一、操作软件

在系统中安装plink和Beagle软件进行数据清洗，后续使用Rstudio进行模型预测

二、数据清洗

1.使用plink把vcf文件转为PCA的主成分：

plink --vcf filter\_269\_indel.vcf --double-id --allow-extra-chr --recode --out 269\_indel1

2.检查基因型和样本缺失情况

plink --file 269\_indel1 --allow-extra-chr --missing

3.先对基因型进行过滤

plink --file 269\_indel1 --allow-extra-chr --geno 0.1 --recode --out 269\_indel2

4.接着对个体进行缺失控制

plink --file 269\_indel2 --allow-extra-chr --mind 0.2 --recode --out 269\_indel3

5.次等位基因频率（MAF）控制，首先对基因进行检验

plink --file 269\_indel3 --allow-extra-chr --freq

6.过滤基因频率小于0.05的位点

plink --file 269\_indel3 --allow-extra-chr --maf 0.05 --recode --out 269\_indel4

7.将plink文件转化为vcf后进行基因型填充

plink --file 269\_indel4 --allow-extra-chr --recode vcf-iid --out 269\_indel\_GENO\_I

8.填充

Java -Xmx512g -jar beagle.22Jul22.46e.jar gt=269\_indel\_GENO\_I.vcf out=imputed\_269\_indel nthreads=64

beagle.22Jul22.46e.jar

9.解压imputed\_269\_indel.vcf.gz

gunzip imputed\_269\_indel.vcf.gz

10.对vcf文件转化为012文件(raw格式)

plink --vcf imputed\_269\_indel.vcf --double-id --recodeA --allow-extra-chr --out 269\_indel\_geno0.1\_maf0.05\_mind0.2\_imputed\_01

三、环境配置

在相应文件夹中安装anaconda环境，使用jupyter notebook运行与调试代码

以下为所需的库



需要注意的是：XGBoost是一个独立的算法库，在使用时需要单独安装xgboost，此外，Mac系统不支持GPU加速运算，只有Linux系统支持多个GPU联合运算。

关键参数



四、模型预测

环境安装好后在文件夹中打开任务管理器





输入active激活环境



后输入jupyter notebook进入操作界面进行操作

后续将对应snp、indel、SV对应文件分别放入以下对应代码中，可得所需模型预测数据

五、备注

XGBOOST代码位于XGBOOST.txt

scaler.joblib文件为数据预处理后产生的数据文件

modle.joblib文件为模型训练文件

