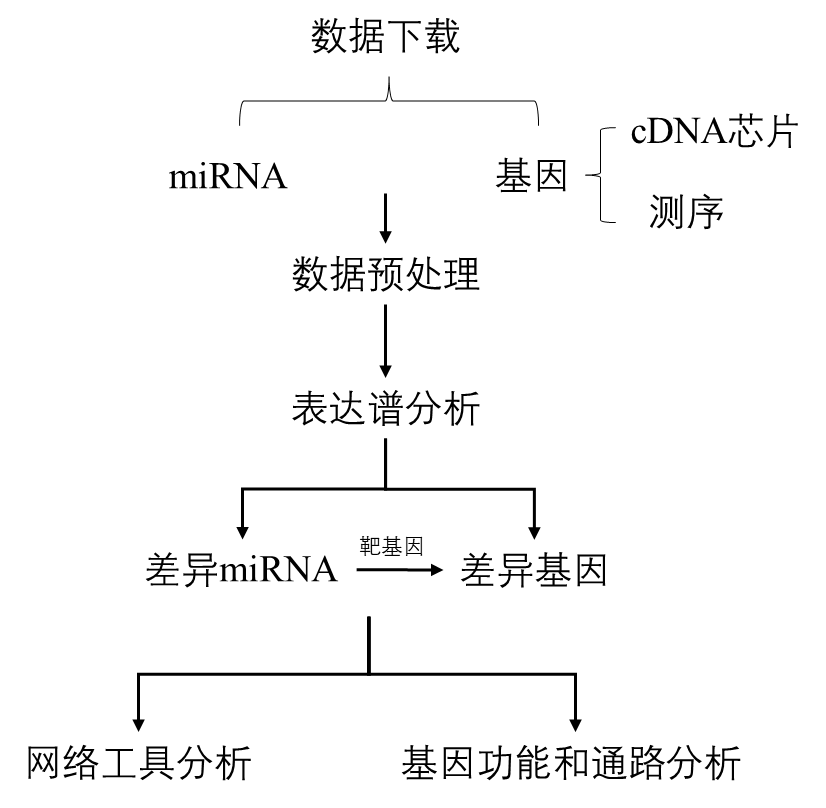
**转录组分析学习资料**

**分析流程**

****

1. 数据下载
   1. 基因表达数据
   2. miRNA数据
   3. 常用数据库整理
2. 数据预处理
   1. 样本合并

Json文件下提取file\_name和对应的sample\_id

获取文件夹下的文件名，

提取文件中的miRNA和read\_counts列，列名为文件名对应的sample\_id

合并read\_counts列，其中的列名为文件

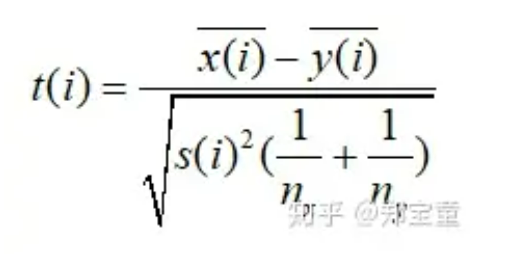
* 1. 数据筛选（去除空值、取平均值等）
  2. 标准化
  3. 样本分析

1. 差异表达分析
   1. FC

FC(Fold Change)算法是最早用于识别**两种不同的实验条件下基因表达水平存在差异**的算法，其算法的原理是**计算基因在两类样本中平均表达水平的倍数值**，若该值达到预先设定的阈值(一般设置为2，在以2为底的对数表达比中为大于1或小于-1)，则判定基因为差异表达（DE，different expression）基因，计算公式如下，其中，mean(X(i))与mean（Y(i)）代表基因i在两类样本中的平均表达值：

* 1. t检验(t-test)

常用来识别两类样本中DE基因的算法。其主要原理为：对每一个基因计算一个t统计量来衡量两类样本中基因表达的差异，然后根据t分布计算显著性p值来衡量这种差异的显著性。计算公式如下，其中，分子代表基因i在两类样本中的平均表达差值，分母代表基因i在所有样本中的标准误：



