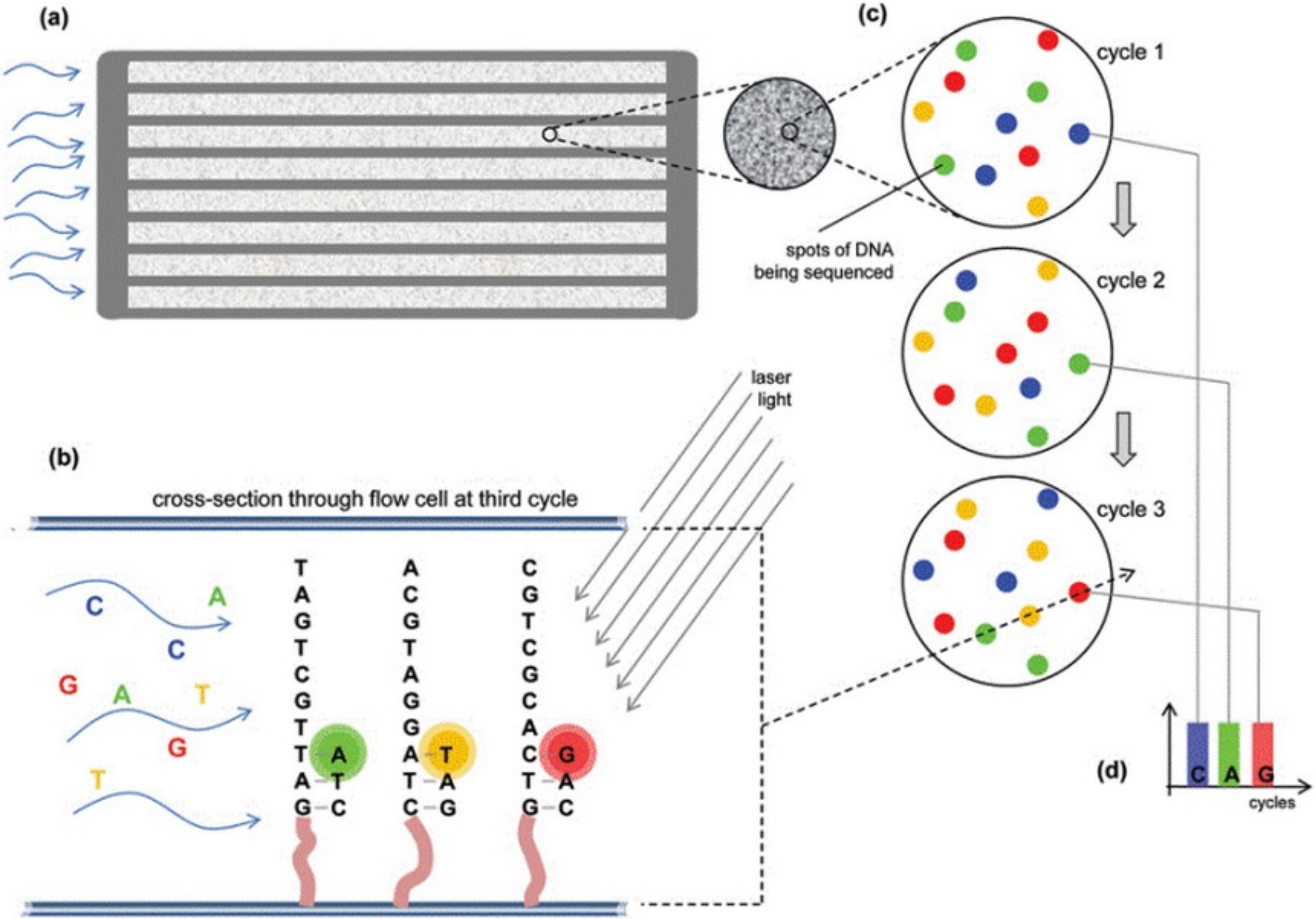
