Outline

- Why sequence transcriptome?
- From RNA to sequence
- The most common way: reference based analysis pipeline
- What about de-novo assembly of transcriptomes?
- And what about scRNA-seq?
- Introduction to exercises
Why sequence transcriptome?
An RNA sequence mirrors the sequence of the DNA from which it was transcribed.

Consequently, by analyzing transcriptome we can determine when and where each gene is turned on or off in the cells and tissues of an organism.
### What can a transcriptome tell us about?

- gene sequences in genomes
- gene functions
- gene activity / gene expression
- isoforms and allelic expression
- fusion transcripts and novel transcripts
- SNPs in genes
- co-expression of genes
- cell-to-cell heterogeneity (scRNA-seq)
Transcriptomes are:

*dynamic, that is not the same over tissues and time points*

*directly derived from functional genomics elements, that is mostly protein-coding genes, providing a useful functionally relevant subset of the genome, translating into smaller sequence space*
Overview

- **Experimental design** (biology, medicine, statistics)
- **RNA extraction** (biology, biotechnology)
- **Library preparation** (biology, biotechnology)
- **High throughput sequencing** (engineering, biology, chemistry, biotechnology, bioinformatics)
- **Data processing** (bioinformatics)
- **Data analysis** (bioinformatics & biostatistics)
From RNA to sequence
From RNA to sequence:

Workflow:

1. **Exon**
   - Genome
   - mRNA

2. **5'**
   - ATG...
   - 3'

3. **Fragmentation/cDNA synthesis**

4. **Adding adapters**

5. **Sequence library**

6. **Sequencing**

**Fastq files**:

- @HWUSI_EAS100R:6:73:941:1973#0/1
  - AGTCCCCTCAGTCGATTCATCAGTTTTATATTTAT
  - $'*(((***)++)(%%%%).1*CCCCFFFF$

- @HWUSI_EAS100R:6:73:941:1973#0/2
  - AGTCCCCTCAGTCGATTCATCAGTTTTATATTTAT
  - $'*(((***)++)(%%%%).1*CCCCFFFF$

- @HWUSI_EAS100R:6:73:941:1974#0/1
  - TACCCCCGTCGAAAAATTCATCAGTTTTATATTTAT
  - $'*(((***)++)(%%%%).1*CCCCFFFF$

- @HWUSI_EAS100R:6:73:941:1974#0/2
  - TACCCCCGTCGAAAAATTCATCAGTTTTATATTTAT
  - $'*(((***)++)(%%%%).1*CCCCFFFF$
.fastq

@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFFGGGGGGGGGGGHHHHHGHGGFHFFHAG

- Line1:
.fastq

@MISEQ:233:00000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFFFGGGGGGGGGGGGGGHFFFFHGHGFHFFFFFFFFHAG

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description
- Line2:
.fastq

@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBFFFFFFFFFGGGGGGGGGGGGGGGGHHHHHHHHHAG

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description
- Line2: is the raw sequence letters
- Line3:
.fastq

@MISEQ:233:000000000-AGJP2:1:1101:15260:1358
CTGTAAATTGCCTGACTTGCTAATTGTGATTAACTTAGTTT
+
BBBBBBBBBBBBBBBBBBBBBBFHHHHHHHHHHHAG

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description

- Line2: is the raw sequence letters

- Line3: begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again

- Line4:
● Line1: begins with a '@' character and is followed by a sequence identifier and an optional description

● Line2: is the raw sequence letters

● Line3: begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again

● Line4: encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence
**Phred Quality Score**

- **Q = -10 x log P**
- where:
  - *P*, probability of base calling being incorrect
  - High *Q* = high probability of the base being correct

- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...
**Phred Quality Score**

- $Q = -10 \times \log P$
- where:
  - $P$, probability of base calling being incorrect
  - High $Q = $ high probability of the base being correct

- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...10 times.
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of...
Phred Quality Score

- \[ Q = -10 \times \log P \]
- where:
  - \( P \), probability of base calling being incorrect
  - High \( Q \) = high probability of the base being correct

- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...10 times.
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of...100 times.
- A Phred quality score of 30 to a base, means that the base is called incorrectly in 1 out of...
Phred Quality Score

- \( Q = -10 \times \log P \)
- where:
  - \( P \), probability of base calling being incorrect
  - High \( Q \) = high probability of the base being correct

