

A Heuristic Algorithm for Inferring Possible Ancestral Recombination Graphs from Haplotype Data

Mark Minichiello and Richard Durbin ¹

Keywords: Ancestral recombination graph, haplotypes, SNPs, recombination rates, disease gene mapping.

1 Introduction

The Ancestral Recombination Graph (ARG) describes the history of mutations and recombinations giving rise to a sample of chromosomes. If the ARG for a sample were known, many population genetic analyses would become trivial. For example, recombination rates could be estimated by counting the number of recombination events in an interval and dividing by the total time over which they occurred. But the ARG is unknown, and instead ARGs compatible with the data have to be inferred. We present a heuristic algorithm for inferring ARGs from haplotype data. It can be applied to the order of 1000 sequences and 100 markers with a few hours' computing power. The algorithm is stochastic, so an ensemble of ARGs compatible with the data can be generated. We explore how such an ensemble of inferred ARGs can be used to estimate recombination rate variation and to map disease genes. Unlike linkage disequilibrium mapping, our technique is based purely on the inferred locations of recombinations. It does not depend on disease mutations co-occurring with marker mutations, and in principle can handle allelic heterogeneity.

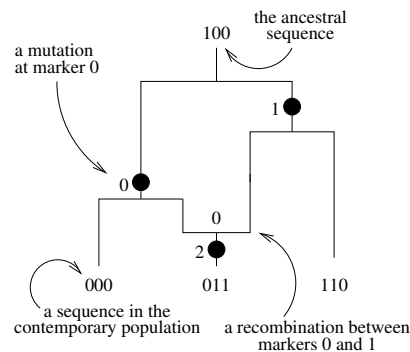


Figure 1: A plausible ARG for three sequences. Moving back in time (up the ARG), a black dot represents a mutation at the given position. The joining of two lineages represents a coalescence to the common ancestor for some individuals in some chromosomal region. Two lineages splitting is a recombination event, and the markers to the left (resp. right) of the recombination breakpoint (occurring immediately to the right of the marker denoted by the integer written above the split) are inherited from the left (resp. right) parent.

¹Wellcome Trust Sanger Institute, Cambridge, UK. E-mail: {mjm,rd}@sanger.ac.uk

2 Inferring Ancestral Recombination Graphs

A plausible ARG for the three haplotype sequences: $S_1 : 000$, $S_2 : 011$, $S_3 : 110$, is given in Figure 1; it shows how these sequences are ultimately descended from a common ancestor and related by recombination and mutation. To infer plausible ARGs, we consider tracts of sequence identity called shared segments; for the three sequences these are: $\{S_1, S_2\}[0, 0]$, $\{S_1, S_3\}[2, 2]$, $\{S_2, S_3\}[1, 1]$. The shared segment $\{A, B\}[x, y]$ means that sequences A and B share the same configuration of SNPs between markers x and y inclusively. A maximal shared segment terminates because of an ancestral recombination or mutation event; we use this insight to build ARGs, progressively combining shared segments and adding mutations in a stochastic fashion. The order in which shared segments are combined affects the resulting ARG, and heuristics can be designed for inferring, say, ARGs with fewer recombinations (ARGs for the three sequences can be built that do not have any recombinations).

3 Fine Mapping of Disease Genes

In case-control association studies we have two sets of individuals: those affected (cases) and those unaffected (controls) by some disease. The goal is to find disease causing SNPs. The genealogical trees of linked markers are correlated, and these trees (called marginal trees) are embedded in the ARG. We can locate a potential disease locus by finding the marker whose marginal tree best fits the assignments of cases and controls, so correlating with the unknown tree for the causative SNP. Our measure of correlation is called the ARG-Cut score, which counts the number of times a marginal tree has to be partitioned in order to arrive at disjoint clusters of case and control individuals (Figure 2).

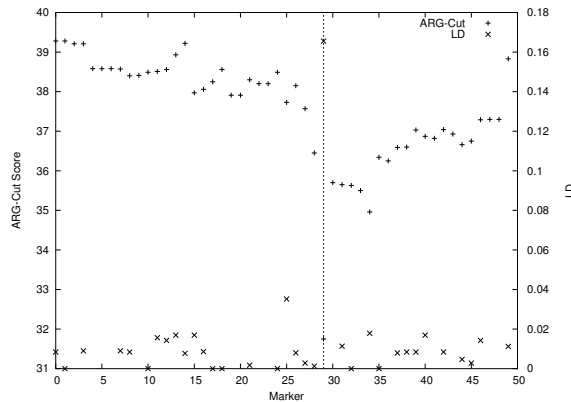


Figure 2: A simulated 1Mb region of 80 case and 80 control chromosomes. The vertical line denotes the location of the causative SNP, and the disease penetrances are: $P(\text{disease}|\text{homozygous wildtype}) = 0.01$; $P(\text{disease}|\text{heterozygous mutant}) = 0.085$; $P(\text{disease}|\text{homozygous mutant}) = 0.16$. The ARG-Cut score (the mean over 100 inferred ARGs is shown) is compared with the standard r^2 measure of linkage disequilibrium (LD). A lower ARG-Cut score implies a stronger disease association, whereas a higher r^2 implies stronger association.