

# The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits

Anthony D. Long<sup>1,3</sup> and Charles H. Langley<sup>2</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, University of California at Irvine, Irvine, California 92697-2525 USA;

<sup>2</sup>Center for Population Biology, University of California at Davis, Davis, California 95616 USA

The statistical power of five association study test statistics (two haplotype-based tests, two marker-based tests, and the Transmission Disequilibrium Test–Q5) to detect single nucleotide polymorphism (SNP)/phenotype associations in a linkage–disequilibrium-based candidate gene scan employing a number of SNPs is examined. Power is estimated as a function of realistic parameters expected to affect the likelihood of detecting a significant association: the number of SNPs examined, the scaled recombination size of the region examined, the proportion of variance in the trait attributable to a hidden causative polymorphism within the region, and the number of individuals or families examined. For the different combinations of parameter values, power is estimated from a large number of realizations of a simulated coalescent describing a single random mating population with mutation, random genetic drift, and recombination. This explicit population genetics model results in a distribution of DNA marker heterozygosities and linkage disequilibria that are likely to resemble those expected in actual population samples. The study concludes that (1) marker-based permutation tests are more powerful than simple haplotype-based tests, (2) there is sufficient power to detect the presence of causative polymorphisms of small effect if on the order of 500 individuals are sampled, (3) greater power is achieved by increasing the sample size than by increasing the number of polymorphisms, (4) association studies are generally more powerful than transmission disequilibrium-based tests, and (5) for the range of parameters considered association studies have a low repeatability unless sample sizes are on the order of 500 individuals. Estimates of  $4Nc$  for a number of gene regions and human populations will be of use in determining the density of SNPs that are likely to be required for successful association studies.

Great progress in the mapping, positional cloning, and characterization of genes that are responsible for Mendelian diseases in humans has been achieved. These successes provide the methods and optimistic prospectus for the successful identification and characterization of loci at which alleles with large disease effects are unavailable. Both genomics and developmental genetics are providing increasing numbers of potential candidate genes for disease risk. Many genes that contribute significantly to variation in risk of prevalent diseases may only be represented in the human population by common alleles with subtle phenotypic effects. Variation among individuals in complex traits, like the quantitative characters of interest to animal and plant breeders and evolutionary biologists, can be viewed as the sum of a number genes, the interactions between these genes, the nongenetic or environmental influences on the phenotype, and the interactions between the environmental and genetic factors (Falconer and Mackay 1996). The inherent complexity of these traits in combination with the relatively subtle effects of the individual genetic factors make complex dis-

eases difficult to study using traditional linkage-based approaches. Association-based methods, in which the joint distribution of phenotypes and genotypes in population samples are examined, may be more powerful than linkage studies for identification of the genes that contribute to variation in complex traits (Spielman et al. 1993; Risch and Merikangas 1996; Long et al. 1997). Theoretical modeling suggests that a genome wide association scan employing *every polymorphic marker in the human genome* may have greater power to detect complex disease causing polymorphisms than genome wide linkage studies, even after compensating for the increased number of false positives expected from testing such a large number of markers (Risch and Merikangas 1996).

The suggestion that association study approaches may be more powerful than linkage-based approaches for the identification of loci that contribute to disease risk is based on the assumption that one of the markers typed will actually be the DNA polymorphism which contributes to variation in disease (Risch and Merikangas 1996; Long et al. 1997). Until further technological advances are made, it is likely that markers will not be discovered and typed at a density high enough to justify the assumption that one of the typed polymor-

<sup>3</sup>Corresponding author.  
E-MAIL [tdlong@uci.edu](mailto:tdlong@uci.edu); FAX (949) 824-2181.

phisms is likely to be causative. An alternative approach is to detect polymorphic DNA markers that are in linkage disequilibrium with the polymorphism at the disease causing site, as opposed to the disease causing variant itself. In this situation the power of the association study approach to detect marker/phenotype associations will be reduced and will depend on the distribution of linkage disequilibrium. It is also likely that initial association studies will focus on candidate gene regions for which a number of single nucleotide polymorphisms (SNPs) have been previously identified, as opposed to genome wide scans. Thus it is important to determine the power of association-based studies when a dense, but not exhaustive, set of SNPs is available throughout a region likely to harbor a site contributing to variation in disease risk. Such studies will also allow an evaluation of the density of SNPs that are likely to be required for association scans of candidate gene regions to be effective.

Linkage disequilibrium-based methods rely on association between neutral DNA polymorphisms and polymorphisms contributing to disease risk. Many population genetic models predict greater linkage disequilibrium between more closely linked loci. The expected linkage disequilibrium between selectively neutral alleles due to genetic drift is approximated by a simple function of the product of the population size and the rate of recombination between them (Hill and Robertson 1968). Thus, it is plausible that “scans” of candidate genes, or perhaps even genome wide scans, with very dense sets of SNPs, could be used to “association map” the genes that contribute to variation in complex traits. To realistically evaluate the power of association studies, it is important that the pattern of linkage disequilibrium between polymorphic markers resemble that expected in population samples. At present, there are no analytically derived probability distributions describing the expected pattern of disequilibria in candidate gene regions. This is even the case for relatively simple population genetic models involving only mutation, recombination, and random genetic drift. Nonetheless, it is possible to efficiently simulate samples that are the result of such a mechanistically based evolutionary process. This “coalescent” approach for generating samples consisting of closely linked polymorphic DNA sites (e.g., SNPs scattered throughout a candidate gene region) based on an explicit population genetics model has been characterized extensively (Hudson 1983). Other means of generating samples with a number of SNPs spaced throughout a candidate gene region that do not incorporate an explicit model involving the forces that give rise to extant patterns of DNA variation are likely to be misleading.

Work presented here examines the power of association studies to implicate polymorphisms in candi-

date gene regions as contributing to variation in a complex trait. The probability of detecting a SNP/phenotype association is examined for a number of association study test statistics over a range of parameters. These variables can either be controlled by an experimenter, or can be estimated within the context of a given experiment. A hidden causative DNA polymorphism, called the quantitative trait nucleotide (QTN), is assumed to contribute a varying proportion of the total variation in a quantitative trait. The number of individuals sampled from the population and the number of SNPs typed throughout a candidate gene region in each individual are varied and replicated in a large number of simulations. These simulations allow an exploration of the densities of SNPs and numbers of sampled individuals that will be required to detect DNA polymorphisms contributing to a small fraction of the total variation in a complex trait. Currently there is a great deal of interest in developing a set of SNPs spanning the human genome, which can be used for such association studies (Collins et al. 1997; Wang et al. 1998). The work presented here is directly relevant to determining the density of SNPs that are likely to be required to identify genes contributing substantially to variation in complexly inherited disease phenotypes.

