How to align your RNA-seq data?

6e école de bioinformatique AVIESAN-IFB

Rachel Legendre

15/11/2017
RNA-seq applications

- Align reads to genome
- Assemble transcripts de novo
- Assemble transcripts from spliced alignments
- Estimate abundance
- More abundant
- Less abundant

Count k-mers

Transcriptome

RNA-seq reads

Genome
### Ask right question before libraries preparation and sequencing:

<table>
<thead>
<tr>
<th>Prokaryote</th>
<th>Eukaryote</th>
</tr>
</thead>
</table>
| ✓ I don't find a ribo-depletion kit for my organism:  
  → Design yourself your oligo | ✓ I want coding RNA only:  
  → PolyA strategy |
| ✓ I want to identify anti-sens RNA:  
  → Directional RNA-seq | ✓ I want non-coding RNA also:  
  → Ribo depletion |
| ✓ I'm interested by transposons:  
  → Longer read sequencing  
  → Paired-end sequencing | ✓ I'm interested by miRNA profiling:  
  → specific protocol |
| ✓ I want to identify isoformes?  
  → paired-end sequencing | |
RNA-seq: Why ? How ?

Regardless of your organism:

✔ Complexity of your genome and the **biological question**
  ➔ paired-end or single-end, length of reads ?

✔ **Sequencing depth** (multiplexing rate)

✔ More **biological replicates** than more sequencing depth

✔ **Stranded RNA-seq** protocol to assigned reads to a particular strand

For a successful experiment, it's imperative to include bioinformaticians and biostatisticians **BEFORE** the beginning of the RNA extraction
Prerequisites

RNA sample:
- DNAse treatment
- Quantity (adapted protocol)
- Quality (RNA integrity number > 7)
- Stocked at -80°C

Reference Genome:
The complete genomic sequence in fasta format

Annotation file:
All features (genes, CDS, intron, UTR) of genome in GFF format

FASTA file:

> name sequence
ATGCTGATCTCCGC
ATCGTGCCATATCG
GCTAGCGTAGCTGA
ATCGTAGTCGACTG
CTAGCTGCTGTATC

GFF file:

# gff format
chr1 UTR 10 20
chr1 gene 10 90
chr1 mRNA 10 90
chr1 mRNA 10 90
chr1 CDS 10 35
chr1 CDS 52 90
Data processing

- Quality assessment
- Reads mapping
- Reads counting
- Normalization and differential gene expression
RNA-seq mapping specificity

- Mapping on genome or transcriptome?
  - the transcriptome is currently not well characterised enough to serve as a suitable reference for RNA-Seq
  - mapping to a genome is more objective and repeatable
  - get more gene isoforms information through mapping it to the genome

- Take account to reads that come from exon-exon junctions

SAM format

**Header:** information about genome and command line

```
@HD VN:1.0 SO:coordinate
@SQ SN:chrXV LN:1091291
@SQ SN:chrXVI LN:948066
@PG ID:TopHat VN:2.1.0 CL:tophat --num-threads 1 genome
/w/galaxy/galaxy3/database/files/001/dataset_1385.dat
```

**Read**

```
HWI-D00395:147:HMV2HADXX:1:1107:7476:20117 16 chrI 4022 255 51M * 0 0
GTTTGAATATTTATGTAGTGTCAACATCAAATGTGTCTATTTTGTGATGAGG
CCCFFFFFGHHGHJHIIJCHIJJIJJJJIJJJJJJJJJJJJIIIIIIJJJI
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0
MD:Z:51 YT:Z:UU NH:i:1
```

**Mandatory fields**

```
HWI-D00395:147:HMV2HADXX:1:1107:7476:20117 16 chrI 4022 255 51M * 0 0
```

**Read sequence**

```
GTTTGAATATTTATGTAGTGTCAACATCAAATGTGTCTATTTTGTGATGAGG
```

**Read quality**

```
CCCFFFFFGHHGHJHIIJCHIJJIJJJJIJJJJJJJJJJJJIIIIIIJJJI
```

**Optionnal fields**

```
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:51 YT:Z:UU NH:i:1
```

How the read/mate is mapped?
Mapping statistiques

SAMtools Flagstat

168 375 631 + 0 in total (QC-passed reads + QC-failed reads)
133 873 423 + 0 duplicates
143 893 516 + 0 mapped (85.46%:-nan%)
168 375 631 + 0 paired in sequencing
84 186 587 + 0 read1
84 189 044 + 0 read2
143 722 660 + 0 properly paired (85.36%:-nan%)
143 722 660 + 0 with itself and mate mapped
170 856 + 0 singletons (0.10%:-nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Properly paired reads =>
• Insert size ≤ max_insert_size
• R1 / R2 mapped on same chromosome
• R1 ===>        <=== R2
Data used: organism

**Organism**: Arabidopsis thaliana, the plant model

**Genome**: first release in 2000, now v.10

**Genome and annotation file**:

- TAIR10_chr4.fa
- TAIR10_chr4.gtf
Data used: conditions

2 conditions: WT (Columbia, Col-0) vs KO AtPRMT5

AtPRMT5 (AT4G31120) : type II protein arginine methyltransferase (methylates proteins, histones, RNA splicing factors)

What are the genes impacted by the AtPRMT5 deletion ?
Could help the characterization of the function of the protein AtPRMT5 during de novo shoot regeneration in Arabidopsis.
Data used: Experiment

- 3 biological replicates
- "TruSeq Stranded mRNA Library Prep Kit (Illumina, catalog number RS-122-2101)"
- paired-end, 125 bp, 30 $10^6$ reads / sample

<table>
<thead>
<tr>
<th>WT (Col)</th>
<th>KO APRTM5</th>
<th>2 files *.fastq.gz</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERR1592576</td>
<td>ERR1592579</td>
<td>WT1_1 WT1_2 KO1_1 KO1_2</td>
</tr>
<tr>
<td>ERR1592577</td>
<td>ERR1592580</td>
<td>WT2_1 WT2_2 KO2_1 KO2_2</td>
</tr>
<tr>
<td>ERR1592578</td>
<td>ERR1592581</td>
<td>WT3_1 WT3_2 KO3_1 KO3_2</td>
</tr>
</tbody>
</table>
Practice: mapping with STAR

[rlegendre@nz ~] qlogin -pe thread 4
[rlegendre@nz ~] source $CONDA3/activate eba2017_rnaseq_ref
[rlegendre@nz ~] cdprojet
[rlegendre@nz ~] mkdir TP_mapping
[rlegendre@nz ~] cd TP_mapping
[rlegendre@nz ~] cp /projet/sbr/ggb/RNAseq/TP_mapping/run_star.sh .
[rlegendre@nz ~] sh run_star.sh
Practice: Explain STAR quantMode values

STAR outputs read counts per gene into ReadsPerGene.out.tab file with 4 columns which correspond to different strandedness options:

**column 1:** gene ID

**column 2:** counts for unstranded RNA-seq

**column 3:** counts for the 1st read strand aligned with RNA (htseq-count option -s yes)

**column 4:** counts for the 2nd read strand aligned with RNA (htseq-count option -s reverse)

Select the output according to the strandedness of your data.
Practice: get the BAM file

[rlegendre@nz ~] pwd
[rlegendre@nz ~] qdel
[rlegendre@nz ~] logout
[rlegendre@nz ~] scp rlegendre@ssh.sbrRoscoff.fr:/projet/externe/univ/rlegendre/TP_mapping/*bam* .

Or use CyberDuck as previous day

Or download this archive : https://goo.gl/2rgqdr
Visualize alignments

Go to AT4G31120