Transcriptomics and Next Generation Sequencing

Centre for Systems Biology
Technical University of Denmark

January 2010
Outline

1. Introduction and background
2. Next Generation Sequencing Technology
3. Sequencing Data
4. Computer Exercises
1. INTRODUCTION AND BACKGROUND
Introduction

The system approach

• Reductionist vs. system-wide approach
Evolution of Technologies
Gene Transcription

• Northern Blot $\Rightarrow$ Single genes
• RT-PCR $\Rightarrow$ Multiple genes
• H.D. Microarrays $\Rightarrow$ Whole genomes
• Next Generation Sequencing $\Rightarrow$ Populations of genomes
Intuition
Sequencing for Transcriptomics?

- Sample cells
- Purify mRNA*
- mRNA → cDNA
- Cleaving into 35-400nt fragments
- PCR amplification
- Sequence fragments in parallel
- Map reads to genome catalogue
- Quantify mapped reads per gene

* or DNA for genomic sequencing
Transcriptomics
Next Gen. Sequencing

**Unique new possibilities**

- Known as well as *de novo* genomes
- Track specific alleles
- Track SNPs & somatic mutations
- Track multiple species or cell lines
- Analyze alternative splicing
<table>
<thead>
<tr>
<th>Features of Technology</th>
<th>Microarrays</th>
<th>N.G. Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Known genomes</td>
<td>All genomes</td>
</tr>
<tr>
<td></td>
<td>cheaper</td>
<td>getting closer</td>
</tr>
<tr>
<td></td>
<td>Lots of post-proc.</td>
<td>Even more post-processing</td>
</tr>
</tbody>
</table>
Where is it Heading?

- Microarray chips get bigger, better and cheaper, but reagent prices don’t change
- N.G. Sequencing: Speed increases and cost drops exponentially resp.
Progression in N.G. Sequencing
A $1000 genome?
High-Throughput

• Sanger: 96 reads < 800-1000b / run

• Solexa: $1.2 \times 10^6$ reads < 75b / run
2.

Next Generation Sequencing

How it really works.
Next Generation Sequencing
Different flavours

- Illumina/Solexa
- Roche 454
- ABI Solid
- Heliscope

and more....
Next Generation Sequencing
Different flavours – 3 common elements
Technology Overview

1. (c)DNA is fragmented
2. Adaptors ligated to fragments
3. Several possible protocols yield array of PCR colonies.
4. Enzymatic extension with fluorescently tagged nucleotides.
5. Cyclic readout by imaging the array.
Technology (emulsion PCR)

1. Fragmentation of DNA
2. Primers are attached to the surface of a bead
Fragments, with adaptors, are PCR amplified within a water drop in oil.
One primer is attached to the surface of a bead
3’ modification of fragments, to covalently bind bead to chip surface.
Technology
3. Cycles of elongation
Technology
4. Cycles of imaging

Image of each chemistry cycle is captured by the instrument. After laser excitation, image data is collected for each bead/cluster.
Sequence read over multiple cycles
Repeated cycles of sequencing to determine the sequence of bases in a given fragment, a single base at a time.
Technology
A single transect image
Transcriptomics and Next Generation Sequencing

