On the Power to Detect SNP/Phenotype Association in Candidate Quantitative Trait Loci Genomic Regions: A Simulation Study

J.M. Comeron, M. Kreitman, F.M. De La Vega

Pacific Symposium on Biocomputing 8:478-489(2003)

ON THE POWER TO DETECT SNP/PHENOTYPE ASSOCIATION IN CANDIDATE QUANTITATIVE TRAIT LOCI GENOMIC REGIONS: A SIMULATION STUDY

JOSEP M. COMERON[1], MARTIN KREITMAN

Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA
[1]Current address: Department of Biological Sciences, University of Iowa, 212 Biology Bldg. (BB), Iowa City, IA 52242, USA

FRANCISCO M. DE LA VEGA

Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA

We use coalescent methods to investigate the ability of linked neutral 'markers" to reveal in simulated population samples the presence of one or more single nucleotide polymorphisms that is contributing to a trait having a complex genetic basis (QTN: quantitative trait nucleotide). Realistic mutation and recombination rates in our simulations allow us to generate SNP data appropriate for analyzing human variation across short chromosomal intervals corresponding to approximately 100 kilobases. We investigate the performance of both single marker and multiple-marker (haplotype) data for several *ad hoc* procedures. Our results with single SNP markers indicate that (1) the density of SNP markers need not be much higher than 10% in order to achieve near-maximal detection of a QTN; (2) a higher density of markers does not improve much on the ability to localize a QTN within an interval unless the recombination rate is high. Haplotype-based tests were investigated for the case in which more than one QTN is present in the studied interval. Larger sample sizes improve both the probability of detecting the haplotype with the largest number of QTNs, as well as the ability to infer correct haplotypes from genotypic data. Testing a series of short haplotypes across a longer interval can also be beneficial. The rate of false positives (*i.e.,* when the most significant haplotype does not contain the greatest number of QTNs in the sample) can be very high when the contribution of individual QTNs to a trait is small. The elimination of low-frequency haplotypes does not substantially reduce the probability of detecting the haplotype with the largest number of QTNs but it can reduce the rate of false positives.

1 Introduction

Many diseases and traits are influenced by combinations of mutations acting at more than one site in the genome. The genes underlying these traits are generally referred to as Quantitative Trait Loci (QTL). The ultimate goal of association studies is to detect the presence of a causative mutation, deemed for simplicity a QTN (Quantitative Trait Nucleotide), by testing whether or not there is a difference in the frequency of individual mutations or haplotypes (*i.e.,* linked markers) associated with differences in phenotype. Using a population-based approach, we have studied and compared the power to detect marker/haplotype(s)-trait associations using different statistical methods, data analysis approaches and

experimental strategies. Following Long and Langley[1], our study uses coalescent theory to simulate selectively neutral SNP in a sample of chromosomes. This methodology uses realistic genetic and population parameters (neutral mutation rate, recombination rate and evolutionary effective population size) but assumes a somewhat simplified population structure and demographic history compared to those likely for human populations. The samples it generates are reasonable approximations of observed human variation with respect to density and number of SNPs, the frequency spectrum of these SNPs, and magnitude of linkage disequilibrium between them. Thus, despite simplifications in generating simulated samples, these *in silico* data should be useful for studying statistical properties of association measures applicable to human data. In all cases, we use phenotype distributions compatible with the subtle genetic and environmental effects expected for most QTLs.

2 General Methods

2.1 Simulations of neutral genealogies

The coalescent theory[2,3] provides an efficient framework for generating neutral population samples. These simulations assume a random mating population at equilibrium, neutrality and an infinite-sites model. Coalescent theory always looks backward in time, with lineages joining up ("coalescing") as a genealogy extends back into the past. Mutations are "sprinkled" along the lineages according to a Poisson process in proportion to branch lengths of the lineages, producing a sample of SNPs that obey the expected equilibrium distribution and frequency spectrum for selectively neutral mutations.

