Title: Curation, c*haracterization and quantification of a PacBio transcriptome*

Tardaguila Manuel[2] , de la Fuente Lorena[1], del Risco Hector[2], Martí Cristina[1], Pereira Cecile[2], Moreno Victoria[3], Rodríguez Susana[4], Conesa Ana[1,2]

1. Centro de Investigación Príncipe Felipe, Genomics of Gene Expression, Valencia, Spain
2. Institute for Food and Agricultural Sciences, Department of Microbiology and Cell Science, University of Florida, Gainesville, USA
3. Centro de Investigación Príncipe Felipe, Gene Expression and RNA Metabolism, Valencia, Spain
4. Centro de Investigación Príncipe Felipe, Neuronal and Tissue Regeneration, Valencia, Spain

Background:

Alternative splicing, a widespread means of creating functional diversity in higher eukaryotes, entails substantial challenges for its bioinformatic analysis. Paramount among these is the elaboration of the transcriptome to analyse, specially given the high similarity rate between isoforms and the incompleteness and/or variation in the annotation of the 5' and 3' ends of the mRNA. Here we have applied both PacBio (long reads) and Illumina (short reads) sequencing in a murine model of neural stem cell differentiation. PacBio sequencing detects whole transcripts (mean length resolved is 3000 bp) and is ideal to elaborate precise transcriptomes and perform isoform discovery. The trade off is the high error rate (around 5%) and the loss of quantification power. Complementarily, Illumina allows for quantification of expression and for the correction of error-prone long reads.

Results:

Classification of our PacBio transcriptome based on the splice pattern of isoforms reveals 60% of transcripts match annotated references in Refseq and ENSEMBL, 30% show novel splice junctions and 5% map to regions thought to be deprived of coding potential (genic introns and intergenic regions). Analysis of splicing features such as non canonical splicing rate, retrotranscription artifacts or Splice Junction coverage among others revealed that PacBio transcriptome needed further curation. We have developed a classifier to deal with this curation and results show that curated transcripts show better splicing features. Further characterization of the novel isoforms involved the evaluation of their peptide coverage using large databases of mass spectrometry profiles. Lastly, important expression associations can be made from this data: we found that most of multi-isoform genes expressed at least one additional annotated isoform at greater levels, in the majority of the cases it being the so called Principal Isoform, while a reduced subset of genes only expressed the novel isoform.

Conclusions:

Our results prove that the output of the PacBio Isoseq pipeline requires careful curation in order to eliminate isoforms showing abnormal features of splicing. After this curation has been done, the percentage of novel isoforms remains as high as 30% indicating the suitability of PacBio to perform the discovery of novel isoforms that are robust. Besides as we and others have found, the use of a filtered transcriptome instead of a global reference, diminishes the amount of quantification artifacts. Altogether these results shed light into the complex dynamics of alternative splicing and points to the necessity of using restricted transcriptomes to adequately analyze gene expression at the isoform level.