

# Short Read Alignment & Variant Calling



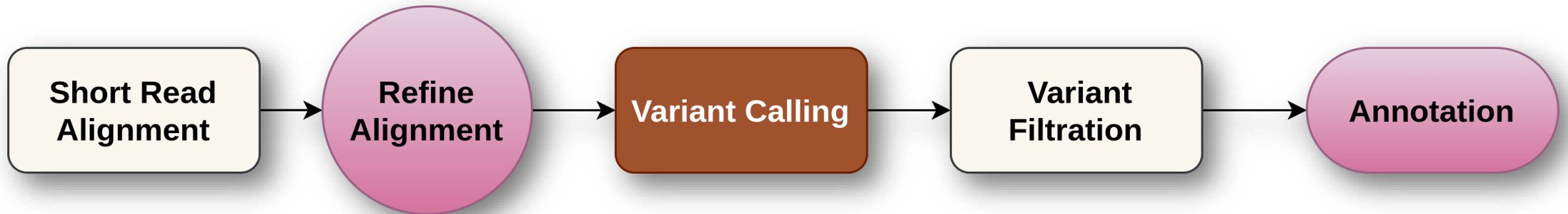
- Loo...ts of **Google/ChatGPT** search
- Proper **data** & **directory** management
- Perform **background study**
- Be **persistent**

