DOI: 10.1534/genetics.105.047472

Inferring Population Parameters From Single-Feature Polymorphism Data

Rong Jiang,* Paul Marjoram,† Justin O. Borevitz‡ and Simon Tavaré*,§,1

**Molecular and Computational Biology Program, University of Southern California, Los Angeles, California 90089, †Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, ‡Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 and §Department of Oncology, University of Cambridge, Cambridge CB2 2XZ, England*

Manuscript received June 29, 2005
Accepted for publication May 3, 2006

ABSTRACT

This article is concerned with a statistical modeling procedure to call single-feature polymorphisms from microarray experiments. We use this new type of polymorphism data to estimate the mutation and recombination parameters in a population. The mutation parameter can be estimated via the number of single-feature polymorphisms called in the sample. For the recombination parameter, a two-feature sampling distribution is derived in a way analogous to that for the two-locus sampling distribution with SNP data. The approximate-likelihood approach using the two-feature sampling distribution is examined and found to work well. A coalescent simulation study is used to investigate the accuracy and robustness of our method. Our approach allows the utilization of single-feature polymorphism data for inference in population genetics.

NATURAL variation is of great interest to a variety of disciplines, such as evolutionary biology, plant and animal breeding, human genetics, and population genetics. For example, it enables us to study the effects of genetic forces, infer population history, and understand the genetic basis of complex traits. There are several types of genetic data available to study natural variation, including restriction fragment length polymorphisms (RFLPs), microsatellites, and single nucleotide polymorphisms (SNPs).

Microarrays, a high-throughput technology for genomic DNA/RNA hybridization, play a key role in functional genomics. Two major applications of DNA microarrays are expression analysis and genotyping. With the development of microarray technology and design, more accurate and robust statistical methods are needed to analyze the massive amount of data produced. Detailed examples may be found in Parmigiani *et al.* (2003) and Speed (2003), for example.

As a model plant, *Arabidopsis thaliana* has been extensively investigated in areas such as plant physiology and crop breeding. Borevitz *et al.* (2003) used Affymetrix arrays to detect the genomic DNA signal in *A. thaliana* and conducted genomewide studies such as gene clustering, deletion identification, and quantitative trait loci (QTL) mapping. Along with new developments of array design, other *A. thaliana* accessions have been studied. Wolyn *et al.* (2004) studied the variation between the accession Kas and the reference Col to map the light response QTL. Werner *et al.* (2005) studied two accessions, Bur-0 and Lz-0, for F_2 extreme array QTL mapping.

The scaled mutation parameter θ and the scaled recombination parameter ρ are important quantities in population genetics studies. Watterson (1975) derived a point estimator of θ based on the number of segregating sites, assuming the infinitely many sites mutation model. For ρ the situation is more complex, but there are several approaches using SNP data, for example, rejection-based methods using summary statistics (Wall 2000) and approximate-likelihood methods (Hudson 2001; Li and Stephens 2003; McVean *et al.* 2004).

In this study, we develop methods to model and analyze single-feature polymorphism (SFP) data. Our focus is on estimating θ and ρ, which will help us explore and utilize this new type of polymorphism data in more general settings. First we revisit the microarray experiment in Borevitz *et al.* (2003) and describe in detail how we model the statistical procedure to call features. We show how the number of features called can be used to estimate θ. An approximate-likelihood approach is proposed to estimate ρ by deriving the two-feature sampling distribution by analogy with the two-locus sampling distribution in Hudson (2001). We assess the accuracy of this estimator using coalescent simulations. Finally, we investigate statistical properties of our maximum-likelihood estimator of ρ and check the robustness of our approach by varying demographic parameters. We also discuss how much information is lost in SFP data compared with SNP data.

[1]*Corresponding author:* Molecular and Computational Biology Program, University of Southern California, 835 W. 37th St., SHS172, Los Angeles, CA 90089-1340. E-mail: stavare@usc.edu

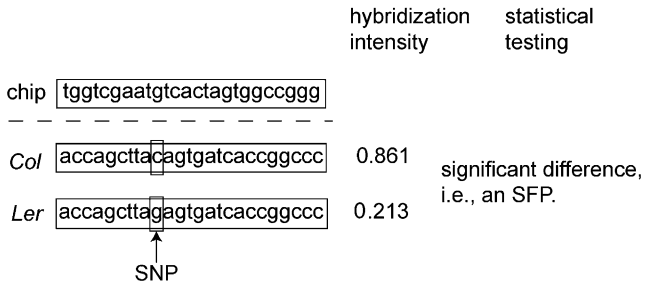

FIGURE 1.—Calling an SFP between the reference Col and the strain L*er*. The chip probe is complementary to the corresponding reference target.

MODELING

Motivation: BOREVITZ *et al.* (2003) studied allelic variation in ecotypes of *A. thaliana*. DNA was prepared independently from three Col and three L*er* plants and hybridized to six Affymetrix expression arrays. We refer to Col as the reference type and L*er* as the accession. The arrays were set up to match the reference genome sequence. After the arrays were scanned, the mean intensity of each probe was spatially corrected, quantile normalized, and analyzed by the SAM method (TUSHER *et al.* 2001). A probe was called as an SFP if its hybridization intensity was detected as being significantly different from that for the corresponding reference probe. An illustration of SFP calling between two strains is given in Figure 1. In BOREVITZ *et al.* (2003), 3806 SFPs were called out of 92,924 unique probes and these were used later for detecting potential deletions and for bulk segregant analysis.

Although comparing accessions one at a time against the reference provides only pairwise diversity, we can combine these pairwise comparisons and produce SFP data across multiple accessions in a similar way to that in which we call segregating sites: a feature is called polymorphic if there is at least one accession probe at this position whose hybridization intensity is called statistically significantly different from that of the reference probe. We can model this scheme exactly in our simulations. At this time, 19 accessions are examined in the Borevitz lab and the SFP data from this group of accessions plus the reference strain serve as the basis for our inference of population history for *A. thaliana*.

We now revisit the procedure for obtaining the SFP data. There are two main steps in the microarray experiments: measuring hybridization intensity for each probe and detecting significant differences of hybridization intensity with respect to the reference, as illustrated in Figure 1. The latter is a purely statistical procedure and we do not attempt to model it at the sequence level. Nevertheless, we use independent sequence confirmation of called SFPs to quantify this step and include it in our simulations. Below, we describe in detail how to model the SFP-calling procedure in coalescent simulations.

The coalescent: Pioneered by KINGMAN (1982), HUDSON (1983), and TAJIMA (1983), coalescent theory continues to prove a useful model for studying neutral population history. In particular, fast and efficient coalescent simulations help us to test the fit of different models to data, especially when analytical results are hard to obtain. For reviews of coalescent theory and related inference methods, see NORDBORG (2001), for example. We use the coalescent to model the genealogy of the accessions and the reference.

