基因组学-高通量测序实验与数据分析

全基因组关联分析(GWAS)

教学目标

01 全基因组关联分析(GWAS)：概念、原理、用途等

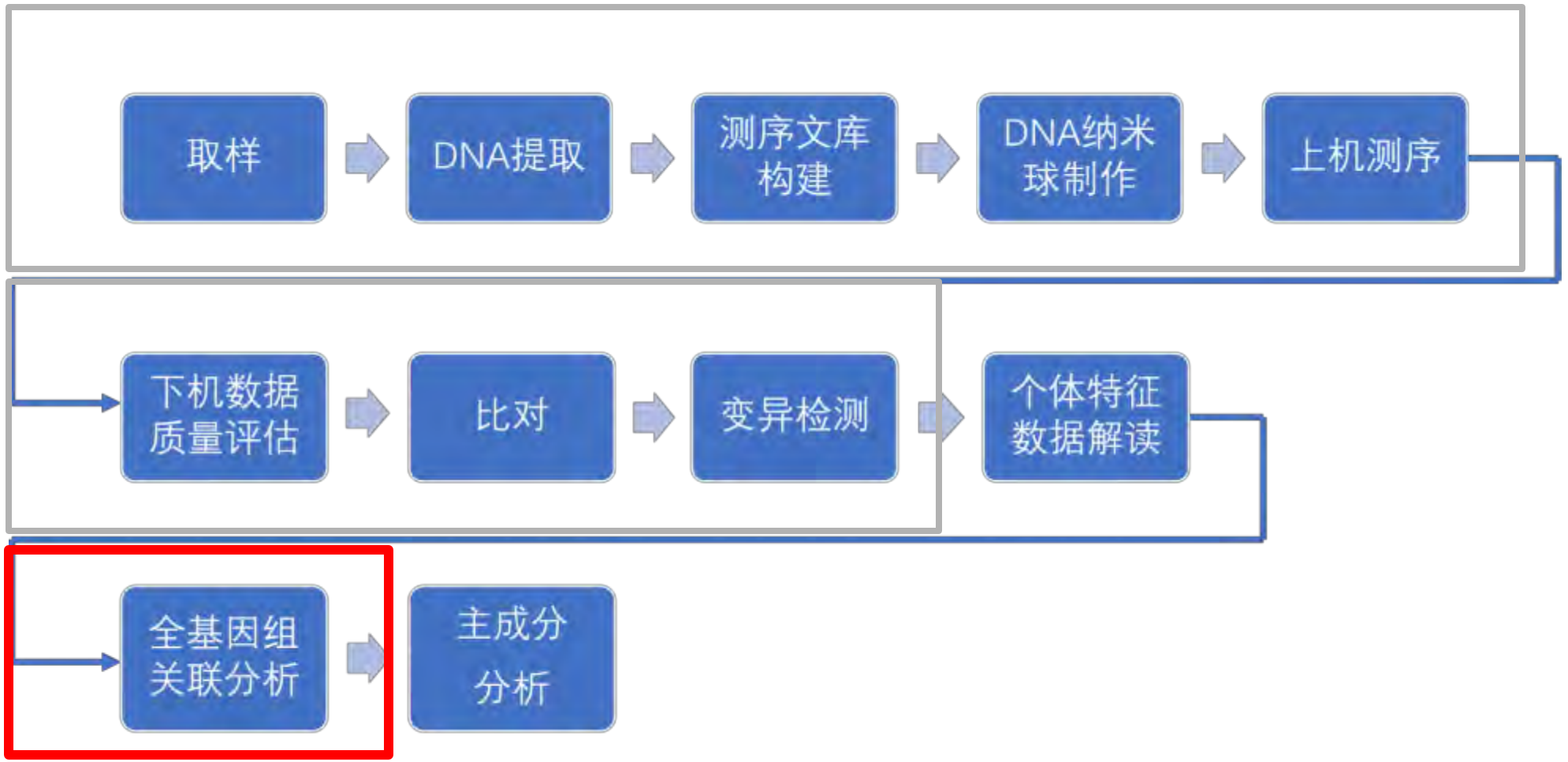
02 GWAS分析环境

03 GWAS分析操作

01 全基因组关联分析(GWAS)

1.1 全基因组关联分析简介

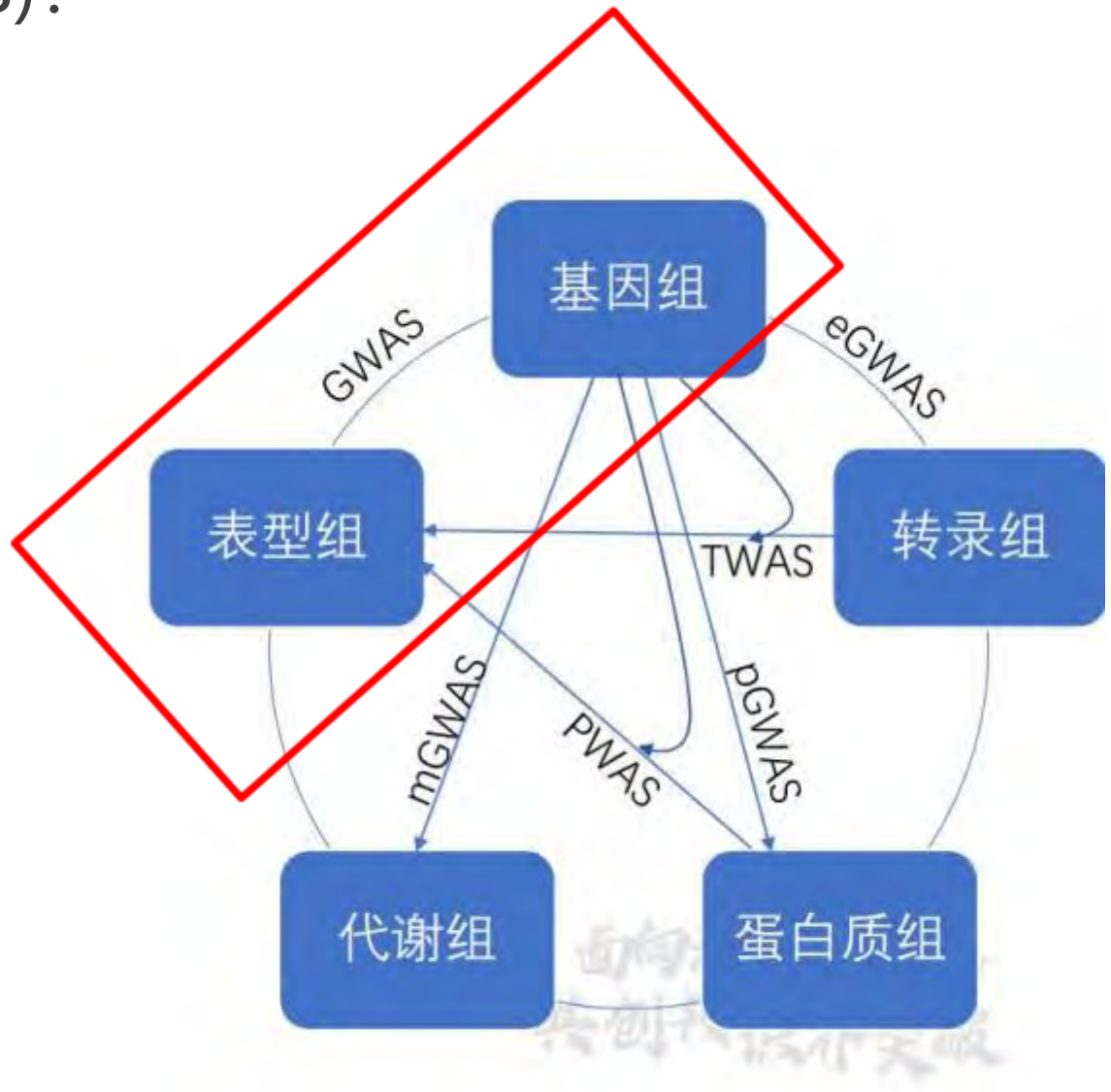
- 测序及分析流程回顾



1.1 全基因组关联分析简介

- 什么是全基因组关联分析(GWAS)?

