

USING THE FLUX CAPACITOR IN GTEx

The reasons to use the transcript quantifications produced by the FluxCapacitor (abbr. as Flux) in the pilot phase of the GTEx project have already been described (<http://liorpachter.wordpress.com/2013/10/31/response-to-gtex-is-throwing-away-90-of-their-data>). Below we describe an evaluation of Flux for the use in GTEx.

1. Performance evaluation of Flux using simulation.

(i) Caveats to benchmarks based on simulated data. While simulations are valuable for evaluating the performance of computational methods, the effect of underlying assumptions going into the simulation will always influence the outcome. The post by Lior shows that for 100M reads of simulated data, Cufflinks has an accuracy of 0.05-0.06 higher than that of the Flux (as measured by the Spearman correlation). The data has been simulated according to the Roberts-Pachter schema. This schema simulates the data using a generative model of the objective function implemented in eXpress, which is also similar in spirit to the objective function in RSEM and Cufflinks. In other words, the data has been generated with the behavior that is exactly expected by eXpress. Hence, one might a priori expect both Cufflinks and eXpress to better match the Pachter simulated datasets than FluxCapacitor (0.95 correlation for Cufflinks vs. 0.89 for the Flux).

(ii) Evaluation of Flux and Cufflinks using Flux Simulator. Given the extraordinary complexity of the human transcriptome it is unlikely that any simulation will comprehensively recapitulate the real biology. We have also performed benchmarking simulations to evaluate Flux, but using a different simulator. We simulated data using the Flux Simulator (<http://www.ncbi.nlm.nih.gov/pubmed/22962361>), a published method that has been widely used by the community—including by scientists involved in the recent development of TopHat (<http://www.ncbi.nlm.nih.gov/pubmed/23618408>.) We have simulated data closely matching that used in GTEx: 50 million 76bp paired end reads. The simulation does not share the objective function optimized by either Flux Capacitor or Cufflinks. Our results show very similar accuracy for Cufflinks and the Flux. Considering all simulated transcripts, Spearman correlation is 0.76 for the Flux, and 0.75 for Cufflinks (Figure 4). These values, although still high, are consistent with the limited available ground truth data based on independent experimental quantifications (Steijger et al., Nature Methods in press*).

