Affymetrix SNP arrays can interrogate thousand of SNPs at the time. This allows us to look at the genomic content of cancer cells and understand the underlying events leading to cancer. Loss-of-heterozygosity (LOH) and copy-number analyses have been done independently but combining both in one analysis is likely to give better results. Here we described one such approach in that we developed a Hidden Markov Model to estimate allelic copy-number changes in tumour cells and the proportion of normal cells in the tumour (mixture proportion).

In a sense the method works on paired normal-tumour samples. It takes as input the genotype call of a normal sample and the allelic SNP intensities of the tumour sample and outputs the estimated copy-number states for each SNP. To limit the state space of the underlying Markov Chain, we restrict the possible copy numbers of each allele to 0, 1, 2 and > 2. The different hidden states estimated by the HMM correspond to different events occurring in the cancer cell (Figure 1).

Many tumour samples contain a large fraction of normal cells and this potentially affects the performance of the method. We therefore included the possibility to estimate the population mixture (proportion of cancer cells) from the data and use this in the analysis.

We show that our method is able to recover the underlying copy-number changes in simulated datasets with high accuracy (above 97.71%). Moreover, although the known hidden states could be well recovered in simulated cancer samples with more than 70% cancer cells (and less than 30% normal cells), we demonstrate that including the mixture proportion in the HMM increased the accuracy of the method. Finally, the method is tested on the Hapmap dataset and on bladder and prostate cancer samples.