TRUFA: User's manual

January 30, 2014

Contents

1 What is this web server for ?

"RNA-seq, also called whole-transcriptome shotgun sequencing, refers to the use of high-throughput sequencing technologies for characterizing the RNA content and composition of a given sample" (Wolf, 2013). Until now, analyzing RNA-seq data remains a Bioinformatic challenge. This web server is designed to make analysis of RNA-seq data in a fast and user-friendly manner, by using cluster computing and reducing the amount of Bioinformatic knowledge necessary.

1.1 Quick background on RNA-seq analysis

Recently and especially with the development of new massively parallel sequencing method, the cost of sequencing whole transcriptomes fell drastically. With a thousand dollars, you can get Gigas worth of transcriptome sequences.

In the case of Illumina sequencing, you will receive as output 1 (single-end) or 2 (paired-end) files per samples in fastq format (.fastq, .fq or compressed fastq such as fq.tar.gz, fq.gz ...).

In order to get high throughput, the original RNA sequences have been shredded before sequencing and so your fastq files are filled with millions to billions of sequences, between 50 and 250 bp long each depending on the Illumina sequencing technology. These sequences are called "reads".

The first step in the RNA-seq analysis will be to clean the reads. Various cleanings can be performed, such as removing duplicates, adapters, poor quality bases/reads, putative contaminants etc...

The second step of the analysis is to reassemble the shredded pieces, the reads, into contigs (named transcripts in the case of RNA-seq) in order to reconstruct the original RNAs. To do so, two main methods are available:

- With a reference genome available, you will use "mapping" methods.

- Without a reference genome: you will use “*de novo* assembly” methods.

So far, TRUFA can only perform *de novo* assembly using the Trinity software. But more is to come ...

Once you obtained the assembly, the next step is to identify those contigs. This is particularly interesting if you are looking for specific genes. For example, if you are interested by venom toxins you could download a venom protein sequences database and then blast your assembly against it. You'll get hits, representing mRNA which could represent precursors of venom proteins. Another way to identify sequences is to scan your assembly for protein profiles (using HMMER for example).

Recently, a specific vocabulary called “Gene Ontology” has been developed in order to help in the identification process. The idea is to link biologically meaningful keywords to nucleotide sequences to identify their biological role, molecular function and location in cellular components. This can be performed using Blast2GO.

Finally, the last main step of an RNA-seq analysis will be to quantify the expression. One of the interest of having high throughput data is that the number of reads sequenced per contigs should reflect the expression levels of the corresponding gene. To quantify the expression, the first step will be to align all reads back to the transcripts (using BOWTIE for example). Then, programs such as CUFFLINKS can provide you with expression values (called RPKM in the case of single-end reads and FPKM in the case of paired-end reads) and then perform differential expression analysis.

For more basics on RNA-seq

Wolf (2013)
Martin and Wang (2011)

1.2 What can be done with TRUFA ?

So far, TRUFA is allowing you to perform the following steps (programs used are specified in the parentheses):

- Reads cleaning:
  - Quality trimming and duplicates removal (Prinseq)

  - Trimming adapters (Cutadapt)
  - Filtering out potential contaminants (Blat)
- De novo assembly of your reads (Trinity)
- Reads mapping (Bowtie)
- Contigs (i.e transcripts) identification:
  - based on sequence alignment (Blat, Blast)
  - based on protein dominions, profiles (HMMER)
  - Annotation with GO terms (Blast2GO)
- Expression quantification:
  - providing FPKMs (cufflinks)

All the steps of the pipeline can be run as a whole or independently, depending on your needs. In the case of making a complete pipeline, your reads files will be cleaned, then assembled into transcripts and finally the transcripts will be identified and quantified.

2 Input:

The input files can be of multiple sources, depending on the kind of analysis you want to perform and the data you have available. So far the accepted inputs are:

- 1 or 2 Illumina reads files (i.e. single-end or paired-end reads) in fastq or compressed (extensions .tar.gz or .gz) fastq format
- 1 file with already assembled reads, i.e contigs in fasta format

⚠ **Important!**
On top of the necessary inputs, you will need to specify some important information related to the library construction:

- Size of the insert
- Adapters sequences
- Strand specific reads ?

3 Output:

So far, TRUFA is providing the classical outputs of a RNA-seq analysis. For a complete analysis, you will get (type of file precised between parenthesis):

- Prinseq report of reads quality (html)
- FastQC report of reads quality (zip and html)
- Reads files after complete cleaning (fastq)
- Assembly file (fasta)
- File with alignment of the reads against the transcripts (bam)
- Results of the Blast2GO annotation (dat)
- Transcript expression quantification (txt)

4 Other useful programs:

Part of Trufa's output can be visualized with programs such as :

- Blast2GO
  http://www.blast2go.com/b2ghome
- IGV
  http://www.broadinstitute.org/igv/

- Tablet
  http://bioinf.scri.ac.uk/tablet/
- RNAseq Viewer
  http://bioinfo.au.tsinghua.edu.cn/software/RNAseqViewer/

Other programs for RNA-seq analysis can be found at:

- http://omictools.com/
- Wikipedia: RNA-seq tools

5 Quick Start:

5.1 Launch a Job

This is a quick demonstration on how the web server can be used.

5.1.1 Upload demo files

- In the Upload area, **Browse** your computer for the reads file “reads_left.fq.tar.gz”
- Before clicking **Send**, select **compressed reads file** in the drop-down menu (instead of **undefined**)
- Once successfully uploaded, do the same for the reads file “reads_right.fq.tar.gz”
- Once both files have been uploaded, you are ready to “start a job”.

