DOI: 10.1534/genetics.105.052720

Association Mapping With Single-Feature Polymorphisms

Sung Kim,*[,1] Keyan Zhao,*[,1] Rong Jiang,* John Molitor,[†] Justin O. Borevitz,[‡] Magnus Nordborg* and Paul Marjoram[†,2]

[*]*Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089-2910,* [†]*Department of Preventive Medicine, University of Southern California, Los Angeles, California 90089-9011 and* [‡]*Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637*

Manuscript received October 22, 2005
Accepted for publication February 21, 2006

ABSTRACT

We develop methods for exploiting "single-feature polymorphism" data, generated by hybridizing genomic DNA to oligonucleotide expression arrays. Our methods enable the use of such data, which can be regarded as very high density, but imperfect, polymorphism data, for genomewide association or linkage disequilibrium mapping. We use a simulation-based power study to conclude that our methods should have good power for organisms like *Arabidopsis thaliana*, in which linkage disequilibrium is extensive, the reason being that the noisiness of single-feature polymorphism data is more than compensated for by their great number. Finally, we show how power depends on the accuracy with which single-feature polymorphisms are called.

In this article we aim to demonstrate that single-feature polymorphisms (SFPs) are a viable alternative to single-nucleotide polymorphisms (SNPs) for genomewide association studies, at least in organisms such as *Arabidopsis thaliana* where extensive linkage disequilibrium (LD) means that noisiness of individual markers can be compensated for by using a higher marker density.

SFPs were first identified in yeast as significant differences in hybridization intensity between strains when genomic DNA was hybridized to high-density oligonucleotide expression arrays (Winzeler *et al.* 1998). Subsequently, the method was used in the considerably more complex context of the *A. thaliana* genome (Borevitz *et al.* 2003). The two main advantages of SFPs are that standard expression arrays are used *in lieu* of specialized genotyping technology and that no prior knowledge of SNPs is required. SFP typing is currently being applied to a wide range of organisms such as mosquito (Turner *et al.* 2005) and barley (Rostoks *et al.* 2005), with some decrease in the signal-to-noise ratio as genome size increases. Replicating the arrays improves the accuracy with which polymorphisms are detected, but with a consequent trade-off in terms of cost.

In our application, each SFP probe corresponds to a 25-bp oligonucleotide on the basis of the published *A. thaliana* reference genome [from the accession (inbred line) *Col-0*]. The distance between probes was ~10 bp on average. Data for each 25-bp probe region are in the form of relative strengths of hybridization for the individual to be genotyped and *Col-0* (the reference). We use a false discovery rate (FDR) threshold of a test statistic to convert the relative hybridization intensities to 0's and 1's, where 0 means "matches the reference genome" and 1 means "does not match the reference genome" (Borevitz *et al.* 2003). Note that our data consist of inbred lines of a selfing organism. Thus, it is overwhelmingly likely that only two genotypes will be observed at any given site, 00 and 11, and this is used as an assumption throughout this article.

There are several interesting asymmetries to the data. First, two individuals that differ from the reference genome at a given probe position, and thus report a 1, may not carry the same mutations. However, unless polymorphisms occur very densely in the genome, we do not expect this to be a major problem. More important is the difference in error rates between 0's and 1's, *i.e.*, sensitivity and specificity. If we want high specificity, *i.e.*, we want to be reasonably certain that a 1 really is a 1, then we are likely to miss several true positives, falsely declaring some 1's as 0's (*i.e.*, we will have low sensitivity).

Thus, SFP data are dense but noisy and characterized by highly asymmetric errors. Indeed, one of the main rationales for our work, and its application to *A. thaliana*, is that the extensive LD in *A. thaliana* not only facilitates LD mapping (Nordborg *et al.* 2005), but it also helps overcome the noise in SFP data because haplotypes can be inferred using multiple, albeit noisy, markers. Although SFP data are much noisier than SNP data, they are also cheaper to generate per genotype, and the extensive LD in *A. thaliana* ensures that lack of quality is compensated by much greater quantity.

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Department of Preventive Medicine, University of Southern California, 1540 Alcazar St, CHP-220, Los Angeles, CA, 90089-9011. E-mail: pmarjora@usc.edu

Our goal here is to explore the properties of *A. thaliana* SFP data with respect to LD mapping by combining SFPs from a pilot study (J. O. Borevitz and J. R. Ecker, unpublished data) with direct sequencing data from another study (Nordborg *et al.* 2005). In other work, we have also explored algorithms for estimating standard population genetics parameters such as the mutation and recombination rates from SFP data (R. Jiang, P. Marjoram, J. O. Borevitz and S. Tavaré, unpublished results). Our methods exploit the LD within the data.

LD is the pattern of nonrandom association between loci. Observed patterns of polymorphism in molecular data reflect the ancestral history of the sample and the effects of mutation on that ancestry. Furthermore, the presence of mutations that affect phenotypes of interest, referred to as functional mutations from here onward, implies increased similarity among individuals with similar phenotypic values in the regions surrounding those mutations. This effect is degraded by the action of recombination over time, the existence of multiple functional mutations, and the existence of other factors that are likely to influence the phenotype of interest (*e.g.*, the environment). Nonetheless, LD has been used as the successful basis for a variety of methods that map functional mutations.

