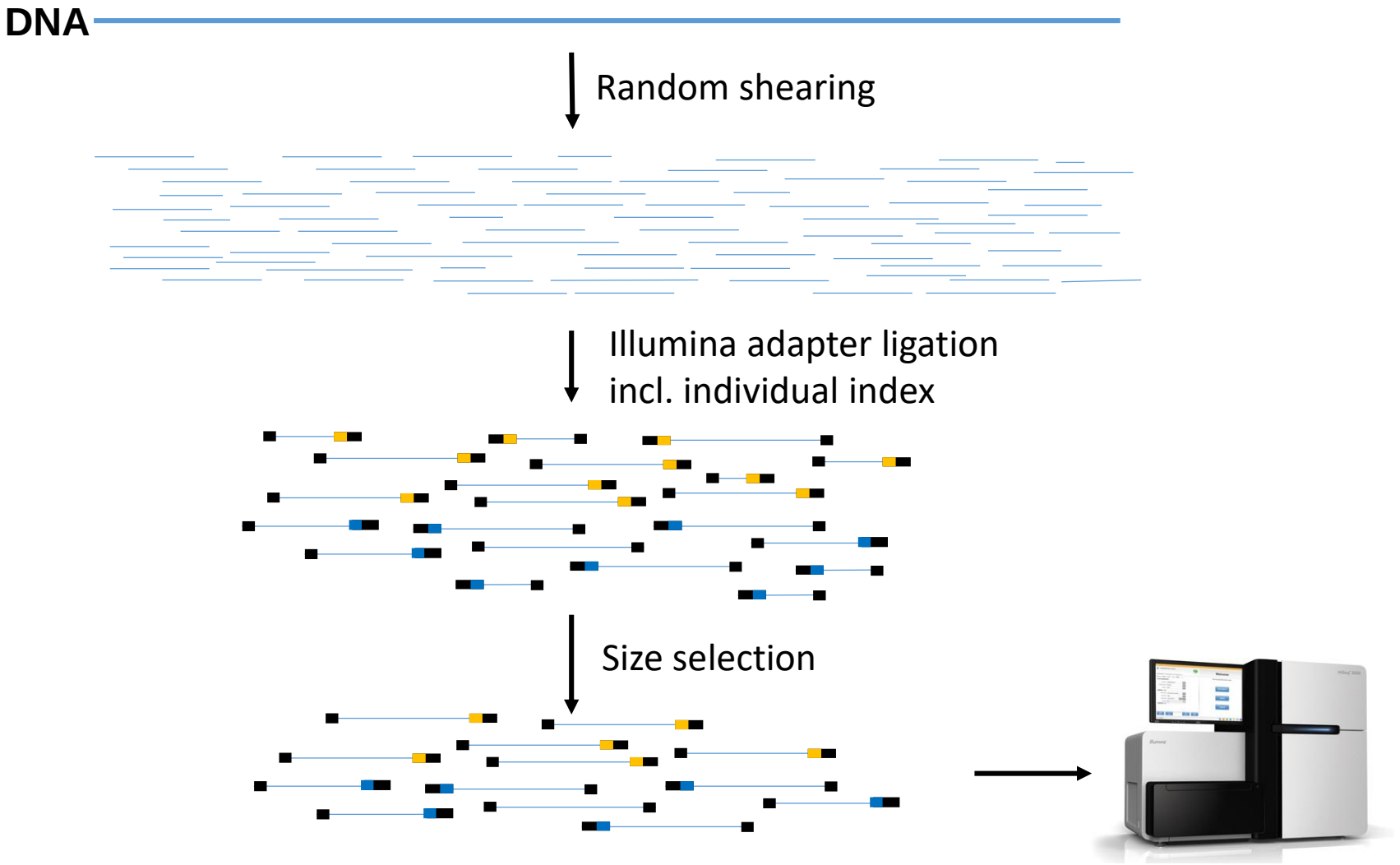
Whole-genome short-read sequencing



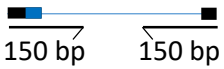
paired-end sequencing

Up to 20 billion read pairs

1. Quality check and trimming raw reads

paired-end Illumina sequencing