**With Great Power**



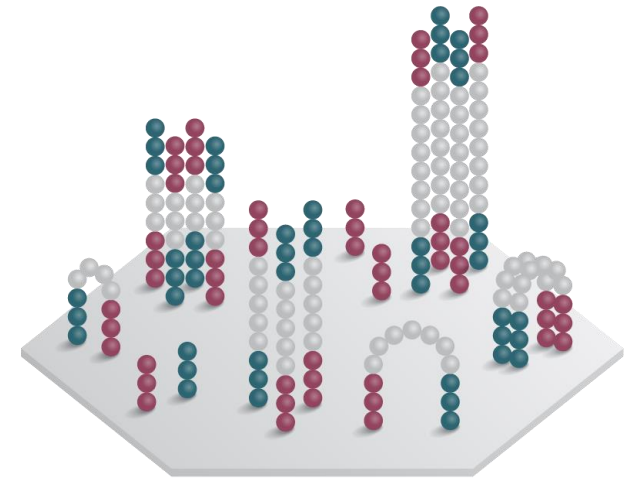
**Comes  
great responsibility**

# Workflow



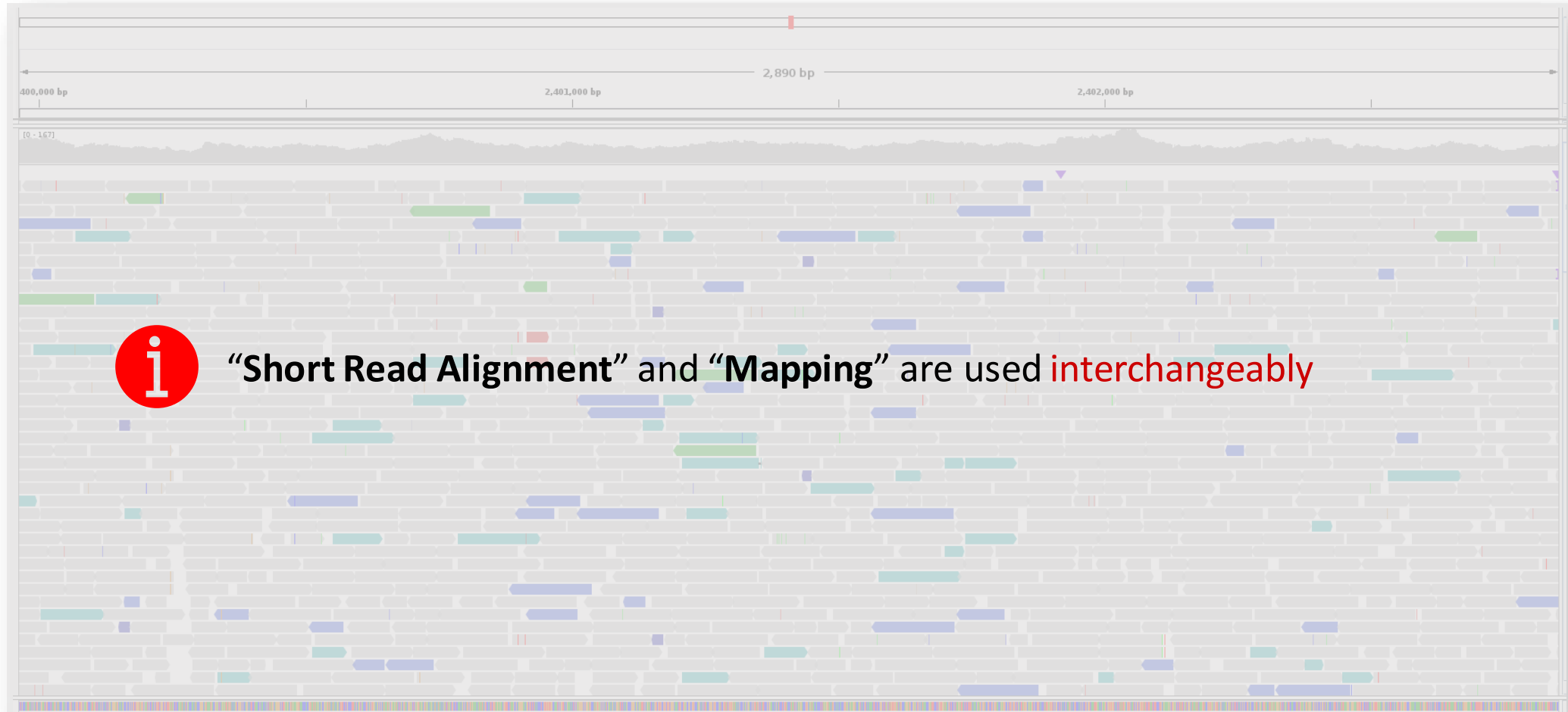
# Short Read Alignment

Basic Concepts and Common Terminologies

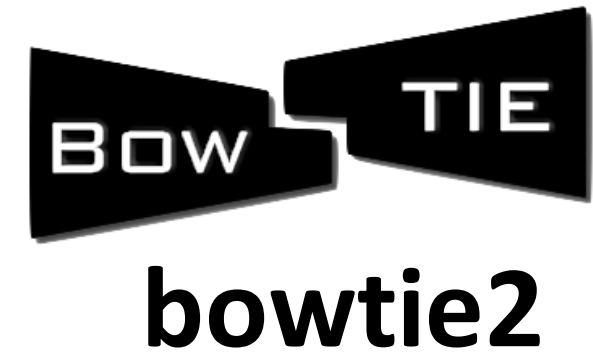
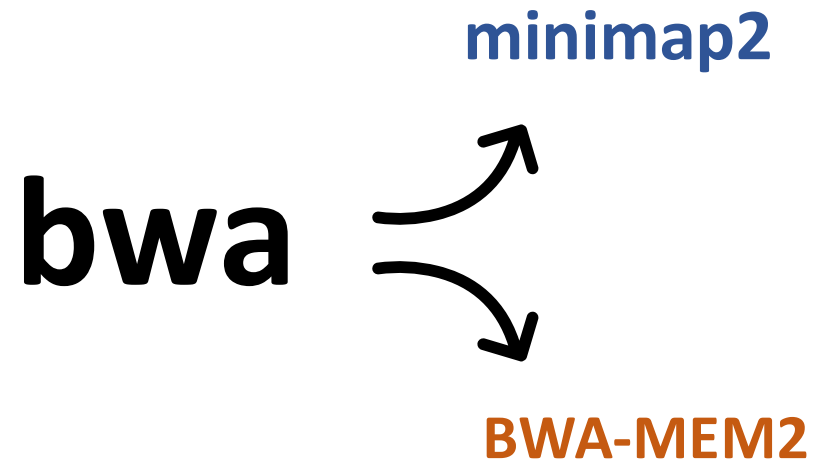


# Short Read Alignment / Mapping

Determine the most likely **location** and **orientation** of each read within the reference sequence.