A genome-wide association study (abbreviated GWAS) is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait.

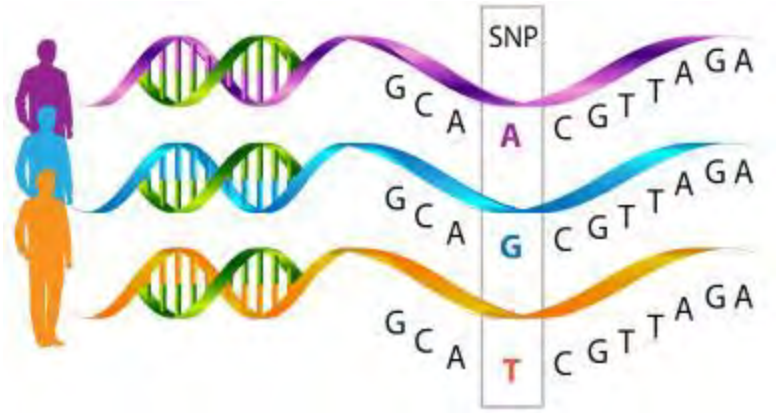


https://www.youtube.com/watch?v=tJjXpiWKMyA

1.1 全基因组关联分析简介

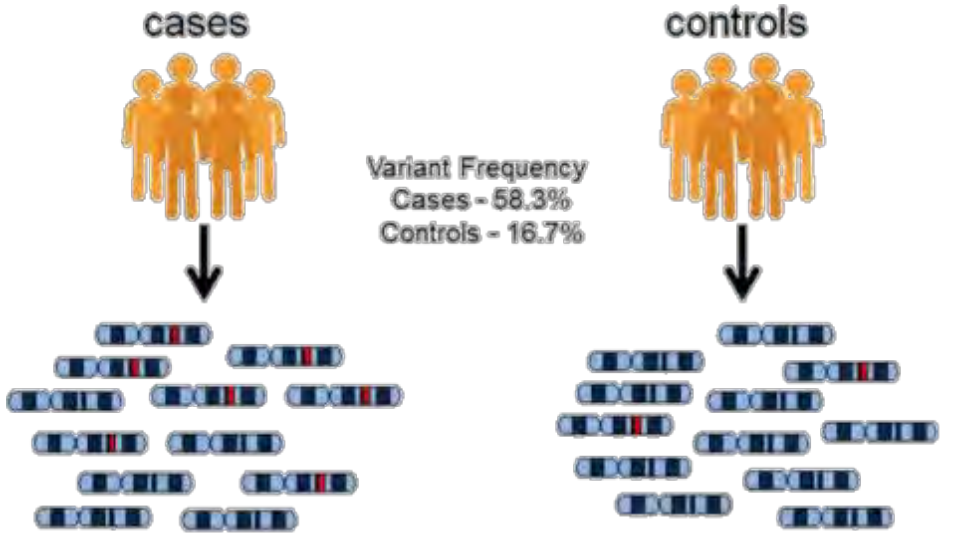
什么是全基因组关联分析(GWAS)?

基因组



表型组

- 全基因组关联研究（GWAS）是识别遗传区域(基因座 loci）和性状（包括疾病）之间关联的分析方法
- 个体之间的遗传变异会导致表型的差异。因此，这些起作用的变异，以及那些与染色体区域紧密相连的变异，在病例（具有该特征的个体）中的出现频率高于对照组（没有该特征的个人）图1.



一个小问题

得肺癌的病人常常是吸烟的，所以吸烟与肺癌有关；

得肺癌的病人常常是吸烟的，而不得肺癌的人常常是不吸烟，
所以吸烟与可能肺癌有关；

肺癌的病人常常吃白米饭，所以吃白米饭与肺癌有关；

先心病人这个位点常常突变，所以这个位点与先心有关；

先心病人这个位点常常突变，而不得先心病的人这个位点常常没有突变，
所以这个位点可能与先心有关。

特别注意，这里有两个问题：
(1) 这里指“有关”，即两者相关，是数据分析需要解决的统计学关联问题；
(2) 同时我们并不清楚两者间何为因何为果，更不清楚这个“因”如何导致“果”，这是基础科研需要解决的问题。

相关分析方法

	肺癌	非肺癌	累计
吸烟	**40**	**4**	44
不吸烟	**10**	**46**	56
总计	50	50	100

```
> chisq.test(c(40,10,4,46))

        Chi-squared test for given probabilities

data:  c(40, 10, 4, 46)
X-squared = 53.28, df = 3, p-value = 1.598e-11
```

基因组位点X	AA	AG	GG
先心病人	**35**	**14**	**3**
非先心病人	**34**	**7**	**20**

A: 35+35+14=84，依次类推

位点X	A	G	Total
病人	**84**	**20**	104
非病人	**75**	**47**	122
	159	67	226

```
> chisq.test(c(84,75,20,47))

        Chi-squared test for given probabilities

data:  c(84, 75, 20, 47)
X-squared = 44.619, df = 3, p-value = 1.115e-09
```

结论：位点X可能与先心有关

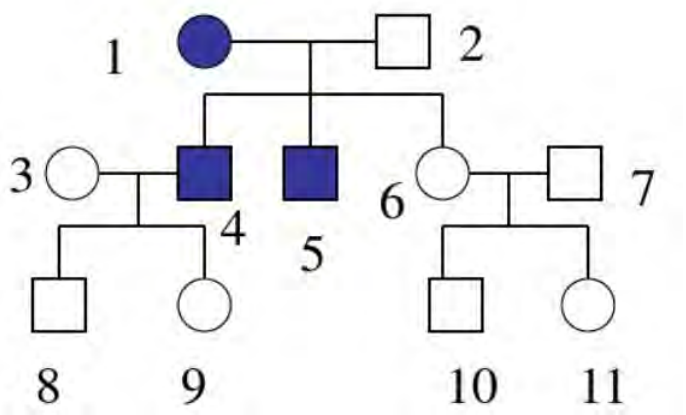
特别注意：检测总数越大，结果越可靠

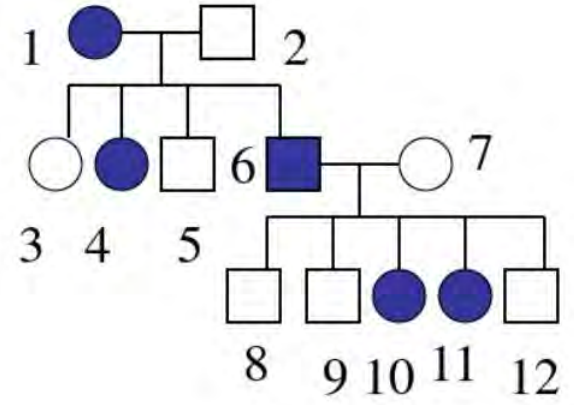
常常面对的临床问题

社会上

临床上

各种疾病





数据这么大怎么办？

人的基因组大约有30亿个碱基，根据 dbSNP 注释可突变位点估计有1.49亿个。
假定这些位点都与先心有关，我们检测了100个先心病人，和100个对照，
那么构成，1.49亿行，和200列的大型矩阵

	patient1	patient2	… …	patient100	control1	control2	… …	control100
rs001	AA	AG	… …	AA	GG	GG	… …	AG
rs002	CC	CA	.. …	CC	AA	AA	… …	CA
rs …	TT	CT	… …	TT	CC	CC	… …	CT
rs …	GG	GA	… …	GG	AA	AA	… …	AA
rs …	AG	AA	… …	AA	GG	GG	… …	AG
… …	… …	… …	… …	… …	… …	… …	… …	… …
rs1.49	AT	AA	… …	AA	TT	TT	… …	AT

GWAS数据分析原理

	patient1	patient2	… …	patient100	control1	control2	… …	control100
rs001	AA	AG	… …	AA	GG	GG	… …	AG
rs002	CC	CA	.. …	CC	AA	AA	… …	CA
rs …	TT	CT	… …	TT	CC	CC	… …	CT
rs …	GG	GA	… …	GG	AA	AA	… …	AA
rs …	AG	AA	… …	AA	GG	GG	… …	AG
… …	… …	… …	… …	… …	… …	… …	… …	… …
rs1.49	AT	AA	… …	AA	TT	TT	… …	AT