- A Phred quality score of 10 to a base means that the base is called incorrectly in 1 out of...10 times.
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of...100 times.
- A Phred quality score of 30 to a base, means that the base is called incorrectly in 1 out of...1000 times etc...
PE, paired-end

- Two .fastq files are created per sequenced library
- The order of reads in files is identical and naming of reads is the same with the exception of the end information
- The way of naming reads are changing over time so the read names depend on software version

@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACC

@61DFRAAXX100204:1:100:10494:3070/2
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATA
+
_^_a^cccegcgghhgZc`ghhc^eggd^_\[d\]defcdfd^Z^OXWaQ^ad
SE
- F: the single read is in the sense (F, forward) orientation
- R: the single read is in the antisense (R, reverse) orientation

PE
- RF: first read (/1) is sequenced as anti-sense (R) & second read (/2) is in the sense strand (F)
- FR: first read (/1) is sequenced as sense (F) & second read (/2) is in the antisense strand (R)
Reference based data analysis pipeline
Reference based data analysis pipeline

Overview

RNA-Seq reads

Align reads to genome

Assemble transcripts de novo

Genome

Assemble transcripts from spliced alignments

More abundant

Less abundant

Align transcripts to genome
Main steps

- Initial processing incl. QC
- Aligning reads to reference genome
- Counting reads
- Differential gene expression
- Further analysis
Initial processing incl. QC

- Demultiplex by index or barcode
- Remove adapter sequences
- Trim reads by quality
- Discard reads by quality/ambiguity

Available tools

FastQC, PRINSEQ, TRIMMOMATICS, TrimGalore, FastX, Cutadapt
Initial processing incl. QC

- filtering reads for quality score, e.g. with avg. quality below 20 defined within 4-base wide sliding window
- filtering reads for read length, e.g. reads shorter than 36 bases
- removing artificial sequences, e.g. adapters
Aligning reads

Reference based data analysis pipeline

Aligning reads

Olga (NBIS)

RNA-seq

November 2017 22 / 49
Aligning reads
Aligning reads: mappers

- important to use mappers allowing for a read to be "split" between distant regions of the reference in the event that the read spans two exons
- lots of different aligners exists based on various algorithms e.g. brute force comparison, Burrows-Wheeler Transform, Smith-Waterman, Suffix tree
- usually there is a trade-off between speed versus accuracy and sensitivity
- usually the "biggest difference" is with default settings, most mappers will allow to optimize settings
- performance vary by genome complexity

https://www.nature.com/articles/nmeth.4106

Available tools
STAR, HISAT, MapSlice2, Subread, TopHat
Aligning reads: reference files

.fasta (download reference genome FASTA file)

.gtfs (download the corresponding genome annotation in GTF or GFF)

Source

ENSEMBL, NCBI
Aligning reads: QC

Post mapping QC, e.g. reads should mostly map to known genes, most splice event should be known and canonical (GU-AG)
Reference based data analysis pipeline

Counting reads

Available tools
HTSeq, featureCounts, R
## Counting reads

<table>
<thead>
<tr>
<th></th>
<th>union</th>
<th>intersection_strict</th>
<th>intersection_nonempty</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>no_feature</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>no_feature</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>ambiguous</td>
<td>gene_A</td>
<td>gene_A</td>
<td>gene_A</td>
</tr>
<tr>
<td>ambiguous</td>
<td>ambiguous</td>
<td>ambiguous</td>
<td>ambiguous</td>
</tr>
</tbody>
</table>

from: [http://www-huber.embl.de/users/anders/HTSeq/doc/count.html](http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)
## Counting reads

