SOLiD Quality Control

Sander Boymans
Contents

• Why quality control

• What to check

• How to make it work

• Future work

• Questions / Suggestions
Why quality control

• Humans make errors

• Sequencing = Chemistry + Biology = Magic

• High quality data = High quality results/conclusions
Why quality control
What to check for

• Pre-mapping
  – Dots in reads
  – Mean quality score for all reads
  – Average quality score per read position
  – Occurrence of clonal reads

• Post-mapping
  – % Reads mapped
  – Unique vs Non-Unique mapped
  – Distribution of mapping quality among reads
  – Percentage of reads with (n) mismatches
  – Percentage of reads with mismatch per read position
  – Target sequence coverage (whole genome/enrichment)
  – Target sequence coverage (enrichment)
  – Occurrence of clonal reads
What to check for

Pre-mapping -- Dots in reads

Percentage reads with (n) Dots

- 0 Total: 548329193
- 1 Total: 2556873
- 1+ Total: 3715352

98.9%

0.3%

2133112.2121.13212
What to check for

Pre-mapping -- Average quality score / reads

<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>

Mean quality/read

- <5 Total: 73770
- 5-10 Total: 34514063
- 10-15 Total: 59186357
- 15-20 Total: 89813306
- 20-25 Total: 120846795
- 25-30 Total: 177037201
- >30 Total: 72308132

- 16.2%
- 10.7%
- 6.2%
- 0%
- 13.1%
- 32%
What to check for

Pre-mapping -- Average quality score / read position
What to check for

Pre-mapping -- Clonality

Premap Clonality overview

1 2,463,718
11-100 33
101-1000 1

100% 0%
What to check for

Post-mapping -- Percentage reads mapped

Mapping Overview

- **Forward Total:** 1,964,259,277 (35.4%)
- **Reverse Total:** 2,196,327,633 (39.6%)
- **Unmapped Total:** 1,385,427,282 (25%)
What to check for

Post-mapping – Unique vs Non-Unique mapped

% Uniquely Mapped Vs Repetatively Mapped

- Uniq. Mapped Total: 353300112
- Rep. Mapped Total: 62758578

84.9%

15.1%
What to check for

Post-mapping – Mapping quality of reads

Mapping quality of reads

% of all reads

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 23 25 37

Mapping Quality
What to check for

Post-mapping – Reads with mismatches

Occurrence of reads with 0-10 mismatch(es)

- 0 Total: 155388289
- 1 Total: 52878872
- 2 Total: 44258575
- 3 Total: 27335101
- 4 Total: 21078326
- 5 Total: 15411670
- 6 Total: 11817977
- 7 Total: 8901227
- 8 Total: 7014667
- 9 Total: 5571041
- 10 Total: 3644367
What to check for

Post-mapping – Percentage reads with mismatches/position

Percentage of mismatches/position
What to check for

Post-mapping – Target sequence coverage (whole genome/enrichment)

Enrichment efficiency (target coverage)

Mean coverage: 4
Median coverage: 5
Uncovered bases: 16.22 %
What to check for

Post-mapping – Target sequence coverage (enrichment)

Enrichment efficiency (on target)

- Bases in targets: 677,786
- Bases in flanks: 201,210
- Bases off-target: 21,614,048

- 96.1%
- 3%
- 0.9%
What to check for

Post-mapping – Clonality

Postmap Clonality overview

- 1: 10,745,134
- 11-100: 140,611
- 101-1000: 11
How to make it work

SOLiD run

Mapping + Pre-map analysis

Pre-map analysis report (PDF)

QC database update

Post-map analysis Report (PDF)

QC database update

Post-map analysis

Downstream analysis

SOLiD Quality Control website
How to make it work
Future work

- Bug fixing
- Add quality scores to mapping graph.
- Unique vs non unique add more categories.
Questions/Suggestions