A number of simulations were carried out to examine different aspects of the power of association studies. First, a set of simulations were carried out to assess the power of a single marker-based permutation test in haploids (the haploid marker permutation test or HMP) over a wide range of parameter values. Significance of the HMP is determined by generating a distribution of *F*-statistics for the marker showing the strongest association with variation in phenotype when phenotypic measures are randomly permuted over individuals with respect to marker haplotypes (Churchill and Doerge 1994; Long et al. 1998). The objective of this initial set of simulations was to determine a refined set of parameter values for which the power of a more extensive set of association test statistics could be measured. A second set of simulations were carried out under this narrower set of parameter values in which three additional test statistics were assessed. These were (1) the diploid marker permutation test (DMP), which is effectively the HMP applied to diploid individuals; (2) the haploid haplotype analysis of variance (ANOVA) test (HHA), a one-way ANOVA which tests if any of the *H* distinct haplotypes observed differ in their average effect on phenotypic variation in haploid individuals; and (3) the haploid haplotype permutation test (HHP), a test of whether any of the *H* observed haplotypes differ from the mean of the other  $H - 1$  haplotypes in haploid individuals assessed using a permutation testing procedure. A third set of simulations were carried out to measure the repeatability of asso-

ciation studies. Repeatability is defined as the probability of a second association study being significant given that an initial association study is significant and the second sample of alleles is drawn from the same population. Because the four test statistics examined had either comparable power, or the marker-based tests performed better than haplotype-based tests, repeatability was only assessed for the case of the HMP. A number of statistics have been proposed that test for the presence of both linkage and association between a polymorphic DNA marker and a quantitative trait (Spielman et al. 1993; Allison 1997). The most powerful of these tests, the Transmission Disequilibrium Test-Q5 (TDT-Q5) (Allison 1997), has an advantage over a simple association study approach in that it greatly reduces the likelihood of false positives when the population being examined has experienced recent admixture. Finally, simulations were carried out to compare the power of the DMP to the TDT-Q5. These two test statistics are directly comparable as they both can be applied to samples of diploid individuals. More detailed descriptions of the test statistics examined and the simulation strategy are provided in the Methods section.

The power of statistical tests for detecting SNP/phenotype associations is reported as a function of the parameter  $4Nc$ .  $4Nc$  is interpreted to be four times the effective size of the population from which the gametes are drawn, multiplied by the recombination rate per gamete per generation between the end points of the region under consideration. If recombinational events are equally probable throughout the region being considered, then  $4Nc$  is proportional to physical distance in that region. This parameter can be estimated from data that consists of a number of randomly ascertained DNA polymorphisms typed in a number of gametes (i.e., haplotype data; Hudson 1987).  $4Nc$  can also be estimated from the variance in the number of heterozygous sites over diploid individuals, although the efficiency of such an estimator of  $4Nc$  is unknown (R. Hudson, pers. comm.). Thus,  $4Nc$  can be directly estimated from data consisting of a number of SNPs typed in a population study. Under the assumptions of selectively neutral mutations and random genetic drift, patterns of linkage disequilibria in a genomic region, and as a result the power of a given association study, are a function of  $4Nc$ . Therefore, although  $4Nc$  is not an intuitive unit of measure for many geneticists, it is a natural measurement scale in a population genetics context, such as association mapping. For example, patterns of disequilibrium will scale with  $4Nc$  but not with recombinational distance derived from a single generation meiotic mapping experiment, among populations that have experienced different historical population sizes.

The results of these simulations indicate that

marker-based permutation tests are more powerful than simple haplotype-based tests for detecting SNP/phenotype associations. The simulations also show that for realistic parameter values there is sufficient power to detect SNP/phenotype associations for QTNs that account for ~5% of the variation in a complex trait provided that ~500 individuals are typed for ~20 SNPs spaced throughout the candidate gene region. In most situations the ability to detect SNP/phenotype association is improved faster by increasing the number of individuals in a study than by increasing the number of SNPs typed. It is shown that association studies are generally more powerful than TDT-based tests in single large randomly mating populations and that association studies generally have a low repeatability unless sample sizes are on the order of 500 individuals. The density of SNPs that are likely to be required for effective association mapping varies with the parameter  $4Nc$ . Estimating  $4Nc$  for representative regions of the genome from various human populations is an essential step in determining the required density of SNPs for association mapping studies.

## RESULTS

### Simulation of Genotypic and Phenotypic Data

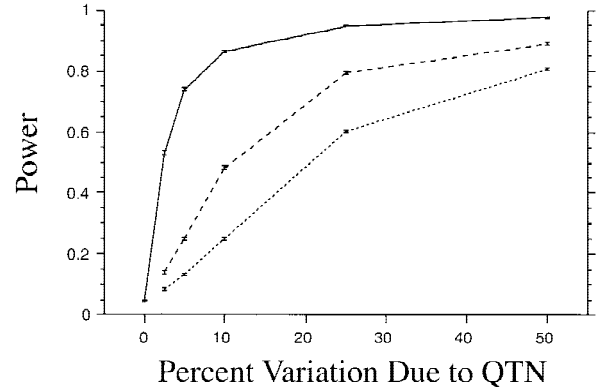
To assess the power of detecting SNP/phenotype associations under a range of population genetic assumptions, a large number of independent replicate gene trees were generated using the selectively neutral coalescent algorithm that incorporates intragenic recombination (Hudson 1983). Each tree defines a topology in which nodes of the tree define a most recent common ancestor of two alleles and branch lengths are proportional to time. Under this model one of the potential ancestries of a sample of selectively equivalent alleles from a finite and randomly mating population of constant size is constructed assuming a fixed and uniform distribution of recombination rates across a genomic region. Then, a specified number of selectively neutral mutations are distributed according to a random uniform distribution across the genomic region and its gene tree, proportional to physical length and number of generations. From the distribution of these mutations, the genetic state of each sampled gamete at each segregating (polymorphic) site is determined. These simulated samples of alleles are the data on which all subsequent analyses were based. The parameters of the simulated coalescent trees were the number of individuals sampled, the number of segregating sites, and the scaled recombinational size of the region being considered ( $4Nc$ ). In the case of the haploid analyses, this sample of alleles is analyzed directly. Alleles are randomly paired to create zygotes in the diploid analyses. Although the time since a mutation arose in the population is related to its sample fre-

quency, the approach used here does not model the age of the mutation explicitly. After the sample of gametes is generated, a QTN is chosen to contribute a percentage of the total variation in a quantitative character. The QTN is not one of the SNPs used to detect associations between marker state and phenotypic variation. The site chosen to be the QTN and the method used to generate phenotypic values is described in Methods.

The goal of this investigation is to examine the power of various test statistics to detect associations between SNPs and variation in a quantitative trait. To this end, the simulation parameters that could potentially affect this power are varied in a fully crossed design (every level of each parameter was tested in combination with every level of the other parameters). The parameters and their values are the number of segregating sites {10, 25, and 50}, the percentage of total variation attributable to the QTN {0, 2.5, 5.0, and 10.0}, the recombinational size of the region in scaled units of  $4Nc$  {5, 10, 25, and 50}, and the number of individuals ( $M$ ) examined {50, 100, and 500}. For each combination of parameter values, 1000 neutral coalescent trees were generated and examined for associations between genotype at the polymorphic sites (SNPs) and phenotype determined by the QTN. The number of SNPs tested for associations with phenotypic variation were typically lower than the number of segregating sites in the simulated data, as only sites with a frequency of  $>5\%$  were included in the calculation of the association test statistic. Polymorphisms that were correlated completely over the  $M$ -sampled individuals are collapsed into a single marker.

### The Power of Different Test Statistics to Detect Marker/Phenotype Associations

Figure 1 depicts the power of the HMP to detect associations between markers and phenotypic variation in haploids. It can be seen that when power is averaged over the 12 parameter sets defined by the number of segregating sites and  $4Nc$ , the percentage of variation attributable to the QTN and the number of individuals sampled have a large impact on the power of this association test. It is apparent that if the QTN accounts for a large fraction of the total variation in phenotype, the presence of the QTN can often be detected regardless of the number of individuals examined. As the variance attributable to the QTN decreases into the range (i.e.,  $<10\%$ ) more likely to represent reality for most complex traits, the power to detect an association decreases, even for samples of 500 individuals. In all cases examined increasing the sample size results in an increased frequency of detecting marker/phenotype associations. When the QTN accounts for  $<10\%$  of the total variation in phenotype, the power to detect



**Figure 1** The power of the HMP as a function of the simulated percentage of variation attributable to the QTN. Dotted, dashed, and solid lines are for experiment sizes of 50, 100, and 500 individuals, respectively. Standard errors are over replicate simulations. Each point is the average power over 12 parameter combinations of the number of segregating sites {10, 25, 50} and  $4Nc$  {5, 10, 25, 50}.

