NGS Data analysis

10th INTERNATIONAL SUMMER SCHOOL ON IMMUNOGENETICS

Stintino, Sardinia, 17 September, 2013

Outline

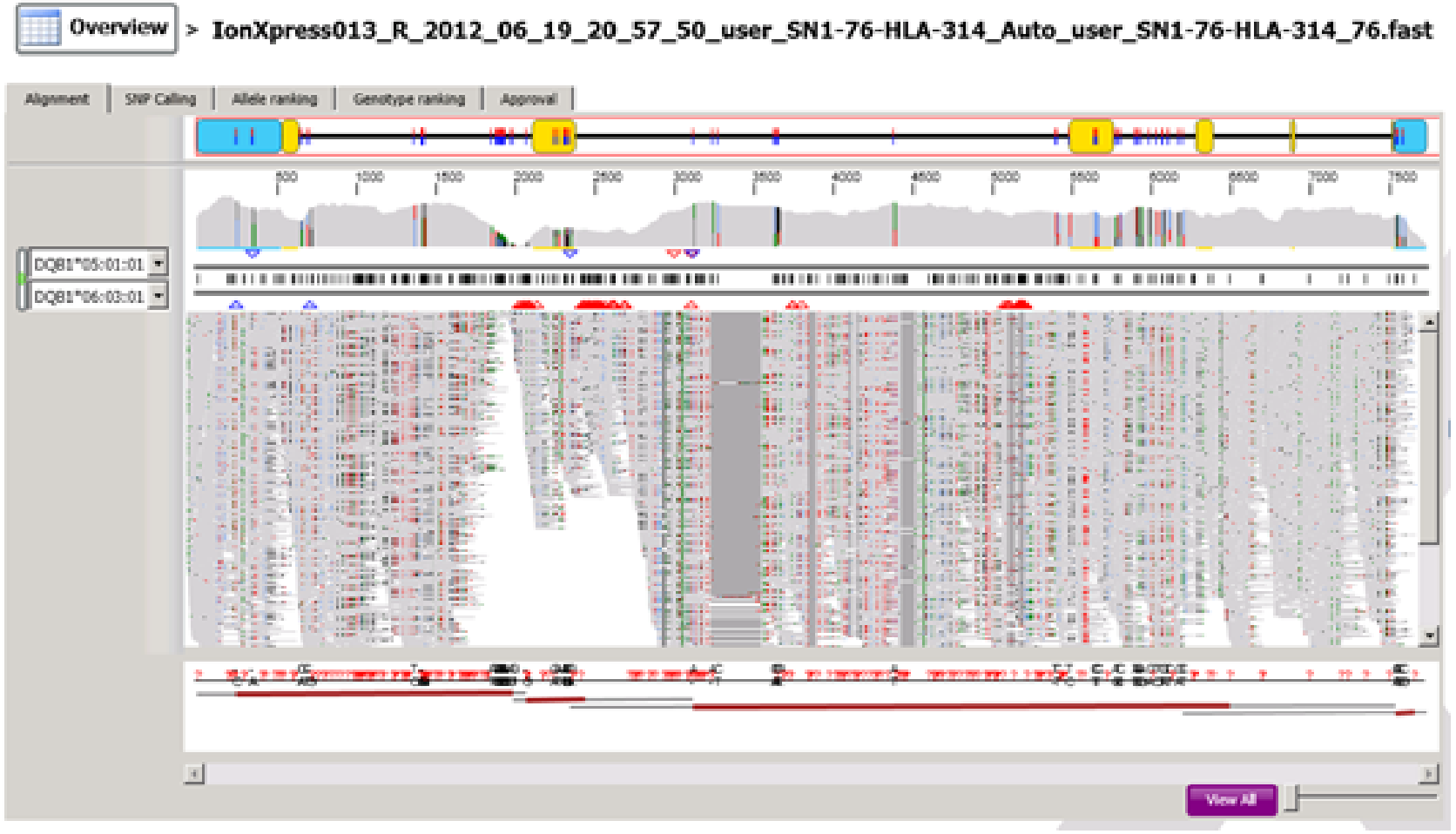
- Pre-Analytical Phase – possible artifacts
- Quality Checks of the device
- Sequencing Artefacts
- Alignments, Assignments, and their main problem

