Differential gene expression analysis using RNA-seq

Applied Bioinformatics Core, March 2018

Friederike Dündar with Luce Skrabanek & Paul Zumbo
Day 1: Introduction into high-throughput sequencing [many general concepts!]

1. RNA isolation & library preparation
2. Illumina’s sequencing by synthesis
3. raw sequencing reads
   • download
   • quality control
4. experimental design
RNA-seq is popular, but still developing.

“RNA-seq is not a mature technology. It is undergoing rapid evolution of biochemistry of sample preparation; of sequencing platforms; of computational pipelines; and of subsequent analysis methods that include statistical treatments and transcript model building.”

ENCODE consortium

"Analysis paralysis"

- basically no generally accepted standard reference (tx definitions often change quarterly)
- myriad tools → highly complex & specialized “pipelines”

“The (...) flexibility and seemingly infinite set of options (...) have hindered its path to the clinic. (...) The fixed nature of probe sets with microarrays or qRT-PCR offer an accelerated path (...) without the lure of the latest and newest analysis methods.”

Byron et al., 2016
What to expect from the class

**Sample type & quality**
- Sequencing
  - Read length
  - PE vs. SR
  - Sequencing errors

**Experimental design**
- Controls
- No. of replicates
- Randomization

**Library preparation**
- Poly-A enrichment vs. ribo minus
- Strand information

**Biological question**
- Expression quantification
- Alternative splicing
- De novo assembly needed
- mRNAs, small RNAs
- ....

**Bioinformatics**
- Aligner
- Normalization
- DE analysis strategy

**NOT COVERED:**
- novel transcript discovery
- transcriptome assembly
- alternative splicing analysis

(see the course notes for references to useful reviews)
RNA-seq workflow overview

Total RNA extraction

- mRNA enrichment
- fragmentation
- cDNA library

Sequencing

- cluster generation
- sequencing by synthesis
- image acquisition

Bioinformatics
Quality control of RNA extraction

RIN = 28S:18S ratio
avoid degraded RNA junk

Use the expertise of the sequencing facility staff!
They’ve seen it all!

Figure from Griffith et al., PLoS Comp Biol (2015). doi: 10.1371/journal.pcbi.1004393
RNA-seq library preparation

RNA extraction

mRNA enrichment

fragmentation (~200 bp)

cDNA synthesis

cDNA plus sequencing adapters

rRNA depletion/mRNA enrichment

fragmentation

random priming and reverse transcription

second strand synthesis

end repair, A-addition, adapter ligation

PCR

classical Illumina protocol (unstranded)

Influence of the RNA enrichment strategy

which transcripts are you interested in?

what type of noise can you tolerate?

- Total RNA
- rRNA depletion
- mRNA selection
- cDNA capture

Figure from Griffith et al., PLoS Comp Biol (2015). doi: 10.1371/journal.pcbi.1004393
RNA-seq library preparation: pick one!

RNA extraction

mRNA enrichment

fragmentation (~200 bp)

cDNA synthesis

cDNA plus sequencing adapters

rRNA depletion/mRNA enrichment

fragmentation

random priming and reverse transcription

second strand synthesis

end repair, A-addition, adapter ligation

PCR

classical Illumina proto (unstranded)

QC!
Size selection

Size selection or exclusion (e.g. PAGE, SPRI magnetics beads, etc.)

Small RNAs lost in both cases

cDNA

more efficient sequencing

Figure modified from Griffith et al., PLoS Comp Biol (2015). doi: 10.1371/journal.pcbi.1004393
RNA-seq workflow overview

Total RNA extraction

RNA

fragments

cDNA with adapters

Sequencing

flowcell with primers

http://informatics.fas.harvard.edu/test-tutorial-page/
Cluster generation

bridge amplification

denaturation

cluster generation
removal of complementary strands \(\rightarrow\) identical fragment copies remain

http://informatics.fas.harvard.edu/test-tutorial-page/
Sequencing by synthesis

1. extend 1\textsuperscript{st} base
2. read
3. deblock

repeat for 50 – 100 bp

generate base calls
Typical biases of Illumina sequencing

- Sequencing errors
- Miscalled bases
- **PCR artifacts (library preparation)**
  - Duplicates (due to low amounts of starting material)
  - Length bias
  - GC bias