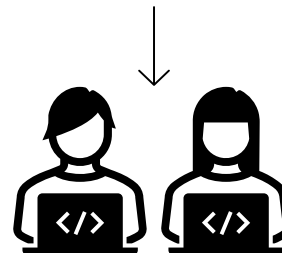


# Commonly Used Tools...



**HISAT2**

**STAR**



# SAM File

@SQ SN:NC\_003198.1 LN:4809037

@PG ID:bwa PN:

2.fastq.gz

VH00293:1:AAAM7FVM5

AAATGATTTGCGTCAGGCC

GCGCGGAGGCAACGATTT

CCCCCCCCCCCCCCCCCCCC

VH00293:1:AAAM7FVM5

AGTTGACCTGCTGGCCGCA

GAATACAGATCCCCACCGC

CCCCCCC;CCCCCCCC;CC

VH00293:1:AAAM7FVM5

CCCCCCCCCACACCACCC

AGTGAGCGTGCCTTACCGA

CCCCCCCC;CCCCCCCCC

VH00293:1:AAAM7FVM5

TTCACTTTCCAGAAATTC

CCGTTTGTGTTATAACAGC

C;C--CC-C-CC-;C;CC

VH00293:1:AAAM7FVM5

AGCAGAAGACGGCATACGA

GTAACCCAACATATGGATC

CCCCCCCCCCCCCCCC;CCC

VH00293:1:AAAM7FVM5

TGCCTCCCCTTGTAAGGGC

CTCGGTGGGGGCCGTATCAT

TAA CCCCCCCCCCCCCCCCCC



এখান থেকে কিছুই নেই, হিজিবিজি!

.fastq.gz data/STY\_0014\_R

2329299 -168 CTGGCCGC

GTGCGCAGAATACAGATCCCCACCG

CCCCCCCCCCCCCCCCCCCC

AS:i:150 XS:i:0

2329317 168 ATCAACAA

ATATTGACACCATCAATAGTGCGCA

CCCCCCCCCCC;C-CCCCCCCCC

AS:i:151 XS:i:0

2873318 -35 CCCCCCCC

CCAGAAATTCTTATTGTGAGAAT

CCCCCCCCCCCCCCCCCCCC

XS:i:0

= 2873318 35 G

CTATAGGTGTAGATCTCCGTGGTCG

CCCCCCCCCCCC---CC--C---C

AS:i:36 XS:i:0

3459458 -75 CTTTTTCA

GGTAGCATATTGTGCTAAGAC

CCCCCCCCCCC-CCCCCCCCC-CC

XS:i:0

3459458 75 AAAAAATA

TTTGTAACTATCGGTGTAGAT

CTCGGTGGGGGCCGTATCAT

TAA CCCCCCCCCCCCCCCCCC



# SAM (Sequence Alignment/Map) File

```
@HD VN:1.5 SO:coordinate
```

```
@SQ SN:ref LN:45
```

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Header  
section

Alignment  
section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; \* meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

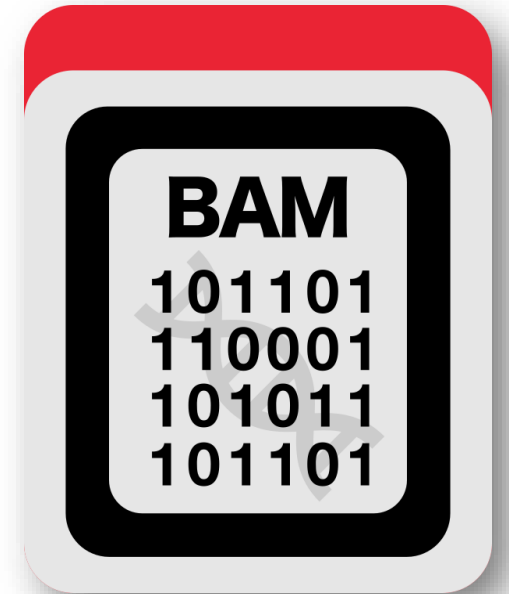
FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID



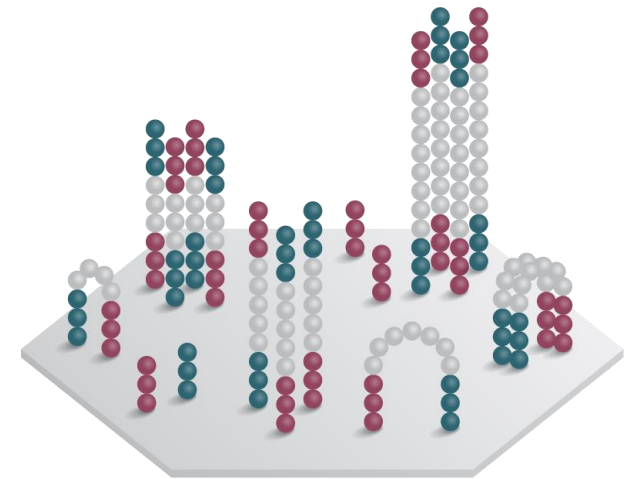
# BAM (Binary Alignment/Map file) File

- **Binary**, compressed (and almost always sorted) representation of the SAM information
  - Sorting by coordinate allows fast query on the information by location.



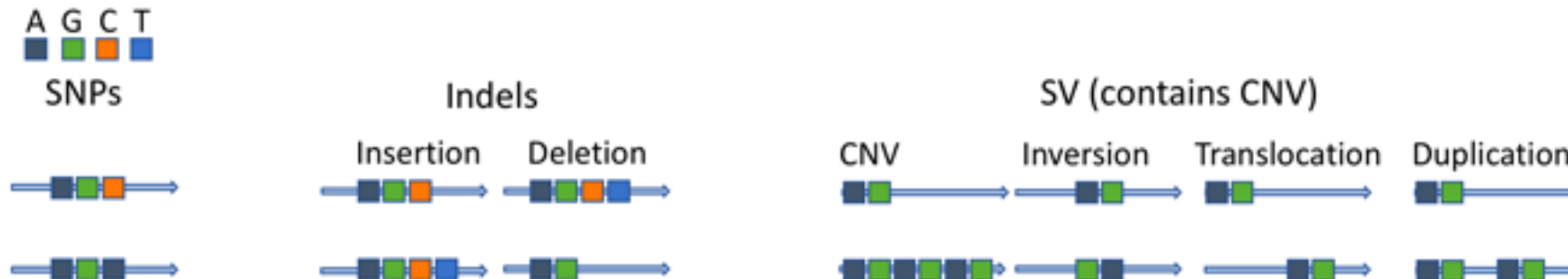
# Variant Calling

Basic Concepts and Common Terminologies



# Variant Calling

Variant calling is the process of identifying genetic variations or differences, such as single nucleotide polymorphisms (**SNPs**) or insertions/deletions (**indels**), in a subject's genome compared to a reference genome.