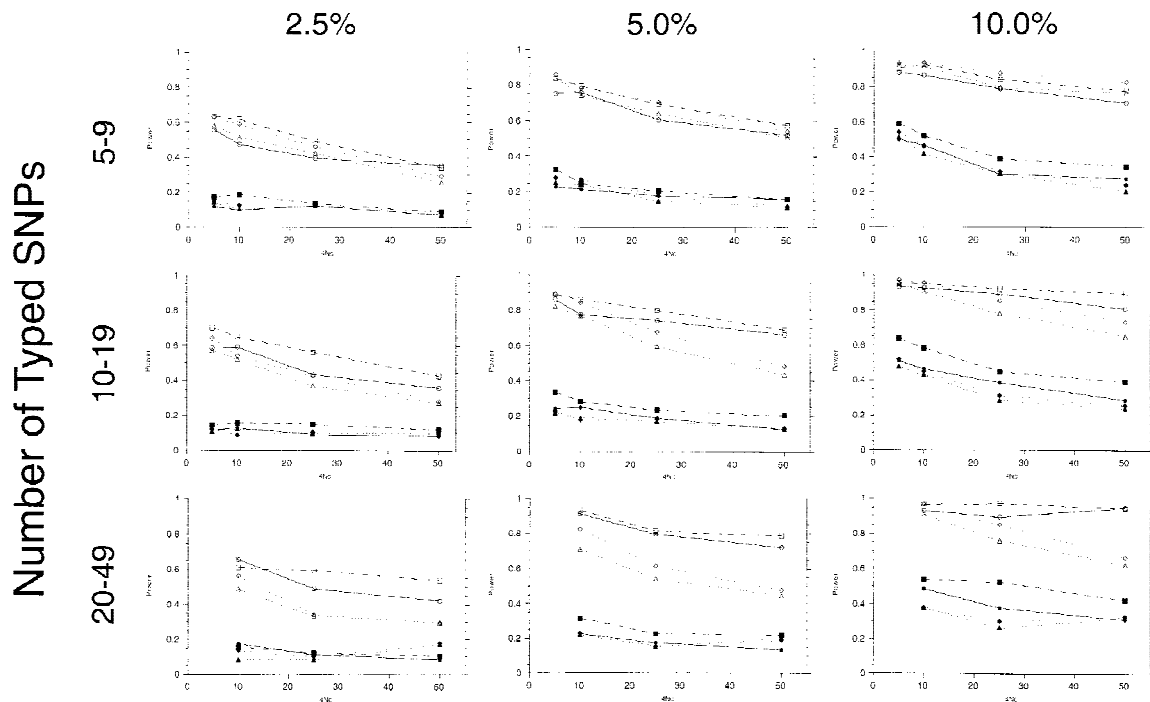
marker/phenotype associations can often be quite low. This will be examined in more detail in the next section.

Figure 2 shows plots of power as a function of simulated  $4Nc$  for the four tests of association between SNPs and phenotypes. Plots are separated based on the number of typed SNPs throughout the surveyed region and the proportion of variance due to the QTN. A number of important relationships between parameters of, or estimates from, the simulations and the power to detect associations between markers and phenotypic variation are presented in Figure 2 and are detailed in the following paragraphs.

Greater power to detect SNP/phenotype associations is consistently observed as the QTN accounts for more of the phenotypic variance and as the sample size increases. In all cases, the marker-based permutation tests had similar or greater power to detect SNP/phenotype associations than the haplotype-based tests. This difference became more pronounced as both the number of typed SNPs increased and the size of the region examined became large. This difference in power between marker-based and haplotype-based tests will become more important as association studies begin to scan larger genomic regions and greater numbers of SNPs become available.

If a QTN accounts for a large portion of the total phenotypic variation and a large number of individuals are surveyed, the power to detect associations is quite high almost irrespective of the size of the region examined or number of markers employed. This suggests that approaches aimed to increase the variance attributable to a QTN will result in more efficient detection of QTNs. Such approaches include the examination of individuals in sensitizing environments or genetic backgrounds and the careful definition of the

## Variance Due to QTN



**Figure 2** The power of detecting an association between SNPs and variation in a quantitative trait for four different statistics designed to detect such associations. The four statistic are represented by squares for HMP, diamonds for HHA, triangles for HHP, and circles for DMP. Solid symbols are for 100 individuals examined; open symbols are for 500 individuals examined. Power was estimated for QTNs accounting for three different percentages of the total phenotypic variation and three different ranges of the number of typed SNPs in the study. In each cell the power is plotted as a function of the total size of the candidate gene region scanned for an association expressed in units of  $4Nc$ . Power estimates for the case of 20–49 typed SNPs when  $4Nc$  is equal to 5 are omitted, because of a small number of replicate measures of power for this combination of parameter values.

phenotype so that observed phenotypic variation is likely to be due to a few segregating factors.

Increasing the number of typed SNPs results in a small increase in the power to detect SNP/phenotype associations when a small region is examined. If a larger region is examined, then increasing the number of typed SNPs results in modest-to-large increases in power regardless of the variance attributable to the QTN. It is generally far more efficient to increase the number of individuals surveyed than to increase the number of SNPs. The implication is that if a region shows weak evidence for a SNP/phenotype association it will generally be more prudent to increase the number of individuals examined than to score additional SNPs on those individuals already typed. However, if it is difficult (or expensive) to recruit more individuals into a study, typing more SNPs will increase power.

Under an additive model of quantitative genetic variation, the marker-based permutation test has comparable power for a given sample size of either diploid or haploid individuals. The power of the haplotype-based tests in diploids was not examined, as this would have necessitated inferring haplotypes from the dip-

loid marker scores. Because this would have added greatly to the computational burden of this study, and the power of HMP and DMP was comparable, this approach was not pursued. An advantage of marker-based tests is that they do not require haplotypes to be inferred. This could result in a considerable savings in genotyping costs, as parents do not have to be genotyped if haplotypes are not inferred (although it may be desirable to genotype parents for use in the TDT as discussed below).

Errors bars are not given in Figure 2 because they were generally small and tended to obscure the trends apparent in each graph. In the vast majority of cases standard errors were between 0.01 and 0.04. Errors associated with the four different statistics designed to detect SNP/phenotype associations appear similar to one another.

### Type I Error

This work assessed the power of four different statistics designed to detect SNP/phenotype associations within the context of an association study. It is important to ensure that any observed increases (or patterns) in

power as a function of variables that can be estimated or experimentally manipulated are not simply the result of increases in type I error. The four statistics employed in this study are designed to hold this error to a constant rate of 5%. The probability of detecting an association that is not present was estimated when the number of individuals sampled was 500, the variance due the QTN was 0, and the observed number of typed SNPs and 4Nc were binned as in Figure 2. In the case of the two marker-based statistics the false-positive rates were 4.0 and 4.2% for HMP and DMP, respectively. For the haplotype-based statistics the false-positive rates appeared to be larger, with average false-positive rates of 6.5 and 6.4% for the HHA and HHP, respectively.

