Abstract: Alternative pre-mRNA Splicing (AS) is a mechanism that increases the protein diversity in vertebrates. AS is considered to be an important source of functional complexity and disease; and about 40-60% of human genes have alternative splice forms. New exon junction microarrays provide a tissue distribution of certain splice forms for several thousand genes and systematic statistical analysis can help use to better understand the biological role of AS. I will discuss a semiparametric Bayesian model for identifying the number of gene products produced from important genetic loci and to obtain clues regarding the regulation of AS across different tissues, with the aim of aiding future exploratory analysis for isolation of splice forms and providing information on specific targets.

The work focused on the analysis of the Rosetta gene expression dataset from five Agilent chips covering over 10,000 multi-exon human genes in 52 tissues and cell lines. We work in a multiplicative measurement error framework and for a given gene, the specific binding effects are modeled by mixture of Dirichlet process (DP) priors that allow simultaneous estimation of clusters as well as the number of clusters. The non-specific binding effect is elicited by a conjugate prior that with prior knowledge about the number of significant non-specific bindings from sequence analysis. The use of conjugate priors and efficient Gibbs Sampling schemes for DP priors makes the computation straightforward.

The results indicate more alternatively spliced (70-75%) human multi-exon genes than reported by previous studies. In addition, there are a number of splice forms that appear to be unique to cancer. These new forms may be potential new targets for oncology drug discovery teams.