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

Hybrid Molecules

Tissue Antigens 80:1 2012


Overview
> IonXpress013_R_2012_06_19_20_57_50_user_SN1-76-HLA-314_Auto_user_SN1-76-HLA-314_76.fast
Alignment
SNP Calling
Allele ranking
Genotype ranking
Approval
DQB1*05:01:01
DQB1*06:03:01
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
View All

First Analysis

Run Summary: R_2012_07_12_14_19_30_user_SN1-91-HLA_314_OT2_200_20130823

Reports: Auto_user_SN1-91-HLA_314_OT2_200_20130823_122 (135)

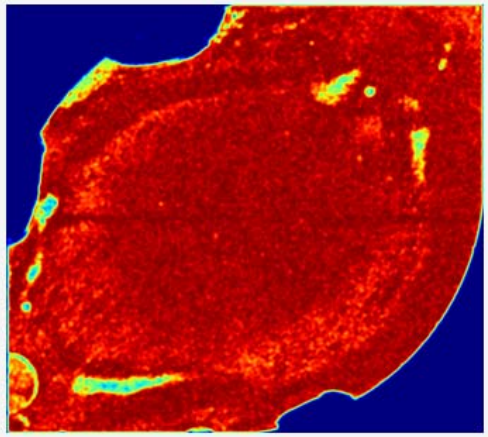
Unaligned

113 M
Total Bases

118
Key Signal

90 %
ISP Loading

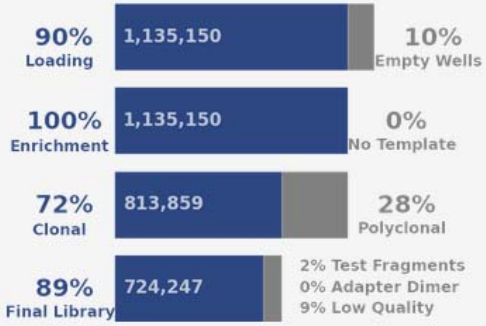
ISP Density



724,247
Total Reads

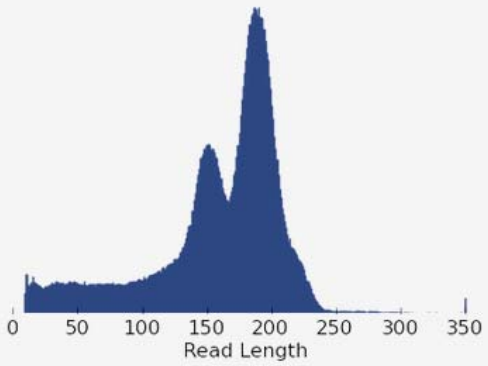
65 %
Usable Reads

ISP Summary



157 bp
Mean Read Length

Read Length



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
- ACCGGGAGACACAGATCTGCAAGGCCAAGGCGACAGACTGACCGAGAGAACCTGCGGATCGCGCTCCGCTACTACAACCAGAGCGAGGCCGGTGAGTGACCCCGGCCCGGGGCGCAGGTCACGACCGCCCATCCACGTACGCGGCGCCCGATC
- +
- @@9>>4;;;45276649/3.307.73;592++,9:8;97<:=;AABBBB>B=@?;;4944188889399=;<8:97396>>>>?@99393929>9430,,00&,&-0&-00&.86893332..-.'-25+6,,(+1011--,2.4047*001+
- @YSEVQ:4:28
- CAAGGCCAAGGCACAGACTGACCGAGAGGACCTGCGGATCGCGCTCCGCTACCTACAACCAGAGCGAGGCCGGTCGAGTCGACCCGGCCCGGGGCGCCAGGTCA
- +
- DE>C>C>E@E@CCCECCA@@=?8=DDDD<DE?DDC@:@...7987:4777-.',742/74,,+65033&0(+/2063545989=1:589+166(043.00+00+