However, the use of LD is complicated by its variability along the genome. The possible block structure of the human genome has been the subject of a growing number of articles (*e.g.*, Daly *et al.* 2001; Patil *et al.* 2001; Stephens *et al.* 2001; Gabriel *et al.* 2002; Phillips *et al.* 2003; Stumpf and Goldstein 2003; Marchini *et al.* 2004; Jeffreys *et al.* 2005; Hinds *et al.* 2005; International HapMap Consortium 2005; Myers *et al.* 2005). Furthermore, a fundamental property of LD is that it is highly variable. This variability reflects the randomness inherent in the underlying evolutionary processes that gave rise to the data. This observed pattern of LD is but a snapshot of a rapidly evolving pattern. The high level of variability means that it is not uncommon, in fact it is usual, to see nearby loci with completely different patterns of LD. We also sometimes observe a high degree of LD between loci that are far apart, even though intermediate loci show little LD. Thus, it is highly likely that some degree of modeling or smoothing will be required to extract the underlying signal from the superficial noise. To some extent, the variability is due to the pairwise nature of many common measures of LD. If we smooth the measure of LD over more loci we expect the level of variability to decrease. This, in part, motivates other LD measures based upon haplotype structure (*e.g.*, Maniatis *et al.* 2004, 2005; Rinaldo *et al.* 2005).

As we have discussed, the pattern of LD is the result of the action of evolutionary events such as recombination and mutation over the ancestral history of the sample. In principle this ancestral history can be described by a stochastic process known as the coalescent (Kingman 1982). Coalescent models have proven to be extremely powerful in many applications; however, these have primarily been in contexts in which recombination is absent and where the data can be assumed to have evolved without selective pressure (see Tavaré 1984; Hudson 1990; Nordborg 2001, for reviews). Although it is possible to develop appropriate coalescent models in the presence of almost any complicating factor, the complexity of such models is enormous, and it is therefore entirely plausible that it is counterproductive to include these details in analyses. The vast increase in computational effort required to include the model can substantially outweigh the theoretical gain in power.

Given the complexity of full evolutionary models such as the coalescent, there has recently been a move to consider fine-mapping approaches that approximate key features of such models while avoiding most of the computational complexity. Some of these methods attempt explicitly to approximate aspects of the underlying coalescent process (*e.g.*, Graham and Thompson 1998; Morris *et al.* 2000, 2002; Liu *et al.* 2001; Roeder *et al.* 2005; Zollner and Pritchard 2005), while others use an approach that is more abstract in nature, in which the coalescent process is replaced by some other method of clustering of the data into clades (*e.g.*, Templeton *et al.* 1987, 2005; Templeton 1995; Molitor *et al.* 2003a,b; Durrant *et al.* 2004; Sillanpaa and Bhattacharjee 2005; Tzeng 2005). Analyses of the latter type are less complex in nature than those of the former type. Thus, while they might lose some power due to the use of a more abstract approximation to the underlying ancestry of the sample, they gain by imposing a smaller computational burden and are therefore likely to be able to analyze larger data sets. Many of these methods are discussed in Molitor *et al.* (2004).

We believe that both mapping methods based on explicit evolutionary models and those based upon more abstract summaries of those models are valid, but that the greater computational complexity of the former approaches will often make them intractable for the analysis of genomewide data. Thus, our approach here is to focus on the more abstract methods.

METHODS

SFP data, because of their imperfect sensitivity, s_n, and specificity, s_p, can be viewed as SNP data to which noise has been added. Given that an individual chromosome i differs from the recognition pattern for probe j, at $\geq$1 bp within that probe, the SFP data report a 1 with probability s_n; otherwise they report a 0. Given that the individual matches the recognition pattern, a 0 is reported with probability s_p; otherwise a 1 is reported. (Note that, for convenience we assume that s_n and s_p are constant across SFPs and chromosomes.) We have developed a range of methods for fine mapping using SNP data in the context of *A. thaliana* or the human

genome (Molitor *et al.* 2003a,b, 2005; Hagenblad *et al.* 2004). These methods rely upon the concept of haplotype sharing. Intuitively speaking, if individuals share a mutation at a location x, they are also expected to share mutations near to x, this effect being broken down over time, by recombination, as we move away from x. While there is an established theory for the sharing of SNPs in regions local to functional mutations, it is less clear how strong this signal remains when one is looking at SFP rather than SNP data. We now elaborate on why this is so.

Assume that a given sample of individuals have SFP data constructed by comparison to an external reference sequence. For convenience we make the approximating assumption that there is no recurrent mutation, so that each SNP in the (unknown) underlying data is the result of a unique mutational event (this assumption is easy to relax). We further assume that at most one base will have mutated within each SFP probe. Empirical results for *A. thaliana* show less than one in six SFPs have multiple alleles (J. O. Borevitz and J. R. Ecker, unpublished results). Allowing for the case in which more than one mutation might occur results in nontractable mathematical derivation in the results that follow. Furthermore, in this article we simulate test data using evolutionary parameters appropriate for *A. thaliana*. In these simulated data, the proportion of probes containing more than one mutation is 18% (comparable to the results of one in six above). Thus, we believe that the results in this article show that our methods work despite these simplifying assumptions. Informally speaking, as long as enough of the SFPs are giving a clean signal, the presence of LD allows us to overcome the effects of SFPs that do not accurately reflect the state of the underlying sequence data.

SFP data give imperfect information: even if the SFP information for two "haplotypes" is identical, the underlying sequences may be different (due to imperfect sensitivity and specificity of the probes). Using our FDR threshold for SFP detection results in the following values for these parameters: $s_n = 0.5$ and $s_p = 0.98$ (Borevitz *et al.* 2003). These values may alter if the FDR threshold or sample size is changed. This introduces considerable noise into the analysis. In the next section we derive specific results used to conduct our fine-mapping analysis of SFP data. In particular, we derive the probability that each possible configuration of pairwise SFP data corresponds to an identity of the underlying sequence data (given the simplifying assumptions above). We then use these probabilities as the basis of a similarity measure that is analogous to haplotype sharing and that reflects the likely sharing in the unobserved underlying sequence information.