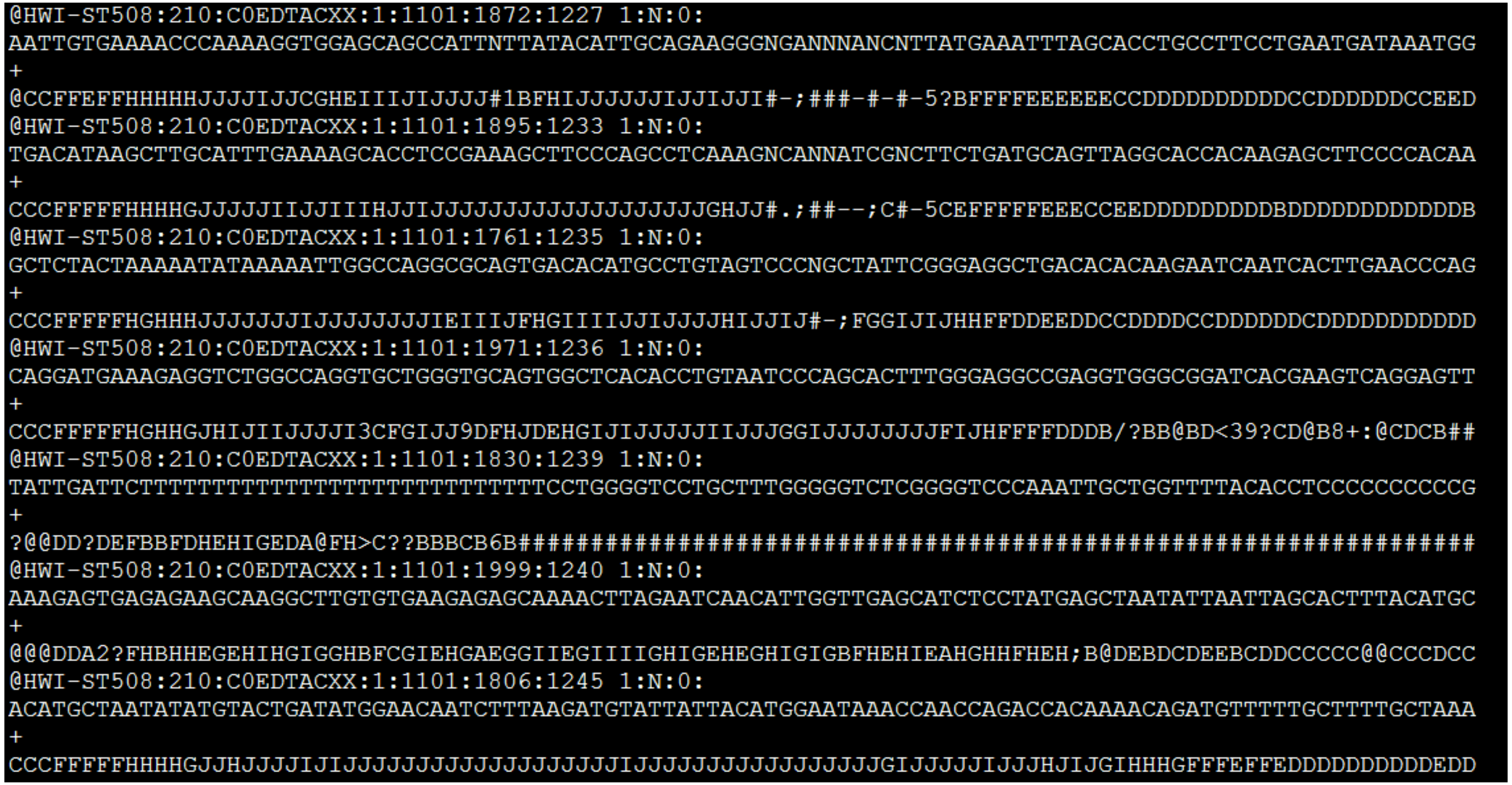
Sequencing by synthesis



