VCF格式解析

VCF是用于描述SNP，INDEL和SV结果的文件，下面所记录的是以GATK软件结果的VCF文件，与SAMtools的结果有点不同

VCF文件可以分为两部分看，最上面#号注释的的部分是对一些参数的解释（看英文能懂的话，下面的解释就不用看了），而下面没#号注释的部分则是各个参数对应的具体的值 一般先关注以下几列信息，从左到右为：

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO
1   17538   rs200046632 C   A   99.60   PASS
1   54421   rs146477069 A   G   88.60   PASS
1   55299   rs10399749  C   T   867.60  PASS
1   61442   rs74970982  A   G   593.03  PASS
1   63268   rs75478250  T   C   357.60  PASS
1   64310   rs367969174 A   G   59.60   PASS
```

- CHROM ： 染色体编号
- POS：参考基因组上variant碱基的位置，如果是INDEL，则该位置是INDEL第一个碱基的位置
- ID：variant的ID，如果在dbSNP中有SNP的id，则显示其id，不然以点表示novel variant。
- REF：参考序列上该位点对应的碱基
- ALT：与参考序列上的碱基相比发生了突变的碱基，即Variant的碱基
- QUAL：Phred格式(Phred_scaled)的质量值，表示该位点存在Variant的可能性，值越高表示突变可能性越大
- FILTER：理想情况下，QUAL这个值应该是用所有的错误模型算出来的，这个值就可以代表正确的变异位点了，但是事实是做不到的。因此，还需要对原始变异位点做进一步的过滤。无论你用什么方法对变异位点进行过滤，过滤完了之后，在FILTER一栏都会留下过滤记录，如果是通过了过滤标准，那么这些通过标准的好的变异位点的FILTER一栏就会注释一个PASS，如果没有通过过滤，就会在FILTER这一栏提示除了PASS的其他信息（如：LowQual）。如果这一栏是一个“.”的话，就说明没有进行过任何过滤。

以上几列是最先关注的，接下来还有两列也是蛮重要的FORMAT和最后一列（最后一列一般为样品名），两者和一起则为基因型信息，前者为格式，后者为对应的数据，如：

```
GT:AD:DP:GQ:PL  0/1:6,5:11:99:138,0,153
```

- GT：表示样品的基因型，对于二倍体生物，GT值表示的是样本在这个位点所携带的两个等位基因。0表示跟REF一样，1表示跟ALT一样，2表示有第二个ALT；当只有一个ALT等位基因时：0/0表示纯合子并跟REF一致；0/1表示杂合子，有两个allele，一个是ALT，另一个是REF；1/1表示纯合子并都为ALT
- AD：两个以逗号分隔的值，分别表示覆盖到REF和ALT碱基的reads数，也就是REF和ALT对应的测序深度
- DP：表示覆盖在这个位点的总reads数，也就是这个位点的测序深度（并不是指具体有多少个reads数量，而是大概满足一定质量值要求的reads数）
- PL：三个逗号分隔的值，分别对应该位点的三个基因型0/0，0/1，1/1的没经过先验的标准化Phred-scaled似然值（L），L=-10lgP，P为支持该基因型的概率，3个概率总和为1；因此，L这个值越小，支持概率就越大，也就是说是这个基因型的可能性越大。
- GQ：表示基因型的质量值，Phred格式(Phred_scaled)的质量值，Phred值 = -10 * log (1-p) p为基因型存在的概率，表示该位点基因型存在的可能性。

最后则是INFO列所包含的信息：

```
AC=1;AF=0.500;AN=2;BaseQRankSum=0.748;ClippingRankSum=0.000;DB;DP=34;ExcessHet=3.0103;FS=3.424;MLEAC=1;MLEAF=0.500;MQ=31.07;MQRankSum=-0.087;QD=11.87;ReadPosRankSum=-1.349;SOR=2.636
AC=2;AF=1.00;AN=2;DB;DP=14;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=31.60;QD=29.36;SOR=5.421
```

- AC：表示该Allele的数目，Allele数目为1表示双倍体的样本在该位点只有1个等位基因发生了突变
- AF：表示Allele的频率，Allele频率为0.5表示双倍体的样本在该位点只有50%的等位基因发生了突变
- AN：表示Allele的总数目

即：对于1个diploid sample而言：则基因型 0/1 表示sample为杂合子，Allele数为1(双倍体的sample在该位点只有1个等位基因发生了突变)，Allele的频率为0.5(双倍体的 sample在该位点只有50%的等位基因发生了突变)，总的Allele为2； 基因型 1/1 则表示sample为纯合的，Allele数为2，Allele的频率为1，总的Allele为2。

- DP：样本在这个位置的reads覆盖度，是一些reads被过滤掉后的覆盖度（跟上面提到的DP类似）
- FS：使用Fisher’ s精确检验来检测strand bias而得到的Fhred格式的p值，值越小越好
- MQ：表示覆盖序列质量的均方值RMS Mapping Quality
- BaseQRankSum：Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities
- ClippingRankSum：Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases
- ExcessHet：Phred-scaled p-value for exact test of excess heterozygosity
- MLEAC：Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each ALT allele, in the same order as listed
- MLEAF：Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each ALT allele, in the same order as listed
- MQRankSum：Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
- QD：Variant Confidence/Quality by Depth
- ReadPosRankSum：Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
- SOR：Symmetric Odds Ratio of 2x2 contingency table to detect strand bias

参考：

https://www.biostars.org/p/187068/

http://www.bio-info-trainee.com/863.html

http://blog.sina.com.cn/s/blog_74cbb8e80101f8ic.html

http://blog.sina.com.cn/s/blog_12d5e3d3c0101qv1u.html