

DATA FORMATS & METHODS IN VARIANT CALLING

Javier Perales-Patón
jperales@cniio.es



Bioinformatics Unit
CNIO. Madrid, Spain.


Fátima Al-Shahrour
[\[falshahrour@cniio.es\]](mailto:falshahrour@cniio.es)
Elena Piñeiro-Yáñez
[\[epineiro@cniio.es\]](mailto:epineiro@cniio.es)
Pedro Fernandes
[\[pfern@igc.gulbenkian.pt\]](mailto:pfern@igc.gulbenkian.pt)



Today

09:30 - 10:00	Introduction to the course and self presentation of the participants. Personalized medicine.
11:30 - 12:30	NGS I : Variant detection.
14:00 - 16:00	Playing with the data and the methods.
16:30 - 18:00	Practical : Running the pipeline.

Data formats cheat sheet

Format	Uses	Example	File type	Software Management	File Extension
Fasta	Human genome Define biological sequences (DNA, RNA, cDNA, proteins).	human_genome.fa	Plain text	samtools, picard-tools	.fa; .fasta
FastQ	Raw sequencing data Single-end sequencing → 1 file Paired-end sequencing → 2 files (R1 and R2 for each end, respectively)	DNAseq_raw_data.fastq (DNAseq_R1.fastq and DNAseq_R2.fastq)	Plain text	samtools, picard-tools Aligners	.fq; .fastq
SAM	Define read alignments. Store alignment meta-info (reference, methods, one- or multi-sample).	mapped_reads.sam	Plain text	samtools, picard-tools	.sam
BAM	VISUALIZE ALIGNMENTS (IGV) The same as SAM, but compressed and indexed. Also to store UNMAPPED reads (compressed).	mapped_reads.bam unmapped_reads.bam	Binary	samtools, picard-tools, IGV (Integrative Genome Viewer)	.bam
VCF 	SNV & Indels calls Indicate genomic variations. Store Variant calling meta-info (reference, methods, one- or multi-sample).	detected_pointvariants.vcf	Plain text	bcftools, vcftools, Unix	.vcf
BED	Intervals Delimit genomic regions (i.e. intervals) w or w/o annotations.	targeted_regions.bed intervals.bed	Plain text	bedtools, Unix GATK , picard-tools	.bed
TSV or CSV	Create data matrix (rows X Columns)	CONTRA_output.tsv annotated_variants.tsv	Plain text	Unix, Microsoft Excel, OpenOffice	.tsv; .csv; .txt
XML	RUBioSeq configuration file (internal) Define software internal configuration.	config.xml	Plain	Unix, Firefox	.xml

FASTA & FASTQ formats

- FASTA format: simple sequences

Each sequence is composed by at least two consecutive lines.

- ">" Sequence name
- Multiple lines with the whole sequence.

Possible chars: DNA → ATCGatcgNRY...,
Protein → 1 letter amino acid code

https://en.wikipedia.org/wiki/FASTA_format

```
>DNA_SEQUENCE_1
NNNNNNCTCTGGGGGACAGAACCCTATGGTGGCCCCGGCTCCTCCCAGTATCCAGTCCT
CCGTGAAGATGGAGCCATATTCC
```

60 chars

- FASTQ format: raw sequencing data.

Each sequence read is composed by 4 lines:

- "@" read name
- Sequence
- "+" (optionally: repeat the read name)
- Base Quality Score: Phred scale (0,40)
transformed in ASCII chars.

1 unique sample could have:

If Single-end Seq → 1 file (suffix ".fastq")

If Paired-end Seq → 2 files (suffixes "_R1.fastq" "_R2.fastq")

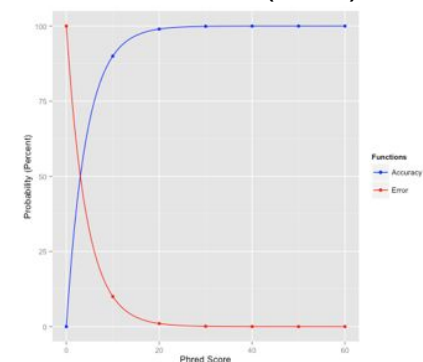
_R1

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTAAATGGTGCCATGCTAGGAGGCCGTGCCCTTCTTGAAAGTTGTATGTGAA
+
BBBFFFFFFBFFFIIIIIFI<FFIIIIIFIIIIIFBFIIIIIIIFFFFIIIIIFI
```

_R2

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 2:N:0:12
CATTTTCGACGTGTGTTAATAAGCTCTGCGTACTTGCAAGCTATCTGCGCGAACG
+
BBBFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFF
```

