RNA Sequencing

MMG 835, SPRING 2016
Eukaryotic Molecular Genetics

George I. Mias
Department of Biochemistry and Molecular Biology
gmiash@msu.edu
RNA-Sequencing

- High throughput method
- Count the number of reads (transcripts)
- Can probe entire genome for expression (non-targeted approach)
- Splicing information
- Alternative splicing and novel isoforms
- Possible to phase (algorithmically/new longer read technology)
- Can get variant information
- Now cheaper (almost comparative to microarray, cost mainly library prep)
<table>
<thead>
<tr>
<th>Technology</th>
<th>Tiling microarray</th>
<th>cDNA or EST sequencing</th>
<th>RNA-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology specifications</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principle</td>
<td>Hybridization</td>
<td>Sanger sequencing</td>
<td>High-throughput sequencing</td>
</tr>
<tr>
<td>Resolution</td>
<td>From several to 100 bp</td>
<td>Single base</td>
<td>Single base</td>
</tr>
<tr>
<td>Throughput</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Reliance on genomic sequence</td>
<td>Yes</td>
<td>No</td>
<td>In some cases</td>
</tr>
<tr>
<td>Background noise</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Application</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simultaneously map transcribed regions and gene expression</td>
<td>Yes</td>
<td>Limited for gene expression</td>
<td>Yes</td>
</tr>
<tr>
<td>Dynamic range to quantify gene expression level</td>
<td>Up to a few-hundredfold</td>
<td>Not practical</td>
<td>&gt;8,000-fold</td>
</tr>
<tr>
<td>Ability to distinguish different isoforms</td>
<td>Limited</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ability to distinguish allelic expression</td>
<td>Limited</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Practical issues</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Required amount of RNA</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Cost for mapping transcriptomes of large genomes</td>
<td>High</td>
<td>High</td>
<td>Relatively low</td>
</tr>
</tbody>
</table>

RNA-Sequencing

Keep in mind

• Quantitation Issues
• Biases
• Transcribed region Complexity
• Sample quality (degradation)
• Aligner dependent alignment
Data Considerations

Steps

• Experiment Design
• Library Prep
• Sequencing
• Data
• Quality Control (QC)
• Assembly
• Annotation
• Quantification
• Differential Expression
• Biological Significance
Rna-Seq Protocols

TruSeq RNA

- mRNA
- polyA select
- Fragment
- Random hexamer
- First and second strand synthesis

illumina.com
Rna-Seq Protocols

TruSeq RNA Stranded

Reverse transcriptase

Preserve strandness information

Novel transcripts/annotation

TSS

First strand (dNTPs)

Note: dUTP

End repair

Phosphorylate 5'

A-overhang 3'

Adaptor ligation

Denature and amplify

Product ready for cluster generation

illumina.com
Rna-Seq Protocols

Possibilities

• mRNA Poly(A)-selection
  ‣ hybridization to oligo-dT beads
  ‣ N.B. 3’ bias

• rRNA depletion

• small RNA
  ‣ size selection

• Paired end

• Low RNA amounts - amplification
Rna-Seq Protocols

Normalization

• Long transcripts
• Coverage differences
Rna-Seq Protocols

**RPKM**, *Reads per Kilobase per Million mapped reads*

**FPKM**, *Fragments per Kilobase per Million mapped reads* (paired-end)

errors in first strand cDNA

Mismatches

• biological variation
  ‣ genomic DNA
  ‣ RNA-editing
• errors in library preparation
  ‣ hexamer mispriming
  ‣ PCR errors
• sequencing errors [id of bases]

van Gurp, McIntyre & Verhoeven, PLoS ONE 8(12): e85583 (2013)
errors in first strand cDNA