### Simulated vs. Estimated 4Nc as a Predictor of Power

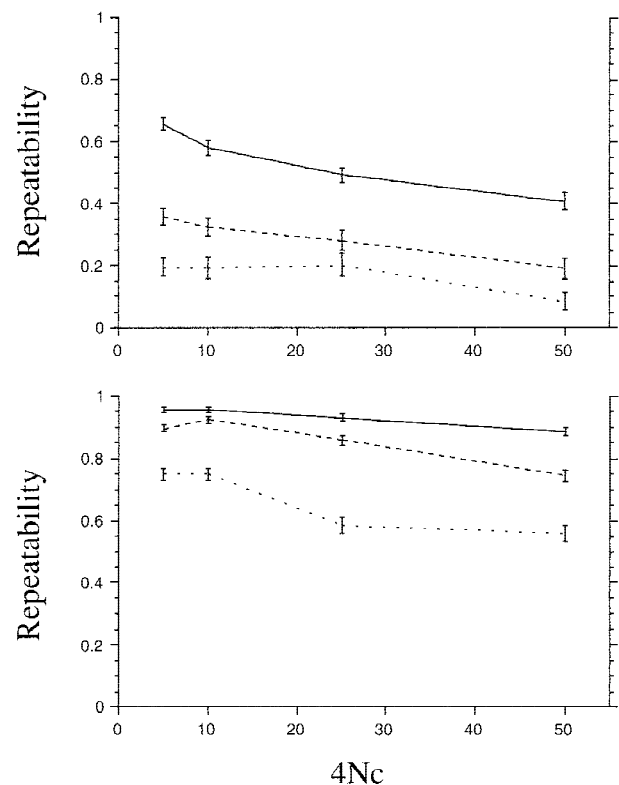
Figure 2 shows the power of the four association study statistics as a function of the simulated parameter value of 4Nc. In reality the experimenter will not know the value of 4Nc but can estimate it from haplotype data (Hudson 1987). It is of interest to examine the power of the different association statistics as a function of estimated (as opposed to simulated) 4Nc. Plots comparable to Figure 2 were generated with estimated values of 4Nc substituted for the simulated parameter (Hudson 1987). Estimated 4Nc values were binned into five classes. These plots (not shown) appear very similar to those shown for the case of simulated 4Nc in Figure 2. Estimates of power for binned estimates of 4Nc generally exhibit larger errors than those based on simulated values of 4Nc, as the binning resulted in greater variation in the number of replicate simulations used to estimate power for given parameter combinations. In addition, the average number of replicate simulations used to estimate power for a given set of parameter combinations was smaller for estimated than for simulated 4Nc, as a greater number of bins were needed to accommodate very small (near 1) and very large (>100) values of 4Nc when it was estimated from the data.

In addition to graphically examining the power of association studies as a function of estimated versus simulated 4Nc, power was examined statistically. ANOVA was carried out with each of the four statistics used as a measure of association as the dependent variable and the percentage of variation attributable to the QTN, the number of individuals examined, the number of typed SNPs, and either 4Nc or estimated 4Nc as predictor variables. Ninety-five percent confidence intervals were constructed for the ratio of the true mean square error from the two models in determine if a model including simulated 4Nc was better at predicting power than a model that included estimated 4Nc. A ratio of 1 indicates that the two predictors explain equal amounts of variation in the model, whereas ratios of >1 indicate the model using simulated 4Nc as a

predictor explains more variation. Ninety-five percent confident intervals on the ratio of the error when estimated 4Nc was used as a predictor relative to simulated 4Nc were [0.982, 1.032], [0.980, 1.030], [1.001, 1.052], and [1.000, 1.051], for HMP, DMP, HHA, and HHP, respectively (Hogg and Craig 1978). These confidence intervals indicate that for the marker-based tests there is no significant difference in using either the simulated or estimated 4Nc as a predictor of power, whereas for the haplotype-based tests simulated 4Nc is better predictor of power than estimated 4Nc but only slightly so.

### The Repeatability of Detecting a Significant Association

The two panels of Figure 3 are plots of the repeatability of an association study for different combinations of parameters that the experimenter can estimate or manipulate. Repeatability was only examined for the case



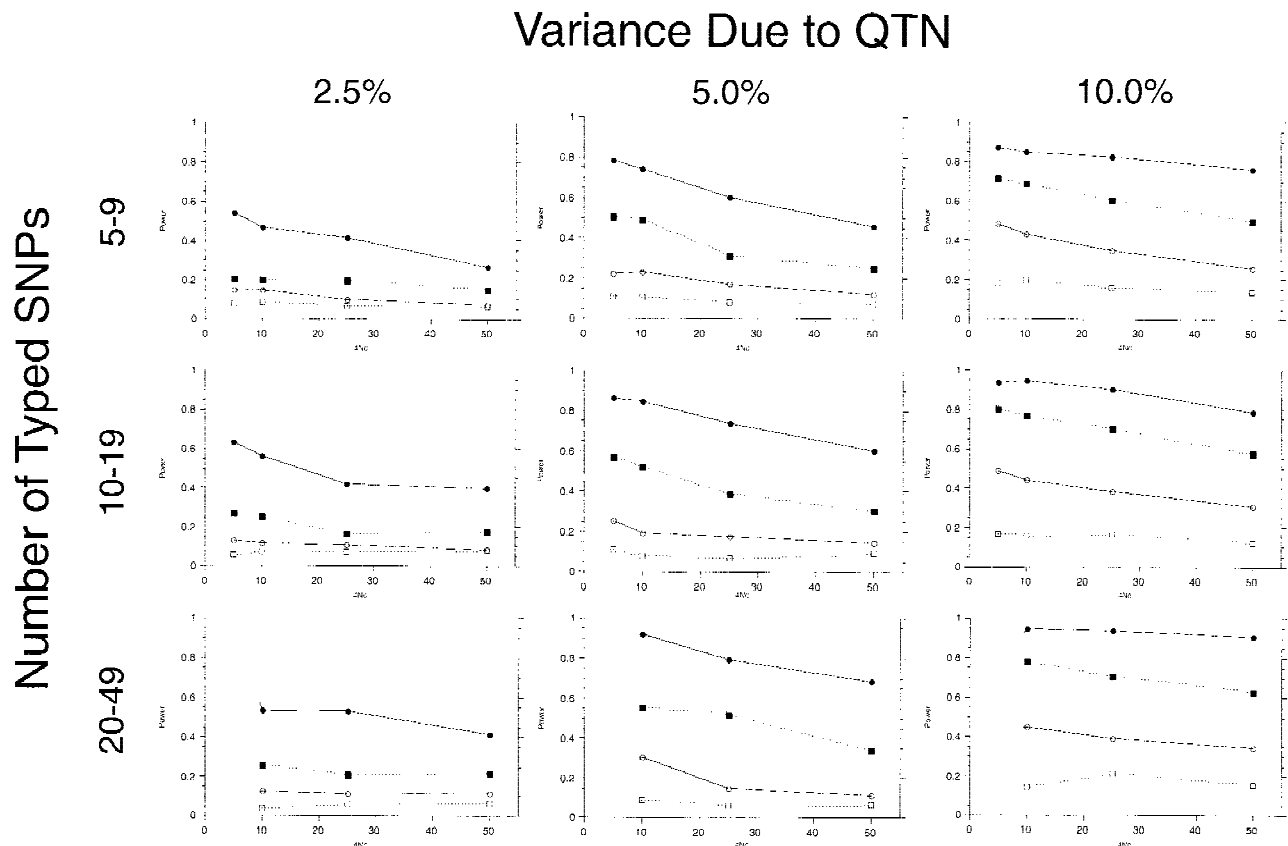
**Figure 3** The repeatability of an association study as a function of 4Nc, the percentage of variation attributable to a QTN, and the number of individuals examined. Repeatability is defined as the probability of a second association study being significant at  $P < 0.05$  using the HMP statistic given an initial association study was also significant at  $P < 0.05$ . Dotted, dashed, and solid lines represent 2.5%, 5.0%, and 10.0% of the total variation attributable to the QTN, respectively. The *top* panel is for 100 typed individuals; the *bottom* panel is for 500 typed individuals. Plotted data are over all realizations of the coalescent, where  $>4$  and  $<20$  SNPs, were typed.

