] diseases. Both population genetic theory[47] and recent empirical studies[48,49] suggest that rare alleles will be enriched for functional and deleterious effects and will thus be disproportionately represented among disease alleles. However, sequencing studies of rare variation face various challenges in their design and statistical analyses, most obviously that of low power because one major determinant of statistical power is allele frequency. This fact has prompted various methodological innovations to improve power (see below).

Here, 'rare' refers to a population frequency of less than 1% (although some have used 0.5% as the cutoff), whereas 'low' refers to a frequency between 1% and 5%. In large sequencing studies, the majority of variants discovered will be rare. For example, in 10,000 haploid chromosomes most discovered alleles will be observed

Box 5 | Power calculation for association studies

Power calculation for case–control association studies (BOX 2) involves calculating the genotype (or allele) frequencies in cases and controls, and substituting these into the formula for the chi-squared test for equality of proportions to give the non-centrality parameter (NCP, which is denoted as λ) for the test. This has been implemented in the Genetic Power Calculator (GPC) software[35]. GPC also provides a power calculator for the transmission disequilibrium test for case–parent trio data. However, for rare diseases and minor-effect loci there is a convenient method to simplify power calculation for family-based association designs. The rules are that a case–parent trio is approximately equivalent to a pair of unrelated case–control individuals, and that a case with k unaffected siblings is approximately equivalent to $k/(k+1)$ pair of unrelated case–control individuals[92]. This allows both case–parent and case–sibling families to be converted to case–control equivalents. For samples with a mixture of unrelated cases and controls, case–parent families and case–sibling families, the case–control equivalents and NCPs of the three data types can be calculated; the overall NCP can be used to evaluate statistical power.

For quantitative traits, the NCP for an association analysis of a random population sample is given by:

$$\lambda = N \times \frac{\beta^2 \mathrm{Var}(X)}{\text{Residual variance of } Y}$$

In this formula, Y is the trait; X is the allele count at a genetic locus (coded 0, 1 or 2) so that under Hardy–Weinberg equilibrium the variance of X ($\mathrm{Var}(X)$) is given by $2p(1-p)$, where p is the allele frequency at the locus, and β is the regression coefficient of Y on X. For a minor-effect locus, the residual variance of Y is not much smaller than the total variance of Y, so that the NCP is given by proportion of trait variance explained by the quantitative trait locus (QTL) (that is, V_A) multiplied by the sample size (N). If the trait is not measured perfectly but subject to measurement error, then this reduces the NCP for association by attenuating the proportion of trait variance explained by the QTL relative to the ideal situation of perfect measurements. Taking the average of several repeated trait measurements provides a more accurate measure of the trait. The measurement error variance is estimated by $1-r$, where r is the test-retest correlation; taking the average of k repeated measurements inflates the proportion of trait variance explained by the QTL from V_A to V_A':

$$V_A' = V_A\left(\frac{k}{1+(k-1)r}\right)$$

For randomly recruited sibpairs from the population, association analysis of a quantitative trait can be partitioned into a between-sibpair component (the NCP of which is denoted as λ_B) and a within-sibpair component (the NCP of which is denoted as λ_W)[93], which are given approximately by:

$$\lambda_B \approx \frac{\frac{3}{2}V_A + \frac{5}{4}V_D}{2V_S + V_N} \quad \text{and} \quad \lambda_W \approx \frac{\frac{1}{2}V_A + \frac{3}{4}V_D}{V_N}$$

In these formulae, V_A is the additive QTL variance for the trait, V_D is the dominance QTL variance, V_S is the residual shared variance, and V_N is the residual non-shared variance[40]. Power calculations using these analytic NCP expressions have been implemented in the GPC[35].

For quantitative traits, one can improve the power of association analyses by oversampling individuals whose trait values are either far below or far above the general population average, and by omitting individuals whose trait values are near the general population average. Compared with the same number of randomly selected, unrelated individuals, this extreme selection design increases the NCP by a factor of V'/V, where V' is the trait variance of the selected sample and V is the trait variance of the general population. Thus, the statistical power of any selected sample can be calculated from V' as long as V is known. For example, assuming a normally distributed trait, ascertaining individuals only from the top 5% and bottom 5% of the distribution yields $V'/V \approx 4.4$, which means that less than one-quarter of the number of individuals need to be genotyped to preserve power, relative to the sample size required from an unselected sample. For sibpair data, appropriate selection according to the trait values of the siblings can also increase statistical power[94].