In our simulated samples we choose one SNP to be contributing to a phenotypic trait, and call it the QTN. That is to say, an individual carrying the derived mutation at this SNP site is given a phenotypic score that deviates from the mean by a certain amount. The magnitude of this deviation is described in the "Phenotypic Distribution" section below. One consequence of this scheme is that the frequency distribution of QTNs follows a neutral frequency spectrum, and will therefore be skewed towards low frequency variants.

The selectively neutral SNPs in the sample, excluding the QTN, are taken to be "markers". Then, by virtue of when each one occurred in the history of the genealogy relative to that of the QTN, which lineages they occurred on relative to the QTN, and how much recombination occurred between the marker and the QTN, each marker (and each haplotype) will have a certain informative value in predicting the presence of the QTN. The statistical properties of the associations between

markers and QTN are what we investigate. Many of our analyses of QTN - SNP associations use haploid data. Similar results are expected to obtain for diploids when there is no dominance. In our analyses of QTN – haplotype associations where haplotypes are reconstructed (inferred) from genotypic data, we form genotypes by randomly pairing chromosomes produced by the simulations.

Two properties of neutral samples are worth mentioning with implications for our study: 1) population genetics theory predicts that the expected mean frequency of a new mutation in a sample decreases with the number of sequences (n), and 2) the number of haplotypes—ordered combinations (phase) of genotypic variants (e.g., SNPs), which may or may not be closely linked or inherited together—constituted by a number of SNPs (S) in a sample is an increasing function of S, the number of sequences (n), and the recombination rate, but the number of expected haplotypes is *always* much smaller than 2^S.

2.2 Mutation and recombination for human populations

The ratio of polymorphisms or SNPs : recombination events has a strong influence on the power to detect SNP/phenotype associations. We used realistic values of mutation and recombination rates for human populations to assure the relevance and applicability of the simulation studies to complex traits in our species. The expected number of polymorphisms (S) per physical distance can be estimated from published studies of nuclear sequence variation[4]. These studies suggest an average of $\approx 1 \times 10^{-4}$ mutations or differences per site when comparing two randomly chosen sequences (popularly referred to as the nucleotide diversity per site). Nucleotide diversity may not be constant across all regions of the human genome, but a large fraction of the genome is expected to have densities of SNP near this average value. The number of SNPs in a sample, in contrast, is not constant, but is an increasing, nonlinear function of the sample size[5] (*n*); for instance, in a study of a 10 kb region, an average S≈60 and ≈36 is expected when the number of sequences is 200 and 20, respectively. The recombination rate in the humans varies across the genome; in regions of normal recombination 1cM (1% recombinants/generation) corresponds approximately to 1 Mb of DNA. The evolutionary relevant recombination rate between adjacent sites (4 *Ne c*) in humans might be on the order of ≈0.0001-0.001 in regions of low and high recombination, respectively. These are the range of values we investigate.

2.3 Phenotype distribution

A core element in this kind of study is the choice of an appropriate phenotype distribution congruent with QTLs. We used as phenotype distribution *Y* a modification of the distribution proposed in [1],

$$Y_i = z(1-\pi)^{1/2} + 1.96\, Q_i(\pi)^{1/2}$$

where Q_i=1, 0 represents presence and absence of the QTN in chromosome i, respectively, π is the proportion of phenotype variation attributable to the QTN, and z is a random normal deviate (mean=0, variance =1). Unlike the somewhat more elaborate formula for generating phenotypic scores given in [1] ours has no dependency on allele frequency[6,7]. Thus low- and high-frequency QTNs have the same average individual contribution to phenotype.

Most QTLs in humans might be compatible with our simulated scheme with π = 10-25%. Therefore, unless indicated, we have used for most of the subsequent analyses a conservative $\pi = 0.1$. The statistical power estimated will be conservative and, more importantly, the methodological approaches and experimental strategies shown to be adequate will be qualitatively accurate when the genetic contribution is stronger. At a practical level, for instance establishing the optimum density of markers in a study, our scheme might be easily modified based on external information about the phenotype distribution and the underlying genetics.