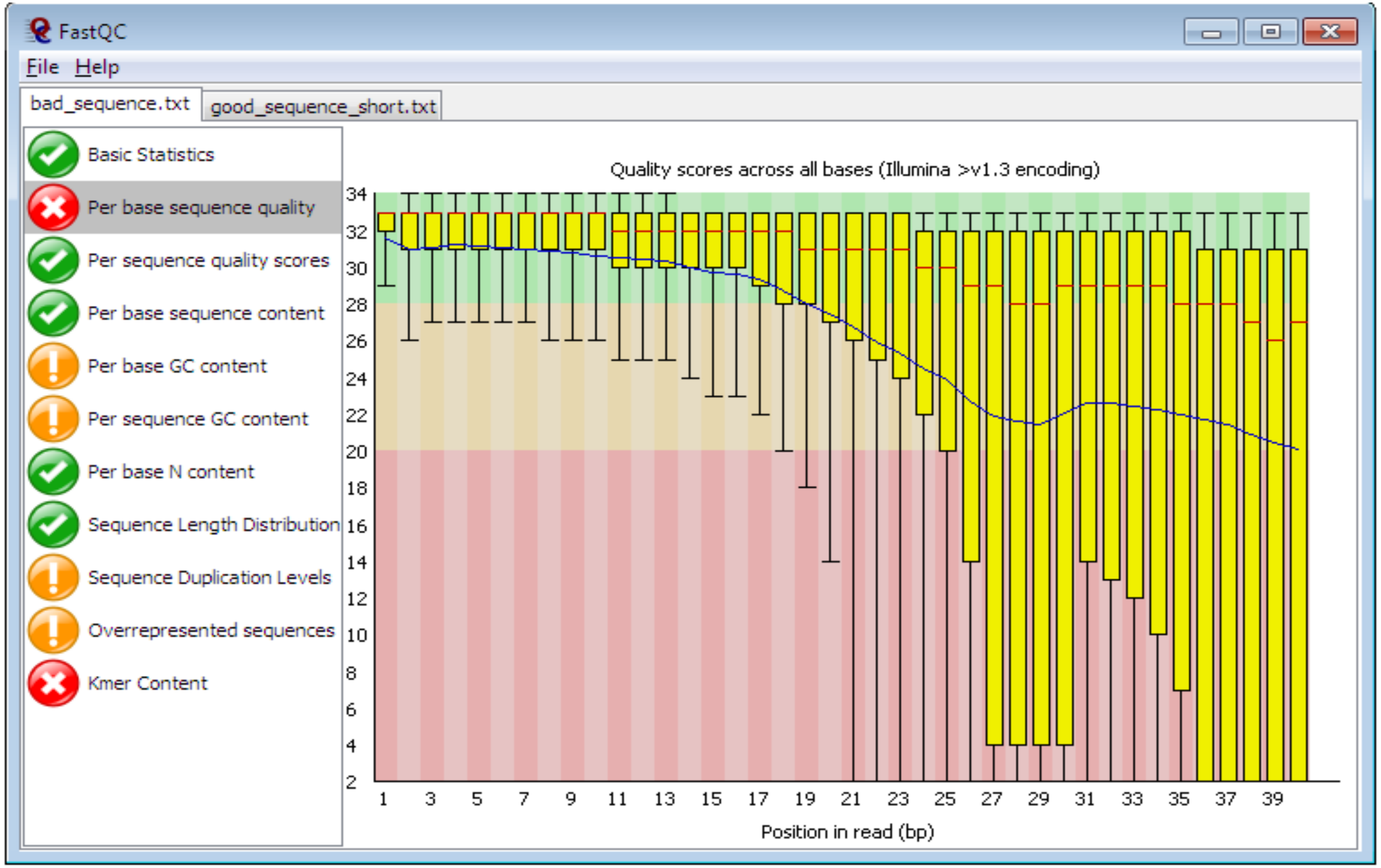
FASTQ files

Line1: Sequence identifier
Line2: Raw sequence
Line3: meaningless
Line4: quality values for the sequence



```
@HWI-ST508:210:C0EDTACXX:1:1101:1872:1227 1:N:0:
AATTGTGAAAACCCAAAAGGTGGAGCAGCCATTNTTATACATTGCAGAAGGGNGANNNANCNTTATGAAATTTAGCACCTGCCTTCCTGAATGATAAATGG
+
@CCFFEFFHHHHHJJJJIJJCGHEIIIJIJJJJ#1BFHIJJJJJJIJJIJJI#-;###-#-#-5?BFFFFEEEEEECCDDDDDDDDDDCCDDDDDDCCEED
@HWI-ST508:210:C0EDTACXX:1:1101:1895:1233 1:N:0:
TGACATAAGCTTGCATTTGAAAAGCACCTCCGAAAGCTTCCCAGCCTCAAAGNCANNATCGNCTTCTGATGCAGTTAGGCACCACAAGAGCTTCCCCACAA
+
CCCFFFFFHHHHGJJJJJIIJJIIIHJJIJJJJJJJJJJJJJJJJJJJGHJJ#.;##--;C#-5CEFFFFFEEECCEEDDDDDDDDDBDDDDDDDDDDDDB
@HWI-ST508:210:C0EDTACXX:1:1101:1761:1235 1:N:0:
GCTCTACTAAAAATATAAAAATTGGCCAGGCGCAGTGACACATGCCTGTAGTCCCNGCTATTCGGGAGGCTGACACACAAGAATCAATCACTTGAACCCAG
+
CCCFFFFFHGHHHJJJJJJIJJJJJJJIEIIIJFHGIIIIJJIJJJJHIJJIJ#-;FGGIJIJHHFFDDEEDDCCDDDDCCDDDDDDCDDDDDDDDDDD
@HWI-ST508:210:C0EDTACXX:1:1101:1971:1236 1:N:0:
