Chapter 7

Additional Applications of the Algorithm

In this chapter I describe some additional applications of the method to questions in population genetics. These descriptions are not intended to be rigorous, but rather give illustrative examples towards how inferred ARGs could be used.

7.1 Detecting Selective Sweeps

A popular technique for identifying selective sweeps from population SNP data is to look for tracts of unexpectedly long haplotype sharing, called extended haplotype homozygosity (Sabeti et al., 2002; Nielsen et al., 2005; Hanchard et al., 2006; Voight et al., 2006). When a favoured allele increases rapidly in frequency, it will tend to reside on an unusually long haplotype of low diversity. This is because there has not been sufficient time for recombination to break down the LD. Meanwhile, chromosomes that do not carry the selected allele will tend to have levels of diversity and LD that are more typical of the genome as a whole. By comparing the rate of haplotype breakdown between one haplotype at a locus and the others at the same locus, it is possible to detect recent positive selective sweeps where the selected alleles have not yet reached fixation.

However, Reed and Tishkoff (2006) show that allele-specific recombination hotspots can leave similar LD patterns to those just described, resulting in false positive signals. Specifically, an allele-specific recombination hotspot could result in reduced haplotype sharing for those chromosomes with that allele, which might then be interpreted as evidence for positive selection on the alternative allele.

If the true ARG for the data were known, it would be possible to distinguish between selective sweeps and allele-specific recombination hotspots. Three measures of an allele are encoded in the ARG: its age, its frequency and the number of recombination events occurring around it. A selective sweep would show up as a young allele with high frequency, while an allele-specific recombination hotspot would show up as an increased intensity of recombination events under the mutation.

Of course, the true ARG is unknown, and my current ARG inference algorithm uses haplotype homozygosity to order recombination and coalescence events, and hence is unlikely to powerfully distinguish between allele-specific recombination hotspots and selective sweeps.

Nevertheless, below I show that as it currently stands, the algorithm detects a signal indicative of positive selection at the Lactase gene, a known example of strong positive selection. This is done by comparing the frequencies of SNPs to their ages, where the relative ages of alleles are estimated using ARGs.

In most human populations, except those of European descent, the ability to digest lactose contained in milk disappears in childhood. Strong signals of selection have been identified in the Lactase gene (LCT) in European populations (Bersaglieri et al., 2004), agreeing with the hypothesis that selective advantage was gained by those with the ability to digest lactose as adults. Indeed, the selective signal at LCT is one of the strongest in the whole genome (The International HapMap Consortium, 2005; Voight et al., 2006).

In order to test the preliminary approach, I applied it to two approximately 1 Mb regions: 232 SNPs around the LCT gene on Chromosome 2 from the Phase 1 CEU HapMap, and a randomly selected region on Chromosome 2, again of 232 SNPs from the Phase 1 CEU HapMap. I then inferred 100 ARGs for the regions, and for each SNP calculated its frequency count (the number of the 120 CEU independent haplotypes possessing that allele) divided by its average age order (age 1 corresponds to the most recent SNP to be mutated in the ARG,

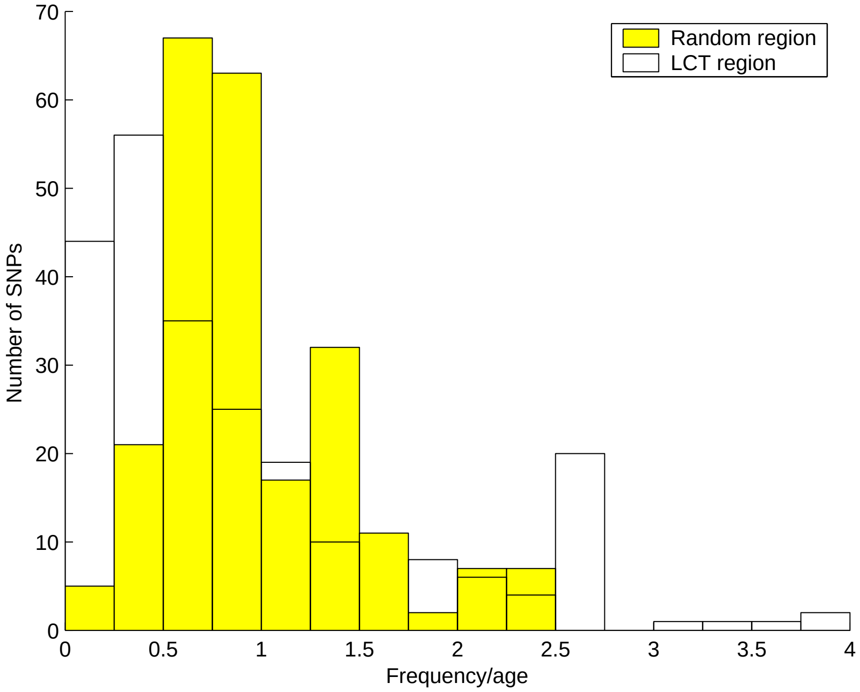

Figure 7.1: 232 SNPs in two 1 Mb regions (yellow = random region in Chromosome 2; clear = LCT region), their frequency count divided by their age ordering (averaged over 100 ARGs) in 60 unrelated CEU individuals.

and working backwards in time, age 232 corresponds to the SNP mutated highest up in the ARG).