2.4 Statistical methods

Several authors have proposed the F-statistic ANOVA test (in particular a Model II ANOVA for two groups) to study association between phenotype and SNP variation[8,9]. We first compared this approach to nonparametric tests, namely the Mann-Whitney (MW) U-test and the Kruskall-Wallis (KW) H-test. In all cases, the significance of the estimated statistic is obtained by comparison to an empirically derived null distribution of this statistic in samples in which the phenotypic scores have been randomly permuted among individual sequences with the same number of markers. This assures that both multiple tests and the non-independence of tightly linked markers are taken into account. Our results reveal unambiguously that the two nonparametric tests (MW and KW) have greater power than the F-statistic for single QTNs with contribution to the phenotype (π) of 50% or lower. Therefore, we have used KW, which may be applied to both marker-based and haplotype-based association studies, in our analyses.

3 Results

3.1 Effect of single causative mutations (QTNs)

3.1.1 Optimal density of markers in the light of plausible recombination rates for humans

Clearly, increasing the density of marker SNPs within a candidate genomic region will increase the likelihood of including the QTN itself. On the other hand,

increasing the density of SNPs studied will increase the number of tests, hence reducing the probability of detecting a statistically significant association of any one marker with a QTN. This is especially true when the QTN has a subtle phenotypic effect unless the sample size is very large. Also, the low levels of effective recombination evident in human population genetic data further suggest that studies that exhaustively include every SNP in a region of interest may not increase dramatically the power compared to those studies analyzing only a fraction of the total variability.

To investigate these possible tradeoffs, we studied a case equivalent to analyzing a human genomic region of 100 kb and a sample of n=200, with average level of polymorphism (i.e., 600 SNPs), and for a range of recombination rates typical of the human genome. SNPs used in the analysis were randomly distributed across the region and only one of the 600 SNPs is the QTN. The results reveal that beyond 10%, increasing the density of studied SNPs will only slightly increase the overall power to detect genetic association between one SNP and the phenotype. That is, for realistic recombination rates observed in the human genome, 5-10 % density almost gives the maximum power, and greater power is achieved by increasing sample size than by increasing the density of markers. The optimal density mostly depends on the number of samples and the recombination environment of the candidate genomic region. Such a density is the one that might indicate that the genomic region under study likely includes a QTN although it does not necessary allow the precise specification of its location. Higher density of SNPs (such as complete resequencing) will increase the probability of localizing the QTN (although see below), especially in regions of high recombination.

3.1.2 Power *vs*. location

We investigated the average distance between the QTN and the SNP with the greatest (significant) association with the phenotype. Again, our simulation design assumes a genomic region with a total of 600 SNPs and the analysis of a varying percentage (density) of these SNPs (1 – 40%). In all cases under scrutiny the average distance between the 'significant' SNP and the QTN is considerable (e.g., always greater than 50 SNPs apart). Increasing the density of studied SNPs helps to locate the QTN with more accuracy, but this is mostly noticeable in regions of high recombination. Again, densities higher than 10-15% will not substantially increase the localization of the QTN, even in regions of high recombination, unless the sample size is very large.

Studies in regions of high recombination in the human genome will give overall reduced power to detect association between *any* SNP and phenotype variability compared to regions with low recombination. But when a significant association SNP/phenotype is detected, there is a higher probability that the detected SNP might be the QTN or close to it than when the region has low rates of recombination. This

result follows because tighter linkage will enhance the probability that SNPs with significant association may be more physically distant from the QTN. Moreover, the lower the recombination rate, the higher the probability of finding multiple SNPs with similarly high statistical significance, also due to tight linkage.

