**Data Analysis**

**甲基化分析**

**Getting Started**

source("http://www.bioconductor.org/biocLite.R")

biocLite("minfi") ###安装minfi程序包###

biocLite("minfiData") ###安装minfi数据###

library(minfi) ###载入minfi程序包###

library(minfiData) ###载入minfi数据###

getwd() 查看工作目录

setwd(“ ”)

**Reading Data**

将IDAT文件读入为RGChannelSet，读取芯片检测的两种光信号强度（IDAT代表两种光通道）

IDAT文件：芯片原始数据，记录红、绿两种光通道

RGChannelSet 文件：原始信号强度

getwd()

setwd(“ ”)

baseDir <- system.file("extdata", package = "minfiData") ###获取数据路径###

targets <- read.450k.sheet(baseDir) ##读取csv文件（sample sheet）##

RGset <- read.450k.exp(targets = targets) ##读取IDAT文件## （表格文件里包含了IDAT路径， 。。。。。。 所以两个语句是一样的）

## RGset2<-read.450k.exp(file.path(baseDir,”5723646052”)) ## input：sample sheet / IDAT文件路径##

## output：RGChannelSet 文件##

pd <- pData(RGset) ## 获取RGset中的表型数据

## 其内容与target/ sample sheet相似##

**Quality Control （检测数据是否达标）**

检测芯片数据的质量

qcReport(RGset, sampNames = pd$Sample\_Name, sampGroups = pd$Sample\_Group, pdf = "qcReport.pdf")

Pdf文件，代表会有一个pdf生成保存到之前的路径里去

每个函数在Pdf上都有写

## densityPlot(RGset, sampGroups = pd$Sample\_Group, main = "Beta", xlab = "Beta")

## densityBeanPlot(RGset, sampGroups = pd$Sample\_Group,sampNames = pd$Sample\_Name)

**Preprocessing**

数据预处理:数据均一化，减去背景信号，

甲基化信号转化：RGChannelSet→MethylSet

MSet.raw <- preprocessRaw(RGset) ##只做甲基化信号转化##(只把甲基化信号转化为光信号）

MSet.norm <- preprocessIllumina(RGset, bg.correct = TRUE,normalize = "controls", reference = 2)

##信号转化；减背景信号；均一化##

## preprocessSWAN :数据均一化，减少不同的探针设计带来的测量误差（对探针标准化）

## preprocessQuantile：分层均一化，减少样品内部变异 （比较适合样品内部的标准化）

**Multi-dimensional scaling(MDS) plot**

类似于主成分分析（PCA）:找到数据主要变化方向

mdsPlot(MSet.norm, numPositions = 1000, sampGroups = pd$Sample\_Group, sampNames = pd$Sample\_Name)

Mds主成份分析，可以确定样品的差异的主要原因

**Finding Differentially Methylated positions (DMPs)**

寻找不同表型的样本间的甲基化差异位点

dmpFinder( )

dmpFinder可以对两种不同的表形分析

离散型变量：F检验

连续型变量：最小二乘法构建回归模型

mset <- MSet.norm[1:20000,] ##截取部分数据，加快运行##

**Categorical phenotypes**

算CpG位点的甲基化程度，有两种，一种是M值，一种是β值

M <- getM(mset, type = "beta", betaThreshold = 0.001) ##计算M值：M=（）##

beta<-getBeta(mset) ## Beta值：β= ##

dmp <- dmpFinder(beta, pheno=pd$Sample\_Group, type="categorical")

##基于各自的Beta值，对两组表型样品每个CpG位点

进行F检验##

##计算代表差异显著性的p值，找到甲基化差异位点##

class(dmp)

dim(dmp)

head(dmp) ## 查看前6个结果 ##

write.csv (dmp,file= "dmp.csv ") ## 保存为csv文件##

**plot**

cpgs <- rownames(dmp)[1:4] ##选取所有差异位点中的前4个CpG##

————————————

par(mfrow=c(2,2)) ##画图参数：图像呈2x2分布##

plotCpg(mset, cpg=cpgs, pheno=pd$Sample\_Group) ##将计算结果可视化##

## input：RGChannelSet / MethylSet,

1.计算Beta /M值，默认为Beta值

2.画出不同表型之间的甲基化差异位点

—————————————

pdf(file=”dmp.pdf”) ##保存为pdf文件##

dev.off()

##用M值找差异位点##

## dmp2 <- dmpFinder(M, pheno=pd$Sample\_Group, type="categorical")

## cpgs2 <- rownames(dmp2)[1:4]

## head(M[rownames(dmp2),]) ##确定y轴范围##

##par(mfrow=c(2,2))

## plotCpg(mset, cpg=cpgs2, pheno=pd$Sample\_Group,measure = "M",ylim =c(-10,3))

## head(getM(mset)[rownames(dmp2),]) ## M值中存在Inf，所以选用Beta值##

**Continuous phenotypes**

不同: type=”continuous”;算法：构建线性回归模型

dmp <- dmpFinder(mset, pheno=continuousPheno, type="continuous")