Figure 7.1 shows the frequency/age distributions for these two regions. There is an excess of SNPs with high frequency/age in the LCT region, indicative of a positive selective sweep.

This analysis could be conducted for the whole genome, and then regions with SNPs in the tail of the frequency/age distribution could be considered as candidates for selection.

The comparison of frequency against age is what is required for identification of positive selection. In extended haplotype homozygosity tests, haplotype sharing is used as a proxy for age, as haplotype sharing decays with recombination over time. Hence, it may be expected that the direct comparison of frequency and age described above will have at least as much power to detect selection. However, it should be noted that the ARGs are inferred on the basis of shared segments, and so the differences between extended haplotype homozygosity and this approach may not be so marked.

It may be a good idea to not only consider the frequency/age distribution of typed SNPs, but also of all branches of the inferred ARGs. This may have greater power to detect selection events when the alleles under selection are not typed, similar to Chapter 3 with fine mapping.

7.2 Sequence Imputation

It may soon be routine to resequence individuals for the purposes of association studies and other population genetic analyses (Balding, 2005; Romeo et al., 2007). However, full resequencing of many individuals is redundant because of LD structure, and while low coverage resequencing results in missing data and errors, it should be possible to impute missing genotypes and correct errors using an enhanced Margarita system. This could reduce the amount of genotyping effort required per individual, allowing more individuals from a greater number of populations to be sampled.

Imputing missing genotype data has been explored in Chapter 6. However, imputing sequence data is a slightly different problem for the following reasons:

- Resequencing data consists of nucleotides with quality scores attached to them, quantifying how accurate they are;
- For an individual, there will be contiguous tracts (reads) of observed nucleotides, and similar tracts which are entirely missing;
- The data will not be biallelic;
- There will be additional complexities such as copy number polymorphisms and rearrangements, which can lead to alignment errors.

To deal with these issues a sequence imputation system would need to incorporate sequence quality data, otherwise sequencing errors may mislead the ARG inference algorithm by giving evidence for recombination. One way to do this would be to adapt the shared segment calculation so that some mismatches are permitted when the quality score for a mismatch is low compared with surrounding evidence for a shared segment, suggesting that it is more likely to be an error than a genetic difference. The probability of a match or a mismatch is dependent on the quality scores of the two sequences at that position. So for each pair of