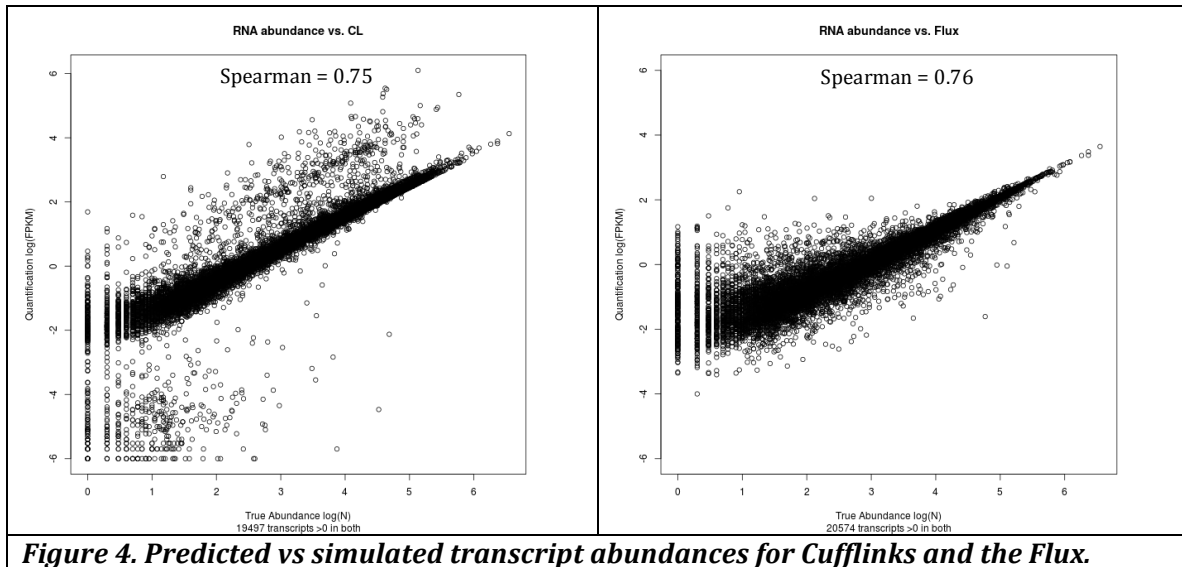


Figure 4. Predicted vs simulated transcript abundances for Cufflinks and the Flux.

Spearman correlation, however, may not be the most appropriate measure of transcript quantification accuracy, given the large dynamic range of transcript abundances in the cell. Spearman measures correlation of ranks, and ignores the absolute values of transcript expression levels. For instance, let's assume a transcriptome with two genes with measured abundances of 1 and 10^6 . If method A predicts 1 and 10^6 , and method B predicts 10.1 and 10.2, both have Spearman correlation of 1, despite the fact that method A is much more accurate than method B. To address this limitation, as well as limitations inherent to other correlation-based methods, Ribeca and Sammeth have developed an alternative approach (<http://algorithms.cnag.cat/pool/pearson.pdf>) that allows to compute a meaningful Pearson coefficient after an optimized power-law transformation of the data.

(iii) Consistency of simulation results with realistic expectations. As mentioned above, the Flux simulation produces Spearman values consistent with available experimental evaluations. However, the Pachter simulation produces very high correlations, even for very low coverage sequencing. Pachter reports a prediction accuracy of 80% for eXpress at 1M (10^6) reads. Assuming that the average number of fragments per expressed gene is 67 at this depth of sequencing (1M) (#) and given the dynamic range of gene expression of 7 orders of magnitude observed in GTEx, it is unlikely that such a small number of reads would give good transcript-level quantifications. Realistic assumptions about the underlying biology are necessary when assessing computational methods, and flawed simulations can lead to incorrect methodological decisions that may be suboptimal given real data.

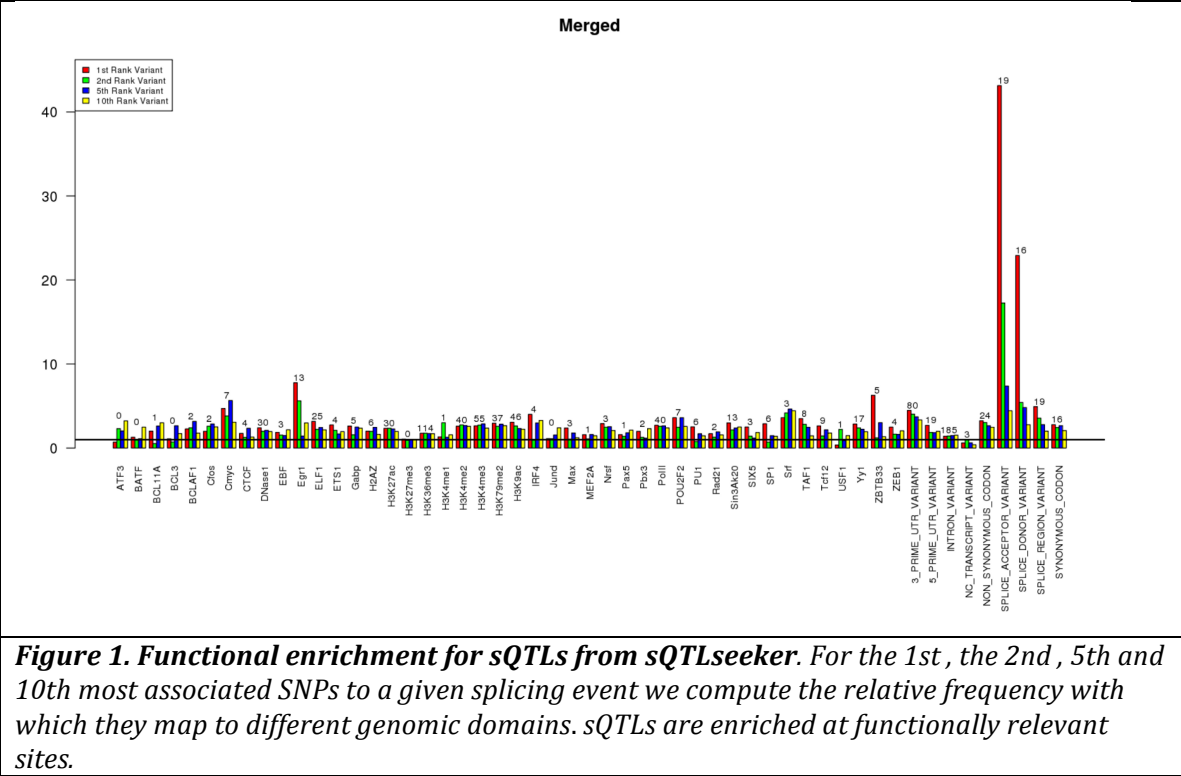
(*) The Steijger et al. analysis is part of RGASP, the RNASeq Assessment Project, carried out within the ENCODE project. In Steijger et al., methods to reconstruct transcript isoforms were evaluated. Programs like the Flux, MISO, RSEM and others that produce quantifications only on annotated elements were, therefore, not included in the evaluation. As part of the analysis, a few of the evaluated programs that also produced quantifications were evaluated against independent

