**#版本：PGG.Selection.Burden**

**1.前期准备**

在进行分析之前需要对数据进行预处理和质控，推荐使用GATK进行jointcalling和VQSR的过滤。

⼀般我们在下游遗传结构的分析中，⼤多采⽤经过质控的biallelic SNP，因此还需⼀步操作来提取这些位点。

bcftools view -f PASS -m 2 -M 2 -v snps your.jointcalling.VQSR.variants.vcf.gz |bgzip -@ 16 -c > your.jointcalling.VQSR.variants.PASS.biallelic.SNPs.vcf.gz

tabix -p vcf your.jointcalling.VQSR.variants.PASS.biallelic.SNPs.vcf.gz

此外，我们还需要对数据进行phasing, 需分染色体进行，如用shapeit5处理，可参考：

https://github.com/Shuhua-Group/PGG.Panel/blob/main/PGG.Panel数据操作文档.pdf step4

如果数据已经经过相应处理，请忽略上述步骤。

**2. 正式运算**

一、准备工作

基础软件安装:

在开始前，请先安装以下工具： bcftools, vcftools, plink, bgzip

# 安装miniconda（用户目录）

wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86\_64.sh

bash Miniconda3-latest-Linux-x86\_64.sh -b -p $HOME/miniconda

source $HOME/miniconda/bin/activate

# 通过conda安装

conda install -c bioconda bcftools vcftools plink -y

conda install bioconda::plink2

1. 流程步骤
2. prepare

# 创建项目目录并进入

mkdir PGG\_selection\_burden\_pipeline

cd PGG\_selection\_burden\_pipeline

准备输入文件

在父目录PGG\_selection\_burden\_pipeline下建立输入文件的存放目录并进入

mkdir 00.input; cd 00.input

将以下文件放入'00.input/'：

- 样本信息文件

准备自然人群样本信息文件group.txt，按民族分组，第一列为民族，若为汉族则为民族加地区，第二列为样本名，以制表符分隔，格式如下：

Han\_Shanghai HG00123

Han\_Shanghai HG00124

Hui HG00125

Zhuang HG00126

- 原始VCF文件（.vcf.gz）、按地区分组提取子vcf文件(.vcf.gz和.vcf)及，根据原始vcf按地区分组提取子vcf文件的脚本如下。

info='group.txt' #替换为你的样本信息文件

for group in `awk '{print $1}' $info|sort -u`

do

mkdir -p $group

for k in {1..22}

do

vcf="your.chr$k.phased.vcf.gz" #替换为你的vcf文件

awk '$1=="'$group'"{print $2}' group.txt > $group/$group.list

bcftools view -S $group/${group}.list $vcf -Oz -o $group/${group}.chr${k}.vcf.gz

bgzip -d -c $group/${group}.chr${k}.vcf.gz > $group/${group}.chr${k}.vcf

done

done

注：如样本间存在亲缘关系，请在除了genetic burden以外的分析去除二级及以上亲缘关系后进行分析，并提供去除样本的list。可参考:

https://github.com/Shuhua-Group/PGG.Panel/blob/main/PGG.Panel数据操作文档.pdf step2

##########################以下脚本可同时运行，无先后关系###########################

1. Frequency

在父目录PGG\_selection\_burden\_pipeline下建立frq的工作目录

mkdir 01.frq; cd 01.frq

将以下脚本下载到工作目录中

https://github.com/Shuhua-Group/PGG.Selection.Burden/blob/main/frq.sh

对chr1-22执行脚本

for k in {1..22}

do

sh frq.sh $k

done

检查各个子目录中的结果文件（\*.chr\*.afreq），示例如下：

#CHROM ID REF ALT ALT\_FREQS OBS\_CT

22 22:10519276 G C 0 206

22 22:10519325 G A 0 206

22 22:10519389 T C 0 206

1. Theta\_D\_H

在父目录PGG\_selection\_burden\_pipeline下建立Theta\_D\_H的工作目录

mkdir 02.Theta\_D\_H; cd 02.Theta\_D\_H

将以下脚本下载到工作目录中

https://github.com/Shuhua-Group/Theta\_D\_H.Est/blob/master/Theta\_D\_H.Est

https://github.com/Shuhua-Group/PGG.Selection.Burden/blob/main/Theta\_D\_H.sh

对chr1-22执行脚本

for k in {1..22}

do

sh Theta\_D\_H.sh $k

done

检查各个子目录中的结果文件（\*.chr\*.gz），示例如下：

regionID chr start end #sequence #marker #singleton ThetaPI ThetaK #segregating #haplotype Hap\_diversity Hfaywu norm\_Hfaywu Ffuli Dfuli Dtajima Dtajima\_P Dtajima\_adj.P