PHRED

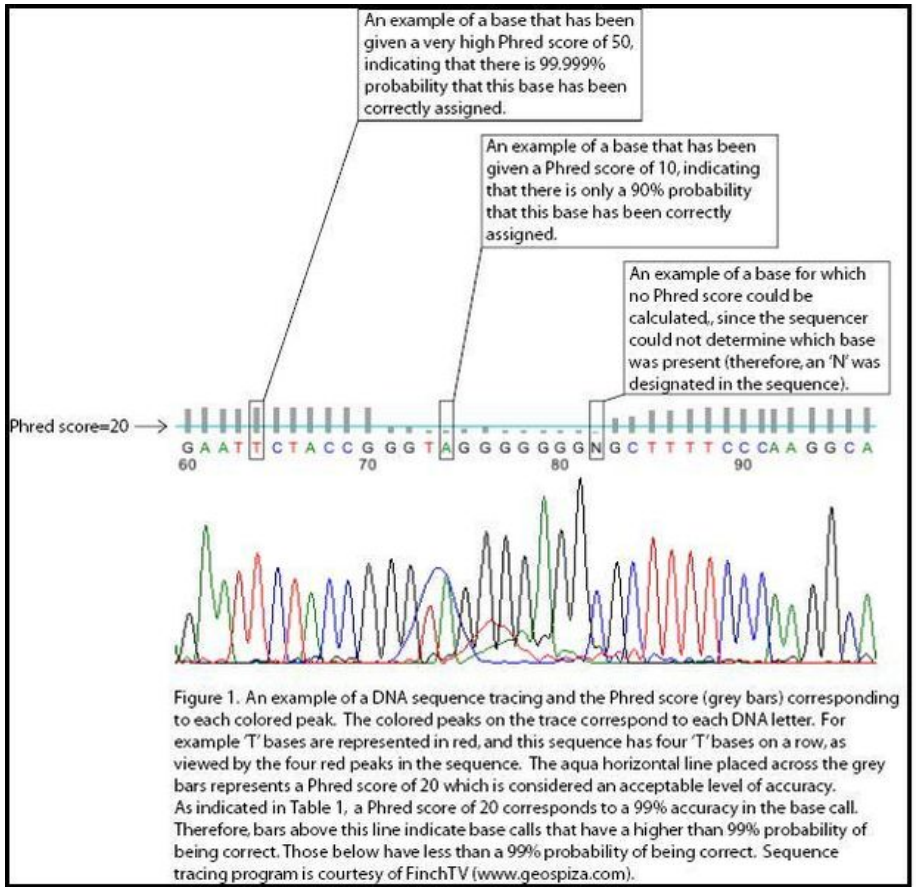
