RNA-SEQ

Francesca Bertolini
DTU-Bioinformatics

08-06-2018
SUMMARY

1. Definition
2. List of major applications
3. RNA classifications
4. Sample collections and RNA integrity
5. Library preparation
6. Data analyses
   – Long reads alignment[assembly
   – Read count
   – Normalization
   – Differential expression
   – Gene enrichment
   – Short RNA
1. TRANSCRIPTOMICS

• Complete set of transcripts in a cell and their quantity, for a specific developmental stage or a physiological condition.

• Accurate profiling depends massively to the quality of the annotation of the reference genome.
2. RNA-seq applications

- Abundance estimation/differential expression
- Alternative splicing
- RNA editing
- Novel transcripts
- Allele specific expression
- Fusion transcripts
2. RNA-seq applications

- Abundance estimation/differential expression
- Alternative splicing
- RNA editing
- Novel transcripts
- Allele specific expression
- Fusion transcripts
3. RNA classification

• Ribosomal RNA (rRNA): catalytic component of ribosomes (about 80-85%)
• Transfer RNA (tRNA): transfers amino acids to polypeptide chain at the ribosomal site of protein synthesis (about 15%)
• Coding RNA (mRNA): carries information about a protein sequence to the ribosomes (about 5%)
• Other Non coding regulatory RNAs
3. RNA classification

Other non coding regulatory RNAs

<table>
<thead>
<tr>
<th>Class</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>MicroRNAs</td>
<td>Protein translation inhibition (2)</td>
</tr>
<tr>
<td>Piwi RNAs</td>
<td>Repression of transposable element expression (6)</td>
</tr>
<tr>
<td>Small nuclear RNAs</td>
<td>mRNA translation (13)</td>
</tr>
<tr>
<td>Small nucleolar RNAs</td>
<td>rRNA processing (through snoRNP action) (5)</td>
</tr>
<tr>
<td>Natural antisense transcripts</td>
<td>Regulation of transcript sense (21, 24, 25)</td>
</tr>
<tr>
<td>Transcribed ultra-conserved regions</td>
<td>Poorly described (26, 29)</td>
</tr>
<tr>
<td>Long intergenic ncRNAs</td>
<td>Guide / Protein scaffolding (HOTAIR) (31, 32)</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>Regulation of mRNA stability (34, 35)</td>
</tr>
<tr>
<td>Circular RNAs</td>
<td>RNA binding protein and miRNA decoy (37)</td>
</tr>
</tbody>
</table>

Delpu et al. 2016. *Drug Discovery in Cancer Epigenetics*
3. RNA classification

Long RNAs: splicing

DNA

RNA

mRNA

Translation

Protein A

Translation

Protein B

IncRNA
4. Samples collection and RNA integrity

**Before RNA extraction**

RNA is more unstable than DNA, therefore higher precautions are needed to avoid degradation

**TISSUE COLLECTION:**

- Liquid nitrogen → -80°C storage
- RNA later (for solid tissues) → -20°C short term
- Tempus/Pax tubes (for blood) → -80°C long term
After RNA extraction

**RIN** (RNA integrity number): algorithm for assigning integrity values to RNA measurements.

- **RIN>7 ok**
- 10: maximum
- 0: minimum
4. Samples collection and RNA integrity

RNA quality (RIN) and quantification: Bioanalyzer
5. Library preparation

**Before Library preparation**

- Total RNA seq (Ribosomal depletion, DNase treatment, fragmentation, library preparation)
- mRNA+Inc (polyA+) RNA seq (polyA enrichment, DNase treatment, fragmentation, library preparation)
- shortRNA seq (Size selection, DNase treatment, library preparation)
1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

4) Second strand cDNA synthesized

5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

7) Ligated fragments PCR amplified
5. 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)
5. Library preparation

1) Ligate 3’ and 5’ adapters

2) RT-PCR

- sequence required for amplification on flow cell
- sequence primer hybridization site
- barcode sequencing primer hybridization site
- barcode sequence
6. Data analyses

WORKFLOW FOR LONG READS

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

Martin et al. 2011, Nature Review Genetics
6. Data analyses

Transcriptome assembly strategies

❖ Reference-based
❖ De novo
❖ Combined (Reference based + de novo)
❖ Pseudoalignment
Reference-based: Most common tools

- **Unspliced read aligner**
  - ✓ BWA
  - ✓ Bowtie2
  - ✓ Novoalign

- **Spliced read aligner**
  - ✓ Tophat2/Hisat2
  - ✓ STAR

- Splice-junction not considered
- Ideal for mapping against cDNA databases
- Novel splice-junction detected
- Better performance for polymorphic regions and pseudogenes
Splice junctions view through IGV (Integrative Genomics Viewer)
6. Data analyses

**De novo assembly:** Most common tools

- **Velvet**
  - ✓ Genomics and transcriptomics

- **Trinity**
  - ✓ Transcriptomics

- **Cufflinks**
  - ✓ Reassemble pre-aligned transcripts to find alternative splicing based on differential expression
6. Data analyses

