# QUIZ-1 exSEEK

致理-生 21 贺昶 2022012374

2024 年 8 月 12 日

# 目录

# 图目录

# 1 研究背景

extracellular RNA(**exRNA**) 或 cell free RNA(**cfRNA**) 指的是存在于细胞外的 RNA。它们通过与 RNA 结合蛋白结合或被包裹而维持相对稳定。在细胞内翻译的 exRNA 的含量与修饰状态可以反映细胞本身乃至于人体的生理状态。在精准医疗的应用上，可以通过液体活检等方式无创收集唾液、尿液、血清等样本，收集其中的 exRNA，进而对病人情况进行分析。Heitzer et al., 2019

本次 Quiz 的目标即为利用生物信息学工具从二代测序数据中提取特征，使用一些机器学习方法，评估长、短两种 exRNA/cfRNA 对于健康人和癌症患者的区分能力，并给出一个有预测能力的特征组合。

# 2 数据集描述

两个数据集分别由血浆 long RNA 双端测序和 small RNA 单端测序产生。原始数据已经进行了预处理 (去除 adapter、去除低质量序列等) 和 mapping。

- Long RNA 数据集包含 373 个样本，样本来源于结肠癌 (CRC)，胃癌 (STAD)，肺癌 (LUAD)，食管癌 (ESCA) 和肝癌 (HCC)5 种癌症患者，以及健康人 (HD)。Short RNA 数据集
- Short RNA 数据集来自公共数据GSE71008。包括 192 个样本，来源于结肠癌 (CRC)，前列腺癌 (PC) 和胰腺癌 (PAAD) 以及 HD。每一个样本包含九种 RNA 类型的 bam 文件。

# 3 计数矩阵的生成

## 3.1 Long RNA

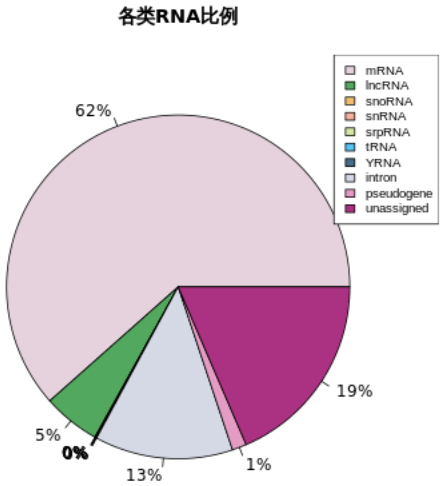
读取 metadata 文件中对样本采集时试剂盒版本的描述，调用 featurecount，选择 forward stranded 或 reverse stranded 的方式对每个样本进行 feature 计数。而后调用 summarize-table.py 脚本对所有样本进行整合，生成完整的计数矩阵。代码主要部分如下：

```
#!/bin/bash

# 建立一个 id 文件为 merge 做准备
touch ./long/sample_ids.txt

# 根据 library 属性进行 featurecount
awk 'NR > 1 {print $1,$5}' ./exRNA-long/metadata.txt | while read line
do
    if [[ -n "$(echo $line | grep "forward")" ]];then
        di="1"
    else
        di="0"
    fi
    sample_id=`echo "$line" | cut -d " " -f 1`
    echo $sample_id >> ./long/sample_ids.txt

    # 提高了线程数
    featureCounts --countReadPairs -O -M -s $di \
    -p -t exon -g gene_id -a ./exRNA-long/genome/gff/gencode.v38.annotation.gff3 \
    -o ./long/counts/${sample_id}.txt -T 24 ./exRNA-long/bam/${sample_id}.bam
done

python3 /data/2022-bioinfo-shared/data/quiz-I/scripts/summarize-table.py \
 --indir ./long/counts \
 --formatter '{}' --sample-ids ./long/sample_ids.txt \
 --row-field 0 --row-name gene_id --first-line 2 --value-field 6 --fillna \
 --output ./long/count.matrix.txt
```