- mismatches mainly in the first seven nucleotides of first strand cDNA

Jiang L et al., Genome Res 21: 1543–1551 (2011)
errors in first strand cDNA

random hexamer mispriming

van Gurp, McIntyre & Verhoeven, PLoS ONE 8(12): e85583 (2013)
errors in first strand cDNA

van Gurp, McIntyre & Verhoeven, PLoS ONE 8(12): e85583 (2013)
Rna-Seq Considerations

Li, Tighe et al., Nature Biotechnology 12(9), p. 915 (2014)
Data Considerations

Mapping Reference
- Genome
- Transcriptome

Alignment
- Short read alignment
- Long read alignment
- Spliced Vs. Non Spliced Alignment
- Mismatches & Duplications
- Non-Uniqueness
- Variant Calls

De Novo Assembly
- Genome based
- non-genome based
.fastq format

Four lines repeating:
1. @title and optional description / Sequence Identifier
2. Sequence
3. + (and optional repeat of title line)
4. Quality scores corresponding to Sequence (2)
.fastq format

Four lines repeating:
1. @title and *optional* description / Sequence Identifier
2. Sequence
3. + (and *optional* repeat of title line)
4. Quality scores corresponding to Sequence (2)

**Example:**
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAAAAAA9#:<#<;<<<<???##=
## .fastq format

Each sequence identifier line starts with `@`.

<table>
<thead>
<tr>
<th>Element</th>
<th>Requirements</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>@</code></td>
<td><code>@</code></td>
<td>Each sequence identifier line starts with <code>@</code></td>
</tr>
<tr>
<td><code>&lt;instrument&gt;</code></td>
<td>Characters allowed: a–z, A–Z, 0–9 and underscore</td>
<td>Instrument ID</td>
</tr>
<tr>
<td><code>&lt;run number&gt;</code></td>
<td>Numerical</td>
<td>Run number on instrument</td>
</tr>
<tr>
<td><code>&lt;flowcell ID&gt;</code></td>
<td>Characters allowed: a–z, A–Z, 0–9</td>
<td></td>
</tr>
<tr>
<td><code>&lt;lane&gt;</code></td>
<td>Numerical</td>
<td>Lane number</td>
</tr>
<tr>
<td><code>&lt;tile&gt;</code></td>
<td>Numerical</td>
<td>Tile number</td>
</tr>
<tr>
<td><code>&lt;x_pos&gt;</code></td>
<td>Numerical</td>
<td>Run number on instrument</td>
</tr>
<tr>
<td><code>&lt;y_pos&gt;</code></td>
<td>Numerical</td>
<td>X coordinate of cluster</td>
</tr>
<tr>
<td><code>&lt;read&gt;</code></td>
<td>Numerical</td>
<td>Read number. 1 can be single read or Read 2 of paired-end</td>
</tr>
<tr>
<td><code>&lt;is filtered&gt;</code></td>
<td>Y or N</td>
<td>Y if the read is filtered (did not pass), N otherwise</td>
</tr>
<tr>
<td><code>&lt;control number&gt;</code></td>
<td>Numerical</td>
<td>0 when none of the control bits are on, otherwise it is an even number. On HiSeq X systems, control specification is not performed and this number is always 0.</td>
</tr>
<tr>
<td><code>&lt;sample number&gt;</code></td>
<td>Numerical</td>
<td>Sample number from sample sheet</td>
</tr>
</tbody>
</table>

[http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm](http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm)
Phred Scores

- Phred Algorithm
- Used with Sanger Data.
- Algorithm assigns probability of error in calling a base

Quality score (Q-Score)
Q = -10 \times \log\text{(error probability)}
\quad P = 10^{-Q/10}

Q-Scores

<table>
<thead>
<tr>
<th>Quality Score</th>
<th>Error Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q40</td>
<td>0.0001 (1 in 10,000)</td>
</tr>
<tr>
<td>Q30</td>
<td>0.001 (1 in 1,000)</td>
</tr>
<tr>
<td>Q20</td>
<td>0.01 (1 in 100)</td>
</tr>
<tr>
<td>Q10</td>
<td>0.1 (1 in 10)</td>
</tr>
</tbody>
</table>

