

# Bioinformatics

[www.cambridgecancer.org.uk/research/coreresources/bioinformatics](http://www.cambridgecancer.org.uk/research/coreresources/bioinformatics)

Head **Matthew Eldridge**

## Senior Computational

### Biology Analysts

Ben Davis\*

Kevin Howe

Roslin Russell\*

Rory Stark

Sarah Vowler

## Computational Biology

### Analysts

Mark Dunning\*

Stewart MacArthur\*



The Bioinformatics core provides expertise in statistics, data analysis and software development to support CRI research groups.

We assist with the computational aspects of research by providing support and consultation in relation to data management and statistical data analysis, focussing primarily on high-throughput genomics technologies including microarrays and next generation sequencing.

High-density DNA arrays composed of many thousands of oligonucleotide probes are used to measure changes in gene expression levels, to investigate transcription factor binding using ChIP-chip assays, and to detect single nucleotide polymorphisms (SNPs) and copy number variation. We are developing analysis workflows for these experiments using the statistical programming language R and various packages from the open source Bioconductor project, supporting Illumina, Affymetrix and Agilent microarray platforms. A microarray analysis project consists of a series of stages from experimental design to quality assessment, data pre-processing, normalization, and statistical inference.

A typical analysis will identify a large number of differentially expressed genes. We have applied a number of downstream methods to help interpret such gene lists, including gene set enrichment and over-representation of biological functions using Gene Ontology terms. We also use pattern discovery tools to look for shared sequence motifs in upstream regions of co-regulated genes, and Cytoscape for visualization of gene expression profiles in the context of biochemical pathways.

We are also responsible for the deployment of a processing pipeline for the sequencing-by-synthesis platform (Solexa) from Illumina. These high-throughput instruments generate tens of millions of short sequence reads every 2–3 days and are used in a range of applications including: ChIP-seq using chromatin immunoprecipitation to identify binding sites of

DNA-associated proteins such as transcription factors and histone marks; small RNA discovery and profiling; RNA-seq transcriptome analysis to detect novel and rare transcripts, alternative splice sites and allele-specific expression; and re-sequencing to study genetic variation, such as SNPs, copy number changes and chromosomal rearrangements common in many cancers.

We have implemented an automated pipeline for transferring the terabytes of raw image data generated to a high performance computing cluster for image analysis, base calling and alignment of sequence reads to a reference genome, followed by transfer to long-term storage. The analysis is configured using information contained in a LIMS system we have developed to support sequencing request submissions, sample tracking and laboratory workflow within the Genomics core.

Increasingly, CRI researchers are relating gene-level experimental data to clinical information for cancer patients. We have been performing survival analysis on these datasets to identify genes that are potentially useful as prognostic indicators.

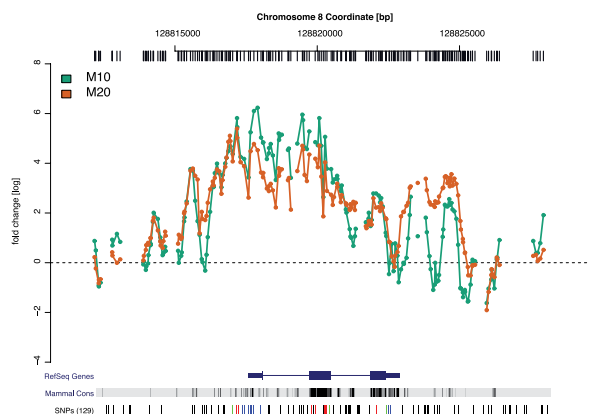


Figure 1. Use of a DNase hypersensitivity array to determine areas of open chromatin. The area around the *Myc* proto-oncogene is shown, indicating a large area of open DNA in the MCF-7 breast cancer cell line.

**Publications listed on page 63**

\*Joined during 2008