



Computational Biology and Statistics

www.cambridgecancer.org.uk/simontavare

Group Leader **Simon Tavaré**

Associate Scientist

Natalie Thorne[†]

Postdoctoral Scientists

Nuno Barbosa-Morais

Christina Curtis

Andrew Lynch

Matthew Ritchie[‡]

Shamith Samarajiwa*

Irene Tieman-Boege[†]

Graduate Students

Jeremy Darot

Mark Dunning[†]

Leonard Goldstein

Thomas Hardcastle^{†‡}

Sergii Ivakhno

Andrea Sottoriva*

Doug Speed

Christiana Spyrou

Julie Woolford

Visiting Workers

Daniel Goodman*

Alexandra Jauhiainen^{*†}

Ian Saunders^{*†}

Summer Placement Student

Jonathan Cairns^{*†}

Our work has focused on five main areas: statistical methods for microarray data; analysis of Solexa resequencing data; somatic cell mitotic clocks; statistical genetics and analysis of methylation data.

A major focus of our research continues to be the development of statistical methods for the analysis of Illumina bead array technologies. Illumina's technology uses randomly assembled arrays of beads, each of which has a probe for a specific genomic feature. The random nature of the arrays leads to a number of novel statistical problems that we have addressed. Illumina's software can now report full bead-level data, thus opening up their data to analysis by other techniques. Our open-source Bioconductor software, *beadarray*, which has been downloaded more than 6,000 times this year, provides a statistical environment for such analyses.

Access to the raw data allows for more detailed quality assessment and flexible statistical analyses. However, quality assessment on summarised data may miss spatial artefacts present in the raw data. With their many replicates and their random layouts, Illumina BeadArrays provide greater scope for detecting spatial artefacts than do other microarray technologies. We addressed the lack of tools that could perform these tasks for Illumina by developing BASH as a tool for this purpose (Figure 1). Using bead-level data, spatial artefacts of various kinds can thus be identified and excluded from further analyses.

We have also developed statistical tools for Illumina's two-channel GoldenGate technology. It retains the desirable

properties of Illumina's BeadArrays in that the probes (in this case 'beads') are randomly arranged across the microarray, there are multiple instances of each probe and many samples can be processed simultaneously. As for other Illumina technologies, however, these properties are not exploited as they might be. We have studied various common adaptations of the GoldenGate platform, reviewed the associated analysis methods and suggested some improvements that can be made over the default analysis. These methods are also available in the *beadarray* software.

We have continued our work on inferring human stem cell dynamics from variations in their epigenomes. The billions of cells within an individual can in principle be organised by genealogy into a single somatic cell tree that starts from the zygote and ends with present-day cells. In theory, this tree can be reconstructed from replication errors that record divisions and ancestry. Such a 'molecular clock' approach is currently impractical because somatic mutations are rare, but it is feasible to substitute instead the 5' to 3' order of epigenetic modifications such as CpG methylation. Potentially the genealogy of any human cell may be reconstructed without prior experimental manipulation by reading histories recorded in their genomes. The three main aspects to this work are the identification of a small number of suitable CpG loci, the generation of data on CpG island methylation at these loci in single cells, and inference of genealogical history using computational inference methods. We have begun a study of colorectal tumours using this approach.

Measuring methylation status in single molecules is a focus of our wet-lab work, where we have developed an emulsion PCR-based method to do this as well as tools to analyse such data. Unlike standard bisulfite sequencing methods, which resequence the entire region of interest, this