The issue of sample size and selection of the reference sample requires some thought in coalescent simulations. Suppose that 19 accessions plus the reference (Col) are in the microarray experiment. We simulate only one haplotype for each ecotype (either accession or reference), since *A. thaliana* is highly selfing and therefore its genome is essentially homozygous (see NORDBORG *et al.* 2005 for a discussion of this issue). Therefore, we set the sample size as 20 when we deal with 20 different ecotypes. Moreover, the reference is chosen randomly among the simulated haplotypes, since the strain Col is chosen as the reference merely because it was the first strain completely sequenced. In other words, there is nothing special about this strain compared with other strains. Last, the sample of haplotypes is simulated using the ms program (HUDSON 2002) and then postprocessed to obtain SFP data. We simulate samples without population structure.

SFP comparison phase: For the simulated data, we are able to measure the sequence polymorphism at each probe position for every accession compared to the reference, instead of needing to measure hybridization intensity in microarray experiments. For simplicity, we restrict our mutation model to consider only single-nucleotide polymorphisms. There are several types of sequence polymorphism within a 25-bp probe when comparing an accession to the reference. We classify them as follows: no difference, one-SNP difference, and more-than-one-SNP difference. Noting that the chance of having more than one SNP within a 25-bp probe is rather small, we take it as an assumption later and categorize the sequence polymorphism as either "no difference" or "one-SNP difference."

In simulations, each accession is compared with the reference at every probe position. If in fact there is more than one SNP in a given probe, it is treated as having a one-SNP difference. Consequently, every probe of each accession is labeled 0 (no difference) or 1 (one-SNP difference) after comparisons.

SFP-calling phase: In microarray experiments, high or low hybridization intensities are called significant by a prespecified statistical procedure, involving results from multiple hypothesis testing. It is not easy to model this step without taking into account the array setup and the distribution of test statistics. On the other hand, we can characterize the statistical procedure by summarizing the four types of possible outcome: true positives, true

negatives, false positives, and false negatives. Borevitz *et al.* (2003) confirmed the SFPs called by comparing available sequence data between the accession (L*er*) and the reference (Col). They found that there were 117 SFPs called out of 340 polymorphic probes (*i.e.*, true positives), and 4 SFPs called out of 477 nonpolymorphic probes (*i.e.*, false positives). Here, polymorphic probes refer to those accession probes that differ from the reference probe. We use these numbers to give us the rates at which false positives (negatives) occur in our modeling of the SFP-calling procedure.

Different calls can be characterized by two probabilities: *sensitivity* and *specificity*, which sometimes are called the *true positive rate* and the *true negative rate*, respectively. Another useful probability is the false discovery rate (FDR) (*cf.* Storey and Tibshirani 2003). For example, we have a sensitivity of 0.34, specificity of 0.99, and FDR of 0.03 in the above scenario. Unlike the multiple-hypothesis testing situation in microarray experiments, in our simulation scheme we know the exact truth about the polymorphism at every probe position. Therefore, we can make calls of SFPs according to the sensitivity and specificity appropriate for real experiments by assuming that each call is independent and identically distributed. Furthermore, as SFPs are quantitative we can evaluate the effect of different thresholds (trading specificity for sensitivity, for example) on estimates of population genetics parameters.

After the SFP comparison phase, each accession can be viewed as a series of 0's (no difference) and 1's (one SNP difference). If we focus on one particular probe, we obtain a column of 0's and 1's, where the reference is always labeled 0. After statistical calls, we label an accession probe 1 if it is called polymorphic with respect to the reference probe; otherwise we label it 0. In this way, we can restate the four types of calls: true positives as $1 \to 1$, true negatives as $0 \to 0$, false positives as $0 \to 1$, and false negatives as $1 \to 0$, where the first number indicates no difference or one-SNP difference in the true sequence comparison and the second number indicates "called polymorphic" or "called nonpolymorphic" by the analytic software. The probabilities for the calls can be calculated on the basis of the sensitivity and specificity. We say that there is an SFP at the current probe position if not all accession probes are labeled 0 (by analogy with the definition of segregating sites).

Simulations: In coalescent simulations, we simulate a sample of 20 chromosomes (haplotypes) with scaled mutation parameter θ and recombination parameter ρ.

In Borevitz *et al.* (2003), the arrays were defined using probes for the reference strain. Probes were clustered at the 3′ end of known and predicted genes. Later on, they used a new technology to tile more probes on arrays and hence enlarge the probe coverage over the genome. For simplicity in both the analysis and the simulation study, we consider only nonoverlapping probes, and we generate evenly distributed probes with a coverage of 3.3%, the same as that in the experiment in Borevitz *et al.* (2003). Moreover, we take the length of the region to be 100,000 bp, the largest value we can use while still allowing for fast simulations in our inference method.

In the SFP-calling phase, we denote the sensitivity by s_n and the specificity by s_p. Because SFP genotyping is quantitative, various stringencies are explored. We consider three combinations of (s_n, s_p): (0.34, 1.0), (1.0, 1.0), and (0.68, 0.75). The first one is similar to the estimates from the sequence confirmation in Borevitz *et al.* (2003); the second one is the ideal case; and the last one considers intermediate values of sensitivity and specificity.

Assumptions: To simplify the analysis while still modeling the essentials of the microarray experiments, we make several assumptions, some of which are already described above. Here we emphasize some important assumptions in modeling the statistical calling procedure. First, calls are independent across the probes for each accession. Second, at every probe position, we propose two alternatives for dealing with calls across accessions: (1) *independent* calling, *i.e.*, each accession probe is called independently from other probes; and (2) *dependent* calling, *i.e.*, samples with the same true state have the same call across sequences for a given probe. An example illustrating the calling of SFPs is given in the appendix.

The intuition for independent calling comes directly from the setup of the microarray experiments, in which each accession is hybridized to the reference probes. On the other hand, dependent calling comes from the biological intuition that probes of the same sequence tend to have similar hybridization intensities as well as similar calls. The truth is likely somewhere between the two.

SOME ANALYTICAL RESULTS

Expected number of SFPs: Our main result about the expected number of SFPs is that it is approximately linear in the mutation parameter θ. As a result, the number of SFPs called serves as a good candidate summary statistic to estimate the mutation parameter. Standard coalescent simulations show that the probability of having more than one SNP within a 25-bp probe is rather small for small θ, *e.g.*, $\sim$2% for $\theta = 2$/kb and 5% for $\theta = 4$/kb. Assuming the above, we can derive an approximate formula for the expected number of SFPs, given the experimental parameters such as the sensitivity and specificity, dependent/independent calling, etc. See the appendix for more details.