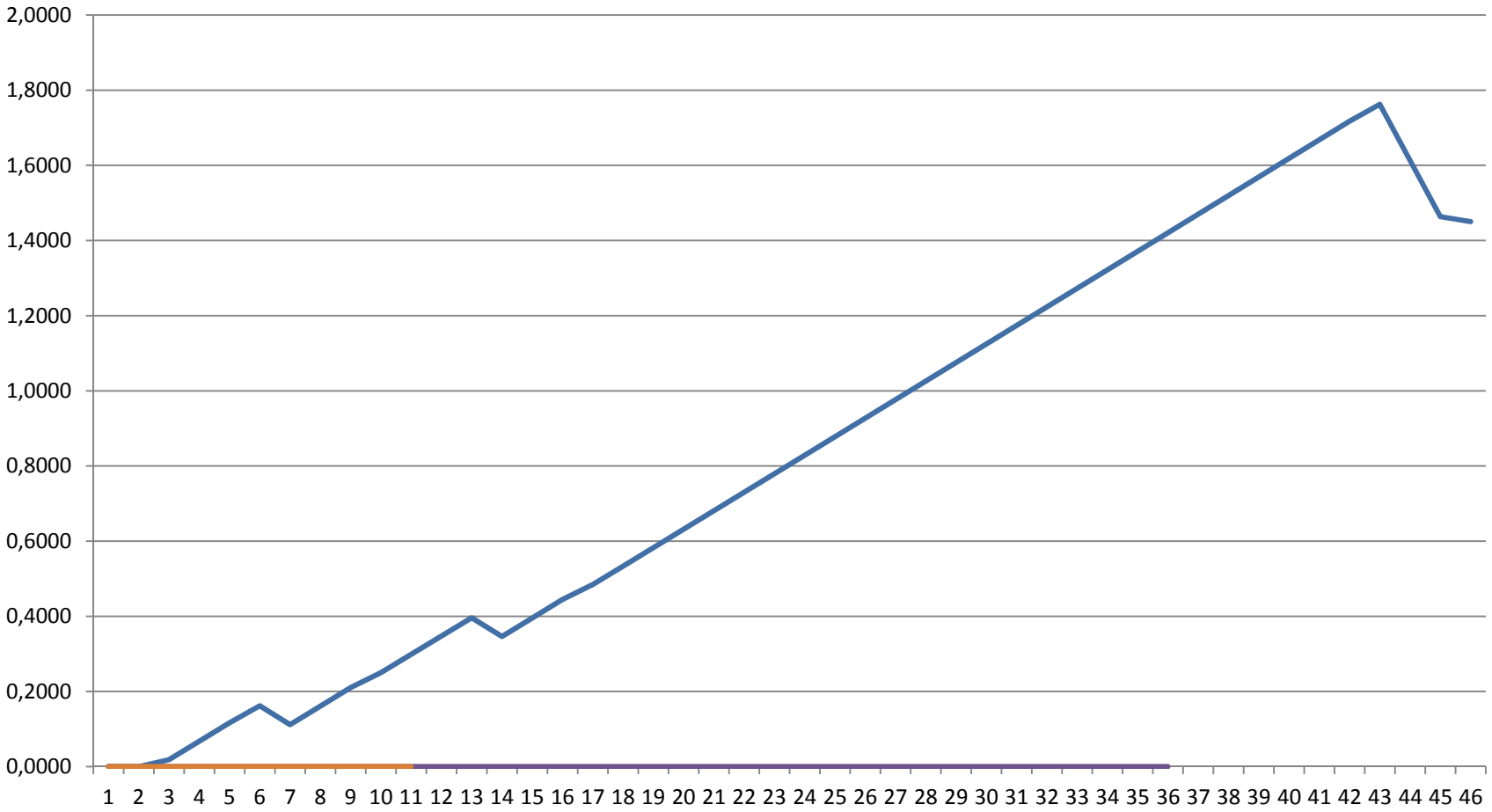


Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks 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

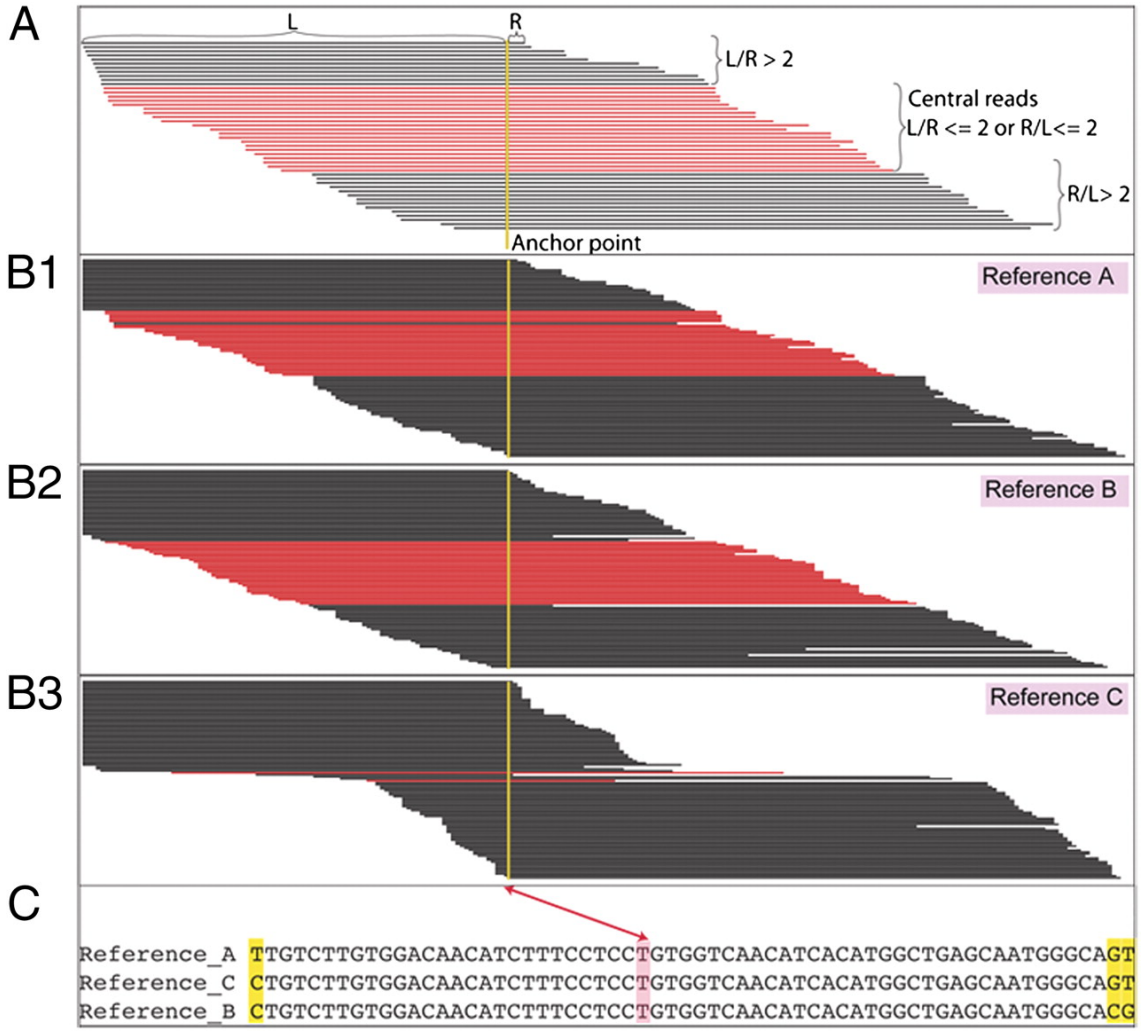
```
Ref: CCCCGA
 #1: CCTCGA    Mutation
 #2: CCC.GA    Deletetion
 #3: CCCCCGA   Insertion
```

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.



Wang C et al. PNAS 2012;109:8676-8681

IMGT/HLA Database ver 12.0

- cDNA sequences n=9291
- gDNA sequences n=640

IMGT/HLA ver 12.0, n Sequences

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

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

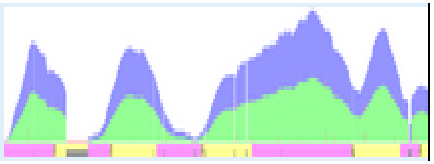
Coverage and Statistics
HLA-B*44:02:01:01 and B*44:138Q

Each Allele Summary	Allele 1	Allele 2
Total Read Count:	76681	74946
Cross Allele Read Count:	48373	
Filtered Read Count:	10050	10418
Forward Read Count:	32689	31158
Averaged Forward Read Length:	147	146
Reverse Read Count:	33942	33370
Averaged Reverse Read Length:	146	146

Software Parameters	Setting
Mismatch Limit	5
Insertion Limit	3
Deletion Limit	3
Minimal Read Length	100
Minimal Read Depth	20
Background Ratio Cutoff (%)	20

