VCF Format

(modified from: https://learn.gencore.bio.nyu.edu)

File extensions : file.vcf

Variant Calling Format is a tab-delimited text file that is used to describe single nucleotide variants (SNVs) as well as insertions, deletions, and other sequence variations. This is a bit limiting as it is only tailored to show variations and not genetic features (that'll be covered on the next page).

There are 8 required fields for this format:

1. Chromosome Name
2. Chromosome Position
3. ID
   - This is generally used to reference an annotated variant in dbSNP or other curate variant database.
4. Reference base(s)
   - What is the reference's base at this position
5. Alternate base(s)
   - The variants found in your dataset that differ from the reference
6. Variant Quality
   - Phred-scaled quality for the observed ALT
7. Filter
   - Whether or not this has passed all filters – generally a QC measure in variant calling algorithms
8. Info
   - This is for additional information, generally describing the nature of the position/ variants with respect to other data.

Example VCF File

(a) VCF example

Header

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##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">
##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
```

Body

```
1        1   .      ACG  A,AT     40   PASS    .                   GT:DP      1/1:13    2/2:29
1        2   .      C    T,CT     .    PASS    H2;AA=T             GT         0|1       2/2
1        5   rs12   A    G       67   PASS    .                   GT:DP      1|0:16    2/2:20
X      100   .      T    <DEL>    .    PASS    SVTYPE=DEL;END=299  GT:GQ:DP   1:12:.    0/0:20:36
```

(b) SNP

Alignment

```
1234
ACGT
ATGT
 ^
```

VCF representation

```
POS  REF  ALT
2    C    T
```

(c) Insertion

```
12345   POS  REF  ALT
AC-GT   2    C    CT
ACTGT
  ^
```

(d) Deletion

```
1234   POS  REF  ALT
ACGT   1    ACG  A
A--T
 ^^
```

(e) Replacement

```
1234    POS  REF  ALT
ACGT    1    ACG  AT
A-TT
 ^^
```

(f) Large structural variant

Alignment

```
   100       110       120       290       300
    .         .         .         .         .
ACGTACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT------------------------[...]----------GTAC
```

VCF representation

```
POS  REF  ALT    INFO
100  T    <DEL>  SVTYPE=DEL;END=299
```

(g) Resolving ambiguity

Alignment

```
1234567890
TTTCCCTCTA
CTTACCT--A
^  ^   ^^
```

Possible representation

```
POS  REF        ALT
1    TTTCCCTCT  CTTACCTA
```

Possible representation

```
POS  REF  ALT
1    T    C
4    C    A
7    TCT  T
```

Recommended VCF representation

```
POS  REF  ALT
1    T    C
4    C    A
5    CCT  C
```

What software use VCF?

- Output of SNP detection tools such as [GATK](https://software.broadinstitute.org/gatk/) and [Samtools](http://samtools.github.io/)
- Input for SNP feature detection like [SNPeff](http://snpeff.sourceforge.net/)
- [VCF Tools](https://vcftools.github.io/index.html)
- Also the required format for [dbSNP](https://www.ncbi.nlm.nih.gov/projects/SNP/)

How are these files generated?

- SNP callers generate these files as output.
- Haplotyping software also report in this format.
- Any database holding variant information will generally have this format available for download.