Am. J. Hum. Genet. 75:424–435, 2004

Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genomewide Association Studies

Frank Dudbridge[1,2] and Bobby P. C. Koeleman[3]

[1]MRC Rosalind Franklin Centre for Genomics Research, and [2]MRC Biostatistics Unit, Cambridge, United Kingdom, and [3]Department of Medical Genetics, University Medical Centre Utrecht, Utrecht

Large exploratory studies, including candidate-gene–association testing, genomewide linkage-disequilibrium scans, and array-expression experiments, are becoming increasingly common. A serious problem for such studies is that statistical power is compromised by the need to control the false-positive rate for a large family of tests. Because multiple true associations are anticipated, methods have been proposed that combine evidence from the most significant tests, as a more powerful alternative to individually adjusted tests. The practical application of these methods is currently limited by a reliance on permutation testing to account for the correlated nature of single-nucleotide polymorphism (SNP)–association data. On a genomewide scale, this is both very time-consuming and impractical for repeated explorations with standard marker panels. Here, we alleviate these problems by fitting analytic distributions to the empirical distribution of combined evidence. We fit extreme-value distributions for fixed lengths of combined evidence and a beta distribution for the most significant length. An initial phase of permutation sampling is required to fit these distributions, but it can be completed more quickly than a simple permutation test and need be done only once for each panel of tests, after which the fitted parameters give a reusable calibration of the panel. Our approach is also a more efficient alternative to a standard permutation test. We demonstrate the accuracy of our approach and compare its efficiency with that of permutation tests on genomewide SNP data released by the International HapMap Consortium. The estimation of analytic distributions for combined evidence will allow these powerful methods to be applied more widely in large exploratory studies.

Introduction

The rapid increase in genomic data available to researchers, coupled with sharply decreasing experimental costs, has created opportunities for exploratory genetic studies of unprecedented size. For example, it is now possible to screen thousands of candidate genes for disease associations at either the sequence or the expression level (Risch 2000; Schulze and Downward 2001). Also, the prospect of a genomewide linkage-disequilibrium map will allow genome scans for association to become as routine as they already are for linkage (International HapMap Consortium 2003). Several initial scans have recently been completed (Ophoff et al. 2002; Ozaki et al. 2002; Sawcer et al. 2002).

A serious problem for these studies is that a large number of nominally significant results are expected, even when there is no true association, because of stochastic variation in the data and the large number of tests that are performed. Furthermore, one cannot be certain at the outset whether there are any true associations to be found at all. For this reason, strict significance thresholds have been recommended that control the familywise type I error (Risch and Merikangas 1996). At such low error rates, very large samples must be obtained to achieve adequate power, making the cost of ascertainment an important factor in study design (Service et al. 2003).

Although the traditional burden is to reject the hypothesis that there are no true associations at all, in practice, some—perhaps many—true associations are anticipated. Furthermore, it can be more economical to perform an initial screening stage with weaker error control to reduce the candidate loci to a smaller set to which stronger error control can be applied (Brown and Russell 1997). For these reasons, attention has turned to methods that are sensitive to multiple associations while retaining weak control of the familywise error (Hoh and Ott 2003; Storey and Tibshirani 2003). One approach that has shown promise is to combine the strongest evidence from multiple tests. Motivated by principles of meta-analysis, the idea is to identify a subset of tests showing a trend of significance exceeding that of the

Received May 5, 2004; accepted for publication June 24, 2004; electronically published July 19, 2004.

Address for correspondence and reprints: Dr. Frank Dudbridge, MRC Biostatistics Unit, Robinson Way, Cambridge CB2 2SR, United Kingdom. E-mail: frank.dudbridge@mrc-bsu.cam.ac.uk

0002-9297/2004/7503-0008$15.00

individual tests. Hoh et al. (2001) proposed forming sums of the k largest test statistics, comparing the sums with their null distributions, and identifying the most significant sum. Zaykin et al. (2002) proposed forming the product of all P values less than a fixed threshold. Dudbridge and Koeleman (2003) suggested a hybrid of these methods, namely, the product of the k smallest P values.

Each of these methods has been shown to give improved power in realistic situations (see also Wille et al. [2003] and Hao et al. [2004]). However, the lack of analytic distributions is a major obstacle to their more widespread use. Although distributions for products of P values are known when the tests are independent (Zaykin 1999; Zaykin et al. 2002; Dudbridge and Koeleman 2003), they do not apply to correlated tests, which arise in genomewide association studies, particularly when dense maps of SNPs are used. We will show that, unlike for Bonferroni-like procedures, increased type I errors can result from incorrectly assuming independence. Monte Carlo procedures are often recommended (Zaykin et al. 2002)—for example, by random permutation of phenotypic labels (Churchill and Doerge 1994)—but can become extremely time-consuming when applied on such a large scale. Furthermore, follow-up analysis may require further levels of permutation—for example, when identifying subsets of the data showing increased association (Johnson et al. 2002).

With the advent of large-cohort studies (Austin et al. 2003) and the prospect of common marker panels for genomewide association scans (International HapMap Consortium 2003), reusability becomes an important issue. If several investigators are conducting the same tests on different samples, it is inefficient for each investigator to generate the same permutation distribution. Yet, if a single supplier generates and distributes the permutation distribution, it is useful only if critical values are supplied for every combination of tests and every significance level. The volume of information involved, reminiscent of traditional statistical tables, makes this model unlikely to be adopted.

Instead of relying on permutation tests, some methods have been suggested to adjust the tests in a way that allows independence to be assumed. One approach assumes an effective number of independent tests, such that the actual tests are an oversampling of some underlying independent variables. The effective number could be estimated by bootstrap resampling (Bailey and Grundy 1999) or by appealing to principal-components analysis (Cheverud 2001; Nyholt 2004). This approach is a restricted case of the modeling we propose below, and we will show that the restriction does not sufficiently capture the full correlation structure, except in some particular cases that are unlikely to occur on the genomewide scale.

Another approach is to sequentially decorrelate the tests. This can be done by application of a single transformation derived from the correlation matrix (Zaykin et al. 2002) or by successive greedy transformations (Wille et al. 2003). The former approach is sensitive to the ordering of the tests, whereas the latter may lose some advantages of combining evidence because it favors the tests with stronger marginal significance. Alternatively, a step-up linear model may be constructed, starting from a single variable and adding covariates one at a time (Cordell and Clayton 2002). However, these approaches all encounter difficult computational problems as the number of tests becomes large. For example, methods using a correlation matrix require inversion of large sparse matrices, and linear modeling requires likelihood maximization over many covariates, with an aggregation of missing data as more covariates are added.