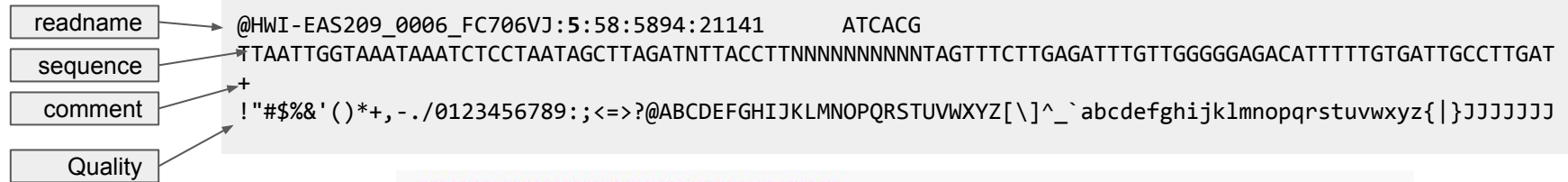
Phred Score (0,40)



https://en.wikipedia.org/wiki/FASTQ_format

FASTQ format in detail

Just to show how one single read is: 4 consecutive lines.



Encoding:

Different platforms
have been using
different encodings.

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
!"#$%&'()*+,-./0123456789;<=>?@ABCDEFGHIJKLMN
OPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                                     |
33                               59   64   73                               104       126
0.....26...31.....40
      -5....0.....9.....40
          0.....9.....40
              3.....9.....40
0.2.....26...31.....41
```

```
S - Sanger      Phred+33, raw reads typically (0, 40)
X - Solexa     Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)
```

WARNING: There are many codifications for the **Phred Score**, which overlap each other. Therefore, you must know the actual codification.

https://en.wikipedia.org/wiki/FASTQ_format

FASTA & FASTQ formats