COMBINED ASSEMBLY

Martin et al. 2011, Nature Review Genetics
6. Data analyses

Pseudoalignment - Kallisto

N. Bray et al., Nature Biotechnology (2016)
6. Data analyses

READ COUNT

Count the number of reads aligned to each known transcripts/isoform

- Need an annotation file

E.g HTSeq-count
6. Data analyses

NORMALIZATION

• Longer genes will have more reads mapping to them (within samples)
• Sequencing run with more depth will have more reads mapping on each gene (between samples)
6. Data analyses

NORMALIZATION

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Rep 1</th>
<th>Rep 2</th>
<th>Rep 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (2Kb)</td>
<td>10</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>B (4Kb)</td>
<td>20</td>
<td>25</td>
<td>60</td>
</tr>
<tr>
<td>C (1Kb)</td>
<td>5</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>D (10Kb)</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Raw counts

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

\[ \text{sequencing depth (million)} \]
\[ \text{transcript length (kilobase)} \]

- TPM (Transcripts Per Million)

\[ \text{transcript length} \]
\[ \text{sequencing depth} \]
6. Data analyses

DIFFERENTIAL EXPRESSION

Differential gene expression based on the **negative binomial distribution**, to account for the variability of gene expression (e.g. lower expressed genes have higher variance than higher expressed genes)
6. Data analyses

DESeq2

Input: Read count table (e.g. HTSeq but also from pseudoalignment)

Output: Table containing statistics for whether a gene is differentially expressed between condition

```
## log2 fold change (MAP): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 6 rows and 6 columns
```

<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>
Differentially expressed gene Categories

Gene Ontology project provides an ontology of defined terms representing gene product properties. The ontology covers three domains:

- **Molecular function**
  molecular activities of gene products
- **Cellular component**
  where gene products are active
- **Biological process**
  pathways and larger processes made up of the activities of multiple gene products.
Some GO and pathway analysis tools

Enrichr
http://amp.pharm.mssm.edu/Enrichr/

GORILLA
Gene Ontology enrichment analysis and visualization tool
http://cbl-gorilla.cs.technion.ac.il/

DAVID
https://david.ncifcrf.gov/

IPA
https://www.qiagenbioinformatics.com/
6. Data analyses

Micro RNA (miRNA)

- 18-22 nucleotide length
- One of the most studied among the small RNAs
- miRNA-mRNA interaction
- www.mirbase.org
The short sequence length makes small RNA difficult to map in large and complex reference genome. Common aligner for long RNA are therefore not accurate for short RNA mapping (Ziemann et al. 2016, RNA)

### TABLE 1. SmRNA/microRNA-seq analysis pipelines in common use

<table>
<thead>
<tr>
<th>Tool</th>
<th>Alignment engine</th>
<th>Reference sequence</th>
<th>Limited species</th>
<th>Local computer</th>
<th>Open source</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRExpress</td>
<td>Smith-Waterman</td>
<td>miRbase</td>
<td>All miRbase</td>
<td>Yes</td>
<td>Yes</td>
<td>Wang et al. 2009</td>
</tr>
<tr>
<td>DSAP</td>
<td>Smith-Waterman</td>
<td>miRbase</td>
<td>All miRbase</td>
<td>Web-server only</td>
<td>NA</td>
<td>Huang et al. 2010</td>
</tr>
<tr>
<td>MiReNA</td>
<td>MEGABLAST</td>
<td>Whole genome</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>Mathelier and Carbone 2010</td>
</tr>
<tr>
<td>miRDeep</td>
<td>MEGABLAST</td>
<td>Whole genome</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>Friedländer et al. 2008</td>
</tr>
<tr>
<td>miRDeep2</td>
<td>Bowtie1</td>
<td>Whole genome</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>Friedländer et al. 2012</td>
</tr>
<tr>
<td>miRanalyzer</td>
<td>Bowtie1</td>
<td>miRbase and whole genome</td>
<td>34 species</td>
<td>Web-server only</td>
<td>No</td>
<td>Hackenberg et al. 2011</td>
</tr>
<tr>
<td>Shorttran</td>
<td>Bowtie1</td>
<td>Whole genome</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>Gupta et al. 2012</td>
</tr>
<tr>
<td>mirTools2</td>
<td>SOAP2</td>
<td>Whole genome</td>
<td>32 species</td>
<td>Yes, and web-server</td>
<td>Yes</td>
<td>Wu et al. 2013b</td>
</tr>
<tr>
<td>MiRNAkey</td>
<td>BWA</td>
<td>miRbase</td>
<td>All miRbase</td>
<td>Yes</td>
<td>Yes</td>
<td>Ronen et al. 2010</td>
</tr>
<tr>
<td>UEA sRNA workbench</td>
<td>PatMaN</td>
<td>Whole genome</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>Stocks et al. 2012</td>
</tr>
<tr>
<td>ShortStack</td>
<td>Any</td>
<td>Whole genome</td>
<td>Any</td>
<td>Yes</td>
<td>Yes</td>
<td>Axtell 2013</td>
</tr>
</tbody>
</table>

List is nonexhaustive.
THANK YOU!

