Objective Bayesian Hidden Markov models for detecting regions of copy number variation in the human genome using SNP genotyping data

Chris Holmes,
Oxford University, UK

Recent discoveries suggest that regions of copy number variation (CNVs) in the human genome are much more widespread than previously thought. A CNV is defined as a segment of DNA > 1 kb that is present at a variable copy number in comparison to a reference genome. It is believed that up to 10% of the human genome maybe copy number variable (contributing to around 10% of genetical transcription variation) and copy number polymorphisms have been linked to a number of diseases. In recent work we have developed an objective Bayesian Hidden Markov model to detect regions of copy number variation from Illumina BeadChip (SNP) data. In our model the hidden states refer to unobserved copy number variants at a locus (SNP) and the transitions between states capture the persistence within CNV states across chromosomal regions. Parameters in the hyper-priors of the model are set "objectively" using prior training data of known CNV and the Bayes factor thresholds (for calling regions of CNV) are calibrated, via a simulation stage, to user set false positive rates. Predictions from the model have been experimentally validated on a number of samples; and the method is to be employed in a number of genome wide association studies. This is joint work with the Wellcome Trust Centre for Human Genetics.