BioCloud Result Documentation
Documentation

Release 0.0.1

NOBODY

August 02, 2014
1 NGS Technology Introduction 3
  1.1 RNA-Seq ................................................................. 3
  1.2 DNA-Seq ................................................................. 3

2 RNA-Seq (Star) Pipeline Overview 5
  2.1 Quality Control .......................................................... 5
  2.2 Genome Alignment using Star .......................................... 5
  2.3 Gene Expression Quantification using Cufflink ....................... 5

3 Result Explained 7
  3.1 Quality Control .......................................................... 7
  3.2 STAR (Spliced Transcripts Alignment to Reference) .................. 9
  3.3 Cufflink ................................................................. 11

4 Indices and tables 13
NGS Technology Introduction

NGS is a awesome technology, everyone should give a hand on it.

1.1 RNA-Seq

1.2 DNA-Seq
CHAPTER 2

RNA-Seq (Star) Pipeline Overview

2.1 Quality Control

2.2 Genome Alignment using Star

2.3 Gene Expression Quantification using Cufflink
3.1 Quality Control

Next generation sequencing (NGS) technologies provide a revolutionary tool for numerous applications and are capable to generate several gigabases of sequence data in a single experimental run. These technologies are being increasingly used for various genome and transcriptome sequencing related applications due to their speed, cost-effectiveness and high-throughput nature. However, several sequence processes, including read errors (base calling errors and small insertions/deletions), poor quality reads and primer/adaptor contamination are quite common in the NGS data, which can impose significant impact on the downstream sequence processing/analysis. The quality of data is very important for various downstream analyses, such as sequence assembly, single nucleotide polymorphisms identification and gene expression studies. Most of the programs available for downstream analyses do not provide the utility for quality check and filtering of NGS data before processing. Therefore, these sequence artifacts need to be removed before downstream analyses, otherwise they may lead to erroneous conclusions.

The quality of data may be affected by several factors regardless of the NGS platform. Although the commercial vendors for all the sequencing platforms provide a quality control (QC) pipeline for filtering of sequencing output, several sequence processes still remain in the dataset. Therefore, it is advisable to perform QC and filtering of high-quality (HQ) sequencing data at the end-user level.

Figure 1. Babraham Bioinformatics Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (A & B) Shows an overview of the range of quality values across all bases at each position in the FastQ file. (C & D) Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called. (E & F) Measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content.

Chapter 3. Result Explained

Good Quality

Bad Quality

[Images of quality assessment graphs for good and bad quality]
3.1.1 Reference

3.2 STAR (Spliced Transcripts Alignment to Reference)

STAR \(^4\) is an alignment tool for RNA-seq, developed by Alexander Dobin et al, Cold Spring Harbor Laboratory, NY, USA, published on *Bioinformatics* in 2012: “STAR: ultrafast universal RNA-seq aligner”. STAR is implemented as a standalone C++ code. STAR is free open source software distributed under GPLv3 license and can be downloaded from [http://code.google.com/p/rna-star/](http://code.google.com/p/rna-star/). The latest version is 2.3.0.

STAR can identify the alternative splice junction in RNA-seq, and have almost the same accuracy of alignment as Tophat. The advantage of STAR is that time of alignment can be shorten into few hours or even in an hour.

3.2.1 Running STAR

Generate Genome for STAR

STAR build its own reference genome first. By analyzing the standard genome (ex, hg19 or mm10), STAR produce a suffix array index for accelerating the alignment step in the next step. The genome only needs to be generated once.

**Run**

```
/pathToStarDir/STAR                      # the location of STAR.exe
--runMode genomeGenerate                 # set up the runMode
--genomeDir /path/to/GenomeDir          # the location of output STAR genome
--genomeFastaFiles /path/to/genome/fasta1 /path/to/genome/fasta2
# the location of reference genome (ex, hg19)
--runThreadN <n>                        # the numbers of processing CPU
--sjdbGTFfile <FileName>.              # the location of annotation file(.gtf)
```

**Alignment**

STAR will output Aligned.out.sam as default. Samtools is needed after alignment step for transform Aligned.out.sam into Aligned.out.bam. As we will go through Cufflinks for further differential expression analysis in RNA-seq, following code option should be cautioned.

```
/pathToStarDir/STAR                      # the location of STAR.exe
--genomeDir /path/to/GenomeDir           # the location of genome built in the previous step
--readFilesIn /path/to/read1 [/path/to/read2]
# the location of input file
--runThreadN <n>                         # the numbers of processing CPU
--outFilterIntronMotifs RemoveNoncanonicalUnannotated
# remove the non-canonical junction.(unannotated junction)
--outSAMstrandField intronMotif         # for un-stranded sequence. If your input data is stranded
# sequence, please contact us.
```

3.2.2 Result

read the Log.final.out file.