由于t检验要求数据呈现正太分布，所以公式中基因的表达值为测量值经过标准化后的值，反应的是两类样本间基因表达的倍数变化，也存在FC方法同样的偏向性。此外，对基础表达量低的基因来说，一个微小变异程度（标准误）可能导致一个大的绝对t统计值,从而被识别为DE基因，即使在两类条件下这个基因的平均表达水平的差异很小。低表达的基因比高表达的基因更容易产生大的t统计量。已有研究指出，数据的信噪比会随着基因表达量的增高而降低，这就意味着，低表达的基因更容易受到噪声的影响而产生误差。因此，t检验同样倾向于识别表达水平低的基因作为DE基因。

* 1. limma分析
     1. 原理
     2. 代码思路
     3. 上调和下调分析
     4. 火山图绘制
  2. DESeq分析
     1. 原理
     2. 代码思路
     3. 上调和下调分析
     4. 火山图绘制
  3. Edge分析
     1. 原理
     2. 代码思路

1. 生存分析
   1. 生存分析介绍
   2. 生存分析思路
2. Cox分析
3. 生物学意义分析

截至2023年10月30日，与TCGA数据库下载miRNA在胃癌样本中的表达，数据量为446个tumor和45个normal

1.2.数据标准化

首先使用R语言对这491个样本进行整合，得到miRNA矩阵和样本信息数据框

然后使用DESeq2进行miRNA矩阵标准化和差异分析

1. DEseq2的安装：

if (!requireNamespace("BiocManager", quietly = TRUE))

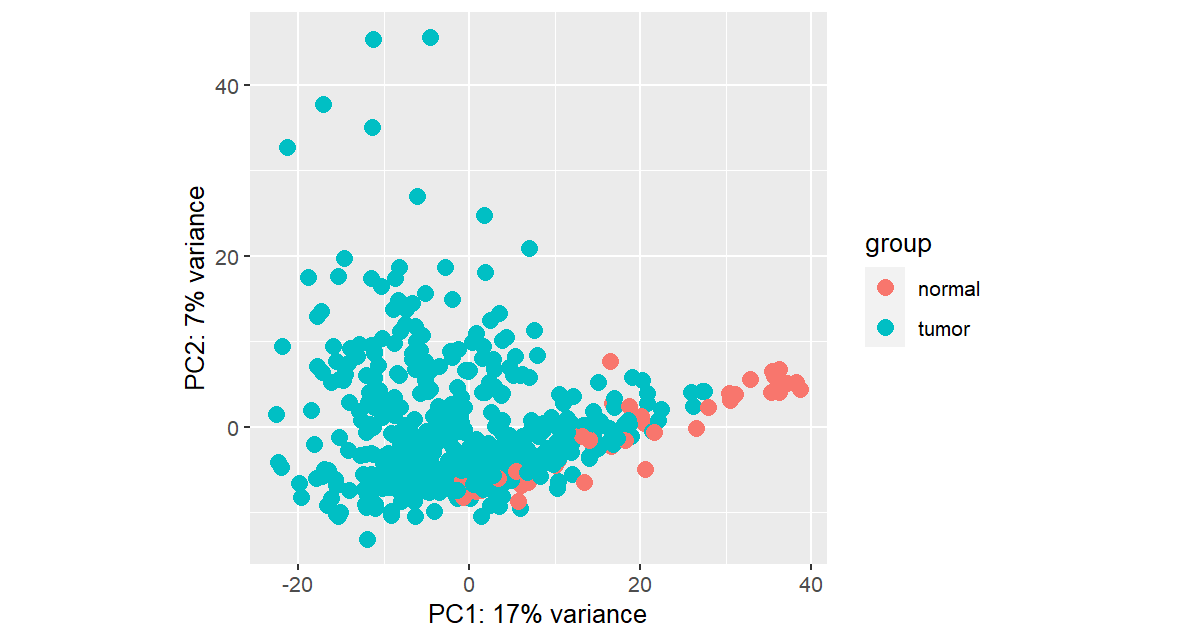
install.packages("BiocManager")

BiocManager::install("DESeq2")

library(DESeq2)

