Supplementary Materials: CLIIQ: Accurate Comparative Detection and Quantification of Expressed Isoforms in a Population

Yen-Yi Lin[1,†], Phuong Dao[1,†], Faraz Hach[1,†], Marzieh Bakhshi[1], Fan Mo[2], Anna Lapuk[2], Colin Collins[2], and S. Cenk Sahinalp[1]

[1] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[2] Vancouver Prostate Centre & Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada
† These authors contributed equally to this work.
* Corresponding Author cenk@cs.sfu.ca

1 Performances on Mapping Results Using TopHat

In previous experiment we assume that all reads can be uniquely mapped to the correct location. In other words, even a read is split into several parts due to junctions, we still provide information in the mapping results such that all methods do not have to handle with problems of ambiguous or disgarded reads. However, in real datasets we may have several problems with split reads: such reads might be missed due to short anchor size, or possess multiple locations of mapping since different exons have similar prefixes/suffixes. These issues, which come from the difficulty of splice junction mapping, will decrease the performance of CLIIQ. To reflect these problems of splice mapping, we use the mapping results generated by TopHat (Version 1.4.1), instead of the perfect mapping results used before, and provide the performances of all the methods below. As described in text, we select the error tolerance, ϵ, by running the CLIIQ on 100 randomly selected genes with different ϵ values and select the ϵ value which has the highest precision and F-score. Table **??** provides the F-score and precision for different values of ϵ of CLIIQ. We select 0.5 for the remained experiments as CLIIQ has the best F-score and precision.

	0.3	0.35	0.4	0.5	0.6
ID Precision	0.5758	0.5777	0.5799	0.6001	0.5990
ID F-score	0.6294	0.6285	0.6287	0.6451	0.6419

Table 1. Performance of CLIIQ on isoform identification for test data with different ϵ values for real mapping results.

We first consider the performance of isoform identification given mapping results of TopHat. We provide precision, recall, and F-score of isoform identi-

fication in Table **??**. In both experiments, single sample formulation of CLIIQ basically achieves comparable results to Cufflinks, and better than IsoLasso. Multiple sample formulation of CLIIQ provides better results than Cufflinks.

	First Experiment				Second Experiment			
	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)
Precision	0.7630	0.6282	0.7331	0.7929	0.7759	0.6203	0.7779	0.7818
Recall	0.6348	0.5903	0.6309	0.6995	0.6996	0.6394	0.7209	0.7642
F-Score	0.6930	0.6087	0.6782	0.7433	0.7358	0.6297	0.7483	0.7729

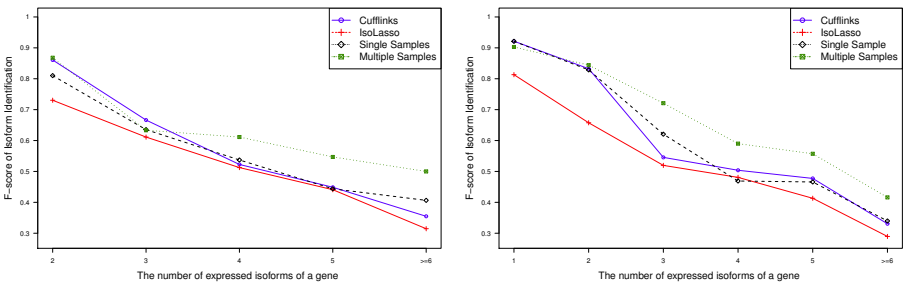
Table 2. Performance of various methods on isoform identification of ϵ=0.5 based on mapping results of TopHat.

For isoform quantification, single sample and multiple sample formulation of CLIIQ perform better than other tools in the first experiement. For the second experiment, CLIIQ performs better than Cufflinks, but similar to IsoLasso.

	First Experiment				Second Experiment			
	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)	Cufflinks	IsoLasso	CLIIQ (Single Sample)	CLIIQ (Multiple Samples)
Precision	0.3139	0.3221	0.3672	0.3935	0.3530	0.4371	0.4460	0.4537
Recall	0.2611	0.3026	0.3160	0.3472	0.3182	0.4505	0.4318	0.4435
F-Score	0.2851	0.3121	0.3396	0.3689	0.3347	0.4437	0.4482	0.4485

Table 3. Performance of various methods on isoform identification and quantification with error 0.1 based on mapping results of TopHat.

We then examine the performance of different methods with respect to the number of expressed isoforms in a sample. In Figure **??**, we see that in both experiments, single sample formulation of CLIIQ performs similar to Cufflinks. Moreover, as in perfect mapping case, for samples with higher number of expressed isoforms, multiple formulation of CLIIQ outperforms other tools.



(a) (b)

Fig. 1. The F-score of isoform identification of the first experiment(left) and the second experiment(right) with respect to isoform number in a gene