实验 4.1 表型测量

（1 学时）

一、实验背景

表型，是指生物体的各种特征。以人类个体为例，我们眼皮的单双、先天肤色的深浅，体检项目当中心、肝、脾、肺、肾的各项检测指标，血常规检测的各项生理、生化、代谢指标，以及体能测试检测的速度、力量等功能指标全部都是个体表型。大量研究表明，表型是由基因、表观遗传学、共生微生物、饮食和环境暴露等复杂因素相互作用产生的一系列可测量特征，包括个体和群体的物理、化学及生物学特征。

表型测量，目的是获取各种特征的量化数据。以疾病研究为例，当我们研究肥胖症表型时，我们会测量体重或 BMI；研究糖尿病表型时，则会关注患者血糖值的高低。

二、教学目标

1、人个体特征表型采集量表设计

2、表型数据采集

三、实验原理

第一部分 表型测量有哪些难点？

表型测量难在表型的细分和表型指标量化。我们以糖尿病表型测量为例加以说明。说起糖尿病表型，大家第一印象可能是高血糖，但高血糖其实只是糖尿病众多临床症状之一。仅仅测量血糖，只能对已经患病的人进行确诊，但没办法找出潜在的或未来可能发展为糖尿病患者的人群。因此，对糖尿病进行更加全面的表型研究，就要求我们充分了解引发糖尿病的复杂成因，能对除了糖类之外的脂类、蛋白类等代谢指标也进行细分监测，对人体内数百甚至数千个不同的分子成分进行测定和分析，从而将临床上的高血糖表型划分为不同的亚型，帮助我们更早识别不同类型的发病前人群，并针对不同类型的人群使用不同的策略和药物进行干预和治疗。

第二部分 表型组学研究有哪些深远意义？

表型组学，就是对表型进行测量和研究的学科。相比于基因组，表型组的研究对象是更加多维和跨尺度的生物指征。通过对人类表型组进行精密的跨尺度测

量，进而积累起海量的生物信息大数据，再对这些数据进行分析建模，就能够发现全新的基因-表型-环境以及宏观-微观表型之间的关联机制，最终支撑起对于疾病和健康更加精准的诊断、干预、调控和治疗。

中国科学院院士、“国际人类表型组计划（一期）”项目首席科学家、复旦大学金力教授曾经这样描述表型组学研究的重要意义：“打个比方，如果说人类基因组计划让科学家手里有了探索生命奥秘的‘指南针’，那么人类表型组计划就将打造未来遨游生命科学世界的‘全球卫星定位系统’。”

在国际生命科学界，表型组已经被公认为继基因组研究之后，生命科学研究的“最先一公里”——未来对大多数生物医学研究的重大突破、实现疾病的精准治疗与下一代新药研发，都将依赖表型组学领域的原始创新。

第三部分 大型国际表型组学研究计划

2003 年，美国科学家弗雷梅尔（Nelson Freimer）和萨巴蒂（Chiara Sabatti）首次在《自然·遗传学》杂志上发文提出“人类表型组计划”，并倡议建立国际人类表型数据库。

但倡议提出后并没有被完全付诸行动，英美等国的人类表型组科研活动在规模、深度、组织性等方面仍长期处于探索阶段。

2015 年，以“国际人类表型组研究”为主题的香山科学会议第 525 次学术讨论会在北京香山举行。英国著名科学家、“代谢组学之父”尼科尔森（Jeremy Nicholson）院士、国际人类基因变异组学会主席科顿（Richard Cotton）教授和中国科学家金力院士、赵国屏院士、王辰院士等近 50 名专家学者在会上一致建议发起“国际人类表型组计划”，全面开展人类表型组研究。

此后，中国开始率先布局人类表型组大科学计划。2015-2016 年，科技部国家科技基础性工作专项和上海市科委基础重大专项相继立项人类表型组相关研究项目。“国际人类表型组”也被正式列入上海建设全球科创中心需布局的重大科学基础工程。2018 年 3 月，在中国上海，“国际人类表型组计划（一期）”正式启动，由金力院士担任首席专家，旨在首次建立国际领先的人类表型组学研究平台。

与此同时，美英等发达国家近年对表型组学的关注也明显加强。2018 年，美国公开招募作为人类表型组研究基础的“我们所有人（All of Us）”百万人队列项目，计划跟踪数十年。2019 年 6 月，英国也启动了国家表型组资源项目。

2019 年 7 月，总投入 2 亿元、建筑面积约 4000 平方米的世界首个跨尺度、多维度人类表型组精密测量集成平台在上海张江正式建成。目前，这个平台已经可以对人体从微观到宏观全尺度的近 15 类表型进行高精度、高灵敏、高通量测量，具备每年对 1000 人次进行在体测量、每个个体 2 万个表型指标的测量能力。同时，该平台的离体检测模块还具备每年检测数 10 万人级人体生物样本、每个样本 1 万余个表型指标的检测能力。