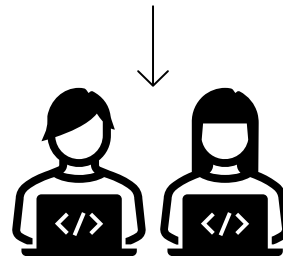
# Commonly Used Variant Callers

**bcftools**

**FreeBayes**

**VarScan2**

**GATK**



# VCF file format

Data representation format used to describe **variations** in the genome

```
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-value"
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate"
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2"
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of reads with each base pair"
##bcftools_callVersion=1.16+htslib-1.16
##bcftools_callCommand=call -v -p -t -c -o - -ploidy 1 -vc - -threads 1
#CHROM POS ID REF ALT QUAL FILTER
NC_003198.1 22623 . T C 225.00
NC_003198.1 28115 . A G 225.00
DP=59;VDB=0.332436;SGB=-0.693146;RPBZ=-2.37244;MQBZ=0;
NC_003198.1 35122 . T C 225.00
DP=89;VDB=0.385487;SGB=-0.693147;RPBZ=3.67751;MQBZ=0;
NC_003198.1 44391 . A G 225.00
NC_003198.1 53460 . G A 225.00
NC_003198.1 61892 . A G 225.00
NC_003198.1 63831 . G A 127.00
NC_003198.1 68011 . T G 225.00
NC_003198.1 69513 . T C 225.00
NC_003198.1 70739 . C T 225.00
NC_003198.1 76149 . G A 225.00
NC_003198.1 76422 . A G 225.00
NC_003198.1 80983 . A G 225.00
NC_003198.1 87828 . T C 225.00
NC_003198.1 90658 . T C 225.00
NC_003198.1 93158 . A G 225.00
NC_003198.1 98494 . GTTTGCCATGCACT GT
INDEL;IDV=49;IMF=0.731343;DP=67;VDB=0.00494681;SGB=-0.
NC_003198.1 117735 . T C 225.00
NC_003198.1 136763 . T C 225.00
NC_003198.1 139560 . G A 5.45858
NC_003198.1 139563 . T C 32.0075
NC_003198.1 156259 . C T 225.00
DP=80;VDB=0.948898;SGB=-0.693147;MQSBZ=0;FS=0;MQOF=0;AF1=1;AC1=1;DP4=0,0,32,23;MQ=60;FQ=-999 GT:PL 1:255,0
DP=1;SGB=-0.379885;FS=0;MQOF=0;AF1=1;AC1=1;DP4=0,0,1,0;MQ=60;FQ=-999 GT:PL 1:34,0
DP=3;VDB=0.1;SGB=-0.453602;FS=0;MQOF=0;AF1=1;AC1=1;DP4=0,0,2,0;MQ=60;FQ=-999 GT:PL 1:62,0
DP=103;VDB=0.985918;SGB=-0.693147;MQSBZ=0;FS=0;MQOF=0;AF1=1;AC1=1;DP4=0,0,32,36;MQ=60;FQ=-999 GT:PL 1:255,0
```

```
">
4=0,0,37,31;MQ=60;FQ=-999 GT:PL 1:255,0
,29,15;MQ=60;FQ=-999;PV4=1,1,1,0.17597 GT:PL 1:255,
,29,25;MQ=60;FQ=-999;PV4=0.0521856,1,1,0.145836 GT:PL
,0,32,32;MQ=60;FQ=-999 GT:PL 1:255,0
0,0,37,22;MQ=60;FQ=-999 GT:PL 1:255,0
=0,0,21,28;MQ=60;FQ=-999 GT:PL 1:255,0
1=1;AC1=1;DP4=0,0,33,42;MQ=18;FQ=-999 GT:PL 1:157,
=0,0,37,33;MQ=60;FQ=-999 GT:PL 1:255,0
0,0,23,21;MQ=60;FQ=-999 GT:PL 1:255,0
=0,0,44,44;MQ=60;FQ=-999 GT:PL 1:255,0
=0,0,36,37;MQ=60;FQ=-999 GT:PL 1:255,0
0,0,25,38;MQ=60;FQ=-999 GT:PL 1:255,0
P4=0,0,19,23;MQ=60;FQ=-999 GT:PL 1:255,0
=0,0,39,33;MQ=60;FQ=-999 GT:PL 1:255,0
0,0,25,20;MQ=60;FQ=-999 GT:PL 1:255,0
=0,0,38,30;MQ=60;FQ=-999 GT:PL 1:255,0
=0;MQOF=0;AF1=1;AC1=1;DP4=0,0,23,25;MQ=60;FQ=-999 GT:PL
0,0,29,31;MQ=60;FQ=-999 GT:PL 1:255,0
0,0,32,23;MQ=60;FQ=-999 GT:PL 1:255,0
GT:PL 1:34,0
GT:PL 1:62,0
GT:PL 1:255,0
```

# VCF file format

Data representation format used to describe **variations** in the genome

## Example

**VCF header**

```
##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">
```

**Mandatory header lines**

**Optional header lines** (meta-data about the annotations in the VCF body)

**Body**

#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

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Deletion**

**SNP**

**Large SV**

**Insertion**

**Other event**

**Phased data** (G and C above are on the same chromosome)





Happy Learning