Here, we propose a more efficient application of permutation sampling, in which analytic distributions are fitted to the permutation samples. We apply the method to the sum of k largest $-\log P$ values from a larger number of tests. Assuming that k is fixed and relatively small, we fit extreme-value distributions to the empirical distribution of the sums. When many values of k are considered, we fit a beta distribution to the most significant sum. Although permutation sampling is required to generate the empirical distributions, this can be completed more quickly than a simple permutation test. Conditional on the correlation structure, the procedure need be performed only once and subsequently allows fast and accurate computation of significance levels for sum statistics. The correlation structure is a property of the study population and the ascertainment criteria, so the reusability is most applicable to prospective cohort studies, although it can also be relevant to retrospective samples. As a special case, our procedure gives a more efficient method for a standard permutation test.

We illustrate the method on chromosomewide SNP genotypes released by the International HapMap Consortium (2003). We compare the efficiency of our approach with simple permutation tests, demonstrate the effect of assuming independence, and study some operating characteristics of the combined-evidence approach.

Material and Methods

Fixed Number of Combined Variables

Suppose a study generates statistics for m hypothesis tests. Denote the P values by P_i and their order statistics by $P_{(i)}$. Let k be an integer, $1 \leqslant k \leqslant m$, termed the "length." We consider the partial sum of $-\log P$ values,

equivalent to a truncated product of P values (Dudbridge and Koeleman 2003)

$$S_k = -\sum_{i=1}^{k} \log P_{(i)} .$$

This is an example of a sum statistic, suggested by Hoh et al. (2001). Those authors used χ^2 statistics rather than P values, which should have similar power, but P values should offer greater flexibility in the context of genome scans, as we suggest in the "Discussion" section.

The partial sum S_k can be regarded as the maximum of all sums of length k, which suggests an application of extreme-value theory. For a large set of independent identically distributed variables, the minimum and maximum have distributions of a general form for a wide class of parent distributions (Gumbel 1958). Here, the sums are proportional to χ^2 variables (Fisher 1932). This holds for correlated variables, provided that they can be regarded as a stationary process (one whose stochastic properties are invariant) with independence in the limit of large distances (Coles 2001). The rapid decay of linkage disequilibrium (LD) ensures long-range independence (Reich et al. 2001); we do not investigate stationarity, but we argue that the model is empirically accurate for genomic data. The theory has been applied elsewhere to molecular sequence analysis (Karlin and Altschul 1990) and to microarray data (Li and Grosse 2003).

When m is large and $k \ll m$, the extremal types theorem predicts that S_k follows an extreme-value distribution (for background, see Coles [2001]). We fit the generalized extreme-value distribution, which has location, scale, and shape parameters. For our purposes, these parameters are just mathematical variables that permit model fitting, but a heuristic interpretation can be made by observing their behavior when fitted to sums of independent $-\log P$ values. The location is, roughly, proportional to the total number of tests m, conditional on the sum length k. The scale is proportional to k, conditional on m, and the shape summarizes the correlation structure among the single tests and the sums of fixed length.

To fit the distribution of the sum statistics, we generate a large number of permutation replicates and fit the distribution by maximum likelihood. In general, the replicates are obtained by a nonparametric combination of dependent permutation tests (Pesarin 2001). In association studies, this can take the well-known form of shuffling trait values among subjects while keeping genotypes fixed. If $s_k^{(i)}$ is the value of S_k in the ith replicate, the log likelihood is

$$l(\mu,\sigma,\xi) = \sum_i \left\{ -\log\sigma - (1 + 1/\xi)\log[1 + \xi(s_k^{(i)} - \mu)\sigma^{-1}] - [1 + \xi(s_k^{(i)} - \mu)\sigma^{-1}]^{-1/\xi} \right\}$$

for location μ, scale σ, and shape ξ. Maximum-likelihood estimates are obtained by numerical optimization; we used the EVD add-on package to R (r Development Core Team 2003; Stephenson 2002). The significance level of the observed data is then calculated from the fitted analytic distribution. In contrast to a standard permutation test, we use the actual values of the replicate statistics to calculate the significance, whereas a permutation test counts only how often they exceed the value in the observed data. By using more information from the permutation replicates, we expect to achieve greater accuracy, in addition to parameterizing the permutation distribution with an analytic model.

Variable Number of Combined Variables

The above procedure applies to a fixed-size subset of the variables. For best power, the length k might be chosen to somewhat exceed the anticipated number of true associations. When this is unknown, one can form sums for many lengths and identify which length has the most significant sum (Hoh et al. 2001). To obtain the overall significance, we need the distribution of the smallest P value of the sums of varying length. We propose a beta distribution, motivated by the fact that order statistics of independent uniform (0,1) variables have known beta distributions. The P values of the sums are highly correlated, but we assume that the smallest value follows a two-parameter beta distribution. We generate permutation replicates and, within each replicate, form the partial sums for a range of lengths and calculate their significances from the previously fitted extreme-value distributions. The smallest of those P values is then the statistic for that replicate. If $p_{\min}^{(i)}$ is the smallest P value in the ith replicate, the log likelihood is

$$l(\alpha,\beta) = \sum_i \left\{ \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + (\alpha - 1)\log p_{\min}^{(i)} + (\beta - 1)log(1 - p_{\min}^{(i)}) \right\}$$

for shape parameters α and β. Maximum-likelihood es-

timates are obtained by numerical optimization, and the significance level of the observed data is obtained from the fitted distribution.

Effective Number of Tests

We wish to test the assertion that there is an effective number of independent tests that could be used in traditional adjustments (Bailey and Grundy 1999; Cheverud 2001; Nyholt 2004). This can be done by fitting beta distributions, as above. If there is an effective number of tests, a real number m' can be found, such that the minimum P value follows the Šidák (1967) correction $1-(1-p_{\min})^{m'}$, which is the beta $(1,m')$ distribution. As the alternative, we allow the minimum P value to follow the two-parameter beta distribution, and we test whether the first parameter is 1 by applying the likelihood-ratio test for nested models. A significant likelihood ratio rejects the hypothesis that one parameter is sufficient and that there is an effective number of independent tests.

Power of Variable Number of Combined Variables

The choice of the sum length k can be problematic. The optimal length depends on the number of true associations and their effect sizes, both of which are unknown. Often, a reasonable prior estimate can be made. For example, genome scans are currently designed under an assumption of no more than, say, 20 true associations, since models with more loci specify effect sizes so small that prohibitive sample sizes are required (Pritchard 2001; Schliekelman and Slatkin 2002). A relevant question is whether misspecification of a fixed length has a greater cost than does estimation of the most significant length.

