major/minor/reference/alternative/risk/effect allele 概念解析

这些名词很容易混淆而引起不必要的错误或误解，早期的遗传统计学软件，例如plink并没有很重视allele概念上的明确区分，但近年新出的软件或旧软件的新版本为保证统一性已经开始注意此问题。

本文内容

```
第一组 频率上的 major 与 minor allele
第二组 参考基因组的 reference （ref） 与 alternative （alt） allele
第三组 关联检验的 reference （non-risk 或者 non-effect）与 risk/effect allele
```

GWASLab

首先第一组概念 **major 与 minor allele**

major allele 与 minor allele 通常针对**某一大小确定的特定群体**而言，频率最高的allele为该群体的major allele， 频率次高的为 minor allele，对于最常见的bi-allelic SNP来说，两个allele频率一高一低，就是这个群体中这个snp的major和minor allele，对于tri- 或者quad-allelic SNP （位点有三种或四种碱基的SNP）而言，minor allele则是**频率第二高的那个allele**

注意点：

区分major与minor的依据是 **某一大小确定特定群体的** allele 频率

plink1.9目前采用的是major与minor allele的概念，软件会自动计算频率，对原始数据进行操作时会自动改变allele的排序，如果你使用plink1.9 的—frq选项计算频率，你会发现输出的文件中是MAF ， minor allele frequency，**不会高于0.5。**

PLINK1.9中,A1为minor，A2为major allele，所以这里MAF是指A1（minor allele）的频率。。

```
CHR    SNP    A1   A2          MAF  NCHROBS
1      SNP1    T    C       0.1258    10000
1      SNP2    A    G       0.1258    10000
```

第二组 **reference （ref） 与 alternative （alt） allele**

reference allele 在这里是指**某一参考基因组**上该位点的allele，该位点上其他的allele则称为alternative allele。注意，**这里reference 与 alternative allele与频率无关，唯一的决定因素是所选的参考基因组。**参考基因组上的allele多为major allele，但这只是巧合，不能以此为依据将major和 reference allele划上等号，也有部分reference allele在该群体中为minor allele。

与plink1.9不同，plink2使用的概念则是reference 与 alternative allele，进行操作时不会自动依据频率而改变ref与alt的排序，使用plink2 的—frq选项计算频率，你会发现输出的文件中是alternative allele frequency (不是MAF)，**取值范围为[0,1]。**

PLINK2中则明确区分了reference 与 alternative allele的概念，例如上述的两个SNP，根据参考基因组对齐后，SNP1在参考基因组中的ref为T，那么alt就为C，这里计算的alt的频率为0.8742，按概念来说在该群体中，SNP1的T为ref allele，但却又是minor allele ， 而C为alt，却又是major。 对于SNP2来说ref 则为major，alt 为minor。

```
#CHROM  ID  REF ALT ALT_FREQS   OBS_CT
1   SNP1        T   C   0.8742  10000
1   SNP2        G   A   0.1258  10000
```

小窍门：使用plink2可以将自己手头数据的ref与alt allele与对应参考基因组对齐，示例代码如下：

```
plink2 \
       --bfile testfile \
       --ref-from-fa -fa hg19.fasta \  从参考基因组的fasta文件来决定plink文件中的ref
       --make-bed \
       --out testfile_fa
```

第三组 reference 与 risk/effect allele

在这里的概念再次改变，同样的reference allele，在与 risk/effect allele并列时，则**指的是GWAS关联检测中的reference allele （non-risk 或者 non-effect）**，也就是效应量beta（或odds ratio）估计时的参考，概念上与上述ref与alt的组合无关，但为了保持统一性，近年来研究中关联检验的reference 也会与 reference genome保持一致，以避免混淆等。（注意：早期多以minor allele为关联检验的ref allele，这也是容易产生混淆的点）

risk allele 则很好理解，就是对疾病发生有贡献的那个allele，在复杂疾病的研究中，一般情况下risk allele经常为minor allele，但也会有例外。effect allele的概念也类似，就是我们想要研究其对疾病或表型效应的allele，所以通常是对表型或疾病有贡献的allele，关联检验结果中effect一栏指的就是effect allele的效应。

理解了以上概念后，我们在分辨allele时就能得心应手了。

相关：

GWAS Sumstats Harmonization GWAS数据的协调统一— 背景介绍 在进行meta分析之前，我们首先要对gwas的sumstats进行预处理，这一步看似简单，...
2021年12月20日
在“04 数据质控 Quality Control”中

$$, = \frac{\sum_j I_j(p_{kj} - p_{k'j})}{\sum_j I_j}$$

MR-MEGA 跨族裔GWAS荟萃分析 Meta-analysis
2021年6月17日
在“13 荟萃分析 Meta analysis”中

使用GWAS数据估计跨族裔遗传相关 popcorn – Cross-ancestry genetic correlation popcorn 背景介绍 回顾（在单一族裔中估计遗传相关的方法）： 通过Bivariate LD Sc...
2022年3月28日
在“07 LD score regression”中

2021年8月21日 Yunye 00 群体遗传学基础 population genetics

上一
多基因风险分数 PRS(Polygenic risk score)系列之二：使用PLINK计算PRS（C+T方法）

下一
多基因风险分数 PRS(Polygenic risk score)系列之三：使用PRSice计算PRS（C+T方法）