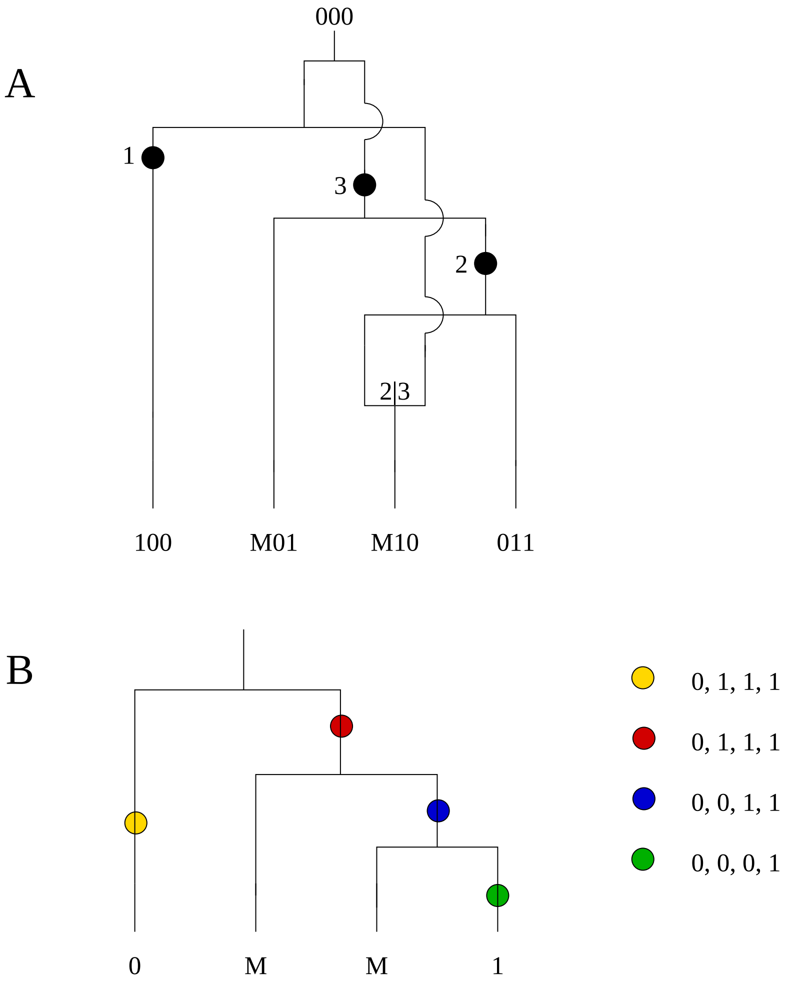

Figure 7.2: The positioning of the mutation affects the imputation.

sequences at a marker, a score could be calculated, such as the log odds of a match versus a mismatch. Maximal scoring shared segments could then be found (Ruzzo and Tompa, 1999) and used to guide the recombination and coalescence operations, as in the original algorithm.

When there are a large number of missing nucleotides, it may also be useful to enhance the way in which mutations are placed on the ARG. Figure 7.2 shows an example ARG where there are missing data. Given the ARG (Figure 7.2.A), both missing data points, denoted as Ms, would be imputed as 0s. However, considering the marginal tree (Figure 7.2.B) for the position with missing data, we see that there are four places where the mutation could be placed on the tree (denoted by the different coloured mutations), yielding four different imputations.

All four of these mutations fit legally into the ARG. The algorithm as described in Chapter 2 would give either the imputation represented by the yellow mutation, or the imputation represented by the red mutation. This is because a mutation is placed as early as possible, but after all missing data at that position is resolved. Nevertheless, the other two imputations are also valid, and it may be possible to increase imputation accuracy by considering them.

In collaboration with D. Carter at the Wellcome Trust Sanger Institute (Carter et al., 2007), the Margarita system has been extended. It now accommodates sequencing errors—early on in the ARG construction—by permitting mismatches in the shared segments. We also compute branch lengths for the ARG, and allow the implicit repositioning of mutations via use of the Felsenstein algorithm (Felsenstein, 1981) for calculating the probability of each nucleotide at the leaves.

In order to estimate branch lengths, we:

1. Throw out the explicit representation of mutation nodes in the ARG and then count the number of mutation events between coalescence and recombination events—or "nodes". This gives the number of mutation events on each edge of the ARG.

2. Calculate the active region for each edge of the ARG, that is the region of genetic material which is defined for that edge.

3. Estimate the ages of nodes in the ARG using a molecular clock assumption. This is done by maximising the likelihood of the number of mutations seen on each edge, subject to retaining the same global order of coalescence and recombination events. If an edge connects nodes with ages t_1 and t_2, and has an active region of length L, then the likelihood of it having k mutations is Poisson with mean $(t_2 - t_1)L$. An initial guess of the node ages is made from the coalescent-with-recombination, and ages are then updated by taking random horizontal "slices" through the ARG and allowing the ages of nodes above the slice to vary by a constant t, where t is chosen to maximise the joint likelihood of the mutation counts on all the edges cut by the slice. This update is performed several thousand times.

We then use the Felsenstein algorithm to assign a posterior probability to each nucleotide on the leaf sequences, thereby correcting sequencing errors, as follows:

1. The marginal tree is extracted for each genomic interval between recombinations in the ARG.

2. We then use the Felsenstein algorithm (Felsenstein, 1981) to calculate the probability of a particular nucleotide at each locus and leaf, given the tree and branch lengths. Since this only uses the tree topology and lengths, it implicitly integrates over all permitted repositionings of mutations, as suggested in Figure 7.2. (In fact, it allows for multiple mutation events, relaxing the infinite sites assumption.)

3. The imputed nucleotide for a particular sequence and locus is the one with the highest posterior probability, averaged over multiple inferred ARGs; and the probability of error is the sum of the mean probabilities for the other possible nucleotide values.

We tested the quality of sequence imputation using *S. cerevisiae* resequencing data. Sanger shotgun sequencing was undertaken on 37 haploid strains at a coverage depth of 0.7 to 4.1, with mean depth of 1.64. We held out at random 10% of the read pairs and imputed their values using the above imputation procedure and the remaining data. Because the data is of varying quality, we only compared the imputed values to the assayed values with high quality scores. The system gives a mean error rate of 0.00126 on this data at polymorphic sites (sites which are monomorphic are trivial to impute, although some sites may appear monomorphic at low coverage which are in fact polymorphic).

