Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
18 Dec 2017
Manuscript Category

Genome Analysis

Visualizing Transcript Coverage

Akshat Singhal [1],*and Avi Srivastava [1]

[1] Computer Science Dept., Stony Brook University, New York, 11790, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Salmon is a tool for quantifying the expression of transcripts using RNA-seq data. However, quantification does not provide information on position coverage of a transcript and users/biologists are, often, also interested in knowing the position coverage of a transcript/gene. A visual representation of the transcripts can provide this information.
Raw data is already present in Salmon which can be processed to create overage vectors of transcripts and create plots using those vectors for further analysis.
Results: Coverage vectors have been created from Salmon's data using very simple and efficient methods and transcript plots created from these vectors clearly show position biasing in transcripts.
Contact: aksinghal@cs.stonybrook.edu

1 Introduction

Salmon is a tool for quantifying the expression of transcripts using RNA-seq data. Salmon performs its inference using an expressive and realistic model of RNA-seq data that takes into account experimental attributes and biases commonly observed in real RNA-seq data.

However, quantification does not provide information on position coverage of a transcript and biologists are, often, also interested in knowing the position coverage of a gene more than just the value of transcript expression.

This project, as an add-on to Salmon, uses the raw data already present in Salmon to create coverage vectors for transcripts. These vectors can be easily plotted to view a transcript's position coverage and infer its biases.

2 Approach

Fig. 1 shows the process to create the plot of a transcript quantified by Salmon. A file is generated through Salmon which contains starting position of every mapped read to every transcript.

This file is, then, used to create coverage vectors of transcripts.

Finally, these coverage vectors are used to create transcript expression plots.

3 Methods

Position mapping of read to transcript is already present in Salmon. Hence, very simple methods are used to create coverage vectors efficiently. These methods are explained in detail below.

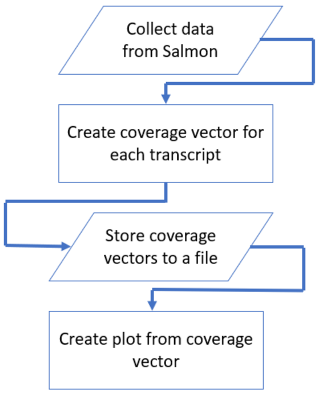
Collect data from Salmon
Create coverage vector for each transcript
Store coverage vectors to a file
Create plot from coverage vector

Fig. 1. Flow diagram

3.1 Collecting data from Salmon

Salmon uses new algorithms (specifically, coupling the concept of quasi-mapping with a two-phase inference procedure) to provide accurate expression estimates. This means there is data in Salmon which can be exploited.

During quasi-mapping, each read(short nucleotide sequences) is mapped to zero or more transcripts.

If a read is mapped to a transcript, then the read's ID along with transcript's ID and mapped position are redirected to a file.

3.2 Creating Transcript Coverage

The data collected from Salmon acts as input for creating coverage of transcripts. Once the vectors are created, they are saved to a file for creating plots.
A very basic algorithm is used to create coverage which is as follows:

1. Empty transcript coverage vectors for all transcripts are created.
2. Each read is assigned a weight 1. Lets call it W.
3. For each read(r_j), expression of all transcripts are added. Lets call it Sum_{r_j}.
4. For every transcript(t_i) to which r_j was mapped, its expression value(E_{t_i}) is divided by Sum_{r_j} and this fraction is added to the position k at which r_j was mapped to t_i.

$$E_{t_{i_k}} = \sum_{j=1}^{r} \frac{E_{t_i}}{Sum_{r_j}} * W$$

Since, W = 1, $E_{t_{i_k}}$ reduces to,

$$E_{t_{i_k}} = \sum_{j=1}^{r} \frac{E_{t_i}}{Sum_{r_j}}$$

5. Repeating steps 2. to 4. gives us coverage vectors for all transcripts.

3.3 Plotting Transcript Coverage

Creating a plot from a coverage vector is a just a 2-step process:

1. Take transcript ID whose plot is to be created.
2. Read coverage vector of the transcript and create its plot.

4 Result

4.1 Plots

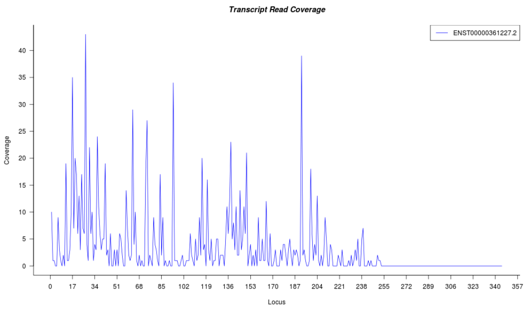

Fig. 2. Moderately expressed transcript

Fig. 2 and Fig. 3 show plots created for two different transcripts.

- Fig. 2 is a moderately expressed transcript. It is more expressed at the initial positions.
- Fig. 3 is a densely expressed transcript as compared to transcript in Fig. 2. This transcript is more expressed on its right end.

4.2 Time and Space Analysis

In general, if total read count is m, total transcript count is n, and maximum length of transcript is l, then time complexity of each step is :

1. $O(m)$
2. $O(m+n)$
3. $O(l)$

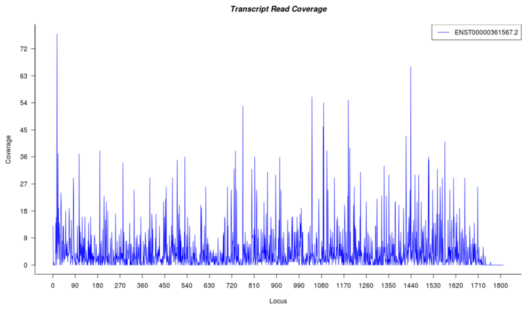

Fig. 3. Highly expressed transcript

Table 1. Time and Space Analysis

Step	Time	Disk Space
Collecting data from Salmon	50 sec	5 GB (intermediate)
Creating transcript coverage	120 sec	600 MB (intermediate)
Plotting transcript coverage	0 sec	70 KB

Data collected for 2290063 reads and 199612 transcripts
Step 3 shows result for only 1 transcript

5 Future Work

- Visual transcript coverage provides a lot of scope for analysis of other biological structures. One such example is analysis of isoforms of genes. Since, transcripts are components of genes, coverage of all transcripts together at one place can help in visually analyzing a gene.
- Transcript coverage may also be modified to provide color based coverage of exons in a transcript. Exon coverage can help in comparing transcripts containing similar exons.
- Time and space complexity of the methods used can be further improved by using binary files instead of text at intermediate stages.
- Plots created for transcripts can be made interactive as a part to improve visual tools.

6 Conclusion

Although Salmon provides very accurate result of quantification of transcripts, biologists are also interested in position coverage of transcripts for better analysis.
Through this project, position coverage in the form of coverage vectors have been successfully created using very simple and efficient methods and plots of these vectors clearly show position biasing in transcripts.
Further, in this project, transcript coverage can also help in gene and exon analysis through visual tools.

References

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.