RNA-sequencing

Next Generation sequencing analysis 2016

Anne-Mette Bjerregaard
Center for biological sequence analysis (CBS)
Terms and definitions

TRANSCRIPTOME

The full set of RNA transcripts and their associated abundance in a sample

RNA-seq

High-throughput sequencing technology used for probing the transcriptome of a sample

EXPRESSION
Data generation
RNA-seq Protocol

Library preparation

Differs primarily from the standard sequencing protocol by an additional reverse transcription step (RNA->cDNA)

- ① mRNA or total RNA
- ② Remove contaminant DNA
- ③ Fragment RNA
- ④ Reverse transcribe into cDNA
- ⑤ Ligate sequence adaptors
- ⑥ Select a range of sizes
- ⑦ Sequence cDNA ends

Poly-A selection

Strand-specific RNA-seq?

PCR amplification?

Poly-A

Martin et al., Nature Genetics Review (2011)
RNA-seq: Strand-specific library preparation

The standard library preparation protocol do not preserve information about which strand was originally transcribed.

Strand-specific sequencing (e.g. dUTP method)

Martin et al., Nature Genetics Review (2011)
More terms

Gene

Transcripts

Wiki: https://en.wikipedia.org/wiki/Alternative_splicing
Data Analysis
Data analysis

1. Raw reads
2. Remove artefacts
3. Correct errors (optional)
4. Assemble into transcripts
5. Post-process transcripts
6. Align reads to transcripts to quantify expression

FastQC
CutAdapt
Trimming

Martin et al., Nature Genetics Review (2011)
Assemble into transcripts

Reads

Transcripts
Transcriptome assembly strategies

- Reference-based
- *De novo*
- Combined
- Pseudoalignment
Reference-based assembly

Align reads

Graph

Traverse graph

Assemble
Alignment tools

• NGS common alignment program:
  – BWA
  – Bowtie (Bowtie2)
  – Novoalign

• Take into account splice-junction
  – Tophat / Cufflinks
De novo assembly

Kmers

De Bruijn graph

Collapse

Traverse graph

Assemble
De novo assembly tools

- Velvet
  - Genomic and transcriptomic

- Trinity
  - Transcriptomic

- Cufflinks
  - Transcriptominc, reassemble pre-aligned transcripts to find alternative splicing based on differential expression
Combined assembly

- RNA-seq reads
  - Align-then-assemble
    - Align reads to the genome
    - Reference-based assembly of aligned reads
      - Unaligned reads
        - De novo assembly
          - Comprehensive assembly
      - De novo assembly
  - OR
  - Assemble-then-align
    - De novo assembly
      - Scaffold contigs
        - Unassembled reads
          - Extend contigs
Pseudoalignment - Kallisto

N. Bray et al., Nature Biotechnology (2016)
## Advantages and disadvantages

<table>
<thead>
<tr>
<th></th>
<th>Novel transcripts</th>
<th>Existing reference</th>
<th>Fast</th>
<th>Trans-spliced genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference-based</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>De Novo</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Pseudo</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>
Expression analysis
Expression analysis

Transcripts

Normalized expression
Within sample normalization

Compare expression levels of different transcripts / genes within the same sample.
Within sample normalization

- RPKM (Reads Per Kilobase Million)
  - Single end reads
- FPKM (Fragments Per Kilobase Million)
  - Paired end reads
- TPM (Transcripts Per Million)

1. Sequencing depth (million)
2. Transcript length (kilobase)

1. Transcript length
2. Sequencing depth

<table>
<thead>
<tr>
<th>RPKM / FPKM</th>
<th>TPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
**Between sample normalization (BSN)**

Compare expression levels of the same transcript between samples
HTSeq-count

Simply counts the number of reads that map to each feature

- Feature: Can be a gene or transcript annotation
- Overlapping features: Union or intersection
Statistics

Poisson and neg. binomial parameter

Additional negative binomial parameter. when overdispersion = 0 
eg. binom = Poisson

mean = 5.81 overdispersion = 1.397

http://www.ats.ucla.edu/stat/stata/seminars/count_presentation/count.htm
DeSeq2

Differential gene expression analysis based on the negative binomial distribution

- **Input:** Read count tables (HTSeq)
- **Output:** Table containing statistics for whether a gene is differentially expressed between two conditions

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 6 rows and 6 columns
##
##   baseMean  log2FoldChange  lfcSE   stat  pvalue  padj
##   <numeric>       <numeric> <numeric> <numeric> <numeric> <numeric>
##
## # FBgn0039155  453       -3.71    0.160   -23.2  4.01e-119  3.11e-115
## # FBgn0029167  2165      -2.08    0.104   -20.1  6.68e-90   2.59e-86
## # FBgn0035085  367       -2.23    0.137   -16.3  1.89e-59   4.87e-56
## # FBgn0029896  258       -2.21    0.159   -13.9  5.85e-44   1.13e-40
## # FBgn0034736  118       -2.57    0.185   -13.9  8.07e-44   1.25e-40
## # FBgn0040091  611       -1.43    0.120   -11.9  1.11e-32   1.44e-29
```

<table>
<thead>
<tr>
<th>Gene id</th>
<th>Mean read count</th>
<th>Log2 fold change and standard error</th>
<th>Test statistic</th>
<th>P-value and adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBgn0039155</td>
<td>453</td>
<td>-3.71 0.160</td>
<td>-23.2</td>
<td>4.01e-119 3.11e-115</td>
</tr>
<tr>
<td>FBgn0029167</td>
<td>2165</td>
<td>-2.08 0.104</td>
<td>-20.1</td>
<td>6.68e-90 2.59e-86</td>
</tr>
<tr>
<td>FBgn0035085</td>
<td>367</td>
<td>-2.23 0.137</td>
<td>-16.3</td>
<td>1.89e-59 4.87e-56</td>
</tr>
<tr>
<td>FBgn0029896</td>
<td>258</td>
<td>-2.21 0.159</td>
<td>-13.9</td>
<td>5.85e-44 1.13e-40</td>
</tr>
<tr>
<td>FBgn0034736</td>
<td>118</td>
<td>-2.57 0.185</td>
<td>-13.9</td>
<td>8.07e-44 1.25e-40</td>
</tr>
<tr>
<td>FBgn0040091</td>
<td>611</td>
<td>-1.43 0.120</td>
<td>-11.9</td>
<td>1.11e-32 1.44e-29</td>
</tr>
</tbody>
</table>
Thank you!
Within sample normalization – more details?


2. [https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/](https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/)
Between sample normalization – more details?

https://haroldpimentel.wordpress.com/2014/12/08/in-rna-seq-2-2-between-sample-normalization/