quantifications obtained using the Nanostring technology. The accuracy reported for Cufflinks (measured as Pearson correlation of log transformed values) was 0.68-0.74. Since the Nanostring benchmark targeted mostly relatively straightforward cases, this numbers should be considered over estimates.

(#) Considering that the transcribed coding genome is 50 Mbps (~25,000 genes), and assuming that for any given tissue about 60% of genes are expressed (30 Mps), then 1M fragments would result in average coverage of 5X (150 Mbps at 150 bps per fragment given 75bp paired end reads), or on average 67 fragments per expressed gene.

2. SpliceQTLs found based on Flux quantifications reflect biological expectations. As part of the GTEx pilot project, we used Flux to quantify transcripts so as to identify SNPs that associate with splicing changes (splicing QTLs, sQTLs). To that end, in addition to more standard exon-centric methods, we developed a new method (called sQTLseeker) that identifies the genetic variants that associate with a gene’s splicing phenotype defined as the multivariate distribution of the relative abundances of the gene’s splice isoforms (Monlong et al., currently under revision). We applied the method to the transcript quantifications obtained with the Flux and obtained results that are consistent with biological expectations:

(i) Flux based sQTLs were enriched for variants that would be expected to have the greatest impact on splicing (e.g. splice acceptor and donor sites) (Figure 1).



