

IDENTIFYING COPY NUMBER ALTERATIONS AND LOSS-OF-HETEROZYGOSITY IN TUMOUR SAMPLES FROM SNP GENOTYPING DATA IN THE PRESENCE OF STROMAL CONTAMINATION AND INTRA-TUMOUR HETEROGENIETY

Christopher Yau and Christopher C. Holmes

Department of Statistics, University of Oxford

Disruptions to the normal mechanisms of cell division and DNA replication in cancer can lead to the gain or loss of genetic material. As a consequence, an improved understanding of the genetic alterations that characterise different forms of cancer forms an integral part of the effort to tackle these diseases. The genome-wide identification of copy number alterations and loss-of-heterozygosity (LOH) events in tumour DNA samples is therefore an important part of modern cancer research. Copy number and LOH analysis can provide insight into the function of genes, in either a tumour suppressing or oncogenic capacity, and may also have important clinical applications as prognostic biomarkers.

Traditionally, tumour studies have been performed using the tools of molecular cytogenetics, such as Spectral Karyotyping (SKY) and Fluorescent in-situ Hybridisation (FISH). Microarray based Comparative Genomic Hybridisation (aCGH) have also been used extensively to obtain genome-wide genetic profiles (10,000-32,000 probes) and greatly increase resolution to identify chromosomal alterations in the <1MB range. More recently, the application of SNP genotyping microarrays has been of interest in cancer research. Current generations of SNP genotyping microarrays provide extremely high-density genomic coverage (up to 2 million probes) and, more importantly, provide allele-specific information that has been critical in the identification of LOH events, such as uniparental disomies (UPD) that are undetected with aCGH.

A variety of computational tools are currently available to cancer researchers that can detect copy number alterations and LOH events from SNP genotyping data. However, these tools are typically limited to studies where extensive purification of tumour DNA samples has taken place, using techniques such as laser-capture microdissection. These purification processes are required in order to remove normal cells in the sample (stromal contamination) and obtain a homogeneous tumour cell population. Often, it is not possible to completely remove all contaminating normal cells from the tumour sample and, moreover, the tumour may itself be heterogeneous and contain a mixture of cells harbouring different genetic alterations. This sample heterogeneity can lead to cryptic structures in the SNP genotyping data that conventional computational tools cannot identify as they assume genetic homogeneity.

We have developed a Bayesian statistical inference framework that allows the deconvolution of SNP genotyping data obtained from heterogeneous tumour samples. The model utilises a mixture of Hidden Markov models to capture the hidden copy number states of each cell population in the heterogeneous sample. In combination with a generative model that explains the cryptic data structures observed under heterogeneous sample conditions, we are able to identify and characterise copy number alterations and loss-of-heterozygosity events even in the presence of intra-tumour heterogeneity and stromal contamination. As SNP genotyping datasets are often very large, Monte Carlo-based inference is often impractical; instead we have used suitable simplifying assumptions to devise a fast expectation maximisation algorithm for approximated model fitting. We demonstrate our approach on a variety of simulated, cell-line and clinical tumour samples and using artificial mixtures of tumour and normal DNA.