NGS data formats and Quality Control

Jacqui Keane

@drjkeane
drjkeane@gmail.com

Adapted from slides provided by Petr Danecek
petr.danecek@sanger.ac.uk

Data Formats

Data Formats

FASTQ

- Unaligned read sequences with base qualities

SAM/BAM

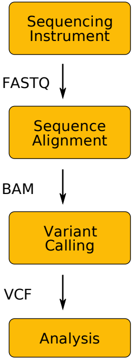
- Unaligned or aligned reads
- Text and binary formats

CRAM

- Better compression than BAM

VCF/BCF

- Flexible variant call format
- Arbitrary types of sequence variation
- SNPs, indels, structural variations

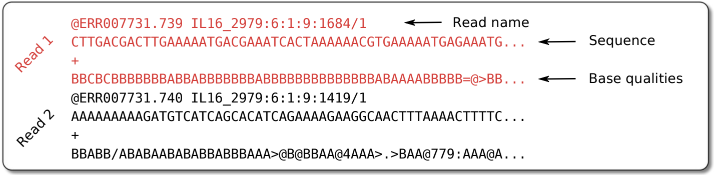


Specifications maintained by the Global Alliance for Genomics and Health

FASTA - reference genome

```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TATTCAAAAAATTGAGAATTTCTGACCACTTAACAAACCCACAGAAAATCCACCCGAGTG
CACTGAGCACGCCAGAAATCAGGTGGCCTCAAAGAGCTGCTCCCACCTGAAGGAGACGCG
CTGCTGCTGCTGTCGTCCTGCCTGGCGCCTTGGCCTACAGGGGCCGCGGTTGAGGGTGGG
AGTGGGGGTGCACTGGCCAGCACCTCAGGAGCTGGGGGTGGTGGTGGGGGCGGTGGGGGT
GGTGTTAGTACCCCATCTTGTAGGTCTGAAACACAAAGTGTGGGGTGTCTAGGGAAGAAG
>2
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
AAAAGCATTTATGCTACAAATTACTATGGTAATTATGCTACAAATTTATGGTACCATAAA
TTACCATAGTAATTTGTAGCATAAATTTGTACTATGGTACAAATTACATGGGAGAGTGAA
GGTGGGTTAAAACATTCATATTAAAGAACTTCCACTCAGATTGCAAGAAAAGAGAGAGGA
ATGGAGATGGTAGCACAAGTCCCTACAATAAAAGTAGATGTTTTGAGATCAGTTCTATTT
```

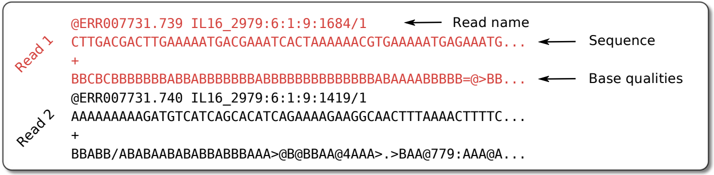
FASTQ



```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG...
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>BB...
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC...
+
BBABB/ABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A...
```

- Simple format for raw unaligned sequencing reads
- Paired-end sequencing: two FASTQ files or one interleaved file

FASTQ



Read 1

```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG...
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>BB...
```

Read 2

```
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC...
+
BBABB/ABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A...
```

- Simple format for raw unaligned sequencing reads
- Paired-end sequencing: two FASTQ files or one interleaved file

- Quality encoded in ASCII characters with decimal codes 33-126
  - ASCII code of "A" is 65, the corresponding quality is Q $= 65 - 33 = 32$

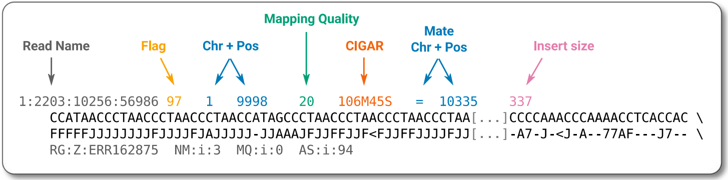
```
Base quality encoded as character
     ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J
                                |                                     |
Numeric ASCII value
    33 . . . . . . . . . . . . . 47 . . . . . . . . . . . . . . . . . 65 . . . . . . . .

                                |                                     |  (65-33 = 32)
Base quality value
     0 . . . . . . . . . . . . . 14 . . . . . . . . . . . . . . . . . 32 . . . . . . . .
```