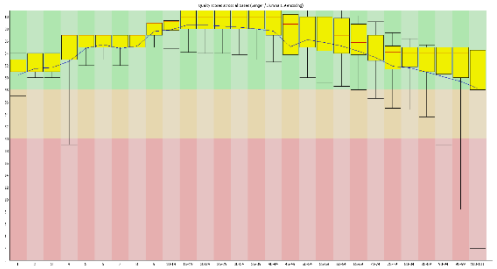
Up to 20 billion read pairs



Forward reads

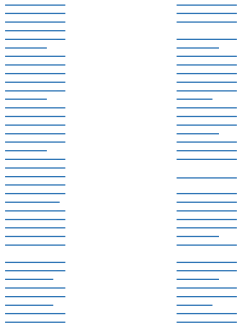
Reverse reads

Fastqc quality plots



fastp

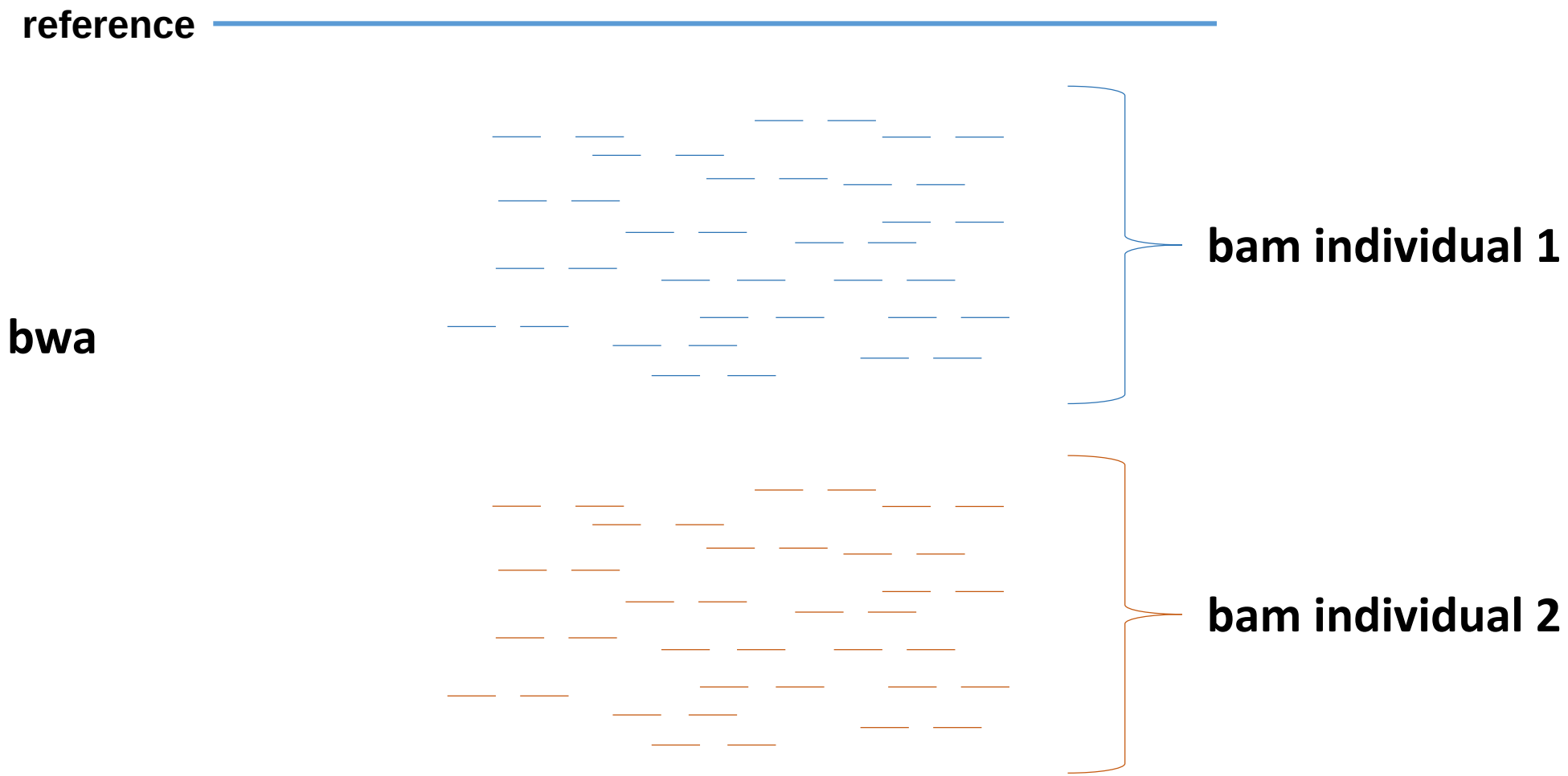
Trimmed and filtered reads of good quality