Measuring similarity between SFP haplotypes: We measure similarity between SFP haplotypes using a window containing $2k$ SFPs centered around a particular location of interest. Denote the two SFP sequences by

TABLE 1

Transition probabilities for SFP data

		$(h_{1,i}, h_{2,i}) =$		
		(0, 0)	(0, 1)	(1, 1)
$(j_1, j_2) =$	(0, 0)	s_p^2	$2s_p(1-s_p)$	$(1-s_p)^2$
	(0, 1)	$s_p(1-s_n)$	$s_p s_n + (1-s_p)(1-s_n)$	$(1-s_p)s_n$
	(1, 1)	$(1-s_n)^2$	$2s_n(1-s_n)$	s_n^2

$$h_1 = \{h_{1,-k}, \ldots, h_{1,-1}, h_{1,0}, h_{1,1}, \ldots, h_{1,k}\}$$

and

$$h_2 = \{h_{2,-k}, \ldots, h_{2,-1}, h_{2,0}, h_{2,1}, \ldots, h_{2,k}\},$$

respectively, where $h_{1,0}$ and $h_{2,0}$ are the SFP status at the position x of interest. Suppose we are considering the SFP configuration at position i. Let $p((j_1, j_2), (h_{1,i}, h_{2,i}))$ denote the probability of obtaining (unordered) SFP pair $(h_{1,i}, h_{2,i})$ from (unordered) SNP pair (j_1, j_2) at position i, $(i = -k, \ldots, 0, \ldots, k)$. We can write the transition probability $p((j_1, j_2), (h_{1,i}, h_{2,i}))$ in terms of sensitivity and specificity as shown in Table 1.

It is then natural to define a similarity measure $s(h_{1,i}, h_{2,i})$ between SFP haplotypes h_1 and h_2 at locus i as

$$s(h_{1,i}, h_{2,i}) = \sum_{j_1=j_2} \mu(j_1, j_2) p((j_1, j_2), (h_{1,i}, h_{2,i})),$$

where $\mu(\cdot, \cdot)$ is a prior distribution for the unordered SNP status. We use a naive prior of $\mu(0, 0) = \mu(1, 1) = 0.25$ and $\mu(0, 1) = 0.5$. Note that the underlying SNP state contributes to the similarity score only when both haplotypes have the same SNP status. In intuitive terms, the score is the probability that the unobserved underlying sequence data match at this locus. The total score for the comparison of these two haplotypes at x is then given by

$$S(x, k) = \sum_{i=-k}^{k} s(h_{1,i}, h_{2,i}) w(d(0, i)), \qquad (1)$$

where $d(0, i)$ is the distance between x and the ith SFP and $w(d(0, i))$ is a weight function. A natural choice for the weight function is $w(x) = \exp(-Rx)$, where R is a parameter chosen to reflect the recombination rate. Thus the score is a weighted average of the probabilities that the underlying sequence data match for these two individuals, where the weights decrease as we move away from the location of interest.

We must also define k, the point at which we stop considering SFPs. While it is possible to include the parameter as part of the state space and mix over it within the algorithm, we choose to use a k-value corresponding to 50 kb, which is expected to take us past the limits of LD. Thus, we include on the order of 1000 SFPs. As the value of k increases, we increase the computational burden of

the algorithm. Smaller values of k will improve computational efficiency, at the expense of loss of some signal.

Are SFPs called in a dependent way? While we assume that, conditional on the SNP information, calling of SFPs is independent across probe positions, the degree of dependency across individuals at any given probe position is less clear. The best-case scenario is that SFPs are called completely dependently. As we argue below, the result of errors when SFPs are called completely dependently across individuals is akin to thinning the SNP data (removing those SNPs at which no polymorphism is detected by the SFP-calling procedure). Since the mapping methods we exploit here are based on sharing in a local region, it is intuitively clear that methods that work on SNP data should also have the potential work on SFP data called dependently (assuming a sufficient density of probes). We have confirmed this using simulation (see RESULTS). We now give a detailed argument for this view.

Assume for convenience of explanation that the reference individual contains the "wild-type" allele for a SNP at a probe position. When SFPs are called completely dependently, if a SNP is detected on one individual (an event with probability s_n) it will also (with probability 1) be detected on all other individuals that also contain the polymorphism. Alternatively, if the SNP is not detected on an individual [an event with probability $(1 - s_n)$] it will also fail to be detected on all other individuals at that location (and all individuals will therefore be assumed to be wild type). In this situation, the SFPs report accurately on the underlying SNP status with probability $s_n s_p$ since the polymorphism is detected, in all individuals that carry it, with probability s_n, and all individuals that do not carry the polymorphism are correctly indicated as not containing the polymorphism with probability s_p. This represents a case in which we have perfect information (assuming there is only one SNP within the probe region). It is also possible that the SFP data indicate no polymorphism, an event that occurs with probability $s_n(1 - s_p) + (1 - s_n)s_p$ [using a similar argument to that given above and allowing for the somewhat less likely additional case in which individuals with the mutation are correctly called (an event with probability s_n), while those that are wild type are incorrectly called as being mutant (an event with probability $1 - s_p$)]. Note that there is one other case, involving a situation in which we get a false-positive call, but this occurs with low probability (the false-positive rate being of the order of 0.02) and so we omit details for the sake of simplicity of exposition.