Quality = Phred-scaled probability of an error

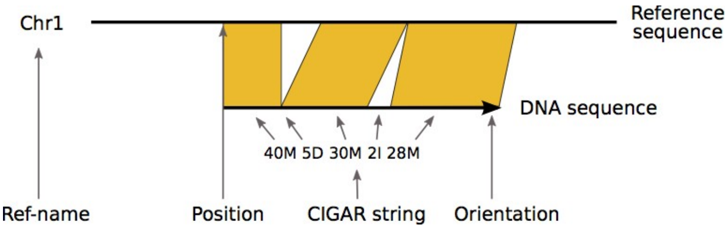
Quality	Probability of error	Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%

$$Q = -10 \log_{10} P \qquad \ldots \qquad P = 10^{-Q/10}$$

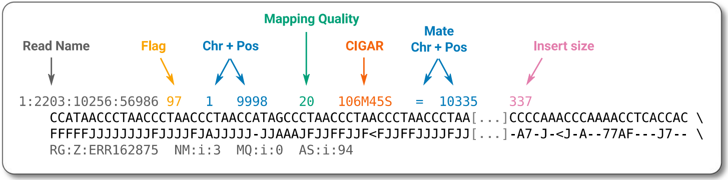
SAM / BAM: Sequence Alignment/Map format



```
1:2203:10256:56986 97   1   9998    20   106M45S   =  10335    337
    CCATAACCCTAACCCTAACCCTAACCATAGCCCTAACCCTAACCCTAACCCTAA[...]CCCCAAACCCAAAACCTCACCAC \
    FFFFFJJJJJJJJFJJJJFJAJJJJJ-JJAAAJFJJFFJJF<FJJFFJJJJFJJ[...]-A7-J-<J-A--77AF---J7-- \
    RG:Z:ERR162875  NM:i:3  MQ:i:0  AS:i:94
```



SAM / BAM: Sequence Alignment/Map format



```
1:2203:10256:56986 97   1   9998   20   106M45S   =  10335    337
    CCATAACCCTAACCCTAACCCTAACCATAGCCCTAACCCTAACCCTAACCCTAA[...]CCCCAAACCCAAAACCTCACCAC \
    FFFFFJJJJJJJJFJJJJFJAJJJJJ-JJAAAJFJJFFJJF<FJJFFJJJJFJJ[...]-A7-J-<J-A--77AF---J7-- \
    RG:Z:ERR162875  NM:i:3  MQ:i:0  AS:i:94
```

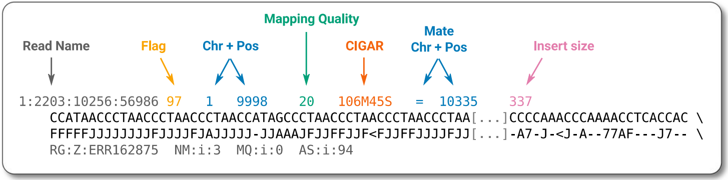
Flag

Hex	Dec	Flag	Description
0x1	1	PAIRED	paired-end (or multiple-segment) sequencing technology
0x2	2	PROPER_PAIR	each segment properly aligned according to the aligner
0x4	4	UNMAP	segment unmapped
0x8	8	MUNMAP	next segment in the template unmapped
0x10	16	REVERSE	SEQ is reverse complemented
0x20	32	MREVERSE	SEQ of the next segment in the template is reversed
0x40	64	READ1	the first segment in the template
0x80	128	READ2	the last segment in the template
0x100	256	SECONDARY	secondary alignment
0x200	512	QCFAIL	not passing quality controls
0x400	1024	DUP	PCR or optical duplicate
0x800	2048	SUPPLEMENTARY	supplementary alignment

Bit operations made easy

- samtools flags
  0xa3 163 PAIRED,PROPER_PAIR,MREVERSE,READ2
- python
  0x1 | 0x2 | 0x20 | 0x80 .. 163
  bin(163) .. 10100011