Listing 1: counts_long.sh

进一步调用 reads-assignment.py 分析 long RNA 中各 RNA 类型比例。整合后可以发现：

- mRNA 占据其中的大多数，比例超过六成
- 有接近五分之一的 long RNA 未能匹配到 RNA 类型
- snoRNA, snRNA, tRNA 等几种相对较小的 RNA 比例相当低




### 3.2 Short RNA

对于 Short RNA，关注其中的 miRNA 与 piRNA 两种类型，对每一个样本调用课程提供的 count-transcripts.py 进行计数。具体代码与 long RNA 类似。而后编写 R 代码对每个样本的计数结果进行整合：

# 4 数据预处理

## 4.1 低表达基因过滤

调用 edgeR::DEGlist() 与 edgeR::filterByExpr 对两个矩阵进行了低表达基因过滤。

```
library(edgeR)
y = DGEList(counts=count.matrix)
keep = filterByExpr(y)
y = y[keep, , keep.lib.sizes=FALSE]
```

## 4.2 数据归一化

分别采取 TMM 与 RLE 两种方式对数据进行归一化，通过画图可知两组归一化方式结果区别不大，最终选择 TMM 的归一化方法。

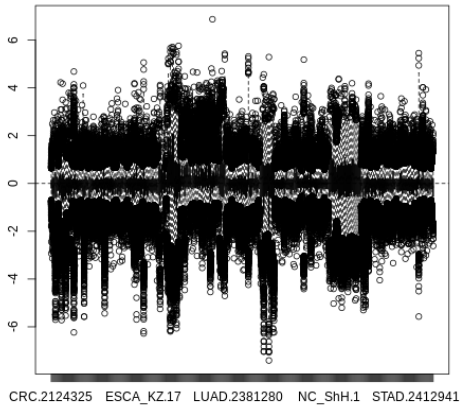



图 1: Long RNA, RLE 法

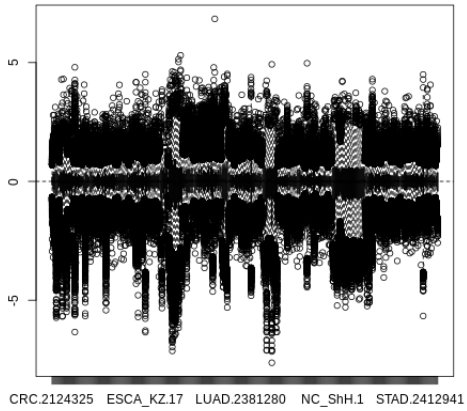