- **Number of input reads** the reads of input file, it is associated with coverage and depth when sequencing.
• **Average input read length** the sequencing read length.

**Unique reads**

• **Uniquely mapped reads number** unique reads indicates the reads is only mapped on one spot. please compare with “Multi-mapping Reads” below.

• **Uniquely mapped reads %** the percentage of uniquely mapped reads. (Uniquely mapped reads number / Number of input reads) This value should not be lower than 80%, due to the mutations and sequencing errors usually won’t be more than 20%. If this value is under 80%, please checkout the parameter of alignment or contact with us.

• **Average mapped length** some of the reads might be missing or bad quality, and these reads will be filterd out before alignment. average mapped length should be as close as possible to the average input read length to make sure the quality of data is qualufued.

• **Number of splices(Total)** the number of alternative splicing.

• **Number of splices: Annotated(sjdb)** the splicing location which is known in the .gtf file(download from iGenome or UCSC)

• **Number of splices: GT/AG** the numbers of splicing location is between GT and AG

• **Number of splices: GC/AG** the numbers of splicing location is between GC and AG

• **Number of splices: AT/AC** the numbers of splicing location is between AT and AC

• **Number of splices: Non-canonical** the number of unannotated splices

• **Mismatch rate per base** the average mismatch rate in one nucleotide. With good library, this value should be 0.5%-0.8%

• **Deletion rate per base** the average deletion rate in one nucleotide

• **Deletion average length** the average of each deletion length

• **Insertion rate per base** the average insertion rate in one nucleotide

• **Insertion average length** the average of each insertion length

**Multi-mapping reads**

• **Number of reads mapped to multiple loci** the reads mapped on more than on spot, ex: poly-A tail

• **% of reads mapped to multiple loci** the percentage of mutli-mapping read. (Number of reads mapped to multiple loci/Number of input reads). Typically, this value should be 5%.

• **Number of reads mapped to too many loci** the default value is 10, which means if the reads mapped more than 10, it will be categorized in “too many loci”.

• **% of reads mapped to too many loci** the percentage of too many loci (Number of reads mapped to too many loci/Number of input reads)

**Unmapped reads**

• **% of reads unmapped: too many mismatches** the reads that contain too many mismatch, the default value is more than (0.3*read length) or 10 bases will be categorized in “too many mismatches”.
% of reads unmapped: too short  the percentage of reads too short to mapped. The default value of too short is 2/3*read length, which means if the reads are less than this value, it will be categorized in “too short”. This output value is usually associated with sequencing quality.

% of reads unmapped: other  the percentage of unmapped due to other causes. This situation might happen on short-RNA, because of the criteria of “too many mismatches” and “too short” can’t filter out the unmapped reads. The reason of unmapped might be contamination, suggest BLAST the reads for more details.

3.2.3 Reference


3.3 Cufflink

Cufflinks is an integrated analysis tool from transcripts assembles, quantification to comparison between different conditions. Cufflinks was developed by the Laboratory for Mathematical and Computational Biology, led by Lior Pachter at UC Berkeley, Steven Salzberg’s computational genomics group at the Institute of Genetic Medicine at Johns Hopkins University, and Barbara Wold’s lab at Caltech.

The first step of Cufflinks is to assemble all possible transcripts with the results of alignment from TopHat. To accurately estimate the abundance of each transcript, any possible isoforms might be considered in the algorithm. However, it is difficult to know how many splicing events exist in current data. Therefore, Cufflink reports few full-length transcript fragments or ‘transfrags’ which enough to explain the all possible isoform from the data. The quantification of expression level of transfrags is calculated after the phase of assemble.

To compare the different expression level of different conditions, Cuffmerge is the tool to merge previous assembles of individual conditions. With the step of merging assembles, the whole gene can be recovered, and the de novo transcripts also can be integrated into the complete the gene model.

The differential analysis is conducted by Cuffdiff. In addition to test the significant difference among two or multiple conditions, Cuffdiff also accept the multi replicate in each condition. Cuffdiff report the expression comparison in different levels, including genes, transcripts, and transcription start sites (TSS). In the final report, Cuffdiff provides the value of fold change, p-value, and multiple-test adjusted q-value.
CHAPTER 4

Indices and tables

- genindex
- modindex
- search