Best Practice and Quality Issues
On Large NGS Dataset Analysis

Dawei Lin, Ph.D.
Director of Bioinformatics Core
Genome Center

July 11, 2010, Boston, ISMB
New Ideas, New People

- **De Novo Sequencing**
  - Focused sequencing
  - Large number of small genomes

- **SNP Discovery**
  - No reference genome sequences

- **Transcriptome Profiling**
  - Unknown transcriptomes

- **Metagenomics**
  - Special interests

- **Novel Use of New Capabilities**
  - High-throughput mapping
Initial Data Processing

- Raw Data Examination
  - Data might not be what is specified
- Reference Source Verification
  - Version and sites
- Quick Preliminary Analysis
  - 50%-80% mapping to reference

Bioinformatics Core, http://bioinformatics.ucdavis.edu
Common Tools

- Velvet/Oases
  - Run several parameter sets and integrate the outputs
- BWA/samtools
  - Short and long reads alignment
  - .....
Quality Trimming

starting at 5' end, find first window with mean quality < 20 (e.g.),
then trim starting at first base in that window with quality < 20

ACAGTTGTAAGGTCTGGTTTGTCTTTGTTGGTAGGACTGGTATTTTTTTACTTGTGGGT
BCBCBCCCBACCBBCCCBBCCBCCBBB7@9+8>0@; ; 5%+@57; )?=6134?-.8A@496.6;<

phred-like
quality score
36–40
31–35
26–30
21–25
16–20
11–15
6–10
1–5
Quality Checking

- Independent or Orthogonal Validation
- Cherry Pick Cases
- Biological Significance
**RNAseq Experimental Designs**

- Unreplicated Data
- Technical Replicates
- Biological Replicates
- Pooling (most popular choice)

<table>
<thead>
<tr>
<th>gene</th>
<th>wt 1</th>
<th>wt 2</th>
<th>wt 3</th>
<th>mut 1</th>
<th>mut 2</th>
<th>mut 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>214</td>
<td>240</td>
<td>190</td>
<td>120</td>
<td>124</td>
<td>137</td>
</tr>
<tr>
<td>$g_2$</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>120</td>
<td>82</td>
<td>93</td>
</tr>
<tr>
<td>$g_3$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$g_4$</td>
<td>2</td>
<td>0</td>
<td>400</td>
<td>120</td>
<td>82</td>
<td>93</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Experimental Validation

- Small Scale Test, Large scale gain
- Experimenters are not always right
  - Samples have different genetic background
  - Rare events (Dr. Joe Fass, Poster A57)
Mutation Discovery Pipeline

Integrated Mutation Discovery Operation At UC Davis Genome Center

1. Scientists retrieving or Reanalyzing data
2. Drosophila DNA Samples
3. DNA Sequencing Core
4. Bioinformaticians + Computing Clusters

Other sequences sources
Computing pipelines (MagGene, Blast ...)

Bioinformatics Core, http://bioinformatics.ucdavis.edu
Cloud and Portable Computing

- Amazon Cloud (http://aws.amazon.com/)
- Portable Ubuntu (run Ubuntu on Windows without rebooting)
Workshops at UC Davis Genome Center

3rd Intensive Next Generation Data Analysis & Cloud Computing Workshop


Bioinformatics Geeks’ Cake

```
echo QSVJRYHFCEXDILPKZTO
BWA | grep -o | sort -r | grep
  ![CKQW] | perl -ne "chomp
  p; print "$\lt";" | cut -f 2,4
  -l perl -ane "\$x = \""KPFFAO
JECKNP AIGM BMHBM CLPDD"
for each \$y (split(/,\$x)) { print
  $F[ord($y) - 65]}; print "\n"
```