**Discovery and validation of splicing at the proteome level in cancer**

Advisors: Alvis Brazma (EBI) Jyoti Choudhary (Sanger) and Ultan McDermott (Sanger)

Alternative splicing is believed to be an important contributor to protein diversity. In the human genome there are fewer than 22,000 protein coding genes, however, these are transcribed into over 140,000 different transcripts (Ensembl 66), the majority of which have protein coding potential. Nearly every multi-exon gene has been associated with more than one expressed alternative splice-form and most of these have been found to be expressed as RNA under some conditions. However, recent analysis of RNASeq data across human tissues and cell lines carried out by the Brazma group has revealed that despite this diversity most protein-coding genes express a single predominant transcript in a particular condition. For about 50% of genes the major transcript appears to be ubiquitously expressed. Only for a relatively small number of genes there is a clear *switch* between two strongly dominant isoforms in different tissues, while the overall expression of the gene remains roughly the same. We have also found such *switch* events between normal and cancer samples of the same patient by analysing RNAseq data from 50 kidney cancers. A critical question is whether this *switch* is reflected in the cell proteome. Or more generally – if there is a correlation between the relative transcript abundance (for a given gene) and the presence of the respective protein in the sample? Modest correlation between proteins and RNA has been demonstrated on gene level, however on transcript level this question is wide open.

The Choudhary group has demonstrated the presence of different isoforms of the Synaptic GTPase-Activating Protein (SynGAP), a key synaptic protein, in brain. This work also showed that protein function is determined by the combination of the protein amino-terminal sequence with its carboxy-terminal sequence. This was the first demonstration that activity-dependent alternative promoter usage and splicing modulates the function of a synaptic protein at excitatory synapses. Promising preliminary results showing a correlation between the gene isoform expression in RNA and protein levels have emerged from studies of human cell lines using siRNA knockouts of specific splicing factors (an on-going collaboration between Brazma/Marioni/Venkitraman/Aabersold groups). However, in these experiments, which were not designed to address this specific question, it was possible to look only at a small number of genes (manuscript in preparation).

Overall, these findings have a number of important implications. Firstly, proteome analysis requires a protein database for peptide spectrum matching, assessment of the use of actual transcripts over current protein and integration with genome datasets to provide an accurate and compact representation of the true coding potential of each gene. Using a transcript based search space in proteomics would circumvent current practice of peptide apportioning or protein grouping. A further impact of this would be to enhance quantitative analysis. Currently, when peptides are shared over various isoforms the associated quantitation has would either be neglected or distributed evenly over the variants. Secondly, alternative transcripts can have critical functional implications.

The ESPOD fellow will address key questions arising from a combined transcriptomic and proteome analysis. The focus will be on cancer cell lines that have distinct phenotypes such as drug resistance/sensitivity that have been characterised in the Cancer Genome Project at the Sanger Institute (McDermott group). RNASeq data on these samples will be generated and analysed using the bioinformatics approaches developed in the Brazma group. The emphasis here will be to look for the most pronounced switch events between pairs of cell lines in the RNASeq data. For those transcripts where the dominant switch is identified we will identify those peptides that distinguish between the dominant forms (either individually or in combination). Our preliminary studies show that finding individual peptides that could discriminate two isoforms of the same gene in the proteomics data are difficult, therefore we will develop Bayesian methods relying on combinations of peptides. Proteomics data will be collected on the same cell lines in the Choudhary group at the Sanger Institute. There will be the opportunity to develop systematic approaches to measure the protein isoforms on a proteomic scale, thus overcoming the low coverage of current shotgun proteomics datasets. The ESPOD Fellow would be able to implement new data collection approaches to improve the proteome coverage.

The Fellow will have the opportunity to pursue experimental and computational work. In particular, the emphasis will be to develop new methods and functional validation approaches. Thus, novel methods for data analysis devising effective ways to bring together the transcriptome and proteome analysis will be explored. This will help validate the findings of the transcriptome and to explore fundamentals of gene expression regulation by possibly linking to other genomics data.

The switching of isoforms is likely to be fundamental to function and biological effect. Selected candidates will be studied in detail using targeted follow-up such as epitope tagging viral recombination systems or immunoaffinity methods. These approaches enable detailed inspection of primary structure include post translational modification as well as means to study protein interactions. In addition, they can also be used for cellular localisation and other functional assays. These studies will provide a better understanding of the effect of isoform switching and mechanisms of regulation.

This project will result in

1) New bioinformatics approaches for joint analysis of RNASeq and proteomics data and their implementation as software modules;
2) The applications of these methods to the RNASeq and proteomics data generated on the selected cell lines;
3) The testing of the hypothesis that the predominant transcripts of the given gene is predominantly translated into the respective protein;
4) An assessment of the human protein diversity created by alternative transcription;
5) A study of functional implications of isoform switching in cancer and the mechanism of their regulation.