of the HMP and then only for cases where >4 and <20 SNPs were typed. Neither the number of typed SNPs (not shown) nor 4Nc has a large effect on the repeatability of an association study. The repeatability of an association study is very dependent on the number of individuals examined (upper vs. lower panel) and the variance attributable to the QTN. The repeatability of an association study is generally >80% only if the proportion of the phenotypic variance attributable to the QTN is >5% and the sample size is 500. If the number of individuals surveyed is 500 or the QTN accounts for 10% of the total variation, then the repeatability is generally >50%. Hence, if a significant SNP/phenotype association is not replicable in a second study, even when the second study is carried out on the same population, one cannot necessarily conclude that the original association was spurious.

**TDT-Q5 vs. Association Studies**

Figure 4 shows plots of power as a function of simulated 4Nc for the DMP and the TDT-Q5 test statistics. Plots are separated based on the number of typed SNPs in the surveyed region and the proportion of variance

due to the QTN. Two different levels of the number of individuals examined are displayed in each panel. For both the association study approach and TDT-Q5, increasing either the number of individuals examined or the total proportion of the variation due to the QTN has a great effect on power. For both statistics, increasing the number of typed SNPs results in more modest gains, although additional SNPs can help a great deal if the size of the region examined is large. The most striking feature of Figure 4 is that the power associated with the TDT-Q5 is much less than the power associated with the single-marker-based association study approach. This effect is likely due to the fact that for *any given marker* used in the TDT-Q5, a large and varying proportion of the families in the study are potentially uninformative. For QTNs contributing 5% to the total variation in a quantitative trait and with 20–49 markers, even for a very small region of 4Nc equal to 10, the power of the TDT-Q5 only approaches 60% (compared to the association study approach having >90% power). In the more likely case of the scan of the entire candidate gene region (i.e., 4Nc = 50) the power of the TDT-Q5 is <40% (compared to the association study statistic having power approaching 70%).



**Figure 4** The power of detecting an association between SNPs and variation in a quantitative trait for DMP and TDT-Q5 statistics. Circles represent the DMP; squares represent the TDT-Q5 test statistic. Open symbols are for 100 individuals examined; solid symbols are for 500 individuals examined.

## DISCUSSION

The simulation study presented here shows that if QTNs act in an additive manner and DNA polymorphisms throughout a candidate gene region are evolving as neutral mutations under the sole influence of random genetic drift, then reliable detection of association between markers in a candidate gene region and the phenotypic effects of the QTN can be achieved in experiments of large but realistic size. Such associations become easier to detect as QTNs account for increasing phenotypic variation. The power to detect associations is also improved by increasing the number of individuals surveyed and by increasing the number of SNPs in a candidate gene region. In general, it is more powerful to increase the number of individuals examined than the number of typed SNPs. Increasing the number of typed SNPs is only helpful if the region being considered is recombinationally large. Over the entire parameter space examined in this work and under the simple population genetic model considered, single-marker-based, permutation-based tests are either of similar or greater power than haplotype-based tests. However, under different models relating genotype to phenotype or under different demographic scenarios this conclusion may not be valid.

The haplotype-based tests considered in this study are based on haplotypes constructed solely on the basis of unique combinations of markers. It is possible that other haplotype-based tests that make use of the evolutionary history of the markers are more powerful than the simple haplotype-based tests examined here (Templeton et al. 1987; Templeton and Sing 1993). One problem with these evolutionary-based tests is that they are difficult to use when the region under consideration has experienced recombination. These tests may be particularly inappropriate for some of the "large 4Nc" candidate gene scanning resolutions examined in this work (i.e., 4Nc >10). An additional problem is that the cladistic approach is difficult to automate and implement. In future work it may be fruitful to directly compare the evolutionary haplotype-based tests to simple marker-based tests over a large number of simulated data-sets in order to determine in what situations one test outperforms the other.

In *Drosophila*, 4Nc has been estimated for a number of gene regions, with a typical value being one 4Nc equal to 200 bp. Comparable studies generally have not been carried out in humans. An analysis of six published data sets suggests that one 4Nc represents between 1 and 50 kb, with values favoring the former more commonly observed. One study directly estimated 4Nc to be 680 bp (Clark et al. 1998). Three of the studies published genotypic data, allowing direct estimates of 4Nc using Hudson's (1987) method: 1.3 kb for the insulin-receptor region (Elbein 1992), 2.2 kb for the

low-density lipoprotein region (Leitersdorf et al. 1989), and 6.1 kb for the phenylalanine hydroxylase region (Chakraborty et al. 1987). In two other studies the correlation coefficient was plotted as a function of physical distance, so the physical distance represented by one 4Nc was interpolated from these figures: 3 kb for the von Willebrand factor region (Watkins et al. 1994), and 40 kb for the adenomatous polyposis coli region (Jorde et al. 1994). The latter five studies may be a biased sample as they represent studies that showed a significant correlation between linkage disequilibrium and physical distance, although studies not observing such a trend generally examined a smaller sized region or fewer markers (Jorde et al. 1994). These estimates are also necessarily approximate, as much of the human population genetic data tends to be selectively presented and results from variable ascertainment of polymorphic markers (but see Clark et al. 1998).

Population genetic studies of different human populations with the objective of estimating 4Nc would be of great value in determining the power of future association studies in these populations. 4Nc will ultimately determine the number of markers required to reliably detect associations between DNA markers and phenotypic variation. It is likely that estimates of 4Nc per kilobase will be different in different regions of the human genome and in different human populations. If we are willing to assume that one 4Nc in humans is equal to 5 kb, then the simulated case of 4Nc equal to 50 of this study would correspond to a rather large candidate gene region of ~250 kb, and the 4Nc equal to 5 of this study would correspond to a few candidate exons or promoter and upstream regulatory region of a gene, or ~25 kb.

In this work it is assumed that DNA polymorphisms within a candidate gene region and the QTN itself are evolving as selectively neutral variants under the influence of genetic drift. This may be true of QTNs that act late in life or of those manifested only recently in human evolution. It is also possible that QTNs may act at a number of developmental stages and influence a number of phenotypic characters in ways that give rise to natural selection. It is difficult to determine how violations of the assumption of selective neutrality would affect the conclusions of this work. Unfortunately, distinguishing between neutral and non-neutral patterns of disequilibrium has not received a great deal of research attention. No general statistical tests that directly test for departures from the neutral theory predictions of the joint distributions of linkage disequilibria have been developed (see below). There are various tests based on other distributional properties of DNA sequence polymorphism and divergence (Hudson 1994).