For a quantitative trait (Y) and a random sample from the general population, the interaction between a quantitative environmental variable (X) and a dichotomous genotype (G; for example, AA and AB combined, versus BB) is assessed by a test of the equality of regression coefficients (β_1 and β_2) of Y on X in the two genetic groups[95]. If both X and Y have been standardized to have a variance of one, then the NCP for this test is given approximately by:

$$\lambda = Np_1p_2(\beta_1 - \beta_2)^2$$

In this formula, N is the overall sample size; p_1 and p_2 are the proportions of individuals in the two genotype groups. If an additive genetic model is assumed, the genetic effect coded as allele count (0, 1 and 2) and the change in regression coefficient of Y on X per extra allele is I (that is, $\beta_{AA} - \beta_{AB} = I$, and $\beta_{AA} - \beta_{BB} = 2I$), then the NCP is given by:

$$\lambda = N2p(1-p)I^2$$

In this formula, p is the allele frequency, and Hardy–Weinberg equilibrium is assumed. Quanto is a program for power calculations of gene–gene and gene–environment interactions[96,97].

only once and will not be expected to recur in a further 10,000 chromosomes. In 1,000 case–control pairs, for a disease of 1% population risk, there is moderate power (70%) to detect a low-frequency (1 in 50) allele with a dominant threefold increased risk, at genome-wide significance ($P < 5 \times 10^{-8}$) (see GPC: Case–control for discrete traits). For a considerably rarer allele (for example, an allele with a frequency of 1 in 2,000), power will only be maintained if the effect size is correspondingly greater (in this case, a nearly 40-fold increased risk), which implies a high penetrance of nearly 40%, as the absolute disease risk in people without the variant will be only slightly less than the population risk of 1%.

Unlike GWASs, rare-variant sequencing studies typically have to detect variable sites before testing those sites for association with disease. It is possible to separate detection and association, for example, by first sequencing a modest number of samples to detect variants that can be followed up by genotyping in a larger sample. Whether a variant is detected at the first stage depends on the sample size, and the coverage and sensitivity of the sequencing and variant calling. One can increase the likelihood of detecting disease susceptibility variants by sequencing cohorts of affected individuals, although this will induce bias if the same case samples are then used in subsequent association testing[50,51]. The two-stage logic of sequencing followed by genotyping is embodied in the recently designed Exome Chip, in which exome sequence data on more than 10,000 individuals were used to compile a list of recurrently observed nonsynonymous mutations, most of which are rare or of low frequency.

Mendelian diseases. Mendelian diseases are characterized by genetic architectures with rare alleles of high penetrance and fairly low levels of genetic heterogeneity, at least at the level of the locus if not the specific allele. Almost all individuals who carry a risk mutation will have the disease, and a large proportion of affected individuals will share either the same rare mutation or rare mutations in the same gene. Typically, disease mutations will be 'smoking guns' with clear and deleterious effects, such as premature termination, nonsense frameshift insertion or deletion, or missense single-nucleotide change leading to substitution with an amino acid that has very different properties. Given these assumptions, various 'filtering' approaches[52] have been successfully applied to an increasing number of diseases, particularly recessive ones[53]. Given a small sample of related or unrelated cases, variants are generally filtered by novelty (or extreme rarity); by clear functional impact (for example, those that disrupt genes) and optionally by mode of inheritance (for example, rare homozygous genotypes for recessive disease); and by identity-by-descent sharing and co-segregation with disease in families. Disease-causing genes can be identified as those harbouring more putative disease mutations than that expected by chance after filtering[45]. As the genetic complexity of the disease increases (for example, reduced penetrance and increased locus heterogeneity), issues of statistical power quickly become paramount. Simulation studies have examined the relative effect of sample size, locus heterogeneity, mode of inheritance, gene and total target size, filtering efficiency and sequencing accuracy on the power of exome sequencing studies in Mendelian diseases[54]. However, a full treatment that models the sensitivity and specificity of the filtering procedure, as well as the genetic architecture of the Mendelian disease, has yet to be developed.