We plot the expected number of SFPs *vs.* the mutation parameter in Figure 2, using probe coverage of 3.3%. We present four cases: different combinations of sensitivity and specificity in dependent calling (cases 1, 2, and 3) and independent calling (case 4). When the mutation parameter is small the simulation results fit well

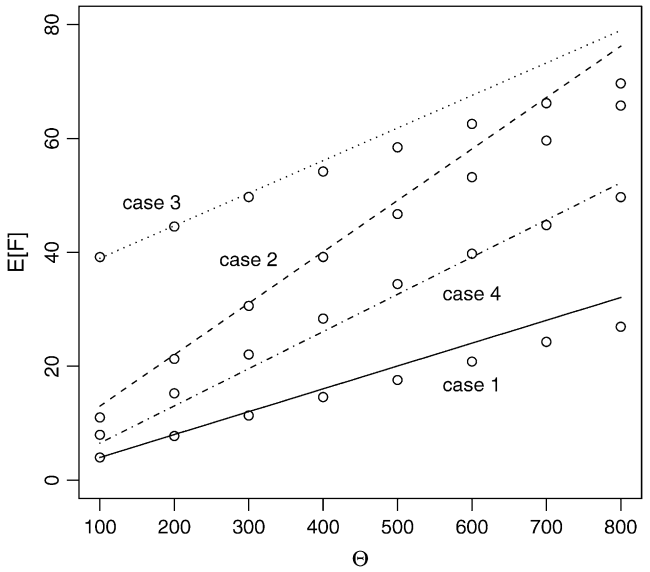

FIGURE 2.—The expected number of SFPs, $\mathbb{E}[F]$, plotted against θ for 3.3% probe coverage. Circles denote the estimated values from simulations. Theoretical lines are drawn according to formula (A6) (see the APPENDIX). Four cases are shown: dependent calling with (s_n, s_p) as (0.34, 1.0) (case 1, solid line), (1.0, 1.0) (case 2, dashed line), and (0.68, 0.75) (case 3, dotted line) and independent calling with (s_n, s_p) as (0.34, 1.0) (case 4, dashed-dotted line).

with the theoretical predictions for each case. The independent calling case with ideal sensitivity and specificity (both = 1.0) is not shown because it is essentially the same as dependent calling case 2. The independent calling case with sensitivity as 0.68 and specificity as 0.75 is not shown because the number of SFPs called in every data set is equal to the number of probes, since the probability to call each probe as an SFP is ≈1 when $s_n = 0.68$ and $s_p = 0.75$ in a sample of size 20. In other words, such a poor specificity introduces too many false positives, which makes the number of SFPs uninformative for θ.

Two-feature sampling distribution: HUDSON (2001) studied the two-locus sampling distribution and derived an approximate-likelihood approach to estimate ρ using SNP data. We denote by $\{A_1, A_2, \ldots, A_S\}$ the configuration vector at the segregating sites 1, 2, …, *S*, where *S* is the total number of SNPs. Analogously, we denote by $\{B_1, B_2, \ldots, B_F\}$ the configurations in the SFP sample, B_i being the configuration vector for the *i*th SFP called, and *F* is the total number of SFPs called. We derive an approximate likelihood for our SFP sample and use a maximum-likelihood approach to infer ρ. For simplicity, we derive the two-feature sampling distribution under dependent calling. (The derivation under independent calling is straightforward, although there are more possibilities involved. We do not present it here.)

First, we review the two-locus sampling approach. HUDSON (2001) considers pairs of segregating sites, whose configuration is denoted by $\mathbf{n} = (n_{00}, n_{01}, n_{10}, n_{11})$, where n_{ij} is the number of times the combination *ij* ($i = 0, 1$ and $j = 0, 1$) is observed at the two given segregating sites. Recall that 0 stands for wild type and 1 for mutant in the SNP matrix. The probability of this pair conditional on both loci being segregating is denoted by $q_c(\mathbf{n}; \rho, \theta)$, and the limit for small mutation rate is denoted by $q_c(\mathbf{n}; \rho) = \lim_{\theta \to 0} q_c(\mathbf{n}; \rho, \theta)$. The approximate likelihood of the sample is then given by

$$L_s(\rho) = \prod_{i=1}^{S-1} \prod_{j=i+1}^{S} q_c(\mathbf{n}_{ij}; \rho), \tag{1}$$

assuming that pairs are independent of each other.

Similarly, we define $p_c(\mathbf{m}; \rho, \theta)$ to be the probability of a pair of SFPs, **m**, having configuration (m_{00}, m_{01}, m_{10}, m_{11}), with similar interpretation to that in the SNP setting. Here the subscript c indicates that it is conditional on two features with one segregating site in each. Unlike the derivation of the two-locus sampling distribution, there are several possible two-locus SNP pairs (**n**) that give rise to the same two-feature SFP pair (**m**). These are illustrated in the APPENDIX.

By summing over all possibilities of SNP configurations, reference types, and calling operations we obtain

$$p_c(\mathbf{m}; \rho) = \sum_{\mathbf{n}} \sum_{r} q_c(\mathbf{n}; \rho) \mathbb{P}(r) b(\mathbf{m} \mid \mathbf{n}, r). \tag{2}$$

Here *r* denotes the reference type, *i.e.*, 00, 01, 10, or 11, so $\mathbb{P}(00) = n_{00}/s$, where *s* is the sample size, for example. Moreover, $b(\mathbf{m} \mid \mathbf{n}, r)$ is the probability of obtaining the SFP pair **m** from the SNP pair **n** and reference type *r*. Note that if the SNP pair **n** can be transformed to the SFP pair **m** then in general there is one and only one combination of calling operations at the two probe positions. Its probability is given by $\mathbb{P}(C_i)\ \mathbb{P}(C_j)$ if the calling operation at the first site is of type C_i and the one at the second site is of type C_j, *i*, *j* = I, II, or III. For example, $\mathbb{P}(C_1) = \mathbb{P}(1 \to 1, 0 \to 0) = \mathbb{P}(1 \to 1)\mathbb{P}(0 \to 0) = s_p s_n$.

We can now write the approximate likelihood of the SFP sample as

$$L_f(\rho) = \prod_{i=1}^{F-1} \prod_{j=i+1}^{F} p_c(\mathbf{m}_{ij}; \rho), \tag{3}$$

where $\mathbf{m}_{ij}$ is the configuration vector for the pair of the *i*th feature and the *j*th feature. Once the two-feature sampling probability $p_c(\mathbf{m}; \rho)$ is tabulated, we can compute the approximate likelihood rapidly for a grid of ρ-values and report the maximum-likelihood estimator. Note that the above approximate-likelihood calculation assumes independent pairs of features, while HUDSON (2001) assumes independent pairs of segregating sites.

Modification of the two-feature sampling distribution: Given the SFP data alone, we are not able to tell if

there is a SNP within a particular feature or not, due to the existence of false positives. However, by studying the pattern of the feature pairs where there is no SNP inside (*i.e.*, false positives), we can modify the above two-feature sampling distribution.

