

Ancestral Reconstruction for Association Mapping

Jotun Hein & Rune Lyngsø

July 5, 2010

Background

One of the greatest promises of modern genetics is the ability to identify loci responsible for, or at least highly correlated with, important physical traits and disease susceptibility. This will enable benefits ranging from improved disease risk assessment and crop design to ultimately a better understanding of how genotype variation effects variations in phenotypes. A key component in this task is association mapping where correlations between the clustering in the trait of interest and the clustering observed in the genealogy of a set of sampled sequences has the potential to localise causative genomic locations with relatively high precision.

Association mapping is commonly based on single nucleotide polymorphism (SNP) data, where individuals have been typed on a number of marker loci where two different nucleotides are present with reasonably high frequency in a population. However, more complex data including multi-allelic loci and insertion-deletion variation, in particular microsatellite repeat count variation can also be utilised. The human HapMap project International HapMap Consortium [2005] identified and continues to refine a dense set of SNP marker locations, and similar data is becoming available for other organisms. The availability of high density chips harnessing this information Barrett and Cardon [2006] allows genome-wide scans to be performed in large sets of individuals, and with the constant decrease in cost and increase in fidelity of sequencing and chip technologies it is to be expected that high density association mapping on large data sets becomes standard for the study of within-population phenotypic variation.

The two fundamental features of SNP data, and similar types of population variation data, that makes association mapping feasible is a relatively low mutation rate and the presence of recombinations in the evolutionary history of the data across the region covered. The low mutation rate causes both variation in the trait of interest and the type of a marker locus to be clustered into one or at most a few clades in the local genealogy of a sample. Recombination causes the genealogy to

change across the data, but with a high degree of correlation between positions separated by only a limited number of recombinations. For simple traits this allows association mapping based on simple marker-by-marker approaches, just considering the correlation between the observed marker variation and the trait variation across samples. However, this is not feasible for more complex traits, and even when the genealogy is unchanged between a marker locus and a trait affecting locus we cannot use the observed variation in the marker locus as a proxy for the variation in the trait affecting locus as the relevant mutations in most cases will have occurred in different parts of the genealogy. As observed in Zöllner and Pritchard [2005], when the trait affecting variants are not present in the data “...the best information that we could possibly get about association is to know the full coalescent genealogy of our sample at that position”. The availability of good methods for inferring the genealogical history of large data sets is thus crucial for harvesting the maximum amount of information from present day population variation data.

Traditionally the Kingman coalescent Kingman [1982] combined with an infinite sites model of mutations has proved a robust and useful model for SNP type data. Recursions describing data probability under this model in the presence of recombination are known Griffiths and Marjoram [1996], but even when applying a discretising approximation it is only infeasible to solve these recursions for relatively small data sets even when heuristics are applied Zöllner and Pritchard [2005]. To address data sets of realistic sizes, current methods either only consider the combinatorial structure of the problem Lyngsø et al. [2005], Minichiello and Durbin [2006], Wu [2008] or restrict genealogy reconstruction to blocks without recombination Mailund et al. [2006], Ding et al. [2008]. One interesting problem to investigate is how well these methods reconstruct local phylogenies and in particular how performance depends on parameters such as data set size and recombination rates.

When working with species that can be subjected to laboratory cultivation where we can create inbred stocks and use programmed interbreeding of these founders to create a heterogeneous stock of individuals with genomes that constitutes fine grained mosaics of the (known) genomes of the inbred stocks Mott et al. [2000], Durrant and Mott [2010]. In this particular context, we would expect the genome variation causing phenotypic variation to be present in the founder sequences. Hence, we don't need to reconstruct the full genealogical structure back to the common ancestor of the inbred founder lineages – just knowing the founder lineages each individual is derived from at a particular locus should provide sufficient information about their clustering.

As the genomes of the founders are known, at least to a reasonable degree, identifying the founder ancestry of an individual can be done by matching local sequence to these genomes. Rather than making a discrete call, it is preferable to

work within a probabilistic framework and compute probabilities of a particular loci being inherited from any particular pair of founders. This can be done using an HMM framework Mott et al. [2000], where states correspond to a particular pair. This works well when all individuals are typed at most markers, but with modern sequencing techniques it is usually more cost-effective to obtain sequence data from random marker positions. It will therefore be useful to develop methods that can perform inference of founder ancestry without considering positions not sequenced for the individual in question. Moreover, the types of uncertainties involved in both founder genome and individual sequence data should be investigated to fully incorporate this in the model developed.

References

- J. C. Barrett and L. R. Cardon. Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38(6):659–662, 2006.
- Zhihong Ding, Thomas Mailund, and Yun S. Song. Efficient whole-genome association mapping using local phylogenies for unphased genotype data. *Bioinformatics*, 24(19):2215–2221, 2008.
- Caroline Durrant and Richard Mott. Bayesian QTL mapping using inferred haplotypes. *Genetics*, 184:839–852, 2010.
- Robert C. Griffiths and Paul Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3, 1996.
- International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.
- J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.
- Rune B. Lyngsø, Yun S. Song, and Jotun J. Hein. Minimum recombination histories by branch and bound. In Rita Casadio and Gene Myers, editors, *Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI)*, number 3692 in Lecture Notes in Bioinformatics, pages 239–250. Springer, 2005.
- Thomas Mailund, Søren Besenbacher, and Mikkel H. Schierup. Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, 7:454, 2006.

Mark J. Minichiello and Richard Durbin. Mapping trait loci by use of inferred ancestral recombination graphs. *American Journal of Human Genetics*, 79(5): 910–922, 2006.

Richard Mott, Christopher J. Talbot, Maria G. Turri, Allan C. Collins, and Jonathan Flint. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(23):12649–12654, 2000.

Yufeng Wu. Association mapping of complex diseases with ancestral recombination graphs: Models and efficient algorithms. *Journal of Computational Biology*, 15(7):667–684, 2008.

Sebastian Zöllner and Jonathan K. Pritchard. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169:1071–1092, 2005.