NGS Data analysis

10th INTERNATIONAL SUMMER SCHOOL ON IMMUNOGENETICS
Stintino, Sardinia, 17 September, 2013
Outline

• Pre-Analytical Phase – possible artifacts
• Quality Checks of the device
• Sequencing Artefacts
• Alignments, Assignments, and their main problem
Pre-Analytical Phase

- Conventional PCR
  - Preferential Amplification – Loss of Alleles
  - ’Unspecific’ Amplification
  - Hybrid Molecules
  - Chimerism
- (Enzymatic Fragmentation)
  - random?
  - intensity
- Single Molecule PCR
  - Various amplification efficacy – uneven coverage
First Analysis

Run Summary: R_2012_07_12_14_19_30_user_SN1-91-HLA_314_OT2_200_20130823

Unaligned

113 M
Total Bases

118
Key Signal

90 %
ISP Loading

ISP Density

724,247
Total Reads

65 %
Usable Reads

ISP Summary

90% Loading: 1,135,150
10% Empty Wells

100% Enrichment: 1,135,150
0% No Template

72% Clonal: 813,859
28% Polyclonal

89% Final Library: 724,247
2% Test Fragments
0% Adapter Dimer
9% Low Quality

157 bp
Mean Read Length

Read Length

0 50 100 150 200 250 300 350
To be sequenced

- Sanger Exon sequencing: 270 nt
- NGS Exon: 27 000 nt
- NGS whole gene (5 kb): 500 000 nt
Fastq

- @YSEVQ:4:21
  - ACCGGGAGACACAGATCTGCAAGGCCAAGGCACAGACTGACCGAGAGAACCT
    GCGGATCGCGCTCCGCTACTACAAACCAGAGCGAGGCCGGTGAGTGACCCCGGCCCG
    CGGGGCGCAGGTCACGACCACGCCCATCCACGTACGCGCGCCCGATC
- +
- @@9>>4;;;45276649/3.307.73;592++,9:8;97<=:;AABBBB>B=@?;;4944188
  889399=;<8:97396>>>>?@99393929>9430,,00&,-0&-0&.86893332..-.'-
  25+6,,(+1011--,2.4047*001+
- @YSEVQ:4:28
  - CAAGGCCAAGGCACAGACTGACCGAGAGGACCTGCGGATCGCGCTCCGCTACCT
    ACAACCAGAGCGAGGCCGGTCCGAGTCCGACCACGGCCGGGGGCCAGGTCA
- +
- DE>C>E@E@CCCECCA@@=78=DDDD<DE?DDC:@@...7987:4777-
  ',742/74,,+65033&0(+/2063545989=1:589+166(043.00+00+
PHRED

Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example, T bases are represented in red, and this sequence has four T bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).
PHRED Quality Score
Sequencing Artefacts

- Mismatches (+/-)
- Insertions and Deletions (mainly homopolymers) (++)
Sequencing Errors

Ref: CCCCCGA

#1: CCTXCGA  Mutation
#2: CCC.GA    Deletetion
#3: CCCCCC CGA  Insertion
Map/Assemble, Assign

- Auto-Assembly
- Align references against autoassembly
- Assign HLA-Type

- Invent a reference
- Align clones against this reference
- Assign HLA Type
Mapping patterns of sequencing reads on correct and incorrect references.

IMGT/HLA Database ver 12.0

- cDNA sequences  n=9291
- gDNA sequences  n=640
### IMGT/HLA ver 12.0, n Sequences

<table>
<thead>
<tr>
<th>Exon</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>DRB1</th>
<th>DQB1</th>
<th>DPB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>270</td>
<td>437</td>
<td>208</td>
<td>72</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>2241</td>
<td>2930</td>
<td>1785</td>
<td>1249</td>
<td>305</td>
<td>149</td>
</tr>
<tr>
<td>3</td>
<td>2241</td>
<td>2929</td>
<td>1784</td>
<td>117</td>
<td>113</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>739</td>
<td>945</td>
<td>414</td>
<td>78</td>
<td>27</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>229</td>
<td>360</td>
<td>200</td>
<td>72</td>
<td>322</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>206</td>
<td>285</td>
<td>185</td>
<td>69</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>205</td>
<td>277</td>
<td>182</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>201</td>
<td>145</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Software

• Open source
  – Blast
  – Samtools
  – Bowtie
  – ...