```
@9VMIV:01561:02299
TGGAGGAGTTGAAGTTTGGAGGAGGGGAGAGTGGTACTCCCTGTCGTCAACTCCTCAGCAGAAGCTGATAGTTGTTGGGCTGCAGCAAGGGCTGGGCGATATCCCCTTGCTCGCAGC
+
@C?DA?B@ABCCB?ACAC>A@?B?==>C@C@A?AA@B@?>BC@?9A@EC@C@>B@D@?E@AB@BAAB@?@CCACC>?<AB>D@?DB@>?>DD??=AC?A?>>>CCC>A@8;>;
@9VMIV:00358:01397
GCTGCGAGCAAGGGGATATCGCCACGCCCTTGCTGCAGCCCAACAACATCAGTTCGTCTGCTGAGGAGTTGACGACAGGGAGTACCACTCTCCCCCTCTCCAAACTTCAACTCTCCAGACCAAGTGAAGGAGTACACT
+
A>BB? :A@>CB@>>>CA@A? :>;B@>>>BBC>AA=D@>?>DB?EC@CAC@B?@B@BC>BBA@?C@C@CBC@9B?C?=?@C@C@?;B>?@A@>>>A?>C@>DCA?A?B=====9;;:A@A@;B@QAA@>?CB?@47821/14:0403
@9VMIV:01056:02786
GCTGCGAGCAAGGGGATATCGCCACGCCCTTGCTGCAGCCCAACAACATCAGTTCGTCTGCTGAGGAGTTGACGACAGGGAGTACCACTCTCCCCCTCTCCAAACTTCAACTCTCTCCC
+
A>DD?<B@?B==<<<@A@A@:>>=B@>>>BBC>BC>E@>?>DB@DB@CAC@E@AC@BC?CCC@?B@CBA?<A?B@=?B@?@B?<@<==7==?:?:@:<==B==@;;<8
@9VMIV:03590:01718
GCTGCGAGCAAGGGGATATCGCCACGCCCTTGCTGCAGCCCAACAACATCAGTTCGTCTGCTGAGGAGTTGACGACAGGGAGTACCACTCTCCCCCTCTCCAAACTTCAACTCTCCAAAGC
+
A>BB?A@>B?@?<@A@A?8<?<A@?><A@=<;;:B?>>5;;?B>@: :9A: ?8: :9>?AC@?B@ABCA@<@A?>=>B? :>=87;9:9:@>><;==: ?B?<==B=4323914:753
@9VMIV:00229:02357
GGAGGAGTTGAAGTTTGGAGGAGGGGAGAGTGGTACTCCCTGTCGTCAACTCCTCAGCAGAAGCTGATAGTTGTTGGGCTGCAGCAAGGGCTGGGCGATATCCCCTTGCTCGCAGC
+
A?DA@A@ABCCB@ACCC?B@?B@>?C@C@A>AA@B@?=BBA?9A@EB@C@>A@B@>D@A@>CCCB@?@BCABC>? ?AB>D@>DC@??>DC>?<?@=8;<8;<<>A?8;>;<
@9VMIV:00282:01752
GGAGGAGTTGAAGTTTGGAGGAGGGGAGAGTGGTACTCCCTGTCGTCAACTCCTCAGCAGAAGCTGATAGTTGTTGGGCTGCAGCAAGGGCTGGGCGATATCCCCTTGCTCGCAGC
+
@>DA@C@ABC@B?@ABC?C@?B?>=?B@B@AC?AA@B@?>BC@?9A@BB@B@>C@D@?D@B@?BBBB@C@C@A@AA>>?=@>E@>CC?>?>BC>?=:AA?>;=:;<?><<<@;
@9VMIV:00290:02125
GGAGGAGTTGAAGTTTGGAGGAGGGGAGAGTGGTACTCCCTGTCGTCAACTCCTCAGCAGAAGCTGATAGTTGTTGGGCTGCAGCAAGGGCTGGGCGATATCCCCTTGCTCGCAGC
+
A?DA?C@ACCAB@BAB?B@?C@>?A@A@AC?@C@B@>=CCA?:A@DB@B@>B@E@=D@B@BBBBA?@BA@AA>?<BB>D@>CA?>?BB?>=9BC?<:>>>ABC?A<37654
@9VMIV:00398:00013
GGAGGAGTTGAAGTTTGGAGGAGGGGAGAGTGGTACTCCCTGTCGTCAACTCCTCAGCAGAAGCTGATAGTTGTTGGGCTGCAGCAAGGGCTGGGCGATATCCCCTTGCTCGCAGC
+
A>DA@C@C@C@C@CBB?B@?A?>A@A@AC?@C@?B@?>BCA@:A@B@>C@>B@D@>D?>=BA@B@:79578558<?>><
@9VMIV:00571:02383
GGAGGAGTTGAAGTTTGGAGGAGGGGAGAGTGGTACTCCCTGTCGTCAACTCCTCAGCAGAAGCTGATAGTTGTTGGGCTGCAGCAAGGGCTGGGCGATATCCCCTTGCTCGCAGC
+
@>AA?AAACCAB@ACBC?C@?B@>>>C@C@AC?@C@??<=?C?<A@DB@A@>B@B@>B@AA@CAAA@C@?>@<5<584645B>=>EB@>;<;?>597: :8:9<<198<7@?054<<
@9VMIV:00585:00910
GGAGGAGTTGAAGTTTGGAGGAGGGGAGAGTGGTACTCCCTGTCGTCAACTCCTCAGCAGAAGCTGATAGTTGTTGGGCTGCAGCAAGGGCTGGGCGATATCCCCTTGCTCGCAGC
+
A?B@B@ABCCB@AAACA?C@?C@? ? ?C@B@C@AA@B?>=CB@:@A@DB@B@;A@D@?E@BB?AAAA<>?CCABC?<=?>E@>EB@?==4B;;;23322::1453;10.244<;