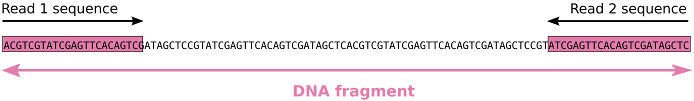
SAM / BAM: Sequence Alignment/Map format



```
1:2203:10256:56986 97   1   9998    20    106M45S    =  10335     337
    CCATAACCCTAACCCTAACCCTAACCATAGCCCTAACCCTAACCCTAACCCTAA[...]CCCCAAACCCAAAACCTCACCAC \
    FFFFFJJJJJJJJFJJJJFJAJJJJJ-JJAAAJFJJFFJJF<FJJFFJJJJFJJ[...]-A7-J-<J-A--77AF---J7-- \
    RG:Z:ERR162875  NM:i:3  MQ:i:0  AS:i:94
```

Insert size

length of the DNA fragment sequenced from both ends by paired-end sequencing:

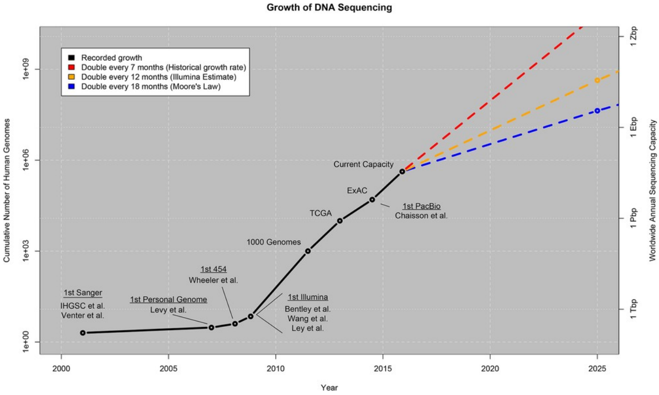

Read 1 sequence
Read 2 sequence
ACGTCGTATCGAGTTCACAGTCGATAGCTCCGTATCGAGTTCACAGTCGATAGCTCACGTCGTATCGAGTTCACAGTCGATAGCTCCGTATCGAGTTCACAGTCGATAGCTC
DNA fragment

CRAM: Reference based Compression

BAM files are too large

- ~1.5-2 bytes per base pair

Increases in disk capacity are being far outstripped by sequencing technologies



Zachary D. Stephens, *et al*, Big Data: Astronomical or Genomical? DOI: 10.1371/journal.pbio.1002195

CRAM: Reference based Compression

BAM files are too large

- ~1.5-2 bytes per base pair

Increases in disk capacity are being far outstripped by sequencing technologies

BAM stores all of the data

- Every read base
- Every base quality
- Using a single conventional compression technique for all types of data

```
Reference sequence: ACGTACGTACGTACGTACGTACGTACGTACGTAC
read 1:             ACGTACGTACGTACGTACGTGC
read 2:                TACGTACGCACGTACGTGCGTA
read 3:                  CGTACGCACGTACGTACGTACG
read 4:                    TACGTACGTACGTGCGTACGTA
read 5:                      CGCACGTACGTACGTACGTACG
read 6:                            TACGTGCGTACGTACGTAC
```

CRAM: Reference based Compression

BAM files are too large

- ~1.5-2 bytes per base pair

Increases in disk capacity are being far outstripped by sequencing technologies

BAM stores all of the data

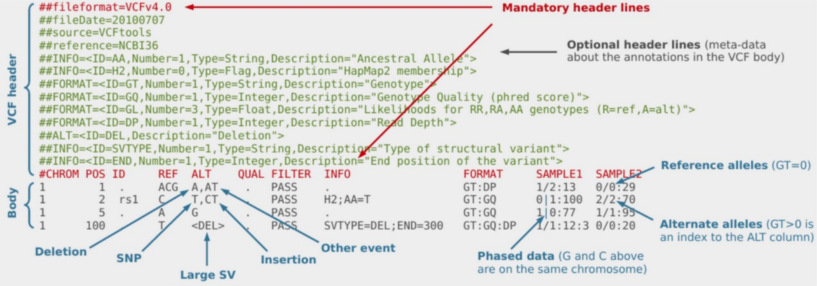
- Every read base
- Every base quality
- Using a single conventional compression technique for all types of data

```
Reference sequence: ACGTACGTACGTACGTACGTACGTACGTACGTAC
read 1:             ....................G.
read 2:                ........C.............
read 3:                  ......C...............
read 4:                    .............G........
read 5:                      ..C...................
read 6:                            .....G.............
```

CRAM: in lossless mode 60% of BAM size

- Reference based compression
- Controlled loss of quality information
- Different compression methods for different type of data

