Algorithms in Bioinformatics
Fall 2014

Hands on practices

Week 8: RNA-seq data analysis

RNA-seq is a technology to extract mRNAs from a sample followed by high-throughput sequencing. RNA-seq is a very powerful technology to study the gene expression and regulation. RNA-seq data can be extremely large. In this week, we will learn some steps in RNA-seq data analysis, especially the sequence mapping and consequence analysis steps.

1. Datasets. We have three types of data in this lab.

   1). RNA-seq data

   Lung_cell_1: 3

   Lung_cell_2: 3

   2). Human genome (hg19):

   hg19.fa, chr22.fa

   3). Gene annotation (GENCODE V19):

   protein_coding_chr22.gtf

   Datasets are under the following directories:

   /share/home/ccwei/pab/2014/week8/lung_cell_1/sra

   /share/home/ccwei/pab/2014/week8/lung_cell_2/sra

   /share/home/ccwei/pab/2014/week8/hg19_index

2. Softwares. We need 6 software packages to do this lab. They have been installed in our server under directory /share/home/ccwei/tools.

   Trimmomatic-v0.32

   Bowtie

   Tophat2

   Cufflinks

   FastQC2.

   Sratoolkit

3. Data analysis

3.1 Set the Environment variable PATH

Add the tools directories to PATH

Edit .bash_profile

>>> emacs ~/.bash_profile

Add the following line:

PATH=$PATH:/share/home/ccwei/tools/bowtie-1.1.1:/share/home/ccwei/tools/cufflinks-2.2.1.Li

nux_x86_64:/share/home/ccwei/tools/FastQC:/share/home/ccwei/tools/sratoolkit.2.4.2-centos_linux64/bin:/share/home/ccwei/tools/tophat-2.0.13.Linux_x86_64

save and quit emacs.

Then source the file to update your PATH environment variable:

>>> source ~/.bash_profile

3.2 Run the pipeline:

3.2.1.Transfer .sra format to .fastq format

>>> mkdir your_home_directory/week8

>>> cd your_home_directory/week8

>>> mkdir -p lung_cell_1/fastq

>>> mkdir -p lung_cell_2/fastq

For each raw file (for example lung_cell_1/raw_sra/SRR1033875.sra):

>>> fastq-dump –gzip --split-3 -o lung_cell_1/fastq/ /share/home/ccwei/pab/2014/week8/lung_cell_1/sra/SRR1033875.sra

3.2.2. Data statistics of raw data

>>> mkdir lung_cell_1/raw_sta

>>> mkdir lung_cell_2/raw_sta

For each raw fastq file (for example lung_cell_1/fastq/SRR1033875_*.fastq.gz):

>>> fastqc –o lung_cell_1/raw_sta lung_cell_1/fastq/SRR1033875_1.fastq.gz

>>> fastqc –o lung_cell_1/raw_sta lung_cell_1/fastq/SRR1033875_2.fastq.gz

3.2.3.Quality control of raw fastq files

>>> mkdir lung_cell_1/high-quality

For each raw fastq file (for example lung_cell_1/fastq/SRR1033875.fastq.gz):

>>>mkdir lung_cell_1/high-quality/SRR1033875

>>> java –jar /share/home/ccwei/tools/Trimmomatic-0.32/trimmomatic-0.32.jar PE -phred33 lung_cell_1/fastq/SRR1033875_1.fastq.gz lung_cell_1/fastq/SRR1033875_2.fastq.gz lung_cell_1/high-quality/SRR1033875/paired_1.fastq.gz lung_cell_1/high-quality/SRR1033875/single_1.fastq.gz lung_cell_1/high-quality/SRR1033875/paired_2.fastq.gz lung_cell_1/high-quality/SRR1033875/single_2.fastq.gz ILLUMINACLIP:/share/home/ccwei/tools/Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 HEADCROP:14 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:40

Note: you need six files for trimmonmatic program input: xxx_2.fastq.gz, xxx_1.fastq.gz paired _1.fastq.gz single_1.fastq.gz paired_2.fastq.gz and single_2.fastq.gz

3.2.4.Data statistics of high-quality data

```
>>> mkdir lung_cell_1/hq_sta
```

For each fastq file (SRR1033875):

```
>>> fastqc –o lung_cell_1/hq_sta lung_cell_1/high-quality/SRR1033875/paired_1.fastq.gz
>>> fastqc –o lung_cell_1/hq_sta lung_cell_1/high-quality/SRR1033875/paired_2.fastq.gz
>>> fastqc –o lung_cell_1/hq_sta lung_cell_1/high-quality/SRR1033875/single_1.fastq.gz
>>> fastqc –o lung_cell_1/hq_sta lung_cell_1/high-quality/SRR1033875/single_2.fastq.gz
```

3.2.5.Map reads to reference genome

5.1 Construct the Bowtie database for hg19.

```
##############################################
###  Note, don't do this step in the class.  ###
##############################################
>>>mkdir hg19_index
>>>bowtie-build /share/home/ccwei/pab/2014/week8/hg19_index/hg19.fa hg19_index/hg19
>>>bowtie-build /share/home/ccwei/pab/2014/week8/hg19_index/chr22.fa hg19_index/chr22
```

5.2 For each sample (SRR1033875):

```
>>>mkdir lung_cell_1/mapping
>>>mkdir lung_cell_2/mapping
>>>mkdir lung_cell_1/mapping/SRR1033875
>>>tophat2 --no-coverage-search -o lung_cell_1/mapping/SRR1033875 --bowtie1 –p 4 --segment-length 20 -G /share/home/ccwei/pab/2014/week8/hg19_index/protein_coding_chr22.gtf /share/home/ccwei/pab/2014/week8/hg19_index/chr22 lung_cell_1/high-quality/SRR1033875/paired_1.fastq.gz lung_cell_1/high-quality/SRR1033875/paired_2.fastq.gz,lung_cell_1/high-quality/SRR1033875/single_1.fastq.gz,lung_cell_1/high-quality/SRR1033875/single_2.fastq.gz
```

3.2.6.Compute expression levels

```
>>>mkdir lung_cell_1/exp
```

For each sample (SRR1033875):

```
>>>mkdir lung_cell_1/exp/SRR1033875
>>>cufflinks -o lung_cell_1/exp/SRR1033875 -G /share/home/ccwei/pab/2014/week8/hg19_index/protein_coding_chr22.gtf lung_cell_1/mapping/SRR1033875/accepted_hits.bam
```

3.2.7.Compute differential expression

>>>mkdir diff

>>>cuffdiff –o diff /share/home/ccwei/pab/2014/week8/hg19_index/protein_coding_chr22.gtf lung_cell_1/mapping/SRR1033875/accepted_hits.bam, lung_cell_1/mapping/SRR1033876/accepted_hits.bam, lung_cell_1/mapping/SRR1033878/accepted_hits.bam lung_cell_2/mapping/SRR1033884/accepted_hits.bam, lung_cell_2/mapping/SRR1033886/accepted_hits.bam, lung_cell_2/mapping/SRR1033888/accepted_hits.bam