Complex diseases. For complex diseases, sequencing studies face more difficult challenges than those for Mendelian diseases[46,55], as many carriers of a given rare susceptibility variant are likely to be unaffected and, conversely, the variant is likely to be present only in a small proportion of affected individuals. Simple filtering based on either the frequency of a variant or co-segregation in families will have low sensitivity and low specificity to screen for true susceptibility variants. Statistical tests for association are therefore needed, despite the low frequencies of many of the detected alleles. In response to the low power for single-variant tests and the inapplicability of Mendelian filtering strategies for common diseases, the gene (rather than the individual allele) has become a primary unit of analysis for exome sequencing studies. This approach aims to combine weak signals from multiple mutations in the same gene to give a more powerful test for association.

The simplest gene-based test compares the frequency of rare mutations in a gene in cases to that in controls. This is statistically and conceptually similar to a single-locus test, except that it is based on the aggregated allele frequency and averaged effect size for all mutations in the gene. A large number of extensions have been proposed[56–59], reviewed[60,61] and quantitatively compared[62–65]. Gene-based tests differ in whether they assume that associated rare mutations will always increase risk as opposed to allowing a mixture of risk and protective rare alleles in the same gene (for example, the C-alpha test[66] and the sequence kernel association test (SKAT)[67]). For studies in which super normal controls have been ascertained, the application of such two-sided tests may be desirable if a priori rare variants with opposite effects are likely to coexist in the same gene. For many case–control studies, in which controls are broadly representative of the unselected population, there will be low power to detect rare protective effects, and the benefit of two-sided tests will therefore be greatly diminished. Tests may also differ in the extent to which their power is reduced when there is a substantial proportion of null alleles; an appropriately constructed score test such as SKAT may have an advantage in this regard. Other differences between gene-based tests involve the way that they weight and group variants, practical differences (such as the applicability to quantitative traits and to related individuals) and the ability to include covariates.

Strategies to increase power of exome studies

Despite a growing body of literature, the precise methodological detail of the adopted gene-based approach is likely to be a less important determinant of success than the effect of the underlying genetic architecture of the disease. So far, the few published applications of exome sequencing in moderately sized case–control sequencing

C-alpha test
A rare-variant association test based on the distribution of variants in cases and controls (that is, whether such a distribution has inflated variance compared with a binomial distribution).

Sequence kernel association test
(SKAT). A test based on score statistics for testing the association of rare variants from sequence data with either a continuous or a discontinuous genetic trait.

studies have yielded broadly negative results[68]. It therefore seems that, similar to GWASs, large sample sizes will be necessary for exome sequencing studies to identify rare or low-frequency risk alleles. Recent predictions from population genetic theory in fact suggest that very large sample sizes, probably more than 25,000 cases, will be required[69]. Nevertheless, there are various other ways in which power could be increased.

Study design considerations. One can restrict association testing to variants that are a priori more likely to be causal both to reduce the multiple-testing burden and to avoid the dilution of signal by including neutral variants. By filtering or weighting, one could assign priority to nonsense alleles over missense alleles or to probable deleterious missense alleles (for example, as predicted by tools such as PolyPhen2 (REF. 70)) over probable benign missense alleles. Alternatively, very rare alleles can be assigned a higher priority than more common ones, as highly deleterious mutations are likely to be subjected to negative selection. However, current methods for characterizing the functional importance of variants are relatively crude, and there is an urgent need for novel, more accurate and comprehensive approaches for identifying variants that are more likely to increase disease risk than others. One potentially powerful approach is to assess conservation across and within multiple species as whole-genome sequence data become more abundant.