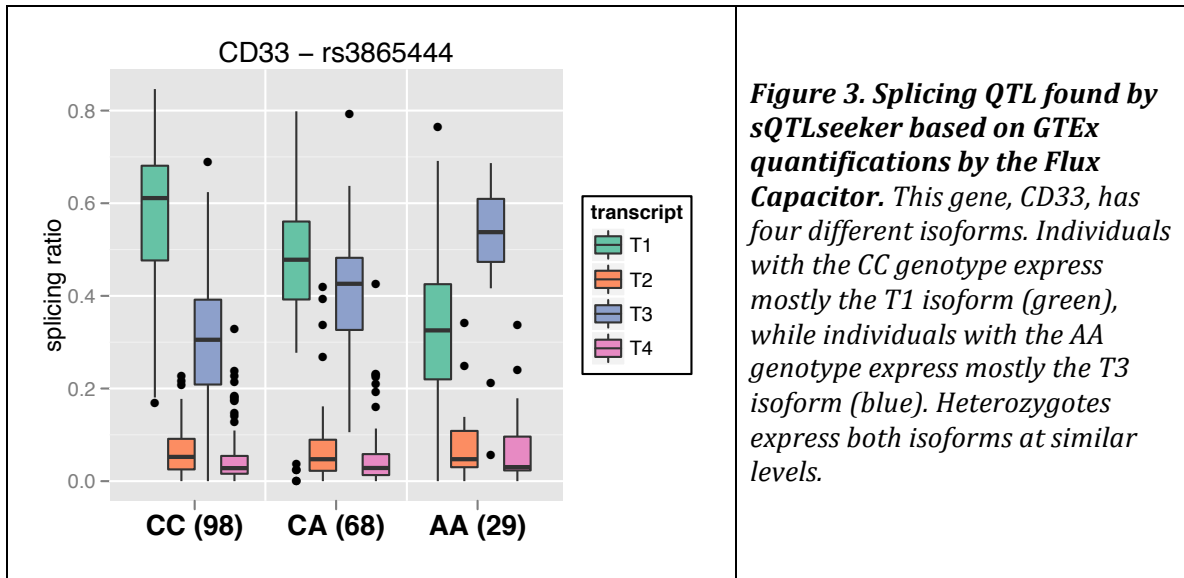
(ii) We tested whether sQTLs were enriched in sites that are predicted to have an effect on splicing based on their splicing motif disruption (evaluated using a position weight scoring matrix). We observed that sQTLs were more likely to be consistent with predicted splicing disruption. Mutations disrupting splice site sequences have the expected effect on the usage of the splice site much more often in sQTLs than in non sQTLs (Figure 2)

Method / Tissue	non-sQTLs		sQTLs	
	Consistent	Inconsistent	Consistent	Inconsistent
sQTLseeker				
Adipose Subcutaneous	0.089 2630/29490	0.075 2224/29490	0.373 82/220	0.136 30/220
Artery Tibial	0.08 2273/28578	0.066 1890/28578	0.385 84/218	0.083 18/218
Heart Left Ventricle	0.103 2961/28764	0.093 2683/28764	0.423 63/149	0.128 19/149
Lung	0.08 2446/30615	0.057 1754/30615	0.492 62/126	0.056 7/126
Muscle Skeletal	0.089 2474/27874	0.072 2007/27874	0.471 64/136	0.044 6/136
Nerve Tibial	0.092 2745/29857	0.084 2497/29857	0.424 72/170	0.318 54/170
Skin Sun Exposed Lower leg	0.096 2888/29971	0.068 2052/29971	0.425 51/120	0.092 11/120
Thyroid	0.081 2428/30159	0.065 1952/30159	0.346 71/205	0.18 37/205
Whole Blood	0.087 2440/28094	0.074 2083/28094	0.5 29/58	0 0/58

Figure 2. Frequency of consistent (positively correlated) and inconsistent (negatively correlated) changes between splice site strength (as measured by a Position Weight Matrix like method) and usage (as measured by RNASeq) for SNPs occurring in splicing sites. The frequency of consistent and inconsistent changes is similar in non-sQTLs SNPs. In contrast, in sQTL SNPs the frequency of consistent changes is several fold that of inconsistent changes.

These results are consistent with the simulation results, which suggest that Flux accurately reflects real biology.

Finally, in Figure 3 we show an example of an sQTL found by sQTLseeker. This is a GWAS hit for Alzheimer disease for which, until the recent publication of Battle et al. (<http://www.ncbi.nlm.nih.gov/pubmed/24092820>) no known eQTL existed.



3. Scalability and reproducibility testing of Flux. Flux ran smoothly on the 1,800 RNASeq samples, and results were 100% reproducible at the CRG and at the Broad.

In conclusion, we believe that Flux is an appropriate method for isoform quantification in GTEx. It performs comparably to Cufflinks in simulation and produces splice QTL results consistent with biological expectations. Flux also integrates well into a high throughput production setting because of its consistent resource requirements.