Assuming that a high degree of dependency in SFP calling across individuals is sensible as each probe has an innate hybridization affinity, good detectors will have high sensitivity across all strains, while poor probes will generally miss the call. SNPs suffer from a similar fate as some are more error prone than others in a dependent fashion across individuals. However, it is also instructive to consider the worst-case scenario in which SFPs are called independently across individuals at each locus. Given the high specificity and the relatively low sensitivity, to a first order of approximation this is conceptually similar to taking the SNP data and randomly (and independently) switching a proportion of the 1's into 0's. It is intuitively clear that this will lead to potentially damaging loss of signal. However, our simulation studies (see RESULTS) indicate that the problem is not as severe as one might expect. Furthermore, we now show that SFPs do appear to be called with a high degree of dependency, as expected.

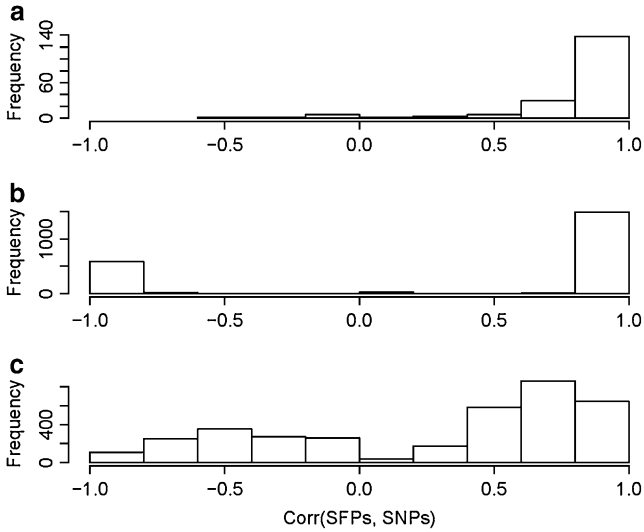

FIGURE 1.—Correlations between SFPs and SNPs in (a) real data, (b) simulated data with completely dependent errors across polymorphisms, and (c) simulated data with independent errors across polymorphisms.

We explore the issue of dependency informally by examining data for which we have SFP calls as well as the underlying SNP information (from direct sequencing, see NORDBORG *et al.* 2005). We have data for 406 such probes at which an SFP has been detected (*i.e.*, evidence of a polymorphism is detected). We consider each SFP position and calculate the correlation coefficient between the true underlying SNP state and the observed SFP calls. We then plot a histogram of these correlation coefficients calculated across all 406 SFPs. The results are shown in Figure 1a. Analogous plots are then constructed from simulated data in which SFPs are called either completely dependently or completely independently (Figure 1, b and c). These plots indicate that SFPs are clearly not called independently and, in fact, appear to be called with a high degree of dependence.

Mapping via spatial clustering algorithms: Our methods aim to cluster haplotypes into groups that are consistent with the genealogical relationship at each point along the chromosome. In a classic article, TEMPLETON *et al.* (1987) noted that if the relationship between a set of haplotypes could be described via a genealogical tree, then it would make sense to incorporate that tree

when testing for phenotypic associations. One could, for instance, successively test more inclusive clades. This idea is not directly helpful for LD mapping because the effects of recombination mean that the relationship between haplotypes is not tree-like (it might be used to test for phenotypic associations in nonrecombining data such as Y chromosomes, for example; see Kittles *et al.* 1999). We require methods for estimating the tree-like relationship at each point in a set of haplotypes (*i.e.*, along the genome). In this article we use two simple methods introduced in Aranzana *et al.* (2005), which we describe here for completeness.

The simplest possible approach for incorporating haplotype structure is to consider windows that are narrow enough for recombination to be rare. Given that the polymorphism data generated by Nordborg *et al.* (2005) were in the form of short sequence fragments, it was natural to treat the resulting short multi-SNP (and indel) haplotypes as multiallelic markers at a single locus. When doing so we always removed singleton polymorphisms, which causes singleton haplotypes to become identical to more common ones. One reason for this procedure is to reduce the degrees of freedom; another is the logic used by Templeton *et al.* (1987) because the procedure effectively collapses singleton branches of the haplotype tree into larger clades. To ensure robustness we analyzed the data using the Kruskal–Wallis nonparametric test (Siegel and Castellan 1988), but one could also use a regular ANOVA with the haplotype clusters as factors. We refer to this method as fragment-based Kruskal–Wallis (or simply Kruskal–Wallis) in what follows.

A more sophisticated approach is to compute a measure of relatedness between all pairs of haplotypes, with respect to a particular point or marker in the haplotypes. These distances can then be used in standard clustering algorithms and the resulting clusters can be used when testing for associations. We have termed this approach cladistic association (CLASS). Using the haplotype-sharing metric described earlier, we can generate a "cladistic" representation of the derived distance matrix at every marker position through a standard hierarchical clustering algorithm, such as neighbor joining, using a distance metric defined as inversely proportional to the score function given in Equation 1. We then heuristically search for the clades (cluster of individuals) that are most strongly associated with the phenotype. We again used the Kruskal–Wallis test to measure strength of association. Approaches similar to ours have recently been proposed by several researchers (Durrant *et al.* 2004; Templeton *et al.* 2005; Tzeng 2005). Our algorithm finds clades as follows. First search all clades and choose the one that gives the lowest *P*-value in a Kruskal–Wallis test with 1 d.f. Then search the tree obtained by removing this clade for the clade that gives the lowest *P*-value in a Kruskal–Wallis test with three factors (and 2 d.f.): the target clade, the clade identified in the previous step, and the remaining individuals. We repeat this step, increasing the degrees of freedom by 1 each step, until the *P*-values no longer decrease.

