VCF format

MSc in Genomic Medicine

Lucy Crooks

26/1/2016

http://samtools.github.io/hts-specs/VCFv4.1.pdf

Detected variants are stored in VCF files

There are multiple header lines which start with #

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER
INFO    FORMAT          1308350-S1001586-03

Chr1  155205669   .       G       T       9007.77    .
AC=2;AF=1.00;AN=2;DP=262;FS=0.000;MLEAC=2;MLEAF=1.00;MQ0=0
GT:AD:DP:GQ:PL    1/1:0,262:262:99:9036,645,0
```

1 line per variant

Can have multiple samples

VCF format

position of variant, allele in the reference and the variant allele are reported

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 1308350-S1001586-03

Chr1 155205669 . G T 9007.77 .
AC=2;AF=1.00;AN=2;DP=262;FS=0.000;MLEAC=2;MLEAF=1.00;MQ0=0
GT:AD:DP:GQ:PL 1/1:0,262:262:99:9036,645,0

VCF format

There is a Phred scaled quality score which tells you the probability that the site is not variant

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 1308350-S1001586-03

Chr1 155205669 . G T 9007.77 . AC=2;AF=1.00;AN=2;DP=262;FS=0.000;MLEAC=2;MLEAF=1.00;MQ0=0 GT:AD:DP:GQ:PL 1/1:0,262:262:99:9036,645,0

VCF format

The info column provides a lot of quality values

#CHROM POS ID REF ALT QUAL FILTER
INFO FORMAT 1308350-S1001586-03

Chr1 155205669 . G T 9007.77 .
AC=2;AF=1.00;AN=2;DP=262;FS=0.000;MLEAC=2;MLEAF=1.00;MQ0=0
GT:AD:DP:GQ:PL 1/1:0,262:262:99:9036,645,0

MQ is mapping quality

DP is the total number of reads

VCF files

The FORMAT column tells you what is in the following column(s)
These give the actual genotype call and details about the call

#CHROM POS ID REF ALT QUAL FILTER
INFO FORMAT 1308350-S1001586-03

Chr1 155205669 . G T 9007.77 .
AC=2;AF=1.00;AN=2;DP=262;FS=0.000;MLEAC=2;MLEAF=1.00;MQ0=0
GT : AD : DP : GQ : PL 1/1 : 0,262 : 262 : 99 : 9036,645,0

GT is the genotype 1 = alternative allele, 0 = reference
AD is the depth of good quality reads split by allele
DP is the sum of these
GQ is genotype quality (Phred scaled) PL are genotype likelihoods