We studied this by simulating a large exploratory study consisting of 10,000 independent tests. For true associations, we simulated $\chi^2_{(1)}$ variables with the noncentrality parameter (NCP) chosen to give maximum power of ~80% (at $\alpha=0.05$) over the range of sum lengths $1 \leqslant k \leqslant 500$. For the remainder, we generated uniform P values on (0,1). This approach avoids making assumptions about genetic effects and LD, but it assumes that sufficiently powerful studies can be designed. We first generated 10,000 replicates of the null distribution with no true associations, and we fitted extreme-value distributions to the fixed length sums and a beta (0.767,1.939) distribution to the smallest P value of the sums. We then simulated studies with 5, 10, 50, and 100 true associations, with 10,000 replicates for each, and plotted the power for each fixed-length sum together with the power for variable length.

Comparison with Permutation Test

Since permutation sampling is required to fit the extreme-value and beta distributions, we compare efficiency with standard permutation tests. Here, a permutation test consists of randomly assigning traits among subjects, while keeping the genotype data fixed. Sum statistics are calculated in each replicate and compared with the statistic in the original data. The significance level is the proportion of replicate statistics exceeding the observed statistic.

We compare efficiency for a fixed P value by considering how many replicates are needed to achieve a given accuracy, measured by the length of the 95% CI for P. Since a permutation test is equivalent to an estimation of a binomial probability, the CI is given by normal theory. For our approach, we obtain the CI by a parametric bootstrap. We generate a number of random deviates from a fixed extreme value or beta distribution, each deviate representing one permutation sample. We refit a distribution to the deviates and from it calculate the significance of the P-quantile point of the generating distribution. The CI is estimated by repeating this procedure a large number of times. From this, we obtain the number of binomial trials needed to achieve the same accuracy, given as

$$\frac{4Z^2_{0.975}P(1-P)}{(CI)^2},$$

where $Z_{0.975}=1.96$ is the 97.5th percentile point of the standard normal distribution and CI is the length of the bootstrap CI for our procedure.

Data Sets

We applied these methods to chromosomewide genotypes from release 6 of the International HapMap Consortium (2003). We obtained genotypes for 20,243 SNPs on chromosome 18 and 13,620 SNPs on chromosome 21, which were the two most densely genotyped chromosomes. Genotypes were available for 90 subjects from CEPH pedigrees. For the proof of concept, we tested each SNP individually, without performing any blocking or grouping of SNPs, although such grouping is likely to occur in real scans. We regarded the subjects as unrelated, for the purpose of generating null distributions. This does not bias the distribution of P values, which is our main interest, but does increase the correlation between loci, since the length of shared haplotypes is greater than in a population of unrelated subjects. Therefore, these data represent a more problematic case for correlated SNP data than would usually be the case but could also be regarded as a reduced-scale copy of whole-genome data.

For both chromosomes, we generated empirical null distributions for the sum statistics by randomly assigning "case" status to half of the subjects and "control" status to the other half. Most software packages for genetic association have severe memory and time requirements when working with this volume of data. We wrote a custom program using compact data structures, to calculate likelihood-ratio tests of allelic association, form sum statistics, and permute affection status. The program is available from the authors on request. We performed 10,000 random permutations and fitted extreme-value distributions to the partial sums for $1 \leqslant k \leqslant 100$ and a beta distribution to the smallest P value of the sum over that range of k.

We were interested in the relationship between the parameters of the fitted distributions and the parameters of the sum statistics. To explore this, we plotted the location, shape, and scale parameters of the fitted distribution against the sum length. We used quantile-quantile plots to show that the analytic distributions gave a good fit to the empirical distributions. We used the HapMap data to study the effect of assuming independence when the data are correlated, by applying the exact distribution for independent tests to the 95th percentile of the empirical distribution. If the exact distribution is accurate, it will give a P value of .05 for each sum length. If it is conservative, the P value will be >.05; if it is liberal, the P value will be <.05.

Results

Accuracy of Fitted Distributions

In figure 1, we show the location, scale, and shape parameters of the fitted distribution plotted against the length of the sum. The location and scale parameters have a strong log-linear correlation with the length, deviating from unity only by sampling variation. The relationship is less clear for the shape, but it does show a continuous form. Therefore, the sum statistics can be modeled as a whole by a smooth family of extreme-value distributions, for which the location and scale can be parameterized by the slope and intercept of fitted lines. For the best accuracy, the shape should be specified explicitly, although a polynomial spline might provide an acceptable approximation. The distributions of the sum statistics for the chromosome-18 SNPs can be summarized as follows: Fixed length sum has the generalized extreme-value distribution. Location $= 9.526 \times k^{0.898}$. Scale $= 1.269 \times k^{0.8205}$. Shape $= \{-0.0453$ for $k = 1, -0.0521$ for $k = 2, -0.0524$ for $k = 3, \ldots\}$.

Figure 2 shows quantile-quantile plots for the empirical and fitted distributions for lengths 10 and 100. The extreme-value distribution gives a good fit, although there is a liberal deviation in the far tail. The fit is adequate for declaring chromosomewide significance at conventional levels. We can obtain a more accurate fit by noting that the distribution for independent variables is constructed from beta and gamma distributions (Dudbridge and Koeleman 2003) and therefore fitting a similar construction to the correlated variables. However, this turns out to be very computationally intensive and numerically unstable; furthermore, it applies only to sums based on P values and not to those of χ^2 statistics, whereas the extreme-value distribution is more generally applicable and accurate for our present purposes.

Figure 3 shows quantile-quantile plots of the empirical and fitted distributions of the smallest P value for $1 \leqslant k \leqslant 100$. The beta distribution gives a good fit over the whole range, confirming that it is an appropriate model for the smallest P value. The linear correlation coefficients for these plots were .9995 and .9997. For chromosome 18, the fitted parameters of the beta distribution were (.8032,1.377), and, for chromosome 21, the parameters were (.7932,1.342).

Effect of Ignoring Correlation

We studied the effect of assuming independence when the tests are correlated. Although this assumption is conservative for the Bonferroni adjustment, it is not generally true for sum statistics. Figure 4 shows the significance of the 95th percentile of the permutation distribution for the chromosome-18 data, by use of the exact distribution under the assumption of independence. The sum of length 1 is less significant than the nominal level, which is consistent with the conservative property of the Bonferroni adjustment. However, the longer-length sums show an increasingly liberal bias as the sum length increases. In contrast, the analytic significance of the fitted distribution is close to the nominal level across the range of lengths. From these data, we can conclude that assuming independence for sum statistics of length >1 can lead to liberal as well as conservative tests, in a manner that is dependent on the sum length, significance level, and number of tests. It is therefore important to have accurate distributions available for the sum statistics.