```

A single sample could have Millions of reads. My experience is that...

Panel of 50 genes → 70,000 reads

Whole-exome seq → 30,000,000 to 70,000,000 reads

Quality assessment of raw sequencing data

Performed by  **FastQC software**: easy-to-use software.

Input : FastQ files. **Output report**: html (web-like) with plots

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Additional tests: over-represented contaminants (adaptor/vector), technical biases, etc.

Integrated in RUBioSeq! Easy-to-use.

Manual: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/>

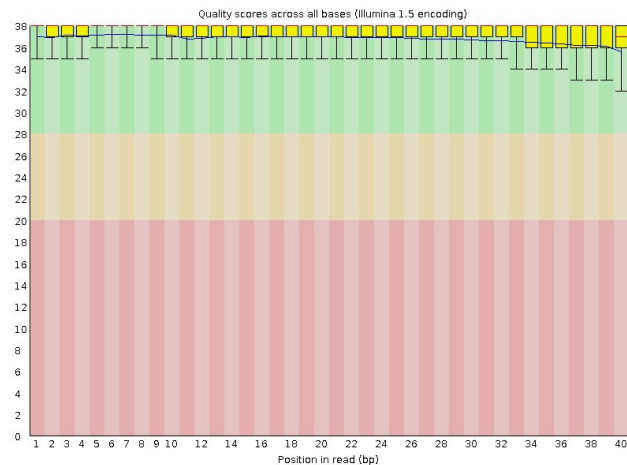
Webpage: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Quality assessment of raw sequencing data

Performed by  **FastQC software**: easy-to-use software.

Input : FastQ files. **Output report**: html (web-like) with plots

✔ **Per base sequence quality**



GOOD SAMPLE

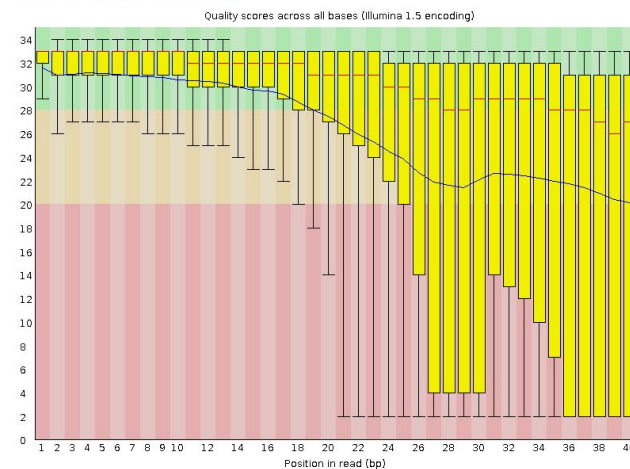
Additional tests: over-represented contaminants (adaptor/vector), technical biases, etc.

Integrated in RUBioSeq! Easy-to-use.

Manual: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/>

Webpage: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

✖ **Per base sequence quality**



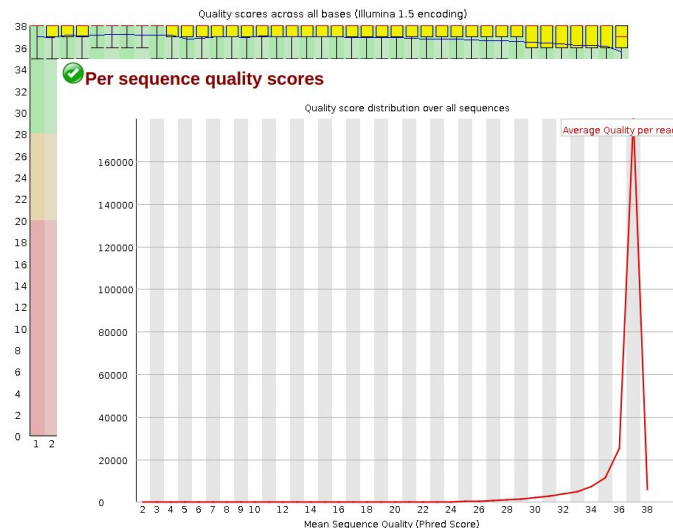
BAD SAMPLE

Quality assessment of raw sequencing data

Performed by  **FastQC software**: easy-to-use software.

Input : FastQ files. **Output report**: html (web-like) with plots

✔ **Per base sequence quality**

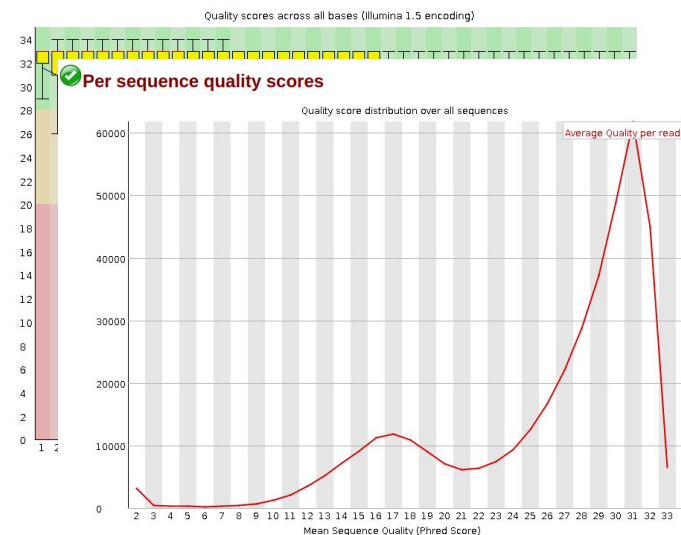


GOOD SAMPLE

Additional tests: over-represented contaminants (adaptor/vector), technical biases, etc.

Integrated in RUBioSeq! Easy-to-use.

✘ **Per base sequence quality**



BAD SAMPLE

Manual: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/>

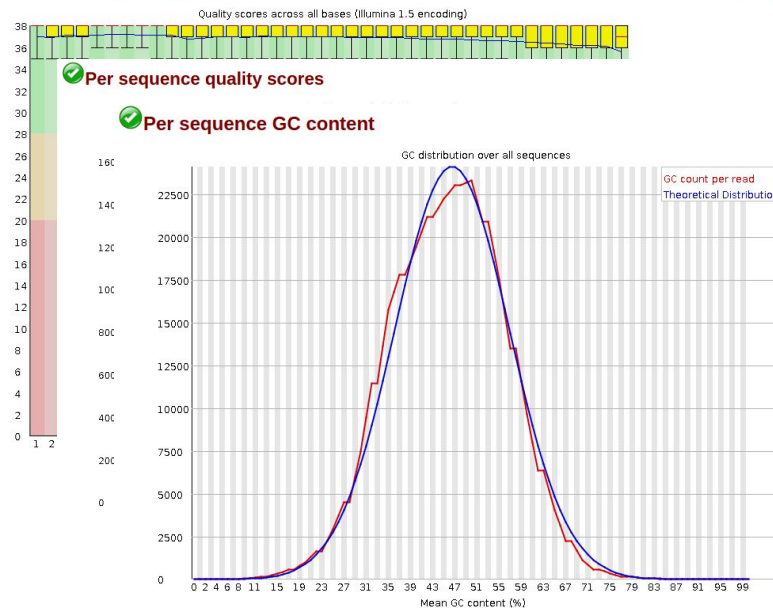
Webpage: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Quality assessment of raw sequencing data

Performed by  **FastQC software**: easy-to-use software.

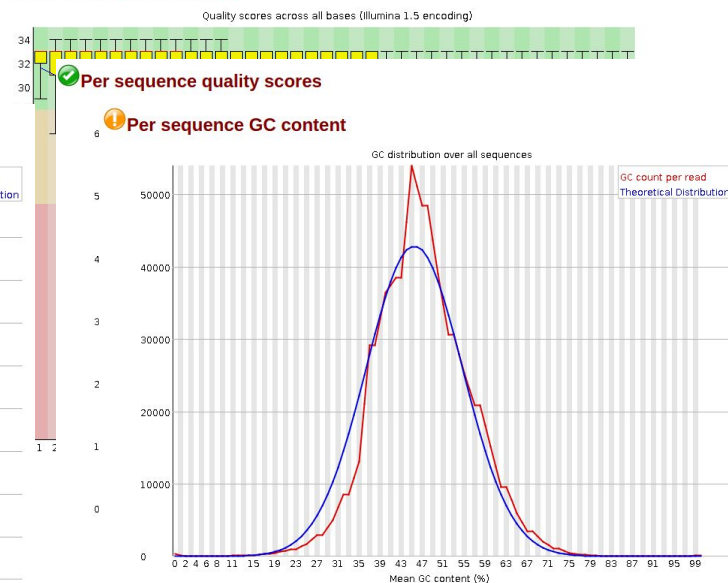
Input : FastQ files. **Output report**: html (web-like) with plots

✔ **Per base sequence quality**



GOOD SAMPLE

✘ **Per base sequence quality**



BAD SAMPLE

Additional tests: over-represented contaminants (adaptor/vector), technical biases, etc.

Integrated in RUBioSeq! Easy-to-use.

Manual: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/>

Webpage: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

SAM & BAM FORMAT :: Part 1 - header

Terminal

