Single-cell RNA-Seq expression analysis

John J. Trombetta,David Gennert,Diana Lu,Rahul Satija,Alex K. Shalek,Aviv Regev

Abstract

This is a Single-cell RNA-Seq expression analysis *Support Protocol* for Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing.

Citation: John J. Trombetta,David Gennert,Diana Lu,Rahul Satija,Alex K. Shalek,Aviv Regev Single-cell RNA-Seq expression analysis. **protocols.io**

dx.doi.org/10.17504/protocols.io.pbidike

Published: 27 Jun 2018

Guidelines

Following sequencing of the cDNA libraries on an Illumina sequencer, data is generated as a series of files in the FASTQ format. For each unique sample specified in the sequencing sample sheet, four files are generated: one containing the “left-hand” read data (one end of the paired-end reads), one containing the “right-hand” read data (the other end of the pair), one containing the “left-hand” Nextera indexing read data, and one containing the “right-hand” Nextera indexing read data. RNA-Seq analysis uses computational tools to match each read pair, align the read pair to the genome sequence, and quantify the number of reads that align within each annotated gene.

The GenomeSpace web portal was developed to assist researchers with minimal computational analysis experience. Using its drag-and-drop interface, data sets and modules of pre-built analytic tools can be organized into customizable pipelines for numerous applications. Despite its ease of use, GenomeSpace uses cloud storage and computing power, making it less efficient for a large number of sequencing analyses or if a researcher has access to higher computing power at their own institution; as an alternative strategy for higher throughput, we also provide a Unix-based workflow.

Protocol

Step 1.

Please select **GenomeSpace** or **Unix command line** for expression analysis.

Step 2 - Unix command line (Protocol for using Unix command line for expression analysis.).

Ensure that the following programs are installed and ready to use on the computer or server that will run the analysis:

- TopHat – http://tophat.cbcb.umd.edu/
- Bowtie (or Bowtie2) – http://bowtie-bio.sourceforge.net/
- Samtools – http://samtools.sourceforge.net/
- Picard tools – http://picard.sourceforge.net/
- Integrative Genomics Viewer (IGV) – http://www.broadinstitute.org/igv/

- Cufflinks – http://cufflinks.cbcb.umd.edu/

Step 3 - Unix command line (Protocol for using Unix command line for expression analysis.).
Run the program TopHat to match each of the paired-end reads with its mate and align the reads to the desired reference genome.

- Files required:
- Reference genome index
  (download at ftp://ftp.ccb.jhu.edu/pub/data/bowtie_indexes/)
- Transcriptome reference annotation file (.GTF)
  (see http://tophat.cbcb.umd.edu/igenomes.shtml for up-to-date reference annotations)
- Left reads FASTQ file (either compressed or uncompressed)
- Right reads FASTQ file (either compressed or uncompressed

cmd COMMAND

```
% tophat [options] [genome index base]
    [~/LeftReads.fastq.gz] [~/RightReads.fastq.gz]
    For full list of [options], see
    http://tophat.cbcb.umd.edu/manual.shtml
```

TopHat will create several files, including ~/accepted_hits.bam. The data in this file contains the alignment and pairing information for all reads that successfully paired and aligned in the BAM file format

Step 4 - Unix command line (Protocol for using Unix command line for expression analysis.).
The Picard suite of command-line tools contains programs that provide important metrics regarding the sequencing data. Run various tools on the /accepted_hits.bam file to analyze the sequencing quality. Helpful tools include the following:

- CollectAlignmentSummaryMetrics.jar – view statistics on the number of reads that correctly align to the reference genome
- CollectInsertSizeMetrics.jar – view statistics on the lengths of the sequenced cDNA fragments

cmd COMMAND

```
    % java –jar ~/picard/CollectAlignmentSummaryMetrics.jar
    [options] I=~/accepted_hits.bam O=~/AlignmentMetrics.txt
    % java –jar ~/picard/CollectInsertSizeMetrics.jar [options]
    H=~/InsertSizeHistogramChart.txt I=~/accepted_hits/bam
```

NOTES

Anita Bröllochs 08 Apr 2018

For full list of tools and [options], see http://picard.sourceforge.net/command-line-overview.shtml

Step 5 - Unix command line (Protocol for using Unix command line for expression analysis.).
Sort the /accepted_hits.bam file using the Picard tool SortSam.jar, which organizes the BAM file data based on the aligned reads' locations in the reference genome.

cmd COMMAND

```
% java –jar ~/picard/SortSam.jar I=~/accepted_hits.bam
```

```
      O=~/accepted_hits.sort.bam SO=coordinate
```

Step 6 - Unix command line (Protocol for using Unix command line for expression analysis.).
Index the sorted BAM file using the Samtools suite of command-line tools.

cmd COMMAND

```
% samtools index ~/accepted_hits.bam
```

Step 7 - Unix command line (Protocol for using Unix command line for expression analysis.).
For graphical visualization of the data, load the aligned sequencing reads into the program IGV (Integrative Genomics Viewer). Open the IGV program, select the appropriate reference genome, and load the /accepted_hits.sort.bam file. This will display a graph for each file loaded with the location within the genome on the horizontal axis and the number of reads on the vertical axis.

Step 8 - Unix command line (Protocol for using Unix command line for expression analysis.).
For a quantitative measure of gene or transcript abundance, run the program Cufflinks on the /accepted_hits.sort.bam file. The two primary output files generated by Cufflinks represent quantified expression level estimates at the gene and transcript isoform level. Each of these presents the expression level estimates as a table relating each gene or transcript to its relative abundance in fragments per kilobase of exon per million mapped fragments (FPKM).

- Files required:
  - BAM data file
  - Transcriptome reference annotation file (.GTF)
    (see http://cufflinks.cbcb.umd.edu/igenomes.html for up-to-date reference annotations)

cmd COMMAND

```
  % cufflinks [options] ~/accepted_hits.sort.bam
```

NOTES

Anita Bröllochs 08 Apr 2018

For full list of [options], see http://cufflinks.cbcb.umd.edu/manual.html

Step 2.
Create an account at