High-Throughput Genomics and the Sword of Damocles

Jerry Taylor
University of Missouri

Missouri Regional Life Sciences Summit, Kauffman Conference Center KC, March 9th 2010
Rapidly Evolving
- MU GAIIX (v4.0 chemistry & v1.4 base calling) yields 2.8 Gb/lane
- New platforms with enhanced capabilities

Cost/Mb is decreasing

Numerous Applications
- RNAseq
- Genome resequencing
  - De novo assembly
  - Mutation discovery
- ChIPseq
- Small non-coding RNAs
- DNA methylation (esp. Pac Biosciences)

Democratizing Genomics!
- Small labs can become genome centers!!!
# MU DNA Core Service Costs

## Library Preparation

<table>
<thead>
<tr>
<th>Sample Preparation</th>
<th>USD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Sequencing Sample Preparation* (Paired-End Adapters)</td>
<td>$225</td>
</tr>
<tr>
<td>Digital Gene Expression (DGE)- small RNA Sample Preparation</td>
<td>$445</td>
</tr>
<tr>
<td>cDNA Sample Preparation*</td>
<td>$225</td>
</tr>
<tr>
<td>ChIP-Seq Sample Preparation*</td>
<td>$225</td>
</tr>
<tr>
<td>mRNA-Seq Sample Preparation**</td>
<td>$500</td>
</tr>
</tbody>
</table>

## Analysis

<table>
<thead>
<tr>
<th>Sample Type</th>
<th>USD/Flow Cell</th>
<th>Raw Data Output/Flow Cell (Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Read (42 bases)</td>
<td>$828.57/lane</td>
<td>5-6Gb</td>
</tr>
<tr>
<td>Single Read (84 bases)</td>
<td>$1,414.29/lane</td>
<td>11-12Gb</td>
</tr>
<tr>
<td>Paired-End Read (2x42 bases)</td>
<td>$1,657.14/lane</td>
<td>11-12Gb</td>
</tr>
<tr>
<td>Paired-End Read (2x84 bases)</td>
<td>$2157.14/lane</td>
<td>22-23Gb</td>
</tr>
</tbody>
</table>

$3000 for 1X mammalian genome sequence coverage!!!
RNASeq pileup over VIPR1

Liver High Feed Efficiency Steer

Liver Low Feed Efficiency Steer
Rapid Data Generation

Whole genome sequence on African, Asian, European cattle breeds will soon be publicly available

- Whole genome sequencing of 120 animals from 12 breeds

<table>
<thead>
<tr>
<th>Library</th>
<th>GAIIx Lanes</th>
<th>Post-Filter Reads</th>
<th>Total Bases</th>
<th>Genome Coverage (2.685 Gb = 1X)</th>
<th>Average Read Length (bp)</th>
<th>Unfiltered SNPs and Indels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brahman (10)</td>
<td>20</td>
<td>335,962,867</td>
<td>23,485,537,782</td>
<td>8.74</td>
<td>69.91</td>
<td>19,901,328</td>
</tr>
<tr>
<td>Hanwoo (15)</td>
<td>32</td>
<td>497,458,572</td>
<td>35,220,941,134</td>
<td>13.11</td>
<td>70.80</td>
<td>18,662,313</td>
</tr>
<tr>
<td>Angus (3)</td>
<td>54</td>
<td>877,918,406</td>
<td>67,112,512,237</td>
<td>24.99</td>
<td>76.45</td>
<td>14,797,621</td>
</tr>
<tr>
<td>AN626</td>
<td>18</td>
<td>310,024,359</td>
<td>23,565,486,109</td>
<td>8.77</td>
<td>76.01</td>
<td></td>
</tr>
<tr>
<td>AN828</td>
<td>18</td>
<td>289,562,373</td>
<td>22,621,184,219</td>
<td>8.42</td>
<td>78.12</td>
<td></td>
</tr>
<tr>
<td>AN1717</td>
<td>18</td>
<td>278,331,674</td>
<td>20,925,841,909</td>
<td>7.79</td>
<td>75.18</td>
<td></td>
</tr>
<tr>
<td>All libraries</td>
<td>106</td>
<td>1,711,339,845</td>
<td>125,818,991,153</td>
<td>46.85</td>
<td>73.52</td>
<td></td>
</tr>
</tbody>
</table>
Repetitive Element

Mean coverage
9X
Drivers

- Relatively inexpensive and fast
- Enormous information content
- Species independent technology
- Will make redundant many other forms of technology
  - Microarrays/Capture arrays/CNV arrays
  - Genotyping

Issues

- Suites of analysis tools only just becoming available
- Require major compute horsepower and data storage
- Analysis is non-trivial – requires training/experience
The Message

- Must prepare for the storage and analysis requirements of petabytes of data
  - It will take longer to analyze data than to generate it
  - MU compute and data storage infrastructures are at capacity
  - Limited depth of expertise

- Hardware, Software and IT/Informatics Personnel
  - Monsanto informatics support is 300 people

- Data Integration:
  - RNAseq and DNA methylation
  - Clinical records, histology, genotypes, gene expression, DNA methylation

- The world will be defined by who can and who cannot adapt (evolution)