Rare and low-frequency variants and predisposition to complex disease

Patrick K. Albers

Wellcome Trust Centre for Human Genetics
Big Data Institute
Medical Sciences Division
Green Templeton College
University of Oxford

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

Hilary 2017

Advances in high-throughput genomic technologies have facilitated the collection of DNA information for thousands of individuals, providing unprecedented opportunities to explore 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 recent explosive growth of the human population afforded unexpectedly large amounts of rare variants with potentially deleterious effects, suggesting that rare variants may play a role in disease predisposition. But, importantly, rare variants embody a source of information through which we may learn more about our recent evolutionary history. In this thesis, I developed several statistical and computational methods to address problems associated with the analysis of rare variants and, foremost, to leverage the genealogical information they encode.

First, one constraint in genome-wide association studies is that lower-frequency variants are not well captured by genotyping methods, but instead are predicted through imputation from a reference dataset. I developed the *meta-imputation* method to improve imputation accuracy by integrating genotype data from multiple, independent reference panels, which outperformed imputations from separate references in almost all comparisons (mean correlation with masked genotypes $r^2 > 0.9$). I further demonstrated in simulated case-control studies that meta-imputation increased the statistical power to identify low-frequency variants of intermediate or high penetrance by 2.2–3.6%.

Second, rare variants are likely to have originated recently through mutation and thereby sit on relatively long haplotype regions identical by descent (IBD). I developed a method that exploits rare variants as identifiers for shared haplotype segments around which the breakpoints of recombination are detected using non-probabilistic approaches. In coalescent simulations, I show that such breakpoints can be inferred with high accuracy ($r^2 > 0.99$) around rare variants at frequencies $\leq 0.05\%$, using either haplotype or genotype data.

Third, I show that technical error poses a major problem for the analysis of whole-genome sequencing or genotyping data, particularly for alleles below 0.05% frequency (false positive rate, FPR = 0.1). I therefore propose a novel approach to infer IBD segments using a Hidden Markov Model (HMM) which operates on genotype data alone. I incorporated an empirical error model constructed from error rates I estimated in publicly available sequencing and genotyping datasets. The HMM was robust in presence of error in simulated data ($r^2 > 0.98$) while non-probabilistic methods failed ($r^2 < 0.02$).

Lastly, the age of an allele (the time since its creation through mutation) may provide clues about demographic processes that resulted in its observed frequency. I present a novel method to estimate (rare) allele age based on the inferred shared haplotype structure of the sample. The method operates in a Bayesian framework to infer pairwise coalescent times from which the age is estimated using a composite posterior approach. I show in simulated data that coalescent time can be inferred with high accuracy (rank correlation > 0.91) which resulted in a likewise high accuracy for estimated age (> 0.94). When applied to data from the 1000 Genomes Project, I show that estimated age distributions were overall conform with frequency-dependent expectations under neutrality, but where patterns of low frequency and old age may hint at signatures of selection at certain sites. Thus, this method may prove useful in the analysis of large cohorts when linked to biomedical phenotype data.