第四部分 表型测量要素

以东亚人先天肤色为例，测量要素包括：

1. 测量位置（如大臂内侧、下巴下缘、大腿内侧等）

2. 测量工具（CR-400 采集或相机拍摄，4 个角及中心共 5 个点 rgb 值取平均值）

3. 测量次数（每取样点 3 次取值）

4. 数值处理方法（每取样点 3 次取平均值）

先天肤色深浅的测量方法，可以大臂内侧作为测量点，每人每测量点测量 3 次取平均值，进行记录。

四、实验步骤

1、表型采集量表设计（以东亚人先天肤色深浅为例）

Num	Pos	Test			Ave	Tools

2、人个体特征表型采集（以先天肤色深浅为例，完成以上采集表）

实验 4.2 全基因组关联分析（GWAS）

（4 学时）

一、实验背景

全基因组关联分析（Genome-wide association study，简称 GWAS），是对多个个体在全基因组范围内的遗传变异多态性进行检测，获得基因型，进而将基因型与可观测的性状，即表型，进行群体水平的统计学分析，根据统计量或显著性 p 值筛选出与该性状显著相关的遗传变异，挖掘与性状变异相关的基因。本章节将基于实验 2.3 获取的 SNP 信息完成 GWAS 分析。

二、教学目标

理解 GWAS 分析的基本思路和操作方法；

掌握 GWAS 分析模型和相关数据库。

三、实验原理

第一部分 GWAS 的定义：

GWAS 是指在人类全基因组范围内找出存在的序列变异，即单核苷酸多态性（Single Nucleotide Polymorphism, SNP），从中筛选出与疾病相关的 SNPs。GWAS 为人们打开了一扇通往研究复杂疾病的大门，将在患者全基因组范围内检测出的 SNP 位点与对照组进行比较，找出所有的变异等位基因频率，从而避免了像候选基因策略一样需要预先假设致病基因。同时，GWAS 研究让我们找到了许多从前未曾发现的基因以及染色体区域，为复杂疾病的发病机制提供了更多的线索。GWAS 研究方法的要点是：选择足够多的样本，一次性地在所有研究对象中对目标 SNP 进行基因分型，然后分析每个 SNP 与目标性状的关联，统计分析关联强度。(补充阅读链接：https://www.sciencedirect.com/science/article/pii/S0092867418310328)

第二部分 GWAS常用模型：

（1）广义线性模型（GLM）

在统计学上，广义线性模型 (generalized linear model，缩写作 GLM) 是一种应用灵活的线性回归模型。该模型允许因变量的偏差分布有除正态分布之外的其它分布.

GLM 的手动计算 GWAS 分析的主要步骤：

（i）将 SNP 的分型转化为 0-1-2（0 位次等位基因），数字格式（x 变量）

（ii）将性状观测值作为 y 变量（GLM 一般分析连续性状）

（iii）对 y~x 做回归分析，计算 x 的回归系数（Effect）和显著性（P-value）

(2) 一般线性模型（MLM）

MLM 模型中，将每个 SNP 作为固定因子进行回归分析，将亲缘关系矩阵（kinship 或者 A 矩阵）作为随机因子，进行 SNP 的显著性检验，P 值就是 GWAS 分析的 p-value，effect 就是 SNP 的 effect 值

(3) Logistic 回归模型

（i）将 SNP 的分型转化为 0-1-2（0 位次等位基因），数字格式（x 变量）

（ii）将性状观测值作为 y 变量（Logistic 一般分析二分类性状）

（iii）对 y~x 做 Logistic 回归分析，计算 x 的回归系数（Effect）和显著性（P-value）

第三部分 GWAS 常见数据库

(1) GWAS ATLAS 数据库

该数据库收录了来自 4756 个人类不同表型的 GWAS 结果，提供了 risk loci, LDSC, MAGMA 基因集关联分析，多种表型间的遗传相关性分析等结果，数据

库网址为 https://atlas.ctglab.nl/

(2)ExAC 数据库

EXAC（the Exome Aggregation Consortium），中文为外显子组整合数据库，该数据库旨在汇总和协调各种大规模测序项目的外显子组测序数据，整合了多个研究项目的外显子集合协作组（ExAC）分析了来自不同祖先的共 60706 位个人的高质量外显子测序数据,通过深度分析制作的人类遗传变异数据库 ExAC 并制定了每个序列变异的致病性的精确度量标准。研究鉴定了根据选择压力区别的突变类型；鉴定了 3230 个基因截短突变，其中 72%的基因没有与已知的人类疾病表型建立关系。网址为 http://exac.broadinstitute.org

(3)百万中国人基因数据库（CMDB, Chinese Millionome Database）

华大基因发布了百万基因组数据库目前所包含样本总数已超过两百五十万，样本来源覆盖我国所有省份，具有非常好的代表性。数据库存放于国家基因库(CNGB, China National GeneBank)，此次发布的为第一期——基于十四万人的基因多态性位点 (SNP) 和频率信息，共包括充分代表中国人群的八百五十余万高质量 SNP，其中 24% 是从未在已公开发布的中国人群基因数据库中被发现过的新 SNP。与千人基因组项目比较，CMDB 数据库解决了人群数目不足而导至的基因数据库精确度低以及抽样误差等问题，访问网址为 https://db.cngb.org/cmdb。