3.2 Effect of multiple causative mutations (QTNs)

So far, we have investigated statistical properties of markers associated with one and only one QTN, and with the ability to detect the QTN in a population-based sample. However, common disease may be influenced by combinations of causative mutations at a candidate locus or gene (multiple QTNs). In these cases, common sense dictates that association studies should focus on haplotype-based tests, as they can better capture the presence of these combinations of mutations. As indicated, the number of haplotypes depends on the number of sequences, the number of SNPs, the recombination rate, and population structure. We investigate the probability of detecting haplotypes with the largest number of QTNs in a sample (i.e, the most extreme haplotype). Note that because most SNPs segregate at low frequency, the number of QTNs in this most extreme haplotype usually will not represent the totality of QTNs present in the sample that might be influencing the phenotype, only a detectable subset. To carry out this study, we designated several SNPs in a region as QTNs, each contributing independently and equally to phenotype.

In diploid organisms the ability to detect association between individual haplotypes and phenotype will be influenced by our ability to discern the haplotype structure from heterozygous individuals. Further, in individuals with heterozygous QTNs, the phenotype is most likely to be less conspicuous (we assume additivity). As a first approach, we compare the probability of detecting the most extreme haplotype in three cases: 1) in a haploid case, 2) in a diploid case where haplotypes are known with certainty, and 3) in a diploid cases where no effort is made to discern the actual haplotypes, and hence haplotypes are constructed randomly from the genetic information. The results show that increasing the sample size causes (as expected) an overall increase in the probability of detecting a significant association for the haplotype with the most QTNs. The results also reveal that the difference between knowing or not the actual haplotype becomes critically important as sample size increases. In other words, in order to have a high probability of detecting association, a large sample is required and diploid genotypes need to be resolved as haplotypes (either by inference or by experiment).

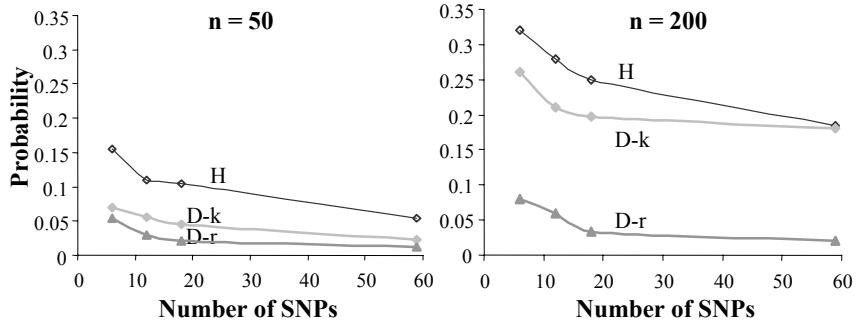

Figure 1. Probability of detecting the haplotype with the largest number of QTNs (i.e, the most extreme haplotype) as a function of the number of SNPs under study. 3 QTNs are segregating in the sample, each with a contribution to phenotype (π) of 25%. Results are shown for the case of no recombination when n= 50 and 200. Three cases are compared: Haploid (H), Diploid with known (D-k) haplotypes, and Diploid with randomly constructed (D-r) haplotypes.

3.2.1 Inference of haplotypes from diploid populations

We investigated a widely used method proposed by A. Clark[10]. Clark's method uses a parsimonious approach to infer the minimum number of haplotypes in a sample: it utilizes information from homozygous or single-site heterozygous individuals to sequentially resolve multiply-heterozygous genotypes into haplotypes. Figure 2 shows two extreme cases. Clark's method always performs well when the number of SNPs is small.