Effective Number of Tests

We tested whether an effective number of independent tests sufficiently describes the correlation structure. We compared the maximum likelihood of the minimum P value under the two-parameter beta distribution with the likelihood with the first parameter set to 1 (see the "Material and Methods" section). In both data sets, the likelihood ratio was highly significant ($\chi^2 = 360$ and 322, respectively, on 1 df). Furthermore, the two-parameter beta distribution gave an excellent fit (data not shown). Therefore, we reject the hypothesis that there is an effective number of tests that allows the use of

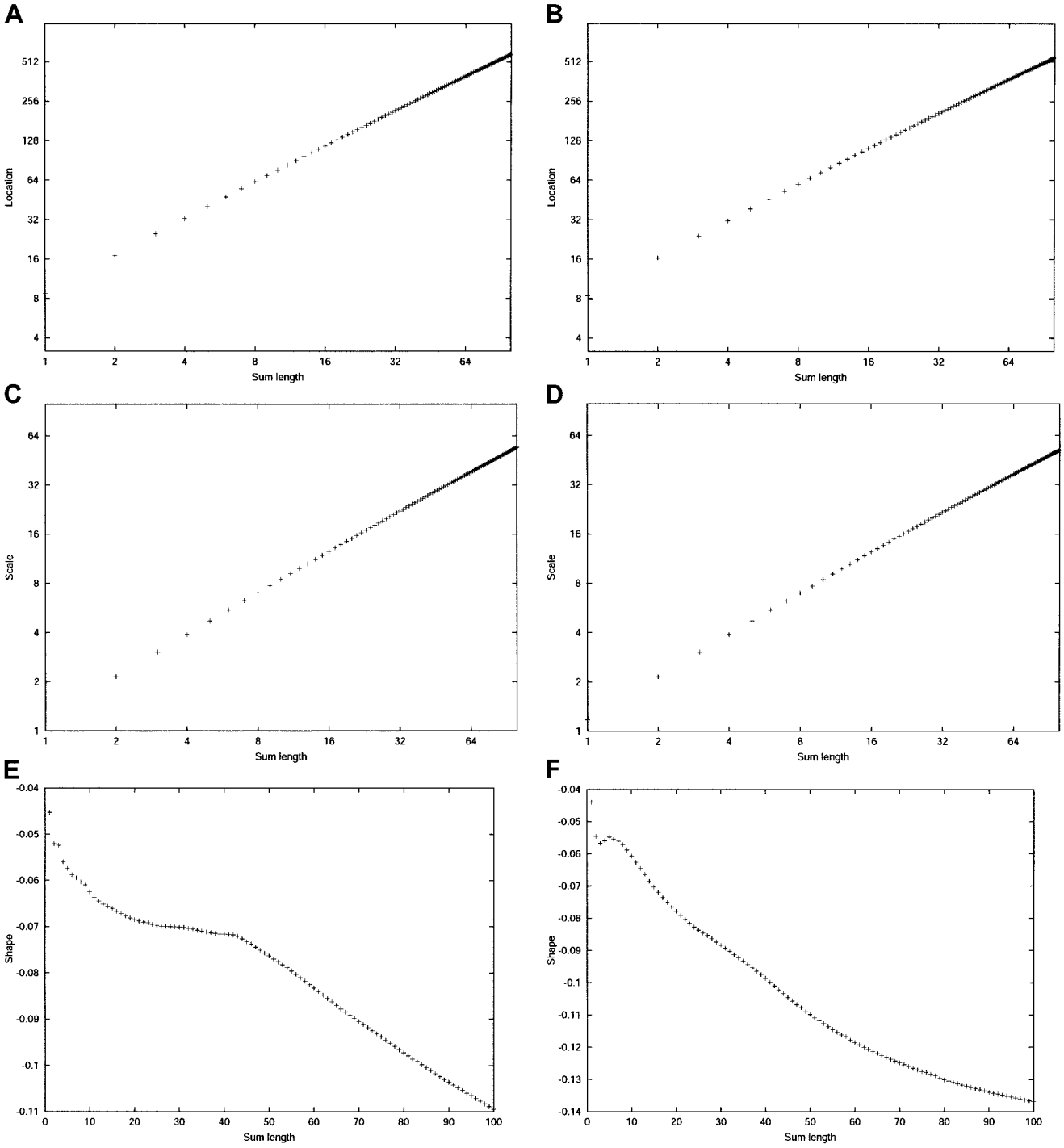

Figure 1 Parameters of the extreme-value distribution for S_k as function of length k. *A*, Location, chromosome 18. *B*, Location, chromosome 21. *C*, Scale, chromosome 18. *D*, Scale, chromosome 21. *E*, Shape, chromosome 18. *F*, Shape, chromosome 21.

traditional corrections. We can interpret this by regarding the minimum P value to be the sum of length $k = 1$ when the tests are independent and concluding that the correlation structure implies an effective sum length as well as an effective number of tests.

Power of Variable Number of Combined Variables

In figure 5, we compare the power that is the result of optimization of the sum length for significance with the power when the sum length is fixed. This shows that