Under the completely dependent calling assumption, the false positives have one common pattern in the resulting SFP matrix, *i.e.*, all 1's but one 0 for the reference probe. Consequently, to help avoid false positives being included in the likelihood computation, we require that there are at least two 0's in the corresponding column in the resulting SFP matrix for every feature considered.

We can readily utilize this to find estimators of ρ. We label this approach *modified*. At the same time, we also compute approximate maximum-likelihood estimates (MLEs) from two other approaches using SFP data: one using all SFPs called without any restriction (labeled *simple*) and the other using only SFPs having exactly one SNP inside (labeled *real*). Comparing these two with the modified likelihood approach, we show that the modified version yields better performance and good robustness under various combinations of sensitivity and specificity.

RESULTS

In this section we provide results from simulation studies.

Estimating θ using the number of SFPs: Here we use an approximate-likelihood approach similar to that of Weiss and von Haeseler (1998). We are interested in estimating the mutation parameter θ on the basis of the observed number, $F = f$, of called SFPs. For a particular value θ we generate a large number of coalescent replicates and approximate the likelihood $L(\theta)$ by the proportion of replicates that have approximately f called SFPs,

$$L(\theta) \approx B^{-1} \sum_{i=1}^{B} I(F_i, f),$$

where B is the number of replicates, F_i is the summary statistic in the ith run with parameter θ, and the indicator $I(F_i, f)$ is 1 if $|F_i - f| < \delta$. In estimating θ we set $B = 250{,}000$ so that we obtain reliable estimates of the likelihood. We repeat this scheme for a range of θ-values chosen on an evenly spaced grid and then report the one with the maximal (approximate) likelihood as the estimate of the true θ. The results depend on the predetermined tolerance δ. In this approach the recombination rate ρ is a nuisance parameter. We accommodate this by simulating each data set with ρ chosen uniformly over [0, 100].

To assess the adequacy of this procedure, we consider three combinations of (θ, ρ), namely (200, 20), (400, 40), and (600, 60). For each (θ, ρ) combination, we simulate 1000 data sets and record the summary statistic f for each of them. Then we apply our method to obtain the corresponding estimate of θ. For each parameter set, we calculate the mean and the root (relative) mean square error (RMSE) to assess the performance of our approach. Table 1 shows the results, assuming 3.3% probe coverage.

TABLE 1

Estimates of θ

Calling	(s_n, s_p)	$\theta_0 = 200$	$\theta_0 = 400$	$\theta_0 = 600$
Dependent	(1.0, 1.0)	(200, 0.29)	(401, 0.23)	(596, 0.18)
Independent	(0.34, 1.0)	(200, 0.33)	(396, 0.25)	(597, 0.23)
Dependent	(0.34, 1.0)	(205, 0.46)	(409, 0.31)	(608, 0.28)
Dependent	(0.68, 0.75)	(203, 0.59)	(408, 0.34)	(613, 0.28)

The two numbers within the parentheses denote the mean and the root (relative) mean square error (RMSE) from the sample of posterior estimates, respectively. Probe coverage is 3.3%.

Overall, the estimates in Table 1 show no bias. However, for a given specificity the smaller the sensitivity is, the larger the RMSE becomes. Moreover, the specificity at a level of 0.75 has the largest RMSE and decreases the power of the inference method. Also we note that the RMSE tends to be smaller as θ increases, but it is not clear why the independent case with sensitivity 0.34 and specificity 1.0 has smaller RMSE than the corresponding dependent case.

Estimating ρ using two-feature sampling distributions: Once more we compared three scenarios, $(\theta, \rho) = (200, 20)$, $(400, 40)$, and $(600, 60)$, and simulated 1000 data sets for each combination. For each simulation, we calculate the composite approximate likelihood of the SFP sample using Equation 3 for a range of ρ-values chosen on an evenly spaced grid between 0 and 100 and report the one with maximal likelihood as the estimate of ρ. We report the mean, the RMSE, and the fraction of times the estimate falls within a factor of two of the true parameter (*cf.* Wall 2000). We see from Table 2 that we obtain reasonable estimates of ρ. Due to many false SFPs being called, the simple approach underestimates ρ when the specificity is moderately large, *e.g.*, 75%. On the other hand, the real approach is impossible to apply in reality since the state of SNPs within each called feature is not observed. The modified approach excludes false positives and hence makes the estimator more robust to changes in sensitivity and specificity. In the following section, we use the modified version when we refer to the likelihood approach using the two-feature sampling distribution.

Robustness of approach: To investigate the robustness of the likelihood approach using the two-feature sampling distribution, we consider the following three scenarios: (a) population growth, (b) population substructure, and (c) gene conversion. For a, we set the starting time point of population expansion to 0.44 (in coalescent units) and the exponential growth rate to

TABLE 2

Estimates of ρ

Type	$\rho_0 = 20$	$\rho_0 = 40$	$\rho_0 = 60$
Modified[a]	(25, 0.90, 0.85)	(46, 0.57, 0.91)	(63, 0.38, 1.0)
Modified[b]	(25, 1.1, 0.81)	(43, 0.62, 0.89)	(59, 0.41, 1.0)
Modified[c]	(27, 1.1, 0.79)	(48, 0.62, 0.88)	(65, 0.40, 1.0)
Simple[b]	(27, 1.2, 0.80)	(44, 0.60, 0.89)	(57, 0.41, 1.0)
Simple[c]	(3, 0.85, 0.99)	(16, 0.77, 0.98)	(41, 0.56, 1.0)
Real[b]	(22, 1.1, 0.84)	(38, 0.62, 0.91)	(50, 0.46, 1.0)
Real[c]	(20, 0.95, 0.87)	(37, 0.60, 0.94)	(52, 0.46, 1.0)

The three numbers within the parentheses denote the mean, the root (relative) mean square error, and probability that the estimate falls within a factor of two of the true parameter, respectively. Three combinations of (sensitivity, specificity) are examined:

[a] (1.0, 1.0).

[b] (0.34, 1.0).

[c] (0.68, 0.75).

2.1, the maximum-likelihood estimates for this scenario obtained in Plagnol *et al.* (2006). For b, we choose two settings to test: two equal-sized subpopulations with scaled migration parameter 1.5 (results not shown) and four equal-sized subpopulations with scaled migration parameter 3.0, chosen to achieve average F_{st}-values of ~0.25 (Hudson *et al.* 1992). In each case the sample was assumed to be divided equally between the subpopulations. For c, we use the estimates in Plagnol *et al.* (2006) to set the mean length of a conversion tract to 100 bp and the ratio of conversion to crossing-over rates to 4.

Table 3 lists estimates of ρ for these three scenarios. We see that we can still estimate ρ if there is exponential growth, gene conversion, or population structure, provided that we have good estimates of the demographic parameters. Moreover, we can see that we tend to overestimate ρ when population growth or gene conversion is present, while there is little bias in the case of population structure.