Sampling cases or controls from the extremes of an appropriate quantitative distribution can often increase power[71]. Such distributions may be either directly observable (such as body mass index and biomarker level) or inferred from statistical models (for example, identification of controls with unusually high levels of non-genetic risk factors on the basis of the residuals from a logistic regression model and the assumption that these subjects will be enriched for protective genetic factors[72]). However, it should be acknowledged that the extensive sampling and phenotypic assessment required for extreme selection may be costly and reduce the overall effectiveness of this approach. Studying families with multiple affected members can increase the power to detect rare variants of moderate to high penetrance (which are enriched in familial cases compared with sporadic cases). Ascertaining cases with a positive family history of the disease increases power even if only one member from each family is genotyped and analysed in a case–control context, and this may indeed be an efficient strategy, even though co-segregation within families of a haplotype with disease can provide some additional information[55,73,74].

Power might also be increased by including types of variants other than single-nucleotide variants (SNVs). Although they are harder to call and annotate, insertion or deletions, multinucleotide variants and structural variants (including copy-number variants, translocations and inversions) constitute a smaller set of variation (in terms of the number of discrete events an individual is expected to carry) relative to all SNVs and are more likely to have functional effects. For example, in autism, researchers have demonstrated rare variant burdens for copy-number variants[75], *de novo* point mutations[76] and complete knockout mutations that arise from rare recessive and compound heterozygous loss-of-function mutations[77].

Analytical approaches. Integrative models that consider multiple classes of genetic variants provide another avenue by which power can be increased. For example, this has been done by combining genotype data for common variants with rare-variant data obtained from sequencing, and by formulating more powerful joint tests[78,79]. Similarly, an integrative model has been proposed to combine *de novo* and inherited variants in a gene-based likelihood framework, in which parameters (such as variant frequencies and penetrance) are estimated by a hierarchical Bayes strategy that uses information on genes across the genome[80]. Testing larger collections of variants across sets of functionally related genes is a further option to increase power, as demonstrated in copy-number variant studies[75]. Sets of genes can be defined, either explicitly or implicitly, by generic annotations such as Gene Ontology terms, by considering proximity or relatedness in protein–protein interaction or gene expression networks, or by other domain-specific knowledgebases[81].

There is currently a lack of consensus in power calculations for rare-variant sequencing studies on several key points. The first issue is how to parameterize multi-locus genetic models with appropriate joint distributions of effect sizes and frequencies to restrict attention to a limited number of realistic scenarios (ideally informed by population genetic theory[69]). The second is how to account for the technical features of sequencing, such as incomplete and biased coverage or variant calling, which introduce differences in data quality and therefore power across the genome. The third concern is how to assess the appropriate multiple-testing thresholds in light of the different ways of grouping variants in either gene-based or pathway-based analyses; a complication is that many tests may be carried out that could never achieve nominal significance, such as those involving genes with only one or two singleton mutations. Some existing and evolving tools address aspects of these problems. For example, software such as SimRare[82] and SEQPower have been developed to contrast the relative powers of different analytic approaches under different assumed true models. Even when standard test statistics (for example, Pearson chi-squared tests or likelihood ratio tests) are used, analytic power calculations that involve rare variants can give rise to biased results, as the assumptions that ensure the large-sample (that is, asymptotic) properties of the tests may not hold. It may therefore be necessary to validate analytic power calculations by computer simulations at least for a few key scenarios. Another issue that requires further work is the potential impact of population stratification on rare-variant association tests; a recent study has shown type 1 error inflation and power reduction in these tests compared with analyses of common variants[83].

Similar to GWASs for common variants, meta-analyses of multiple studies are likely to be necessary for the detection of rare-variant effects in exome sequencing studies. The wide varieties of gene-based tests for rare variants and the availability of several options — such as the choice of weighting scheme or minor allele frequency cutoff — are likely to complicate meta-analyses of exome sequencing studies. However, recent proposals of using multivariate score tests, which are constructed from single-variant test statistics, promise to provide a flexible framework for meta-analyses of exome sequencing studies[84,85]. Furthermore, the scores statistics and their covariance matrix allow many popular rare-variant gene-based tests to be constructed, thus providing a unified analytic framework for single data sets and meta-analyses. It has been shown that the combination of summary single-variant statistics from multiple data sets, rather than the joint analysis of a combined data set, does not result in an appreciable loss of information[85], and that taking into account heterogeneity in effect size across studies can improve statistical power[84].