Quality score (Q-Score)

\[ Q = -10 \times \log(\text{error probability}) \]

\[ P = 10^{-\frac{Q}{10}} \]
# Q-Scores

**Q-Score TABLE**

Encode Q-Score to characters Using ASCII

<table>
<thead>
<tr>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>[NULL]</td>
<td>32</td>
<td>20</td>
<td>[SPACE]</td>
<td>64</td>
<td>40</td>
<td>@</td>
<td>96</td>
<td>60</td>
<td>`</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>[START OF HEADING]</td>
<td>33</td>
<td>21</td>
<td>!</td>
<td>65</td>
<td>41</td>
<td>A</td>
<td>97</td>
<td>61</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>[START OF TEXT]</td>
<td>34</td>
<td>22</td>
<td>`</td>
<td>66</td>
<td>42</td>
<td>B</td>
<td>98</td>
<td>62</td>
<td>b</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>[END OF TEXT]</td>
<td>35</td>
<td>23</td>
<td>#</td>
<td>67</td>
<td>43</td>
<td>C</td>
<td>99</td>
<td>63</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>[END OF TRANSMISSION]</td>
<td>36</td>
<td>24</td>
<td>$</td>
<td>68</td>
<td>44</td>
<td>D</td>
<td>100</td>
<td>64</td>
<td>d</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>[ENQUIRY]</td>
<td>37</td>
<td>25</td>
<td>%</td>
<td>69</td>
<td>45</td>
<td>E</td>
<td>101</td>
<td>65</td>
<td>e</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>[ACKNOWLEDGE]</td>
<td>38</td>
<td>26</td>
<td>&amp;</td>
<td>70</td>
<td>46</td>
<td>F</td>
<td>102</td>
<td>66</td>
<td>f</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>[BEL]</td>
<td>39</td>
<td>27</td>
<td>'</td>
<td>71</td>
<td>47</td>
<td>G</td>
<td>103</td>
<td>67</td>
<td>g</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>[BACKSPACE]</td>
<td>40</td>
<td>28</td>
<td>(</td>
<td>72</td>
<td>48</td>
<td>H</td>
<td>104</td>
<td>68</td>
<td>h</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>[HORIZONTAL TAB]</td>
<td>41</td>
<td>29</td>
<td>)</td>
<td>73</td>
<td>49</td>
<td>I</td>
<td>105</td>
<td>69</td>
<td>i</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>[VERTICAL TAB]</td>
<td>43</td>
<td>2B</td>
<td>+</td>
<td>75</td>
<td>4B</td>
<td>K</td>
<td>107</td>
<td>6B</td>
<td>k</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>[FORM FEED]</td>
<td>44</td>
<td>2C</td>
<td>,</td>
<td>76</td>
<td>4C</td>
<td>L</td>
<td>108</td>
<td>6C</td>
<td>l</td>
</tr>
<tr>
<td>13</td>
<td>D</td>
<td>[CARRIAGE RETURN]</td>
<td>45</td>
<td>2D</td>
<td>-</td>
<td>77</td>
<td>4D</td>
<td>M</td>
<td>109</td>
<td>6D</td>
<td>m</td>
</tr>
<tr>
<td>14</td>
<td>E</td>
<td>[SHIFT OUT]</td>
<td>46</td>
<td>2E</td>
<td>.</td>
<td>78</td>
<td>4E</td>
<td>N</td>
<td>110</td>
<td>6E</td>
<td>n</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>[SHIFT IN]</td>
<td>47</td>
<td>2F</td>
<td>/</td>
<td>79</td>
<td>4F</td>
<td>O</td>
<td>111</td>
<td>6F</td>
<td>o</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>[DATA LINK ESCAPE]</td>
<td>48</td>
<td>30</td>
<td>0</td>
<td>80</td>
<td>50</td>
<td>P</td>
<td>112</td>
<td>70</td>
<td>p</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>[DEVICE CONTROL 1]</td>