基因组位点X	AA	AG	GG
先心病人	**35**	**14**	**3**
非先心病人	**34**	**7**	**20**

A: 35+35+14=84，依次类推

位点X	A	G	Total
病人	**84**	**20**	104
非病人	**75**	**47**	122
	159	67	226

```
> chisq.test(c(84,75,20,47))

        Chi-squared test for given probabilities

data:  c(84, 75, 20, 47)
X-squared = 44.619, df = 3, p-value = 1.115e-09
```

结论：位点X可能与先心有关

	CHISQ	pvalue	OR	ExAC_ALL	… …	gnomAD_exome_ALL	SIFT_pred	Polyphen2_HDIV_pred	Polyphen2_HVAR_pred
rs001	44	1.15e-09	…	…	…	…	…	…	…
rs002	40	5.20e-09	…	…				…	…
rs …	…	…	…						…
rs …	…	…	…						…
rs …	…	…	…						…
… …	…	…	…						…
rs1.49	…	…	…						…

从p值、基因注释、突变类型、人群突变频率、致病潜力等多方面综合考虑，选择相应的位点进行后续研究。

如何确定候选的基因呢？

	CHISQ	pvalue	OR	ExAC_ALL	……	gnomAD_exome_ALL	SIFT_pred	Polyphen2_HDIV_pred	Polyphen2_HVAR_pred
rs001	44	1.15e-09	…	…	…	…	…	…	…
rs002	40	5.20e-09	…	…					…
rs…	…	…	…						
rs…	…	…	…						
rs…	…	…							
……	…	…							
rs1.49	…	…							

从p值、基因注释、突变类型、人群突变频率、致病潜力等多方面综合考虑，选择相应的位点进行后续研究。

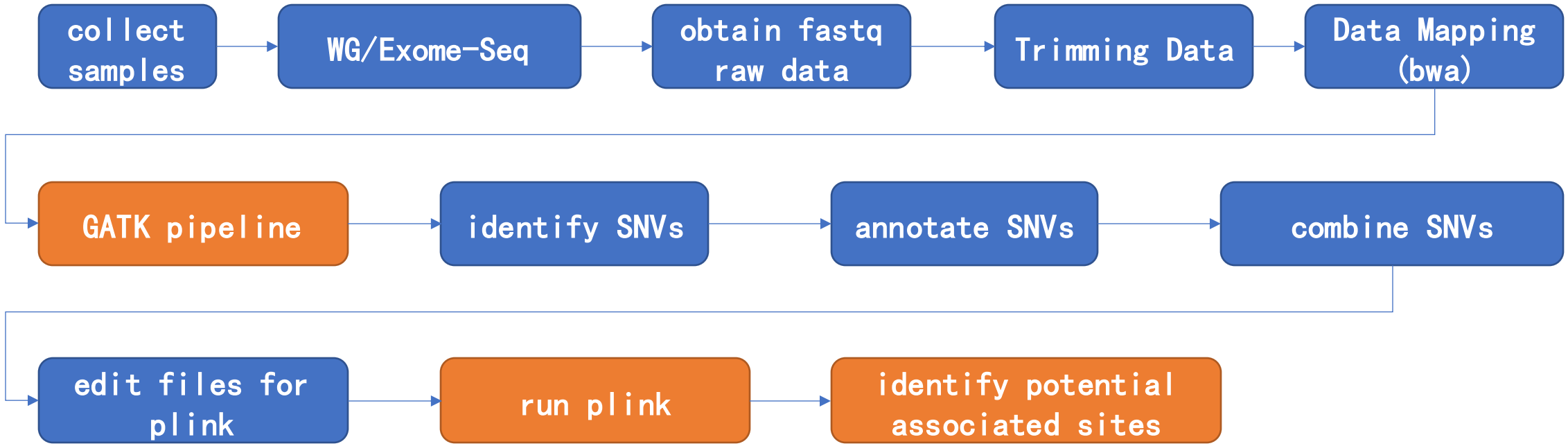
产生关联的可能原因有三：

(a) 是这个位点就是致病位点，
(b) 这个位点与疾病位点存在连锁不平衡（LD），
(c) 这种关联性是由于混杂因素造成的虚假的联系。所谓的混杂因素指的是这一因素与疾病和检测位点均有关联，如遗传背景不同的人群混杂在一起造成的虚假联系，又称为不同分层人群。

如何综合考虑呢？

(1) 首先看p值是否显著？
(2) 然后看突变类型：是不是有意义突变，或者移码突变等可能导致氨基酸序列改变的基因组变异？
(3) 相关基因目前已知有什么功能？是否与发育或细胞生长等相关信号通路有关？
(4) 该位点在正常人群中发生的频率如何？如果发生频率过高，在正常人中普遍存在，则不太可能是致病基因。
(5) 致病潜力预测：用SIFT等软件的预测结果；
(6) 该位点在进化上的保守性；
(7) 该突变位点在蛋白序列中的结构域。

生物信息分析流程


collect samples
WG/Exome-Seq
obtain fastq raw data
Trimming Data
Data Mapping (bwa)
GATK pipeline
identify SNVs
annotate SNVs
combine SNVs
edit files for plink
run plink
identify potential associated sites

