实验记录　　序号：02

摘要： 基于已有的重测序 SNP 数据，练习使用 python 脚本筛选浙江产地三叶青特有的 SNP 位点，用于后续分子标记开发。

1. 实验准备

从老师处获得 29 个产地 139 个体的 SNP 数据，了解 vcf 文件的存取格式。

2. 实验步骤

2.1. vcf 文件处理

连接远程服务器，使用 linux 命令行从 vcf 文件中获取每个个体的基因型文件。

对 139 个样本的 vcf 文件进行处理：

```
#将前面的注释信息去掉，只保留矩阵数据
cat 139_文件名.vcf | tail -n +245 >xxx.vcf

#生成gt.txt文件（使用脚本vcf_keep_gt_only_1.py）
cat xxx.vcf | python vcf_keep_gt_only_1.py >xxx.gt.txt

#将矩阵中的所有的斜杠（/）替换为竖线（|）
sed -i 's/\//\|/g'  xxx.gt.txt

#提取出vcf文件中的1、2、4、5行
awk '{print $1,$2,$4,$5}' xxx.vcf >site.txt

#用awk 和paste 把 ref和alt 列合并到gt.txt文件中
awk '{print $3,$4}' site.txt>site2.txt
paste site2.txt xxx.gt.txt >139.gt.txt
```

2.2. 利用 python 脚本筛选浙江产地三叶青特有的 SNP 位点。

基于所有个体基因型文件，利用 python 脚本选浙江产地三叶青特有的 SNP 位点。

```

#然后使用脚本将txt文件转换为csv文件（方便后面直接得到表格结果）
python txt_transform_csv.py 139.gt.txt 139.gt.csv
#筛选出位点
python vcf数据处理.py
```

3. 实验结果

最后鉴定到 19 个浙江产地特有的 SNP 位点，用于开发能够鉴定浙产三叶青的分子标记。

实验记录　　序号：02

```
→  jianding python "/home/shanshan/wanghaotian/三叶青项目/vcf数据处理.py"

读取的139.SNP文件的形状： (27105626, 143)
筛选出的以ZJ开头的列： (27105626, 15)
其他剩余个体列:(27105626, 124)
以ZJ开头的个体里基因型全是1|1的位点： (25176, 143)
其他剩余个体里面全是0|0的位点： (512, 143)
同时满足的位点情况： (0, 143)
ZJ开头的个体里面基因型全是0|0 (13070151, 143)
其他个体里面是1|1 (108, 143)
同时满足： (0, 143)
ZJ里面全是1|1： (25176, 143)
其他个体里面没有1|1这种情况: (12442992, 143)
同时满足： (5, 143)
ZJ里面是0|0： (13070151, 143)
其他个体里面没有0|0的： (291438, 143)
同时满足： (14, 143)
总的输出结果:       REF     ALT #CHROM       POS AHQM01 AHQM02 AHQM04 AHQM06 AHQM08
JSC12
0         G       T  chr05   27776741    0|0    0|0    0|0    0|1    0|0    0|0  ...
1         A       G  chr05   33176151    0|1    0|1    0|1    0|1    0|1    0|0  ...
2     GAAAA  GAAAAA  chr19   64739874    0|1    0|1    0|1    0|0    0|0    0|0  ...
3         G       A  chr19   64739910    0|1    0|1    0|1    0|0    0|1    0|0  ...
4         A       G  chr19   64739919    0|1    0|1    0|1    0|0    0|1    0|0  ...
0         G      GA  chr02   27491153    1|1    1|1    0|1    1|1    0|1    1|1  ...
1         A       C  chr02   34086968    0|1    0|1    0|1    0|1    0|1    1|1  ...
2         A       G  chr05   27651111    0|1    0|1    0|1    0|1    1|1    1|1  ...
3         T       C  chr05   27651359    0|1    0|1    0|1    0|1    1|1    1|1  ...
4         C       T  chr05   27651453    1|1    1|1    0|1    0|1    1|1    1|1  ...
5         A       G  chr05   27664954    0|1    0|1    0|1    0|1    1|1    1|1  ...
6         A       G  chr05   27664980    1|1    1|1    0|1    0|1    1|1    1|1  ...
7         A       G  chr05   27682024    1|1    1|1    0|1    0|1    1|1    1|1  ...
8         T       C  chr05   74876883    0|1    0|1    0|1    0|1    0|1    1|1  ...
9         C       T  chr12   78244735    1|1    0|1    1|1    0|1    0|1    1|1  ...
10        T       G  chr20    5095086    1|1    1|1    1|1    1|1    1|1    1|1  ...
11        C       T  chr20    5095427    0|1    1|1    1|1    0|1    1|1    1|1  ...
12        C       T  chr23  102019144    1|1    1|1    1|1    1|1    .|.    1|1  ...
13        A       G  chr26   17497735    0|1    0|1    0|1    0|1    1|1    1|1  ...

[19 rows x 143 columns]
```