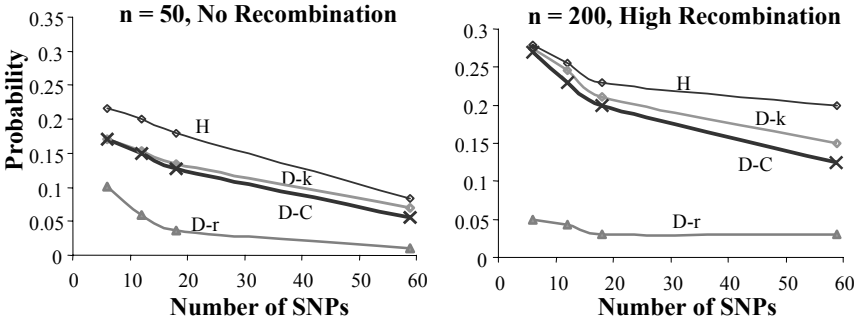

Figure 2. Probability of detecting the haplotype with the largest number of QTNs as a function of the number of SNPs under study. 3 QTNs are segregating in the sample, each with a contribution to phenotype (π) of 25%. Results are shown for four cases: Haploid (H), Diploid with known (D-k) haplotypes, Diploid with randomly constructed (D-r) haplotypes, and Diploid after applying Clark's method[10] (D-C) to infer haplotypes.

As expected, recombination produces unresolved or wrongly inferred haplotypes, a problem that is enhanced by increasing the number of SNPs. The 'sequential' problem (i.e., the fact that a different order in the sequential solution of haplotypes might cause different solutions) also increases with the number of SNPs. Large sample sizes increase the power to detect associations and also the probability of observing homozygous individuals, required in Clark's method. But large sample sizes will also increase the probability of observing individuals with two or more SNPs present only once in the sample, causing unresolved haplotypes (but see below the small practical consequences of this problem).

Another method, proposed by Stephens, Smith and Donnelly[11] applies a Markov-chain Monte-Carlo (MCMC) algorithm with population genetics assumptions. The results (data not shown) indicate that for the conditions studied in Figure 2, it performs similar or a little worse than Clark's method when recombination occurs and the number of SNPs is high.

3.2.2 Number of SNPs used to define haplotypes: Partial haplotypes

We investigated the effect of haplotype 'size' --the number of adjacent SNPs used to define a haplotype-- on the probability of detecting significant association. *A priori*, we expect that the use of more SNPs will increase the likelihood of haplotypes including more QTNs. Also, the use haplotypes based a small number of markers (henceforth called 'small' haplotypes) will increase substantially the number of tests performed with the consequent reduction of statistical power. We studied a scenario equivalent to a situation in humans in which all QTNs are restricted to 1 kb (≈ 6 SNPs) but SNP markers are dispersed across a 10 kb interval (59 SNPs).

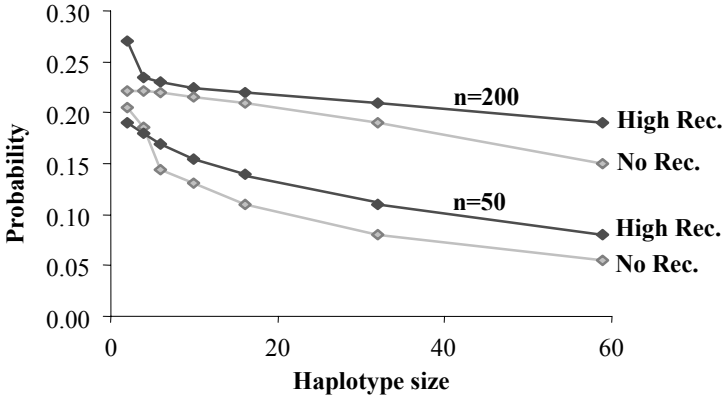

Figure 3. Relationship between the number of adjacent SNPs used to construct the haplotypes and the probability of detecting the haplotype with the largest number of QTNs. 5 QTNs are segregating in the sample, each with a contribution to phenotype (π) of 25%.

The use of haplotypes constructed by a small number of SNPs increases the probability of detecting the haplotype with the highest number of QTNs. Interestingly, the number of adjacent SNPs used to construct the haplotype with the highest probability of detecting the most extreme haplotype is smaller than the actual number of QTNs, in agreement with the idea that it is highly unlikely to observe a haplotype with all QTNs. Note also that small haplotypes are those more accurately predicted by most algorithms.

