Comparing Kallisto and Salmon RNAseq quantification tools

By- Shashank Katiyar

Purpose-

High-throughput sequencing has transformed genomics by enabling rapid, cost-effective transcriptomic data analysis. Kallisto and Salmon, popular RNA-seq tools, utilize pseudo alignment instead of aligning reads to the genome, ensuring speed and low memory requirements. Recognizing their similarities and differences is crucial for informed tool selection in research analyses.

The primary objective of this project is to compare Kallisto and Salmon sequencing tools in terms of their algorithmic differences, speed, and accuracy in quantifying transcript abundances. Specifically, I aim to:

1. Investigate the algorithmic differences between Kallisto and Salmon, focusing on their underlying methodologies for transcript abundance estimation.
2. Evaluate the speed differences between Kallisto and Salmon in generating index files and quantifying transcript abundances using commands like time (terminal).
3. Assess the accuracy of Kallisto and Salmon by comparing their results to publicly available processed data from the ENCODE project.

Approach-

To compare the tools, I needed data and corresponding standard results (transcript quantifications) to which I could compare the results from the two tools. I extracted data from ENCODE (experiment id: ENCSR813BDU). The data consisted of two raw fastq files (paired-end reads), and the quantification tsv file, which was obtained by the people who originally conducted the experiment using the STAR aligner followed by the RSEM quantifier.

My plan was to first generate the index for the two tools (the index stores the information about the reference transcriptome in a format that is suitable for the tool and helps in doing quantification efficiently). Then, I would run the quant commands (actual commands below) for the tools to get the quantification tsv files. Lastly, I would process the tsv files in a jupyter notebook (comparing_quantifications.ipynb) to compare and contrast them.

Commands:

Kallisto quantification command:

```
time kallisto quant -i index_files/kallisto_index/index.idx -o kallisto_output ENCFF765REV.fastq.gz ENCFF339UJI.fastq.gz
```

Salmon quantification command:

```
time salmon quant -i index_files/salmon_index -l A -1 ENCFF765REV.fastq.gz -2 ENCFF339UJI.fastq.gz --validateMappings -o salmon_output
```

Results-

Basic Algorithm differences

Kallisto creates a De Bruijn Graph, which efficiently stores the possible k-mers of the reference transcriptome. It maps the reads to the transcripts using the possible k-mers from the reads. Finally, it employs an iterative Expectation-Maximization algorithm to determine the transcripts' quantification, which has the maximum likelihood given the set of reads we have.

Salmon, on the other hand, employs a two-step approach to quantify transcripts. The first step is an online step in which the initial expression levels of the transcripts and model parameters are estimated. The next step is the offline step, in which the model parameters and expression estimates are fine-tuned. Salmon builds a probabilistic model of the data, considering various factors that influence the probability of a read arising from a particular transcript. Kallisto doesn't take these factors into account.

Speed and Accuracy-

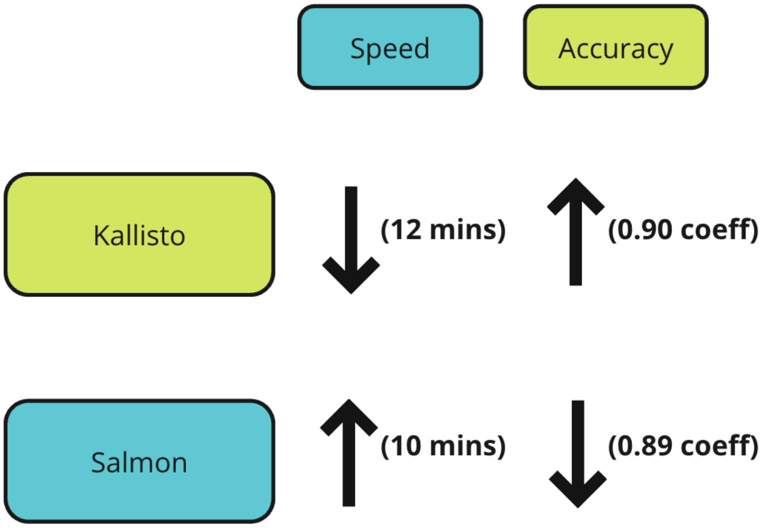

Fig1. Summary of the speed and accuracy comparisons

Kallisto was the slower method of the two in both, making the index and quantifying the transcript abundances. It took around 11 minutes to build the index, unlike the 2 minutes taken by salmon. Transcript quantification using kallisto took around 12 minutes, whereas salmon took around 10 minutes to do the same.

Although Salmon claims to be more accurate than Kallisto, I found it slightly less similar to the quantifications from a proper alignment strategy (alignment using STAR followed by quantification using RSEM). Hence, at least in my analysis, Salmon was the less accurate method.

The correlation (measured using the .corr() method of pandas dataframe, which defaults to calculating Pearson correlation) was the statistic used to measure the similarity between the results of the different methods.

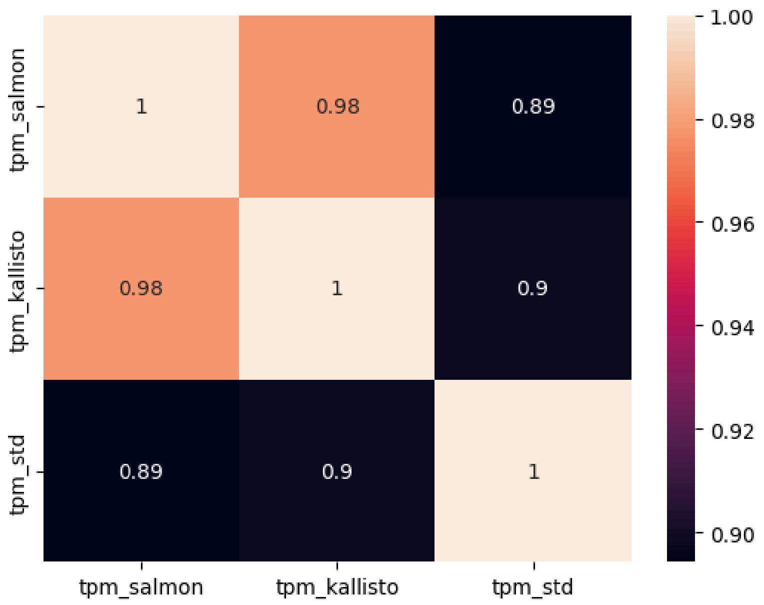

Fig 2. Correlation Coefficients between the three quantifications

Further Improvements

The following were the major drawbacks of my analysis and can be improved further.

- In this analysis, I have just considered one data point (total RNA seq sample from the HepG2 cell line in Homo Sapiens) to compare Salmon and Kallisto. However, to get more confident results, more data points are required.
- There are many more modes of RNA sequencing (paired-end, single cell, etc.). In this case, it is paired-end bulk sequencing, but the other modalities may yield different results when the two tools are compared.
- Also, since a significant advantage of Salmon is that it considers many different types of biases in the fastq files in the quantification, which Kallisto doesn't, if the file has little bias to begin with, Salmon is at a disadvantage in the comparison.

References

- https://salmon.readthedocs.io/en/latest/salmon.html
- https://pachterlab.github.io/kallisto/manual
- https://www.encodeproject.org/