We emphasize that these scenarios are intended to be exploratory rather than exhaustive. A more extensive test of robustness would examine the effects of changes to mutation, migration and recombination parameters, sample size, and different minor allele frequencies used in obtaining the two-locus sampling distribution.

To address the model misspecification question, *i.e.*, what happens if we analyze data under the wrong model or with the wrong sensitivity and/or specificity, we show the effects of using incorrect sensitivity or specificity in Table 4. The main observation is that both our approaches suffer in this case. We show that dependent modeling is more appropriate for SFP experiments (*cf.* Figure 3) in the DISCUSSION.

We generated test data under $\theta = 200$ and $\rho = 20$ for a 100-kb region with particular sensitivity and specificity (Table 4, a and b, column 1) and then estimated either θ or ρ with different sensitivity and specificity. For example, given the true value of $\theta = 200$ we estimated θ with three different combinations of sensitivity and specificity, and we see that only inference using the true parameters (numbers on the diagonal in Table 4a) provides accurate estimates. On the other hand, we also generated the two-feature sampling distribution with parameters similar to the real experimental setting, *i.e.*, sensitivity = 0.50 and specificity = 0.98, and then used this lookup table to estimate ρ in the test data sets generated with different sensitivity and specificity. For the data sets generated under $s_n = 0.34$ and $s_p = 1.0$, most of the estimates returned are 0, and we label this case "NA," *i.e.*, not applicable. For the other two cases, the estimates seem to be unaffected.

In summary, one needs to obtain reliable estimates of sensitivity and specificity, for example, by sequence comparison, for successful inference.

TABLE 3

Robustness of estimates of ρ

Scenario	$\rho_0 = 20$	$\rho_0 = 40$	$\rho_0 = 60$
Growth[a]	(27, 1.0, 0.81)	(45, 0.60, 0.90)	(55, 0.43, 1.0)
Growth[b]	(26, 0.80, 0.84)	(48, 0.58, 0.90)	(61, 0.35, 1.0)
Growth[c]	(29, 1.0, 0.80)	(49, 0.62, 0.88)	(62, 0.38, 1.0)
Gene conversion[a]	(28, 1.0, 0.79)	(48, 0.65, 0.86)	(60, 0.40, 1.0)
Gene conversion[b]	(28, 0.85, 0.81)	(51, 0.60, 0.87)	(64, 0.38, 1.0)
Gene conversion[c]	(30, 1.0, 0.76)	(52, 0.65, 0.84)	(66, 0.40, 1.0)
Subpopulation[a]	(22, 1.1, 0.86)	(41, 0.58, 0.92)	(55, 0.42, 1.0)
Subpopulation[b]	(21, 0.74, 0.90)	(41, 0.49, 0.95)	(58, 0.36, 1.0)
Subpopulation[c]	(23, 0.92, 0.86)	(43, 0.57, 0.91)	(59, 0.38, 1.0)

Notation is as in Table 2. See text for details. Simulations are run under dependent calling, with probe coverage of 3.3% and three combinations of (sensitivity, specificity):

[a] (0.34, 1.0).

[b] (1.0, 1.0).

[c] (0.68, 0.75).

TABLE 4

Model misspecification effect in dependent model with 3.3% probe coverage

	Inference		
Simulation	(0.34, 1.0)	(1.0, 1.0)	(0.68, 0.75)
	a. True $\theta = 200$		
(0.34, 1.0)	(204, 0.46)	(609, 2.2)	(987, 3.9)
(1.0, 1.0)	(60, 0.71)	(201, 0.27)	(446, 1.2)
(0.68, 0.75)	(14, 0.92)	(8, 0.96)	(207, 0.54)

	Inference	
Simulation	True (s_n, s_p)	(0.5, 0.98)
	b. True $\rho = 20$	
(0.34, 1.0)	(25, 1.2, 0.81)	NA
(1.0, 1.0)	(25, 0.90, 0.85)	(25, 0.91, 0.85)
(0.68, 0.75)	(27, 1.1, 0.79)	(22, 0.80, 0.88)

Test data are generated with sensitivity and specificity listed in the Simulation column and estimates are listed under different (s_n, s_p) used in Inference. See text for details. Notation in a is the same as in Table 1 and notation in b is the same as in Table 2.

DISCUSSION

We have shown that single-feature polymorphism data can be used effectively to infer aspects of the population history of *A. thaliana*. Single-feature polymorphisms, as a new type of polymorphism data, have certain advantages over other types of polymorphism data. By exploiting the high-throughput nature of arrays, large amounts of polymorphism data can be produced in an economic and efficient way. Given recent advances in the technology, the probe coverage over the genome of interest can be easily increased (Borevitz and Ecker 2004). New commercially available tiling arrays cover the reference strain Col with very small spacing (10 bp) between probes. This will help expedite the exploration of natural variation at the genomewide scale.

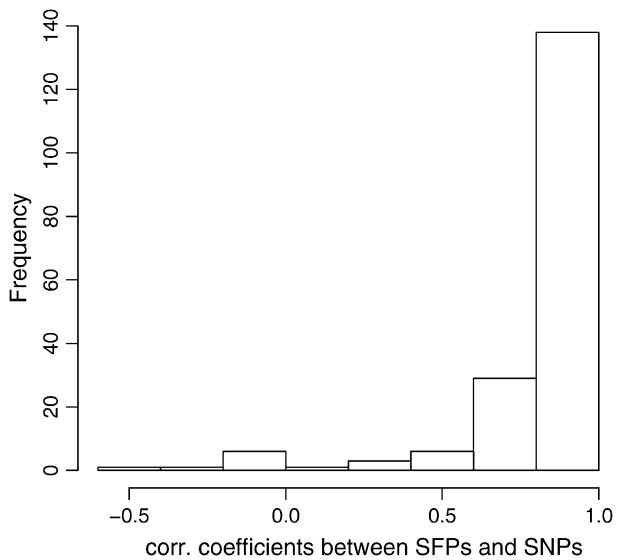

Figure 3.—Histogram of correlation coefficients between SFPs and underlying SNPs for the 406 called features in a real experiment.

By taking account of the transformation from SNPs to SFPs, one could apply most of the methods developed for SNP data to SFP data. For example, the product of approximate conditional-likelihood (PACL) approach in Li and Stephens (2003) could be applied by adding one more step reflecting SFP calling after simulating the SNP sample. Moreover, we could use sliding windows across the genome to study variation in mutation rate and recombination rate using SFP data, as well as to find recombination hot spots, provided that there is a reasonably dense probe coverage of the region of interest (Fearnhead *et al.* 2004). Other statistics defined for SNP data, such as Tajima's *D* and the inbreeding coefficient, might be derived in an analogous way and used to study natural selection and/or population structure. One appealing and challenging task is to develop methods for using SFPs for fine-mapping purposes. For example, methods such as the spatial clustering scheme in Molitor *et al.* (2003), based on haplotype sharing, can be adopted by defining a new metric that takes into account the uncertainty due to SFP calling (Kim *et al.* 2006).