CAGGATGAAAGAGGTCTGGCCAGGTGCTGGGTGCAGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGCCGAGGTGGGCGGATCACGAAGTCAGGAGTT
+
CCCFFFFFHGHHGJHIJIIJJJJI3CFGIJJ9DFHJDEHGIJIJJJJJIIJJJGGIJJJJJJJJFIJHFFFFDDDB/?BB@BD<39?CD@B8+:@CDCB##
@HWI-ST508:210:C0EDTACXX:1:1101:1830:1239 1:N:0:
TATTGATTCTTTTTTTTTTTTTTTTTTTTTTTTTTTCCTGGGGTCCTGCTTTGGGGGTCTCGGGGTCCCAAATTGCTGGTTTTACACCTCCCCCCCCCG
+
?@@DD?DEFBBFDHEHIGEDA@FH>C??BBBCB6B##################################################################
@HWI-ST508:210:C0EDTACXX:1:1101:1999:1240 1:N:0:
AAAGAGTGAGAGAAGCAAGGCTTGTGTGAAGAGAGCAAAACTTAGAATCAACATTGGTTGAGCATCTCCTATGAGCTAATATTAATTAGCACTTTACATGC
+
@@@DDA2?FHBHHEGEHIHGIGGHBFCGIEHGAEGGIIEGIIIIGHIGEHEGHIGIGBFHEHIEAHGHHFHEH;B@DEBDCDEEBCDDCCCCC@@CCCDCC
@HWI-ST508:210:C0EDTACXX:1:1101:1806:1245 1:N:0:
ACATGCTAATATATGTACTGATATGGAACAATCTTTAAGATGTATTATTACATGGAATAAACCAACCAGACCACAAAACAGATGTTTTTGCTTTTGCTAAA
+
CCCFFFFFHHHHGJJHJJJJIJIJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJJJJJJGIJJJJJIJJJHJIJGIHHHGFFFEFFEDDDDDDDDDDEDD
```

FastQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Position in read (bp)