VCF: Variant CallFormat



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID    REF  ALT    QUAL FILTER  INFO                FORMAT    SAMPLE1   SAMPLE2
1        1 .     ACG  A,AT   .    PASS    .                   GT:DP     1/2:13    0/0:29
1        2 rs1   C    T,CT   .    PASS    H2;AA=T             GT:GQ     0|1:100   2/2:70
1        5 .     A    G      .    PASS    .                   GT:GQ     1|0:77    1/1:95
1      100       T    <DEL>  .    PASS    SVTYPE=DEL;END=300  GT:GQ:DP  1/1:12:3  0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Phased data (G and C above are on the same chromosome)
Deletion
SNP
Large SV
Insertion
Other event

File format for storing variation data

- tab-delimited text, parsable by standard UNIX commands
- flexible and user-extensible
- compressed with BGZF (bgzip), indexed with TBI or CSI (tabix)

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO           FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535  .   G   A     243   PASS   DP=221;AF=0.5  GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Row-oriented, tab-delimited file with eight mandatory columns (CHROM-INFO)

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO             FORMAT  SAMPLE1     SAMPLE2   SAMPLE3
11     24535 .   G   A     243  PASS    DP=221;AF=0.5    GT:AD   0/1:73,15   0/0:48,0  0/1:71,14
```

Genomic coordinates

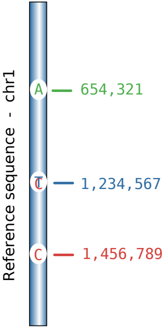
VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT   QUAL FILTER  INFO           FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535  .   G   A     243   PASS   DP=221;AF=0.5  GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Arbitrary string, typically a dbSNP RefSNP id. Dot for missing value.

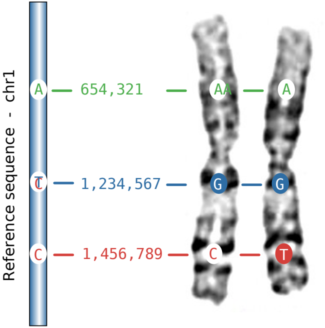
VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT   QUAL FILTER  INFO           FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535  .   G   A     243   PASS   DP=221;AF=0.5  GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```



VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT   QUAL FILTER  INFO           FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535  .   G   A     243   PASS   DP=221;AF=0.5  GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```



VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1     SAMPLE2    SAMPLE3
11    24535  .   G   A    243   PASS   DP=221;AF=0.5 GT:AD   0/1:73,15   0/0:48,0   0/1:71,14
```

Although in theory phred-scaled probability, don't expect truly probabilistic interpretation in practice.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER INFO            FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535   .   G   A     243   PASS  DP=221;AF=0.5   GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Soft-filter variants with e.g. low quality, low depth, etc.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO            FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11     24535 .   G   A     243  PASS    DP=221;AF=0.5   GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Per-site annotations. Here **DP** is the cumulative read depth across all samples and **AF** allele frequency of the allele in general population.

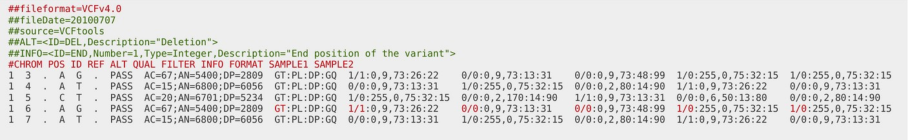
VCF vs BCF

VCFs can be very big

- compressed VCF with 3781 samples, human data:
  - 54 GB for chromosome 1
  - 680 GB whole genome

VCFs can be slow to parse

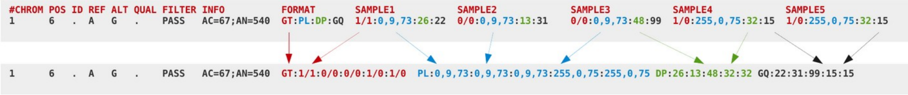
- text conversion is slow
- main bottleneck: FORMAT fields



