Chapter 8

Conclusions

In this thesis I have introduced a new method for constructing Ancestral Recombination Graphs from population genetic data. The ARG describes the history of a sample of sequences according to mutation, recombination and coalescence, and many analyses based on those features could be assisted by use of inferred ARGs. I have demonstrated this principally for disease mapping.

Analysing current case-control association data, involving thousands of individuals typed across the genome, requires an analysis method to be computationally efficient. This has been achieved by taking a heuristic approach to ARG inference—Margarita does not sample statistically from the coalescent-with-recombination, nor does it search for the ARG with the minimum number of obligate recombination events, but it does appear to capture some worthwhile features. This makes the method the first ARG mapping approach that can be applied to realistically sized data sets.

In order to detect disease associations, the local genealogies for loci, which are embedded in the ARG, are examined for a clustering of disease cases under a particular branch in the genealogy. Such a clustering suggests that a causative mutation may have occurred on that branch. In simulation studies, I found that this approach gives increased power to detect causative alleles after correcting for multiple testing by case-label permutation, and more accurate positioning of them, compared with the single marker chi-square test and a haplotype based method. Part of the increase in power after correcting for multiple testing is due to the genealogies at nearby loci being more strongly correlated than r^2 linkage disequilibrium, meaning that nearby tests using my mapping method are more strongly correlated than tests with the chi-square. This results in a reduction in the multiple testing burden. This also gives more accurate positioning of causative loci; the decay in association with the ARG test is due to recombination events, whereas for the chi-square test, association is also affected by the relative timing of mutation events.

In addition to any increase in power and localisation, having an explicit estimate of the genealogy of a locus allows properties of untyped causative alleles to be inferred. On simulation studies, I showed that the inferred ARGs could be used successfully to infer the frequencies of untyped alleles.

In collaboration, I applied the ARG mapping method to two case-control studies. One of association between a 300kb region and Graves disease (Ueda et al., 2003), and the other of association between an 800kb region and prostate cancer (Yeager et al., 2007). In both cases, it was the additional interpretative power given by the ARG that proved to be the most useful aspect of the approach. For the Graves disease data, it was observed that there are multiple clusterings of disease individuals on the ARG, suggesting that those different clusters correspond to different causative alleles. This led to the identification of a weak signal of possible epistatic interaction. In the prostate cancer data the inferred ARGs helped show that the genealogies of two nearby association peaks are decorrelated due to a recombination hotspot, and thus correspond to independent signals. Analysis of the haplotypes extending around the association peaks showed a possible signal of selection.

Detecting disease association with the ARG approach is a missing data problem, where the untyped causative polymorphism is being imputed. In order to take a more direct approach, I considered using the ARG inference algorithm to impute missing values at known loci, specifically, loci which are untyped in the case-control data, but which are typed in a denser sample, such as the HapMap. The performance of the ARG approach was shown to be competitive with fastPHASE (Scheet and Stephens, 2006) as far as the quality of genotype imputation is concerned. However, it is not obvious from the experiments that this approach is useful in practice. There have been a number of signals detected in the 8q24 region showing association with prostate cancer, and when I attempted to impute these using the Yeager et al. (2007) data combined with the CEU HapMap, the additional signals that had not previously been seen in the data set were not found.

However in the near future it may become routine to resequence individuals for disease studies (Balding, 2005; Romeo et al., 2007), in which case, any increases in power and localisation achieved by the ARG methods described here, over single marker tests, will be limited. When this situation arises, it is likely that the search for heterogeneity and epistatic interactions will begin in earnest, hence, the ARG approach will still be useful, as shown in my analysis of the CTLA4 data in Chapter 4. In fact, for that study (Ueda et al., 2003), the region was searched for new polymorphisms, and all of them were typed, hence it is arguable that this data is very similar to resequencing data; and still, the ARG approach was able to draw additional interesting inferences.

Indeed, as resequencing technology undergoes maturation, an important additional application of ARG inference could be to imputing missing nucleotides in resequencing data. Depending on the sequencing coverage there will be tracts of contiguous nucleotides which are observed, and other tracts that are completely unobserved. After appropriate modification, and coupled with some further enhancements, Margarita was applied to a population of *S. cerevisiae* sequences. The experiments show that resequencing can be performed at low coverage, and linkage disequilibrium relied upon in order to fill in the missing data. This is important because allowing resequencing to take place at lower coverage means that more individuals, from a wider range of populations, can be resequenced, allowing more complete SNP discovery.

Furthermore, many population genetic questions will still require sophisticated methods even when all nucleotides are assayed. As mentioned above, inferred ARGs may be useful in many analyses which rely on interpreting the recombination and mutation history. I gave some initial suggestions indicating how inferred ARGs could potentially be used to detect selective sweeps and population substructure.