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

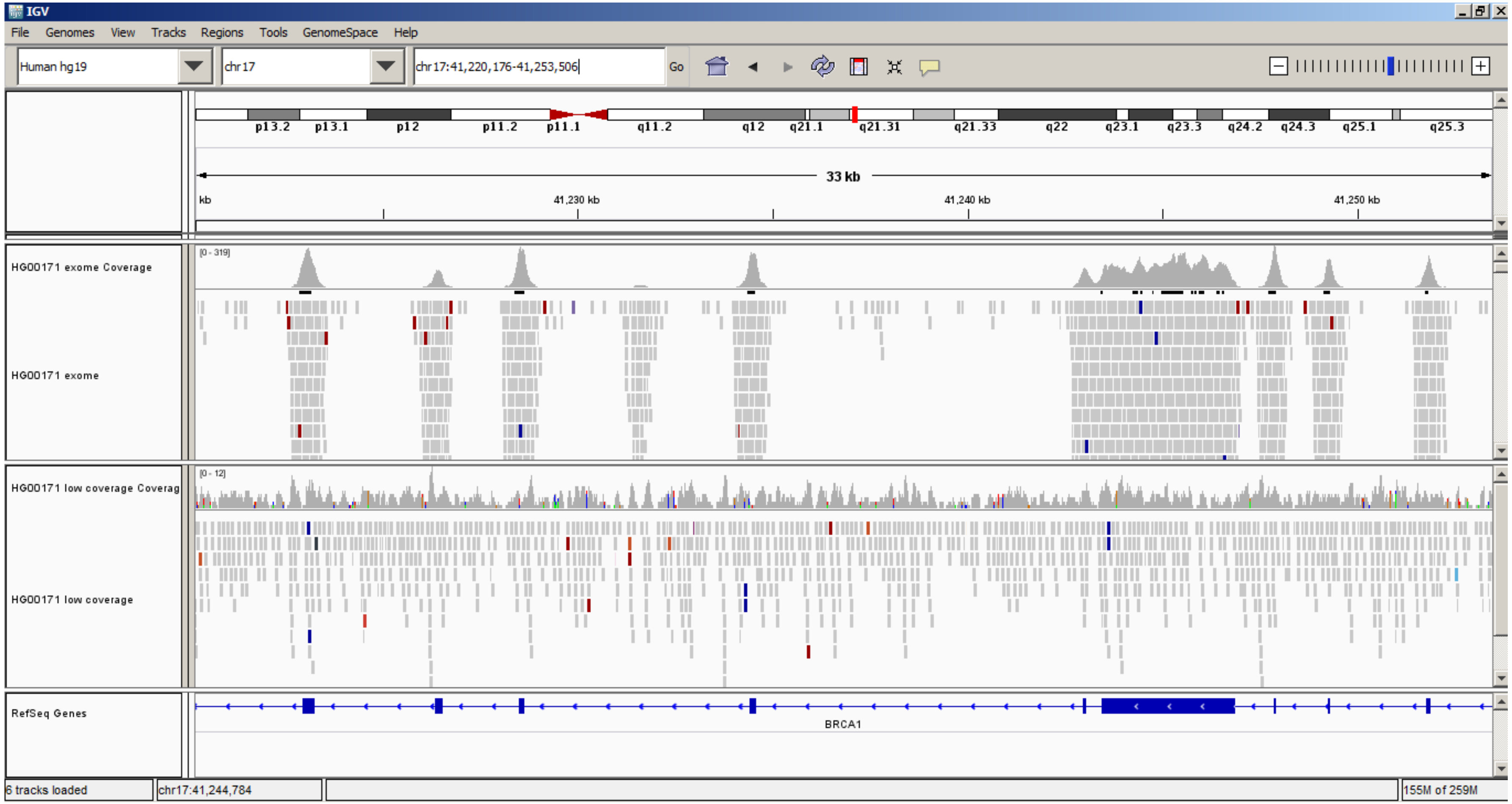
SAM format

- Standard format for short-read alignements.

