Supplementary Material for "Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows"

Elsa Bernard[1,2,3], Laurent Jacob[4], Julien Mairal[5], Jean-Philippe Vert[1,2,3]

September 10, 2013

1 Sparsity of the ℓ_1-penalized estimator

We illustrate here the fact the flow decomposition returns a solution of the ℓ_1-penalized estimator (problem 3 in the main paper) which is sparse in the number of transcripts. Figure S1 shows the final number of predicted transcripts after flow decomposition and model selection for genes with a particular number of expressed transcripts.

2 Gene size influence on isoform recovery

In the main paper we stratified precision and recall for isoform recovery by the number of expressed transcripts for each gene (Figure 3) . The number of exons of a gene is also a parameter that affects greatly the difficulty of the problem. Indeed, the more exons the bigger the set of candidate transcripts. Figure S2 shows similar experiments as the ones presented in Figure 3 of the main paper with the only difference being the exon stratification instead of the transcript stratification. The number of exons varies from 2 to 116 and we compare FlipFlop, Cufflinks and IsoLasso.

For both single-end and paired-end reads, FlipFlop performance increases greatly compared to Cufflinks and IsoLasso when the read length increases (Figure S2(a) and Figure S2(b)). For 300bp read length FlipFlop outperforms Cufflinks and IsoLasso for all genes with between 2 and 20 exons. Similarly to what we observed on simulations by transcript levels, and because FlipFlop predicts its transcripts by using both read alignment positions and read density without any filtering, an increase in coverage leads to better results for all exon levels (Figure S2(c)).

[1]Centre for Computational Biology – CBIO, Mines ParisTech, Fontainebleau, France, [2]Institut Curie, Paris, France, [3]INSERM U900, Paris, France, [4]LBBE, Lyon, France, [5]LEAR Project-Team, INRIA Grenoble - Rhône Alpes, France

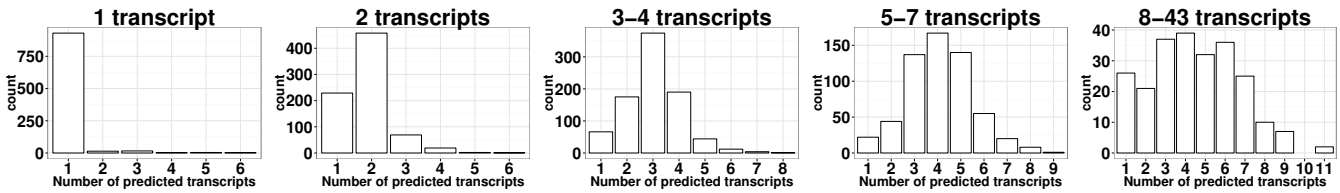
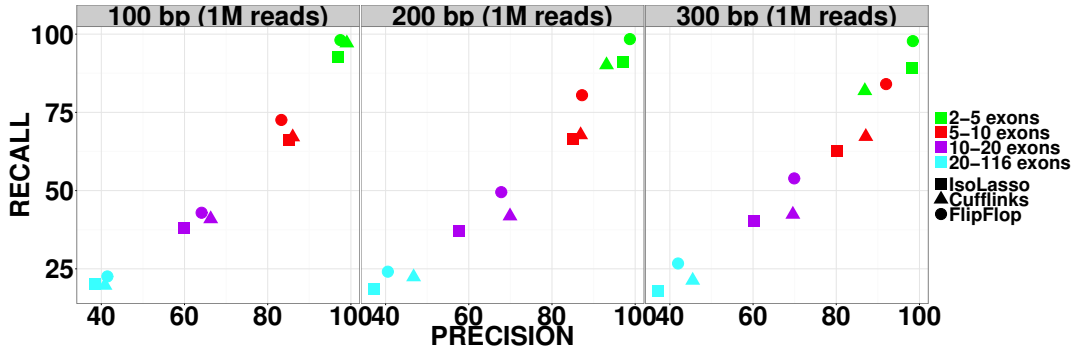
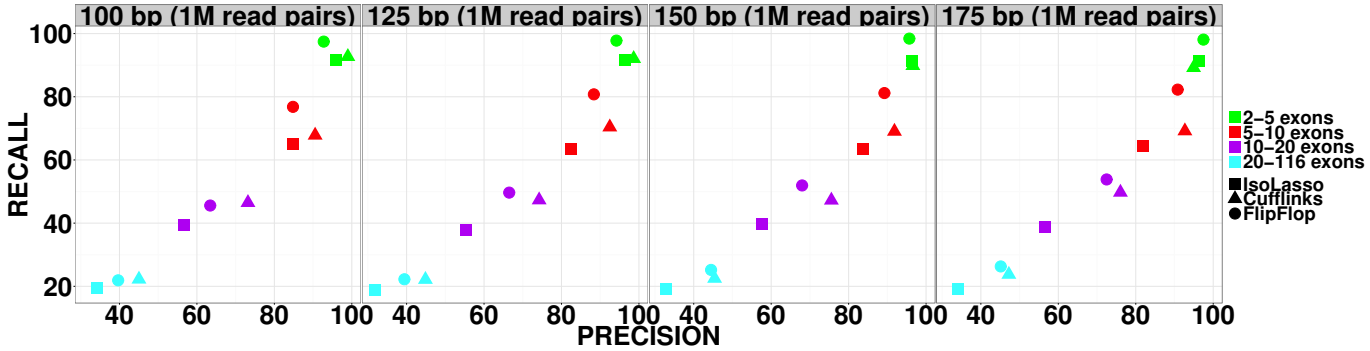


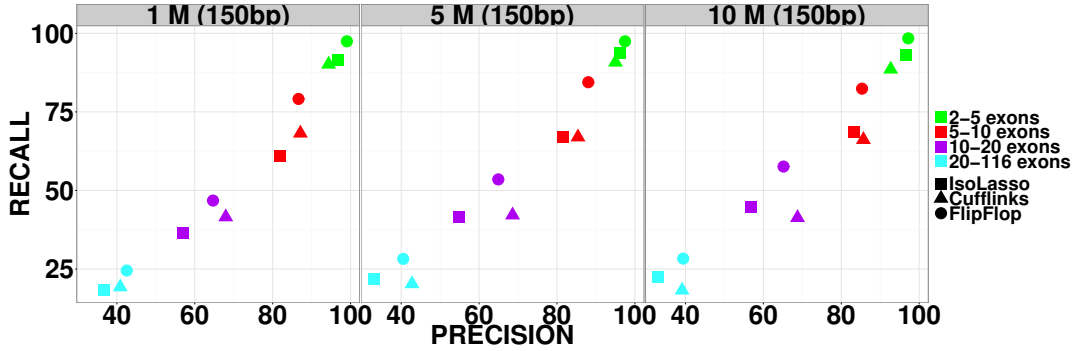
Figure S1: Number of predicted transcripts for human RNA-Seq simulations with 150bp long single-end reads and 1 million reads by expressed transcript levels.



(a) Single-end reads with different lengths (100, 200, 300bp) and 1 million reads by exon level.



(b) Paired-end-end reads with different lengths (100, 125, 150, 175bp), 400bp fragment length and 1 million read pairs by exon level.



(c) Single-end reads with a fixed 150bp length and an increasing amout of material (1, 5, 10 million)

Figure S2: Precision and recall on simulated reads from the UCSC annotated human transcripts with an exon stratification.