The TDT has received a great deal of attention as a means of detecting associations between DNA poly-

morphisms and phenotypic variation (Spielman et al. 1993; Risch and Merikangas 1996). The TDT has been extended to test for associations and linkage between a polymorphic DNA marker and variation in a quantitative trait (TDT-Q5) (Allison 1997). The TDT (and TDT-Q5) requires that a DNA polymorphism both be in linkage disequilibrium and that it segregates with disease status to be significant (Spielman et al. 1993; Allison 1997). As a result, the TDT (and TDT-Q5) is unlikely to give spurious associations due to very recent population admixture or population samples that are stratified with respect to genetically differentiated groups. Use of the TDT (and TDT-Q5) gives the researcher confidence that an observed SNP/phenotype association is not simply a sampling artifact (Lander and Schork 1994). Tests of association that also require demonstrable linkage have a cost associated with their use relative to population level association studies: For any given polymorphic DNA marker a percentage of families sampled will be uninformative and as a result will not contribute to the power of the statistical test. Although the reduction in the ability of the TDT-Q5 to detect associations between SNPs and variation in a quantitative trait relative to a purely association study-based approach can be quite striking, such reductions in power are justifiable if the population being studied is likely to be admixed.

Because the TDT test statistic is not related to  $4Nc$  in a theoretically well-defined manner, it may be difficult to use the TDT statistic to association map QTNs. Population level association studies offer a number of potential benefits, provided the admixture problem can be effectively addressed. In particular, population level association studies allow a single large sample of individuals typed for a number of SNPs to be used in association studies of many different complex traits (and late onset diseases are not particularly excluded from analyses) (Long et al. 1997). In future studies, as more SNPs become available and gene regions are typed for a large number of these markers, it may be possible to directly detect admixture in a gene region. Direct detection of admixture may allow statistical analyses to be stratified by admixture classes, potentially controlling for the effect of admixture and additionally offering the possibility of detecting sites contributing to variation in complex traits with different effects in the populations over which the sample is stratified. Unlike the TDT, statistical control of admixture could control for very "old" admixture, potentially affecting only small genetic regions. Until the ability of statistical control of admixture based on high densities of DNA markers is assessed, it would seem prudent to confirm any population-level associations with a TDT-type statistical test.

If a significant association between a SNP and variation in a quantitative trait is the result of a SNP

showing the association being in linkage disequilibrium with the QTN (as opposed to the QTN itself), it is unlikely that this association will be replicable in a second association study. The results of the simulations show that association studies are unlikely to examine a large enough number of individuals for observed associations to be replicable, even in a sample drawn from the same population. If the sample is drawn from a second population, then the replicability may be lower still, because the pattern of disequilibrium will vary over populations and/or the effect associated with a QTN is expected to vary due to epistatic interactions or genotype by environment interactions. Failure to replicate an association study (either with a second association study or TDT study) has been attributed to admixture in the association study sample (Lander and Schork 1994). It is almost certain that many association studies that are not replicable in a second association study (or a TDT study) result from the initial association being due to population admixture. However, it also seems likely that some of the instances of failure to replicate are due to the inherent problem of the low repeatability of association studies.

The statistics examined in this study test for associations between DNA polymorphisms in candidate gene regions and phenotypic variation in a quantitative trait. If the test statistic is significant as assessed by a permutation test or an ANOVA, the implication is that at least one marker in the candidate region is associated with phenotypic variation. Population genetic theory shows that

$$E[\sigma_{\text{marker}}^2] = E[R^2] \sigma_{\text{QTN}}^2 = \sigma_{\text{QTN}}^2 / 1 + 4 Nc$$

where  $E[x]$  is the expectation of  $x$ ,  $\sigma_y^2$  is the variance of  $y$ , and  $R^2$  is the correlation between the QTN and marker (Hill and Robertson 1968; Lai et al. 1994). That is, the variance associated with a significant DNA polymorphism is proportional to the product of the variance due to the QTN and the inverse of the distance between the marker and the QTN. This relationship suggests that associations between markers and phenotypic variation in random samples of individuals may allow a QTN to be localized. Unfortunately, population genetic theory also predicts a large stochastic variance associated with independent evolutionary realizations. As a result, the location of a disease-causing locus can not be predicted with much accuracy (Hill and Weir 1994). Furthermore, pairs of marker sites are not independent of one another and no theory instructs us how to efficiently combine information from more than one marker to aid in estimating the position of a QTN. This is disappointing, as experimental data will often consist of multiple SNPs spaced throughout a candidate gene region. It seems likely that all of the SNPs provide information pertaining to the location of

the QTN within the gene region. Likelihood-based approaches that simultaneously consider all markers may eventually allow optimal localization of the sites contributing to quantitative variation (Kuhner et al. 1995; Griffiths and Marjoram 1996). These approaches do not seem computationally practical (at the present), nor do they currently address the issue of gene localization. More heuristically derived and computationally practical statistical approaches, although less founded in specific population genetic theory, may be more broadly applicable, at least in the short term.

The ideal experimental design (e.g., random samples, case control, TDT) for detecting SNP/phenotype associations and using marker data to localize a putative QTN is currently unknown. Work presented in this paper suggests that power varies over statistical tests designed to detect SNP/phenotype associations and that power is also likely to vary over demographics from which population samples will be drawn. An equilibrium population genetic model that only incorporates mutation, recombination, and random genetic drift is likely to be unrealistic for actual samples drawn from human populations. The statistical properties of patterns of disequilibria under non-equilibrium demographic conditions have only begun to be explored (Slatkin 1994). But population genetic models based on the coalescent have a mechanistic and flexible basis and can thus be extended to incorporate additional forces such as admixture, population expansion, gene conversion, and selection. The addition of such factors to the evolutionary/demographic model is likely to decrease the power of both the simple association type statistics and TDT-based statistics employed in this work. Thus, it is possible that the power estimates presented in this work represent “upper bounds” for these simple statistical approaches. Simulation studies employing more realistic population genetic models and the coalescent process may ultimately define both the number of SNPs required and the experimental designs necessary to successfully dissect complex traits.

## METHODS

### Generation of Phenotypic Data

A single realization of the coalescent process results in a set of gametes (or haploid individuals) with polymorphic sites numbered from 1 to  $S$  along the genomic region. The site chosen to be the QTN was always equal to the greatest integer less than  $S/4$ . The phenotypes of each (haploid or diploid) individual are

$$Y_i^{hap} = \sqrt{1 - \pi} z_i + Q_i \sqrt{\pi^{-1} p(1 - p)} \text{ and} \\ Y_i^{dip} = \sqrt{1 - \pi} z_i + (Q_{iA} + Q_{iB} - 1) \sqrt{\pi^{-1} 2 p(1 - p)}$$

where  $\pi$  is the desired proportion of variation attributable to the QTN,  $z_i$  is a random normal (mean = 0, variance = 1) de-

viate,  $p$  is the sample frequency of the polymorphism at the QTN,  $Q_i$  is the state (0 or 1) of the QTN in the  $i$ th haploid individual, and  $Q_{iA}$  and  $Q_{iB}$  are the states of the QTN in the  $i$ th diploid individual. In the case of the diploids, a strictly additive model of gene action is assumed. The formulas above show that if a particular realization of the coalescent results in a low frequency QTN, a large effect would be associated with it. If the QTN is at intermediate frequency, it will have a smaller effect. Only realizations of the coalescent yielding a QTN having a frequency of at least 5% are considered. In the cases where low frequency QTNs account for a large enough fraction of variation to be detectable, they would have a large enough effect to be studied using more traditional meiotic mapping methods.