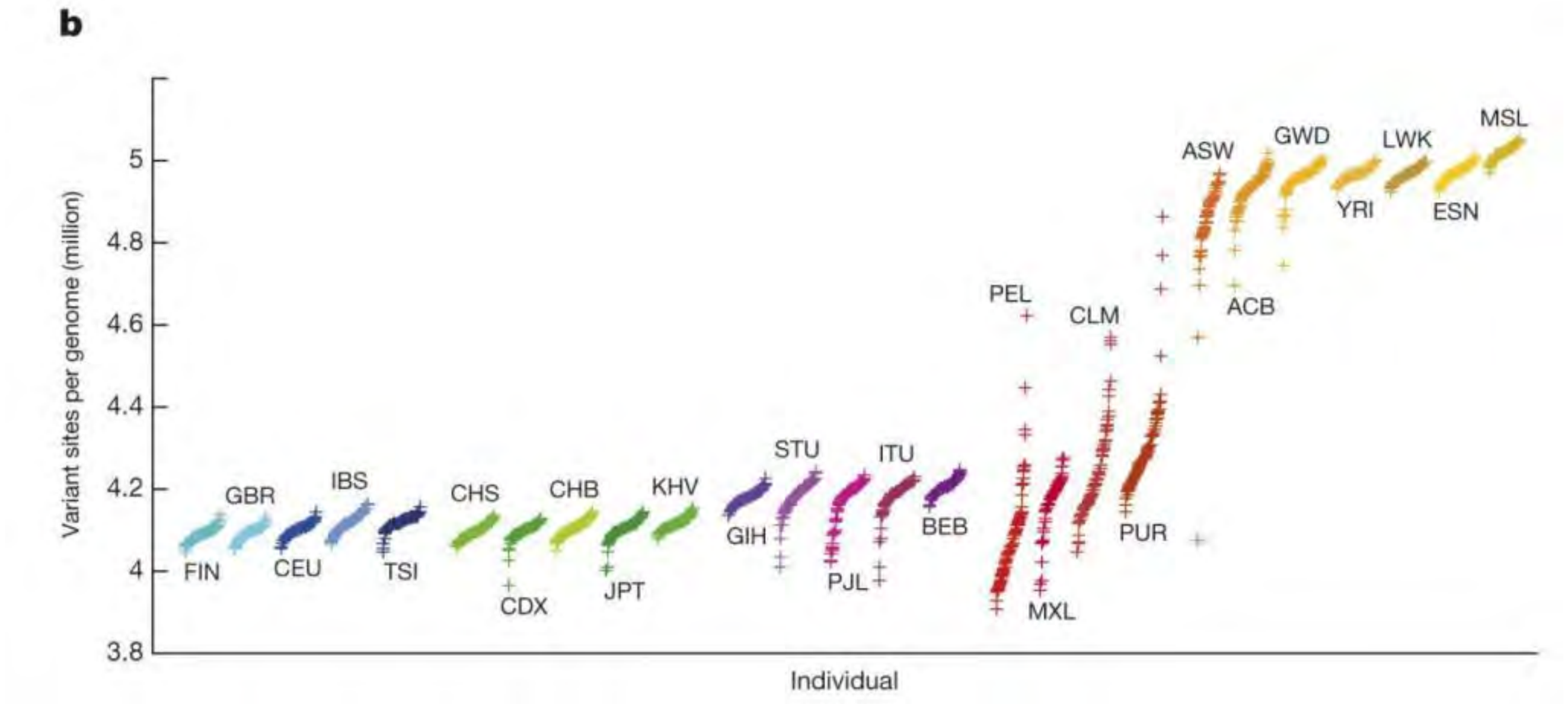
1.2 课程相关性状

- 课程中涉及的25个特征性状

序号	特征性状（25 个）/总共 206 个	影响表型	文献备注
1	rs4988235	乳糖代谢能力	[5][6][7][8][9]
2	rs182549	乳糖代谢能力	[5][6][7][8][9]
3	rs16891982	头发颜色/瞳孔颜色	[10][11]
4	rs28777	头发颜色	[10][11]
5	rs12203592	头发颜色/瞳孔颜色	[10][11]
6	rs4959270	头发颜色	[10][11]
7	rs683	头发颜色	[10][11]
8	rs1815739	肌肉类型	[12][13][14][15][16]
9	rs1042602	头发颜色	[10][11]
10	rs1393350	瞳孔颜色	[10][11]
11	rs12821256	头发颜色	[10][11]
12	rs671	酒精反应	[17][18][19][20][21]
13	rs12896399	瞳孔颜色	[10][11]
14	rs2402130	头发颜色	[10][11]
15	rs1800407	头发颜色/瞳孔颜色	[10][11]
16	rs12913832	头发颜色/瞳孔颜色	[10][11]
17	rs17822931	耳垢类型	[22]
18	N29insA	头发颜色	[10][11]
19	rs1805005	头发颜色	[10][11]
20	rs1110400	头发颜色	[10][11]
21	rs1805008	头发颜色	[10][11]
22	rs885479	头发颜色	[10][11]

1.3 千人基因组计划

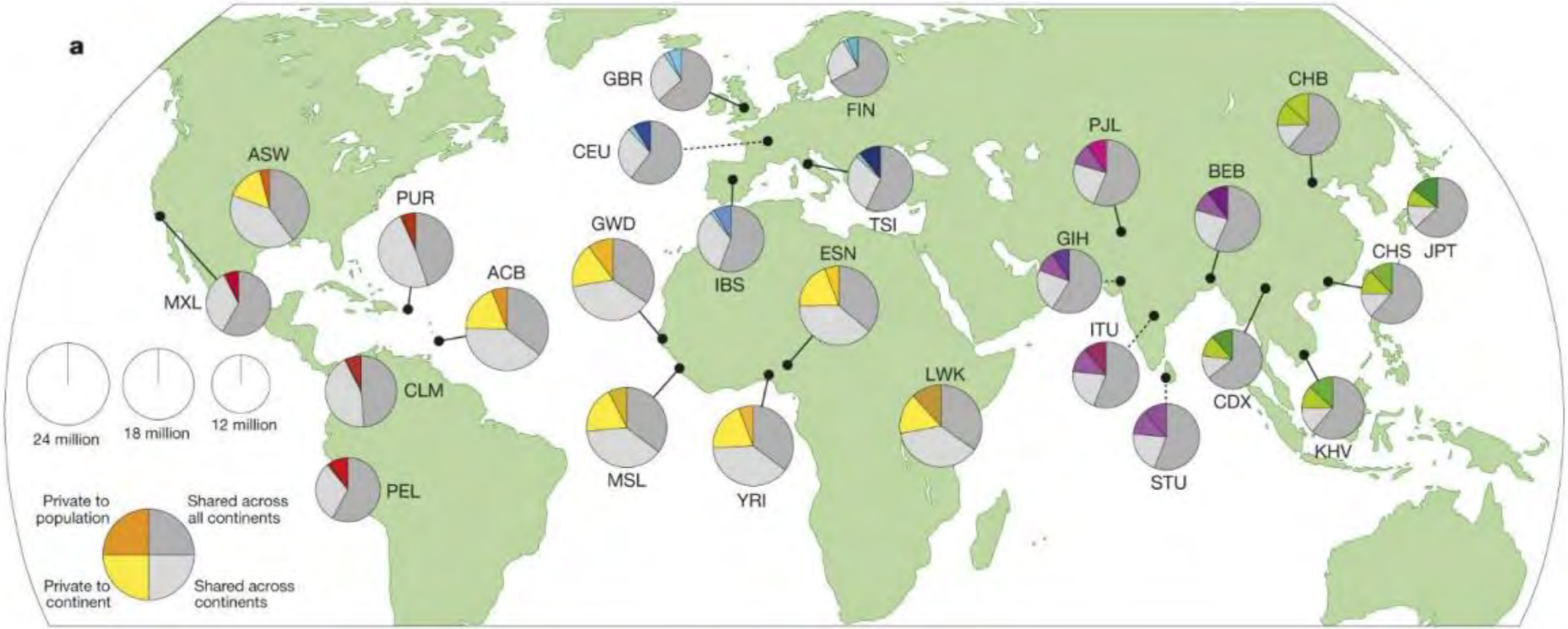
- 相比于参考序列， 410万~500万碱基差异/人，占全基因组约1.67‰



https://www.nature.com/articles/nature15393

1.3 千人基因组计划

- 千人基因组计划



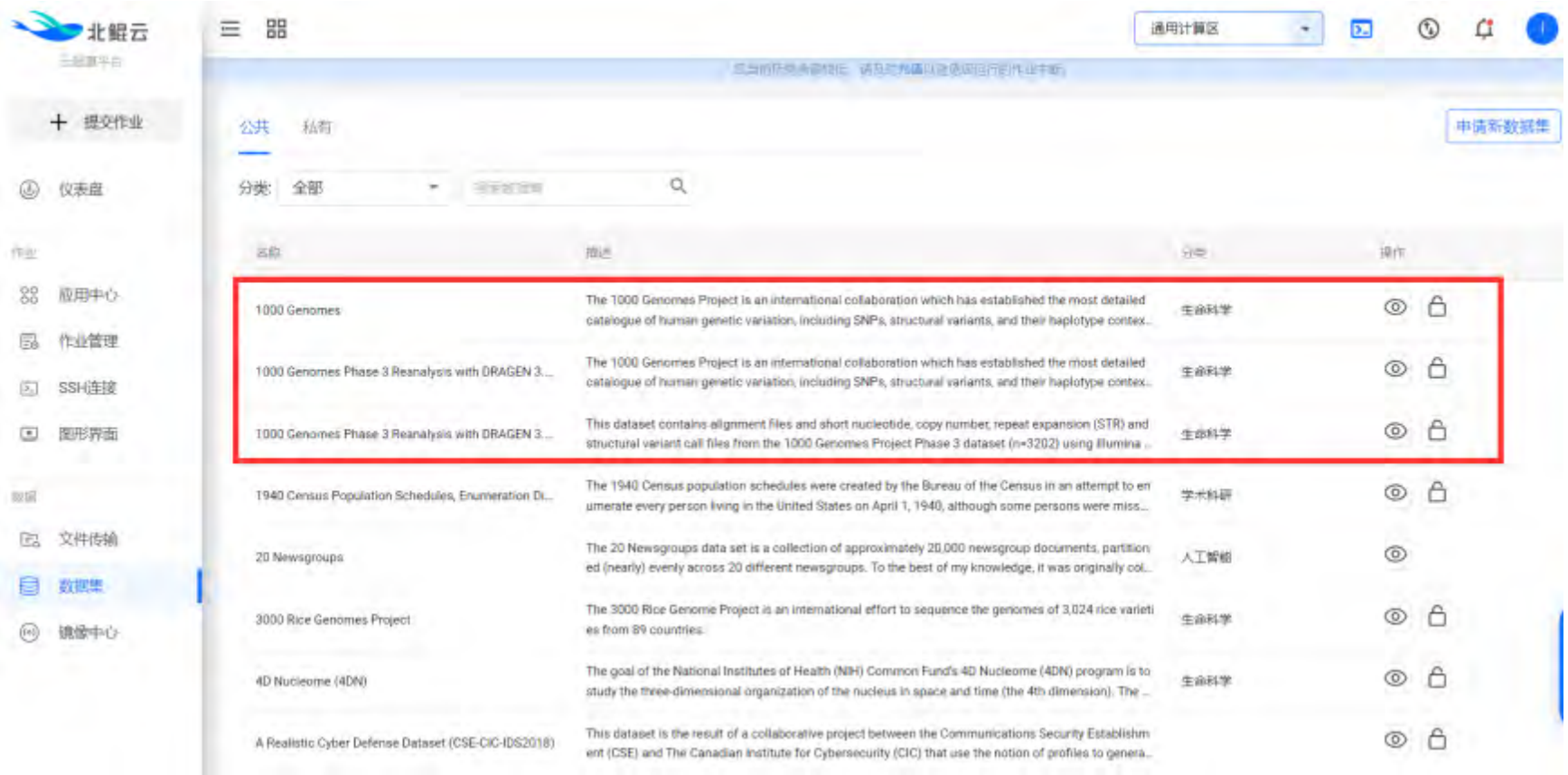
https://www.nature.com/articles/nature15393