Specifying the file type in the dropdown menu is mandatory and will help you by filtering the correct inputs for each steps of the analysis.

5.1.2 Start the job

- Go to **Start a Job**
- Select **Paired-end reads (2 fastq files)**
- Specify the 2 input files: "reads_left.fq.tar.gz" for **left reads** and "reads_right.fq.tar.gz" for **right reads**
- Go down to **RNA-seq steps** section and to the **Reads cleaning** tab.
- Check the boxes **Duplicated reads** and **Quality trimming** in the Prinseq section
- Check the box **FastQC** in the **Post-cleaning quality control** section
- Then go to the **Reads Assembly and Mapping** tab
- Check the boxes **Assemble with Trinity** and **Align reads against contigs with Bowtie2**
- Then go to the **Contigs identification** tab
- Go down to **Blast2GO searches** section
- Check the box **Blast2GO**
- Then go to the **Expression quantification** tab
- Check **Cufflinks**
- Then go down to **Launching Analysis** and hit the **Start** button.

A message confirming that the Job has been sent should appear. You can go back to **Home** in the main menu and here click on the corresponding Job in the Job list. You should be able to see the current state of your job there.

5.2 Check output

Once your job has been completed, here are some ways to check the generated outputs.

- Go to **File Manager** in the main menu
- Put your username and password
- Browse the **Jobs** folder for the corresponding Job name

5.2.1 Check reads quality:

- Go to **STAT** and then **prinseq_report**

Here you should have 4 html files (2 for each reads file, one report after the removal of the duplicates and one report after the trimming process) that you can download and then open locally with your internet browser. Information about Prinseq statistics can be found in the "quality control" section of Prinseq manual:
http://prinseq.sourceforge.net/manual.html#QC

- Go to **STAT** and then **fastqc_report**

Here you should have 2 zip files (1 for each reads file). Download them, extract them and open locally the html files with your internet browser. Information about the FastQC output can be found here:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

5.2.2 Check the assembly:

- Go to **ASSEMBLY_MAPPING** and then **trinity**

Here you will find the assembly file **Trinity.fasta**

- Go to **STAT** and then **assembly_qc**

Here you will find text files displaying statistics related to your assembly which will help you to determine the quality of your assembly.

5.2.3 Check reads mapping against transcripts

- Go to **ASSEMBLY_MAPPING** and then **bowtie2**
- Download the file **aligned_reads.bam** and **aligned_reads.bam.bai**

You can use these two files to visualize the alignments using programs such as Tablet http://bioinf.scri.ac.uk/tablet/.

5.2.4 Check the Blast2GO results:

- Go to **IDENTIFICATION** and then **b2go**
- Download the file **out_b2go.dat**
- Open it by starting Blast2GO program locally and going to **File**, **Load B2GO-Project (.dat)**

Blast2GO can be obtained from http://www.blast2go.com/b2ghome

5.2.5 Check the Cufflinks results:

- Go to **EXPRESSION** and then **cufflinks**

Here you will find two text files ("genes.fpkm_tracking" and "isoforms.fpkm_tracking") with the results of the expression quantification according to genes and isoforms as well as two GTF files. Informations of the text file format used by cufflinks can be found here:
http://cufflinks.cbcb.umd.edu/manual.html#fpkm_tracking_format

6 Running an analysis:

Essentially, all the steps of the RNA-seq pipeline on the web server can be realized with only reads files (1 or 2) as input. For the purpose of testing and parameters tuning, later steps can be as well directly performed with an assembly (fasta file with already assembled reads).

First, you should specify and upload the necessary input files:

- For Reads cleaning: You will need one or two fastq files (i.e. single or paired-end reads files)
- For Assembly: You will need one or two fastq files (i.e. single or paired-end reads files)
- For Identification, you will need either: a fasta file with the reads already assembled OR 1 or 2 fastq reads files and generate an assembly (see How to generate an assembly).
- For Reads mapping: 1 fasta file with the reads already assembled and the corresponding reads in 1 or 2 fastq files OR 1 or 2 fastq reads files and generate an assembly (see How to generate an assembly).
- For Expression quantification: 1 fasta file with the reads already assembled and the corresponding reads in 1 or 2 fastq files OR 1 or 2 fastq reads files and generate an assembly (see How to generate an assembly).

Then, by specifying in the **Run a Job** part the type of input and the input files, different possibilities will be available to you, depending on the type and number of input files you specified.

6.1 Reads cleaning:

6.1.1 Removing "contaminants" using BLAT:

6.2 Visualizing the output:

6.2.1 Cleaning outputs:

6.2.2 Assembly-Mapping outputs:

Assembly and reads alignment can be visualized using IGV by importing the assembly file (Trinity.fas) as a genome and loading the corresponding bam file(s).

6.2.3 Identification outputs:

6.2.4 Expression outputs:

7 Acknowledgments

References

Jeffrey A Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10):671–82, October 2011. ISSN 1471-0064. doi: 10.1038/nrg3068. URL http://www.ncbi.nlm.nih.gov/pubmed/21897427.

Jochen B W Wolf. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular ecology resources*, 13:559–572, April 2013. ISSN 1755-0998. doi: 10.1111/1755-0998.12109. URL http://www.ncbi.nlm.nih.gov/pubmed/23621713.