When considering how best to improve microarray experiment technology and design in our context, we first observe that increasing the specificity leads to larger improvement in accuracy than increasing the sensitivity. This is because the inference is considerably affected by false positive calls. Thus controlling the false discovery rate in any statistical approach that calls SFPs is important. To improve modeling of the SFP calling procedure, one could allow for the existence of multiple SNPs within a probe when deriving analytic formulas. In fact, the proportion of probes having more than one SNP in our simulations ranges from 4 to 12% as θ increases from 200 to 600, while Nordborg *et al.* (2005) reports that less than one-sixth of the true positive SFPs have more than two alleles. Moreover, the rate at which a probe hybridizes is likely affected by the position of the SNP within the 25-bp region (Borevitz *et al.* 2003).

It is also important to study the effect of the dependent/independent calling schemes. To investigate this, we obtained 406 SFP loci and their corresponding SNPs by comparing SFP calls with the available sequence data for the 16 accessions in our experiment. We then calculated the correlation between the SNP and SFP calls at each SFP locus. Figure 3 gives a histogram showing that the degree of correlation is close to 1, which is very similar to the histogram obtained from simulations under dependent calling (histogram not shown). On the other hand, simulations show that for independent calling with parameters matching real experiments, there is substantial mass to both the left and the right of 0 (histogram not shown). Therefore, the mass at 1 tells us that calling is generally quite dependent.

TABLE 5

Estimates of the population recombination rate ρ in a 100 kb region

ρ	SFP			SNP		
20	22	0.59	0.92	22	0.78	0.89
40	45	0.48	0.94	42	0.59	0.91
60	66	0.28	1.00	52	0.44	1.00

The SNP case has density one every 10 kb and the SFP case has 70% coverage, 95% specificity, and 50% sensitivity under a dependent calling scheme. In each of the SFP and SNP categories, first column gives the average estimate, the second the RMSE, and the third the proportion of estimates lying within a factor of 2 of the truth.

Finally, we have performed a further simulation study to address how much information is lost in inference when using SFP data rather than SNP data. We first simulated coalescent samples of 20 haplotypes, in a 100-kb region. Then SFPs were generated from the SNPs under dependent calling, with coverage 70%, sensitivity 50%, and specificity 95%, as in the current experimental setup. Moreover, we selected candidate SNPs from the SNP sample at a density of 1/10 kb with minor allele frequency >10%. To estimate the recombination parameter ρ, we applied the two-locus/feature sampling distributions on the selected SNP pairs and called SFP pairs, respectively. The results in Table 5 indicate that SFPs slightly outperform SNPs in this case: the SFPs have a smaller RMSE and a larger proportion of estimates lying within a factor of two of the true ρ than those of the SNPs.

We thank two anonymous referees for their helpful comments. This work was funded in part by National Institutes of Health grants HG002790, GM67243, and GM069890. S.T. is a Royal Society–Wolfson Research Merit Award holder.

LITERATURE CITED

Borevitz, J. O., and J. R. Ecker, 2004 Plant genomics: the third wave. Annu. Rev. Genomics Hum. Genet. **55:** 443–477.

Borevitz, J. O., D. Liang, D. Plouffe, H. Chang, T. Zhu *et al.*, 2003 Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res. **13:** 513–523.

Fearnhead, P., 2003 Consistency of estimators of the population-scaled recombination rate. Theor. Popul. Biol. **64:** 67–79.

Fearnhead, P., R. Harding, J. Schneider, S. Myers and P. Donnelly, 2004 Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. Genetics **167:** 2067–2081.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stoch. Models **14:** 273–295.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model. Bioinformatics **18:** 337–338.

Hudson, R. R., M. Slatkin and W. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Kim, S., K. Zhao, R. Jiang, J. Molitor, J. O. Borevitz *et al.*, 2006 Association mapping with single-feature polymorphisms. Genetics **173:** 1125–1134.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165:** 2213–2233.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Molitor, J., P. Marjoram and D. Thomas, 2003 Application of Bayesian clustering via Voronoi tesselations to the analysis of haplotype risk and gene mapping. Am. J. Hum. Genet. **73:** 1368–1384.

Nordborg, M., 2001 Coalescent theory, pp. 179–208 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. J. Bishop and C. Cannings. John Wiley & Sons, New York.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Tang *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. **3:** e196.

Parmigiani, G., E. S. Garett, R. A. Irizarry and S. L. Zeger (Editors), 2003 *The Analysis of Gene Expression Data: Methods and Software*. Springer-Verlag, New York.

Plagnol, V., B. Padhukasahasram, J. D. Wall, P. Marjoram and M. Nordborg, 2006 Relative influences of crossing over and gene conversion on the pattern of linkage disequilibrium in *Arabidopsis thaliana*. Genetics **172:** 2441–2448.

Speed, T. P. (Editor), 2003 *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC Press, London/New York/Cleveland/Boca Raton, FL.

Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA **100:** 9440–9445.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tusher, V. G., R. Tibshirani and G. Chu, 2001 Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA **98:** 5116–5121.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

Werner, J. D., J. O. Borevitz, N. Warthmann, G. T. Trainer, J. R. Ecker *et al.*, 2005 Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc. Natl. Acad. Sci. USA **102:** 2460–2465.

Wolyn, D., J. O. Borevitz, O. Loudet, C. Schwartz, J. Maloof *et al.*, 2004 Light-response quantitative trait loci identified with composite interval and eXtreme array mapping in *Arabidopsis thaliana*. Genetics **167:** 907–917.

Communicating editor: G. Gibson

APPENDIX

An illustration of the SFP-calling procedure: Under independent calling, true/false positive/negative calls are independent between accessions at a given probe position, while there are only four possibilities under dependent calling, *i.e.*,

I. $1 \to 1$ and $0 \to 0$;
II. $1 \to 0$ and $0 \to 1$;
III. $1 \to 1$ and $0 \to 1$;
IV. $1 \to 0$ and $0 \to 0$.

Here we illustrate a case in which SFP calling is assumed to be dependent across accessions. Suppose we have four accessions plus the reference and are interested in a region of three distinct probes. We simulate a sample of five haplotypes, noting again that *A. thaliana* is mostly homozygous and there is little within-strain

SNP			*			SFP		
						III	I	II
1	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	0
0	0	0	1	0	0	1	0	1
0	1	0	1	1	0	1	1	1
1	0	0	0	0	0	1	0	1

FIGURE A1.—Example of transforming SNP data to SFP data. Four accessions plus the reference on the top row. Three probes with exactly one SNP within each probe. See text for the meanings of 0 and 1 in different matrices.

variation for each accession. The haplotypes are represented by the five rows in Figure A1, where 0 denotes wild type and 1 denotes mutant. The top row is chosen randomly as the reference, and the other four are accessions. Assuming that there is exactly one SNP within each probe, this results in the haplotypes shown in the left part of Figure A1, labeled "SNP." We then undertake the SFP comparison phase by comparing the accessions with the reference (the top row). In this phase we use 1 to indicate a location that is different from the reference probe and 0 otherwise. By doing so we get the middle matrix, labeled "*." Note that only the first column has been changed since the chosen reference contained the derived mutation in this case. Finally, we apply the SFP-calling phase to the middle matrix to obtain the SFP matrix on the right, where 1 denotes a called significant difference between the accession probe and the reference probe and 0 anything else. This illustrates a calling operation of type III at the first probe position, type I at the second probe position, and type II at the last probe position. Calling operations of type I represent true calls, so every comparison remains the same. Calling operations of type II introduce false negatives and false positives, which flip 0's to 1's and vice versa. Type III introduces only false positives, which flip 0's to 1's. When the sensitivity and specificity are both 1.0, *i.e.*, the ideal case, the middle matrix is the same as the SFP matrix.