图 2: Long RNA, TMM 法

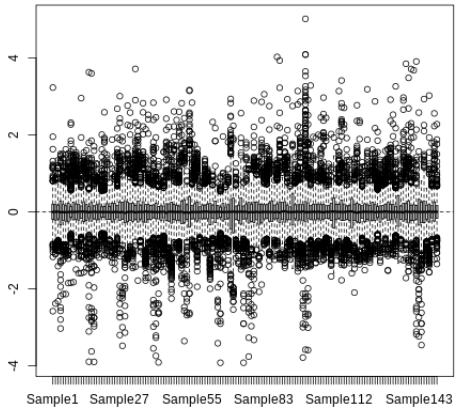



图 3: Short RNA, RLE 法

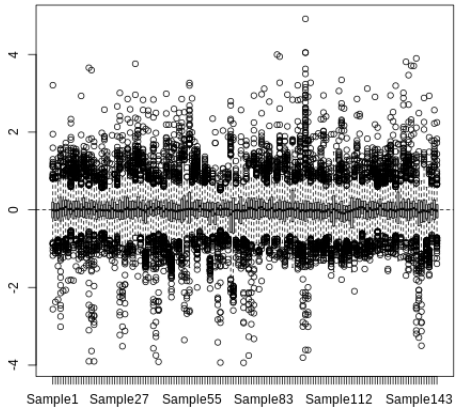



图 4: Short RNA, TMM 法

### 4.3 消除批次效应

对矩阵进行对数处理后，利用 sva::Combat() 方法消除数据的批次效应。通过 PCA 降至 2 维作图可以发现，处理后数据的批次效应有了一定程度的缓解。

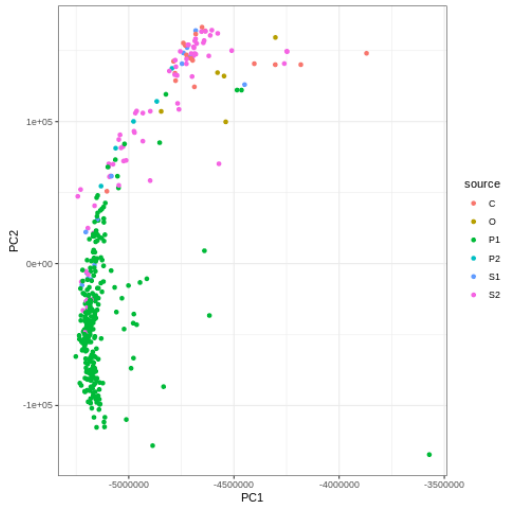



图 5: Long RNA, 消除前

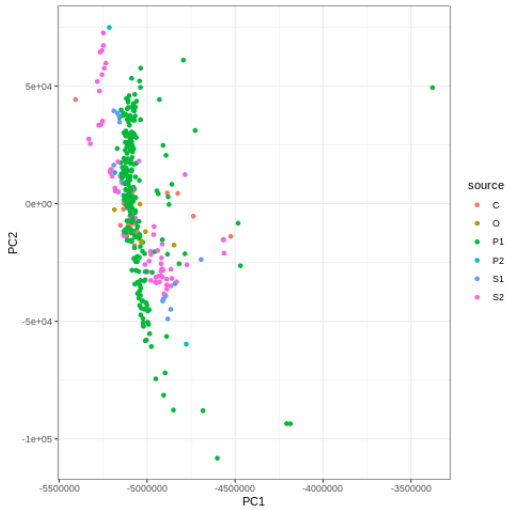



图 6: Long RNA, 消除后

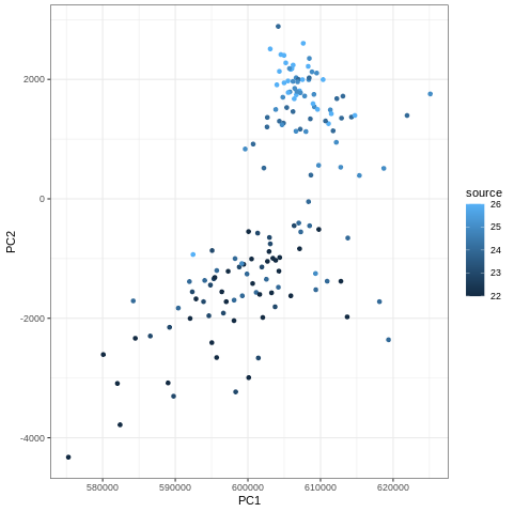



图 7: Short RNA, 消除前

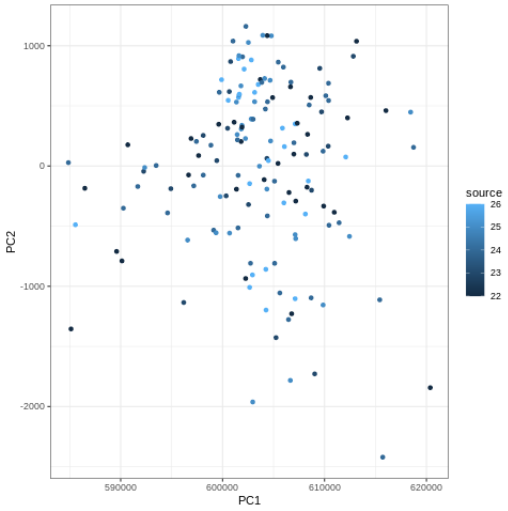



图 8: Short RNA, 消除后

# 5 机器学习

## 5.1 Long RNA

随机分割数据集为测试集与训练集，前者占 80%，基于 Logistic Regression 训练分类器，过程中采用交叉验证尽量规避小样本量带来的问题。

首先尝试采用递归特征消除 (RFE) 的方法进行特征选择，尝试特征个数为 10 到 100 内的每一个整十数及 100 到 900 内的每一个整百数，结果显示 90 个特征个数为最佳。应用这 90 个特征训练得到分类器一。