• HLA-Software
  – Connexio
  – GenDX
  – Omixon
  – ....
Worst Case Scenario

- sequencing of the whole gene (5-10 kb amplicon)
- coverage
- of several loci (A, B, C, DRB1, DQB1, DPB1)
- against non existing reference sequences
Coverage and Statistics

HLA-B*44:02:01:01 and B*44:138Q
## Coverage and Statistics

<table>
<thead>
<tr>
<th>Allele 1</th>
<th>Base difference counts: [M/N*/L*] M in key exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0/0/0] B*44:02:01:01</td>
<td>I4-92 A/G</td>
</tr>
<tr>
<td>[0/0/1*] B*44:02:01:02S</td>
<td>I3-487 T/C</td>
</tr>
<tr>
<td>[0/0/1*] B*44:02:01:03</td>
<td>I3-487 T/C</td>
</tr>
<tr>
<td>[0/1*/1*] B*44:02:27</td>
<td>I5-915 C/T</td>
</tr>
<tr>
<td>[0/1*/2*] B*44:02:25</td>
<td>I1-12.1 G/. I3-168.1 G/. E4-756 C/T</td>
</tr>
<tr>
<td>[0/1*/2*] B*44:19N</td>
<td>E1-5 G/. I1-12.1 G/. I3-168.1 G/.</td>
</tr>
<tr>
<td>[0/1*/2*] B*44:27:01</td>
<td>I1-12.1 G/. I3-168.1 G/. E4-668 T/C</td>
</tr>
<tr>
<td>[0/1*/2*] B*44:66</td>
<td>I1-12.1 G/. I3-168.1 G/. E4-649 C/A</td>
</tr>
<tr>
<td>[0/1*/2*] B*44:118</td>
<td>I1-12.1 G/. I3-168.1 G/. E4-646 C/G</td>
</tr>
<tr>
<td>[1/0*/0*] B*44:02:17</td>
<td>E3-606 G/C</td>
</tr>
<tr>
<td>[1/0*/0*] B*44:49</td>
<td>E2-97 T/G</td>
</tr>
<tr>
<td>[1/0*/1*] B*44:23N</td>
<td>E3-493 C/T</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:06</td>
<td>I1-12.1 G/. E3-402 C/T I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:07</td>
<td>I1-12.1 G/. E3-369 C/T I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:08</td>
<td>I1-12.1 G/. E3-573 A/G I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:09</td>
<td>I1-12.1 G/. E3-486 C/G I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:10</td>
<td>I1-12.1 G/. E2-141 C/T I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:11</td>
<td>I1-12.1 G/. E3-546 C/T I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:12</td>
<td>I1-12.1 G/. E3-399 C/T I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:13</td>
<td>I1-12.1 G/. E2-201 G/T I3-168.1 G/.</td>
</tr>
<tr>
<td>[1/0*/2*] B*44:02:14</td>
<td>I1-12.1 G/. E3-378 C/T I3-168.1 G/.</td>
</tr>
</tbody>
</table>
HLA-B*44:138Q

log:
[0/0/0] B*44:02:01:01, find allele 1: B*44:02:01:01
[M/N/2*] B*44:138Q, find allele 2: B*44:138Q
Crucial Position in Exon 3
Conclusion

• Sequencing artefacts on NGS devices are different from those of conventional devices
• If whole gene is sequenced, the lack of reference sequences is the largest problem
• Open source software for assignment, alignment does exist, but cannot deal with HLA-specific problems
• Specialised software is being developed