1.3 千人基因组计划

- 千人基因组数据集

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/gambian_genome_variation_project/release/20200217_biallelic_SNV/

北鲲云平台：https://www.cloudam.cn/v2/console/cloude-dashboard



1.4 GWAS数据分析模型

- 特定表型与基因型之间的关联关系是如何确定的?

序号	特征性状（25 个）/总共 206 个	影响表型	文献备注
1	rs4988235	乳糖代谢能力	[5][6][7][8][9]
2	rs182549	乳糖代谢能力	[5][6][7][8][9]
3	rs16891982	头发颜色/瞳孔颜色	[10][11]
4	rs28777	头发颜色	[10][11]

- GWAS模型

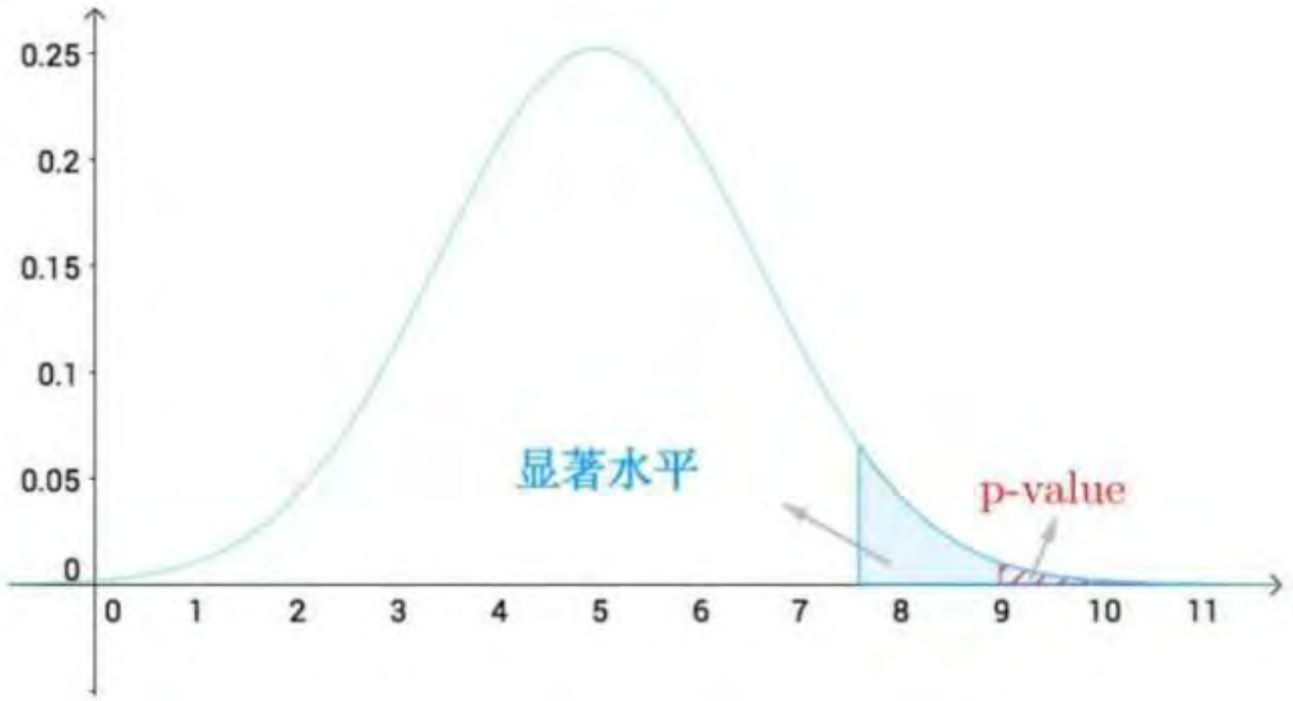
$$\boldsymbol{Y} \sim \boldsymbol{W}\boldsymbol{\alpha} + \boldsymbol{X}_s\boldsymbol{\beta}_s + g + e$$

表型 | 环境效应 | SNP位点效应 | 遗传背景随机效应 | 其他随机效应及误差

https://www.nature.com/articles/s43586-021-00056-9

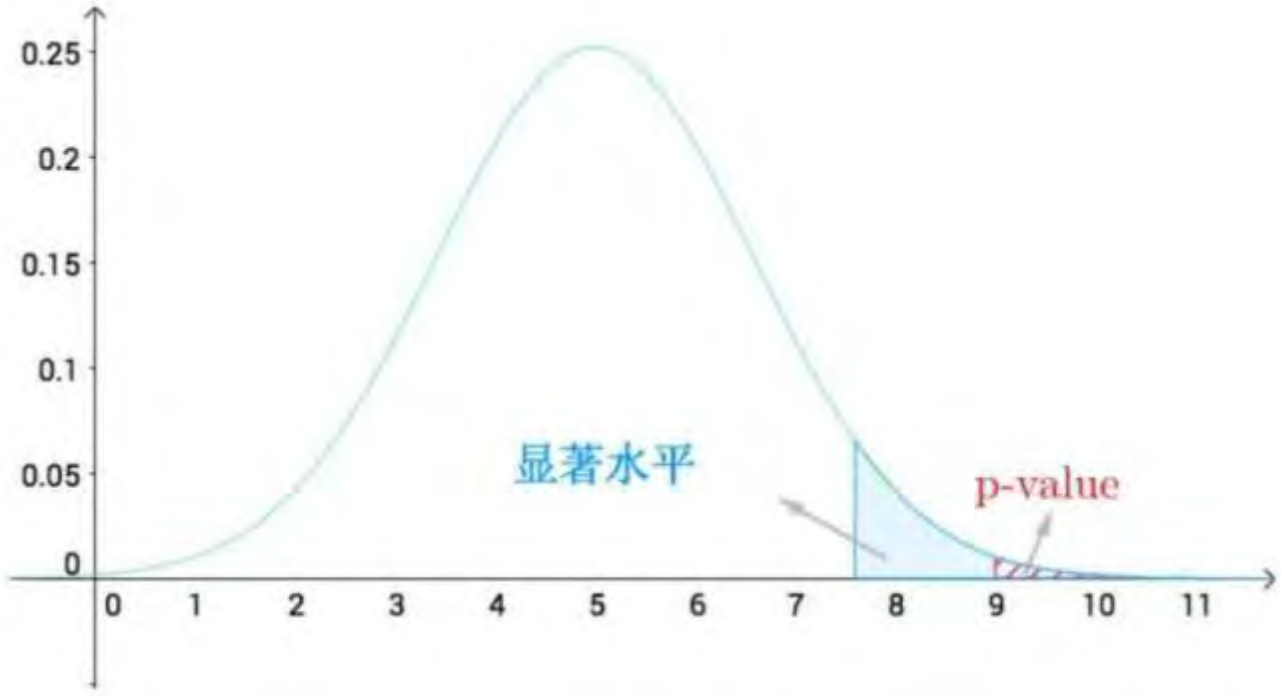
1.4 GWAS数据分析模型

- 0假设：特定基因型（Xsβs ）与表型（Y）无关联
- 选择统计模型，进行检验
- 确定拒绝域（p值< ？ ，则拒绝0假设)
- 求出统计检验的p值
- 查看样本结果是否位于拒绝域
- 判定0假设是否成立，做出决策