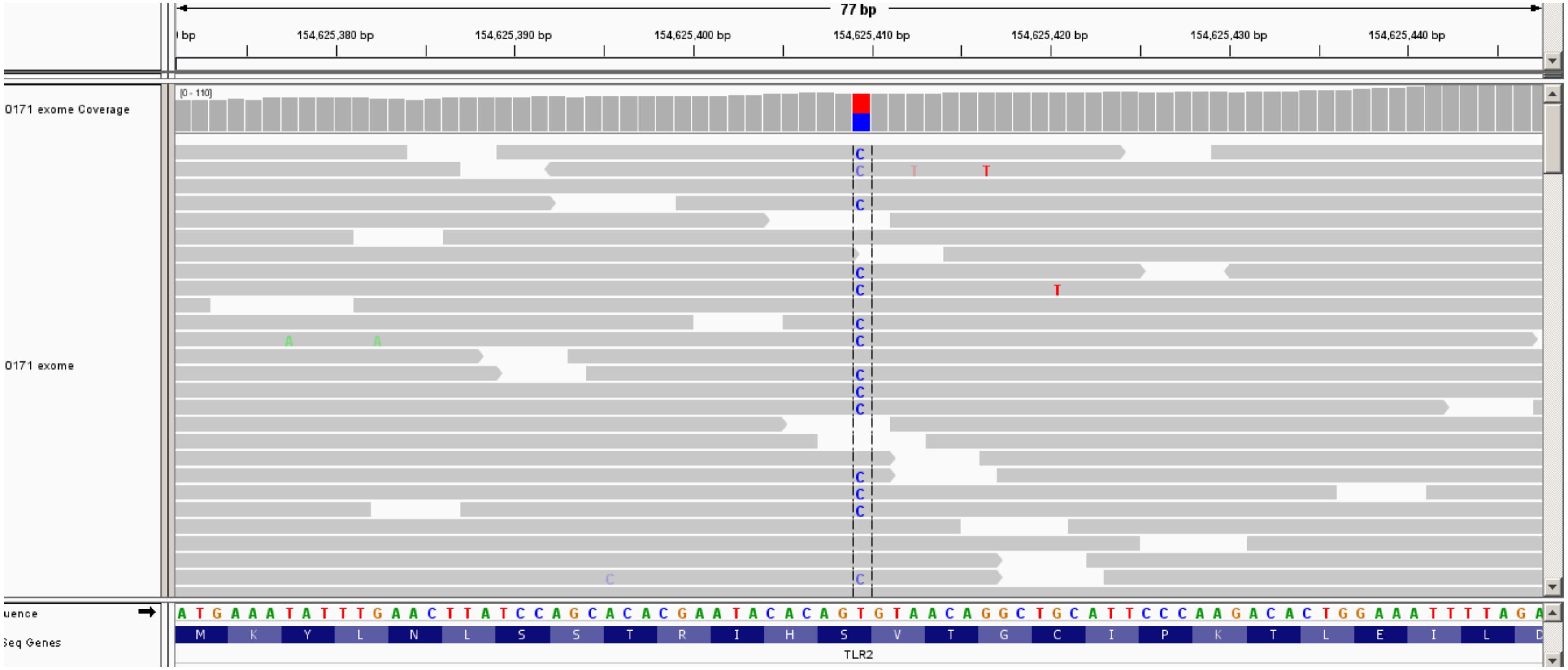
```
 HWI-ST151_106137860:1:67:20248:73945#0    129      chr17    98508    255      40M      =
 
        98849    378      AGGGGTTGGCGGGGCAAGGTGGCTCACGCCTGTCATCCCA

        @B@B@:8A?8>@@80DCCDA@85,C7>7>>AB########
```