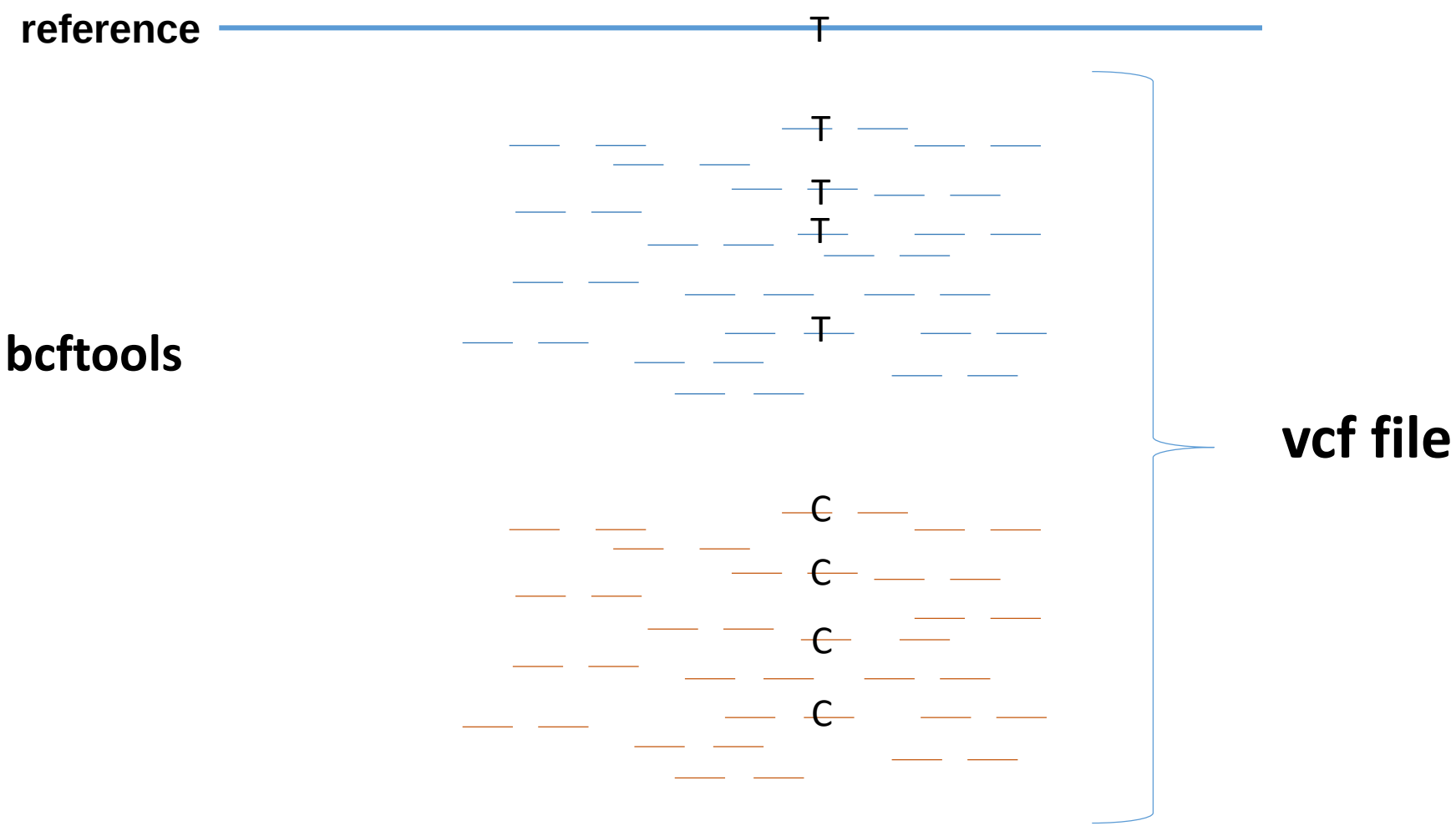
Fastq file of reads

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
```

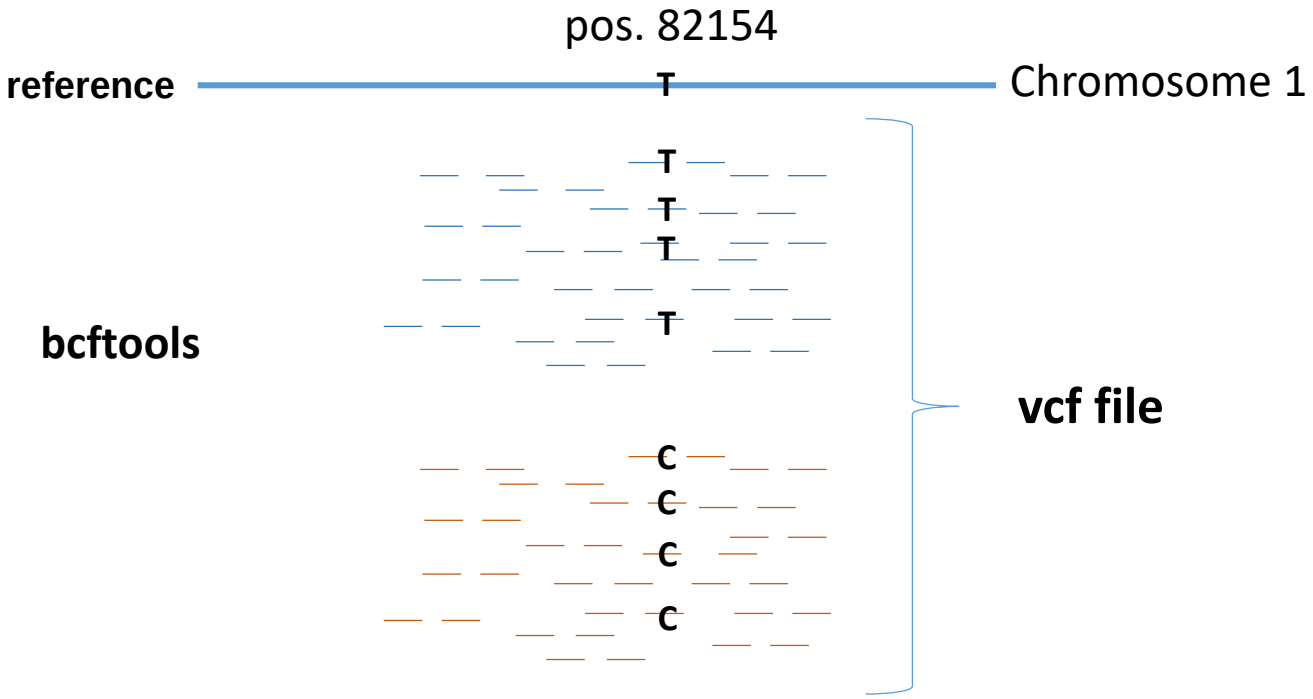
2. Alignment to the reference genome with bwa



3. Variant and genotype calling with bcftools



3. Variant and genotype calling with bcftools



vcf file: Genotypes for each individual at genomic sites

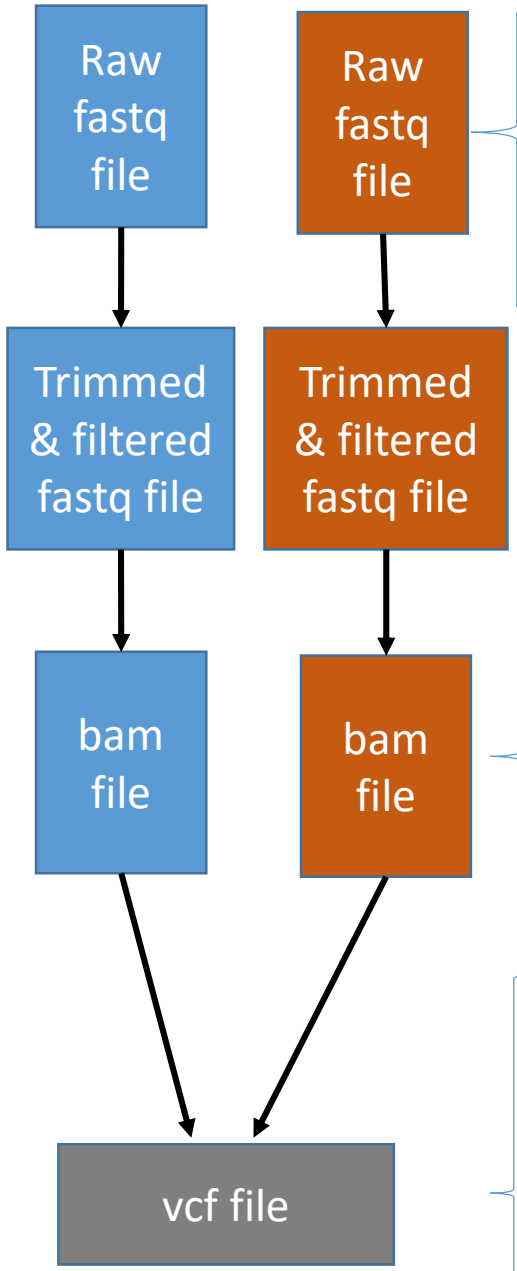
```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS  ID  REF ALT QUAL  FILTER  INFO  FORMAT  GEN
chr1    82154    .    T    C    .    GT  0/0 1/1 0/0 1/1 0/0 1/1
chr1    752566   .    T    .    .    GT  0/0 0/0 0/0 0/0 0/0 0/0
chr1    752721   .    T    C    .    GT  1/1 1/1 1/1 1/1 1/1 1/1
chr1    752721   .    A    .    .    GT  ./. 0/0 0/0 0/0 0/0 0/0
```

Fastqc
Check read quality

fastp
Quality trim
Remove adapter
Remove polyG-tails

bwa
Align to reference

bcftools
Call variants
and genotypes



Fastq: raw reads with sequencing quality information

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
```

