A ROBUST STATISTICAL METHOD FOR THE DETECTION OF ALTERNATIVELY SPLICED MRNAS FROM AFFYMETRIX EXON ARRAYS

George Nicholson[1], Jennifer Taylor[2] and Chris Holmes[1]

[1]Department of Statistics, University of Oxford
[2]Wellcome Trust Centre for Human Genetics, University of Oxford

It is estimated that as many as 75% of genes in the human genome engage in production of multiple alternative splice variants (*mRNA isoforms*) from the same base sequence. Knowledge of which isoforms are created on specific physiological, genetic and environmental backgrounds will expedite the discovery of novel associations - as well as the characterisation of known ones - between genetic polymorphism and disease, and is also an important step in the integration of transcriptomics and proteomics.

The development of microarray technology over the past decade has permitted the study of mRNA transcript abundance on a genome-wide scale. It is now feasible to assess simultaneously the level of expression at all known exons in the human genome, using the Affymetrix Exon 1.0 ST Array platform. Statistical methodology for extracting informative signal from such a wealth of data is required. Here we describe a statistical method for the robust discovery of novel splice variants; the method is designed to attain a low rate of false discoveries.

We first focus on careful annotation: we retain only those oligonucleotide probes that can be mapped to a unique genomic location, and constrain attention specifically to those probes mapped within exons and UTRs in the ENSEMBL database. We then adopt a non-parametric approach to (i) summarise exon expression robustly; (ii) provide quality weights for these exon summaries; (iii) perform weighted K-means-style clustering on exon summaries. Our approach clusters exons into groups that are empirically co-expressed on mRNA isoforms. The true number of clusters underlying the data is estimated greedily. Each cluster's 'mean' represents the combined expression of a unique, identifiable superposition of mRNA isoforms. We describe the application of our method to a publicly available data set measuring expression in HapMap cell lines.