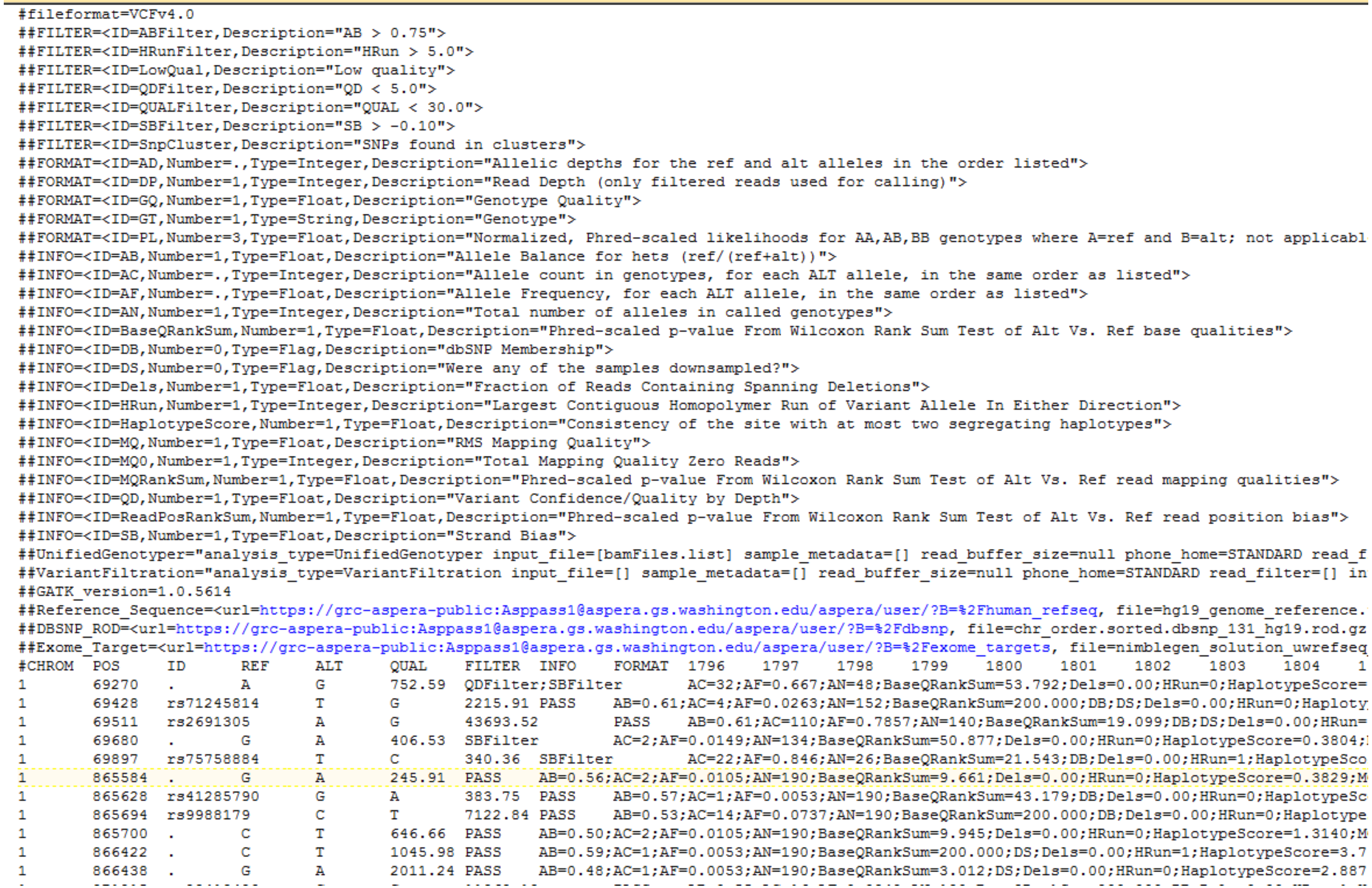
Visualization



Visualization


77 bp
bp
154,625,380 bp
154,625,390 bp
154,625,400 bp
154,625,410 bp
154,625,420 bp
154,625,430 bp
154,625,440 bp
0171 exome Coverage
[0 - 110]
0171 exome
uence
A T G A A A T A T T T G A A C T T A T C C A G C A C A C G A A T A C A C A G T G T A A C A G G C T G C A T T C C C A A G A C A C T G G A A A T T T T A G A
Seq Genes
M K Y L N L S S T R I H S V T G C I P K T L E I L D
TLR2