<table>
<thead>
<tr>
<th>Transcript</th>
<th>P1822_1</th>
<th>P1822_2</th>
<th>P1822_3</th>
<th>P1822_4</th>
<th>P1822_5</th>
<th>P1822_6</th>
<th>P1822_7</th>
<th>P1822_8</th>
<th>P1822_9</th>
<th>P1822_10</th>
<th>P1822_11</th>
<th>P1822_12</th>
<th>P1822_13</th>
<th>P1822_14</th>
<th>P1822_15</th>
<th>P1822_16</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSMUSG0000102693</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000088000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG0000001052655</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000103922</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td>14</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000033845</td>
<td>972</td>
<td>860</td>
<td>878</td>
<td>1085</td>
<td>1058</td>
<td>1099</td>
<td>992</td>
<td>1143</td>
<td>947</td>
<td>1059</td>
<td>970</td>
<td>1147</td>
<td>801</td>
<td>837</td>
<td>1042</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000102275</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000025903</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000104217</td>
<td>16</td>
<td>13</td>
<td>17</td>
<td>16</td>
<td>22</td>
<td>17</td>
<td>12</td>
<td>27</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000033813</td>
<td>2560</td>
<td>2581</td>
<td>2937</td>
<td>3904</td>
<td>2975</td>
<td>3100</td>
<td>3027</td>
<td>3417</td>
<td>2272</td>
<td>2801</td>
<td>2266</td>
<td>3294</td>
<td>2491</td>
<td>2578</td>
<td>2554</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000062588</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000103280</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000020459</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG0000001091305</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000102653</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000085623</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000091665</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000033793</td>
<td>3682</td>
<td>3757</td>
<td>4414</td>
<td>5978</td>
<td>3774</td>
<td>4102</td>
<td>3815</td>
<td>4250</td>
<td>4193</td>
<td>4962</td>
<td>4240</td>
<td>5694</td>
<td>3565</td>
<td>3757</td>
<td>3849</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000104352</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000104046</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000102907</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000025905</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000103936</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000093015</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000093519</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000033774</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG000000103090</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000025907</td>
<td>1816</td>
<td>2087</td>
<td>2088</td>
<td>2820</td>
<td>2012</td>
<td>2236</td>
<td>2065</td>
<td>2727</td>
<td>2586</td>
<td>2931</td>
<td>2813</td>
<td>3667</td>
<td>2410</td>
<td>2739</td>
<td>2479</td>
<td></td>
</tr>
<tr>
<td>ENSMUSG00000090031</td>
<td>43</td>
<td>58</td>
<td>55</td>
<td>73</td>
<td>38</td>
<td>38</td>
<td>57</td>
<td>96</td>
<td>89</td>
<td>107</td>
<td>98</td>
<td>123</td>
<td>76</td>
<td>93</td>
<td>66</td>
<td>69</td>
</tr>
</tbody>
</table>
Gene counts depend e.g. on sequencing depth of a sample and on the sequence length of the gene/transcript. Raw read counts cannot be used to compare gene expression across libraries.

Normalization methods

- CPM, counts per million, accounts for sequencing depth
- RPKM/FPKM, Reads/Fragments Per Kilobase Per Million accounts for sequencing depth and transcript length
- TMM, Trimmed Mean of M-values, accounts for sequencing depth and transcript length and composition of the RNA population
- and few other using scaling factors methods...
Differential gene expression

\[ \text{Outcome}_i = (\text{Model}_i) + \text{error}_i \]

- we collect data on a sample from a much larger population. Statistics lets us to make inferences about the population from which it was derived
- we try to predict the outcome given a model fitted to the data
Differential gene expression

\[ t = \frac{x_1 - x_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]

in RNA-seq case:

- we take the normalized read counts
- and we perform statistical analysis to discover quantitative changes in expression levels between experimental groups
- e.g. to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.
Differential expression

*Usually, reads counts do not follow normal distribution & we work with low number of biological replicates*

**DE methods**

- Discrete distribution models, e.g. edgeR, DESeq2
- Continuous discrete models, e.g. t-test
- Non-parametric model, e.g. SAMseq
The likelihood of observing a significant p-value increases as we do more tests, i.e. testing more than one gene. Modern FDR adjustment techniques take into account of background expectation of a uniformly distributed p-values and adjust their values accordingly to how significantly different things are, so the p-values from multiple testing can be interpreted more accurately.
Available tools
edgeR, DEXSeq
Further analysis

- Annotating the results e.g. with gene symbols, GO terms
- Visualizing the results, e.g. Volcano plots
- Gene set analysis etc...

Available tools

bioMart (R), DAVID, GOrilla, REVIGO, ClustVis...
What about de-novo assembly of transcriptomes?
What about de-novo assembly of transcriptomes?

Overview

RNA-Seq reads

Align reads to genome

Assemble transcripts de novo

Genome

Assemble transcripts from spliced alignments

More abundant

Less abundant

Align transcripts to genome
Building a reference transcriptome

- alternative strategy when well-assembled reference genome from a relatively recently diverged organism is not available
- primary goal: assembling a transcriptome *de novo* to reconstruct a set of contiguous sequences (contigs) presumed to reflect accurately a large portion of the RNAs actually transcribed in the cells

not a trivial task, because

- a limited amount of information about the original gene transcripts is retained in the short reads produced by a sequencer
- genes show different levels of gene expression (uneven coverage)
- more sequencing depth is needed to represent less abundant genes and rare events
- reads from the same transcript must be placed together in the face of variants introduced by polymorphism and sequencing errors
- and the process must assemble reads from different but often similar, paralogous transcripts as separate contigs
What about de-novo assembly of transcriptomes?

