Rare and low-frequency variants and predisposition to complex disease

Patrick K. Albers
Wellcome Trust Centre for Human Genetics
Green Templeton College
University of Oxford

Submitted in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy (DPhil)

Hilary 2017

Recent advances in high-throughput genomic technologies have enabled the collection of DNA information for thousands of individuals, providing unprecedented opportunities to learn about the genetic architecture of complex disease. One important finding has been that the majority of variants in the human genome are low in frequency or rare. It has been hypothesised that the recent explosive growth of the human population afforded unexpectedly large amounts of rare variants with small but deleterious effects, suggesting that rare variants may play a significant role in the predisposition to complex disease. Moreover, properties specific to rare variants embody a rich source of information relating to their evolutionary history.

In this thesis, I develop several statistical methods to address problems associated with the analysis of rare variants in the context of large cohorts linked to biomedical phenotype data, and to leverage the information they encode. Firstly, one constraint in genome-wide association studies is that lower-frequency variants are not captured by genotyping methods, but instead must be predicted through imputation from a reference panel. I develop a method to improve imputation accuracy by integrating genotype data from multiple reference datasets, which outperformed imputations from separate references in almost all comparisons (mean correlation with masked genotypes $r^2 > 0.9$). In a series of simulated case-control experiments, I demonstrate that this approach (meta-imputation) increases power to identify low-frequency variants of intermediate or high penetrance, improving power by 2.2–3.6%. Secondly, I utilise rare variants as identifiers for recently co-inherited shared haplotypes, as rare variants are likely to have originated recently through mutation, making them highly population-specific. I develop a non-probabilistic method to detect shared haplotype segments that are identical by descent (IBD) from patterns of allele sharing and the detection of recombination breakpoints. I show that the latter can be inferred with higher accuracy at very low allele frequencies ($\leq 0.05\%$, $r^2 > 0.99$) using either haplotype or genotype data. Thirdly, I show that genotype error poses a major problem in the analysis of empirical data, for example as obtained through whole genome sequencing or SNP genotyping, in particular towards lower allele frequencies (false positive rate, FPR = 0.1, at frequency $\leq 0.05\%$). I therefore subsequently propose a novel approach to infer IBD from genotype data using a Hidden Markov Model (HMM) under an empirical error model, which I construct by identifying misclassified genotypes in existing datasets, showing that the HMM is robust in presence of error ($\leq 0.05\%$, $r^2 > 0.98$) while previous methods fail ($r^2 < 0.02$). Finally, the age of a rare allele (time since its creation through mutation) may provide evidence about the selective forces that resulted in its observed frequency, and its impact on fitness. I further develop a novel method to estimate rare allele age, based on the inferred IBD structure of a sample. I demonstrate that allele age can be estimated with high accuracy using the HMM-based approach for IBD detection, even in presence of genotype error (Spearman correlation coefficient $r_S = 0.74$, compared to $r_S = 0.82$ when true IBD data is available). I apply this method to data from the 1000 Genomes Project, showing that there are notable age differences between rare alleles of varying predicted phenotypic consequences.