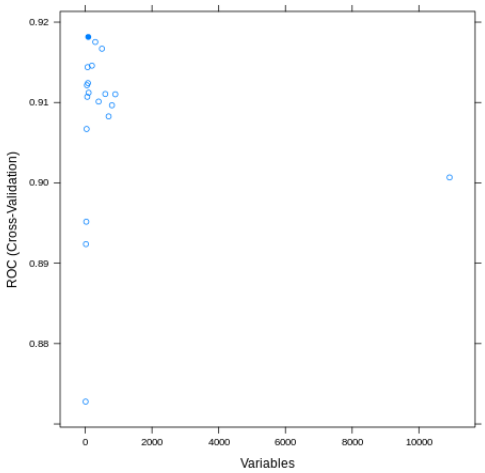



图 9: Long RNA RFE 特征选择结果

同时，考虑到特征选择本质上是一种数据降维，可以通过 PCA 代替进行降维与训练。观察碎石图可以发现，20 维 PCA 足够解释数据矩阵的绝大多数方差。通过 20 维度 PCA 训练得到基于 LR 的分类器二。

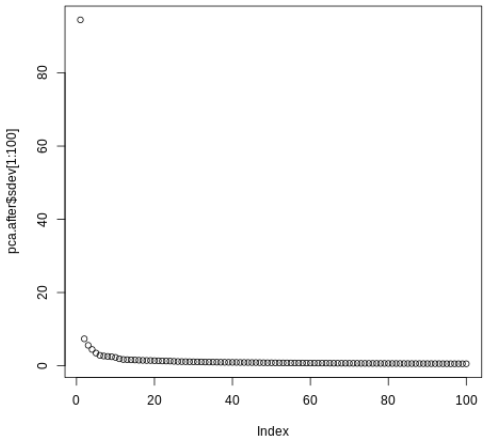



图 10: Long RNA 崖底碎石图

最后，考虑到 RNA 数据矩阵的高维度、稀疏性、相关性，线性模型如 LR 不一定有好的分类效果；而 Ranger 是一个基于随机森林的方法，可以实现分类与回归，对高维数据有很好的适配性。虽然随机森林分类器缺少可解释性，但可以训练 ranger 分类器作为参照。

画出三个模型的 ROC 曲线进行对比发现，以 20 维 PCA 训练的 LR 模型分类效果与 ranger 模型相近，AUC 达到 0.87；RFE 完成特征选择的 LR 模型分类效果相对较差。

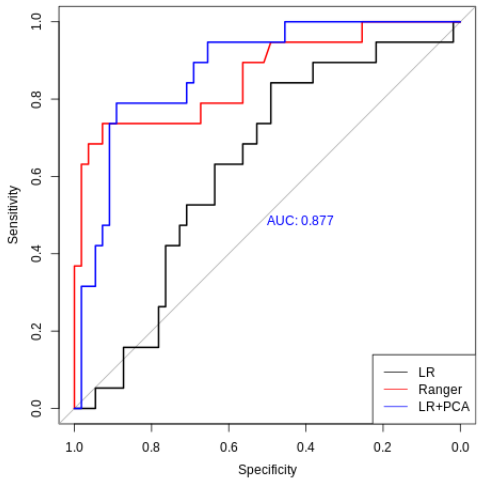



图 11: Long RNA 初步模型评估

但是，PCA 在特征选择上的可解释性相对 RFE 更差，不能够直接给出可靠的特征，而是特征的一组线性组合。因此，考虑选择 20 维 PCA 中总贡献最大的 20 个特征直接进行模型训练。评估分类效果，AUC 达到了 0.86，与 20 维 PCA 训练的分类模型相差无几。

