Notes of RNA-seq

Xun Zhao

The Quality of Sequencing Result

`fastqc`

Analyze the reads and give out an overview.

Main code:

```
fastqc -o $out_dir *.fq.gz
```

Parameters:

- `-t $num`: use `$num` cores
- `-q`: quiet mode

Output:

- `.html`: the summary of the `.fq.gz` file
- `.zip`: the content in `.html` file

Notes:

In RNA-seq, we no not have to remove the duplication.

`fastx_trimmer`

To remove low quality sequences, namely, the head and tail of a read that have low quality value.

Main code:

```
zcat $fastq_1 | fastx_trimmer -f 11 -l 140 -z -o $out_fastq_1
```

Parameters:

- `zcat`: unzip
- `-f, -l`: keep sequence from … to …, cut off other parts
- `-z`: zip the output to `.gz` file

Notes:

This tool can be installed by `conda install fastx_toolkit`.

cutadapt

Cut adapters based on given adapter sequence and some parameters to determine whether a sequence is adapter (filter about the quality of alignment)

Main code:

```
nohup cutadapt --times 1 -e 0.1 -O 3 --quality-cutoff 6
    -m 50 -a AGATCGGAAGAGC -A AGATCGGAAGAGC -o $out_fastq_1
    -p $out_fastq_2 $fastq_1 $fastq_2 > $log_file 2>&1 &
```

Parameters:

- `--times 1`: only cut adapter once. That is, there is only one adapter at one end.
- `-e 0.1`: allowing 10% error rate when aligning the sequence to the reads
- `-O 3`: at least, there should be 3 matches in the alignment
- `--quality-cutoff 6`: common setting
- `-m 50`: if the sequence is shorter than 50bp after cutting, discard it
- `-a, -A`: given adapters
- `-o, -p`: output

bowtie2

Remove rRNA, using the alignment with rRNA-index to describe the quality of reads in RNA-seq experiment. In general, the rRNA should be less than 10%.

Main code:

```
nohup bowtie2 -x $rRNA_index -1  $fastq_1 -2 $fastq_2
    -S $sam_out -p 4 --un-conc-gz $fastq_unmap > $log 2>&1 &
```

Parameters:

- `-x`: input the rRNA-index
- `-1, -2`: input RNA reads
- `-S`: ???
- `-p`: number of cores
- `--un-conc-gz`: the unmapped sequences are what we want, namely, the reads that cannot align with rRNA

Align with Ref-Genome

`topaht2`

Considering the exons and introns, the RNA (mRNA) should be cut into fragments before they are mapped to genome.

Main code:

```
nohup tophat2 -p 8 -o $output_dir $hg19_index
    $fastq_1 $fastq_2 > $log 2>&1 &
```

Parameters:

- `-p`: number of cores
- `-o`: the directory of output
- `(after -o)`: reference genome, input file(s)

Outputs:

- `.bam`: for next step
- ...

Calculate Transcripts

cufflinks

We want to use some value to quantify the expression of a gene. So we use FPKM values, which is the number of mapped reads of one exon per length (kb) of this exon and per amount (M) of the total reads.

Main code:

```
nohup cufflinks -o $cufflinks_dir -p 4
    -G $hg19_gtf $bam_file > $log 2>&1 &
```

Parameters:

- -G: the reference genome

Outputs:

- ...: ???

Compare Difference between Groups

cuffdiff

Calculate the difference between different genes' expression level using a method like *t-test*. First, we estimate teh distribution of the gene, assuming the distribution is Guassian. Then use *t-test* to calculate the significance.

Main code:

```
treat_bam=${treat_1_bam},${treat_2_bam}
ctrl_bam=${ctrl_1_bam},${ctrl_2_bam}
label=hela_ctrl,hela_treat
nohup cuffdiff -o $out_dir -p 8 --labels $label
    --min-reps-for-js-test 2 $hg19_gtf
    $ctrl_bam $treat_bam > $log 2>&1 &
```

Parameters:

- `--labels`: the order of input files
- `--min-reps-for-js-test`: the replication of the experiment