<td>49</td>
<td>31</td>
<td>1</td>
<td>81</td>
<td>51</td>
<td>Q</td>
<td>113</td>
<td>71</td>
<td>q</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>[DEVICE CONTROL 2]</td>
<td>50</td>
<td>32</td>
<td>2</td>
<td>82</td>
<td>52</td>
<td>R</td>
<td>114</td>
<td>72</td>
<td>r</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>[DEVICE CONTROL 3]</td>
<td>51</td>
<td>33</td>
<td>3</td>
<td>83</td>
<td>53</td>
<td>S</td>
<td>115</td>
<td>73</td>
<td>s</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>[DEVICE CONTROL 4]</td>
<td>52</td>
<td>34</td>
<td>4</td>
<td>84</td>
<td>54</td>
<td>T</td>
<td>116</td>
<td>74</td>
<td>t</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>[NEGATIVE ACKNOWLEDGE]</td>
<td>53</td>
<td>35</td>
<td>5</td>
<td>85</td>
<td>55</td>
<td>U</td>
<td>117</td>
<td>75</td>
<td>u</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>[SYNCHRONOUS IDLE]</td>
<td>54</td>
<td>36</td>
<td>6</td>
<td>86</td>
<td>56</td>
<td>V</td>
<td>118</td>
<td>76</td>
<td>v</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>[ENG. OF TRANS. BLOCK]</td>
<td>55</td>
<td>37</td>
<td>7</td>
<td>87</td>
<td>57</td>
<td>W</td>
<td>119</td>
<td>77</td>
<td>w</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
<td>[CANCEL]</td>
<td>56</td>
<td>38</td>
<td>8</td>
<td>88</td>
<td>58</td>
<td>X</td>
<td>120</td>
<td>78</td>
<td>x</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>[END OF MEDIUM]</td>
<td>57</td>
<td>39</td>
<td>9</td>
<td>89</td>
<td>59</td>
<td>Y</td>
<td>121</td>
<td>79</td>
<td>y</td>
</tr>
<tr>
<td>26</td>
<td>1A</td>
<td>[SUBSTITUTE]</td>
<td>58</td>
<td>3A</td>
<td>:</td>
<td>90</td>
<td>5A</td>
<td>Z</td>
<td>122</td>
<td>7A</td>
<td>z</td>
</tr>
<tr>
<td>27</td>
<td>1B</td>
<td>[ESCAPE]</td>
<td>59</td>
<td>3B</td>
<td>;</td>
<td>91</td>
<td>5B</td>
<td>[</td>
<td>123</td>
<td>7B</td>
<td>{</td>
</tr>
<tr>
<td>28</td>
<td>1C</td>
<td>[FILE SEPARATOR]</td>
<td>60</td>
<td>3C</td>
<td>&lt;</td>
<td>92</td>
<td>5C</td>
<td>}</td>
<td>124</td>
<td>7C</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1D</td>
<td>[GROUP SEPARATOR]</td>
<td>61</td>
<td>3D</td>
<td>=</td>
<td>93</td>
<td>5D</td>
<td>`</td>
<td>125</td>
<td>7D</td>
<td>`</td>
</tr>
<tr>
<td>30</td>
<td>1E</td>
<td>[RECORD SEPARATOR]</td>
<td>62</td>
<td>3E</td>
<td>&gt;</td>
<td>94</td>
<td>5E</td>
<td>~</td>
<td>126</td>
<td>7E</td>
<td>~</td>
</tr>
<tr>
<td>31</td>
<td>1F</td>
<td>[UNIT SEPARATOR]</td>
<td>63</td>
<td>3F</td>
<td>?</td>
<td>95</td>
<td>5F</td>
<td>[DEL]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(ASCII, American Standard Code for Information Interchange)

https://simple.wikipedia.org/wiki/ASCII
Q-Scores

ASCII Conversion for different schemes

https://en.wikipedia.org/wiki/FASTQ_format#Encoding
Aligners

• Unspliced
  ‣ Fast
  ‣ Known exons and splice junctions
  ‣ Reference required
  ‣ c.f. DNA alignment
    • Borrows Wheeler Transform based (BWT)
      • BWA
      • Bowtie
      • SHRiMP
      • Stampy