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1  3  .  A  G  .  PASS  AC=67;AN=5400;DP=2809  GT:PL:DP:GQ  1/1:0,9,73:26:22    0/0:0,9,73:13:31    0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15
1  4  .  A  T  .  PASS  AC=15;AN=6800;DP=6056  GT:PL:DP:GQ  0/0:0,9,73:13:31    1/0:255,0,75:32:15  0/0:0,2,80:14:90  1/1:0,9,73:26:22    0/0:0,9,73:13:31
1  5  .  C  T  .  PASS  AC=20;AN=6701;DP=5234  GT:PL:DP:GQ  1/0:255,0,75:32:15  0/0:0,2,170:14:90   1/1:0,9,73:13:31  0/0:0,6,50:13:80    0/0:0,2,80:14:90
1  6  .  A  G  .  PASS  AC=67;AN=5400;DP=2809  GT:PL:DP:GQ  1/1:0,9,73:26:22    0/0:0,9,73:13:31    0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15
1  7  .  A  T  .  PASS  AC=15;AN=6800;DP=6056  GT:PL:DP:GQ  0/0:0,9,73:13:31    1/0:255,0,75:32:15  0/0:0,2,80:14:90  1/1:0,9,73:26:22    0/0:0,9,73:13:31
```

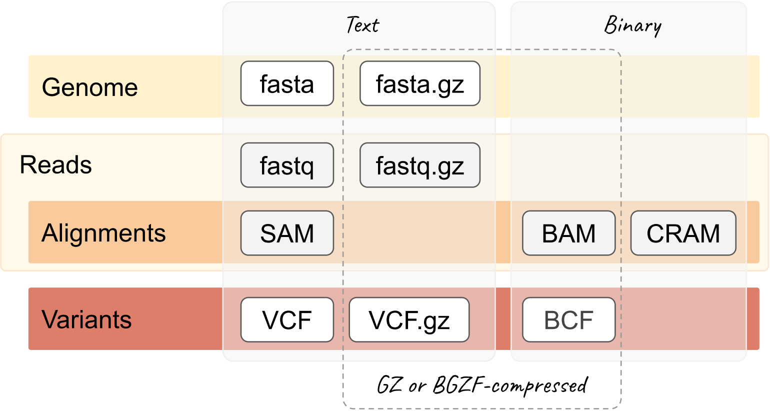
BCF

- binary representation of VCF
- fields rearranged for fast access



```
#CHROM POS ID REF ALT QUAL FILTER INFO          FORMAT       SAMPLE1            SAMPLE2           SAMPLE3           SAMPLE4              SAMPLE5
1      6   .  A   G   .    PASS   AC=67;AN=540  GT:PL:DP:GQ  1/1:0,9,73:26:22   0/0:0,9,73:13:31  0/0:0,9,73:48:99  1/0:255,0,75:32:15   1/0:255,0,75:32:15

1      6   .  A   G   .    PASS   AC=67;AN=540  GT:1/1:0/0:0/0:1/0:1/0  PL:0,9,73:0,9,73:0,9,73:255,0,75:255,0,75 DP:26:13:48:32:32 GQ:22:31:99:15:15
```

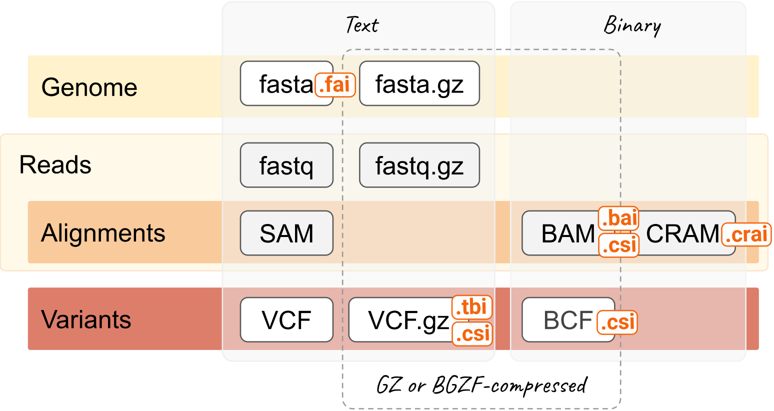
File formats summary



	Text		*Binary*	
Genome	fasta	fasta.gz		
Reads	fastq	fastq.gz		
Alignments	SAM		BAM	CRAM
Variants	VCF	VCF.gz	BCF	

GZ or BGZF-compressed

File formats summary



Text

Binary

Genome

fasta .fai

fasta.gz

Reads

fastq

fastq.gz

Alignments

SAM

BAM .bai .csi

CRAM .crai

Variants

VCF

VCF.gz .tbi .csi

BCF .csi

GZ or BGZF-compressed