The expected number of SFPs: Denote the probe configuration by $\mathcal{P} = (P_1, P_2, \ldots, P_{n_p})$, where n_p is the number of probes, and $P_i = 1$ if the ith probe is called an SFP or 0 otherwise. Recall that an SFP is called at the ith probe if and only if at least one accession is called polymorphic with respect to the reference. Denote the total number of SFPs called by F. Then we have $F = \sum_{i=1}^{n_p} P_i$ and

$$\mathbb{E}[F] = \mathbb{E}\left[\sum_{i=1}^{n_p} P_i\right] = \sum_{i=1}^{n_p} \mathbb{E}[P_i] = \sum_{i=1}^{n_p} \mathbb{P}(P_i = 1), \quad \text{(A1)}$$

where $\mathbb{E}[\cdot]$ stands for mathematical expectation and $\mathbb{P}(\cdot)$ stands for probability.

Conditional on the number of SNPs within a particular probe, the probe set $\mathcal{P}$ can be partitioned into two parts, probes containing one SNP and probes without SNPs, denoted by $\mathcal{P}_1$ and $\mathcal{P}_0$, respectively. That is, $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_0$. Following Equation A1, we have

$$\begin{aligned}\mathbb{E}[F] &= \sum_{i\in\mathcal{P}_1} \mathbb{P}(P_i = 1) + \sum_{j\in\mathcal{P}_0} \mathbb{P}(P_j = 1) \\ &= \alpha\,|\mathcal{P}_1| + \beta\,|\mathcal{P}_0|, \end{aligned} \quad \text{(A2)}$$

where $|\mathcal{A}|$ denotes the size of the set $\mathcal{A}$, $\alpha = \mathbb{P}(P_i = 1 \mid \mathcal{P}_1)$, and $\beta = \mathbb{P}(P_i = 1 \mid \mathcal{P}_0)$. Recall that we assume that probes are called independently.

Next, we compute the probabilities α and β. First, consider a probe with one and only one segregating site inside. The mutation within the probe partitions the sample into two groups, the wild-type group and the mutant group. For example, there are three wild types and two mutants in the first probe in Figure A1 (the first column in the SNP matrix). Let q_{nb} denote the probability of obtaining b mutants out of n individuals, which is given in GRIFFITHS and TAVARÉ (1998). Recall that an SFP is called if and only if at least one 1 appears in the resulting SFP matrix (the right part in Figure A1). In other words, the only way of not calling an SFP is to have true negative calls and false negative calls at the probe position, by applying calling operation IV ($1\to0$ and $0\to0$) on each accession probe (*cf.* the middle section in Figure A1). Conditioning on which group the reference is chosen from, we have

$$\alpha = \begin{cases} \sum_{b=1}^{n-1} q_{nb}[(1-(1-s_n)^{n-b}s_p^{b-1})\frac{b}{n} \\ \quad + (1-(1-s_n)^b s_p^{n-b-1})\frac{n-b}{n}], & \text{independent calling;} \\ 1-(1-s_n)s_p, & \text{dependent calling.} \end{cases} \quad \text{(A3)}$$

For a probe containing no SNP, the chance of being called an SFP is determined by the specificity. Therefore, we have

$$\beta = \begin{cases} 1 - s_p^{n-1}, & \text{independent calling;} \\ 1 - s_p, & \text{dependent calling.} \end{cases} \quad \text{(A4)}$$

Finally, assuming that there is at most one SNP within a probe, we can determine the size of the set of probes with SNPs; *i.e.*, $|\mathcal{P}_1| = S$, where S is the number of segregating sites or SNPs. Following Equation A2, we obtain

$$\mathbb{E}[F \mid S] = \alpha\,|\mathcal{P}_1| + \beta\,|\mathcal{P}_0| = \alpha S + \beta(n_p - S), \quad \text{(A5)}$$

where n_p is the total number of probes. Furthermore, we have

$$\begin{aligned}\mathbb{E}[F] &= \mathbb{E}[\mathbb{E}[F \mid S]] = n_p\beta + (\alpha - \beta)\mathbb{E}[S] \\ &= n_p\beta + (\alpha - \beta)\left(\frac{n_p L_p}{L}\right)\sum_{k=1}^{n-1}\frac{1}{k}\cdot\theta, \end{aligned} \quad \text{(A6)}$$

SNP			*			SFP		
						I	I	
0	0	(3)	0	0	(3)	0	0	(3)
0	1	(5)	0	1	(5)	0	1	(5)
1	0	(7)	1	0	(7)	1	0	(7)
1	1	(5)	1	1	(5)	1	1	(5)

SNP			*			SFP		
						I	I	
0	0	(5)	0	1	(5)	0	1	(5)
0	1	(3)	0	0	(3)	0	0	(3)
1	0	(5)	1	1	(5)	1	1	(5)
1	1	(7)	1	0	(7)	1	0	(7)

SNP			*			SFP		
						I	II	
0	0	(6)	0	0	(6)	0	1	(5)
0	1	(2)	0	1	(2)	0	0	(3)
1	0	(5)	1	0	(5)	1	1	(5)
1	1	(7)	1	1	(7)	1	0	(7)

FIGURE A2.—Example of different two-locus configurations that can lead to the same two-feature configuration. The sample size is 20 and the number within the parentheses indicates how many such combinations are present in the sample. The type of reference is 00 (top), 01 (middle), or 00 (bottom). See text for more details.

where the last term is given under the infinite-sites mutation model (WATTERSON 1975), and L_p is the length of a probe (*i.e.*, 25 bp) and L is the total length of the region in base pairs.

