A platform for the network assembly and visual analysis of transcript isoforms from short-read RNA-sequencing data

Barbara Shih[1], Neil A. Mabbott[1], Tom C. Freeman[1,2]

The Roslin Institute and Royal (Dick) School of Veterinary Studies, the University of Edinburgh, Easter Bush, Midlothian, Edinburgh EH25 9RG, UK[1], Kajeka Ltd., Roslin Innovation Building, Easter Bush, Midlothian, Edinburgh EH25 9RG, UK[2]

RNA-sequencing (RNA-seq) data describes both transcript abundance and exon usage. However, splicing events can be complex, and therefore difficult to visualise and interpret.

Here, we describe a new data pipeline/visualisation platform developed to support the representation of RNA-seq data as networks (RNA-assembly graphs), thereby providing a visual representation of data structure that facilitates interpretation of exon/intron use. The pipeline requires the following files: input BAM files, the corresponding GTF file, and genes of interest. Two approaches are available within the pipeline for the graph generation. The first is based on comparing sequence similarity, e.g. by Blast, whereby the similarity scores are used to define edges between reads (nodes). However, generation of graphs using this approach can be computationally expensive (due to the blast step) and graphs can be very large. In order to circumvent these issues, a second approach is based on mapping reads to specific loci, i.e. regions of the genome, which are represented as nodes, and edges are defined by the number of reads spanning across regions. To visualise both RNA-assembly graph types we have also been developing a new network analysis platform, called Graphia (Kajeka Ltd). This platform not only supports the visualisation of massive graphs (millions of nodes of edges), but supports the overlay of other information, e.g. gene or exon IDs, dynamic filtering on edge weight or source as well as a range of other functionalities.

Using real and simulated short-read RNA-seq data we have demonstrated that different splicing events (alternative start/end, exon skipping and mutually exclusive exons) produce distinct network structures by both approaches. However, the implementation of the loci-based approach drastically simplifies the graphs, enabling direct comparison of exon usage across different samples. We believe this approach significantly improves our ability to interpret the often complex splicing captured by RNA-seq analysis.