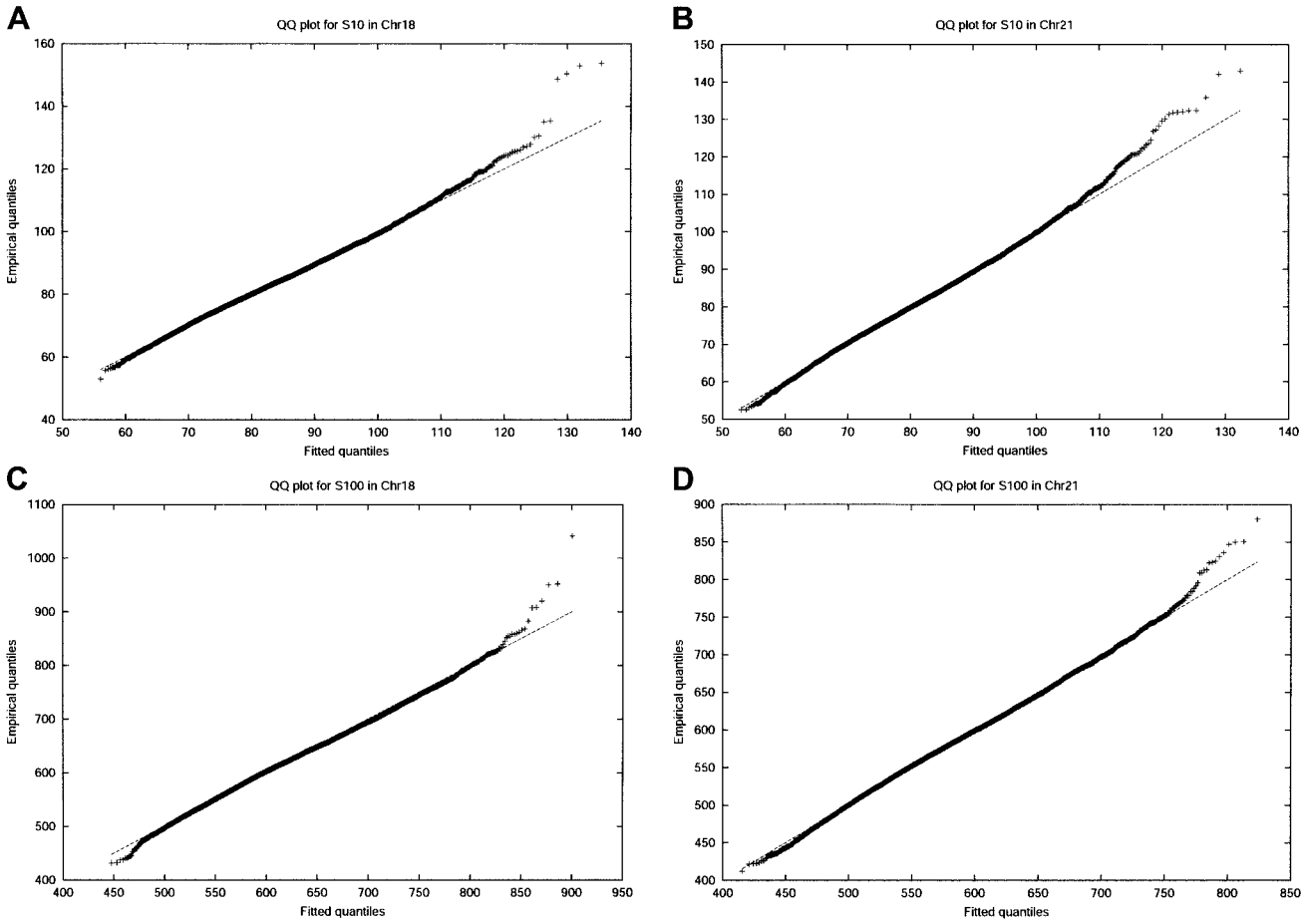

Figure 2 Quantile-quantile plot of extreme-value distribution for S_k. *A*, Length 10, chromosome 18. *B*, Length 10, chromosome 21. *C*, Length 100, chromosome 18. *D*, Length 100, chromosome 21.

very little loss of power results from varying the sum length. For 5, 10, and 50 true associations, the difference in power is only 2%–3%, and it would be even smaller if a shorter range of lengths were considered. The difference is ~10% for 100 true associations, because the most powerful length is outside the range considered, because of the small NCP of the true associations. Note, however, that this scenario is unlikely in genome-association scans. On the other hand, fixed lengths do have higher power over a range of values, and there is some margin for error with the use of fixed lengths, but, outside the optimal range, the power of fixed lengths drops off sharply.

Although there is little cost in overall power, varying the length does not give a reliable estimate of the number of true associations. In each of the situations we considered, the range of lengths for which the minimum *P* value was achieved varied over the whole range of 1–500. The median lengths of the most significant sum, for the four situations in figure 5, in the order shown, were 4, 7, 51, and 133, which is fairly accurate, but the interquartile ranges were 6, 17, 143, and 432, and the mean lengths were 25, 37, 129, and 219, indicating both a high variability and upward bias. Therefore, varying the length is a useful way to improve power when there is very little idea of the number of true associations, but it should not be relied on to estimate the number of true associations, in particular, to determine the number of follow-up loci.

Comparison with Permutation Test

We compared efficiency with a permutation test for the sum of length 10 in the chromosome-18 data. The extreme-value distribution fitted to this sum was used to generate a parametric bootstrap CI for our method, which was matched to the normal theory CI for the binomial permutation test. Table 1 shows the number of binomial replicates needed to achieve the same accuracy as with our method with 1,000 and 10,000 rep-

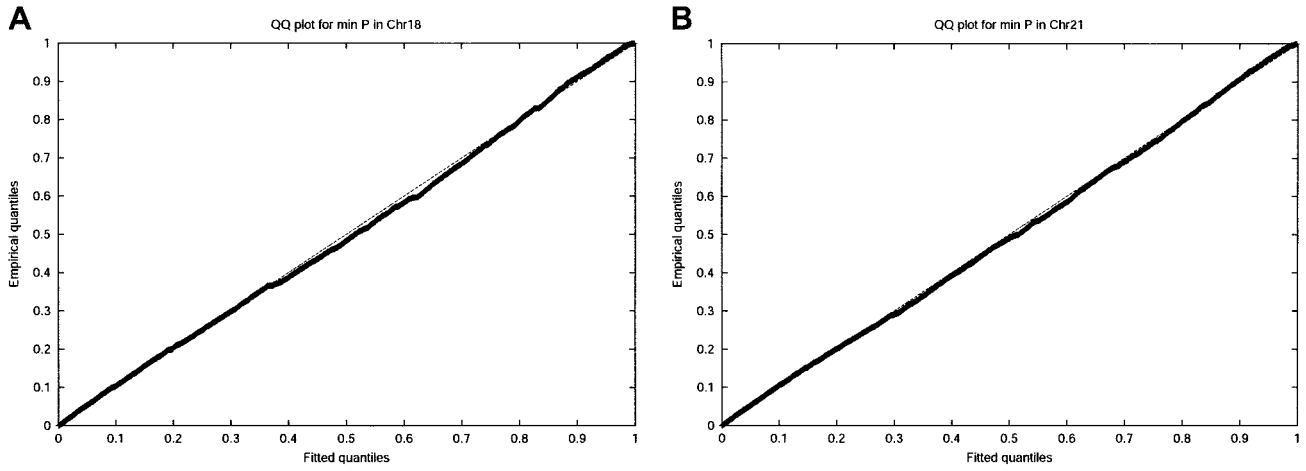

Figure 3 Quantile-quantile plot of beta distribution for the minimum P value for S_k. *A,* Chromosome 18, parameters (.8032,1.3766). *B,* Chromosome 21, parameters (.7932,1.3423).