Allele1	Base Count	Valid Bases	Min. Depth	Max. Depth	Avg. Depth	High Bkg Count (>20%)	Mean Accuracy %
UTR	284	273	0	4552	2883	3	99.1
E-1	73	72	0	3257	2801	0	99.6
I-1	129	128	4	1928	668	0	99.2
E-2	270	270	1955	4308	3586	0	99.7
I-2	243	243	183	2638	778	1	99.3
E-3	276	276	495	3761	2568	2	99.3
I-3	575	575	2655	5937	4650	2	99.4
E-4	276	276	2258	5141	3747	5	99.1
I-4	93	93	1788	2819	2389	3	98.4
E-5	117	117	2002	2862	2474	0	99.8
I-5	441	441	1684	4441	3391	4	99.4
E-6	33	33	2257	2577	2387	0	99.6
I-6	106	106	1255	2206	1692	0	99.5
E-7	44	44	905	1248	1076	0	99.7
UTR	363	351	0	4702	2192	0	99.5
All	3323	3298	0	5937	2947	20	99.4

Allele2	Base Count	Valid Bases	Min. Depth	Max. Depth	Avg. Depth	High Bkg Count (>20%)	Mean Accuracy %
UTR	284	264	0	4452	2762	3	99.2
E-1	73	72	0	3257	2801	0	99.6
I-1	129	128	4	1928	668	0	99.2
E-2	270	270	1955	4306	3586	0	99.7
I-2	243	243	151	2644	769	2	99.2
E-3	273	273	361	3767	2571	2	99.3
I-3	575	575	2528	5937	4639	2	99.4
E-4	276	276	2068	5014	3581	5	99.1
I-4	104	104	1609	2352	1901	1	98.7
E-5	117	117	1753	2234	2018	2	99.3
I-5	441	441	1543	4037	3116	2	99.5
E-6	33	33	1934	2346	2112	0	99.6
I-6	106	106	1046	1911	1414	0	99.5
E-7	44	44	795	1044	917	0	99.7
UTR	363	342	0	4265	1999	0	99.7
All	3331	3288	0	5937	2816	19	99.4

[0/0/0] B*44:02:01:01	1	[0/0/0*] B*44:138Q	1	IonXpress_012	11874	

HLA-B*44:02:01:01

Coverage and Statistics | Match List | Tools and Links

Allele 1	Base difference counts: [M/N*/L*] M in key exor
[0/0/0] B*44:02:01:01	
[0/0/1*] B*44:02:01:02S	I4-92 A/G
[0/0/1*] B*44:02:01:03	I3-487 T/C
[0/1*/1*] B*44:02:27	I3-487 T/C E5-915 C/T
[0/1*/2*] B*44:02:25	I1-12.1 G/. I3-168.1 G/. E4-756 C/T
[0/1*/2*] B*44:19N	E1-5 G/. I1-12.1 G/. I3-168.1 G/.
[0/1*/2*] B*44:27:01	I1-12.1 G/. I3-168.1 G/. E4-668 T/C
[0/1*/2*] B*44:66	I1-12.1 G/. I3-168.1 G/. E4-649 C/A
[0/1*/2*] B*44:118	I1-12.1 G/. I3-168.1 G/. E4-646 C/G
[1/0*/0*] B*44:02:17	E3-606 G/C
[1/0*/0*] B*44:49	E2-97 T/G
[1/0*/1*] B*44:23N	E3-493 C/T I3-218 G/T
[1/0*/2*] B*44:02:05	I1-12.1 G/. E2-285 A/G I3-168.1 G/.
[1/0*/2*] B*44:02:06	I1-12.1 G/. E3-402 C/T I3-168.1 G/.
[1/0*/2*] B*44:02:07	I1-12.1 G/. E3-369 C/T I3-168.1 G/.
[1/0*/2*] B*44:02:08	I1-12.1 G/. E3-573 G/A I3-168.1 G/.
[1/0*/2*] B*44:02:09	I1-12.1 G/. E3-486 C/G I3-168.1 G/.
[1/0*/2*] B*44:02:10	I1-12.1 G/. E2-141 C/T I3-168.1 G/.
[1/0*/2*] B*44:02:11	I1-12.1 G/. E3-546 C/T I3-168.1 G/.
[1/0*/2*] B*44:02:12	I1-12.1 G/. E3-399 C/T I3-168.1 G/.
[1/0*/2*] B*44:02:13	I1-12.1 G/. E2-201 G/T I3-168.1 G/.
[1/0*/2*] B*44:02:14	I1-12.1 G/. E3-378 C/T I3-168.1 G/.