VCF format



```
#fileformat=VCFv4.0
##FILTER=<ID=ABFilter,Description="AB > 0.75">
##FILTER=<ID=HRunFilter,Description="HRun > 5.0">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=QDFilter,Description="QD < 5.0">
##FILTER=<ID=QUALFilter,Description="QUAL < 30.0">
##FILTER=<ID=SBFilter,Description="SB > -0.10">
##FILTER=<ID=SnpCluster,Description="SNPs found in clusters">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered reads used for calling)">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=3,Type=Float,Description="Normalized, Phred-scaled likelihoods for AA,AB,BB genotypes where A=ref and B=alt; not applicabl
##INFO=<ID=AB,Number=1,Type=Float,Description="Allele Balance for hets (ref/(ref+alt))">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Phred-scaled p-value From Wilcoxon Rank Sum Test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=HRun,Number=1,Type=Integer,Description="Largest Contiguous Homopolymer Run of Variant Allele In Either Direction">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Phred-scaled p-value From Wilcoxon Rank Sum Test of Alt Vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Phred-scaled p-value From Wilcoxon Rank Sum Test of Alt Vs. Ref read position bias">
##INFO=<ID=SB,Number=1,Type=Float,Description="Strand Bias">
##UnifiedGenotyper="analysis_type=UnifiedGenotyper input_file=[bamFiles.list] sample_metadata=[] read_buffer_size=null phone_home=STANDARD read_f
##VariantFiltration="analysis_type=VariantFiltration input_file=[] sample_metadata=[] read_buffer_size=null phone_home=STANDARD read_filter=[] in
##GATK_version=1.0.5614
##Reference_Sequence=<url=https://grc-aspera-public:Asppass1@aspera.gs.washington.edu/aspera/user/?B=%2Fhuman_refseq, file=hg19_genome_reference.
##DBSNP_ROD=<url=https://grc-aspera-public:Asppass1@aspera.gs.washington.edu/aspera/user/?B=%2Fdbsnp, file=chr_order.sorted.dbsnp_131_hg19.rod.gz
##Exome_Target=<url=https://grc-aspera-public:Asppass1@aspera.gs.washington.edu/aspera/user/?B=%2Fexome_targets, file=nimblegen_solution_uwrefseq
#CHROM  POS     ID          REF     ALT     QUAL    FILTER  INFO    FORMAT  1796    1797    1798    1799    1800    1801    1802    1803    1804    1
1       69270   .           A       G       752.59  QDFilter;SBFilter       AC=32;AF=0.667;AN=48;BaseQRankSum=53.792;Dels=0.00;HRun=0;HaplotypeScore=
1       69428   rs71245814  T       G       2215.91 PASS    AB=0.61;AC=4;AF=0.0263;AN=152;BaseQRankSum=200.000;DB;DS;Dels=0.00;HRun=0;Haploty
1       69511   rs2691305   A       G       43693.52        PASS    AB=0.61;AC=110;AF=0.7857;AN=140;BaseQRankSum=19.099;DB;DS;Dels=0.00;HRun=
1       69680   .           G       A       406.53  SBFilter        AC=2;AF=0.0149;AN=134;BaseQRankSum=50.877;Dels=0.00;HRun=0;HaplotypeScore=0.3804;
1       69897   rs75758884  T       C       340.36  SBFilter        AC=22;AF=0.846;AN=26;BaseQRankSum=21.543;DB;Dels=0.00;HRun=1;HaplotypeSco
1       865584  .           G       A       245.91  PASS    AB=0.56;AC=2;AF=0.0105;AN=190;BaseQRankSum=9.661;Dels=0.00;HRun=0;HaplotypeScore=0.3829;M
1       865628  rs41285790  G       A       383.75  PASS    AB=0.57;AC=1;AF=0.0053;AN=190;BaseQRankSum=43.179;DB;Dels=0.00;HRun=0;HaplotypeSc
1       865694  rs9988179   C       T       7122.84 PASS    AB=0.53;AC=14;AF=0.0737;AN=190;BaseQRankSum=200.000;DB;Dels=0.00;HRun=0;Haplotype
1       865700  .           C       T       646.66  PASS    AB=0.50;AC=2;AF=0.0105;AN=190;BaseQRankSum=9.945;Dels=0.00;HRun=0;HaplotypeScore=1.3140;M
1       866422  .           C       T       1045.98 PASS    AB=0.59;AC=1;AF=0.0053;AN=190;BaseQRankSum=200.000;DS;Dels=0.00;HRun=1;HaplotypeScore=3.7
1       866438  .           G       A       2011.24 PASS    AB=0.48;AC=1;AF=0.0053;AN=190;BaseQRankSum=3.012;DS;Dels=0.00;HRun=0;HaplotypeScore=2.887
```

VCF

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS  ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:DP:GQ   0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:DP:GQ   0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:DP:GQ   0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:DP:GQ   1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

Primary, secondary and tertiary analysis

Primary analysis

Happens in the instrument, checks during the run for quality control.
Consists of image alignment, color and quality value calls. **FASTQ files**

Secondary analysis

A reference sequence is converted into color space and the data are aligned to the reference sequence.
A consensus sequence may then be constructed from the sequencing reads.
Comparison of the consensus sequence to a reference genome enables the identification of SNPs and structural variations. **SAM/BAM files**

Tertiary analysis

Data analysis that takes place after the reads are mapped.
Visualization of data. Annotation of variants. **VCF files**