RESULTS

We demonstrate the potential of our approach using a simple simulation-based power study. Our goal is to compare the performance of analyses of SFP data with that of corresponding tag SNPs (SNPs that capture the variation in regions of high LD; Johnson *et al.* 2001) and to further investigate how this comparison depends upon the accuracy with which SFPs are detected. SFP data represent an imperfect summary of the SNP data within the corresponding probe region. The imperfect nature of this summary (caused by less than optimal sensitivity and specificity) leads to a degradation of the signal on a *per SNP* basis. However, SFPs are more economical than tag SNPs, which will typically allow the use of a much higher density of SFP probes than would be possible using SNP data. Thus, it is hoped that the increased signal due to the higher density of SFP information exceeds the loss of signal due to the imperfect nature of each individual SFP, giving an SFP analysis the potential to be more powerful than a comparable SNP-based study.

We simulate the underlying SNP data using a coalescent model. From each such data set we construct test data according to three scenarios: first, we select tag SNPs (Zhang *et al.* 2005); second, we construct SFP data using a calling scheme that is completely dependent across sequences; and third, we construct SFP data under a calling scheme that is completely independent across sequences. The latter two represent best- and worst-case comparisons. As we have shown, SFPs appear to be called with relatively strong (but clearly not complete) dependence, so results for actual data are likely to lie between these two extremes.

Using Hudson's ms program (Hudson 2002), we simulate 100 replicate data sets of a 1-Mb region with constant mutation rate $\theta = 0.005$/bp and recombination rate $\rho = 0.0002$/bp (these numbers are chosen to be appropriate for a 1-Mb region in the *A. thaliana* genome; see Nordborg *et al.* 2005). Each replicate set consists of 400 haploid individuals. To model the SFP data, we select one accession at random as the reference sequence and then "call" SFP genotypes (*i.e.*, model the process by which probes detect SNPs), using the empirically derived estimates of specificity of $s_p = 0.98$ and a deliberately conservative sensitivity of $s_n = 0.5$ (*cf.* Borevitz *et al.* 2003). Our simulation leads to 18% of the SFP probes containing multiple mutations, compared to 16.7% in actual data. Two types of statistical SFP calling were modeled: (1) independent calling, which assumes that each probe is called independently across haplotypes, and (2) dependent calling, which assume that samples with the same true underlying SNP

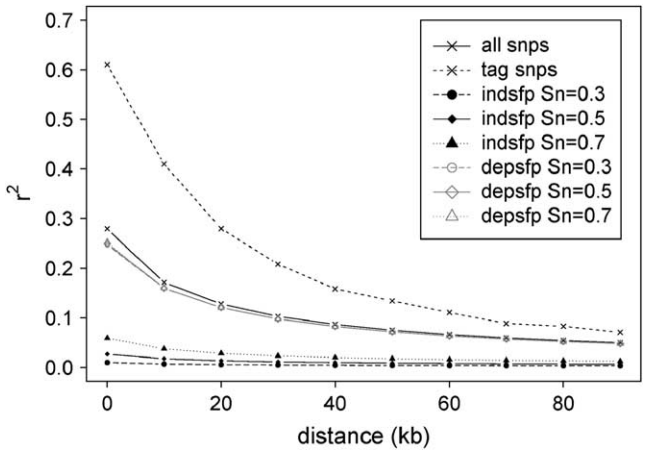

FIGURE 2.—Decay of LD according to type of data. Crosses are SNP data, solid symbols are independently called SFPs, and open symbols are dependently called SFPs.

state will have the same call across sequences for the given probe.

Phenotypes were generated to mimic a scheme analogous to the effects of the vernalization locus *FRI* (see NORDBORG *et al.* 2005, for details) on flowering time. We begin by selecting a single SNP. If an accession contains the minor allele we sample from a "late-flowering" phenotypic distribution of $N(80, 46^2)$. If the accession contains the major allele we sample from an "early-flowering" phenotypic distribution of $N(32, 6^2)$ with probability $1 - p$; otherwise we again sample from a distribution of $N(80, 46^2)$. We refer to p as the heterogeneity parameter. This construction reflects a situation in which multiple other mutations (outside the region of interest) or environmental effects are also influencing the phenotype. The causal SNP was chosen within a specified range of allele frequencies and was required to be located close to the center of the simulated genomic region.

For each replicate set, tag SNPs were selected from a randomly ascertained subsample of 20 individuals (this number is based on an ongoing genomewide resequencing study in *A. thaliana* that would have to serve as source of SNPs for an actual SNP study). We utilize the HapBlock program of ZHANG *et al.* (2005) to identify tag SNPs on the basis of "haplotype diversity" (JOHNSON *et al.* 2001). We force the total number of tag SNPs to be ~50 to create a desired marker density of 1 every 20 kb, which we have argued might give reasonable power given the decay of LD in *A. thaliana* (ARANZANA *et al.* 2005). The average numbers of markers across all 100 replicate sets with singletons removed were 49, 27,959 and 8127, for tag SNPs, independent SFPs, and dependent SFPs, respectively. Tag SNPs gave an average marker density of 1/20 kb while SFPs with independent and dependent calling gave ~1/35 bp and 1/120 bp, respectively. Figure 2 shows the decay of LD measured under the various scenarios.