3.
Sequencing Data
Data Sequencing output

- Fastq

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCAGATT
+

!"*(((((**+))%%%++)(%%%%).1***-+**))*55CCF>>>>>>>CCCCCCCC65
```

- Ascii conversion table

<table>
<thead>
<tr>
<th>Dec</th>
<th>Hx Oct</th>
<th>Char</th>
<th>Dec</th>
<th>Hx Oct</th>
<th>Html</th>
<th>Chr</th>
<th>Dec</th>
<th>Hx Oct</th>
<th>Html</th>
<th>Chr</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 000</td>
<td>NUL (null)</td>
<td>32</td>
<td>20 040</td>
<td>$#32; Space</td>
<td>64</td>
<td>40 100</td>
<td>$#64; #</td>
<td>96</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1 001</td>
<td>SOH (start of heading)</td>
<td>33</td>
<td>21 041</td>
<td>$#33; !</td>
<td>65</td>
<td>41 101</td>
<td>$#65; A</td>
<td>97</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2 002</td>
<td>STX (start of text)</td>
<td>34</td>
<td>22 042</td>
<td>$#34; &quot;</td>
<td>66</td>
<td>42 102</td>
<td>$#66; B</td>
<td>98</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3 003</td>
<td>ETX (end of text)</td>
<td>35</td>
<td>23 043</td>
<td>$#35; #</td>
<td>67</td>
<td>43 103</td>
<td>$#67; C</td>
<td>99</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4 004</td>
<td>EOT (end of transmission)</td>
<td>36</td>
<td>24 044</td>
<td>$#36; $</td>
<td>68</td>
<td>44 104</td>
<td>$#68; D</td>
<td>100</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5 005</td>
<td>ENQ (enquiry)</td>
<td>37</td>
<td>25 045</td>
<td>$#37; %</td>
<td>69</td>
<td>45 105</td>
<td>$#69; E</td>
<td>101</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6 006</td>
<td>ACK (acknowledge)</td>
<td>38</td>
<td>26 046</td>
<td>$#38; &amp;</td>
<td>70</td>
<td>46 106</td>
<td>$#70; F</td>
<td>102</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7 007</td>
<td>BEL (bell)</td>
<td>39</td>
<td>27 047</td>
<td>$#39; '</td>
<td>71</td>
<td>47 107</td>
<td>$#71; G</td>
<td>103</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8 010</td>
<td>BS (backspace)</td>
<td>40</td>
<td>28 050</td>
<td>$#40; (</td>
<td>72</td>
<td>48 110</td>
<td>$#72; H</td>
<td>104</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9 011</td>
<td>TAB (horizontal tab)</td>
<td>41</td>
<td>29 051</td>
<td>$#41; )</td>
<td>73</td>
<td>49 111</td>
<td>$#73; I</td>
<td>105</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>A 012</td>
<td>LF (NL line feed, new line)</td>
<td>42</td>
<td>2A 052</td>
<td>$#42; \</td>
<td>74</td>
<td>4A 112</td>
<td>$#74; J</td>
<td>106</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>B 013</td>
<td>VT (vertical tab)</td>
<td>43</td>
<td>2B 053</td>
<td>$#43; +</td>
<td>75</td>
<td>4B 113</td>
<td>$#75; K</td>
<td>107</td>
<td>€</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>C 014</td>
<td>FF (NP form feed, new page)</td>
<td>44</td>
<td>2C 054</td>
<td>$#44; ,</td>
<td>76</td>
<td>4C 114</td>
<td>$#76; L</td>
<td>108</td>
<td>€</td>
<td></td>
</tr>
</tbody>
</table>
Base Calls
Read quality

• Phred quality: \( Q_{\text{phred}} = -10 \log_{10} p \)

• Solexa quality: \( Q_{\text{solexa pre-1.3}} = -10 \log_{10} \frac{p}{1 - p} \)

<table>
<thead>
<tr>
<th>Phred quality score</th>
<th>Probability that the base is called wrong</th>
<th>Accuracy of the base call</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>
A single Illumina/Solexa run (paired)

- 5 minutes just to count the number of lines in Unix
- 9 million reads of length ~75
- 0.675 billion bases
- 135x coverage @ 5 mio. bp genome
- 1.5 billion characters in one fastq file
- 1.5 gigabyte of data
4. Computer Exercises
Data

Issues with N.G. Sequencing

Sequencing is a stochastic process

• Larger genes make up bigger proportions of the fragments

• Constant highly expressed genes dominates the fragment pool

• No upper limit on quantification of a single gene as in microarrays

• 16S can constitute up to 90% of all sequenced reads

• Other specific biases we will deal with in the exercises
Thank you