licates, for a range of significance levels. For example, at $P = .2$, the permutation test requires 1,511 replicates to reach the accuracy achieved by our method with 1,000 replicates. The efficiency is greater at stronger significance levels: at $P = .001$, the permutation test requires over three times as many replicates as does our method. At each significance level, the relative efficiency is the same for 1,000 and 10,000 deviates. Very similar results were obtained for other extreme-value distributions (data not shown). Table 2 gives the same comparison for the beta distribution fitted to the smallest P value of the sums. A similar pattern is seen, with the efficiency being even greater at strong significance levels. This is because two parameters are fitted rather than three, which results in smaller SEs. At $P = .001$, the permutation test requires >13 times as many replicates as does our method. Thus, even if we fit distributions to permutation replicates only once, our method provides an improvement in efficiency over standard permutation tests.

Discussion

The realization that complex heritable traits require large sample sizes to detect small effects has led to calls for more efficient methodology for detection of multiple associations (Hoh and Ott 2003). Combination of the strongest evidence is an approach that shows promise for identifying the pertinent effects while maintaining familywise error control. However, large-scale genomic studies encounter problems with correlated data that can be satisfactorily overcome only by permutation sampling. Because these procedures are time-consuming and might be duplicated by multiple investigators, we propose analytic distributions that can be fitted to the permutation distributions, which would give both a more efficient alternative to one-off permutation tests and a reusable calibration for repeated use.

Our results suggest that, for $P = .05$, a 40% reduction in computation is possible compared with a standard permutation test. This improvement in efficiency can translate to significant time savings. The results reported here were computed on recent Sun hardware with a 900 MHz UltraSparc III processor. We used a simple test of allelic association for single SNPs, using data on two of the shorter autosomes. Our software took ~5 h to compute 10,000 permutation replicates for the two chromosomes. A genomewide association scan may involve testing 100,000 haplotype blocks of 30 kb, with

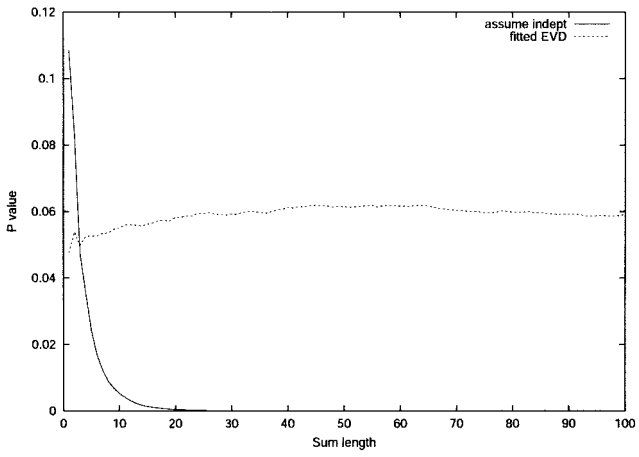

Figure 4 Effect of assumption of independence in correlated tests. Solid line shows the P value of the 95th percentile of the empirical distribution, under assumption of independent tests. Dotted line shows the P value according to the fitted extreme-value distribution.

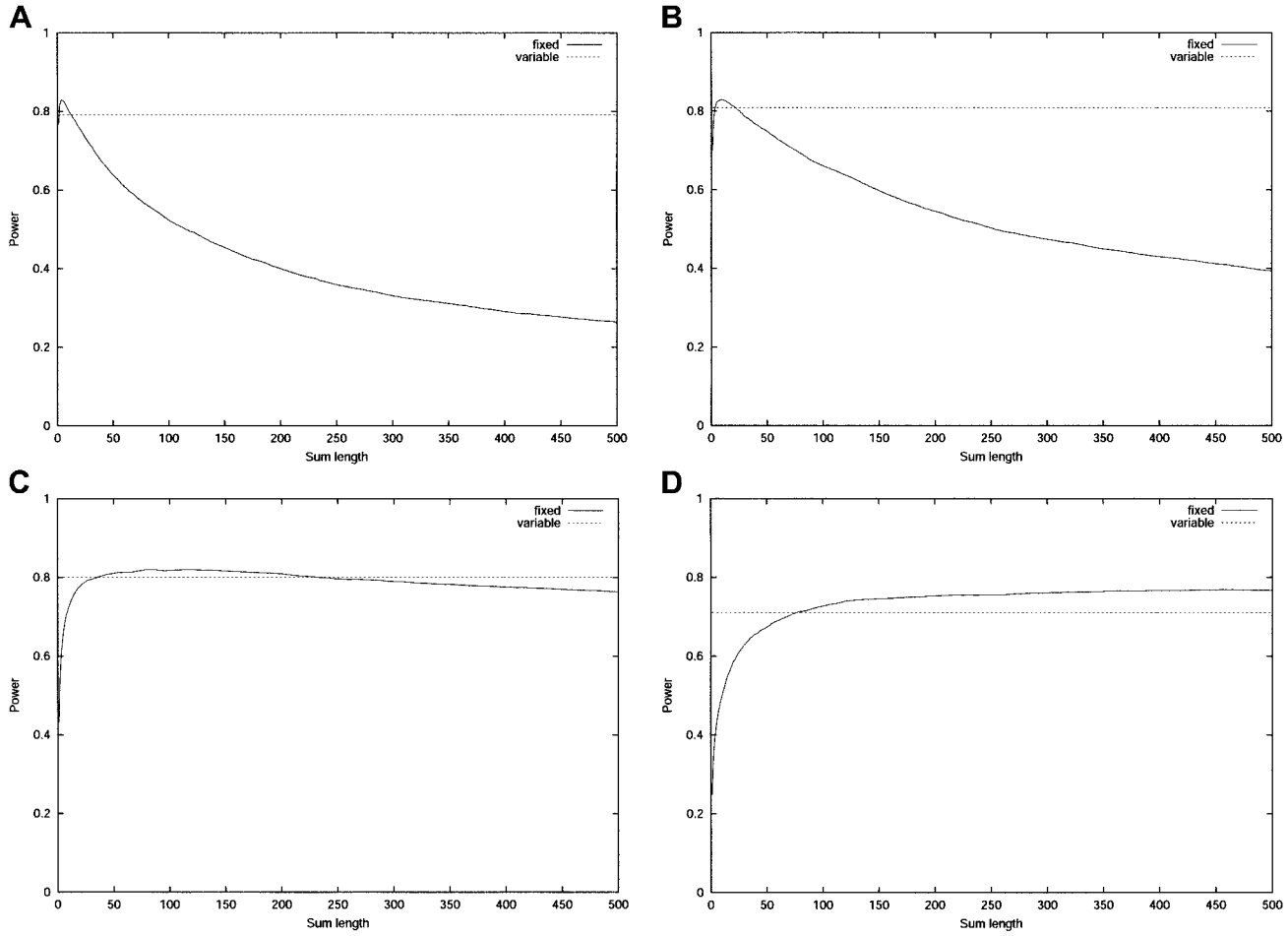

Figure 5 Power of fixed-length sum compared with variable-length sum, for 10,000 tests. *A,* Five true associations with χ^2 NCP 15. *B,* Ten associations with NCP 11. *C,* Fifty associations with NCP 5. *D,* One hundred associations with NCP 3.