1.4 GWAS数据分析模型

- 0假设：rs1800414位点与本次课程涉及到的5种表型无关联
- 选择统计模型，进行检验
- 确定拒绝域（p值< 0.05，则拒绝0假设）
- 求出统计检验的p值
- 查看样本结果是否位于拒绝域（查看p值是不是小于0.05）
- 判定0假设是否成立，做出决策（判断rs1800414位点是否真的与所有5种表型无关？）



1.5 GWAS分析的应用

- GWAS研究使用全基因组SNP阵列(array)收集数据，以便在多个个体中找出基因组中具有或不具有共同特征（如疾病）的共同变异。

- 与疾病相关的变异，通常在病例中发现的频率会高于对照组。然后，可以通过统计分析，来证明这个变异有多大概率是与一个性状（比如：疾病）相关

- 由于GWAS分析的是常见的变异（大多都在商业化的SNP阵列上了（图2）），通常商业化的SNP阵列是不会识别因果变异。

- p值 表示病例和对照组之间检测的SNP频率差异的显著性，即SNP可能与性状相关的概率。
  GWAS结果通常显示在曼哈顿图（图2）中，其中-log10（p值）与基因组中的位置相对应 。

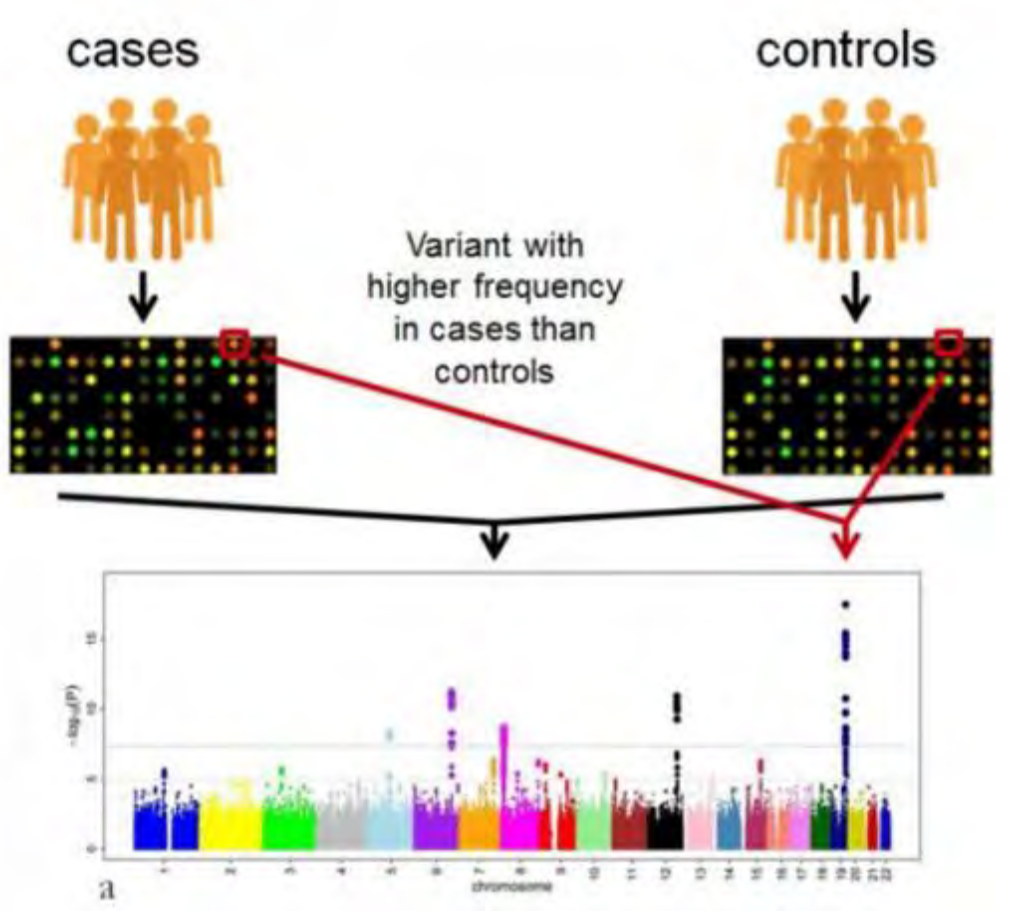


Fig2. Diagram to show the identification of alternative variants in cases and controls using an array-based typing method. Results are subject to statistical analyses to assign a*p-value* to each variant

1.5 GWAS分析的应用

- 全基因组关联分析(GWAS)的应用

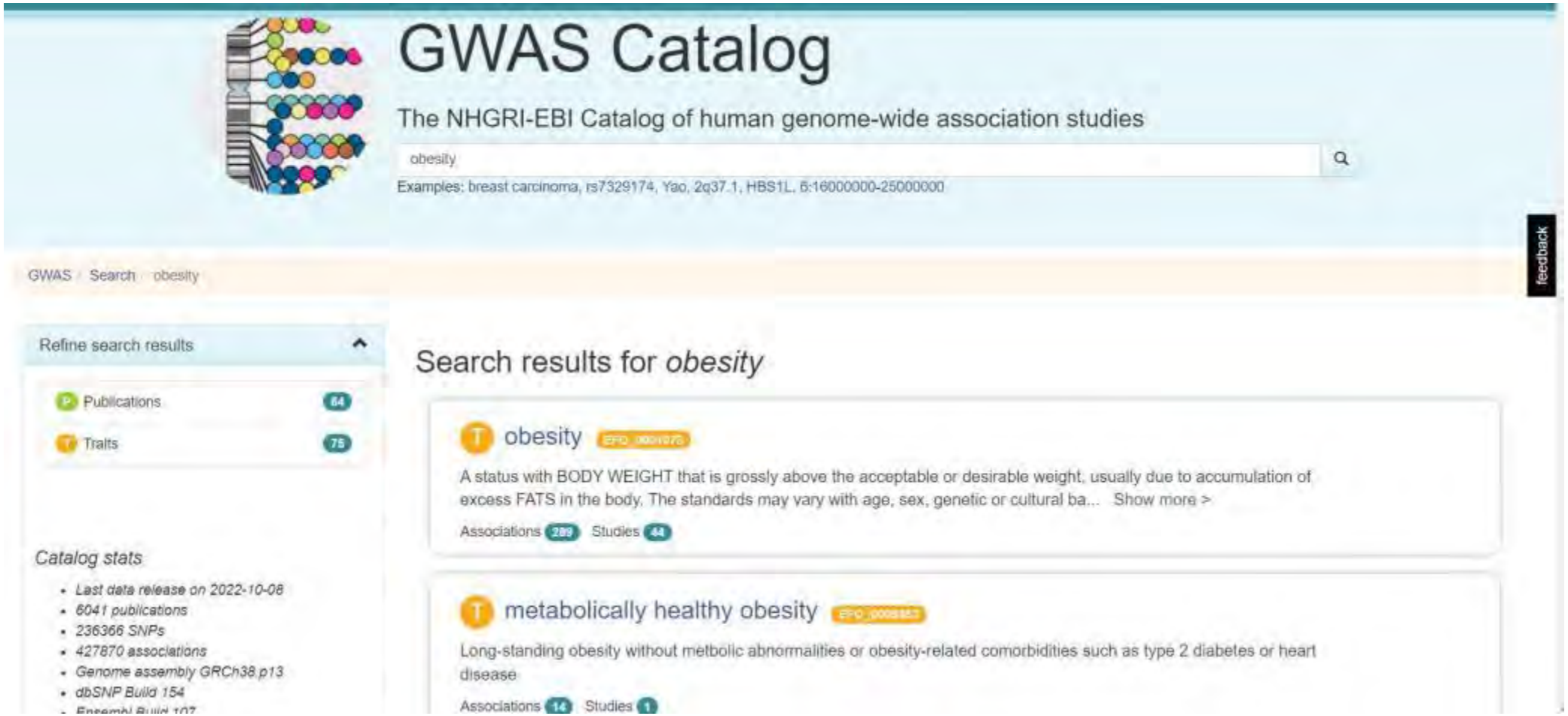
with the trait of interest, known as genomic risk loci. After 15 years of GWAS[1], many replicated genomic risk loci have been associated with diseases and traits[1], such as *FTO*[2] for obesity and *PTPN22* (ref.[3]) for autoimmune diseases. These results have sometimes provided hints into disease biology; for example, a GWAS implicated the IL-12/IL-23 pathway in the development of Crohn's disease[4], which supported subsequent clinical trials for drugs targeting the IL-12/IL-23 pathway[5].