### Assessing Associations between Phenotypic Variation and Marker State

The ability to detect associations between polymorphic markers and variation in a quantitative trait was assessed for four different statistical tests. In the HMP test each marker is examined independently for an association between marker state and variation in phenotype using an ANOVA. To combine information from multiple markers, the largest  $F$ -statistic over all the markers considered is compared to the distribution of the same largest, marker-associated  $F$ -statistic over 1000 shufflings of the simulated data set in which the phenotypic data have been permuted randomly with respect to the marker data (Churchill and Doerge 1994). This approach has been employed successfully to detect molecular marker/phenotypic associations in *Drosophila* (Long et al. 1998). The HMP was applied to data sets consisting of haploid individuals. In the case of the DMP the same test was applied, but to diploid, as opposed to haploid individuals. In the case of the haploid data each marker can take on two possible states (0 and 1), whereas in the case of the diploid data set each marker can take on three states (0, 1, and 2) corresponding to the sum of the marker states over the two gametes that compose a diploid. This permutation testing approach can be extended to arbitrary dominance in a straight forward manner, but that is not the goal here.

It is possible that analyses based on haplotypes that combine information over polymorphic markers have greater power than single marker tests for assessing SNP/phenotype associations. The power of two different haplotype-based test statistics was assessed. The HHA is a one-way ANOVA that tests if any of the  $H$ -observed distinct haplotypes differ in their average effect on phenotypic variation (null hypothesis that  $h_1 = h_2 = \dots = h_H$ , where  $h_i$  is the average phenotype associated with the  $i$ th haplotypic class). The HHP is a test of whether any of the  $H$ -observed haplotypes differ from the mean of the other  $H - 1$  haplotypes (a test of the null hypothesis that  $h_i = h_i'$ , where  $h_i'$  is the average phenotype of all individuals not associated with the  $i$ th haplotypic class, for all  $i$ ). Because the HHP involves  $H$  statistical tests, significance is assessed using the same permutation testing approach described above. That is, it was determined if the haplotypic class with the largest effect on phenotypic variation (assessed by an  $F$ -statistic) is significant when compared to the most significant haplotypic class over 1000 permutations of the simulated data. Although other haplotype based tests exist that are likely to be more powerful than the tests described here, they are generally more difficult to automate and implement.

Haplotypes analyzed in the two haplotype-based tests are constructed from all polymorphic markers with a frequency of >5%. For the two haplotype-based statistical tests of association between haplotypes and phenotypic variation, haplotypic classes with frequencies of <5% are pooled into one class. Weakly polymorphic markers (rarest state <5%) are excluded, whereas rare haplotypes are pooled for two reasons. First, there is little power to detect association between such rare markers (or haplotypes) and QTNs of small effect. The average phenotypic measure associated with a rare class or marker has a large error, as the rare class or marker will only be represented by a small number of individuals. Second, the markers (e.g., SNPs) likely to be employed in future association studies will be ascertained with bias. The rarer the polymorphism, the less likely it is to be initially discovered and included in efficient and "highly informative" genotyping systems. Haplotype-based tests were only applied to haploid data sets so as not to confound the problem of haplotype estimation with the power of association tests based on haplotype information.

### Repeatability of Association Studies

It is of interest to determine the repeatability of association studies as a function of factors within and beyond the control of an experimenter. To estimate the repeatability of an association test, 1000 realizations of the coalescent were generated in which the number of segregating sites {10, 25, and 50}, the percentage variation attributable to the QTN {0, 2.5, 5.0, and 10.0}, the total recombinational size of the gene region ( $4Nc$ ) {5, 10, 25, and 50}, and the number of individuals sampled ( $M$ ) {100, and 500} varied. In each realization of the coalescent,  $2M$  gametes were used to create  $2M$  haploid individuals, which were divided into two groups of size  $M$ , with molecular marker/phenotypic associations being assessed in each group using HMP. Repeatability is defined as the probability of detecting a significant association at  $P < 0.05$  in the second sample from a given replicate given that the first sample is observed to be significant at  $P < 0.05$ . That is, repeatability is the probability of a significant association test being replicable in a second sample of the same size drawn from the same population. Our concept of repeatability assumes an experiment in which all SNPs are typed in both samples, and a single SNP in the second study (not necessarily the same SNP as in the first study) shows a significant association with variation in the phenotype. This is a likely context for replicating association studies if GeneChips or microarrays designed for typing a specific candidate gene region are available to the experimenter.

### TDT-Q5 vs. Population Association

The TDT-Q5 reduces the rate of false positives due to linkage disequilibrium between unlinked markers and phenotypic variation through a test statistic that requires *both* linkage and linkage disequilibrium to be statistically significant. We compared the power of the TDT-Q5 to a pure association study approach under our idealized model of a large random mating population. To accomplish this, 1000 replicates of  $4M$  gametes were generated for all possible combinations of the following parameter values: the number of segregating sites equal to 10, 25, and 50; the percentage of variation attributable to the QTN being 0, 2.5, 5.0, and 10.0;  $4Nc$  equal to 5, 10, 25, and 50; and  $M$  equal to 100, and 500. For each realization of the coalescent process the  $4M$  gametes were randomly paired to create  $2M$  "parents", which in turn were randomly

combined to produce  $M$  "couples". For each couple, one gamete was chosen from each parent to create a diploid "offspring", with the offspring being assigned a phenotypic value in the same manner as described previously. For each combination of parameter values the power of both the TDT-Q5 and DMP were estimated. In the case of DMP only marker information in the offspring were used, and in the case of the TDT-Q5, marker information from both parents and the single offspring were used. For both test statistics examined, significance over the multiple markers tested was assessed using the permutation testing procedure described earlier. It should be noted that for any given polymorphic DNA marker only a fraction of the  $M$  families contributed to the TDT-Q5 test statistic; furthermore, the actual families that contributed varied over markers.

### ACKNOWLEDGMENTS

We thank R. Hudson for providing the computer code to generate coalescent samples with recombination. We thank R. Spielman, A. Clark, and M. Drapeau for comments on an earlier version of this manuscript. This work was supported by the National Science Foundation (grant DEB-9623970) to C.H.L. and University of California at Irvine start-up funds to A.D.L.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

### REFERENCES

- Allison, D.B. 1997. Transmission disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**: 676–690.
- Chakraborty, R., A.S. Lidsky, S.P. Daiger, F. Guttler, S. Sullivan, A.G. Dilella, and S.L.C. Woo. 1987. Polymorphic DNA haplotypes at the human phenylalanine hydroxylase locus and their relationships with phenylketonuria. *Hum. Genet.* **76**: 40–46.
- Churchill, G.A. and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
- Clark, A.G., K.M. Weiss, D.A. Nickerson, S.L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C.F. Sing. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**: 595–612.
- Collins, F., M. Guyer, and A. Chakravarti. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* **278**: 1580–1581.
- Elbein, S.C. 1992. Linkage Disequilibrium among RFLPs at the insulin-receptor locus despite intervening Alu repeat sequences. *Am. J. Hum. Genet.* **51**: 1103–1110.
- Falconer, D.S. and T.F.C. Mackay. 1996. *Introduction to quantitative genetics*, 4th ed. Addison Wesley Longman, Harlow, Essex, UK.
- Griffiths, R.C. and P. Marjoram. 1996. Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.* **3**: 479–502.
- Hill, W.G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**: 226–231.
- Hill, W.G. and B.S. Weir. 1994. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* **54**: 705–714.
- Hogg, R.V. and A.T. Craig. 1978. *Introduction to mathematical statistics*. Macmillan Publishing Co., Inc., New York, NY.
- Hudson, R.R. 1983. Properties of a neutral allele model with intergenic recombination. *Theor. Popul. Biol.* **23**: 183–201.
- . 1987. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**: 245–250.
- . 1994. How can the low levels of DNA sequence variation in regions of the *Drosophila* genome with low recombination rates