22 22 1 50000 206 0 0 0.0 0.0 0 0 0.0 NA NA NA NA NA NA NA

1. IHS

软件安装

selscan下载链接：

https://github.com/szpiech/selscan/blob/master/releases/selscan-linux-2.0.0.tar.gz

selscan安装教程：

https://github.com/szpiech/selscan/blob/master/INSTALL

predictGMAP下载链接：

https://github.com/szpiech/predictGMAP/tree/master/src

predictGMAP安装教程：

https://github.com/szpiech/predictGMAP/blob/master/README

plink map文件下载：

https://bochet.gcc.biostat.washington.edu/beagle/genetic\_maps/plink.GRCh38.map.zip

在父目录PGG\_selection\_burden\_pipeline下建立ihs的工作目录并进入

mkdir 03.IHS; cd 03.IHS

将以下脚本下载到工作目录中

https://github.com/Shuhua-Group/PGG.Selection.Burden/blob/main/ihs.sh

对ihs.sh进行编辑，将以下三行替换为你的predictGMAP路径、selscan路径、plink map路径

predictGMAP\_d='/home/sunyumeng/software/predictGMAP'

selscan\_d='/home/sunyumeng/software/selscan-linux-2.0.0'

map\_d='/home/sunyumeng/data/PLINK\_format\_genetic\_map'

对chr1-22执行脚本

for k in {1..22}

do

sh ihs.sh $k

done

检查各个子目录中的结果文件（\*.chr\*.ihs.out.100bins.norm），示例如下：

chr22:10634429 10634429 0.140777 0.00360501 0.00388136 -0.0320774

检查各个子目录中的结果文件（\*.chr\*.ihs.out.100bins.norm），示例如下：

chr22:10634429 10634429 0.140777 0.00360501 0.00388136 -0.0320774 -1.27343 0

检查各个子目录中的结果文件（\*.chr\*.ihs.out.100bins.norm.50kb.windows），示例如下：

1 50001 0 -1 -1 NA

1. CLR

软件安装

wget http://sco.h-its.org/exelixis/resource/download/software/SweeD\_v3.2.1\_Linux.tar.gz

tar -xzf SweeD\_v3.2.1\_Linux.tar.gz

在父目录PGG\_selection\_burden\_pipeline下建立CLR的工作目录并进入

mkdir 04.CLR; cd 04.CLR

将以下脚本下载到工作目录中

https://github.com/Shuhua-Group/PGG.Selection.Burden/blob/main/clr.sh

对clr.sh进行编辑，将下行替换为你的SweeD下载路径

SweeD\_d='/home/sunyumeng/software/SweeD\_v3.2.1\_Linux'

对chr1-22执行脚本

for k in {1..22}

do

sh clr.sh $k

done

检查各个子目录中的结果文件（SweeD\_Report.\*.chr\*.50kb），示例如下：

//1

Position Likelihood Alpha

10522570 7.786579e-01 1.635287e-03

1. Genetic Burden

在父目录PGG\_selection\_burden\_pipeline下建立Genetic Burden的工作目录并进入

mkdir 05.burden; cd 05.burden

将22个DamageSnp文件下载到工作目录中

百度网盘链接: https://pan.baidu.com/s/1HZY9A6aSbyvGypayl0cRWw 提取码: ky8i

将pathway.input.txt文件下载到工作目录中

https://github.com/Shuhua-Group/PGG.Selection.Burden/blob/main/pathway.input.txt

将以下脚本下载到工作目录中

https://github.com/Shuhua-Group/PGG.Selection.Burden/blob/main/burden.sh

https://github.com/Shuhua-Group/PGG.Selection.Burden/blob/main/burden.py

对chr1-22执行脚本

for k in {1..22}

do

sh burden.sh $k

done

检查各个子目录中的结果文件（\*.chr\*.whole.burden.txt），示例如下：

##Total number of deleterious SNVs: 1

##Total number of loss of function SNVs: 0

#sample geo sum\_Het ... ... CADD\_weigted\_Aloft\_Dominant\_Hom

NA18525 test 1 ... ... 1

... ...

检查各个子目录中的结果文件（\*.chr\*.pathway.burden.txt），示例如下：

##Total number of deleterious SNVs of hsa00020: 1

##Total number of deleterious SNVs of hsa00010: 1

##Total number of deleterious SNVs of hsa00030: 0

... ...

##Total number of deleterious SNVs of hsa05418: 0

#sample geo hsa00020\_Het ... ... hsa05418\_Het hsa05418\_Hom

NA18525 test 1 ... ... 1 0

... ...

检查各个子目录中的结果文件（\*.chr\*.gene.burden.txt），示例如下：

##Total number of deleterious SNVs of 10327: 0

##Total number of deleterious SNVs of 124: 0

##Total number of deleterious SNVs of 125: 0

... ...

##Total number of deleterious SNVs of 57534: 0

#sample geo 10327\_Het ... ... 57534\_Het 57534\_Hom

NA18525 test 1 ... ... 1 0

... ...

**3.输出结果文件**

将整个工作目录，除了输入文件（00.input）外，全部打包

rm -rf PGG\_selection\_burden\_pipeline/00.input

tar -czvf PGG\_selection\_burden\_results.tar PGG\_selection\_burden\_pipeline