*Joined during 2008 †Left during 2008

‡with James Brenton

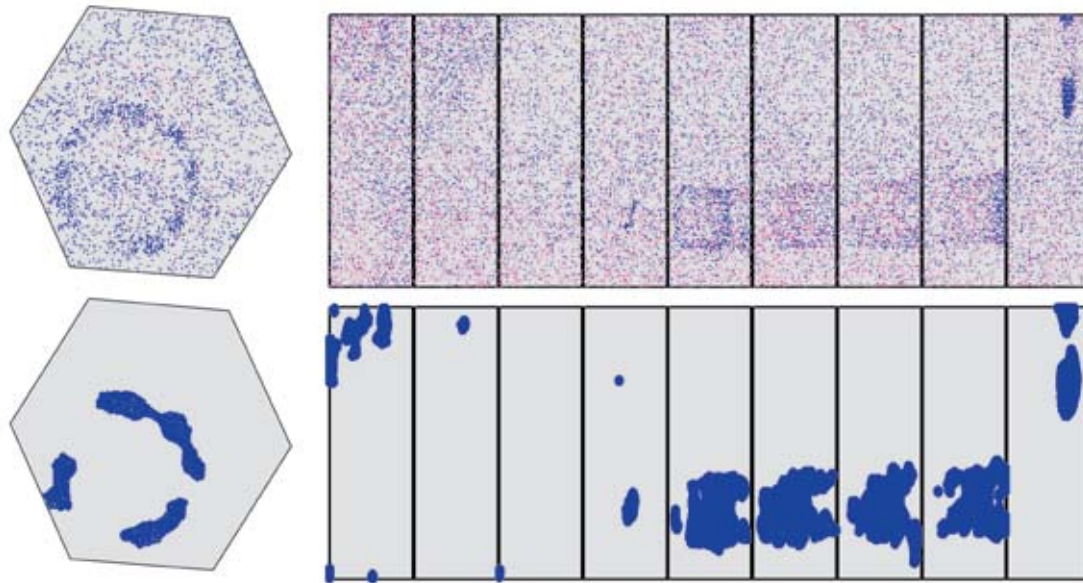


Figure 1. A SAM array (left) and a BeadChip (right), showing the locations of the outliers on the array surfaces (top) and the areas excluded from the analyses by BASH (bottom).

technique identifies just the methylation status of CpGs, and can measure this cheaply in many thousands of molecules simultaneously. This approach can also be exploited to study differential allelic expression, an ongoing collaboration with the Ponder laboratory.

We have collaborated on several projects that involve the statistical analysis of resequencing data, including development of peak-calling algorithms for ChIP-seq experiments, transcript resequencing and image analysis for the Solexa instruments. In a collaboration with the Miska lab at the Wellcome Trust/Cancer Research UK Gurdon Institute, we identified a class of 21-nucleotide RNAs as the piRNAs of *C. elegans*, and are identifying other small RNAs, such as microRNAs, by Illumina sequencing.

In a collaboration with the Odom laboratory, we used hepatocytes from an aneuploid mouse strain carrying human chromosome 21 to determine, on a chromosomal scale, whether interspecies differences in transcriptional regulation are primarily directed by human genetic sequence or mouse nuclear environment. We used several different ChIP-chip experiments to study transcription factor binding site occupancy and trimethylation at transcription start sites, and deduced that, in homologous tissues, genetic sequence is largely responsible for directing transcriptional programs. Ongoing work will study expression in this system as well.

We have continued our work with the Caldas laboratory on the statistical analysis of a number of breast cancer datasets, in particular the METABRIC project that is studying copy number aberrations (CNA) and expression in some 2,000

breast tumours. We developed experimental design and analysis methods for pilot projects designed to choose the appropriate experimental array platform, and we are now working with the Bioinformatics Core on analysis pipelines for the production runs. Similar analysis issues have also arisen in our collaboration with the Tuveson laboratory on the analysis of CNAs in pancreatic cancer:

Shamith Samarajiwa joined us in November. His expertise is in data visualization and data mining, themes he is developing in a collaboration with the Narita laboratory. Andrea Sottoriva joined as a graduate student. He will be working on agent-based models for crypt and tumour evolution. Daniel Goodman joined us in October for the year on a Whittaker pre-doctoral fellowship. He is working on the analysis of methylation data. Ian Saunders visited us for three months from CSIRO in Adelaide. Ian's research focused on identification of interaction among SNPs in association studies. Alexandra Jauhiainen visited us from the Statistics Department at Chalmers University in Gothenburg. Alexandra worked with the Brindle laboratory on the analysis of metabolomic data.

We have had a number of departures from the group. Natalie Thorne and Matt Ritchie have returned to the bioinformatics division at the Walter and Eliza Hall Institute in Melbourne and Irene Tiemann-Boege has taken up a position in the Institute of Biophysics at the Johannes Kepler University in Austria. Mark Dunning and Tom Hardcastle completed their PhDs; Mark is now in the Bioinformatics Core, and Tom is a postdoc in the Baulcombe lab in Plant Sciences, Cambridge.

Publications listed on page 61