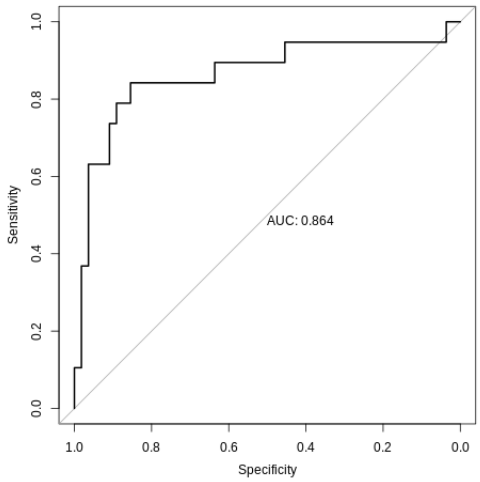



图 12: Long RNA 选择特征训练的模型评估

总结可以认为，这 20 个特征在区分健康与疾病样本上有较好的代表性，可能具有特殊的生物学意义。

## 5.2 Short RNA

对于 Short RNA，采用类似方法分割数据集、特征选择、训练基于 LR 的分类模型。与 Long RNA 不同的是，miRNA 与 piRNA 的表达矩阵相对更加稀疏，筛选低表达基因后仅剩 1277 个特征。因此，在采用 RFE 进行特征选择时选择得到了更少的特征组合：

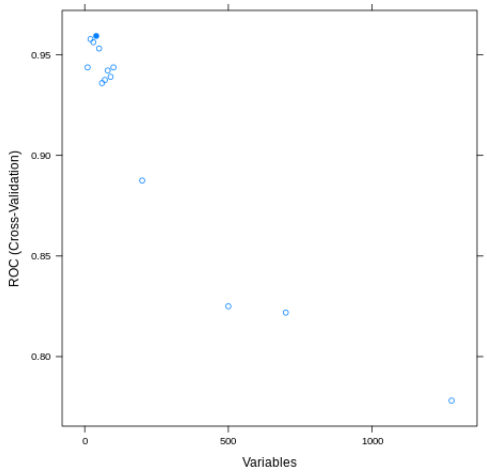



图 13: Short RNA 的 RFE 结果

在评估这 40 个特征训练的 LR 模型发现：与参照的 ranger 模型相比，LR 模型分类效果同样良好，AUC 达到 0.91。这提示了该特征组合可能的疾病相关性。

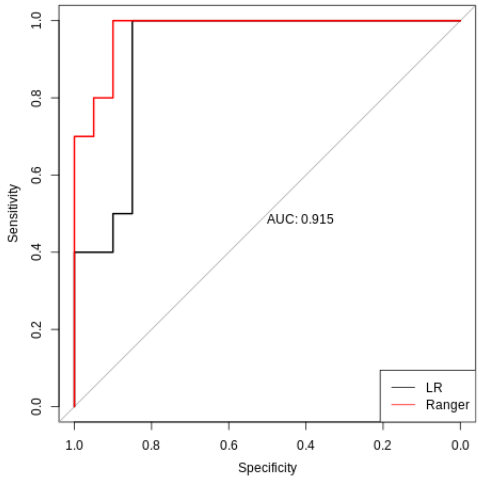



图 14: Short RNA LR 模型评估

## 6 特征解释

以 Long RNA 矩阵中选择的 20 个特征和 Short RNA 矩阵中选择的 40 个特征进行热图绘制。图中 lebel 表示癌症类型或健康（NC, HD）。