multilocus tests conducted on each block. The number of tests is about three times as large, giving an extrapolated run time of 15 h, but multilocus tests require more computation, so that the total run time would be on the order of days. A 40% reduction can therefore have a significant impact. Subsequent tests of the same hypotheses in the same population require just a single analysis followed by reference to an analytic distribution, which takes only a few minutes. However there is a need for efficient software that can handle such large volumes of data. For example, COCAPHASE 2.4 (Dudbridge 2003) took ~4 d to perform the same calculations, whereas SUMSTAT (Hoh et al. 2001) could not, as supplied, input all the marker data. (In smaller data sets, its performance was similar to that of our own program.)

The parameters of the extreme-value and beta distributions can be fitted once and then distributed with other descriptive data for screening panels, such as map locations and allele frequencies. In principle, the supplier of the marker panel could calibrate these distributions to very high accuracy, allowing end users to rapidly calculate significance levels for single experiments. However, it is important to recognize that the calibration is conditional on the correlation structure. Different marker sets and different populations will have different structures, and different samples from the same population can also, in principle, exhibit different correlation. Thus, the reusability is most applicable to cohort studies, but markers can be calibrated for retrospective sampling, if the sample used for calibration is sufficiently large to minimize the variation in correlation structure. Furthermore, correlation can depend on ascertainment, because subjects that are selected according to some genetic features are expected to exhibit more LD in the vicinity of the relevant loci than the general population. It is therefore preferable to calibrate the null distributions by use of unselected subjects, but this may un-

Table 1

Efficiency of Extreme-Value Approximation Compared with Permutation Test

	No. of Binomial Replicates Needed to Achieve Accuracy of Current Study Method when			
	$n = 1{,}000$[a]		$n = 10{,}000$[a]	
P	Length of 95% CI[b]	No. of Binomial Trials with Same CI Length	Length of 95% CI[b]	No. of Binomial Trials with Same CI Length
.2	.0403	1,511	.01264	15,397
.1	.0292	1,623	.00918	16,408
.05	.0209	1,657	.006597	16,769
.01	.00894	1,902	.002813	19,220
.001	.00208	3,543	.000654	35,938

[a] Number of random deviates from extreme-value distribution.
[b] Length of 95% CI for the P value, estimated by parametric bootstrap.

derestimate the correlation around true associations, affecting power.

The extreme value and beta distributions give good fits for the HapMap data used here, but this property should not be automatically assumed. In particular, if the number of tests is not much greater than the sum length, then the extreme-value distribution may not be a good fit, though our experience is that it is a very robust model. Our approach should always be used in conjunction with quantile-quantile plots or goodness-of-fit tests to confirm model accuracy.

Our results show that the assumption that tests are independent can lead to inaccurate and biased results. If one ignores the correlation entirely, both liberal and conservative tests are possible, depending on underlying conditions that may not be understood. If the total number of tests is adjusted downward to an effective number and independence is then assumed, the resultant distribution can be significantly different from the true distribution, so this approach cannot be guaranteed to capture the full correlation structure. In the cases of complete dependence or complete independence, an effective number clearly applies, but we suggest that, in general, this approach is accurate only when the tests can be partitioned into sets with complete dependence within sets and complete independence between sets. This is extremely unlikely on the genomewide scale.

A different approach to detecting multiple associations aims to control the false-discovery rate (FDR) (Benjamini and Hochberg 1995). In some ways, this is complementary to the combined-evidence approach, a salient difference being that the FDR makes assertions about individual tests. Combined-evidence methods identify a candidate set for follow-up but give a significance level only for the whole set of tests. However, if there is an informal aim to proceed with the highest number of true associations and a sufficiently nondistracting number of false associations, then the combined-evidence approach identifies a good quality follow-up set (Wille et al. 2003; Hao et al. 2004). It is tempting to choose the follow-up set from the most significant sum length, but, although this appears to result in little net loss in power, it should not be relied on to predict the number of true associations, because the optimal sum length has high variance. Other methods that estimate the number of true associations from the distribution of P values

Table 2

Efficiency of Beta Approximation Compared with Permutation Test

	No. of Binomial Replicates Needed to Achieve Accuracy of Current Study Method when			
	$n = 1{,}000$[a]		$n = 10{,}000$[a]	
P	95% CI[b]	No. of Binomial Trials with Same CI Length	95% CI[b]	No. of Binomial Trials with Same CI Length
.2	.0397	1,562	.0128	15,032
.1	.0297	1,570	.00959	15,045
.05	.0201	1,809	.00650	17,034
.01	.00667	3,420	.002155	32,745
.001	.00106	13,538	.000340	132,778

[a] Number of random deviates from extreme-value distribution.
[b] Length of 95% CI for the P value, estimated by parametric bootstrap.

(Pounds and Morris 2003; Storey and Tibshirani 2003) also give highly variable estimates when the number of associations is small. It may be more prudent to select follow-up loci according to biological considerations or individually adjusted tests, conditional on a significant sum statistic.

Although FDR methods detect a high average number of associations (Benjamini and Hochberg 1995; Sabatti et al. 2003), the power to detect *at least one* association is not much greater than Bonferroni-like procedures (Simes 1986; Dudbridge and Koeleman 2003). This means that sample sizes for FDR analysis are not significantly smaller than for traditional methods, unless one compromises on the chances of obtaining any result at all. The increased power of sum statistics comes from the combination of evidence while error control for individual tests is forfeited. FDR methods are most applicable when many associations are present, with sufficiently strong effects that the prior power is high. In that situation, the false-discovery proportion has smaller variance, so investigators can be more confident that the target rate is achieved (authors' unpublished data). These conditions are more likely to be met in expression-array experiments than in linkage-disequilibrium scans.

