RNA-seq Data Analysis

Qi Sun
Bioinformatics Facility
Biotechnology Resource Center
Cornell University

• Lecture 1. RNA-seq read alignment

• Lecture 2. Quantification, normalization & differentially expressed gene detection

• Lecture 3. Clustering; Function/Pathway Enrichment analysis
RNA-seq Experiment

mRNA

cDNA Fragments (100 to 500 bp)

Sequencing the end(s) of cDNA fragments
Some experimental aspects relevant to data analysis

**Single End**

**Paired End**

**Stranded**

**Unstranded**
Some experimental aspects relevant to data analysis

Long sequence reads

50 bp

150 bp

Adapter
Experimental design with good reference genome

• **Read length**
  50 to 100 bp

• **Paired vs single ends**
  Single end

• **Number of reads**
  >5 million per sample

• **Replicates**
  3 replicates
RNA-seq Experiments with NO reference genome

• Longer reads (150 bp or longer)

• Paired-end & stranded

• More reads (pooled from multiple samples)
Limitation of RNA-seq 1. Sequencing bias

mRNA

```
AAAAA
AAAAA
AAAAA
```

Not random

cDNA Fragments

```

```

Reads

```

```
Limitation of RNA-seq 1. Sequencing bias

- There are sequencing bias in RNA-seq;
- RNA-seq is for comparing same gene across different samples;
RNA-seq Data Analysis

Step 1. Map reads to gene

Step 2. Count reads per gene, estimate the transcript abundance
RNA-seq Data Analysis

Ambiguous reads placements

1. Between paralogous genes;
2. Between splicing isoforms;

Key:
- Coding sequence
- Introns
- Exons
- Splice junctions
Read-depth are not even across the same gene
Data analysis procedures
Step 1. Quality Control (QC) using FASTQC Software

1. Sequencing quality score

![Quality scores across all bases (Sanger/Illumina 1.9 encoding)](chart)
Diagnose low quality data

1. Low quality reads & reads with adapters
   • Trimming tools (FASTX, Trimmomatic, et al.)

2. Contamination (BLAST against Genbank)
   • Tool in bioHPC: fastq_species_detector

3. Correlation of biological replicates
   • MDS plot
Step 2. Map reads to genome using TOPHAT Software

- Alignment of genomic sequencing vs RNA-seq

About the files

1. Reference genome (FASTA)

>chr1
TTCTAGGTCTGCGATATTTTCTGCTATCCATTTTTGTTAACTCTTTCAATG
CATTCCACAAATACCTAAGTATTTTTAATAATGDTGTTTTTTTTTTTTT
TTTGCACTATGAAGTTTTTTCAAAATTCTTTTTAAGTGCAAAAACCTTGT
CATGTGATCGCTCAATATTTTCTAGTCGACAGCAGCTGTTCGTCAACATG
AAACCGTGCACTCCAGGAAAAAGCAGACACAGACCGGCACTCCCTTTG4GACC
CGGTTTTATACCTTTGAACTGCTCGAGGCCCTCTCAGAACCGTCTCC
CACACCCCCGCTCCAGGCTCCTCTCCGAGTACGCTGCTGAGCC
CCGGGAACCCACGCGGCGTACAGAAGATTGCGGTCTTCCCTACGAGGAGCA
GGAAGCTCCCCGGCACCAGCTGCAGGAAGACCCCGCAGGCTTCCAG
AACCAGCGACCAGCGGCGAAGACGACAGAGTTGCGGAGGCGGAACCAGGACC
CCGGAGACCCCCCCGCAGCTCTGGCGCCCAACAGCCTTGGCTCCCTCTGAGC
GCCCCAGAGCCGCCTGACAGGGCGCGCTCCAGGAGAAGCTCCGGGG
CGACCCCAAGAGCCCTCCCGGGGCGTCGGGCCCACGCGCGCGGGCTGCC
GGAGACGCCACGGCCAGCGCGCCCGAACCAGCAGCTTTGGGCAAGGC
TTCTCGTCAGAGAACGCTCCCGGGCCTCCCGGCCGCTCCTCCAGCAGCC
TCCGGGTCCCTACTTCTGGGCCCAGCAAGCCAGCCTACTCCCTCCGCC
GGCCCGAGACGCCCTCCTACCTCGAGACAGGCGCCCTCCGGAAGCTCCGGCC
GCCGTTCTCGCTTGCGGACGGCTGGCTTCTAGGCCCGGCGCCCGCCAG
TCCGGCGCCGCTCTGGGTCTTAACGCGGCGCCTGCAGCTCAGCC
CTCCGGGAGCGGCTCCAGGACCCCGGTCAGCGCCAGAGCGCTGCTGC
TGGCCCGAGTCGCGGCGCCTGGCCAGAAGCTACGCTCAGCTCCAGGCTCC
GACAGTCACAGGCTCCCACTGCGGTTGGTCTCAGCGGCTAGCGGCCGCCCC
ATACCCCGTGCTTTCTGCTCTGACGCTAGCCCGCCCTCCTTACTGACCTCCCTGCTCTTTGT
About the files