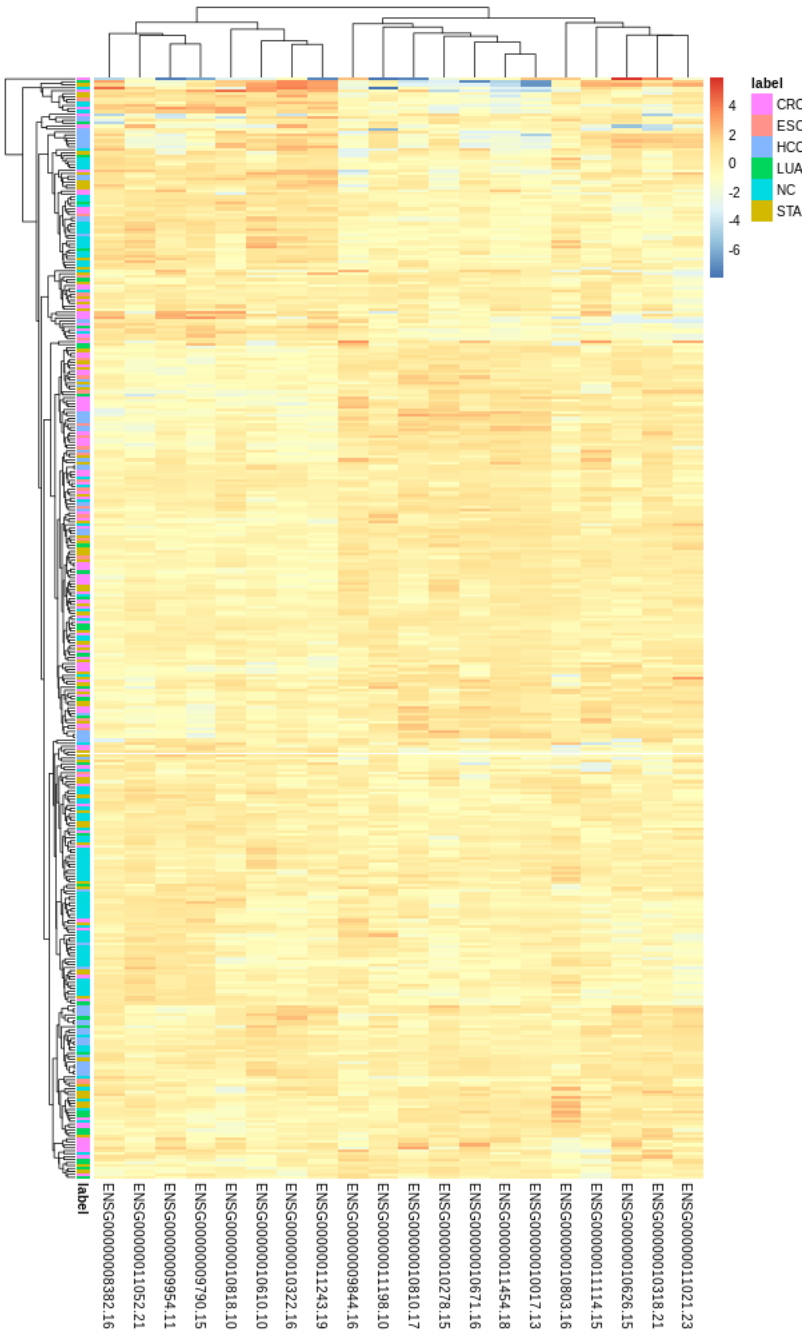



图 15: Long RNA 热图

长 RNA 的热图中可见不明显的区域化，对特征的聚类结果提示了特征之间可能的互作和相关性。

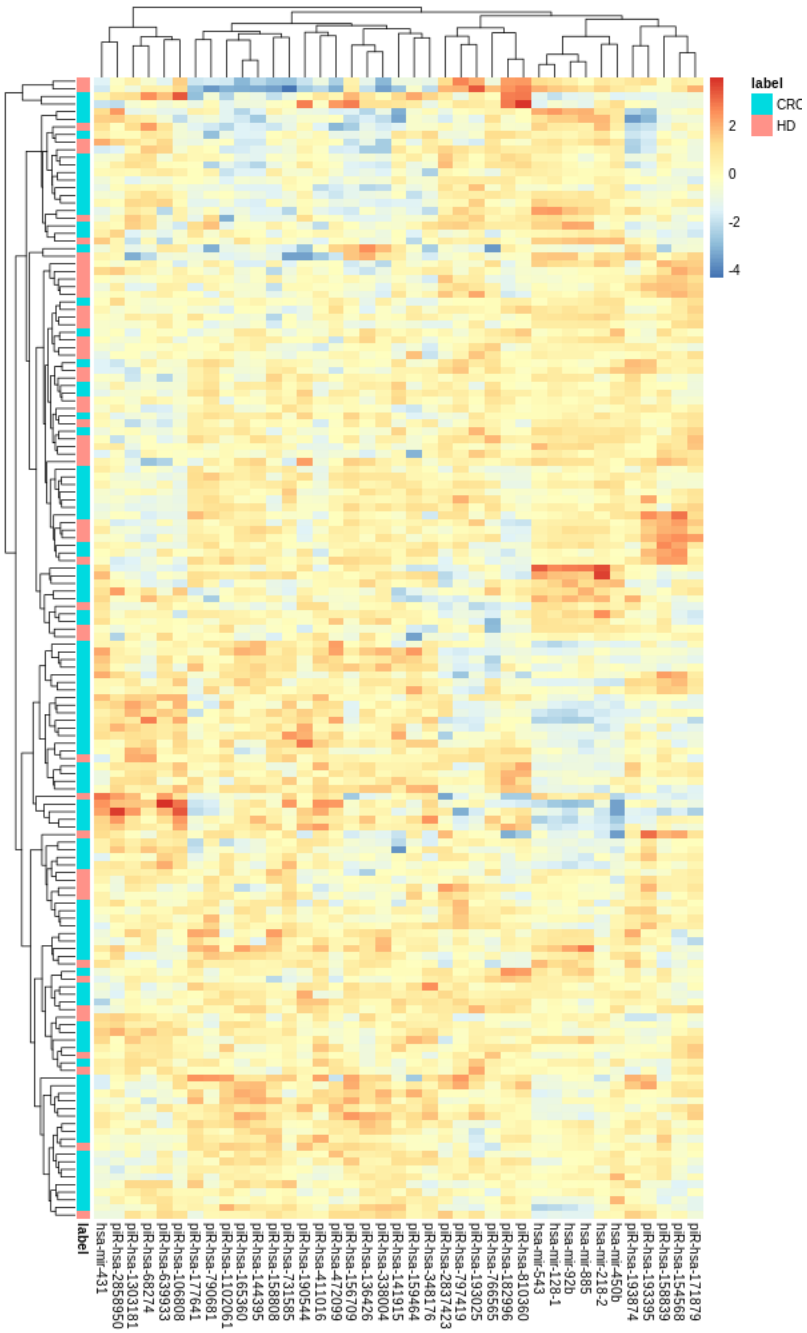



图 16: Short RNA 热图

短 RNA 热图中似乎并无明显的模式，原因可能是短 RNA 的复杂性和在体液中的相对低浓度，以及样本数量较少导致的随机性。

# 7 结论

在本次 QUIZ 中，尝试使用 Long RNA 与 Short RNA 中的 miRNA、piRNA 作为分辨疾病与健康样本的指标，训练基于 Logistic Regression 的分类器，初步证实了这一思路的合理性。但在对特征组合的解释上，各特征的模式并不明显，需要将较多的特征整合分析，这表明少量特征并不能很好的概括癌症样本与健康样本的区别。在模型分类性能方面，可以通过提升样本量以及改变算法，如尝试基于高斯核函数的 SVM（SVM RBF）等非线性分类算法或 ranger 等对高维数据优化的随机森林算法，来提升分类准确性。

# 参考文献

Heitzer, Ellen, Imran S. Haque, Charles E. S. Roberts, and Michael R. Speicher (Feb. 2019). "Current and Future Perspectives of Liquid Biopsies in Genomics-Driven Oncology". In: *Nature Reviews Genetics* 20.2, pp. 71–88. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-018-0071-5. (Visited on 08/07/2024).