Obtaining two-feature SFPs from SNPs: Figure A2 illustrates how different SNP configurations can lead to the same SFP configuration. Suppose that we have a sample of 19 accessions plus the reference, *i.e.*, the sample size is 20. Suppose further that calling is completely dependent across accessions. We focus on a particular pair of features, where we assume there is one and only one segregating site within each feature. First we explain the top part in Figure A2, where there are 3 00's, 5 01's, 7 10's, and 5 11's at the two segregating sites in the sample. Assume that the type of the reference is randomly chosen to be 00. Then after the SFP comparison phase, all the combinations remain the same in the * matrix. Furthermore, if both calling operations at the two features are of type I, *i.e.*, true positive/negative only, we again obtain the SFP configuration $\mathbf{m} = (3, 5, 7, 5)$. For the case where the reference is not of type 00, we look at the middle section in Figure A2. Suppose that we have a different SNP configuration, $\mathbf{n} = (5, 3, 5, 7)$, and the type of the reference is 01. Thus in the SFP comparison phase, the original combination 00 becomes 01 compared to the reference 01, and 01 becomes 00, etc. As can be seen, the * matrix will lead to the same SFP configuration, $\mathbf{m} = (3, 5, 7, 5)$, if we have both calling operations of type I. Similarly, we can see in the bottom part of Figure A2, where the reference is again 00 but now the second feature uses a calling operation of type II, how a different SNP configuration can give rise to such an SFP configuration by taking appropriate calling operations at the two features.

Although there are four possible calling operations altogether, only the first three lead to an SFP in the resulting sample, since, assuming completely dependent calling, operation IV changes the column of that probe to all 0's. For given sensitivity and specificity, we can write down the probability for each calling operation, *e.g.*, $\mathbb{P}(C_I) = s_n s_p$.

We noted earlier that under the completely dependent calling assumption, the false positives have one common pattern in the resulting SFP matrix, *i.e.*, all 1's but one 0 for the reference probe. To see this, start from a probe without any SNP in it. After the SFP comparison phase, the probes at this position are all labeled by 0, since there is no polymorphism among them. In the SFP calling phase, if one of the accession probes is called significantly different, *i.e.*, $0 \rightarrow 1$, then the rest of the accession probes are also called significantly different, under dependent calling. Thus this probe is called as an SFP, however, a false positive. On the other hand, the only SNP configuration that could give rise to the above pattern is $(0, 1, 1, \ldots, 1)$ (without loss of generality, we label the first one as wild type and the rest as mutants). In this case the reference is most likely to be chosen from the mutant group, in which case this column becomes $(1, 0, 0, \ldots, 0)$ after the SFP comparison phase. Furthermore, the only operation that can change this column to the pattern of false positives is of type III, *i.e.*, ($1 \rightarrow 1$ and $0 \rightarrow 1$). This is the only situation that a probe with segregating sites can lead to the same pattern as a false positive. However, the probability of obtaining only one wild type in the infinite-sites model is very small, in particular when the sample size is large.

Consistency of MLE of ρ: The consistency of the maximum-likelihood estimates of ρ cannot be shown for the general setting in the two-feature sampling distribution. However, it can be shown that if we modify the approach by insisting that all pairs of features be within a certain distance we would obtain consistent estimates of ρ, as in the modified approach of the two-locus sampling distribution. The key point of the additional restriction is that linkage disequilibrium decays inversely to the distance between two sites. The proof in FEARNHEAD (2003) can be adopted without substantial change. The adjustment is in Lemma 5, in which we need to sum over all possible two-loci SNP configurations to compute the covariance coefficients between two features. Recall that there are only four possibilities in the modified two-feature sampling derivation and nine possibilities in the simple two-feature sampling derivation. Thus, the conclusion holds in the SFP setting.

Statistical properties of the two-feature sampling distribution: One way to study the statistical properties of the maximum-likelihood estimate $\hat{\rho}$ is to consider the minimal contrast, *i.e.*, the expectation of $\log_{\rho_0}(p_c(\mathbf{m};\rho))$, over the distribution of $\mathbf{m}$ conditional on polymorphism at both features (*cf.* Hudson 2001). That is, we consider

$$\mathbb{E}_{\rho_0}[\log(p_c(\mathbf{m};\rho))] = \sum_{\mathbf{m}} p_c(\mathbf{m};\rho_0)\log(p_c(\mathbf{m};\rho)). \quad \text{(A7)}$$

The plots of the above function (not shown here) show that the likelihood curve has a peak at a position close to the true value ρ_0, which enables the maximum-likelihood approach to work.

To study the asymptotic variance of the maximum-likelihood estimate of ρ_0, we compute the second derivative of the function in (A7) with respect to ρ, evaluated at the ρ_0, assuming k pairs of features are considered. That is, we compute

$$\mathrm{Var}_{\rho_0,k}(\hat{\rho}) \approx \frac{1}{-k(\partial^2/\partial\rho^2)\mathbb{E}_{\rho_0}(\log p_c(\mathbf{m};\rho))\,|_{\rho=\rho_0}}. \quad \text{(A8)}$$

An approximation of the above is plotted in Figure A3. The curves of estimated variance in both the SNP case and the SFP case are similar; however, there is smaller variance in the SNP case than in the SFP case, since there is more randomness in the latter. The asymptotic plot shows that features separated by ρ in the range of 2–15 are best for estimating ρ, as is observed also in Hudson (2001).

Note that the difference between the scales in Figure A3 and the corresponding figure in Hudson (2001) is due to different conditionings on the two sites when obtaining the two-locus sampling distribution. Hudson (2001) restricted to those sites where the minor allele frequency is >10%, where we require just that both sites are segregating. In other words, the minor allele frequency is >5% (*i.e.*, 1 of 20) in our scheme. Since the ages of SNPs are related to their frequencies, SNPs with larger frequency tend to be older than those with smaller frequency, so they are more informative about recombination than rarer SNPs. As a result, conditioning on the minor allele frequency >10% leads to smaller variance of MLEs than those conditional on minor allele frequency >5%.

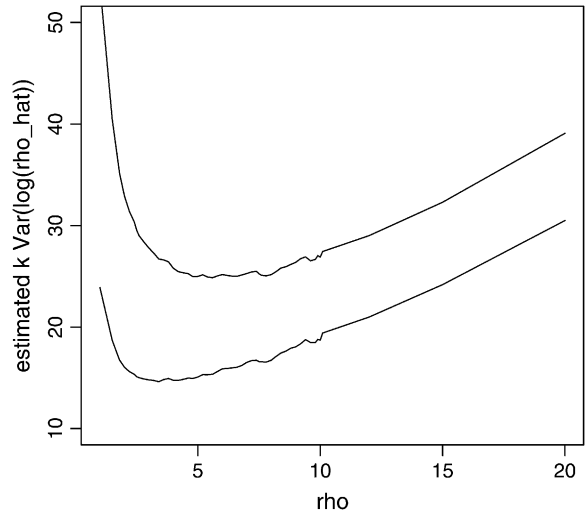

Figure A3.—Estimates of the asymptotic variance of $\log\hat{\rho}$ based on k independent pairs of polymorphic sites. Top line, from the two-feature sampling distribution; bottom line, from the two-locus sampling distribution.