The SKAT program provides a power calculation tool that allows the specification of multiple parameters, including size of the genomic region (or regions) to be tested, the maximum allele frequency of rare variants, the proportion of rare variants that increase or decrease disease risk, the effect size of individual rare variants (which can be constant or can increase with decreasing variant frequency) and the significance threshold[86]. There are currently no guidelines on the appropriate settings for these parameters which, in any case, will depend on the genetic disorder and the analysis (that is, whether it is gene based or pathway based). In general, for tests involving large genomic regions (for example, pathway-based tests), it would be more realistic to assume that fewer variants (for example, 5%) have an effect on risk than tests that involve small genomic regions, when it may be more reasonable to assume a larger proportion of risk-altering rare variants (for example, 20%). The odds ratio for rare variants may be as high as 57 — the largest estimated effect size in an analysis of rare copy-number variants in schizophrenia[87].

Conclusion

Significance testing is the most commonly used method for evaluating statistical hypotheses in genetic studies, including GWASs and exome sequencing studies. However, as described above, *P* values are difficult to interpret without some consideration of statistical power, as an insignificant test can result both from the absence of effect and from inadequate statistical power. Consideration of statistical power is therefore important not only in the design of efficient genetic studies but also in the interpretation of statistical findings.

The latest challenges in significance testing and power calculations come from the analysis of rare variants. Although many methods have been proposed, the evaluation of their performance is not straightforward because of the uncertainties and complexities of the true underlying model. Faced with variants of low frequencies and modest effect sizes, new methods are emerging that attempt to gain statistical power by taking into account knowledge of genome function and organization. Such methods are well developed for coding variants but are still in their infancy for regulatory regions of the genome. Meanwhile, rapid advances in rare-variant gene-based and pathway-based tests are paving the way for more powerful exome sequencing studies for complex diseases.

1. Fisher, R. A. *Statistical Methods for Research Workers* (Oliver and Boyd, 1925).
2. Neyman, J. & Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A* **231**, 289–337 (1933).
3. Nickerson, R. S. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* **5**, 241–301 (2000).
4. Balding, D. J. A tutorial on statistical methods for population association studies. *Nature Rev. Genet.* **7**, 781–791 (2006).
5. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature Rev. Genet.* **10**, 681–690 (2009).
This is a highly readable account of Bayesian approaches for the analysis of genetic association studies.
6. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
7. Ioannidis, J. P. A. Genetic associations: false or true? *Trends Mol. Med.* **9**, 135–138 (2003).
8. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
9. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
10. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
11. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
12. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
13. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
14. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
15. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
16. Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C. & Balding, D. J. Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.* **32**, 179–185 (2008).
17. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).
18. Juran, B. D. *et al.* Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk variants. *Hum. Mol. Genet.* **21**, 5209–5221 (2012).
19. Duggal, P., Gillanders, E. M., Holmes, T. N. & Bailey-Wilson, J. E. Establishing an adjusted *p*-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* **9**, 516 (2008).
20. Nyholt, D. R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.* **74**, 765–769 (2004).
21. Galwey, N. W. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genet. Epidemiol.* **33**, 559–568 (2009).
22. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221–227 (2005).
23. Moskvina, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.* **32**, 567–573 (2008).
24. Li, M. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective number of independent tests and significant *p*-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).
25. North, B. V., Curtis, D. & Sham, P. C. A note on the calculation of empirical *P* values from Monte Carlo procedures. *Am. J. Hum. Genet.* **71**, 439–441 (2002).
26. North, B. V., Curtis, D. & Sham, P. C. A note on calculation of empirical *P* values from Monte Carlo procedure. *Am. J. Hum. Genet.* **72**, 498–499 (2003).
27. Dudbridge, F. & Koeleman, B. P. C. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am. J. Hum. Genet.* **75**, 424–435 (2004).
28. Seaman, S. R. & Müller-Myhsok, B. Rapid simulation of *P* values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.* **76**, 399–408 (2005).
29. Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.* **96**, 434–442 (2004).
30. Panagiotou, O. A., Ioannidis, J. P. & Genome-Wide Significance Project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* **41**, 273–286 (2011).
31. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
32. Wakefield, J. Bayes factors for genome-wide association studies: comparison with *P*-values. *Genet. Epidemiol.* **33**, 79–86 (2009).

33. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
This paper summarizes and interprets GWAS findings on common diseases and quantitative traits.
34. Pawitan, Y., Seng, K. C. & Magnusson, P. K. E. How many genetic variants remain to be discovered? *PLoS ONE* **4**, e7969 (2009).
35. 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).
36. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
37. Zhong, H. & Prentice, R. L. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621–634 (2008).
38. Ghosh, A., Zou, F. & Wright, F. A. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am. J. Hum. Genet.* **82**, 1064–1074 (2008).
39. Zollner, S. & Pritchard, J. K. Overcoming the winner's curse: estimating penetrance parameters from case–control data. *Am. J. Hum. Genet.* **80**, 605–615 (2007).
40. Sham, P. C., Cherny, S. S., Purcell, S. & Hewitt, J. K. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616–1630 (2000).
41. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Including known covariates can reduce power to detect genetic effects in case–control studies. *Nature Genet.* **44**, 848–851 (2012).
42. Li, Q., Zheng, G., Li, Z. & Yu, K. Efficient approximation of *P*-value of the maximum of correlated tests, with applications to genome-wide association studies. *Ann. Hum. Genet.* **72**, 397–406 (2008).
43. González, J. R. *et al.* Maximizing association statistics over genetic models. *Genet. Epidemiol.* **32**, 246–254 (2008).
44. So, H.-C. & Sham, P. C. Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates. *Behav. Genet.* **41**, 768–775 (2011).
45. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Rev. Genet.* **12**, 745–755 (2011).
46. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nature Genet.* **44**, 623–630 (2012).
47. 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).
48. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
49. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
50. Li, B. & Leal, S. M. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet.* **5**, e1000481 (2009).
51. Liu, D. J. & Leal, S. M. Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am. J. Hum. Genet.* **87**, 790–801 (2010).
52. Li, M. X., Gui, H. S., Kwan, J. S. H., Bao, S. Y. & Sham, P. C. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res.* **40**, e53 (2012).
53. Ng, S. B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nature Genet.* **42**, 790–793 (2010).
54. Zhi, D. & Chen, R. Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing. *PLoS ONE* **7**, e31358 (2012).
55. Feng, B.-J., Tavtigian, S. V., Southey, M. C. & Goldgar, D. E. Design considerations for massively parallel sequencing studies of complex human disease. *PLoS ONE* **6**, e23221 (2011).
56. 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).
This is one of the first association tests for rare variants.
57. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
58. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 982 (2010).
59. Lin, D.-Y. & Tang, Z.-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89**, 354–367 (2011).
60. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nature Rev. Genet.* **11**, 773–785 (2010).
61. Stitziel, N. O., Kiezun, A. & Sunyaev, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* **12**, 227 (2011).
62. Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* **35**, 606–619 (2011).
63. Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M. T. & Richards, J. B. The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLoS Genet.* **8**, e1002496 (2012).
64. Ladouceur, M., Zheng, H.-F., Greenwood, C. M. T. & Richards, J. B. Empirical power of very rare variants for common traits and disease: results from Sanger sequencing 1998 individuals. *Eur. J. Hum. Genet.* **21**, 1027–1030 (2013).
65. Saad, M., Pierre, A. S., Bohossian, N., Macé, M. & Martinez, M. Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data. *BMC Proc.* **5**, S33 (2011).
66. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322 (2011).
67. Wu, Michael, C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
This is the original paper that describes the SKAT for rare-variant association.
68. Liu, L. *et al.* Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genet.* **9**, e1003443 (2013).
69. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2013).
This paper presents a framework for power calculation and ways to improve power for rare-variant studies.
70. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
71. Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E. & Conti, D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genet. Epidemiol.* **35**, 790–799 (2011).
72. 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).
73. Bailey-Wilson, J. E. & Wilson, A. F. Linkage analysis in the next-generation sequencing era. *Hum. Hered.* **72**, 228–236 (2011).
74. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur. J. Hum. Genet.* **21**, 1158–1162 (2013).
75. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
76. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
77. Lim, Elaine, T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235–242 (2013).
78. Longmate, J. A., Larson, G. P., Krontiris, T. G. & Sommer, S. S. Three ways of combining genotyping and resequencing in case–control association studies. *PLoS ONE* **5**, e14318 (2010).
79. Aschard, H. *et al.* Combining effects from rare and common genetic variants in an exome-wide association study of sequence data. *BMC Proc.* **5**, S44 (2011).
80. He, X. *et al.* Integrated model of *de novo* and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).
81. Ye, K. Q. & Engelman, C. D. Detecting multiple causal rare variants in exome sequence data. *Genet. Epidemiol.* **35**, S18–S21 (2011).
82. Li, B., Wang, G. & Leal, S. M. SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. *Bioinformatics* **28**, 2703–2704 (2012).
83. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genet.* **44**, 243–246 (2012).
84. Lee, S., Teslovich, Tanya, M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
85. Hu, Y.-J. *et al.* Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *Am. J. Hum. Genet.* **93**, 236–248 (2013).
References 83 and 84 propose powerful and convenient score tests for meta-analyses of rare-variant association studies.
86. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
This paper describes the SKAT power calculation tool.
87. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* **204**, 108–114 (2013).
88. Patnaik, P. B. The power function of the test for the difference between two proportions in a 2 × 2 table. *Biometrika* **35**, 157 (1948).
89. Sidak, Z. Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Statist. Associ.* **62**, 626 (1967).
90. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Application* (Cambridge Univ. Press, 1997).
91. Patnaik, P. B. The non-central χ^2 - and F-distribution and their applications. *Biometrika* **36**, 202 (1949).
92. Whittaker, J. C. & Lewis, C. M. Power comparisons of the transmission/disequilibrium test and sib–transmission/disequilibrium-test statistics. *Am. J. Hum. Genet.* **65**, 578–580 (1999).
93. Fulker, D. W., Cherny, S. S., Sham, P. C. & Hewitt, J. K. Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259–267 (1999).
94. Kwan, J. S. H., Cherny, S. S., Kung, A. W. C. & Sham, P. C. Novel sib pair selection strategy increases power in quantitative association analysis. *Behav. Genet.* **39**, 571–579 (2009).
95. Luan, J. Sample size determination for studies of gene–environment interaction. *Int. J. Epidemiol.* **30**, 1035–1040 (2001).
96. Gauderman, W. J. Sample size requirements for association studies of gene–gene interaction. *Am. J. Epidemiol.* **155**, 478–484 (2002).
97. Gauderman, W. J. Sample size requirements for matched case–control studies of gene–environment interaction. *Statist. Med.* **21**, 35–50 (2002).