The use of haplotypes constructed only by a subset of adjacent SNPs (partial haplotypes) can be put to good advantage in surveys of longer regions by utilizing a 'sliding window' analysis across these regions (or a 5'- vs central vs 3' study). A sliding window approach might better localize the region encompassing QTNs since it would give a quantitative idea of the signal observed in the 'background'. For instance, in a study of 200 sequences with five adjacent QTNs, partial haplotypes give 17% significant detection (false positive) when they are located 50 SNPs apart from the QTN positions when recombination is high, while this percentage jumps to 23% around the QTN positions. Overall, the higher the recombination rate or the more distant the regions, the higher the chance of detecting differences between regions, and hence of localizing the region with more QTNs

3.2.3 Probability of false positives

We studied the probability that the haplotype showing the strongest significant association with phenotype variability is not the one with the highest number of QTNs.

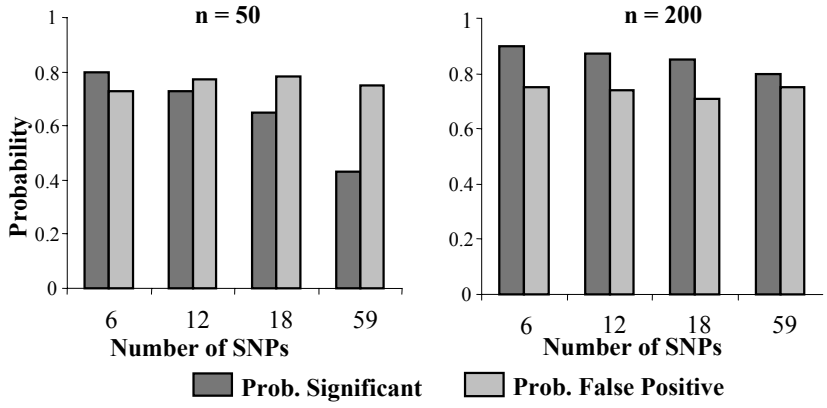

Figure 4. Probability of detecting a significant association between any haplotype and phenotype variability and the probability that the detected haplotype is not the one with the highest number of QTNs in the sample (false positives). 5 QTNs are segregating in the sample, each with a contribution to phenotype (π) of 25%.

As shown in Figure 4, the probability of false positives is very high for QTNs with small contribution to phenotype. Increasing the sample size improves the probability of detecting significant associations between haplotype and phenotype. But the probability of detecting the haplotype with more information about the causative mutations increases less rapidly.

3.2.4 Haplotype frequency

As discussed above, the highest probability of detecting significant association between SNP (or haplotype) and phenotype is attained when QTNs are at intermediate frequency. This is also true when we take into consideration the frequency of false positives, since higher QTN frequencies in the sample does increase the probability of detecting a significant association, but it does not alter the probability of false positive (data not shown). Hence, the use of a sample with high-frequency QTNs will also increase the reliability of the results. We have studied whether the elimination of haplotypes at low frequency in the sample also reduces the probability of false positives.

Indeed, as suspected the elimination of haplotypes at low frequency substantially reduces the probability of false positives. Eliminating low frequency haplotypes has a practical advantage as well: haplotypes at low frequency are also more difficult to infer from diploid data.

