Quality Control and Target Validation in Sequencing

v1.0

Simon Andrews
Laura Biggins
Boo Virk
• Support service for bioinformatics
  – Academic – Babraham Institute
  – Commercial – Consultancy

• Support BI Sequencing Facility
  – HiSeq / MiSeq based sequencing service
  – Data Management / Processing / Analysis
Interests in QC

• Developed QC for in-house sequencing
• Developed QC packages
  – FastQC
  – BamQC
  – FastQ Screen

• Developed application specific QC
  – Bismark (bisulphite methylation)
  – HiCUP (Hi-C genome structure)

• Developed data visualisation QC
  – SeqMonk (generic sequencing visualisation / analysis)
    • RNA-Seq QC
    • Small RNA QC
    • Duplication QC
Areas for today

- How do sequencing experiments go wrong
  - Learn from mistakes of others
- How to construct good QC
  - What should you run
  - What should you look for
  - How should you interpret / act
- What software exists
  - Review of existing QC packages / use cases
An example...
<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arhgef4</td>
<td>Rho guanine nucleotide exchange factor (GEF) 4</td>
</tr>
<tr>
<td>Cflar</td>
<td>CASP8 and FADD-like apoptosis regulator</td>
</tr>
<tr>
<td>Als2</td>
<td>amyotrophic lateral sclerosis 2 (juvenile) homolog (human)</td>
</tr>
<tr>
<td>Cxcr2</td>
<td>chemokine (C-X-C motif) receptor 2</td>
</tr>
<tr>
<td>Col4a3</td>
<td>collagen, type IV, alpha 3</td>
</tr>
<tr>
<td>Sag</td>
<td>retinal S-antigen</td>
</tr>
<tr>
<td>Gpr35</td>
<td>G protein-coupled receptor 35</td>
</tr>
<tr>
<td>Acmsd</td>
<td>amino carboxymuconate semialdehyde decarboxylase</td>
</tr>
<tr>
<td>Qsox1</td>
<td>quiescin Q6 sulfhydryl oxidase 1</td>
</tr>
<tr>
<td>9430070O13Rik</td>
<td>RIKEN cDNA 9430070O13 gene</td>
</tr>
<tr>
<td>Mrps14</td>
<td>mitochondrial ribosomal protein S14</td>
</tr>
<tr>
<td>Scyl3</td>
<td>SCY1-like 3 (S. cerevisiae)</td>
</tr>
<tr>
<td>Illdr2</td>
<td>immunoglobulin-like domain containing receptor 2</td>
</tr>
<tr>
<td>Atp1a2</td>
<td>ATPase, Na+/K+ transporting, alpha 2 polypeptide</td>
</tr>
<tr>
<td>Slamf8</td>
<td>SLAM family member 8</td>
</tr>
<tr>
<td>Wdr38</td>
<td>WD repeat domain 38</td>
</tr>
<tr>
<td>Exd1</td>
<td>exonuclease 3'-5' domain containing 1</td>
</tr>
<tr>
<td>Serf2</td>
<td>small EDRK-rich factor 2</td>
</tr>
</tbody>
</table>
Coverage of Raw Data

Normal Gene

PCA Gene
Using a different read mapper...
Using a different read mapper...
The Evolution of Sequencing Analysis

- Early Data
- Exploratory Tools
- Application specific tools
- Meta-analyses
- High throughput analysis
- Pipeline Development

Babraham Bioinformatics
Good Lessons from exploration and tool development

• General structure of the data
• Quantitation
• Points of commonality
  – Expectations
  – Reference points
  – Normalisation
• Statistics
Bad Lessons from exploration and tool development

• Failure modes
  – Contamination
  – Library failures

• Artefacts

• Biases

• Mis-interpretations
The Evolution of Sequencing Analysis

1. Early Data
2. Exploratory Tools
3. Application specific tools
4. Meta-analyses
5. High throughput analysis
6. Pipeline Development
Areas for today

• How do sequencing experiments go wrong
  – Learn from mistakes of others
• How to construct good QC
  – What should you run
  – What should you look for
  – How should you interpret / act
• What software exists
  – Review of existing QC packages / use cases