Combining *P* values is a more balanced approach than summing the test statistics, because *P* values are usually identically distributed, whereas test statistics may not be. In particular, genomewide association scans will be designed around blocks of varying length and diversity (Weale et al. 2003), and it will be more efficient to perform blockwise tests than individual tests of markers within blocks (Chapman et al. 2003). It makes sense to combine all the evidence within blocks and to regard the blockwise tests as the correlated units in the sum statistics. This strategy inevitably produces statistics with different degrees of freedom—therefore on different scales—whereas the *P* value is on a common scale. On the other hand, accurate *P* values may not always be available, such as when distributional assumptions are not met. In this case, it can be more convenient to work with test statistics, and the extreme-value distribution can still be used to model the permutation distribution.

We have considered the sums of fixed length and the most significant fixed-length sum, but these are not the only possible combinations. For example, sums could be constructed on the basis of all tests that are significant at a nominal level (Zaykin et al. 2002). The number of nominally significant tests follows a binomial distribution, so that the sum can be regarded as the maximum of a set of variables whose distribution is a mixture of distributions for fixed length sums. This also allows the extreme-value distribution to be applied in this case. More generally, our approach is applicable whenever the null distribution falls within a parametric class of analytic distributions.

An economical approach to genome scanning is to screen all markers in the first stage and only the significant markers in the second. This strategy reduces the total genotyping cost and has been proposed both for individual genotyping and for pooling protocols (Sagatopan et al. 2002; Sham et al. 2002). Combined-evidence methods are a natural choice for the first stage, providing a powerful method to identify follow-up loci, while controlling the type I error for the null hypothesis that there are no genetic effects at all. They provide an appealing trade-off between reducing the size of the exploratory space and controlling the error of individual tests, giving confidence and justification for the follow-up testing. The efficient approach we propose for applying these methods on genomewide scales will allow them to be applied more widely in forthcoming studies.

Acknowledgments

F.D. is supported by European Commission grant 503485. B.P.C.K. is supported by funding from the Dutch Diabetes Research Foundation, The Netherlands Organization for Health Research and Development, and The Juvenile Diabetes Research Foundation International grant 2001.10.004.

References

Austin MA, Harding S, McElroy C (2003) Genebanks: a comparison of eight proposed international genetic databases. Community Genet 6:37–45

Bailey TL, Grundy WN (1999) Classifying proteins by family using the product of correlated *p*-values. Paper presented at the Third International Conference on Computational Molecular Biology, Lyon, France, April 11–14

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Brown BW, Russell K (1997) Methods of correcting for multiple testing: operating characteristics. Stat Med 16:2511–2528

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52–58

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

Coles S (2001) An introduction to statistical modelling of extreme values. Springer, London

Cordell HJ, Clayton DG (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to *HLA* in type 1 diabetes. Am J Hum Genet 70:124–141

Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25:115–121

Dudbridge F, Koeleman BP (2003) Rank truncated product of P-values, with application to genomewide association scans. Genet Epidemiol 25:360–366

Fisher RA (1932) Statistical methods for research workers. Oliver and Boyd, London

Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York

Hao K, Xu X, Laird N, Wang X, Xu X (2004) Power estimation of multiple SNP association test of case-control study and application. Genet Epidemiol 26:22–30

Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 4:701–709

Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 11:2115–2119

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

Johnson GC, Koeleman BP, Todd JA (2002) Limitations of stratifying sib-pair data in common disease linkage studies: an example using chromosome 10p14-10q11 in type 1 diabetes. Am J Med Genet 113:158–166

Li W, Grosse I (2003) Gene selection criterion for discriminant microarray data analysis based on extreme value distributions. Paper presented at the Seventh International Conference on Computational Molecular Biology, Berlin, April 10–13

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87:2264–2268

Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769

Ophoff RA, Escamilla MA, Service SK, Spesny M, Meshi DB, Poon W, Molina J, Fournier E, Gallegos A, Mathews C, Neylan T, Batki SL, Roche E, Ramirez M, Silva S, De Mille MC, Dong P, Leon PE, Reus VI, Sandkuijl LA, Freimer NB (2002) Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate. Am J Hum Genet 71:565–574

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654

Pesarin F (2001) Multivariate permutation tests with applications in biostatistics. Wiley, Chichester, United Kingdom

Pounds S, Morris SW (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics 19:1236–1242

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137

R Development Core Team (2003) R: a language and environment for statistical computing. R foundation, Vienna, Austria

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856

Risch NJ, Merikangas KR (1996) The future of genetic studies of complex human disease. Science 273:1516–1517

Sabatti C, Service S, Freimer N (2003) False discovery rate in linkage and association genome screens for complex disorders. Genetics 164:829–833

Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002) Two-stage designs for gene-disease association studies. Biometrics 58:163–170

Sawcer S, Maranian M, Setakis E, Curwen V, Akesson E, Hensiek A, Coraddu F, Roxburgh R, Sawcer D, Gray J, Deans J, Goodfellow PN, Walker N, Clayton D, Compston A (2002) A whole genome screen for linkage disequilibrium in multiple sclerosis confirms disease associations with regions previously linked to susceptibility. Brain 125:1337–1347

Schliekelman P, Slatkin M (2002) Multiplex relative risk and estimation of the number of loci underlying an inherited disease. Am J Hum Genet 71:1369–1385

Schulze A, Downward J (2001) Navigating gene expression using microarrays: a technology review. Nat Cell Biol 3:E190–E195

Service SK, Sandkuijl LA, Freimer NB (2003) Cost-effective designs for linkage disequilibrium mapping of complex traits. Am J Hum Genet 72:1213–1220

Sham PC, Bader JS, Craig I, O'Donovan M, Owen M (2002) DNA pooling: a tool for large-scale association studies. Nat Rev Genet 3:862–871

Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 78:626–633

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Stephenson AG (2002) EVD: extreme value distributions. R-News 2:31–32

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 100:9440–9445

Weale ME, Depondt C, Macdonald SJ, Smith A, Lai PS, Shorvon SD, Wood NW, Goldstein DB (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene *SCN1A:* implications for linkage-disequilibrium gene mapping. Am J Hum Genet 73:551–565

Wille A, Hoh J, Ott J (2003) Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. Genet Epidemiol 25:350–359

Zaykin DV (1999) Statistical analysis of genetic associations. PhD thesis, North Carolina State University, Raleigh

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2002) Truncated product method for combining P-values. Genet Epidemiol 22:170–185