1. 方差平稳标准化（dds为DEseq2构建的包含表达、tumor、normal信息的对象）

vsd <- varianceStabilizingTransformation(dds)

对方差平稳标准化后的数据绘制PCA

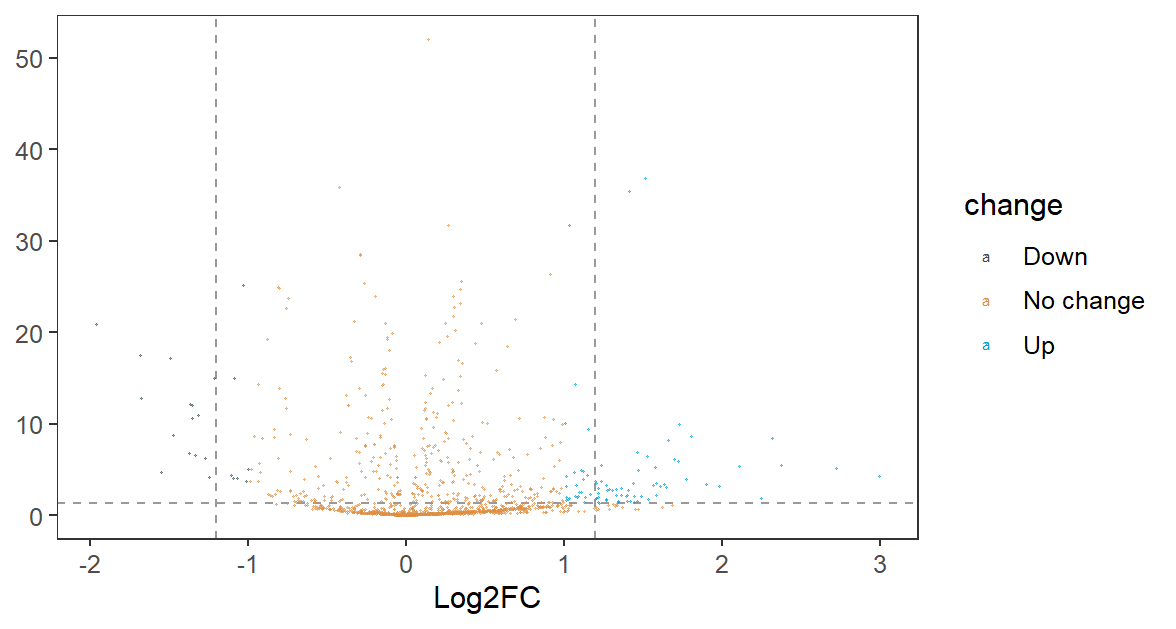
可以看出，基于miRNA为特征并不能将肿瘤组织和正常组织完美分开

* 1. R语言差异分析

去除在所有样本中都不表达的miRNA，然后DEseq标准化

1.3.1 FC+T检验

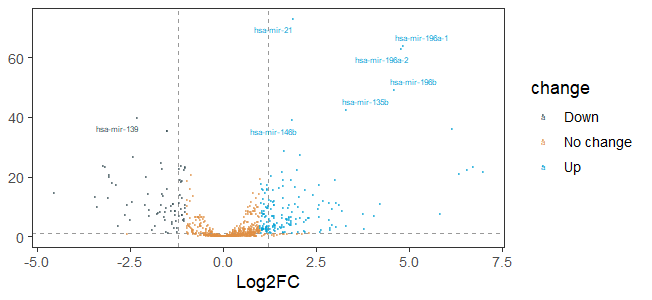
1. 计算差异倍数：差异倍数（fold change）,也叫作radio，简单来说就是基因在一组样品中的表达值的均值除以其在另一组样品中的表达值的均值，用FC表示。如果规定log2FC>1为上调，那么FC<-1为下调
2. T检验，计算p值后经BH法校正
3. 最后得到22个肿瘤样本下调，95个肿瘤样本上调的显著miRNA（FDR）