Solutions to sequence assembly arose from the field of mathematics known as graph theory. These approaches were designed with genome assembly in mind but have been adapted for transcriptome assembly as necessary. Most of them are based on de Bruijn graphs.

Available tools

- **Velvet/Oases**: Velvet constructs de Bruijn graphs, simplifies the graphs, and corrects the graphs for errors and repeats. Oases post-processes Velvet assemblies with different k-mer sizes.

- **Trans-ABySS**: much like the Velvet/Oases model, Trans-ABySS (Robertson et al. 2010) takes multiple ABySS assemblies (Simpson et al. 2009) produced from a range of k-mer sizes to optimize transcriptome assemblies in the face of varying coverage across transcripts.

- **Trinity**: "Inchworm" builds initial contigs by finding paths through k-mer graphs. "Chrysalis" groups these contigs together and builds de Bruijn graphs for these groups, in which the overlaps are nodes and the k-mers connecting edges. "Butterfly" simplifies the graphs when possible, then reconciles the graphs with original reads to output individual contigs representative of unique splice variants and paralogous transcripts.
What about de-novo assembly of transcriptomes?

Building a reference transcriptome

- a) all substrings of length k (k-mers) are generated from each read
- b) each unique k-mer is used to represent a node in the De Bruijn graph, pairs of nodes are connected if shifting a k-mer by one character creates an exact k??1 overlap between the two k-mers.
- The example (5-mers) illustrates a SNP or sequencing error and an example of an intron or a deletion.
- Single-nucleotide differences cause 'bubbles' of length k in the De Bruijn graph, whereas introns or deletions introduce a shorter path in the graph
- c,d) chains of adjacent nodes in the graph are collapsed into a single node when the first node has an out degree of one and the second node has an in degree of one
- e) the isoforms are then assembled. See more http://rdcu.be/zSpz
If a reference genome is available, annotation is relatively straightforward: genomic coordinates from the reference genome are normally associated with various forms of annotation information through databases. A transcriptome assembled de novo, on the other hand, is often annotated from scratch.

**NCBI-supported BLAST**

- "match" query sequences to one or more databases of curated, annotated sequences, using an efficient local sequence alignment approach.
- It may be adequate to blast against a database of known or predicted transcripts from the reference genome of a closely-related organism.
- It may be desirable to blast contigs against all nucleotide sequences in an inclusive database.
- If the annotation emphasis is on protein-coding transcripts, BLASTx, which translates each query sequence (in all six reading frames) to amino acid sequences and uses these to query a protein database, may be an appropriate tool.
And what about scRNA-seq?
And what about scRNA-seq?

- scRNA-seq are affected by higher noise (amplification biases, dropout event, 3’bias, partial coverage, uneven depth, stochastic nature of transcription, multimodality in gene expression)
- read processing steps to generate count matrix are largely the same as for bulk RNA-seq, but the spike-in normalization is a must
And what about scRNA-seq?

Common steps

- Quality control on the cells
- Classification of the cell-cycle phase
- Normalization of cell-specific biases
- Checking for important technical factors
- Modelling and removing technical noise
- Data exploration with dimensionality reduction
- Clustering cells into putative subpopulations
- Detecting marker genes between subpopulations
- see more: Bioconductor simpleSingleCell workflow
Exercises
Main exercise

- checking the quality of the raw reads with FastQC
- mapping the reads to the reference genome using STAR
- converting between SAM and BAM files format using Samtools
- assessing the post-alignment reads quality using QualiMap
- counting reads overlapping with genes regions using featureCounts
- building statistical model to find DE genes using edgeR called from a prepared R script
Bonus exercises

- functional annotation, putting DE genes in the biological context
- exon usage, studying the alternative splicing
- data visualisation and graphics
- de novo transcriptome assembly
Thank you for attention
Questions?

Enjoy the rest of the course

Read more

- RNA-seqlopedia
- RNA-Seq blog
- Conesa et al. Genome Biology, 2016, A survey of best practices for RNA-seq data analysis