```
jperales@machine:recal $ (samtools view -H experiment-output_tumour_recalibrated.bam; samtools view experiment-output_tumour_recalibrated.bam | head -n 1)
```

Header
(start with @)
meta-info.

Contigs (chr names) & lengths

Post-alignment processing program
Aligners (versions)

PL : Platform

SM : Sample name

Alignment of a single read

```
@HD VN:1.4 GO:none SO:coordinate
@SQ SN:chr1 LN:249250621
@SQ SN:chr2 LN:243199373
@SQ SN:chr3 LN:198022430
@SQ SN:chr4 LN:191154276
@SQ SN:chr5 LN:180915260
@SQ SN:chr6 LN:171115067
@SQ SN:chr7 LN:159138663
@SQ SN:chr8 LN:146364022
@SQ SN:chr9 LN:141213431
@SQ SN:chr10 LN:135534747
@SQ SN:chr11 LN:135006516
@SQ SN:chr12 LN:133851895
@SQ SN:chr13 LN:115169878
@SQ SN:chr14 LN:107349540
@SQ SN:chr15 LN:102531392
@SQ SN:chr16 LN:90354753
@SQ SN:chr17 LN:81195210
@SQ SN:chr18 LN:78077248
@SQ SN:chr19 LN:59128983
@SQ SN:chr20 LN:63025520
@SQ SN:chr21 LN:48129895
@SQ SN:chr22 LN:51304566
@SQ SN:chrX LN:155270560
@SQ SN:chrY LN:59373566
@RG ID:Undefinedtumour PL:illumina PU:AA LB:experiment-output
@PG ID:GATK IndelRealigner VN:3.1-1-g07a4bf8 CL:knownAlleles=[(RodBinding name=knownAlleles source=/local/jperales/REFERENCES/bundle_GATK/2.8/dbsnp_138.hg19.vcf), (RodBinding name=knownAlleles2 source=/local/jperales/REFERENCES/bundle_GATK/2.8/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf)] targetIntervals=/home/jperales/D1_overnight/experiment-output/tumour/realign/experiment-output_tumour_forRealigner.intervals LODThresholdForCleaning=5.0 consensusDeterminationModel=USE_READS entropyThreshold=0.15 maxReadsInMemory=150000 maxIsizeForMovement=3000 maxPositionalMoveAllowed=200 maxConsensuses=30 maxReadsForConsensuses=120 maxReadsForRealignment=20000 noOriginalAlignmentTags=false nWayOut=null generate_nWayOut_md5s=false check_early=false noPGTag=false keepPGTags=false indelsFileForDebugging=null statisticsFileForDebugging=null SNPsFileForDebugging=null
@PG ID:bfast VN:0.7.0b
@PG ID:bwa PN:bwa VN:0.7.10-r789 CL:/home/jperales/Soft/bwa-0.7.10/bwa samse -f /home/jperales/D1_overnight/experiment-output/tumour/preProcess/experiment-output_tumour_BWA.sam /local/jperales/REFERENCES/hg19_bwa-0.7.10_bfast-0.7.0b/hg19.fa /home/jperales/D1_overnight/experiment-output/tumour/preProcess/tumour_aln_sa.sai /home/jperales/D1_overnight/Raw_data/patientCHP_tumour.fastq
@PG ID:GATK PrintReads VN:3.1-1-g07a4bf8 CL:readGroup=null platform=null number=-1 sample_file=[] sample_name=[] simplify=false no_pg_tag=false
9VMIV:02395:02034 0 chr1 43814873 8 215M2I1M1D10M3I5M2I7M1I3M2D5M1I2M3D5M5I3M1D4M3I1M1D6M2I2M5I7M4I2M5D112M1D6M * 0 0 CAGCAGGT
CAGAGACGGGGGGCCACAGTCGCCGGGATGCGTAGCTCCCTTCCCGATGTTCTCAACAACCTTGATAGATGTTGCCCTTCAAGGGGCTCGCATCTAGTGTGGGCTCAGCGCCGCTCTGGGCTGCTGCTGCTGAGGTGGCAGTTTCCTGCACACTACAGGTACCGCCC
CCGCCAGGCAGGAGACTGCGCGTGACAGGTG
=B==>=====?>>><8=<=<9==A>@>78=<<8<==>@>8<==>=<8=<=<8?A=====>>>@>==A>;<>>@>;43:4:563<9<=<9<=>@>@>==?<=<=>;==@><8<97<<<<>;7
<7<>;><<<<;<342132110:<2015:==><<<=<=:95:0378811586;:58>=558<;>87,7468.+6856 XA:i:3 MD:Z:2T3^T3G7T1C11^GC2A4^TGA6T1^G5^T1G1C6T6^TGACC1C110^C6 PG:Z:bfast
G:Z:Undefinedtumour IH:i:1 NH:i:3 HI:i:1 NM:i:54 AS:i:2850
```

[...] the rest of reads.

- SAM format specs: <http://samtools.github.io/hts-specs/SAMv1.pdf>
- SAM flags: <http://broadinstitute.github.io/picard/explain-flags.html>

SAM & BAM FORMAT :: Part 2 - alignment