---

**Figure from Love et al. (2016). Nat Biotech, 34(12). More details & refs in course notes (esp. Table 6).**
General sources of biases
(not inherently sample-specific)

• issues with the reference
  • CNV
  • mappability

• inappropriate data processing
RAW SEQUENCING READS

Let the data wrangling begin!
RNA-seq workflow overview

Total RNA extraction
- mRNA enrichment
- fragmentation
- cDNA library

Sequencing
- cluster generation
- sequencing by synthesis
- image acquisition

Bioinformatics
Bioinformatics workflow of RNA-seq analysis

**FASTQC**

- **Raw reads**
  - `.fastq`

- **Aligned reads**
  - `.sam`, `.bam`

- **Read count table**
  - `.txt`

- **Normalized read count table**
  - `.Robj`

- **List of fold changes & statistical values**
  - `.Robj`, `.txt`

- **Downstream analyses on DE genes**

**Base calling & demultiplexing**
- Bustard/RTA/OLB, CASAVA

**Mapping**
- STAR

**Counting**
- featureCounts

**Normalizing**
- DESeq2, edgeR

**DE test & multiple testing correction**
- DESeq2, edgeR, limma

**Filtering**
- Customized scripts
Where are all the reads?

GenBank

DDBJ
http://www.ddbj.nig.ac.jp/intro-e.html

ENA
https://www.ebi.ac.uk/ena/

The SRA is the main repository for publicly available DNA and RNA sequencing data of which three instances are maintained world-wide.
Let’s download!

- We will work with a data set submitted by Gierliński et al.

- they deposited the sequence files with SRA – we will retrieve it via ENA (https://www.ebi.ac.uk/ena/)

- accession number: ERP004763

Course notes @ https://chagall.med.cornell.edu/RNASEQcourse/ of @ http://www.trii.org/courses/rnaseq.html

See Section 2 (Raw Data) for download instructions etc.
Downloading a batch of fastq files

https://www.ebi.ac.uk/ena/ → study ERP004763

1. get link with list of **ftp sites** for every file: right-click on "TEXT" → "copy link location"

2. **download** on server/via CL: copy and paste to `wget` (mind the quotation marks to keep the link intact!):
   
   ```
   wget -O samples_at_ENA.txt "<LINK>"
   ```

get the **sample information**:

   ```
   wget -O ERP004763_sample_mapping.tsv --no-check-certificate "https://ndownloader.figshare.com/files/2194841"
   ```

```
list of links

<table>
<thead>
<tr>
<th>RunAccession</th>
<th>Lane</th>
<th>Sample</th>
<th>BiolRep</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERR458493</td>
<td>1</td>
<td>WT</td>
<td>1</td>
</tr>
<tr>
<td>ERR458494</td>
<td>2</td>
<td>WT</td>
<td>1</td>
</tr>
<tr>
<td>ERR458495</td>
<td>3</td>
<td>WT</td>
<td>1</td>
</tr>
<tr>
<td>ERR458496</td>
<td>4</td>
<td>WT</td>
<td>1</td>
</tr>
<tr>
<td>ERR458497</td>
<td>5</td>
<td>WT</td>
<td>1</td>
</tr>
<tr>
<td>ERR458498</td>
<td>6</td>
<td>WT</td>
<td>1</td>
</tr>
<tr>
<td>ERR458499</td>
<td>7</td>
<td>WT</td>
<td>1</td>
</tr>
<tr>
<td>ERR458500</td>
<td>1</td>
<td>SNF2</td>
<td>1</td>
</tr>
<tr>
<td>ERR458501</td>
<td>2</td>
<td>SNF2</td>
<td>1</td>
</tr>
</tbody>
</table>
```
1. find out which RunAccession numbers belong to the WT and SNF2 samples of BiolRep #1

```bash
awk ' $4 == 1 {print $0}' ERP004763_sample_mapping.tsv
```

2. download individual sample

```bash
awk -F "\t" ' $5 == "ERR458493" {print $11}' samples-overview.txt | xargs wget
```

3. either do this 6 more times individually or write a for-loop

```bash
for i in `seq 3 9`
do
  SAMPLE=ERR45849$i
  egrep ${SAMPLE} samples_at_ENA.txt | cut -f11 | xargs wget
done
```

4. for-loop for SNF2 samples

```bash
for i in `seq 0 6`
do
  SAMPLE=ERR45850$i
  egrep ${SAMPLE} samples_at_ENA.txt | cut -f11 | xargs wget
done
```

5. sort reads into folders

```bash
$ mkdir raw_reads
$ mkdir WT_1
$ mkdir SNF2_1
$ mv ERR45849.gz WT_1/
$ mv ERR4585.gz SNF2_1/
```
FASTQ file format

= FASTA + quality scores

1 read ⇔ 4 lines!