Garber et al., Nature Methods 8, 469–477 (2011)
Aligners

• Align to whole genome, including intron-spanning reads that allow large gaps
• Exon first
  ‣ TopHat, MapSplice
  ‣ Unspliced alignment first
  ‣ Unmapped reads after
• Seed-extend
  ‣ GSNAP, QPALMA
  ‣ Short seeds mapped first
• K-mer mappers
Rna-Seq Alignment

Example 2 Exons

Garber et al., Nature Methods 8, 469–477 (2011)
Rna-Seq Alignment

Example 2 Exons

map full, unspliced reads (exonic reads)

Garber et al., Nature Methods 8, 469–477 (2011)
Rna-Seq Alignment

Example 2 Exons

map full, unspliced reads (exonic reads)

remaining reads divided into smaller pieces & mapped to genome

Garber et al., Nature Methods 8, 469–477 (2011)
store a map of all small words (k-mers) of similar size in the genome in an efficient lookup data structure
Rna-Seq Alignment

- each read is divided into k-mers
- k-mers mapped to the genome via the lookup structure.

Garber et al., Nature Methods 8, 469–477 (2011)
Rna-Seq Alignment

- Mapped k-mers extended into larger alignments
- May include gaps flanked by splice sites.

Garber et al., Nature Methods 8, 469–477 (2011)
Example: gene & associated retrotransposed pseudogene.

**Pseudogenes:**
defunct genomic loci with sequence similarity to functional genes but *lacking coding potential* due to the presence of disruptive mutations such as frame shifts and premature stop codons.

Pei et al., Genome Biology 13:R51 (2012).
Categories of Pseudogenes

1. **Processed pseudogenes** Created by retrotransposition of mRNA from functional protein-coding loci back into the genome

2. **Duplicated (also referred to as unprocessed) pseudogenes**
   Derived from duplication of functional genes

3. **Unitary pseudogenes** Arise through in situ mutations in previously functional protein-coding genes

Pei et al., Genome Biology 13:R51 (2012).

Example: gene & associated retrotransposed pseudogene.
Example: gene & associated retrotransposed pseudogene.

- Exonic reads will map to both the gene and its pseudogene
- Gene placement preferred owing to lack of mutations
- Spliced read could be incorrectly assigned to the pseudogene as it appears to be exonic, preventing higher-scoring spliced alignments from being pursued.

Transcriptome Reconstruction

Reads originating from two different isoforms of the same genes colored black and gray.

Transcriptome Reconstruction

Genome Guided Approach

Align reads to genome

Fragments aligned to genome

Transcriptome Reconstruction

Genome Guided Approach

Align reads to genome

Fragmented reads are used to build a transcript graph, which is then parsed into gene annotations.

Transcriptome Reconstruction

**Genome Guided Approach**

1. Align reads to genome
2. Fragments aligned to genome
3. Assemble alignments
4. Transcript graph
5. Parse graph into transcripts

Transcriptome Reconstruction

Genome Guided Approach

- Reference genome necessary
- Examples: Cufflinks or Scripture

Transcriptome Reconstruction

Genome-Independent Approach

Break reads into k-mer seeds

Transcriptome Reconstruction

Genome-Independent Approach

Break reads into k-mer seeds

Assemble reads

dk-mer graph structure.