We also evaluated this approach on simulated resequencing data, derived from the FREGENE population of Hoggart et al. (2005) used in Chapter 3, and matched to the sequencing characteristics of the *S. cerevisiae* data: read length, coverage and error probability. The results are shown in Table 7.1.

If high quality sequence is considered to have no more than 100 errors per million, then 7.1 gives us some idea of how this can be achieved with low coverage, imputation and error correction. The sequencing capacity which this frees can then be applied to resequencing more individuals, which has the important benefit of allowing more complete SNP discovery.

	Number of haplotypes					
Coverage	6	12	24	50	100	200
0.5	519	403	243	139	85	59
1.0	286	177	109	62	37	26
2.0	110	73	37	24	15	12
3.0	58	38	22	14	9	7

Table 7.1: Number of errors per million nucleotides for simulated resequencing data.

7.3 Detecting Population Substructure

Another question of interest in population genetics is that of identifying population substructure within data (Pritchard et al., 2000), either for the purpose of making demographic inferences, or for correcting for population effects in case-control studies (Clayton et al., 2005).

For the *S. cerevisiae* data, it may be of interest to identify regions of the genome where strains cluster together; for example, where those used for baking cluster, indicating that those parts of the genome may have been selected for. In Figure 7.3, I construct an ARG for 38 strains of *S. cerevisiae*, sequenced for Chromosome 1. In order to calculate the distance between strains, I take the average tree traversal distance between strains. For a particular marginal tree, the tree traversal distance between two strains is the number of coalescence events that must be traversed when travelling the path from the leaf corresponding to one of the strains to the other, divided by the maximum traversal distance, which is $n-1$, where n is the number of sequences. This is averaged over all the polymorphic sites for Chromosome 1. The next step would be to analyse the clustering locally for patterns indicative of selection.

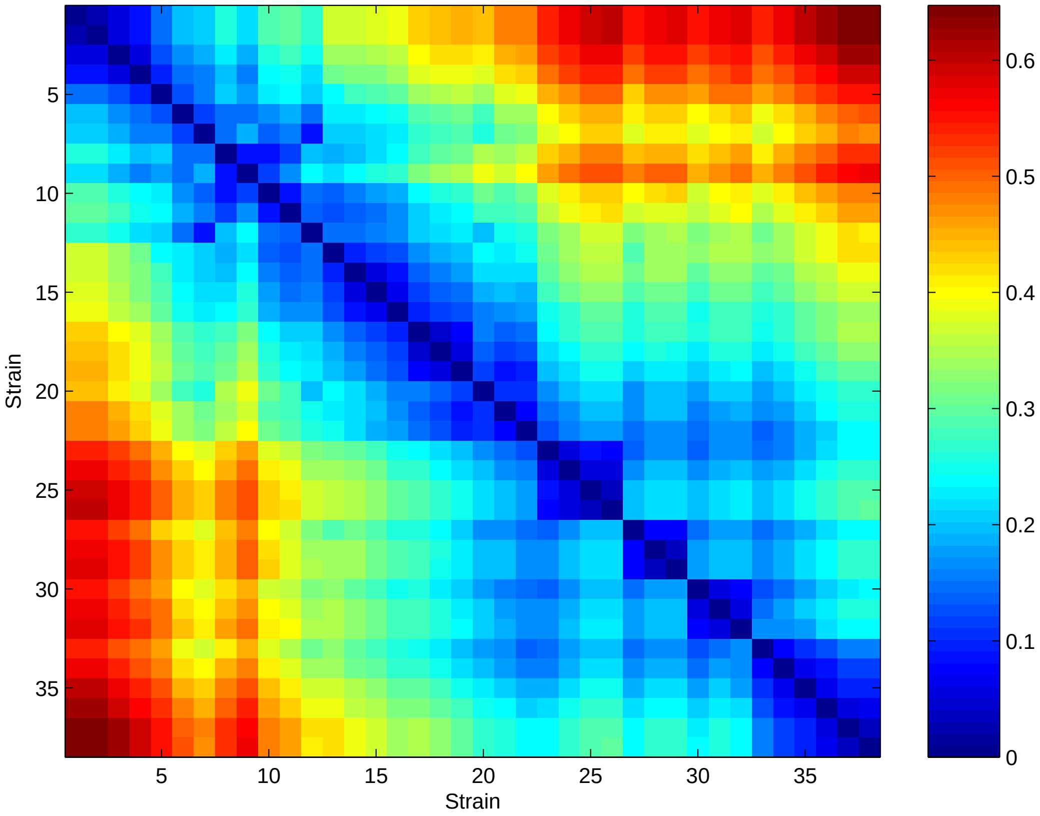

Figure 7.3: Pairwise tree traversal distances for 38 *S. cerevisiae* strains.