1. @Read ID and sequencing run information
2. sequence
3. + (additional description possible)
4. quality scores
Base quality score

@ERR459145.1 DHKW5DQ1:219:DOPT7ACXX:2:1101:1590:2149/1
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAGAGACGGGTTCAGC
+
@7<DBADDDBH?DHHI@DH>HHHEGHIIIGGIFFGIBFAAGAFHA ’5?B@D

<table>
<thead>
<tr>
<th>DEC</th>
<th>OCT</th>
<th>HEX</th>
<th>BIN</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>040</td>
<td>20</td>
<td>00100000</td>
<td>!</td>
</tr>
<tr>
<td>33</td>
<td>041</td>
<td>21</td>
<td>00100001</td>
<td>&quot;</td>
</tr>
<tr>
<td>34</td>
<td>042</td>
<td>22</td>
<td>00100010</td>
<td>#</td>
</tr>
<tr>
<td>35</td>
<td>043</td>
<td>23</td>
<td>00100011</td>
<td>$</td>
</tr>
<tr>
<td>36</td>
<td>044</td>
<td>24</td>
<td>00100100</td>
<td>%</td>
</tr>
<tr>
<td>37</td>
<td>045</td>
<td>25</td>
<td>00100101</td>
<td>&amp;</td>
</tr>
<tr>
<td>38</td>
<td>046</td>
<td>26</td>
<td>00100110</td>
<td>’</td>
</tr>
<tr>
<td>39</td>
<td>047</td>
<td>27</td>
<td>00100111</td>
<td>(</td>
</tr>
<tr>
<td>40</td>
<td>050</td>
<td>28</td>
<td>00101000</td>
<td>)</td>
</tr>
<tr>
<td>41</td>
<td>051</td>
<td>29</td>
<td>00101001</td>
<td>*</td>
</tr>
<tr>
<td>42</td>
<td>052</td>
<td>2A</td>
<td>00101010</td>
<td>+</td>
</tr>
<tr>
<td>43</td>
<td>053</td>
<td>2B</td>
<td>00101011</td>
<td></td>
</tr>
</tbody>
</table>

base error probability $p$, e.g. $10e^{-4}$

$-10 \times \log_{10}(p)$

Phred score, e.g.: 40

turn score into ASCII symbol

“FASTQ score”, e.g.: ( 

http://www.ascii-code.com/
Base quality scores

• each base has a certain error probability \( p \)
• Phred score = \(-10 \times \log_{10}(p)\)
• Phred scores are ASCII-encoded, e.g., “!” COULD represent Phred score 33

| L | -Sanger        Phred+33,  raw reads typically (0, 40) |
| X | -Solexa        Solexa+64, raw reads typically (-5, 40) |
| I | -Illumina 1.3+ Phred+64,  raw reads typically (0, 40) |
| J | -Illumina 1.5+ Phred+64,  raw reads typically (3, 40) |
| L | -Illumina 1.8+ Phred+33,  raw reads typically (0, 41) |

For SOLiD data, the sequence is in color space, except the first position. The quality values are those of the Sanger format. Alignment tools differ in their preferred version of the quality values: some include a quality score (set to 0, i.e. '!') for the leading nucleotide, others do not. The sequence read archive includes this quality score.

For raw reads, the range of scores will depend on the technology and the base caller used, but will typically be up to 41 for recent Illumina chemistry. Since the maximum observed quality score was previously only 40, various scripts and tools break when they encounter data with quality values larger than 40. For processed reads, scores may be even higher. For example, quality values of 45 are observed in reads from Illumina’s Long Read Sequencing Service (previously Moleculo).

also see Table 2 in the course notes
Quality control of raw reads: FastQC

http://www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

$ ~/mat/software/FastQC/fastqc
$ ~/mat/software/anaconda2/bin/multiqc