

Genome-wide association analysis

Supervisors: G.Hellenthal, R.Lyngsoe, J.Hein

October 1, 2007

Genome-wide association (GWA) studies have become overwhelmingly popular in the last few years as a means to elucidate associations between particular alleles in one's DNA and a predisposition to disease, using genetic data from unrelated individuals randomly sampled from a population (Balding, 2006; WTCCC, 2007). The recent availability of large amounts of such *population genetic data* necessitates the need to efficiently and accurately test for disease-susceptibility loci in a computationally viable manner.

Single-marker-analysis (SMA), i.e. testing each DNA marker for disease association without including any information from additional DNA markers, has thus far been the primary tool for many GWA studies. However, ignoring the information from neighboring markers may significantly decrease the power to find associated markers. In particular, it is well known in population genetics that allelic types from nearby markers can be correlated, a phenomenon known as *linkage disequilibrium (LD)*. Scientists have tried to exploit ideas of LD to increase the power for finding markers associated with disease status. The power gained by considering multiple markers in such a manner versus using simple SMA has been demonstrated in much recent literature (Zöllner and Pritchard, 2005; Minichiello and Durbin, 2006; Mailund et al., 2006; Marchini et al., 2007).

The thus far most powerful of these methods (Zöllner and Pritchard, 2005) has used the notion of the *coalescent* to capture LD. The coalescent is a well-developed model that captures how the DNA of "unrelated" individuals may nonetheless be correlated due to the individuals' shared ancestral history far enough back in time. Such methods attempt to reconstruct the *ancestral recombination graph (ARG)* of a sample, which represents the complete evolutionary history of the DNA of all the individuals in the sample. If the ARG can be accurately reconstructed for any particular region of the genome, one can check if the ARG has a tendency to cluster disease individuals and non-disease individuals separately, which would be indicative of a potential disease-associated locus (or loci) in the region.

While the method of Zöllner and Pritchard (2005) has proven highly accurate towards achieving this goal, its techniques for reconstructing the ARG are computationally expensive, so that the method cannot be applied to the large amounts of available genetic data we have today. Therefore, several new methods have been recently developed that less rigorously reconstruct the ARG (Minichiello and Durbin, 2006; Mailund et al., 2006). Such methods attempt

to capture as much of the ancestral information as possible while remaining computationally tractable for large datasets.

The primary aim of this project consists of incorporating a means of detecting genomewide associations with disease status into a newly developed algorithm that efficiently reconstructs ARGs for population genetic data (R.Lyngsoe, unpublished). Furthermore, as all of the aforementioned computationally-efficient ARG-based methods have been developed only recently, the second component of this project involves evaluating how these methods compare to one another and to SMA. This involves conducting an extensive simulation study to compare the newly developed ARG-based method to those of Minichiello and Durbin (2006) and Mailund et al. (2006), as well as to SMA, concentrating on the differences in computation speed and the power to find true disease associations.

Therefore, the main aims of the project entail:

- understanding current GWA research interests and typical datasets
- reviewing current methods in GWA studies, with particular emphasis on the new ARG-based methods of Minichiello and Durbin (2006) and Mailund et al. (2006)
- incorporating GWA framework(s) into a new, unreleased ARG-reconstruction program (R.Lyngsoe, unpublished)
- conducting an extensive simulation study to compare the ARG-based methods of Minichiello and Durbin (2006), Mailund et al. (2006), and the new method (R.Lyngsoe, unpublished), using a new, unreleased simulation program (G.Hellenthal, unpublished)
- assessing performance of the ARG-based methods under a wide variety of simulated conditions:
 - using hundreds vs thousands of individuals, mimicking current genomewide analyses
 - using hundreds vs thousands of markers, varying the density of markers in a region
 - using “resequencing” vs Illumina-based “tagSNP” strategies for marker selection
 - using different modes and manners of disease inheritance (e.g. dominant/additive/recessive modes of inheritance, multiple causative sites, 2-site interactions)
 - using various disease severities (e.g. *relative risks* of 1.2,1.4,1.6,1.8,2.0)
- writing up the main conclusions and findings

Overall this proposal presents an unique opportunity for the researcher to become familiar with the most recent GWA concepts and methodologies, as well as ideas of the extremely prominent coalescent field. Time permitting,

additional exploration might include further extensions to the new ARG-based method, extensions to the new case/control simulation software, and perhaps the development of additional GWA techniques.

The expected duration of this project is estimated as *two months*.

References

- Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7, 781–791.
- Mailund, T., S. Besenbacher, and M. Schierup (2006). Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics* 7, 454–476.
- Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly (2007). A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics* 39, 906–913.
- Minichiello, M. and R. Durbin (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet* 79, 910–922.
- WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678.
- Zöllner, S. and J. Pritchard (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* 169, 1071–1092.