Transcriptome Reconstruction

Genome-Independent Approach

Break reads into k-mer seeds

Assemble reads

d e Bruijn k-mer graph

Parse graph into sequences

The graph is parsed to identify transcript sequences

Transcriptome Reconstruction

Genome-Independent Approach

Break reads into k-mer seeds

Assemble reads

de Bruijn k-mer graph

Parse graph into sequences

Align sequences to genome

Transcript 1

Transcript 2

Sequence-fragmented RNA

RNA 1  RNA 2

Sequences are aligned to the genome.

Transcriptome Reconstruction

Genome-Independent Approach

Break reads into k-mer seeds

Assemble reads

d de Bruijn k-mer graph

Parse graph into sequences

Align sequences to genome

Obtain gene annotations

Transcriptome Reconstruction

Genome-Independent Approach

Break reads into k-mer seeds

Assemble reads

- Computationally intensive (time & resources)
- Sequence can affect assembly
- Examples: Trinity, TransAbyss, Velvet

Transcriptome Reconstruction

Branch points and transcript possibilities

- Spliced reads give rise to four possible transcripts
- E.g. Scripture

Transcriptome Reconstruction

Branch points and transcript possibilities

only 2 transcripts needed to explain all reads (Cufflinks)

Transcriptome Reconstruction

Branch points and transcript possibilities


Select path based on coverage
Expression Quantification

- Transcripts of different lengths & different read coverage levels

Expression Quantification

- Transcripts of different lengths & different read coverage levels
- Different total read counts observed for each transcript

Expression Quantification

- Transcripts of different lengths & different read coverage levels
- Different total read counts observed for each transcript
- FPKM-normalized read counts

Reads from alternatively spliced genes may be attributable to a single isoform or more than one isoform.

• Reads are color-coded when their isoform of origin is clear.
• Black reads indicate reads with uncertain origin.

Expression Quantification

- **Isoform expression methods** estimate isoform abundances that best explain the observed read counts under a generative model.

Expression Quantification

- Samples near the original maximum likelihood estimate (dashed line) improve the robustness of the estimate and provide a confidence interval around each isoform’s abundance.

Expression Quantification

Expression Quantification

- E.g. gene with 2 expressed isoforms
- Exons colored according to the isoform of origin to all gene isoforms.

Expression Quantification

- E.g. gene with 2 expressed isoforms
- Exons colored according to the isoform of origin to all gene isoforms.

Gene models used for quantification

Expression Quantification

- E.g. gene with 2 expressed isoforms
- Exons colored according to the isoform of origin to all gene isoforms.

Gene models used for quantification

Expression Quantification

Comparison of true versus estimated FPKM values in simulated RNA-seq data.

Differential expression analysis


Expression Microarrays
Fluorescence based comparison of intensity
Differential expression analysis

Rna-Seq: Fraction of aligned reads to a given gene

Differential expression analysis

- Example Gene with 2 isoforms with isoform switch in 2 conditions.
- Similar no. of reads
- Different distribution across isoforms

Differential expression analysis

Tuxedo Suite

Tuxedo Suite

Trapnell et al., Nature Biotechnology 28(5) p511 (2010)
Tuxedo Suite

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

Trapnell et al., Nature Biotechnology 28(5) p511 (2010)
Tuxedo Suite

Trapnell et al., Nature Biotechnology 28(5) p511 (2010)
ChIP-seq

- Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq)
- Detect protein–DNA binding
- Detect chemical modifications of histone proteins

HI-C & DNA folding

- Cells cross-linked with formaldehyde; covalent links between spatially adjacent chromatin segments
- Chromatin digested with a restriction enzyme (here, HindIII)
- Resulting sticky ends are filled in with nucleotides - one biotinylated (purple dot)
- DNA is purified and sheared.
- Biotinylated junctions are isolated with streptavidin beads and identified by paired-end sequencing.

HI-C & DNA folding

- Hi-C produces genome-wide contact matrix
- Each pixel represents all interactions between 1-Mb loci
- Intensity as total number of reads