https://www.nature.com/articles/s43586-021-00056-9

»» 1.6 GWAS Catalog检索

- 全基因组关联分析(GWAS)的应用

https://www.ebi.ac.uk/gwas/downloads/summary-statistics



GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

obesity

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

feedback

GWAS / Search / obesity

Refine search results

- Publications 64
- Traits 75

Catalog stats

- *Last data release on 2022-10-08*
- *6041 publications*
- *236366 SNPs*
- *427870 associations*
- *Genome assembly GRCh38.p13*
- *dbSNP Build 154*
- *Ensembl Build 107*

Search results for *obesity*

obesity EFO_0001073

A status with BODY WEIGHT that is grossly above the acceptable or desirable weight, usually due to accumulation of excess FATS in the body. The standards may vary with age, sex, genetic or cultural ba... Show more >

Associations 289 Studies 44

metabolically healthy obesity EFO_0009382

Long-standing obesity without metbolic abnormalities or obesity-related comorbidities such as type 2 diabetes or heart disease

Associations 14 Studies 1

02

GWAS分析环境

1.1 GWAS分析特点

GWAS分析的难点

- 样本量大
- 群体结构复杂
- 大多数表型是受多位点影响的数量性状，导致计算模型复杂度增加
- 表型量化难度大
- ……

本次GWAS实验特点

- 样本量小
- 样本组成简单，无家系样本，无需考虑亲缘关系
- 关注表型明确，分析所涉表型多受单一位点或数量有限的位点影响

2.2 实验分析思路与方法

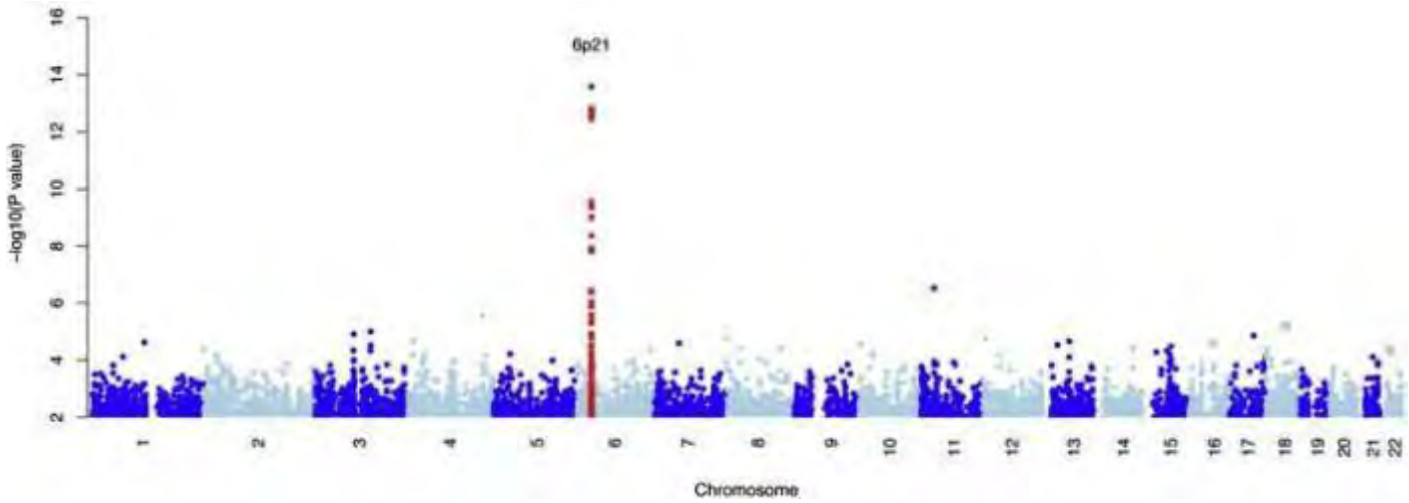
基因型数据准备　　表型数据准备

假设检验（统计模型选择）

p值计算（linux环境、 R语言与工具包）

曼哈顿图绘制

实验结果解读



2.3 服务器R语言环境

R语言代码框架与功能模块解析

登录linux环境，进入R环境　　R

↓

安装R包　　镜像地址选择

↓

读取基因型、表型数据集　　备用数据　or 备用数据+现场整合数据

↓

参数赋值与数据格式整理

↓

统计检验与p值计算Manhattan图绘制

2.3 所需数据文件

基因型数据文件

- 个体基因型详表采用了长窄型表示：<样本编码，SNP名称，基因型>。

SampleCode	SNP_Marker	IP_Genotype
342_57	rs1490413	AA
342_57	rs5745448	TC
342_57	rs3737576	TC
342_57	rs1698647	CT
342_57	rs1343469	AG
342_57	rs11239930	GA
342_57	rs7554936	TT
342_57	rs3829868	CC
342_57	rs2814778	TT
342_57	rs560681	AA
342_57	rs10801520	CT
342_57	rs1106201	CT
342_57	rs2013162	AA
342_57	rs2292564	AA
342_57	rs1294331	CC
342_57	rs10495407	GG

2.3 所需数据文件

表型数据文件

●表型数据也类似：<样本编码，表型名称，表型值>。

245_58	耳垢类型	干
398_59	耳垢类型	干
367_60	耳垢类型	干
345_61	耳垢类型	湿
336_62	耳垢类型	干
235_63	耳垢类型	干
390_64	耳垢类型	干
205_65	耳垢类型	干
296_66	耳垢类型	干
271_67	耳垢类型	干
358_68	耳垢类型	干
342_57	肌肉性状	爆发
245_58	肌肉性状	爆发
398_59	肌肉性状	爆发

03

GWAS分析操作

3.1 所需数据文件

IP: http://192.168.79.142:8787

Username: testp

Password: fudan2024

在Linux服务器上创建工作目录

mkdir yourName_gwas

设置Linux服务器上设置工作目录，请把ecoli改成自己的个人文件夹名

setwd(" ~/yourName_gwas ")

自己配置R语言环境的话可以安装下面的包:

install.packages("openxlsx")

install.packages("qqman")

install.packages("tidyr")

3.1 所需数据文件


```
# 导入读入xlsx的包openxlsx
library(openxlsx)
library(qqman)


# 导入基因型文件
rawgenotype<- read.xlsx("~/data/20211118.R2.SNP.xlsx")
# 导入表型文件
rawphenotype<- read.xlsx("~/data/20211118.phenotype.xlsx", colNames=FALSE) #无标题


# 通过数据框操作， 提取文件的前3列有用信息， 去掉后面的列
mg<- rawgenotype[,c(1,2,3)]
mp<- rawphenotype[,c(1,2,3)]
```

3.2 整合转换数据格式


```
names(mp)<-names(mg) #统一列名称，才能rbind。
mgp<-rbind(mg,mp)


# 第一列是样本代码，获取样本数目
nsample<-length(levels(factor(mg[,1])))


# 基因型数据第二列是SNP位点名称，获取数目
nloci <- length(levels(factor(mg[,2])))


# 表型数据第二列是性状特征名称，获取数目
ntrait<- length(levels(factor(mp[,2])))
```