```
#1      #2      #3      #4      #5      #6      #7      #8      #9      #10     #11     #12
ReadName 99      chr10    2       30      3MD2M1I1M    =       14      20      CATCTAG    jjjjjjj    z:Aligner
```

RefPos: 1 2 3 4 5 6 7 8 9
Reference: C C A T A C T - G A
Read: C A T - C T A G
POS: 2
CIGAR: 3M1D2M1I1M

CIGAR

Op	Description
M	alignment match (can be a sequence match or mismatch)
I	insertion to the reference
D	deletion from the reference
N	skipped region from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
P	padding (silent deletion from padded reference)
=	sequence match
X	sequence mismatch

FIELDS

#Col	Field	Description
1.	QNAME	read name
2.	FLAG	bitwise FLAG* (unmapped, pair unmapped, properly mapped, ...)
3.	RNAME	Reference sequence name (e.g. chr1).
4.	POS	1-based leftmost position.
5.	MAPQ	Mapping Quality (Phred-scaled). Scale 0 to 255.
6.	CIGAR	extended CIGAR string.
7.	MRNM	Paired-end: Mate Reference sequence Name (= if same as RNAME).
8.	MPOS	Paired-end: 1-based Mate position.
9.	TLEN	Paired-end: Insert size
10.	SEQ	Read sequence
11.	QUAL	Base Quality Score from the Read sequence.
12.	OPT	Optional Tags.

- SAM format specs: <http://samtools.github.io/hts-specs/SAMv1.pdf>
- SAM flags: <http://broadinstitute.github.io/picard/explain-flags.html>

BED FORMAT

