Analysis of genetic variation data using Ancestral Recombination Graphs

Mark J. Minichiello,
Gonville and Caius College,
University of Cambridge.

This dissertation is submitted for
the degree of Doctor of Philosophy

04 July 2007

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and Acknowledgements. No part of this dissertation has been submitted for a degree or diploma or other qualification at this or any other university.

Mark J. Minichiello
Cambridge
04 July 2007

Summary

Analysis of Genetic Variation Data using Ancestral Recombination Graphs

Mark J. Minichiello

Large-scale association studies are being undertaken with the goal of uncovering the genetic determinants of complex disease. In this thesis I describe a computationally efficient method for inferring genealogies from population genotype data and show how these genealogies can be used to fine map disease loci and interpret association signals.

These genealogies take the form of the ancestral recombination graph (ARG). The ARG defines a genealogical tree for each locus, and, as one moves along the chromosome, the topologies of consecutive trees shift according to the impact of historical recombination events.

There are two stages to the analysis. First, I infer plausible ARGs using a heuristic algorithm, which can handle unphased and missing data, and is fast enough to be applied to large-scale studies. Second, I test the genealogical tree at each locus for a clustering of the disease cases beneath a branch, suggesting that a causative mutation occurred on that branch. Since the true ARG is unknown, I average this analysis over an ensemble of inferred ARGs.

I characterise the performance of the method across a wide range of simulated disease models. Compared with simpler tests, the method gives increased accuracy in positioning untyped causative loci and can also be used to estimate the frequencies of untyped causative alleles.

I apply the method to Ueda et al.'s association study of CTLA4 and Graves disease, and Yeager et al.'s association study of 8q24 and prostate cancer, showing how it can be used to dissect association signals. With the CTLA4 data, the method suggests a possible signal of allelic heterogeneity and interaction, not identified in the original analysis. With the 8q24 data, the method demonstrates the genealogical independence of two nearby association signals.

I also use inferred ARGs to impute missing data. The performance of the method is compared to a standard method by using genotype data with held-out values, and is shown to be competitive. I evaluate the utility of an approach where case-control studies are merged with more densely typed control sets, such as the HapMap, and the additional loci imputed, allowing them to be tested directly for association.

The thesis concludes with a discussion of further population genetic questions which may be addressable by use of inferred ARGs.

Acknowledgements

Firstly, I would like to thank my PhD supervisor, Richard Durbin, and my PhD adviser Simon Tavaré. I would also like to thank Ralph McGinnis and the members of the Durbin Research Group for many helpful discussions. I also thank: Lachlan Coin for providing code to calculate Extreme Value Distributions. David Balding and Clive Hoggart for providing the FREGENE forward simulator and simulated populations. John Todd and Neil Walker for providing the CTLA4 data, and David Clayton, Chris Lowe and Joanna Howson for helpful discussions on my analysis of that data. Gilles Thomas for providing the 8q24 data and jointly performing the ARG analysis of that data. Inês Barroso and Eleanor Wheeler for providing the Chromosome 20 Ashkenazi data. David Carter for extending the MARGARITA system so that it can be used for nucleotide imputation with resequencing data, and for providing the *S. cerevisiae* Chromosome 1 data. The Wellcome Trust Sanger Institute for my PhD studentship and Gonville and Caius College for funding conference travel.

Mark J. Minichiello,
Wellcome Trust Sanger Institute,
04 July 2007.

Contents