3.2 整合转换数据格式


```
# 导入tidyr包进行数据格式的转换
library(tidyr)

# 第一列为样本名称，转换成行名称
# 第二列为基因型或表型名称，转换为列名称
# 第三列为基因型或表型取值，对应为矩阵项取值
md <- pivot_wider(mgp, names_from=names(mgp)[2], values_from=names(mgp)[3])

# 先单独生成一个公式试试，波浪号~前面为空，表示按出现项数来计数
mc <- xtabs(~ID.rs1490413 + 耳垢类型, md)
```

3.3 对离散性状生成频数列联表和进行关联显著性检验 30

以下开始使用as.formula函数和paste函数来规律地生成公式表达式

第一列是Sample Code，第二列到nloci+1列是SNP，后面直到nloci+ntrait+1列是表型性状

```
dname<-names(md)
```

前面已经保存了nsample、nloci和ntrait三个值，方便我们来用下标进而可循环批量生成公式。

公式为：

```
xf <- paste(" ~ ", dname[2], " + ", dname[nloci+2])
```

as.formula在此很关键，#这样我们就得到了Chi-squared test或fisher's exact test需要的列联表mc

```
mc <- xtabs(as.formula(xf), md)
```

```
# 开始进行Chi-squared test
chisq.test(mc)

# 这样会提示“Chi-squared近似算法有可能不准”。我们换成:
chisq.test(mc, simulate.p.value = TRUE, B = 10000)

# 参数simulate.p.value是指定用Monte Carlo模拟来计算p-value，参数B设置模拟次数。
# 因为样本数目较少，即列联表中频数较小，实际上应该直接用fisher's exact test。
fisher.test(mc)
```

3.4 计算所有位点-性状关联并画曼哈顿图


```
# 计算所有位点-性状关联并画曼哈顿图
# 接下来我们写一个循环来完成所有SNP与所有性状之间的列联表以及显著性检验。
  # 首先造一个矩阵来存储显著性q-values值， 并加上行列名称， 以便筛选后知道是哪个SNP关联上哪个性状。
assocTraitQ  <- data.frame(matrix(rep(0, ntrait*nloci), ncol =ntrait))
names(assocTraitQ) <- dname[(1+nloci+1):(1+nloci+ntrait)]
row.names(assocTraitQ)<-dname[(1+1):(1+nloci)]

#  使用pdf命令开一个pdf画布， 因为实时显示只能有一张图， 而我们有多个性状关联结果， 所以输出到pdf文件中
pdf(file="E5GWAS.pdf", family="GB1")
```

3.4 计算所有位点-性状关联并画曼哈顿图

每个性状画一张曼哈顿图。

```
for(k in 1:ntrait){
        pvalues=array(0)
        for(i in 1:nloci){
                xf <- paste(" ~ ", dname[1+i], " + ", dname[1+nloci+k])
                mc <- xtabs(as.formula(xf), md)
                if(dim(mc)[1]==1){
                        pvalues[i] <- 1
                }else{
                        pvalues[i] <- fisher.test(mc)$p.value
                }
                assocTraitQ[i,k] <- pvalues[i]
        }
```

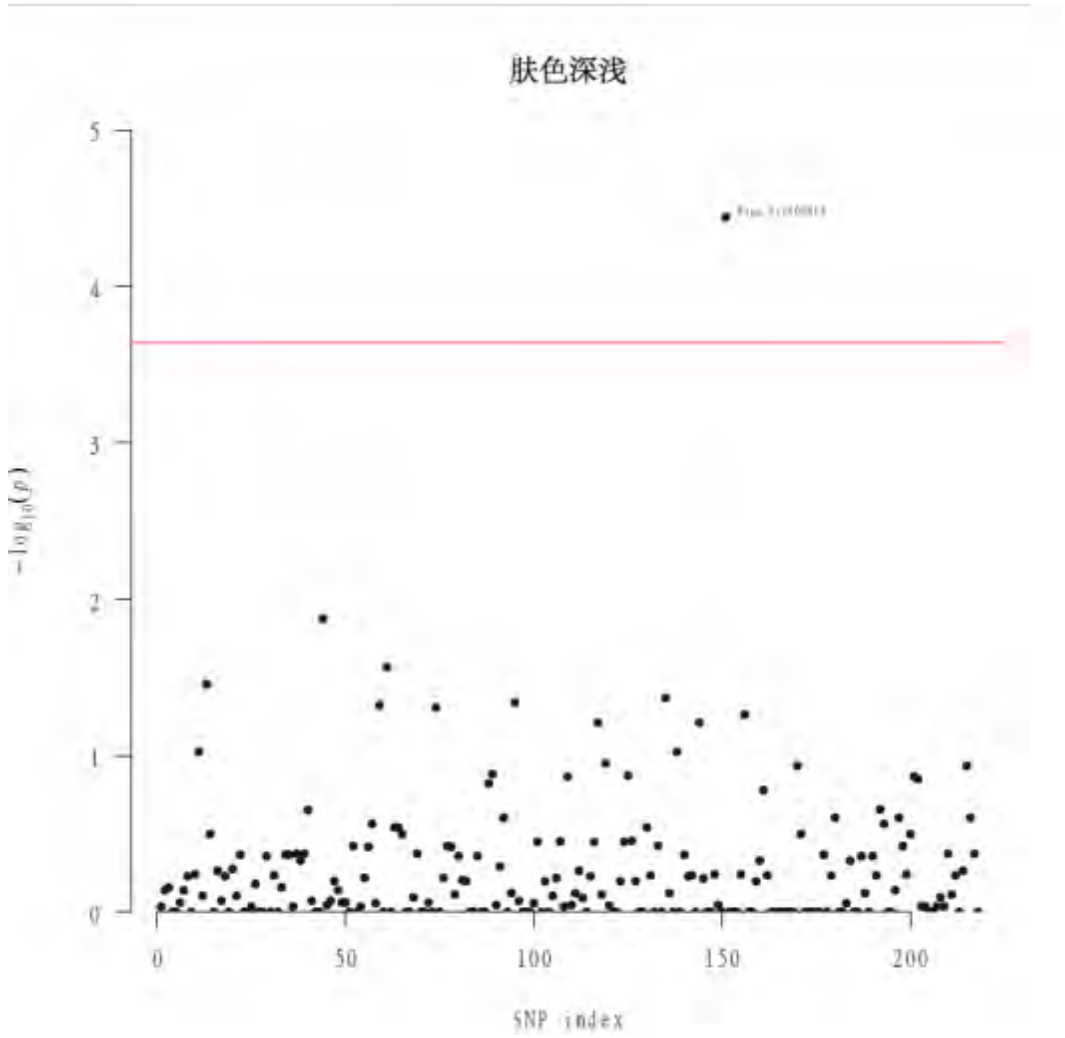
3.4 计算所有位点-性状关联并画曼哈顿图


```
        nSNP <- length(pvalues)
        gwasMatrix <- data.frame(dname[(1+1):(1+nloci)], rep(1,times=nSNP), 1:nSNP, pvalues)
        names(gwasMatrix) <- c('SNP', 'CHR', 'BP', 'P')
        manhattan(gwasMatrix,
                annotatePval = -log10(0.05/nSNP), #标注p值最小的SNP位点
                suggestiveline = FALSE,
                genomewideline = -log10(0.05/nSNP), #画出邦费罗尼校正显著性的阈值线
                ylim = c(0,5),
                xlab = 'SNP index',
                main = dname[1+nloci+k]
                )
}
dev.off() #保存并关闭pdf文件
```

3.4 计算所有位点-性状关联并画曼哈顿图

我们筛选出q-value小于0.05的SNP

assocTraitQ[apply(assocTraitQ,1,min) < 0.05/nSNP,]



教学内容小结

01 全基因组关联分析(GWAS)：概念、原理、用途等

02 GWAS分析环境：分析流程

03 GWAS分析操作：R语言分析和作图

谢 谢！