```
track name="CHP2_Designed" description="Amplicon_Insert_CHP2" type=bedDetail ionVersion=4.0
chr1 43814968 43815086 CHP2_MPL_1 . GENE_ID=MPL
chr1 115252185 115252269 CHP2_NRAS_3 . GENE_ID=NRAS
chr1 115256504 115256584 CHP2_NRAS_2 . GENE_ID=NRAS
chr1 115258689 115258774 CHP2_NRAS_1 . GENE_ID=NRAS
chr2 29432572 29432680 CHP2_ALK_2 . GENE_ID=ALK
chr2 29443607 29443729 CHP2_ALK_1 . GENE_ID=ALK
chr2 209113103 209113206 CHP2_IDH1_1 . GENE_ID=IDH1
chr2 212288904 212288990 CHP2_ERBB4_8 . GENE_ID=ERBB4
chr2 212530051 212530180 CHP2_ERBB4_7 . GENE_ID=ERBB4
```

The first three are required BED fields, the rest optional.

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2).
2. **chromStart** - The starting position of the feature in the chromosome. The first base in a chromosome is **numbered 0**.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold.

Additionally, 9 optional fields:

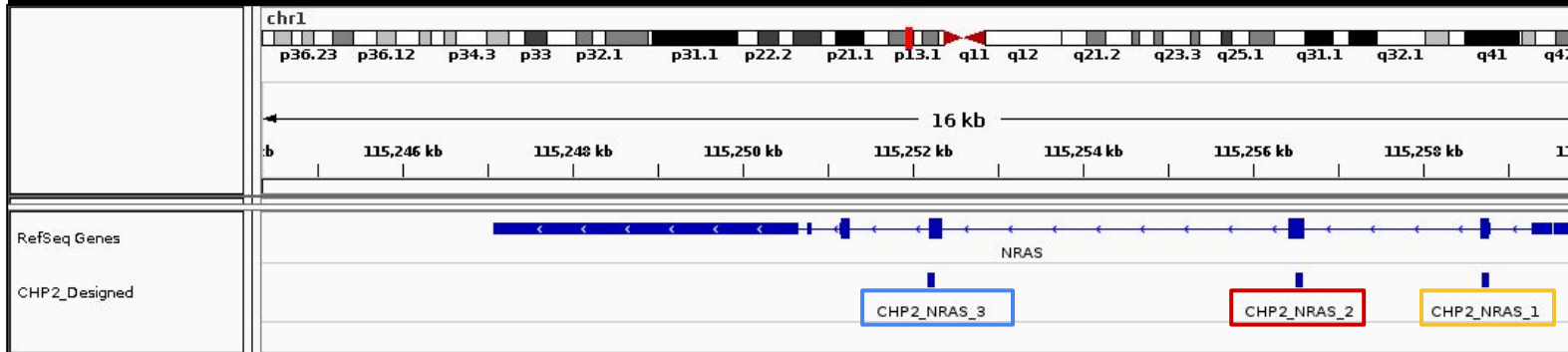
4. **name** - Defines the name of the BED line.
5. Score (. or a number between 0 and 1000).
6. strand (+ forward, - reverse)
7. thickStart
8. thickEnd
9. itemRgb (**255,0,0**).
10. blockCount; 11. blockSizes; 12. blockStarts

The BED format is 0-based.

Official specs: <http://genome.ucsc.edu/FAQ/FAQformat#format1>
<http://bedtools.readthedocs.org/en/latest/content/general-usage.html>

BED FORMAT