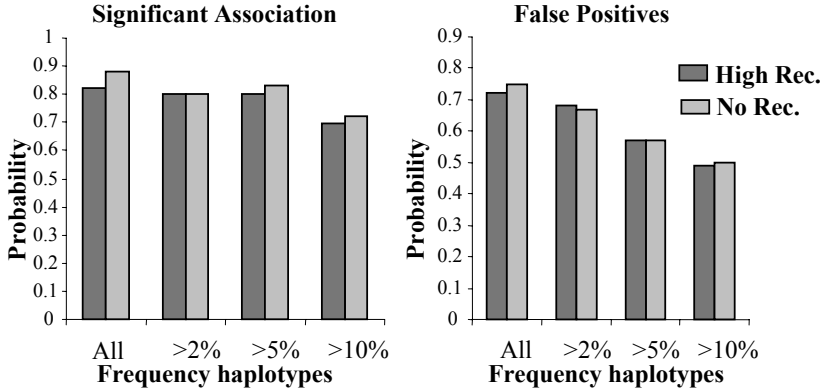

Figure 5. Probability of detecting significant association between haplotype presence and phenotype variability and of false positives (among those cases with the strongest significant association) when the haplotypes used in the analysis are those at frequency higher than 2%, 5% or 10% in the sample. n = 200.

4 Conclusions

Overall, the study shows that the probability of detecting a significant association between nucleotide and phenotype variability is low for most conditions suitable to most QTLs (π<50%). The low, albeit variable, rate of recombination present in the human genome also contributes to a very high percentage of false positives, a percentage that decreases with recombination. On the other hand, empirical investigation of linkage disequilibrium in the human genome suggests a strong haplotype structure, possibly caused by recombination cold- and hot-spots[12]. If true, the study of only a small percentage of all SNPs present in a genomic region gives almost the maximum power to detect association, and greater power is achieved by increasing sample size than by increasing the density of markers. The use of nonparametric tests, the study of extreme phenotypes, and the analysis of common haplotypes based on a small number of adjacent SNPs are all methodologies that increase the chance of detecting and locating a QTN.

The recent molecular evolutionary history of humans almost certainly includes intense selection on many sites across the genome, and many QTNs may be selected mutations. Such a scenario, taken together with relatively low recombination rates, implies a high probability of false positives, a problem that might be exacerbated by population expansion. Overall, recent selection and/or population expansion makes the problem of discerning a QTN among all mutations or SNPs across a small genomic region (i.e., the same exon or gene) a more difficult task. As shown here, some analytical and experimental approaches can improve the chance of being successful. Nothing, however, is likely to overcome the need for very large population sample sizes in order to achieve reasonable power and acceptably low levels of false positives when scanning the whole genome. This will place additional incentive for commercial development of lower-cost, higher-throughput SNP assays.

References

1. A. D. Long, and C. H. Langley, "The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits" *Genome Res.* **9,** 720 (1999)
2. J. F. C. Kingman, "The coalescent" *Stoch. Proc. Appl.* **13,** 235 (1982)
3. R. R. Hudson, "Gene genealogies and the coalescent process" *Oxf. Surv. Evol. Biol.* **7,** 1 (1990)
4. M. Przeworski, R. R. Hudson, and A. Di Rienzo, "Adjusting the focus on human variation" *Trends Genet.* **16,** 296 (2000)
5. G. A. Watterson, "On the number of segregating sites in genetical models without recombination" *Theor. Popul. Biol.* **7,** 256 (1975)

6. D. E. Reich, and E. S. Lander, "On the allele spectrum of human disease" *Trends Genet.* **17,** 502 (2001)
7. J. K. Pritchard, "Are rare variants responsible for susceptibility to complex disease?" *Am. J. Hum. Genet.* **69,** 124 (2001)
8. G. A. Churchill, and R. W. Doerge, "Empirical threshold values for quantitative trait mapping" *Genetics* **138,** 963 (1994)
9. A. D. Long, R. F. Lyman, C. H. Langley, and T. F. Mackay, "Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in Drosophila melanogaster" *Genetics* **149,** 999 (1998)
10. A. G. Clark, "Inference of haplotypes from PCR-amplified samples of diploid populations" *Mol. Biol. Evol.* **7,** 111 (1990)
11. M. Stephens, N. J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data" *Am. J. Hum. Genet.* **68,** 978 (2001)
12. S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, *et al*., "The Structure of Haplotype Blocks in the Human Genome" *Science* **296**, 2225 (2002)