TABLE 2

Power ($Y = 2$, 5, and 10) conditioned on trait minor allele frequency (MAF) for tag-SNPs

	Kruskal–Wallis			CLASS		
MAF	2	5	10	2	5	10
0.05–0.10	0.36	0.53	0.72	0.40	0.51	0.71
0.10–0.15	0.52	0.65	0.76	0.51	0.71	0.80
0.15–0.20	0.56	0.73	0.82	0.64	0.77	0.89
0.20–0.25	0.65	0.79	0.93	0.79	0.86	0.93
0.25–0.30	0.63	0.86	0.90	0.76	0.88	0.96
0.30–0.35	0.77	0.84	0.92	0.83	0.89	0.93
0.35–0.40	0.84	0.93	0.95	0.94	0.94	0.96
0.40–0.45	0.75	0.87	0.93	0.88	0.95	0.97
0.45–0.50	0.85	0.90	0.95	0.91	0.96	0.96
Null rate	0.11	0.25	0.37	0.12	0.28	0.49

We then applied the fragment-based Kruskal–Wallis and CLASS algorithms to these simulated data. We examined the power to detect a causal variant for the three different data types using both tests. In addition, we examine the effects of the trait minor allele frequency. We measure power as the probability that one of the Y markers with the most extreme test statistic values falls within a window of a given size centered around the position of the true functional locus. Our definition is, of course, completely arbitrary, but is sufficient for a proof-of-principle analysis such as this, while also allowing us to compare the relative performance of the different data types. We set $Y = 2$, 5, and 10, and use a window size of 25 kb for tag SNP data type and 10 kb for SFP data types (note that this results in a more generous criterion for the tag-SNP data). Simulations in which no tag SNP, or SFP, is within an appropriate window of the functional mutation are not included in the power calculation for that data type. We restrict our attention to markers with minor allele frequency >5%. Tables 2–4 present results of our power study for tag SNPs and for independent and dependent SFP calling (respectively). We give the power for each test statistic and data type conditioned on minor allele frequency at the QTL. In the row labeled "null rate" we show results from analyses in which phenotypic values were randomly permuted among the samples before analysis. This gives an indication of a null or false-positive rate with which loci with extreme *P*-values are found in a given window size even if there is no QTL present.

We find that SFP data allow us to perform fine mapping successfully and that for our particular choice of scenarios they typically give more power for both methods. It is striking that, despite the very substantial loss of LD in independently called SFP data (see Figure 2), our methods are still able to extract a signal from such data. Given the current, somewhat simplistic forms of our SFP mapping algorithms, these results give cause for some optimism. The relative noisiness of each SFP (compared

TABLE 3

Power ($Y = 2$, 5, and 10) conditioned on trait minor allele frequency (MAF) for independently called SFPs

	Kruskal–Wallis			CLASS		
MAF	2	5	10	2	5	10
0.05–0.10	0.44	0.67	0.78	0.42	0.63	0.73
0.10–0.15	0.54	0.75	0.85	0.58	0.71	0.85
0.15–0.20	0.68	0.82	0.92	0.66	0.83	0.93
0.20–0.25	0.84	0.91	0.96	0.85	0.93	0.97
0.25–0.30	0.86	0.95	0.97	0.83	0.96	0.97
0.30–0.35	0.88	0.97	0.99	0.91	0.95	0.99
0.35–0.40	0.89	0.97	1.00	0.90	0.97	1.00
0.40–0.45	0.87	0.98	1.00	0.90	0.97	1.00
0.45–0.50	0.92	0.99	1.00	0.93	0.98	1.00
Null rate	0.05	0.10	0.18	0.09	0.19	0.33

to SNPs) is more than compensated for by their greater density.

As shown in Figure 3a, similar patterns of power are observed when we vary the heterogeneity parameter p (expressed as a percentage) for minor allele frequencies in the range of 10–15% and $Y = 10$ for all mapping methods and marker types. Dependent SFPs show the greatest power and SFPs consistently outperform tag SNPs. We explore the power as a function of error rates by setting S_n-values from 0.3 to 0.7 incremented by 0.2 for a putative trait minor allele frequency between 10 and 15% and $Y = 10$. Increasing the sensitivity will reduce the noise of the data and thus is expected to provide more power. Figure 3b illustrates this.

Figure 4 shows representative output from a single run using each algorithm and analyzing the data with independently called SFPs and a minor allele frequency of 0.10–0.15. In both cases we clearly see the signal above the level of noise. As we have shown, in reality SFPs are called with a good degree of dependence, so the signal-to-noise ratio will be much higher for real data.

TABLE 4

Power ($Y = 2$, 5, and 10) conditioned on trait minor allele frequency (MAF) for dependently called SFPs

	Kruskal–Wallis			CLASS		
MAF	2	5	10	2	5	10
0.05–0.10	0.64	0.78	0.86	0.67	0.80	0.87
0.10–0.15	0.76	0.85	0.89	0.96	0.99	1.00
0.15–0.20	0.80	0.90	0.96	0.99	0.99	1.00
0.20–0.25	0.92	0.95	0.98	1.00	1.00	1.00
0.25–0.30	0.91	0.93	0.95	0.99	0.99	0.99
0.30–0.35	0.91	0.97	0.97	1.00	1.00	1.00
0.35–0.40	0.94	0.97	0.99	0.99	0.99	0.99
0.40–0.45	0.98	1.00	1.00	0.99	0.99	0.99
0.45–0.50	0.94	0.98	1.00	1.00	1.00	1.00
Null rate	0.04	0.08	0.12	0.02	0.05	0.07