```
track name="CHP2_Designed" description="Amplicon_Insert_CHP2" type=bedDetail ionVersion=4.0
chr1 43814968 43815086 CHP2_MPL_1 . GENE_ID=MPL
chr1 115252185 115252269 CHP2_NRAS_3 . GENE_ID=NRAS
chr1 115256504 115256584 CHP2_NRAS_2 . GENE_ID=NRAS
chr1 115258689 115258774 CHP2_NRAS_1 . GENE_ID=NRAS
chr2 29432572 29432680 CHP2_ALK_2 . GENE_ID=ALK
chr2 29443607 29443729 CHP2_ALK_1 . GENE_ID=ALK
chr2 209113103 209113206 CHP2_IDH1_1 . GENE_ID=IDH1
chr2 212288904 212288990 CHP2_ERBB4_8 . GENE_ID=ERBB4
chr2 212530051 212530180 CHP2_ERBB4_7 . GENE_ID=ERBB4
```



WARNING: You must know what **library designed** was used to sequence your data, each library has particular regions to be sequenced. You will call variants along these regions.

Examples: Agilent SureSelect v5 Human all Exon, IonTorrent comprehensive cancer panel, etc

VCF format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Header

single
Variant
calls

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA000001	NA000002	NA000003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1 1:43:5:..
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Genotype
variables

Individual
Samples

valid?

Genotype specifications

- VCF records are oriented to provide details of single variant calls.
- Not all records in a VCF are true calls, the **FILTER** column specifies those which passed the calling.
- **QUAL** is the score assigned to a given call. The greater QUAL is, the more reliable is. It is in log-scale.
- **ID** is an identifier. E.g. a dbSNP id.

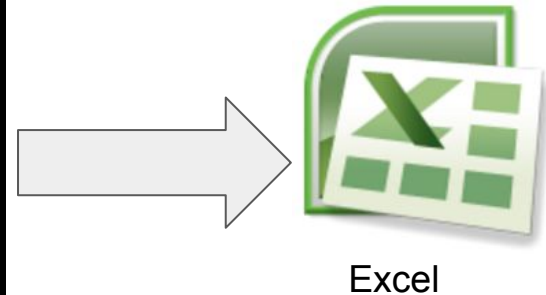
A PDF with the v4.2 specifications: <http://samtools.github.io/hts-specs/VCFv4.2.pdf>

What is a VCF and how to interpret it : <https://software.broadinstitute.org/gatk/guide/article?id=1268>

TSV

Tab-separated Text (tabular structure).

Column1	Column2	Column3	Column4	...	ColumnN
var_r1c1	var_r1c2	var_r1c3	var_r1c4	...	var_r1cN
var_r2c1	var_r2c2	var_r2c3	var_r2c4	...	var_r2cN
var_r3c1	var_r3c2	var_r3c3	var_r3c4	...	var_r3cN
var_r4c1	var_r4c2	var_r4c3	var_r4c4	...	var_r4cN
...
var_rNc1	var_rNc2	var_rNc3	var_rNc4	...	var_rNcN



No format specifications.

There are no limits in terms of the matrix dimension (rows X columns).

However, there are good practices in Bioinformatics:

1. Use intuitive column names without spaces or rare chars. Instead of those, use "." or "_".
2. Data dimension: higher number of rows than columns.
3. Rows → individual observations or records. E.g. genomic positions.
4. Columns → Individual variables for each individual observation. E.g. Chromosome, position, Score, Gene mutated, etc.
5. Do NOT mix data from different observations

Tidy data principles:

<http://vita.had.co.nz/papers/tidy-data.pdf>