1. FASTA

2. RNA-seq data (FASTQ)

3. GFF3/GTF

4. SAM/BAM
About the files

1. FASTA

2. RNA-seq data
   (FASTQ)

3. GFF3/GTF

4. SAM/BAM

---

Single-end: one file per sample
Paired-end: two files per sample
About the files

1. FASTA

2. FASTQ

3. Annotation (GFF3/GTF)

4. SAM/BAM
About the files

1. FASTA

```
HWUSI-EAS525_0042_FC:6:23:10200:18582#0/1       16      1       10      40      35M
*       0       0       AGCCAAAGATTGCATCACGTTCTGCTGCTATTTCCT
agafgaffcfdf[fdccgggccccffagggg MD:Z:35 NH:i:1 HI:i:1 NM:i:0 SM:i:40
XQ:i:40 X2:i:0
```

2. FASTQ

```
HWUSI-EAS525_0042_FC:3:28:18734:20197#0/1       16      1       10      40      35M
*       0       0       AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCCT
hghghghhhhhhhhhhhhhhhhhghhhfhhhh MD:Z:35 NH:i:1 HI:i:1 NM:i:0
SM:i:40 XQ:i:40 X2:i:0
```

3. GFF3/GTF

```
HWUSI-EAS525_0042_FC:3:94:1587:14299#0/1       16      1       10      40      35M
*       0       0       AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCCT
hfhghhhhhhhhhhhhhhhhhhhhhhhhhh MD:Z:35 NH:i:1 HI:i:1 NM:i:0
SM:i:40 XQ:i:40 X2:i:0
```

4. Alignment (SAM/BAM)

```
D3B4KKQ1:227:D0NE9ACXX:3:1305:14212:73591       0       1       11      40      51M
*       0       0       GCCAAAGATTGCATCAGTTCTGCTGCTATTTCCTCCTACATTCTCTCTG
CCCFFGGFHHJGIHHJJJGJJGJJJGJJJGJJJGJJJGJJJGJI
MD:Z:51 NH:i:1 HI:i:1 NM:i:0
SM:i:0 SM:i:40 XQ:i:40 X2:i:0
```

```
HWUSI-EAS525_0038_FC:5:35:11725:5663#0/1       16      1       11      40      35M
*       0       0       GCCAAAGATTGCATCAGTTCTGCTGCTATTTCCTC
hhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehhehheh
MD:Z:35 NH:i:1 HI:i:1 NM:i:0
SM:i:40 XQ:i:40 X2:i:0
```
Running TOPHAT

• **Required files**
  - Reference genome. (FASTA file indexed with **bowtie2-build** software)
  - RNA-seq data files. (FASTQ files)

• **Optional files**
  - Annotation file (GFF3 or GTF)
    * If not provided, TOPHAT will try to predict splicing sites;
Running TOPHAT

tophat -G myAnnot.gff3 myGenome myData.fastq.gz

Some extra parameters

• **--no-novel**: only using splicing sites in gff/gtf file
• **-N**: mismatches per read (default: 2)
• **-g**: max number of multi-hits (default: 20)
• **-p**: number of CPU cores (BioHPC lab general: 8)
• **-o**: output directory

What you get from TOPHAT

• A BAM file per sample
  File name: accepted_hits.bam

• Alignment statistics
  File name: align_summary.txt

Input: 9230201
Mapped: 7991618 (86.6% of input)
of these: 1772635 (22.2%) have multiple alignments (2210 have >20)
86.6% overall read alignment rate.
Visualizing BAM files with IGV

* Before using IGV, the BAM files need to be indexed with “samtools index”, which creates a .bai file.
Exercise 1

• Run TOPHAT to align RNA-seq reads to genome;

• Visualize TOPHAT results with IGV;

• Learn to use Linux shell script to create a pipeline