四、软件安装与数据准备

确认基因型与表型数据格式

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	SNP_Genotype
235_63	ID.rs733164	GA
235_63	ID.rs5749426	CC
235_63	ID.rs987640	TA
235_63	Popu.rs2024566	AA
235_63	ID.rs2040411	GG
235_63	ID.rs1028528	AG
390_64	ID.rs1490413	GA
390_64	ID.rs5745448	TC
390_64	Popu.rs3737576	TT

表型数据也类似：<样本编码，表型名称，表型值>。

```
342_57  gender  male
205_65  gender  male
235_63  gender  male
245_58  gender  male
271_67  gender  female
296_66  gender  female
336_62  gender  female
345_61  gender  female
358_68  gender  female
367_60  gender  female
390_64  gender  female
398_59  gender  female
342_57  BMI     21.1
205_65  BMI     18.7
235_63  BMI     29.3
245_58  BMI     26.4
271_67  BMI     17.9
296_66  BMI     23.6
336_62  BMI     28.0
345_61  BMI     20.9
358_68  BMI     22.8
367_60  BMI     23.4
390_64  BMI     24.5
398_59  BMI     27.2
```

五、实验步骤

1、读取基因型与表型数据

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 ")

我们在 R 里读取 xlsx 文件需要特定的 R 包，搜索一下后选择 openxlsx。

library(openxlsx)

library(qqman)

rawgenotype<- read.xlsx("~/data/20211118.R2.SNP.xlsx")

rawphenotype<- read.xlsx("~/data/20211118.phenotype.xlsx", colNames=FALSE) # 无标题

mg<- rawgenotype[,c(1,2,3)] #实际文件包含了其它列，这里去掉。

mp<- rawphenotype[,c(1,2,3)]

names(mp)<-names(mg) #统一列名称，才能 rbind。

mgp<-rbind(mg,mp)

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

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

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

2、整合转换数据格式

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])

接下来用 xtabs 函数可得到一个 SNP 与一个离散表型特征之间的列联表。

mc<- xtabs(~ ID.rs1490413 +耳垢类型, md)

#波浪号~前面为空，表示按出现项数来计数。

不过上面这样表示，就需要在代码中包含可能冗长而奇怪的列名称（如“ID.rs1490413”），这实在太反人类了！

我们需要 as.formula 函数和 paste 函数来规律地生成公式表达式。

dname<-names(md) #第一列是 SampleCode，第二列到 nloci+1 列是 SNP，后面直到 nloci+ntrait+1 列是表型性状。

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])

mc<- xtabs(as.formula(xf), md) # 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。

然后调用检验函数就是了~

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、离散性状模型与其它检验方法的选择

不过上面这样检验默认将基因型如 AA、GA 和 GG 视为三个无关的类别，实际上对应于 plink 中 Alternate / full model association tests 中的 Genotypic (2 df) test。

实现 basic allelic test

4、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)]
```

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。

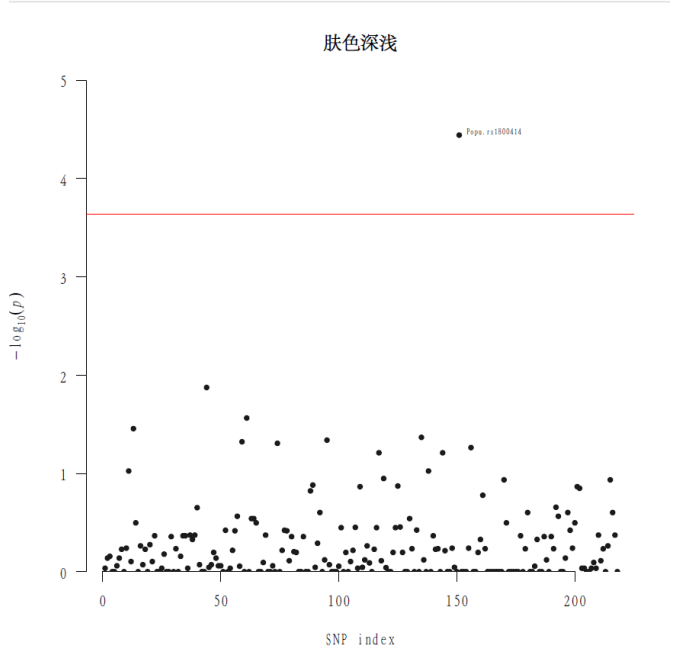
```
pdf(file="E5GWAS.pdf", family="GB1")
```

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
      }
  }
  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
我们筛选出 q-value 小于 0.05 的 SNP
assocTraitQ[apply(assocTraitQ,1,min)< 0.05/nSNP,]
```

六、预期实验结果



这样我们就筛选出与东亚人先天肤色深浅相关的 SNP 位点：rs1800414