sam file: reads mapped to the reference genome -> binary version: bam file

```
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

vcf file: Genotypes for each individual at genomic sites

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  GEN
chr1    82154   .   T   C   .   GT  0/0 1/1 0/0 1/1 0/0 1/1
chr1    752566  .   T   .   .   GT  0/0 0/0 0/0 0/0 0/0 0/0
chr1    752721  .   T   C   .   GT  1/1 1/1 1/1 1/1 1/1 1/1
chr1    752721  .   A   .   .   GT  ./. 0/0 0/0 0/0 0/0 0/0
```

Many tools for variant and genotype calling

- Bcftools – fast, easy and good
- GATK – state of the art but only if the individuals to map are closely related to the reference genome (designed for human data)
- FreeBayes – also widely used
- ANGSD – for low-coverage data (<15x coverage)

Vcf file structure

- For every sequenced position in the genome, there will be a line
- For every individual, there will be a column

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  GEN
chr1    82154   .   T   C   .   GT  0/0 1/1 0/0 1/1 0/0 1/1
chr1    752566  .   T   .   .   GT  0/0 0/0 0/0 0/0 0/0 0/0
chr1    752721  .   T   C   .   GT  1/1 1/1 1/1 1/1 1/1 1/1
chr1    752728  .   A   .   .   GT  ./. 0/0 0/0 0/0 0/0 0/0
```