HLA-B*44:138Q

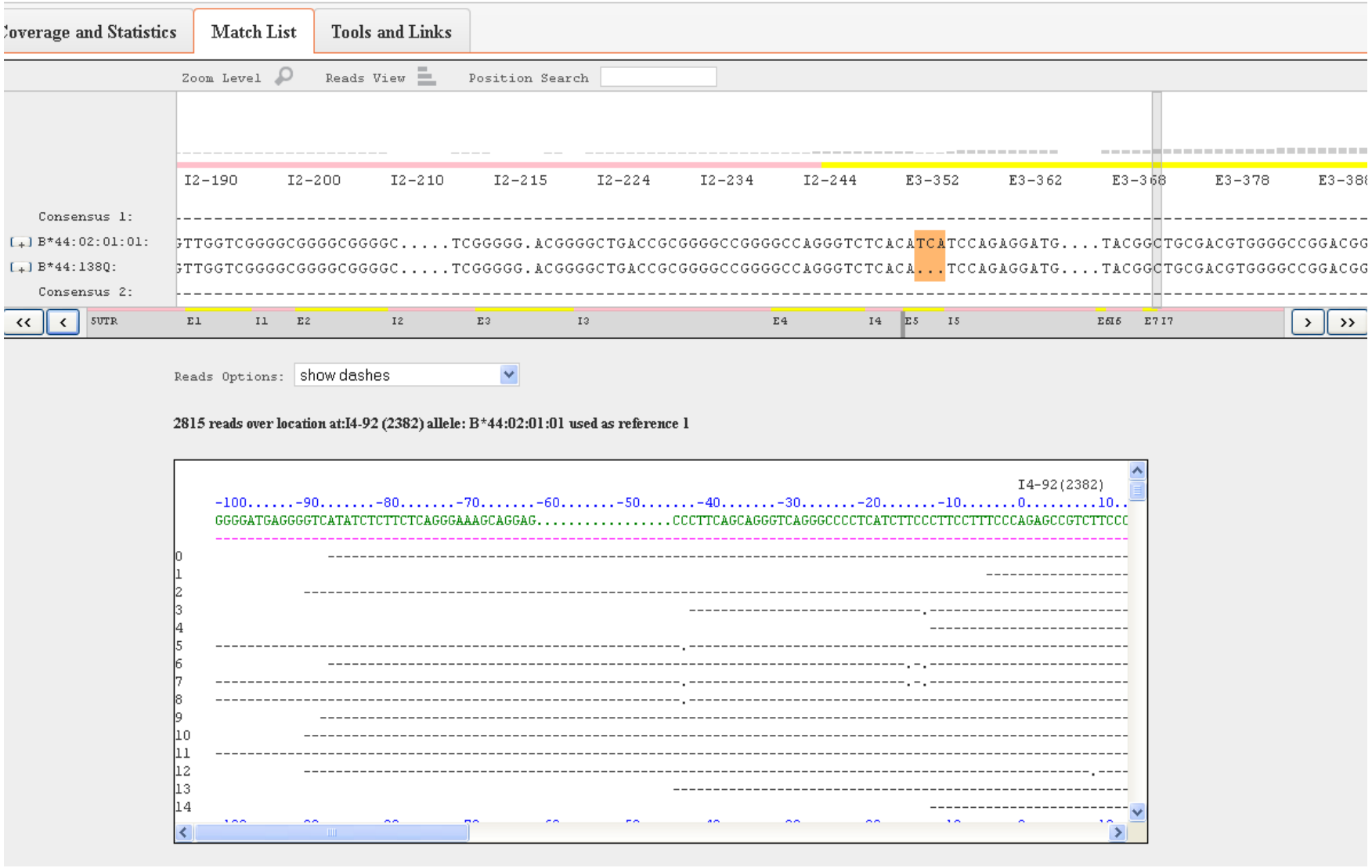
Allele 2	Base difference counts: [M/N*/L*]
[0/0/0*] B*44:138Q	

log:

[0/0/0] B*44:02:01:01 , find allele 1: B*44:02:01:01

[0/0/2*] 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