1.3.2 DESeq筛选差异miRNA

1. 使用DESeq筛选

2. 同样基于log2FC>1-上调，log2FC<-1 下调，得到74个下调，166个上调的显著miRNA（基于p）



补充：DESeq中有三种标准化：

·count标准化

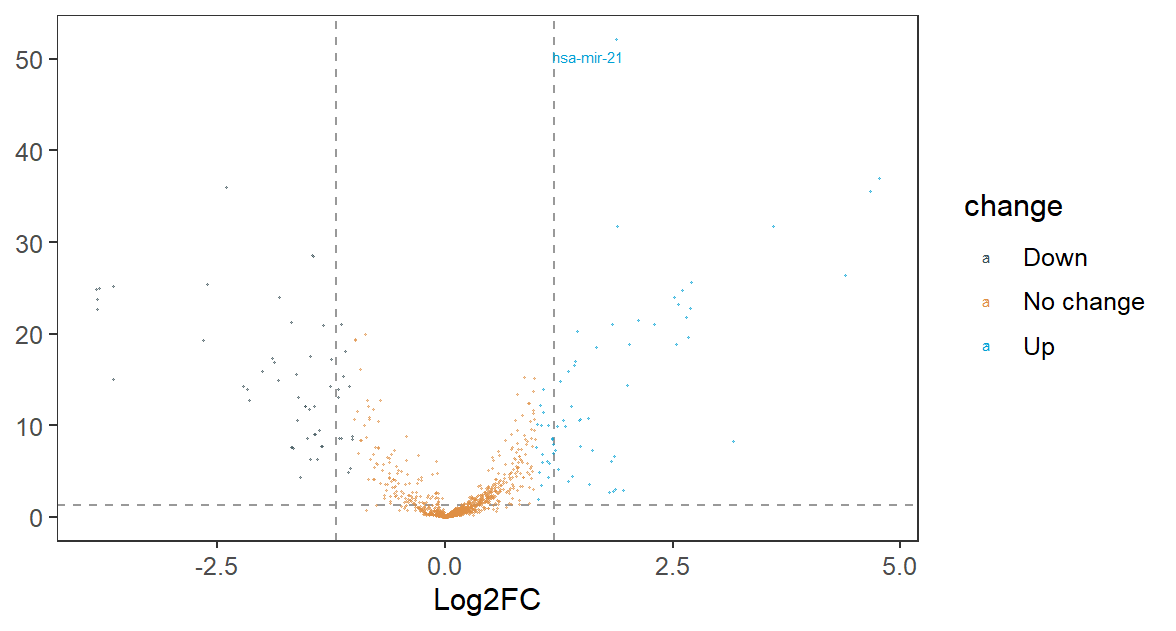
·rlog标准化

·vst标准化

1.3.3 Limma筛选差异基因

1.使用lmfit后进行 整理得到55个下调和65个上调

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Univariate Cox regression | | | | multivariate Cox regression | | | | |
| Gene | HR | L95CI | H95CI | Pvalue | HR | L95CI | H95CI | Pvalue | coef |
| hsa-mir-1292 | 0.8391059 | 0.7132911 | 0.9871127 | 0.034 |  |  |  |  |  |
| hsa-mir-17 | 0.8266464 | 0.6862212 | 0.9958075 | 0.045 |  |  |  |  |  |
| hsa-mir-200b | 0.857469 | 0.7644077 | 0.9618598 | 0.009 |  |  |  |  |  |
| hsa-mir-200c | 0.8571061 | 0.7579141 | 0.9692798 | 0.014 |  |  |  |  |  |
| hsa-mir-2115 | 0.8334078 | 0.7093192 | 0.9792045 | 0.027 |  |  |  |  |  |
| hsa-mir-217 | 1.128803 | 1.017925 | 1.2517585 | 0.022 |  |  |  |  |  |
| hsa-mir-371a | 1.1202289 | 1.0116316 | 1.240484 | 0.029 |  |  |  |  |  |
| hsa-mir-372 | 1.0775952 | 1.0042312 | 1.1563189 | 0.038 |  |  |  |  |  |
| hsa-mir-373 | 1.1034846 | 1.0038362 | 1.2130248 | 0.041 |  |  |  |  |  |
| hsa-mir-4664 | 0.8620057 | 0.7533942 | 0.986275 | 0.031 |  |  |  |  |  |
| hsa-mir-548v | 1.1642311 | 1.0279338 | 1.3186005 | 0.017 |  |  |  |  |  |
| hsa-mir-549a | 0.8418992 | 0.7415246 | 0.9558608 | 0.008 | 0.849055 | 0.7299931 | 0.9875359 | 0.033781766 | 0.1636313 |
| hsa-mir-675 | 1.0869161 | 1.0210574 | 1.1570227 | 0.009 |  |  |  |  |  |
| hsa-mir-7-2 | 0.8121578 | 0.7181114 | 0.9185208 | 0.001 |  |  |  |  |  |
| hsa-mir-7-3 | 0.8418009 | 0.7471886 | 0.9483935 | 0.005 |  |  |  |  |  |
| hsa-mir-708 | 1.1790963 | 1.0324241 | 1.3466056 | 0.015 |  |  |  |  |  |
| hsa-mir-7641-1 | 0.7989516 | 0.6604662 | 0.9664744 | 0.021 | 0.7480588 | 0.6058402 | 0.9236627 | 0.006974628 | -0.2902737 |