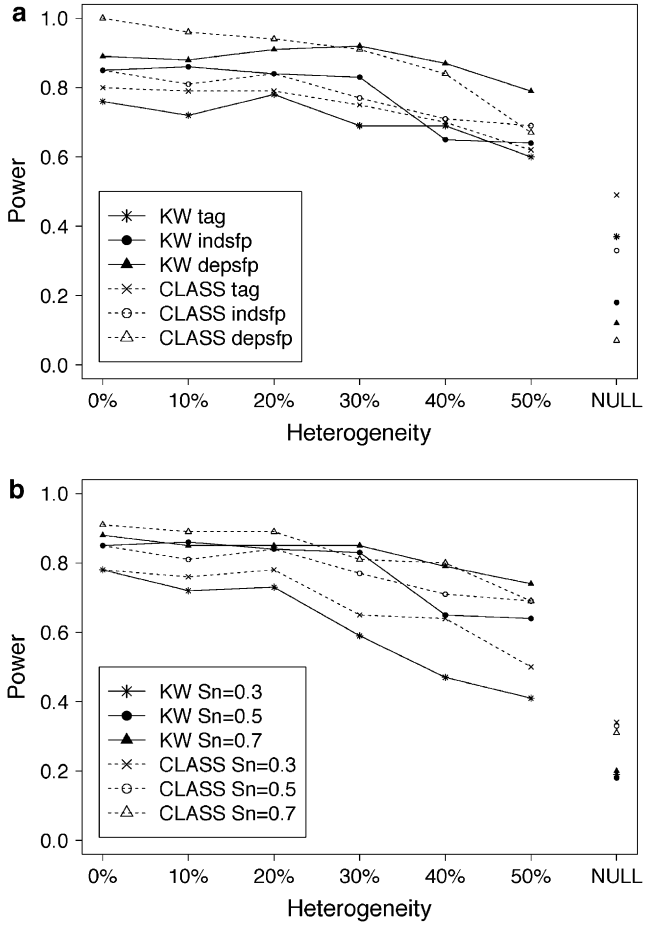

FIGURE 3.—Power conditioned on the heterogeneity parameter p for minor allele frequency in the 10–15% range and $Y = 10$. (NULL indicates the false-positive rate.) (a) Power shown for Kruskal–Wallis (KW) and CLASS methods on tag SNPs, independent SFPs, and dependent SFPs. (b) Power shown for KW and CLASS with varying sensitivity (S_n) on independent SFPs.

DISCUSSION

Our goal here has been to demonstrate that SFPs are a viable alternative to SNPs for genomewide associations studies in *A. thaliana*. Our methodology might also be applied to inbred lines for other organisms, such as yeast or mouse. SFPs are less likely to be effective in organisms like *Drosophila melanogaster*, where LD often decays over hundreds of base pairs, although it should be noted that the SFP data are sufficiently dense that the causal polymorphism itself will often be detected—the ideal situation for association mapping (RISCH and MERIKANGAS 1996).

SFPs have several advantages over SNPs. Perhaps the most important one is cost and/or convenience. No specialized genotyping equipment is required and no SNP development is required. Individuals can be assayed using a single high-density array (the cost per array is currently ~$425). However, in addition, SFPs are unbiased (in the sense that no SNP ascertainment is involved, other than for the reference sequence) and encompass

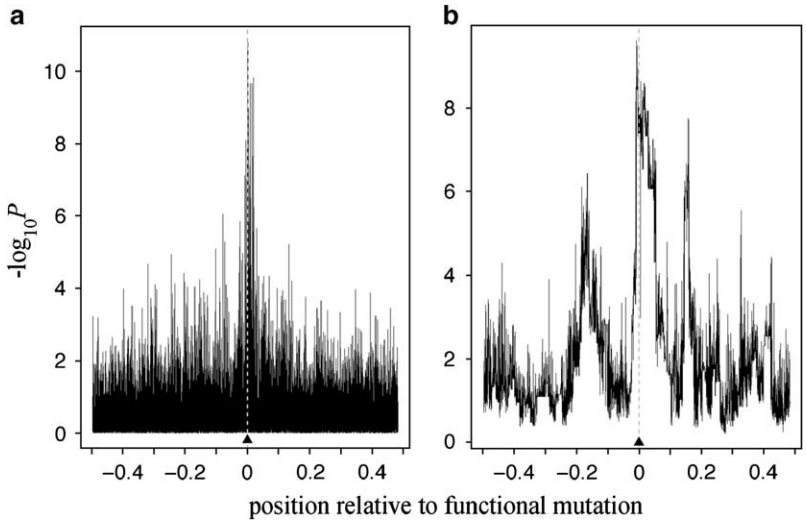

FIGURE 4.—Examples of output from our fragment-based Kruskal–Wallis (a) and CLASS (b) methods when run on simulated SFP data (independent calling). Each plot contains the output of a single analysis. The dashed vertical line and associated arrowhead indicate the position of the functional polymorphism, which in these simulations had minor allele frequency in the 10–15% range. The *x*-axis corresponds to ~1 Mb.

several different types of polymorphisms, including repeat-length polymorphisms and larger insertion–deletions. We recently showed how causative deletions can be revealed by SFP analysis (WERNER *et al.* 2005).

Clearly, the methods in this article, although successful, are somewhat simplistic. For example, in the future, it is also possible that one would have site- or sequence-specific estimates of sensitivity and specificity from an external data source, in which case it would then be a simple matter to extend the methods we present here to reflect those estimates. One might also explicitly allow for the possibility of multiple SNPs occurring within a single SFP probe, although our article suggests that this is not a prerequisite for successful analysis.

The authors also gratefully acknowledge the comments of the reviewers and their role in improving this manuscript. This work was funded mainly by National Institutes of Health grants GM069890-01A1 and HG002790-01A1. The sequencing data were produced by National Science Foundation grant DEB-0115062.

LITERATURE CITED

ARANZANA, M. J., S. KIM, K. ZHAO, E. BAKKER, M. HORTON *et al.*, 2005 Genome-wide association studies in *Arabidopsis thaliana* identify previously known genes responsible for variation in flowering time and pathogen resistance. PLoS Genet. **1:** e60.

BOREVITZ, J., 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.

DALY, M., J. D. RIOUX, S. F. SCHAFFNER, T. J. HUDSON and E. S. LANDER, 2001 High-resolution haplotype structure in the human genome. Nat. Genet. **29:** 229–232.

DURRANT, C., K. T. ZONDERVAN, L. R. CARDON, S. HUNT, P. DELOUKAS *et al.*, 2004 Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. **75:** 35–43.

