

International Society for Bayesian Analysis, 9th World Meeting,
Hamilton Island, Australia, 2008.

INFERRING REGIONS OF COPY NUMBER VARIATION (CNVs) IN
HUMAN DNA FROM SNP GENOTYPING DATA USING OBJECTIVE
BAYESIAN SIGNAL PROCESSING METHODS

Chris Holmes^{1*} and Chris Yau¹

¹ University of Oxford, Oxford, UK

* c.holmes@stats.ox.ac.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 genome-wide single nucleotide polymorphism (SNP) genotyping data (of around 500,000 SNPs). 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. In certain samples, such as from tumour biopsies, tissue heterogeneity introduces additional complications requiring a mixture deconvolution. Predictions from the model have been experimentally validated on a number of samples. We report the findings from a number of large studies including 1500 samples from the 1958 UK birth cohort and a genome-wide association study of childhood malaria risk in an African population. This is joint work with the Wellcome Trust Centre for Human Genetics.