| hsa-mir-877 | 0.8738155 | 0.7681525 | 0.994013 | 0.04 |  |  |  |  |  |
| hsa-mir-96 | 0.8507312 | 0.7543245 | 0.9594592 | 0.008 |  |  |  |  |  |
| hsa-mir-100 | 1.1634062 | 1.04846 | 1.2909544 | 0.004 | 1.4611105 | 1.0225121 | 2.0878423 | 0.037323351 | 0.3791968 |
| hsa-mir-125a | 1.3489568 | 1.1081621 | 1.6420742 | 0.003 |  |  |  |  |  |
| hsa-mir-125b-1 | 1.121569 | 1.0053322 | 1.2512453 | 0.04 |  |  |  |  |  |
| hsa-mir-125b-2 | 1.1182873 | 1.0036428 | 1.2460274 | 0.043 |  |  |  |  |  |
| hsa-mir-133a-1 | 1.07794 | 1.0049005 | 1.1562882 | 0.036 |  |  |  |  |  |
| hsa-mir-137 | 1.1525751 | 1.0346797 | 1.283904 | 0.01 |  |  |  |  |  |
| hsa-mir-145 | 1.1343848 | 1.0275404 | 1.2523389 | 0.012 |  |  |  |  |  |
| hsa-mir-218-1 | 1.1391453 | 1.0105258 | 1.2841355 | 0.033 |  |  |  |  |  |
| hsa-mir-218-2 | 1.1507138 | 1.0215946 | 1.2961522 | 0.021 |  |  |  |  |  |
| hsa-mir-30a | 1.1503553 | 1.0139255 | 1.3051425 | 0.03 |  |  |  |  |  |
| hsa-mir-328 | 1.2967818 | 1.1042767 | 1.5228457 | 0.002 |  |  |  |  |  |
| hsa-mir-365a | 1.3038935 | 1.0584112 | 1.6063118 | 0.013 |  |  |  |  |  |
| hsa-mir-365b | 1.2984939 | 1.0539639 | 1.5997572 | 0.014 |  |  |  |  |  |
| hsa-mir-378i | 0.6588827 | 0.4965745 | 0.8742422 | 0.004 | 0.6392704 | 0.4723773 | 0.8651275 | 0.003749412 | -0.4474278 |
| hsa-mir-605 | 1.2152721 | 1.0319407 | 1.4311738 | 0.019 |  |  |  |  |  |



1.3.4 edge 也筛选出55下调，65上调的miRNA（负二项广义对数线性模型和拟似然负二项广义对数线性模型）

2.1 差异基因筛选（DEseq）

首先基于padj，在1556个miRNA中筛选出452个miRNA，其padj＜0.05

其中以log2FC>1为上调，log2FC<-1为下调，得到74个下调miRNA和166个上调miRNA

2.2 Univariate Cox regression and multivariate Cox regression by survival package(R)

险比例回归分析240个上调/下调miRNA对生存时间和生存状态的影响，得到34个p＜0.05的miRNA，其中4个经多因素Cox回归相关，为