GABRIEL, S. B., S. F. SCHAFFNER, H. NGUYEN, J. M. MOORE, J. ROY *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

GRAHAM, J., and E. A. THOMPSON, 1998 Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am. J. Hum. Genet. **63:** 1517–1530.

HAGENBLAD, J., C. TANG, J. MOLITOR, J. WERNER, K. ZHAO *et al.*, 2004 Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. Genetics **168:** 1627–1638.

HINDS, D. A., L. L. STUVE, G. B. NILSEN, E. HALPERIN, E. ESKIN *et al.*, 2005 Whole genome patterns of common DNA variation in three human populations. Science **307:** 1072–1079.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, London/New York/Oxford.

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

INTERNATIONAL HAPMAP CONSORTIUM, 2005 A haplotype map of the human genome. Nature **437:** 1299–1320.

JEFFREYS, A. J., R. NEUMANN, M. PANAYI, S. MYERS and P. DONNELLY, 2005 Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. **37:** 601–606.

JOHNSON, G. C., L. ESPOSITO, B. J. BARRATT, A. N. SMITH, J. HEWARD *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nat. Genet. **29:** 233–237.

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

KITTLES, R. A., J. C. LONG, A. W. BERGEN, M. EGGERT, M. VIRKKUNEN *et al.*, 1999 Cladistic association analysis of Y chromosome effects on alcohol dependence and related personality traits. Proc. Natl. Acad. Sci. USA **96:** 4204–4209.

LIU, J. S., C. SABATTI, J. TENG, B. J. B. KEATS and N. RISCH, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. **11:** 1716–1724.

MANIATIS, N., A. COLLINS, J. GIBSON, W. ZHANG, W. TAPPER *et al.*, 2004 Positional cloning by linkage disequilibrium. Am. J. Hum. Genet. **75:** 846–855.

MANIATIS, N., N. E. MORTON, J. GIBSON, C.-F. XU, L. K. HOSKING *et al.*, 2005 The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. Hum. Mol. Genet. **14:** 145–153.

MARCHINI, J., L. CARDON, M. PHILLIPS and P. DONNELLY, 2004 The effects of human population structure on large genetic association studies. Nat. Genet. **36:** 512–517.

MOLITOR, J., P. MARJORAM, D. CONTI, D. STRAM and D. THOMAS, 2004 A survey of current Bayesian gene mapping methods. Hum. Genomics **1:** 371–374.

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

MOLITOR, J., P. MARJORAM and D. THOMAS, 2003b Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. Genet. Epidemiol. **25**(2): 95–105.

MOLITOR, J., K. ZHAO and P. MARJORAM, 2005 Fine mapping—19th century style. BMC Genet. **6**(Suppl. 1): S63.

MORRIS, A. P., J. C. WHITTAKER and D. J. BALDING, 2000 Bayesian fine-scale mapping of disease loci, by hidden Markov models. Am. J. Hum. Genet. **67:** 155–169.

Morris, A. P., J. C. Whittaker and D. J. Balding, 2002 Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am. J. Hum. Genet. **70:** 686–707.

Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321–324.

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

Patil, N., A. J. Berno, D. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1719–1723.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding *et al.*, 2003 Chromosome-wide distribution of haplotype blocks and role of recombination hot spots. Nat. Genet. **33:** 382–387.

Rinaldo, A., S.-A. Bacanu, B. Devlin, V. Sonpar, L. Wasserman *et al.*, 2005 Characterization of multilocus linkage disequilibrium. Genet. Epidemiol. **28:** 193–206.

Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science **273:** 1616–1617.

Roeder, K., S.-A. Bacanu, V. Sonpar, X. Zhang and B. Devlin, 2005 Analysis of single-locus tests to detect gene/disease associations. Genet. Epidemiol. **28:** 207–219.

Rostoks, N., J. O. Borevitz, P. E. Hedley, J. Russell, S. Mudie *et al.*, 2005 Single-feature polymorphism discovery in the barley transcriptome. Genome Biol. **6:** R54.

Siegel, S., and N. J. Castellan, Jr., 1988 *Nonparametric Statistics for the Behavioral Sciences*, Vol. 2. McGraw-Hill, New York.

Sillanpaa, M. J., and M. Bhattacharjee, 2005 Bayesian association-based fine mapping in small chromosomal segments. Genetics **169:** 427–439.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

Stumpf, M. P., and D. B. Goldstein, 2003 Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr. Biol. **13:** 1–8.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

Templeton, A. R., 1995 A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics **140:** 403–409.

Templeton, A. R., E. Boerwinkle and C. F. Sing, 1987 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the Apoprotein E locus. Genetics **117:** 343–351.

Templeton, A. R., T. Maxwell, D. Posada, J. H. Stengard, E. Boerwinkle *et al.*, 2005 Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. Genetics **169:** 441–453.

Turner, T. L., M. W. Hahn and S. V. Nuzhdin, 2005 Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. **3**(9): e285.

Tzeng, J.-Y., 2005 Evolutionary-based grouping of haplotypes in association analysis. Genet. Epidemiol. **28:** 220–231.

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.

Winzeler, E. A., D. R. Richards, A. R. Conway, A. L. Goldstein, S. Kalman *et al.*, 1998 Direct allelic variation scanning of the yeast genome. Science **281:** 1194–1197.

Zhang, K., Z. Qin, T. Chen, J. S. Liu, M. S. Waterman *et al.*, 2005 Hapblock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. Bioinformatics **21:** 131–134.

Zollner, S., and J. K. Pritchard, 2005 Coalescent-based association mapping and fine mapping of complex trait loci. Genetics **169:** 1071–1092.

Communicating editor: G. Gibson