- be explained? *Proc. Natl. Acad. Sci.* **91**: 6815–6818.
- Jorde, L.B., W.S. Watkins, M. Carlson, J. Groden, H. Albertson, A. Thlivers, and M. Leppert. 1994. Linkage Disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am. J. Hum. Genet.* **54**: 884–898.
- Kuhner, M.K., J. Yamato, and J. Felsenstein. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**: 1421–1430.
- Lai, C., R.F. Lyman, A.D. Long, C.H. Langley, and T.F. Mackay. 1994. Naturally occurring variation in bristle number and DNA polymorphisms at the *scabrous* locus of *Drosophila melanogaster*. *Science* **266**: 1697–1702.
- Lander, E.S. and N.J. Schork. 1994. Genetic dissection of complex traits. *Science* **265**: 2037–2048.
- Leitersdorf, E., A. Chakravarti, and H.H. Hobbs. 1989. Polymorphic DNA haplotypes at the LDL receptor locus. *Am. J. Hum. Genet.* **44**: 409–421.
- Long, A.D., M.N. Grote, and C.H. Langley. 1997. Genetic analysis of complex diseases. *Science* **275**: 1328.
- Long, A.D., R.F. Lyman, C.H. Langley, and T.F.C. Mackay. 1998. Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics* **149**: 999–1017.
- Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* **273**: 1516–1517.
- Slatkin, M. 1994. Linkage disequilibrium in growing and stable populations. *Genetics* **137**: 331–336.
- Spielman, R.S., R.E. McGinnis, and W.J. Ewens. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**: 506–516.
- Templeton, A.R. and C.F. Sing. 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134**: 659–669.
- Templeton, A.R., E. Boerwinkle, and C.F. Sing. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of *Alcohol Dehydrogenase* activity in *Drosophila*. *Genetics* **117**: 343–351.
- Wang, D.G., J. Fan, S. Chia-Jen, A. Berno, R. Lipshutz, M. Chee, and E. Lander. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.
- Watkins, W.S., R. Zenger, E. O'Brien, D. Nyman, A.W. Eriksson, M. Renlund, and L.B. Jorde. 1994. Linkage disequilibrium patterns vary with chromosomal location: A case study from the von Willebrand factor region. *Am. J. Hum. Genet.* **55**: 348–355.

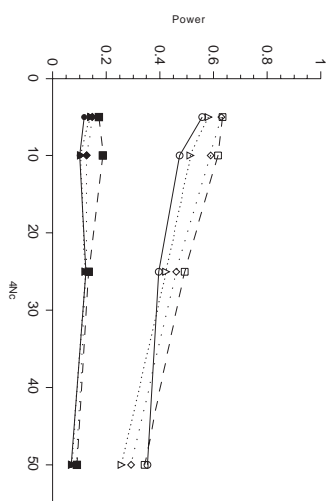
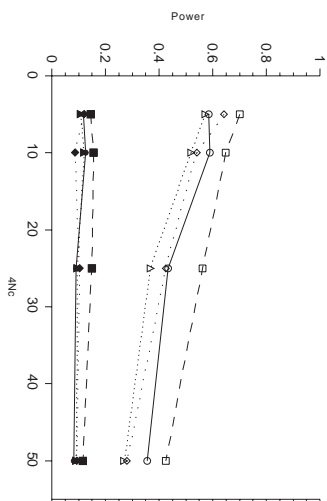
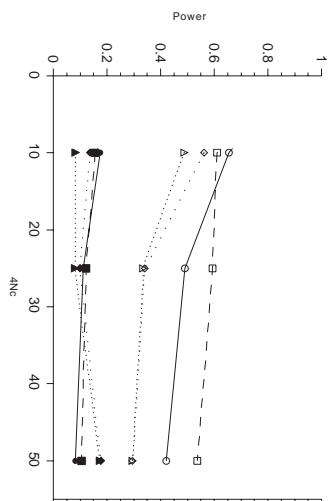
Received April 7, 1999; Accepted in revised form June 7, 1999.

# Number of Polymorphic Markers

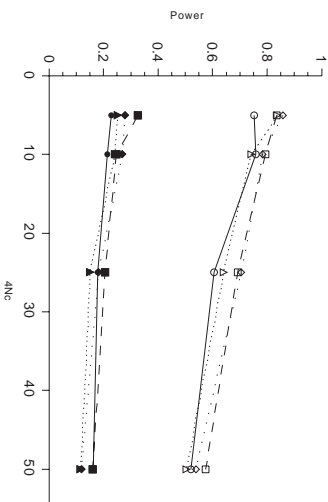
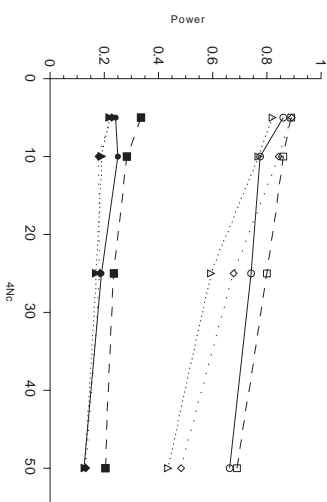
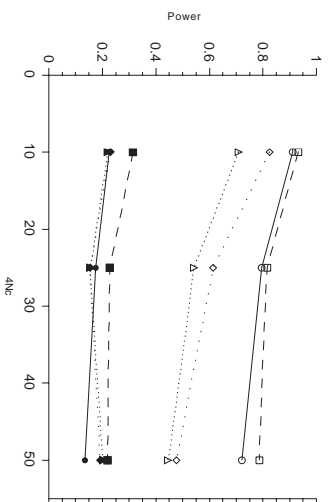
20-49

10-19

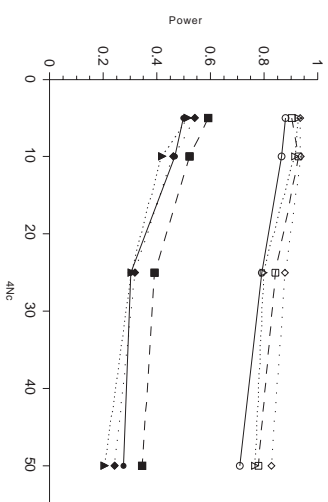
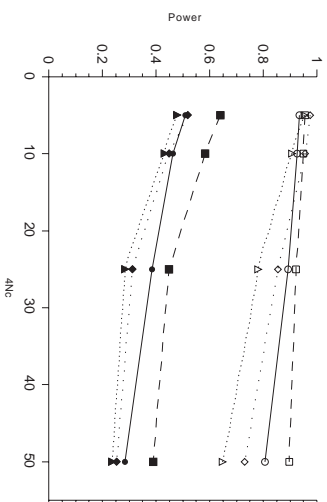
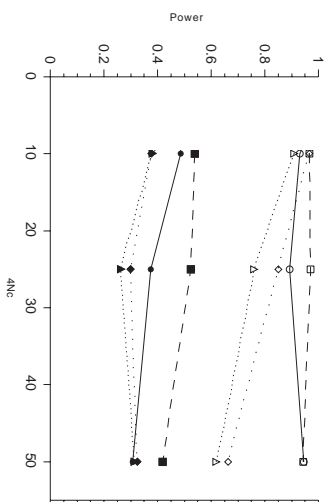
5-9



2.5%



5.0%



10.0%

Variance Due to QTN