Acknowledgements

This work was supported by The University of Hong Kong Strategic Research Theme on Genomics; Hong Kong Research Grants Council (HKRGC) General Research Funds 777511M, 776412M and 776513M; HKRGC Theme-Based Research Scheme T12-705/11 and T12-708/12-N; and the European Community Seventh Framework Programme Grant on European Network of National Schizophrenia Networks Studying Gene–Environment Interactions (EU-GEI); and the US National Institutes of Health grants R01 MH099126 and R01 HG005827 (to S.M.P.). The authors thank R. Porsch and S.-W. Choi for technical assistance with the manuscript.

Competing interests statement

The authors declare no competing interests.

FURTHER INFORMATION

1000 Genomes Project: http://www.1000genomes.org
Case–control for discrete traits: http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html
CaTS: http://www.sph.umich.edu/csg/abecasis/CaTS
Exome Chip: http://genome.sph.umich.edu/wiki/Exome_Chip_Design
Exome Power Calculation: http://exomepower.ssg.uab.edu
Exome Variant Server: http://evs.gs.washington.edu/EVS
GPC: http://pngu.mgh.harvard.edu/purcell/gpc
GWA Test Driver: http://gwatestdriver.ssg.uab.edu
PAWE and PAWE-3D: http://www.jurgott.org/linkage/home.html
PolyPhen2: http://genetics.bwh.harvard.edu/pph2/
Power for Genetic Association Analyses: http://dceg.cancer.gov/tools/analysis/pga
SEQPower: http://www.bioinformatics.org/spower/start
SKAT: http://www.hsph.harvard.edu/skat/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF