

清华 大学

综合 论文 训 练

题目：基于细胞外 RNA 测序数据的
生物标志物鉴定工具开发

系 别：生命科学学院

专 业：生命科学

姓 名：陈旭鹏

指导教师：鲁志 教授

2019 年 6 月 12 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名：_____ 导师签名：_____ 日 期：_____

中文摘要

体液中的细胞外 RNA (exRNA) 可以为癌症鉴定提供大量候选生物标记物，从而实现癌症尤其是早期癌症检测。深度测序的发展使得全面地检测 exRNA 成为可能，然而由于 exRNA 测序数据的独特性质，从生物信息学的角度分析 exRNA-seq 并鉴定可临床使用的生物标志物仍然是一项非常有挑战性的工作。我们为此开发了 exSEEK，作为一种用于 exRNA-seq 分析和鉴定与癌症相关的生物标志物的生物信息学工具，可用于不同来源（细胞外游离、外泌体分泌）以及不同测序方法（小 RNA 测序、长 RNA 测序）产生的 exRNA 数据分析及生物标志物鉴定。

exRNA 测序数据具有高度碎片化、稀疏性、异质性和批次效应等特征。我们通过对 exRNA-seq 进行精细的比对，构建小 RNA-seq 基因表达矩阵时使用“结构域特征”，对表达矩阵进行多种样本库大小归一化和去除批次效应方法进行组合，使用非监督聚类准确性 (UCA) 和 m-K 最近邻 (mKNN) 指标衡量其对计数矩阵的归一化和去除批次效应效果。我们还开发了一个特征选择框架，使用逻辑斯谛回归，随机森林等机器学习模型用于选择区分癌症和正常样本的有效且稳定的特征。我们使用 exSEEK 对三个数据集进行了综合分析并评估了 exSEEK 在肝癌，结肠癌，胰腺癌和前列腺癌等癌症的分类性能，取得了良好的效果。最后我们还将 exSEEK 封装成一个简单易用的软件，并且提供交互性的可视化模块供用户使用。

关键词：液体活检；生物标志物；特征选择；机器学习

ABSTRACT

Extracellular RNAs (exRNAs) in body fluid provide a large repository of biomarker candidates for cancer (early) diagnosis. Deep sequencing makes it possible to monitor exRNAs in a comprehensive way. However, due to the unique properties of exRNA sequencing data, analyzing exRNA-seq data and identifying potential biomarkers for clinical usage remains challenging. Here we developed exSEEK, as a bioinformatics tool for identification of biomarkers associated with certain diseases by analyzing both small and long exRNA sequencing data generated by both cell free or exosome RNA.

exRNA-seq data are difficult to deal with since it is highly fragmented, sparse heterogeneous and has batch effect. In this work, we did careful mapping, constructing expression matrix using domain features. We applied various combinations of normalization and batch removal methods to the count matrix to correct data heterogeneity and batch effects. We evaluate the normalization and batch correction result using unsupervised clustering accuracy (UCA) and m-K-nearest neighbor (mKNN). We also developed a feature selection framework wrapping some machine learning models including logistic regression and random forest to robustly select the most important features that distinguish cancer from normal samples. We performed integrative analysis of three datasets: cell-free small RNA, exosomal small RNA and exosomal long RNA and evaluated the classification performance of HCC, CRC, PRAD, and PAAD. At last we wrapped all the useful functions into an easy-to-use software with interactive visualization modules for end users.

Keywords: liquid biopsy; biomarker; feature selection; machine learning

目 录

第 1 章 引言	1
1.1 液体活检与癌症早期诊断	1
1.2 exRNA 作为生物标志物的研究进展	2
1.3 机器学习与特征选择算法	3
1.4 研究计划概述	4
1.5 研究意义	5
第 2 章 算法与结果	7
2.1 exRNA 数据的收集与预处理	7
2.1.1 exRNA 数据收集	7
2.1.2 长 exRNA 测序数据的处理	8
2.1.3 小 exRNA 测序数据的处理	9
2.2 表达矩阵的构建	11
2.3 表达矩阵的处理	14
2.3.1 过滤与归责处理	15
2.3.2 样本库大小归一化	16
2.3.3 去除批次效应	17
2.3.4 评估表达矩阵处理效果	19
2.4 特征选择和模型评估	20
2.4.1 差异表达分析	21
2.4.2 特征选择策略与机器学习模型	23
2.4.3 特征选择的分类效果和稳健性评估	26
2.4.4 模型分类效果比较总结	27
2.4.5 挑选出的候选生物标志物表现	29
2.5 exSEEK 软件介绍及使用	31
2.5.1 软件基本功能模块与使用	31
2.5.2 软件绘图模块及使用	31

第3章 总结与讨论.....	33
3.1 结论	33
3.2 讨论	33
插图索引	36
表格索引	37
公式索引	38
参考文献	39
致 谢	42
声 明	43
附录A 外文资料的调研阅读报告	44
A.1 Research significance and scientific basis of the project	44
A.2 Current research status internationally	46
A.3 Previous analyzing tools for small RNA-seq	50

第1章 引言

1.1 液体活检与癌症早期诊断

癌症早期诊断的意义和问题 癌症又名为恶性肿瘤（Malignant tumor），指的是细胞不正常增生，且这些增生的细胞可能侵犯身体的其他部分^[1]；是由控制细胞分裂增殖机制失常而引起的疾病。在人类身上，目前已知的癌症超过一百种，大多数癌症未经合理治疗都会导致死亡，其治疗难度也远大于一般疾病^[2]。癌症早期诊断患者的五年生存率要比癌症晚期患者高 5~10 倍^[3]，在晚期癌症治愈手段匮乏的情况下，癌症的早期诊断对于提高治愈率和患者的生存率至关重要。和欧美发达国家相比，中国的癌症患者五年生存率低很多，其主要原因就是癌症早期诊断的技术不够先进，而即使在欧美国家，癌症的早期诊断技术也远没有成熟。

RNA 生物标志物作为液体活检指标 近年来，体液活检（liquid biopsy）受到人们的密切关注，体液活检相比传统的组织活检具有动态性强，无创性，成本低等特点。目前已报道的可以作为癌症检测生物标志物（biomarker）大多是蛋白质分子或者 ctDNA。如 2018 年发表的研究 CancerSEEK^[4]，通过整合 ctDNA 和蛋白质数据，可以从血液中对 8 种可能的癌症进行检出和分类，但在该方法在确定癌症类型上的表现并不完美，准确度最低只有 40%，而检测成本可以达到 500 美元一个样本，成本过高。而由于 RNA 在中心法则中处于特殊地位，与众多的生物学过程相联系，越来越多的研究发现其在疾病发生发展中可以作为一种更有优势的标志物，RNA 标志物与 DNA 和蛋白标志物相比，具有更好的敏感性和组织特异性^[5]。利用简单经济的一般 PCR 技术，便可以高灵敏度、高特异性地捕获和跟踪 RNA 序列。另外由于 RNA 分子在单个细胞中便拥有多个拷贝并且具有多种转录调控形态，RNA 分子标志物具备反映细胞状态与调控过程动态变化的优点。因此，大规模体液 RNA 表达数据的测定可以提供基因组差异与转录组动态变化的双重信息，可以作为准确直接的标志物用来无创地检测人体健康和疾病状态的变化^[6]，尤其是 RNA 的组织特异性克服了 ctDNA 难以从循环血中溯源的天然缺陷，对于检测和鉴定具有组织特异性的癌症具有重大的科研价值和应用前景。

1.2 exRNA 作为生物标志物的研究进展

国际上已有一些规模较大的研究团队和商业组织，开始将 exRNA (extracellular RNA) 作为生物标志物进行研究。美国 NIH 下属的转化科学国家中心 (NCATS) 在 2013 年启动了 exRNA 研究项目 ERCC(Extracellular RNA Communication Consortium)^[7]，研究内容包括 1) exRNA 的原理和功能; 2) exRNA 作为分子标志物的可能性; 3) 基于 exRNA 的癌症治疗方案等。2019 年的十大科技进展之一，用细胞外游离 RNA (cfRNA) 信息来预测孕妇的早产风险的研究也获得了广泛的关注^[8]。在之前的体液 exRNA 研究中，miRNA 受到了最多的研究关注，例如在一项针对肝癌的研究中，科学家使用 miRNA 芯片数据区分肝癌患者和正常人群、慢性乙型肝炎患者、肝硬化患者的血浆样本，得到 7 个分类效果最显著的 miRNA，构建了多元逻辑斯谛回归模型用以区分肝癌患者和其它对照人群^[9]。达到了很高的敏感度 (sensitivity) (约 80%-90%) 和特异性 (specificity) (约 70%-80%)，不过依然有 20% 左右的误诊率提升空间。

exRNA 生物信息学分析方法的发展与挑战 为了解析体液 exRNA-seq 数据，必须针对其特征设计专门的生物信息学工具。针对 exRNA-seq 数据的非均一化、易降解、碎片化、杂音大、具有批次效应、动态性更强等特点，目前尚缺乏专业和完整的生物信息学分析方法。例如，不同批次（如不同实验日期，不同实验条件等）取得体液样本之间存在很大差异。实验条件的不同也导致不同样本的库大小并不一致，因此还需要进行样本库大小的归一化。除此以外，RNA 分子除了能够反映基因组变异的信息，同时后转录调控过程使得 RNA 分子具有广泛的多态性，血液样本中的 RNA 分子不同于或组织中 RNA 的存在形式，往往受到降解作用的影响，多以碎片的形式存在，传统的构建表达矩阵（基因计数矩阵）的方法不够精确。同时由于 exRNA 的高度动态性和微量性，对于挑选出稳健的生物标志物也提出了很大的挑战。

现在已经有几种用于小 RNA-seq 分析的工具：ExceRpt^[10]，用于细胞外 sRNA 的整合基因组分析的工具 (TIGER)^[11]，TIGER 的一个关键改进是能够分析亲本 RNA 和单个片段水平以及宿主和非宿主 sRNA。Chimira^[12] 是一种基于网络的系统，用于小 RNA-seq 数据中 miRNA 的分析。包含自动清理，修剪，大小选择并直接比对到 miRNA 发夹序列等。产生表达矩阵用于随后的统计分析。Chimira 还提供了一套简单直观的工具，用于分析和解释比对结果。miRge^[13] 是一种处理小 RNA-Seq 数据以获得 microRNA 熵的方法。miRge 使用贝叶斯比对方法，按

照顺序与成熟的 miRNA，发夹 miRNA，非编码对齐 RNA 和 mRNA 比对。其他种类 RNA (tRNA, rRNA, snoRNA, mRNA) 也可以比对。miRge 能够在 52 分钟内同时分析 100 个小 RNA-seq 样品，提供关于 miRNA 表达的综合分析。Oasis^[14] 是一个 Web 应用程序，可以快速灵活地在线分析小 RNA-seq 数据。它是专门针对实验室终端用户设计的工具，提供易于使用的 Web 前端和演示视频教程，演示如何一步步地分析 sRNA-seq 数据。Oasis 包含差异表达模块以及用于稳健生物标志物检测的模块以及 GO 和通路富集分析。支持批量提交任务。Oasis 可以生成可下载的交互式 Web 报告，以便在本地系统上轻松可视化和分析数据。一些应用于 RNA-seq 和 single cell 数据的标准化和批次效应校正工具（包括 scImpute^[15], SCnorm^[16] 和 Combat^[17]）也可能用于 exRNA 数据的分析。但是针对 exRNA 测序数据的一些关键问题，如高噪音、碎片化，特征不够稳定等问题，目前还没有一个通用的整合性的工具可以兼顾到这些问题，完成 exRNA 数据的分析以及潜在生物标志物的发掘。

1.3 机器学习与特征选择算法

机器学习可以利用计算机、数学、统计等方法，利用计算机深入挖掘数据的内部分布、潜在特征等，完成对数据模式的学习和识别。机器学习 (machine learning, ML) 的方法可以从复杂的数据分布中发现数据的相关特征，提取出数据中最重要的特征，并且具有很强的数据拟合能力，一些经典的机器学习模型，如支持向量机 (support Vector Machines, SVM)，逻辑斯谛回归 (Logistic Regression, LR)，随机森林 (Random Forest, RF)，决策树 (decision trees, DTs) 等，已经被广泛应用于生物学数据分析中，包括疾病尤其是癌症的分类和检测中。

常见的机器学习方法包括有监督学习和无监督学习。有监督学习中，模型通过获得输入数据以及对应的每个样本的标签，来学习输入数据和输出标签的内在映射，有监督学习包括分类和回归两类，对于标签为离散值的预测问题（如癌症，正常人两类标签）被称为分类问题，标签为连续值的问题为回归问题。与有监督学习不同，无监督学习不提供输入标签，因此模型需要将输入的样本进行聚类，每个类别具有类似的特征。

有监督学习的分类模型要求输入数据集和对应的类别标签对模型进行训练，之后模型可以对新输入的数据进行预测。数据集一般是数值矩阵，如本问题中使用的 exRNA-seq 测序数据构建出的基因表达矩阵，除此之外，基因组，蛋白

质组，代谢组，图像等数据都可以作为数据集输入。类别标签为癌症和正常人，可以被数值化以便输入模型。在训练模型并进行模型预测和模型评估时，还需要进行交叉验证（cross validation, CV），交叉验证通过将数据分为训练集和测试集，在训练集上训练数据，在独立的测试集上测试数据，可以很好地反应模型的泛化能力。最常见的 CV 方法为 K 折交叉验证（K-fold CV），在 K-fold CV 中，数据集被划分为 K 个互斥的子集，分类器每次在 K-1 个子集上训练并在 1 个子集上进行测试，直到每个子集都被用作测试数据集。最终通过计算 K 折上的平均准确度作为模型的准确度。

在机器学习和统计学中，特征选择（feature selection）也被称为变量选择、属性选择。即为了构建模型而选择相关特征（即属性、指标）子集的过程^[18]。使用特征选择技术可以简化模型，增加模型解释性，缩短训练时间，降低过拟合等。如果训练数据包含许多冗余或无关的特征，就可以移除这些特征而不损失信息。对于癌症检测以及挑选生物标志物的任务而言，挑选出少量的生物标志物的同时保证较高的预测准确率、灵敏度和特异性，对于实际应用非常有意义。通过选择不同的评价指标，可以把特征选择算法分为三类：包装类（wrapper）、过滤类（filter）和嵌入类（embedding）方法。嵌入式方法通过在分类算法中构建特征选择来评估最佳特征子集。与过滤方法相比，嵌入式方法更加节省计算资源。特征选择方法对于决定最终的分类效果非常重要，由于基因组学数据往往具有很高的特征维度和很低的样本维度，因此选择少量的足够有代表性的特征非常重要，选择稳定的，有解释性的特征具有相当的挑战性。

不同的 ML 技术和特征选择算法已被广泛应用于癌症的预测和预后。如使用多元逻辑斯谛回归区分肝癌和正常样本的研究^[19]，使用带有正则化的随机森林方法进行不影响预测性能的特征选择用于癌症检测^[19] 等。

1.4 研究计划概述

exSEEK 的主要功能如图 1.1 所示，包括测序数据的清洗，质量控制，序列比对映射，对于小 RNA 数据我们专门探索了顺序比对的方法。对于构建表达矩阵（基因计数矩阵），我们针对细胞外 RNA 测序数据的特点，设计了结构域检测的算法，检测 lncRNA、mRNA、snoRNA、snRNA、srpRNA、tRNA、Y RNA 等 RNA 的结构域特征用于构建表达矩阵，并与 miRNA 合并产生计数矩阵，之后我们使用一系列的矩阵处理模型对表达矩阵进行测序深度的归一化以及去除批

次效应，并且使用针对性的指标衡量矩阵处理的效果。最后我们会构建特征选择的流程来挑选可以用于癌症分类的潜在生物标志物，使用一些机器学习模型进行特征的挑选，并对其分类效果进行评估。最后我们会将代码封装为一个操作简单的命令行软件，并且提供可交互的可视化模块方便终端用户进行数据的可视化和分析。

本课题是 Lu lab 的一项研究的一部分，我负责了测序数据的处理，表达矩阵的处理以及简单的机器学习模型的搭建和可视化分析部分，完成了所有代码中的约 40% 的工作。

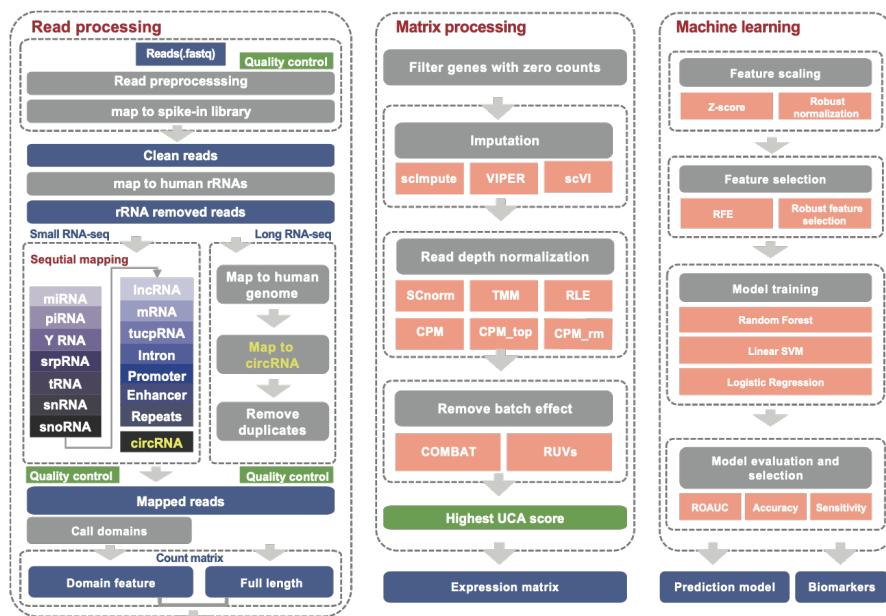


图 1.1 exSEEK 流程图

1.5 研究意义

我们已经总结了体液检测对癌症早期诊断和预后的意义，其中使用 exRNA 作为分析数据鉴定癌症尤其是早期癌症是非常有潜力和意义的研究。目前针对 exRNA-seq 数据的分析工具和方法依然缺乏，大多数都来自于其他类型数据（如 RNA-seq, single cell）分析工具的迁移。这些方法依然无法解决 exRNA-seq 数据中的一些关键问题，如数据的高度碎片化和噪音，数据的微量性和不稳定性。同时也欠缺一个覆盖所有流程的统一的 exRNA-seq 分析工具。我们通过进行精细的序列比对，质量控制，结构域检测，样本库归一化，去除批次效应，以及进行稳定特征的选择等步骤，构建了一个操作简单的 exRNA-seq 分析和潜在生

物标志物挑选工具，可以作为 exRNA-seq 数据分析的有力工具。相信在未来随着 exRNA-seq 测序数据的大量产生，机器学习模型将在矩阵处理和特征选择中发挥更加重要的作用，可以对 exRNA-seq 数据的分析和生物标志物挖掘发挥更加重要的作用。

第2章 算法与结果

2.1 exRNA 数据的收集与预处理

2.1.1 exRNA 数据收集

我们收集整理了实验室自产的一套和两套发表的 exRNA-seq 数据，为了后续方便，我们称细胞外游离的 RNA 测序数据为 cfRNA-seq，外泌体的 RNA 测序数据为 exoRNA-seq。同时我们用 L(long) 和 S(small) 区分长 RNA 和小 RNA 测序数据。一共四种主要的数据类型：S-cfRNA-seq, S-exoRNA-seq, L-cfRNA-seq, L-exoRNA-seq。我们收集的数据的具体信息如下：S-exoRNA-seq 数据来自于 GEO 数据库中收录的 GSE71008^[20]，包含健康供体 (healthy donor, HD) 的 50 个，结直肠癌 100 个 (colorectal cancer, CRC) 样本，前列腺癌 (prostate adenocarcinoma, PRAD) 36 个。S-cfRNA-seq，我们整理了实验室内部产生的数据 GSE123972^[21] 以及 GSE113994^[22]，GSE53080^[23] 和 GSE94582^[24]；GSE123973 有可用的肝癌 (hepatocellular carcinoma, HCC) 样本 30 例，其中包括早期肝癌 (HCC stageA) 16 个，健康供体 (HD) 的 13 个样本。GSE113994 包含 53 健康供体，GSE53080 包含 17 健康供体，GSE94582 包含 20 健康供体。对于 L-exoRNA-seq 数据，我们收集 exoRBase 数据库^[25] 中的数据，其源于多个实验室的合作，包括肝癌样本 14 个，结肠癌样本 12 个，胰腺癌样本 14 个和健康供体 32 个。数据总结如表 2.1 所示

表 2.1 exRNA 数据收集总结

Data type	Sources	Sample class	Sample size
S-exoRNA-seq	GSE71008	CRC, PRAD, HD	186
S-cfRNA-seq	GSE123972	HCC, HD	43
S-cfRNA-seq	GSE113994	HD	53
S-cfRNA-seq	GSE53080	HD	17
S-cfRNA-seq	GSE94582	HD	20
L-exoRNA-seq	exoRBase	HCC, CRC, PRAD, HD	20

exRNA 测序数据的处理 针对 exRNA 数据微量性的特点，我们专门设计了小 RNA 测序数据的顺序比对方法，确定了其比对的先后顺序。完整的流程如图 2.1 所示。包括数据清洗，质量控制，顺序比对（针对小 RNA）或序列比对，以及构建表达矩阵，在构建表达矩阵时，针对小 RNA 测序中的长 RNA（如 Y RNA, lncRNA, snoRNA 等），课题合作成员史斌斌还开发了专门的寻找其片段作为特征的方法。由于 exRNA 的微量性，不同样本间的各 RNA 比例变化很大，不论是确定比对顺序，还是做质量控制（quality control, QC）均需要精细的控制，我们使用丰富的可视化方法获得各类 RNA 的比例的统计结果，以及质量控制的结果总结等。接下来分别简要叙述长 exRNA 测序数据 (L-exoRNA-seq) 和小 exRNA(s-exoRNA-seq, s-cfRNA-seq) 测序数据的处理流程和结果。

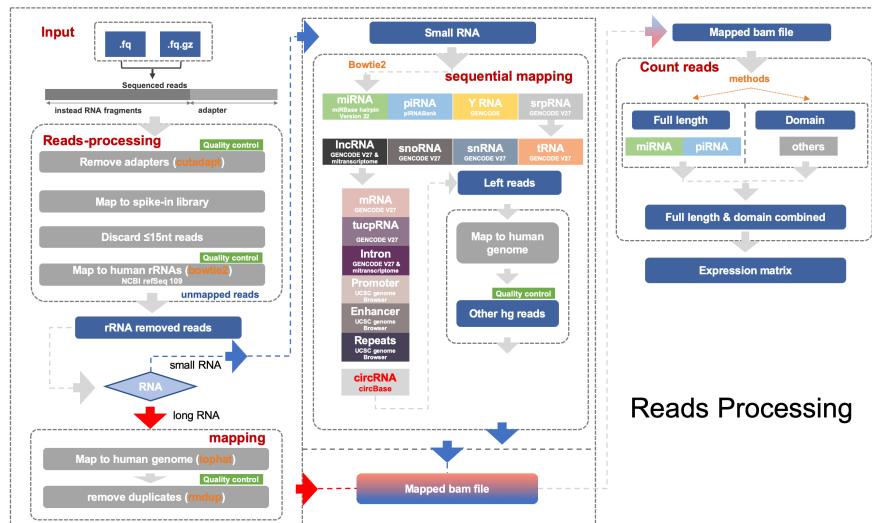


图 2.1 reads 处理流程图

2.1.2 长 exRNA 测序数据的处理

我们采用 L-exoRNA-seq 数据作为长 exRNA 数据，数据来自于 exoRBase 数据库^[25]，测序类型为长 RNA-seq 双端 (paired end) 测序。我们首先用 cutadapt 工具实现对接头的剪切 (trim adapter)，因为是双端测序，我们要求只要其中任意一个 read 的质量分数低于 30 就会被滤除。我们首先将 reads 比对到 rRNA 并且去除掉所有比对上的 reads，因为 rRNA 不会被分泌到细胞外，不符合我们的要求。然后我们将 reads 比对到人类基因组 (version: human genome 38)，除此之外，我们还额外关注了 circular RNA 的信息，因此在比对到人类基因组之后，我们专门又将 reads 比对到 circBase 数据库。由于是双端测序，我们规定需要在互

相配对的两条 reads 均比对成功的前提下才算作比对成功。长 exRNA 测序数据还有较为严重的因为 qPCR 扩增导致的 duplication 问题，因此我们还是用 picard 去除了 duplicates，最后再使用 featurecounts 软件生成基因表达矩阵（基因计数矩阵）。

长 RNA 比对结果统计 exSEEK 可以自动统计出入图 2.2 所示的比对结果，(A) 表示不同 RNA 映射比例的饼图；(B) 表示不同 RNA 的长度分布的三维条形图；(C) 表示不同 RNA 的映射比例的箱线图 (D) 表示每个样本各种 RNA 映射比例的叠加条形图。

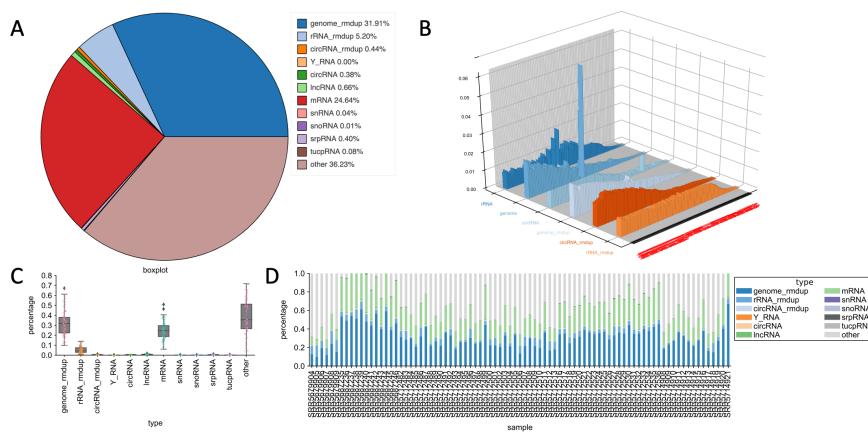


图 2.2 L-exoRNA-seq 数据 mapping 基本情况统计。

可以看到长 exRNA 测序数据的结果中，绝大部分的 reads 比对到了 mRNA 和人类基因组上以及有较大比例没有比对上的 reads。另外还有少量的 circRNA, srpRNA, lncRNA 等。我们在构建表达矩阵时并不使用 miRNA 和 piRNA，因为 exoRBase 建库时使用的是双端 150 (PE150) 测序，自动筛选掉了 RNA 长度较低的 miRNA 和 piRNA。值得注意的是，虽然非编码 RNA 如 circRNA, srpRNA, lncRNA 等的绝对比例不高，但是后续分析可以发现发生差异表达的 RNA 中这些非编码 RNA 可能占有较高的比例，所以它们依然有作为生物标志物的潜力。

2.1.3 小 exRNA 测序数据的处理

我们采用 S-exoRNA-seq 和 S-cfRNA-seq 数据作为小 exRNA 数据，S-exoRNA-seq 数据来自于 GEO 数据库中收录的 GSE71008^[20]，S-cfRNA-seq 来

自于 GSE123972^[21] 以及 GSE113994^[22], GSE53080^[23] 和 GSE94582^[24]，测序类型为小 RNA-seq 单端 (single end) 测序。

我们首先用 cutadapt 工具实现对接头的剪切 (trim adapter)，相比于双端测序，单端测序时只需要去除 3' 端的接头即可，我们同样要求 reads 的质量得分必须要于 30 分，且要求其长度必须在 16 个碱基长度到 50 个碱基长度之间。我们首选可选择地将 reads 比对到 spikein 序列上，然后是 rRNA 数据库以及载体数据库 (UniVec) 上 (以去除被载体污染的 reads)。接下来针对小 RNA 测序长度短，量小的问题，为了关注我们感兴趣的 RNA，我们使用单独比对的方法，分别将 reads 按顺序比对到各个 RNA 注释数据库上，我们发现不同的比对顺序会造成各类 RNA 的比对比例产生较大的变化，因此我们在前期工作中探索了不同的比对顺序的影响，因为使用 Bowtie2 进行一次比对需要的时间很长，我们使用了专门设计的算法可以在一次比对后测定数十种不同的测序顺序各个 RNA 的比例，确定了我们关注的 RNA 如 lncRNA, mRNA 等的比例如何要求，最终确定了 lncRNA、miRNA、mRNA、piRNA、snoRNA、snRNA、srpRNA、tRNA、TUCP RNA、Y RNA 的比对顺序。对于无法比对到人类基因组的 reads，我们将其比对到启动子，增强子，和重复区域 (enhancer, promoter, repeats) 等位置，最终将剩余的 reads 比对到 circRNA 上，对于依然无法比对的 reads，我们将其命名为 unmapped reads。

小 RNA 比对结果统计 exSEEK 可以自动统计出入图 2.3所示的比对结果，(A) 表示不同 RNA 映射比例的饼图；(B) 表示不同 RNA 的长度分布的三维条形图；(C) 表示不同 RNA 的映射比例的箱线图 (D) 表示每个样本各种 RNA 映射比例的叠加条形图。

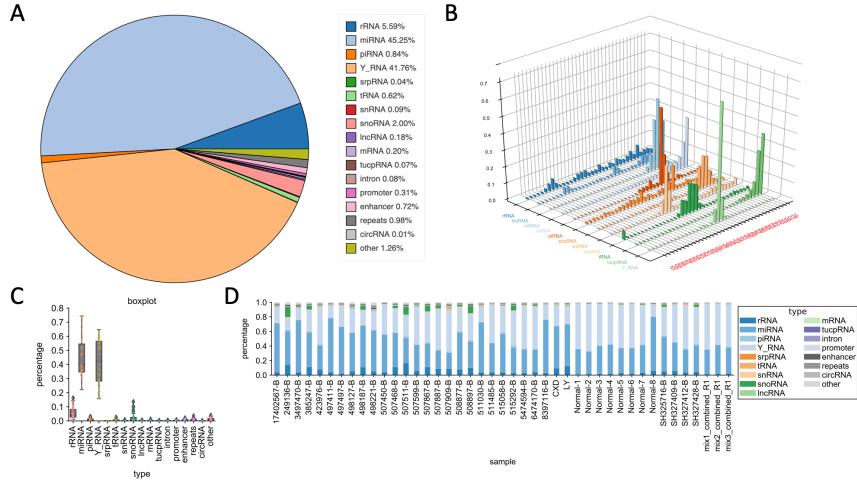


图 2.3 S-cfRNA-seq 数据 mapping 基本情况统计

通过对比结果统计图可以看到，大多数的 reads 比对到了 Y RNA, miRNA 和 rRNA 上，snoRNA, enhancer 和 repeats 也有一定比例的 reads 比对上。我们发现 rRNA, miRNA 和 Y RNA 不但比例较高，而且不同样本间的比例还有较大的差异，这可能是因为不同的样本存在一些批次效应（如 RNA 提取方法等的差异）。Y RNA 的种类很少，只有四条，但是其丰度极大，有百分之四十左右的 reads 会比对到这四条 RNA 上。Y RNA 是小的非编码 RNA。它们是 Ro60 核糖核蛋白颗粒的组分，是系统性红斑狼疮患者的自身免疫抗体的靶标。它们也是通过与染色质和起始蛋白相互作用进行 DNA 复制所必需的。Y RNA 在一些人类中过度表达，是细胞增殖所必需的，其分解产物也会参与到自身免疫和一些其他病理情况中^[26]，因此 Y RNA 也有潜在的区分癌症和正常样本的功能。值得注意的是 circRNA 在小 RNA 测序中的比例极低，不到万分之一，因此可能不能作为潜在的生物标志物使用，这和长 exRNA 测序中的结果不同，与长 exRNA 测序类似的是，虽然非编码 RNA 如 lncRNA, snoRNA, tRNA 等的绝对比例不高，但是后续分析可以发现发生差异表达的 RNA 中这些非编码 RNA 可能占有较高的比例，所以它们依然有作为生物标志物的潜力。

2.2 表达矩阵的构建

完成了 exRNA-seq 数据的收集，预处理和比对后，我们进行了 exRNA-seq 数据的表达矩阵的构建。传统的表达矩阵构建只需要使用如 featurecounts 之类的软

件工具即可。对于 L-exoRNA-seq 数据我们就是这么处理的。对于小 exRNA-seq 测序数据，我们采取了由课题合作成员史斌斌专门设计的结构域检测（domain calling）方法来发现结构域特征，并以结构域特征取代全长特征来构建表达矩阵。

小 exRNA-seq 测序数据的碎片化特征 我们首先从 exRNA-seq 数据的比对部分获得图 2.4，使用三维条状图和折线图展示了 S-cfRNA-seq 数据的各种类型 RNA 的长度分布。可以发现在 S-exRNA-seq 中，碎片化的情况非常严重，大多数的长 RNA 的 reads 长度也集中在 20-30 个碱基长度范围内。为此我们希望可以找到信噪比较高的 reads 覆盖区域更加集中的片段作为特征，以取代其全长特征。我们开发了一种专门针对 exRNA 数据的域检测算法，其总体设计思想与传统的根据 read 起始位置计算其覆盖值的 peak calling 软件 Piranha 类似，但是我们的方法可以做到更高灵的敏度以及找到更加准确的域位置。

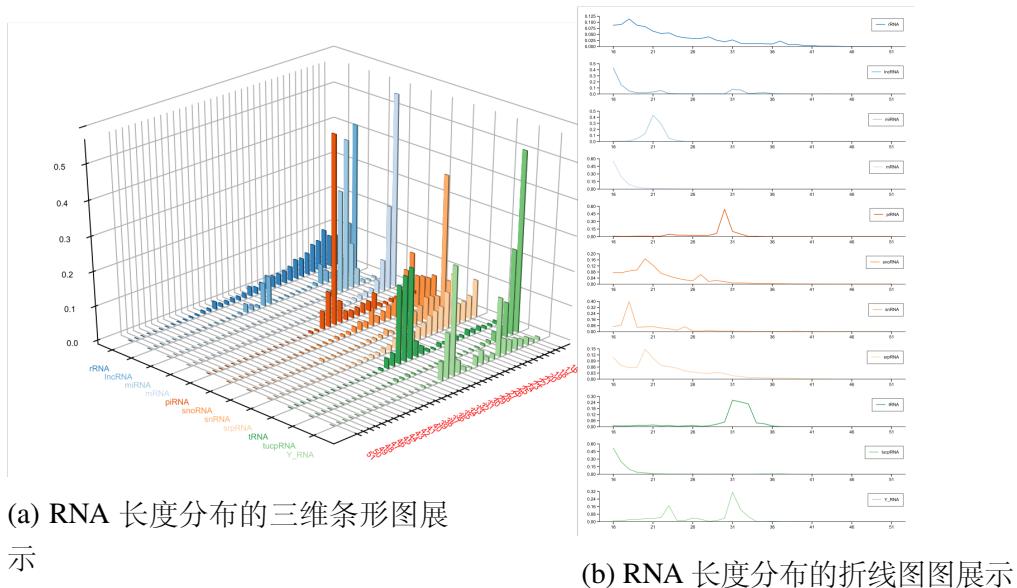


图 2.4 RNA 长度分布

结构域可视化 由于结构域检测算法主要由史斌斌同学完成，这里不再叙述其具体原理和过程，只展示相关的部分分析结果。为了展示 exRNA-seq 数据，尤其是小 exRNA 数据缺失存在明显的碎片化特征，我们比较了四套 RNA-seq 测序数据，两套来自小 exRNA-seq 数据，两套来自于组织 RNA 数据。我们以 S-cfRNA-seq 和 S-exoRNA-seq 的每个峰的中点作为原点坐标进行覆盖度的可视化，

如图 2.5 所示，可以看到对于 exRNA 数据，在峰的周围有非常显著的 reads 覆盖度的凸起，高于周围的区域，而对于组织 RNA 数据则没有明显的峰存在。我们还对 exRNA 的结构域长度统计，发现大多数结构域片段的长度集中在 30 个碱基长度左右，很少超过 100 个碱基长度，这进一步佐证了 exRNA 测序数据的大多数 reads 都以碎片化形式存在，而不是全长形式。

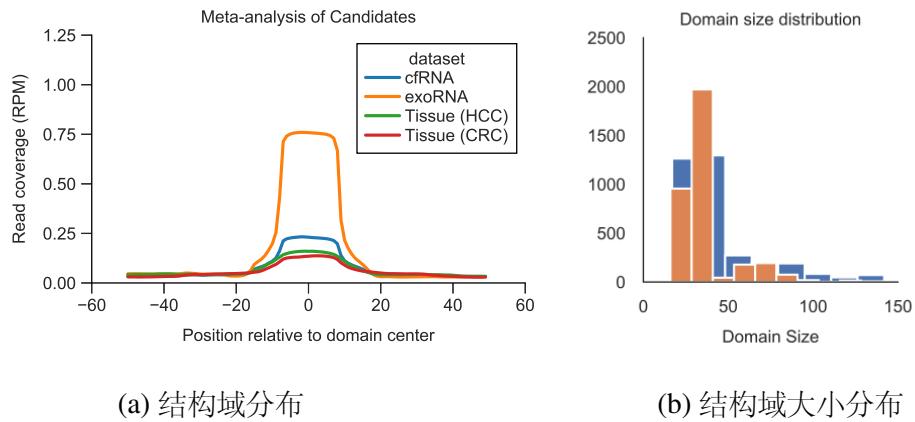


图 2.5 不同 exRNA 和组织数据结构域分布和大小分布

表达矩阵的数据统计 最终对于小 exRNA-seq 测序数据，我们构建了由长 RNA 的结构域特征加 miRNA 全长特征合并的表达矩阵，对于长 exRNA-seq 数据，我们直接适用 featurecounts 构建全长特征，作为下一步表达矩阵处理的输入。对于我们专门使用结构域检测算法特殊处理的小 exRNA 测序数据的表达矩阵，我们还做了如下的可视化分析。如图 2.6 所示，对于 S-cfRNA-seq 数据，我们发现 miRNA，Y RNA 和 lncRNA 占据了较高的丰度，但是由于使用了结构域检测算法，一些非编码 RNA 如 snoRNA、tRNA、tucpRNA、snRNA 和 srpRNA 均可以有较多的种类被检测到，显示出结构域检测算法可以帮助提供非编码 RNA 在表达矩阵中的多样性和种类数量。对于 S-exoRNA-seq 数据同样也可以观察到这样的特征，甚至更加明显，其中几种丰度非常微量的 RNA 如 snRNA、lncRNA、mRNA、tucpRNA、srpRNA 和 snoRNA 经过结构域检测算法，可以得到非常多样的特征。

进一步地为了检测这些特征是否对于区分癌症和正常样本有用，我们绘制了如图 2.7 所示的差异表达分析中各类 RNA 所占比例的饼图。我们发现虽然 miRNA 在 exRNA-seq 数据中的丰度很大，但是其在 S-cfRNA-seq 和 S-exoRNA-

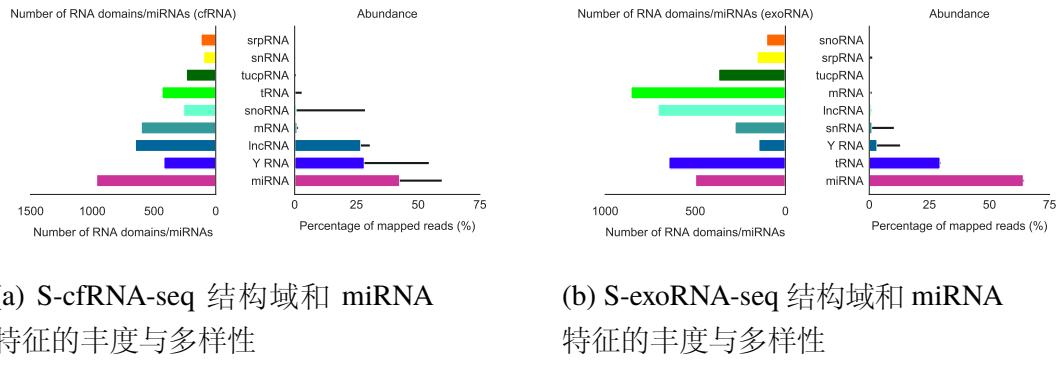


图 2.6 小 exRNA-seq 结构域和 miRNA 特征的丰度与多样性

seq 的差异表达基因中所占的比例只有 20% 左右，这进一步说明了其他丰度较低的 RNA 其实具有很高的多样性，而且有潜力作为区分癌症和正常样本的生物标志物，作为过去被广泛研究的 miRNA 癌症生物标志物的补充。

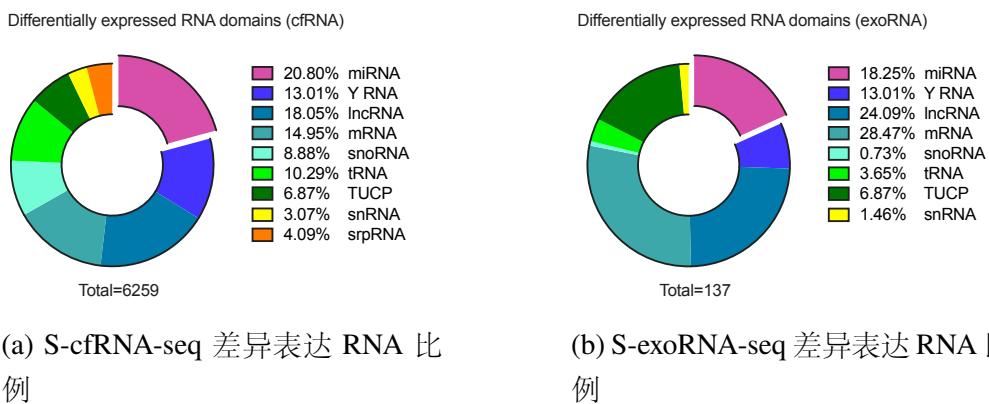


图 2.7 小 exRNA-seq 差异表达 RNA 比例

2.3 表达矩阵的处理

在矩阵处理部分，我们进行了如图 2.8 的顺序处理过程，包括过滤与归责，样本库大小归一化(测序深度归一化)，去除批次效应等步骤。我们将不同的处理方法进行组合，并对每个步骤进行了单独的和综合的评估，使用了非监督聚类准确性 (unsupervised clustering accuracy, UCA) 和 m-K 最近邻 (m-K-nearest neighbor, mKNN) 两个指标综合选取最佳的矩阵处理方法组合。本部分的矩阵处理代码由 R 语言实现，评估指标由 python 实现。

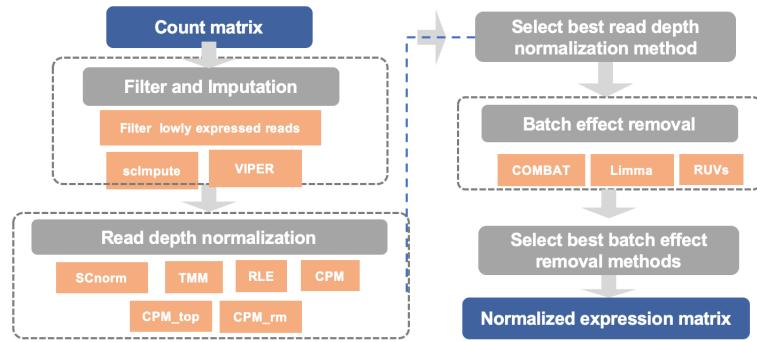


图 2.8 表达矩阵处理流程图

2.3.1 过滤与归责处理

我们首先是用了过滤和归责处理 (imputation) 的方法对表达矩阵进行处理。因为 exRNA 数据的稀疏化特征，使用过滤和归责是很常规的想法。

由于 exRNA 中有大量的缺失值，因此过滤掉一些整体表达值很低的基因是常见的做法。exSEEK 可以自己设置过滤的条件，比如对于原始基因计数矩阵，或者 counts per million(CPM), reads per kb per million reads(PRKM) 进行过滤，默认设置为过滤掉基因计数矩阵中表达值小于 5 的样本超过 50% 的那些基因。

归责处理 归责处理即对数据中的缺失值进行推断和插补，我们使用 scImpute^[15] 进行归责处理，scImpute 的思想如图 2.9：

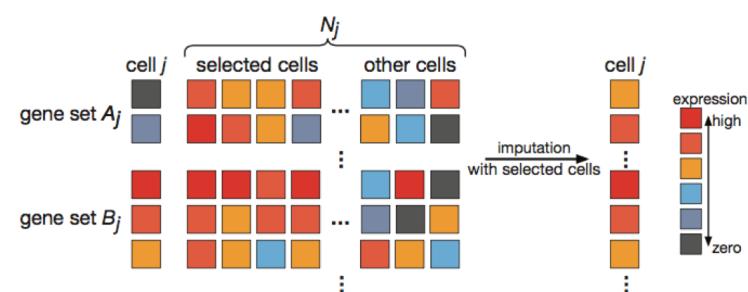


图 2.9 scImpute 原理

- 首先，使用一个混合模型（mixture model）学习每个细胞中每个基因存在缺失的概率。
- 接着，插补很大概率为缺失值的基因。首先将基因分为较大程度受到缺失影响的集合 A_j 和没有受到缺失值较大影响的集合 B_j ，然后根据 B_j 的基因

首先做主成分分析，找到解释性较大的几个主成分，利用这几个主成分的矩阵对样本进行聚类。

- 根据聚类结果，可以将样本分为有较高可能有缺失值和较高可能性表达值正常测定的两类，对表达值正常的样本进行线性回归，并将权重用于较高可能缺失的样本上，就完成了对缺失值的插补。

2.3.2 样本库大小归一化

对样本库大小 (library size) 进行归一化的原因在于实验中不同样本的计数总数会受到实验时的诸多因素影响 (如 PCR 扩增倍数，测序深度等)，造成样本间同种基因的计数有差异。对于 RNA-seq 数据，我们通常认为这种影响因素对于每个基因是等比例的，然而对于量较少的 exRNA-seq 以及单细胞测序，也有理论认为对于样本内部的不同基因，也应该使用不同的归一化因子，单细胞测序的归一化方法之一 SCnorm^[16] 即采取这种策略。基本的样本库归一化方法是将每个样本的所有基因的原始 reads 各乘以一个归一化因子，保证各个样本的样本库大小在归一化后一致。我们选择了多种归一化方法：CPM, CPM_top, TMM, RLE, UQ, SCnorm.

样本库大小归一化方法原理简介

CPM (counts per million) 通过对样本的所有 reads 求和得到样本库大小，每个基因的 reads 数除以该样本总 reads 数再乘以 10^6 即为归一化后该基因的表达值。

我们对 CPM 做了一些修改，CPM_top 算法对于 top20 表达量的基因和其余基因分别使用不同的归一化因子进行归一化，这是考虑到 exRNA-seq 数据中表达量前 20 的基因可以占据 reads 总数的 50% 以上，分别归一化可以避免表达量极大的基因对其余基因的分布的影响，这和 SCnorm 的想法比较类似。

UQ^[27] (upper quartile) 对 CPM 进行一些修改，与 CPM_top 类似。UQ 用在至少一个样本中表达的基因 reads 数的上四分位数 (75% 分位数) 来作为样本库大小进行归一化。这样可以避免一些表达量特别大的基因对剩余基因的影响，一定程度上避免差异表达分析的假阳性。

TMM^[28] (trimmed mean of M-values) 略微复杂。对每个样本，把其他样本作为比较样本，计算一个系数。去掉样本中表达值的很大的极端值，每个比较样本与本样本分别计算对数比例，加权求和得到每个样本各自的归一化因子。可

以看出如果基因没有差异表达，那么系数应该接近 1，事实上大多数基因确实是
没有差异表达的，因此 TMM 可以用于校正技术因素导致的系数偏离 1 的情况。

RLE^[29] (relative log expression) 对每个基因，首先计算每个基因的样本间
几何平均，对每个样本，计算每个基因的 reads 数与对应的几何平均的比值，取
每个样本所有比值的中位数作为 RLE 系数。RLE 也假设了大多数基因是没有差
异表达的，因此 RLE 也接近 1，偏离 1 的 RLE 系数可以被用来做样本库大小的
归一化。

RLE 图可以用于显示样本库大小归一化前后的变化。RLE 图和 RLE 归一化
思想是一样的，对于每个样本，如果大多数基因没有差异表达，上述的 RLE 系
数，即每个样本的 reads 数与基因跨样本几何平均的比值的中位数应该接近 0。
图 2.10 展示了使用 RLE 作为归一化方法的效果：从中可以看到在归一化之前，
数据 GSE123972 的 RLE 系数明显大于 0，不同的数据集的测序深度很不一样，
经过计算可以得到所有数据的 RLE 系数相对于 0 的偏离为 4.25，归一化之后不
同数据的异质性明显减小，而且 variance 分数明显降低为 0.45，体现出 RLE 具
有较好的归一化效果。

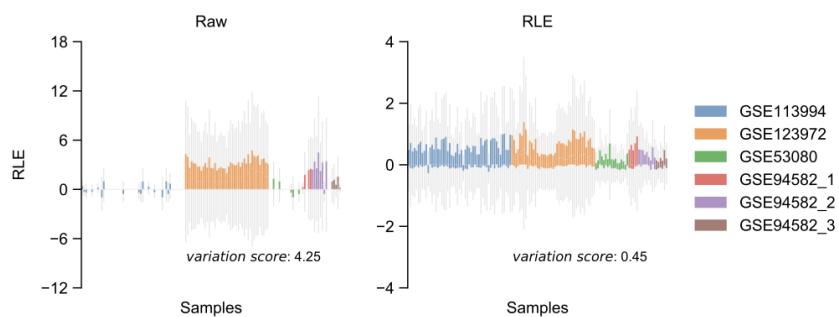


图 2.10 样本库大小归一化效果

2.3.3 去除批次效应

批次效应指人为引入的技术差异 (technical variance) 使得样本不按照真正
的类别信息聚类，而是向人为造成的批次偏移，如不同的实验日期，不同的建
库大小，胶选择长度等都会造成批次效应。我们依据数据来源文献中所指明的
批次效应，采用了 RUV^[30], ComBat^[17], Limma^[31] 等方法进行批次效应的去除。
并且使用 m-K 最近邻 (m-K-nearest neighbor, mKNN) 指标来衡量批次效应的去除
效果。

limma 使用线性回归对每个样本的每个基因的计数值进行校正，并不要求计数必须为整数值，即可以输入经过归一化的数值。事实上 **Limma** 要求输入的数据分布更加紧凑，所以可以取对数作为输入。**Limma** 假设基因的表达值符合固定因素和随机因素的加性效应，固定效应包括批次效应和其他技术因素引入的差异。因此 **Limma** 可以建立一个线性模型，以批次效应的信息作为自变量，以基因表达值作为因变量进行回归，再用基因表达值减去固定效应值，得到校正表达值。所以 **Limma** 要求一定要有批次信息作为输入。

ComBat 生成实现了比 **Limma** 更加适用于小数据的批次效应去除方法。其原理在于线性回归的参数估计可能会受到极端值影响。**ComBat** 使用一种经验贝叶斯的方法，对批次效应值实现层次的估计，即为了估计批次效应值服从的先验分布，而该先验分布的参数也服从另一个分布，由一组超参数控制，经验贝叶斯使用频率学派的估计方法估计超参数，再用贝叶斯方法估计先验分布，对于小样本数据的噪音具有更强的鲁棒性。

RUV 适用于批次效应信息不可获得或者不希望按照已有的批次信息进行完全去除的情况（如批次信息与样本类别信息重合过大）。**RUV** 可以先进行差异表达分析，去除差异表达的因素，即生物学差异后，再进行因子分析来寻找潜在的批次效应，用户可以设置一个 k 值，作为 **RUV** 进行线性回归时的隐变量个数控制，越大的 k 代表 **RUV** 的潜在批次越复杂。

对于批次信息存在且与样本类别信息差异不大的情况，可以使用以上三种批次效应去除方法，以及不使用批次效应去除方法作为对比，否则可以使用 **RUV** 对不输入批次效应信息的数据进行批次效应去除。

批次效应的去除可以通过降维可视化的方式鉴别，我们将在图 2.12 展示，直观上来看，同一批次的样本越分散，也就是不同批次的样本混合越均匀，则批次效应的影响越小，去除批次效应的效果越好。另一种展示方法通过方差分解 (variance decomposition analysis)。将表达值作为因变量，可以将各种因素分别作为自变量回归，如样本信息，批次信息，样本库大小等。对于每个基因作为因变量，都可以计算出不同的自变量解释的方差与总方差的比值，对于每一种自变量，都可以画出来其对每个基因的方差贡献比的分布曲线。图 2.11 所示，以批次效应信息作为自变量时，去除批次效应后，方差贡献分布曲线明显左移（蓝线到红线），说明批次效应对基因表达值的影响变小。而以样本类别信息作为自变量时，虽然方差贡献分布曲线也有左移，但是左移相对于批次效应的曲线明显较小，可以认为我们在去除批次效应的同时较好地保留了不同类别样本的差异性，

即生物学效应带来的组间差异。

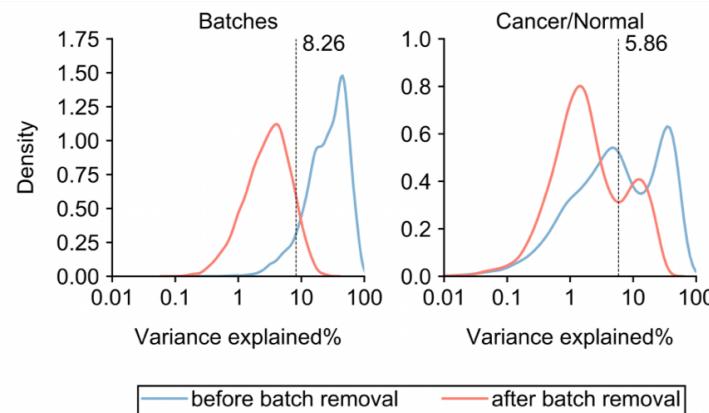


图 2.11 去除批次效应效果

2.3.4 评估表达矩阵处理效果

非监督聚类准确性 (unsupervised clustering accuracy, UCA)^[32], 由一篇单细胞领域研究中首次适用, 用于衡量聚类的效果, 作为 PCA 和 t-SNE 等降维可视化分析的量化评估, 可以使我们预先判断当前处理效果下数据的大致分类效果。UCA 首先只用 K-means^[33] 或者 KNN^[34] 作为聚类的方法, 由于是非监督聚类, 还需要使用线性分配的算法如匈牙利算法^[35] 将聚类标签与真实标签进行匹配。此时该问题可以看做一个有标签的分类问题, 计算出分类的准确率即为非监督聚类准确性。对于多种不同的聚类度量指标, 我们发现 UCA 和有监督的分类度量指标 AUC 具有最强的相关性, 因此采用非监督聚类准确性指标作为矩阵处理效果的度量之一。

为了更好地衡量批次效应的严重性以及批次效应去除效果, 我们参考了过去的指标如对齐分数 (alignment score)^[36], 和 kBET^[37] 两个研究。对齐分数虽然计算简单, 但是只适用于二分类标签, kBET 使用时会遇到受到随机效应影响而导致极端值附近的结果不够稳定的情况。为此我们提出了 m-K 最近邻 (m-K-nearest neighbor, mKNN) 指标, 公式如式 (2-1) 所示, 其中 \bar{x}_b, k, N_b, N, B 分别表示: 批次为 b 的样本周围同类批次的样本平均个数, 自主定义的最近邻取样数, 批次为 b 的样本总数, 所有样本的总数以及批次的种类数。利用 mKNN 指标, 我们可以逐批次逐样本地衡量其批次效应的严重性, 批次效应越严重, 则分子中的 \bar{x}_b 越大, 相应的整个指标越大。为了更好地表示“去除批次效应的效果”,

我们使用 1-mKNN 作为去除批次效应效果的指标。

$$\frac{1}{B} \sum_{b=1}^B \frac{\bar{x}_b - kN_b/(N-1)}{\min(k, N_b) - kN_b/(N-1)} \quad (2-1)$$

UCA 和 mKNN 指标均为 0 ~ 1，数值越大代表聚类效果越好以及批次效应的影响越小。因此我们统一地考虑两个指标，将矩阵处理方法的组合（即测序深度归一化以及去除批次效应方法的组合）加以统一的评估。图 2.12 中每一个点表示一种矩阵处理方法的组合。越靠近右上角的指标代表矩阵处理效果越好。如图所示，我们可以选择 RLE 作为测序深度归一化加上 Limma 作为去除批次效应的方法组合。对于该方法组合，我们还是用了 PCA 对矩阵处理前后的数据进行了可视化，PCA 可以用于降维，选择方差贡献最大的两个主成分表示到二维平面上，点的颜色表示批次，性状表示类别信息。不管是视觉上观察可以看到不同批次的点更好的混在了一起，还是从两个指标的变化（UCA : 0.519 → 0.692; mKNN : 0.055 → 0.792）上，均可以得出矩阵处理方法较好地完成了测序深度归一化以及批次效应去除的任务的结论。

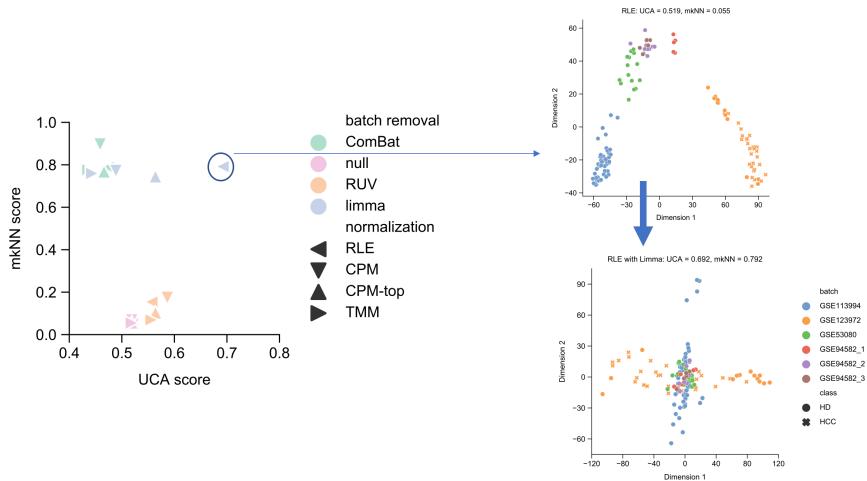


图 2.12 综合使用 UCA 和 mKNN 评估矩阵处理效果

2.4 特征选择和模型评估

由矩阵处理流程处理好的表达矩阵消除了部分技术性差异 (technical variance) 导致的样本库大小不一致和批次效应问题。我们可以对处理过的表达矩阵

应用一些统计学习模型，设计一套特征选择流程，选取对于癌症和正常样本分类效果好的，稳定的，具有生物学意义的特征，作为癌症检测的潜在的生物标志物。

2.4.1 差异表达分析

差异表达分析可以使用统计模型逐基因地寻找在癌症和正常人之间差异表达的特征。因为也可以作为特征选择的基本方法之一。

我们使用 DESeq2 包进行差异表达分析，值得注意的是 DESeq2 专门要求输入的矩阵必须是基因计数矩阵，也就是未经标准化和归一化的表达矩阵 (*un-normalized counts*)^[27]。因为这样的矩阵具有最多的信息和准确性，DESeq2 在内部会自动完成归一化的工作。

DESeq2 为了比较两组样本之间的计数差异建立了一个模型。该模型包含以下参数：(1) 归一化参数，至少可以归一化库大小的差异；(2) 方差参数，也被成为分散系数 (*dispersion*)；(3) 表示组间差异的参数。DESeq2 使用与原始 DESeq 相同的方法拟合 (1)。拟合 (2) 分两步：首先找到使似然 (*likelihood*) 最大的参数值，即完成最大似然估计。查找所有的基因表达值，并将这些值向中间值移动，移动的量由贝叶斯模型给定：如果基因的信息较低，则值更多地移动到中间，如果基因的信息很大，则值移动很少。使用与 (2) 中使用的相同的技术拟合 (3)。(3) 的值是最终需要得到的输出，找到一组组间差异高于某个数值的基因集合。这个阈值一般由错误发现率 (*False Discovery Rate, FDR*) 规定，一般取从小到大排最小的十个 FDR 基因作为差异表达基因。

对于差异表达基因的选取，这里我们使用一个改进的指标以替代 FDR，该指标如公式 2-2 所示，可以综合性考虑 FDR 和 fold change，更加完善。

$$\pi = \|\log_2 FC\| \cdot (-\log_{10} \text{FDR}) \quad (2-2)$$

使用 DESeq2 对原始的基因计数矩阵进行分析，可以得到如图 2.13 和图 2.14 的结果。

对于 2.13，我们使用 S-cfRNA-seq 数据，癌症类型分别为肝癌和早期肝癌。将其分别与正常样本比较，可以分别找出可以区分正常样本与肝癌以及正常样本与早期肝癌的差异基因。(A) 表示每个基因的对数 fold change 和对数调整后 p 值的火山图，图中红色的点为差异表达的基因，越靠近右上和左上的基因其差

异表达越明显; (B) 前十个差异表达基因的 fold change, 表达值以及对数调整后 p 值; (C) 挑选出的差异表达基因的热图以显示其对于癌症和正常样本的分类效果。

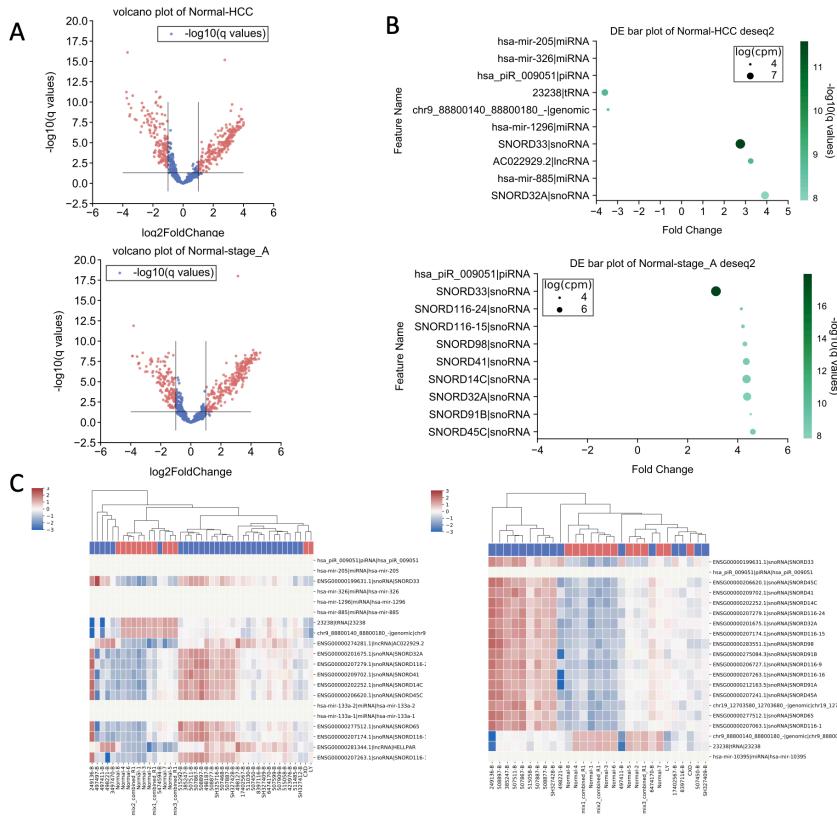


图 2.13 S-cfRNA-seq 数据的差异表达分析

对于图 2.14，我们使用 S-exoRNA-seq 数据，癌症类型分别为结肠癌和早期结肠癌。图的具体信息与图 2.13表示的类似。

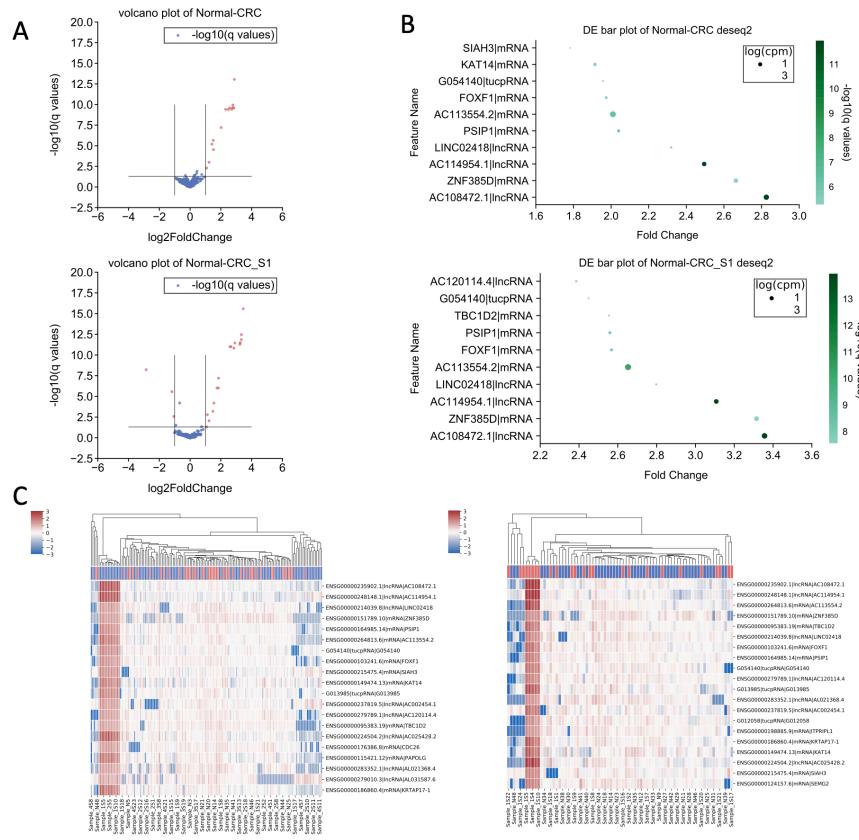


图 2.14 S-exoRNA-seq 数据的差异表达分析

差异表达分析的模型一般会独立考虑每个特征（即基因）的贡献，并不能结合性地讨论特征对分类的共同贡献，因此有局限性。接下来我们将会使用一些经典的机器学习模型来更好地组合不同的特征，以选出一组生物学上更有解释性和代表性的基因作为可能的生物标志物。同时考虑到模型必须具有的泛化能力以及稳定性，我们还会针对性地设计一个特征选择的框架来更好的挑选特征。

2.4.2 特征选择策略与机器学习模型

特征选择算法 特征选择方法通常使用一些经典的机器学习模型用于将不同的样本类别进行区分（如区分肝癌和正常样本）。通过优化机器学习模型使其更好地区分组间差异，获得更好的分类效果的同时，我们往往可以分析模型的参数，得到每个特征对分类效果的贡献，即权重。通过寻找权重大的特征，我们可以找到对模型分类贡献最大的基因，也就是我们希望寻找的潜在的生物标志物。与差异表达的逐个基因判断其贡献相比，机器学习模型的建模假设可以更复杂，可以综合考虑基因间的关联性，甚至建立非线性的模型刻画这种关系，因此对

数据有更好的建模效果，往往也可以获得更好的分类效果。然而由于模型的参数更复杂，对于数据的表示能力更强，往往也会带来严重的过拟合问题^[38]。对于 exRNA-seq 数据的标志物发掘，我们还需要考虑实际的生物学意义，寻找稳定的，解释性强的特征，这要求我们更好地固定机器学习模型，减少特征选择算法的随机性。针对这些考虑，我们设计了如下的特征选择算法：

- Scale each feature independently using robust normalization;
- **for** feature_num $k \in [1, 10]$ **do**;
- Random select 90% samples for 50 times;
- Run a classifier (Random Forest, Logistic Regression or Linear SVM) to select features based on feature importance.
- **for** each classifier \in [Random Forest, Logistic Regression or Linear SVM] **do**
- Optimize classifier's hyper-parameters by 5-fold cross-validation;
- **end for**
- Select top k features that are recurrently selected across resampling runs;
- Calculate AUC mean;
- **end for**
- Rank processing method by AUC, select the best k ;

首先，不同于 DESeq2 要求输入未经归一化的表达值，特征选择算法可以输入经过了归一化和去除批次效应的表达值。我们之前只对样本进行了库大小的归一化，一般而言还要对特征进行逐个标准化。公式 2-3 表示 robust normalization 的具体做法，对于每个特征，通过减去其中位数，除以四分间距 (IQR: Interquartile Range)，即可得到标准化后的表达值，使用四分间距作为分母而不是标准差，可以有效避免极端值对标准化的影响，具有更强的鲁棒性。

$$\text{robust_scale} \left(x'_{ij} \right) = \frac{x_{ij} - \text{median}_k x_{ik}}{Q_{0.75}(\mathbf{x}_i) - Q_{0.25}(\mathbf{x}_i)} \quad (2-3)$$

接着，由于我们对于最终选择的特征的数量尚未确定，我们首先需要测试挑选的特征的数量，即 $k \in [1, 10], 20, 30, 40, 50$ 。为了挑选出稳健的特征，我们对每个 k 都要进行 50 轮的测试，挑选在测试中反复出现的特征。由于样本量本身较小，我们没有采用交叉验证 (cross validation) 的方法，而是直接将全部数据集在每次测试中划分为 90% 的训练集和 10% 的测试集。在每轮测试中，训练集和测试集的样本都有随机不放回抽样决定，保证训练集中的各类样本的比

例与总体类似。我们会使用多种机器学习模型，如逻辑斯谛回归^[39]，支持向量机^[40]，随机森林^[41]等。在每轮测试，我们都会进行五折的交叉验证 (5-fold cross validation) 以优化机器学习模型内部的超参数，这些超参数不能由模型自己优化得到，而必须通过人为的选择，比较不同超参数设定时的结果来得到，比如逻辑斯谛回归中的 C 来控制正则化 (regularization) 的大小的倒数， C 越小代表越强的正则化，随机森林中有树的个数和最大深度作为超参数。最终我们选取 50 轮测试中重复出现的特征，重新使用机器学习模型选择权重贡献最大的 k 个特征作为最终选择的特征。在 50 轮的测试中我们可以计算出 50 轮的平均分类准确度指标：AUC 的平均值。之后我们可以选择最佳的 k ，在尽可能保证较高的 AUC 的情况下减少 k 的值。

机器学习模型 在特征选择框架中我们可以选择多种机器学习模型，包括逻辑斯谛回归，支持向量机，随机森林等。以逻辑斯谛回归为例，对于二分类问题，逻辑斯谛回归的决策模型如公式 2-4 所示，其中 x 为表达矩阵， β 为逻辑斯谛回归中线性模型的权重向量， β_0 为偏置项。公式 2-4 表示输入为 x 时模型判断样本标签为 1 的概率。若设患癌症样本的标签为 1，则代表该样本为癌症患者的概率。

$$P(Y = 1|X = x) = \frac{\exp(\beta_0 + x^T \beta)}{1 + \exp(\beta_0 + x^T \beta)} \quad (2-4)$$

类似地可以得到多分类的逻辑斯谛回归如公式 2-5。当逻辑斯蒂回归模型获得数据输入，进行了训练后，为了最大化分类正确的概率，模型会内部优化出 β_0, β 的最优值，通过寻找 $|\beta|$ 最大的那些值，即可找到我们希望寻找的，对于分类最有效的特征。

$$P(Y = k|X = x) = \frac{\exp(\beta_{k,0} + x^T \beta_k)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l,0} + x^T \beta_l)}, \text{ for } k = 1, \dots, K-1 \quad (2-5a)$$

$$P(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l,0} + x^T \beta_l)}, \text{ for } k = K \quad (2-5b)$$

2.4.3 特征选择的分类效果和稳健性评估

分类效果衡量指标 AUC 为了衡量模型的分类效果，我们使用了曲线下面积 (Area Under Curve, AUC) 来衡量其分类效果，AUC 是一个 0-1 之间的数字，对于二分类模型，完全随机分类的 AUC 为 0.5，完全区分两类样本时的 AUC 为 1，AUC 越大代表分类效果越好。

对于样本的预测，我们使用医学和生物领域常用的混淆矩阵来计算其相应的指标，并进一步获得 AUC 的值。首先我们使用如表 2.2 所示的混淆矩阵来计算真阳性 (True Positive Rate, TPR) 和假阳性 (False Positive Rate, FPR)，其中真阳性率也被称为灵敏度 (Sensitivity)，假阳性率与特异度 (Specificity) 的关系为 $1 - FPR = Specificity$ 。由于机器学习模型可以预测出每个样本分别属于两类别 (癌症和正常人) 的概率，因此对每个样本的二分类预测概率，选取一系列的截断值 (cutoff)，就可以用于构建多个混淆矩阵，并计算出每一个混淆矩阵的 TPR 和 FPR 值。将同一模型每个阈值的 (FPR, TPR) 座标都画在 ROC 空间里，就可以生成特定模型的 ROC 曲线^[42]。ROC 曲线在完全随机预测时就是 $y = x, x \in [0, 1]$ 的直线，ROC 曲线下围的面积就是 AUC，AUC 越接近 1 代表分类效果越好。

表 2.2 混淆矩阵

Predict/True	Positive	Negative	Metric
Positive	True Positive (TP)	False Positive (FP)	$PPV = \frac{TP}{TP+FP}$
Negative	False Negative (FN)	True Negative (TN)	$NPV = \frac{TN}{TN+FN}$
Metric	$Sensitivity = \frac{TP}{TP+FN}$	$Specificity = \frac{TN}{TN+FP}$	

特征稳定性衡量指标 KI 挑选出重复性强的稳定的特征在液体活检领域非常重要，在不同测试轮数中重复出现的特征不仅对分类很重要，而且可能具有泛化能力强的特点。为了衡量挑选的特征的稳定性，我们使用了 Kuncheva index (KI) 来衡量特征的稳定性，其公式如式 (2-6)，其中 f_i, f_j 为两次挑选的特征， d, N 分别为一次挑选的特征数和总特征数量。

$$KI(f_i, f_j) = \frac{|f_i \cap f_j|N - d^2}{dN - d^2} = \frac{|f_i \cap f_j| - d^2/N}{d - d^2/N} \quad (2-6)$$

若做 K 次采样，均采取不放回的策略时，可以获得平均 KI 的公式如式 (2-7)，

即 AKI。

$$AKI = \frac{2}{M(M-1)} \sum_{i=1}^M \sum_{j=1}^M KI(f_i, f_j) \quad (2-7)$$

KI 和 AKI 同样为 0-1 之间的值，其值越接近 1 则代表特征越稳定，AKI 为 1 时代表多轮测试中挑选出的权重最大的特征完全一致，这是最理想的情况。

2.4.4 模型分类效果比较总结

为了更好的比较 exSEEK 挑选出的生物标志物和已知的一些经典的生物标志物，我们总结了各类生物标志物和 exSEEK 挑选出的特征的 AUC 用以比较，我们首先注意到如图 2.15 所示，在挑选 10 个以下的特征时，特征的数量变化并不会造成 AUC 的剧烈变动，挑选 6-7 个甚至更少的特征也可以取得很好的分类效果。

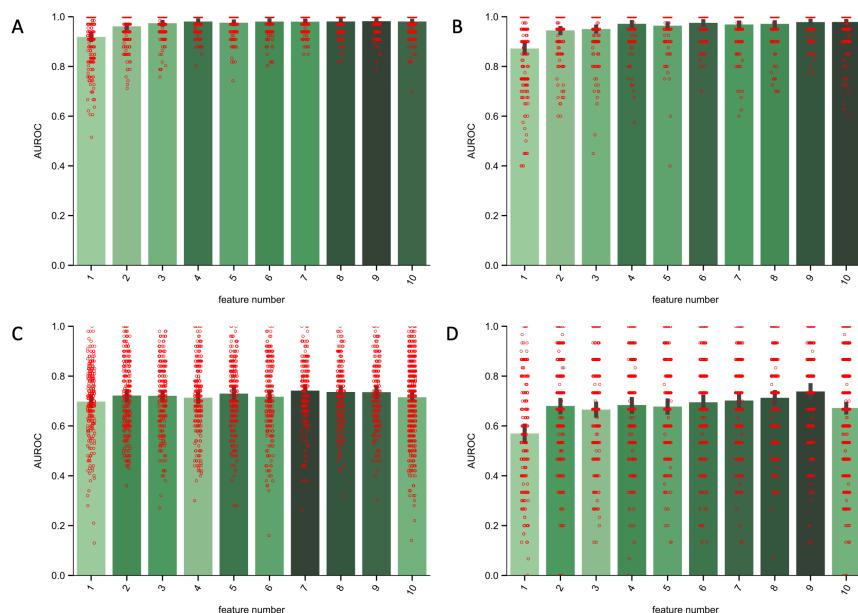


图 2.15 挑选不同数量 feature 时 AUC 的变化

图 2.16 汇总了表 ?? 的具体结果，以条形图加以展示。CRC/HD, CRC_S1/HD, HCC/HD, HCC_S1/HD 分别代表结肠癌和正常人，结肠癌早期和正常人，肝癌和正常人，肝癌早期和正常人的比较。下划线下标明 exSEEK 的标明使用 exSEEK

优化得到的参数进行分类。标志物名称的括号中的数字表示其使用的基因的数量。每个表格的前两个标志物 (miR-20, miRNA6; miR-21, miRNA7) 为已知的生物标志物，使用其原始文献中的参数进行测试。miRNA6, miRNA7 还通过 exSEEK 进行了重新优化，其结果得到了显著提升。ncRNA 和 miRNA_only 分别代表 exSEEK 发现的非编码 RNA 生物标志物组合以及仅包含 miRNA 的生物标志物组合。使用 exSEEK，我们比较了已知的生物标志物和由 exSEEK 优化或挑选出的生物标志物的分类结果。可以看到由 exSEEK 挑选的 ncRNA 取得了最好的分类效果，而经过 exSEEK 优化的已知的生物标志物比该标志物原始参数设定下的分类结果获得了一定的提升。

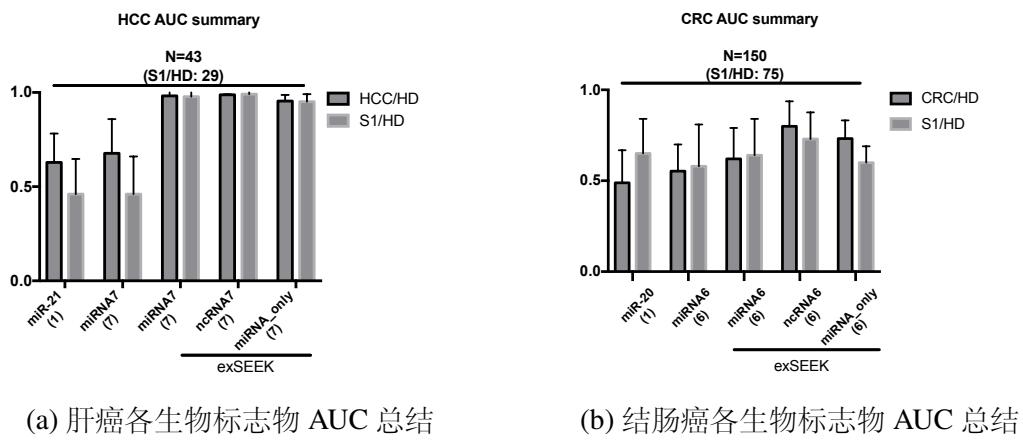


图 2.16 肝癌和结肠癌 AUC 总结

各类癌症在各个数据上的 AUC 汇总 最后，除了结肠癌和肝癌，我们也在前列腺癌 (PRAD) 以及胰腺癌 (PAAD) 上对癌症和正常人进行了分类。通过五十轮交叉验证，我们可以得到对每个样本的归属概率的多轮预测值的平均。对于模型预测值大于 0.5 的样本，我们认为其为癌症，小于 0.5 的样本被归类为正常人，可以得到各个数据集上相关癌症的每个样本的预测概率值箱线如图 2.17。

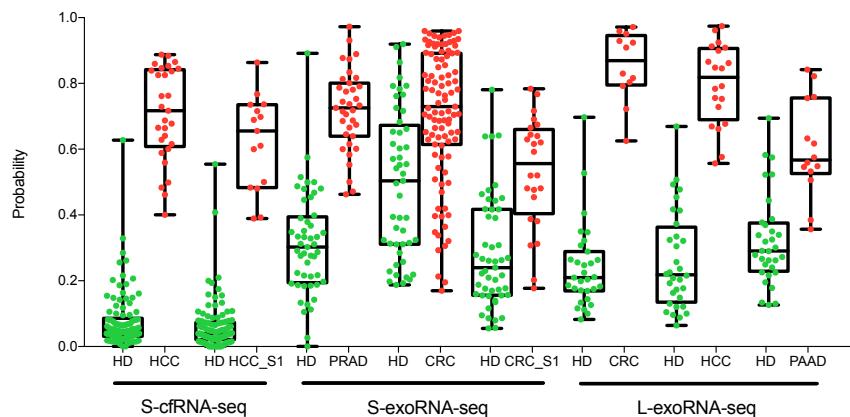


图 2.17 每个样本多轮交叉验证的预测概率箱线图

对于图 2.17 的结果，使用上述获得 ROC 曲线的方法，可以再计算出相应的曲线下面积 AUC，绘制所有类型的癌症各自的分类指标 AUC 如图 2.18。

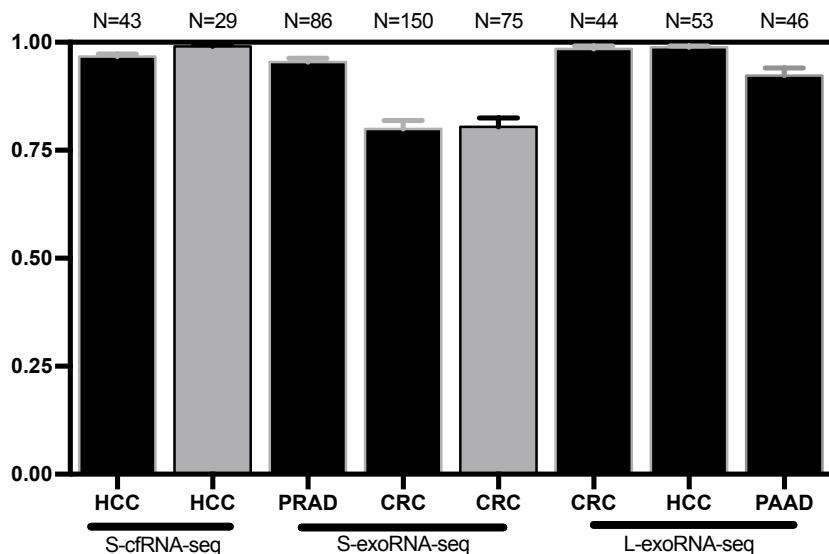


图 2.18 每个样本多轮交叉验证的 AUC 汇总

2.4.5 挑选出的候选生物标志物表现

我们使用 exSEEK 对各生物标志物进行分类测试，测试方法如前文特征选择算法所述，使用 50 轮随机抽样，获得对每个样本的二分类预测概率，选取一系列的截断值（cutoff）用于构建混淆矩阵，并计算出每一轮的一系列 TPR 和 FPR 值。为了更好地展示每种生物标志物的分类效果，我们绘制了如图 2.19 所示的 ROC 曲线，由于样本较少，相应的可用截断值也就较少，为了更好地绘制

ROC 曲线，我们还对曲线进行了线性插值，使其尽量光滑。ROC 曲线所包围的面积（曲线下面积，AUC）越大，则说明该模型/特征的分类效果越好。

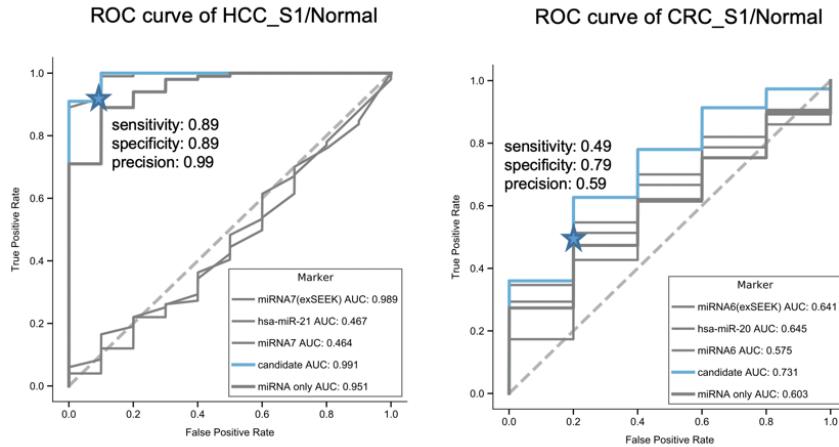


图 2.19 早期肝癌和结肠癌分类的 ROC 曲线

使用 exSEEK，我们比较了已知的生物标志物和由 exSEEK 优化或挑选出的生物标志物在早期癌症上分类结果的 ROC 曲线。

对于 HCC_S1/Normal，已知生物标志物 miRNA7 原始的 AUC 为 0.464，由 exSEEK 优化后可以达到 0.989。已知生物标志物 miR-21 的 AUC 为 0.467，exSEEK 只挑选 miRNA 作为生物标志物的 AUC 为 0.951。candidate 表示 exSEEK 挑选的 7 个非编码 RNA 生物标志物，其 AUC 为 0.991，灵敏度为 0.89，特异性为 0.89，精度为 0.99。

对于 CRC_S1/Normal，已知生物标志物 miRNA6 原始的 AUC 为 0.575，由 exSEEK 优化后可以达到 0.641。已知生物标志物 miR-20 的 AUC 为 0.645，exSEEK 只挑选 miRNA 作为生物标志物的 AUC 为 0.603。candidate 表示 exSEEK 挑选的 6 个非编码 RNA 生物标志物，其 AUC 为 0.731，灵敏度为 0.49，特异性为 0.79，精度为 0.59。

评估挑选的生物标志物 对于挑选出的特征，我们还可以做进一步做如图 2.20 的综合评估，包括使用热图（heatmap）衡量其分类效果，可以看到使用 exSEEK 选取的特征可以很好的区分正常和癌症样本，且选取的基因包含 lncRNA, snoRNA 等非编码 RNA。同时我们使用了右图来分析挑选出的基因的一些特性。如散点图的 x 轴的数值代表其在分类模型中的权重。如图所示，权重绝对值越大则说明

其对模型的分类贡献越大。同时点的大小代表该基因的表达值，以 CPM (counts per million) 表示。其颜色的深浅代表特征在五十轮测试中出现的频率，颜色越深，则频率越接近 1，证明该特征越稳定出现。可以看到 SNORA2C 既是对分类贡献最大的特征，也是相对比较稳定出现的特征。红色的条形图中的 x 轴则表示差异表达分析中每个特征的 fold change，颜色深浅代表其负对数 FDR 值。

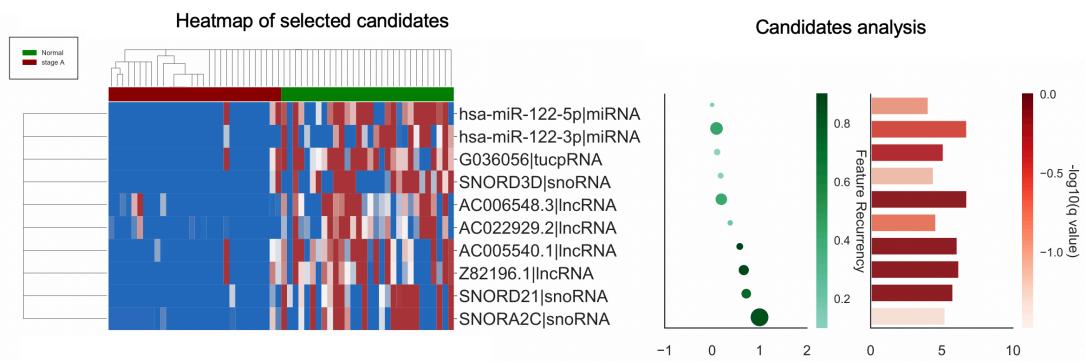


图 2.20 综合评估挑选的生物标志物

2.5 exSEEK 软件介绍及使用

2.5.1 软件基本功能模块与使用

基于上述的工作我们开发了 exSEEK 软件，我们使用 Snakemake 软件包^[43]对分析流程进行控制，这是因为处理过程比较复杂，输出文件互相依赖，处理耗时长且容易出现断点，使用 Snakemake 可以保证样本处理的管理有序。exSEEK 包含了三大模块，如图 2.21 所示，Utilities 和 Preprocess 模块包含了数据前处理，映射和构建表达矩阵等功能，可用于 L-exRNA-seq 和 S-exRNA-seq 数据的映射和表达矩阵构建。exSEEK 模块为核心模块，可以完成差异表达，矩阵处理以及特征挑选以及标志物评估等功能。exSEEK 软件读取一个可以由用户简洁地自定义的 config.json 文件用于设置各个模块的相关参数，运行各个模块的具体步骤时均只需要一行命令即可完成，方便各类用户使用。

2.5.2 软件绘图模块及使用

exSEEK 软件还包含有高度可交互以及高度标准化的绘图模块，如图 2.22 所示，我们使用 jupyter-notebook^[44] 进行绘图模块的构建，用户可以交互式地使用绘图模块完成高质量的结果可视化，只需要选择相应的参数即可自动地产生类

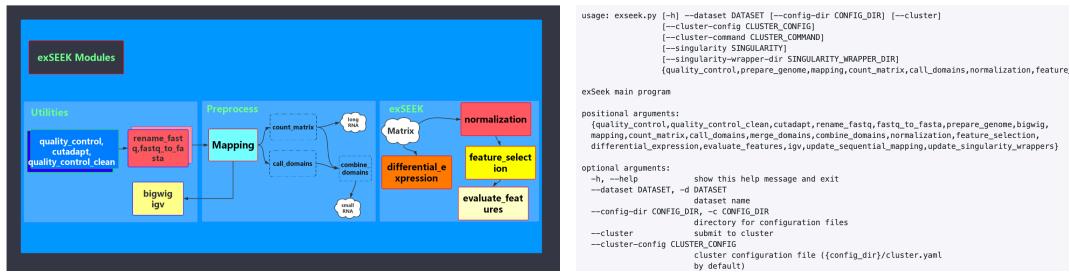


图 2.21 exSEEK 软件模块及使用命令

似本文配图的输出结果图。使用可交互式的绘图方式，使得我们可以尽可能地隐藏代码细节，尤其适合用户测试了多套数据，不同的测序方法，以及配合本工具的多种 matrix processing 以及 feature selection 方法的组合。我们针对 python 绘图比较粗糙的特点，专门配置了符合出版物绘图规范的绘图函数，使得 exSEEK 工具产生的图片都符合标准的绘图要求。

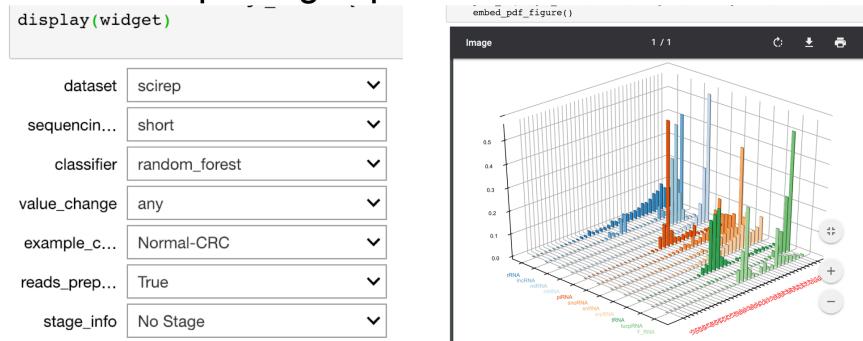


图 2.22 exSEEK 绘图模块

第3章 总结与讨论

3.1 结论

在本研究中，我们完成了 exRNA-seq 的完整的生物信息学分析流程，包括 L-exoRNA-seq 和 S-cfRNA-seq，S-exoRNA-seq 数据的比对（小 RNA-seq 数据采用顺序比对），构建基因表达矩阵（对于小 RNA-seq 数据采用结构域检测的方法获得结构域特征），使用一系列的矩阵处理方法完成样本库的归一化和批次效应的去除，并且使用 UCA 和 mKNN 两个分数衡量处理效果，最后使用一个特征选择框架完成特征的挑选以及在多套数据和癌症类型上进行分类效果的评估以及生物标志物的分析，我们挑选出了一些非编码 RNA 和 miRNA 的组合，可以取得比已知 miRNA 生物标志物更好的分类效果。最后我们将相关代码封装，制作成一个易于使用的工具 exSEEK，方便终端用户使用 exSEEK 完成 exRNA-seq 的全流程分析，并且提供可交互的，可以方便可视化的工具供用户分析和理解处理结果。

3.2 讨论

exRNA 作为生物标志物的生物学原理 细胞外 RNA (exRNA) 可能富集于外泌体 (exosomes)、微囊泡 (microvesicles, MVs)、非囊泡结构的核糖核酸蛋白复合体 (RNPs) 中，与组织细胞中的 RNA 差异较大，细胞外 RNA 已经被证明可能与癌症细胞相关，作用于癌症的发生，转移等过程，并且影响免疫系统。exRNA 由于被分泌到细胞外，且能够稳定存在的 exRNA 往往具有较为稳定的 RNA 二级结构或者与 RNA 结合蛋白一起形成复合物。这种生物学上的选择在原理上证明了 exRNA 作为生物标志物的潜力，通过我们的分析结果，可以发现 exRNA 数据在癌症的鉴定，尤其是早期癌症的检测上，具有较好的准确率，灵敏度以及特异性。

矩阵处理的改进 exRNA 数据存在明显的异质性和批次效应，目前我们使用一些 RNA-seq 和单细胞领域的处理方法，这些方法往往基于一些统计学的假设建模，并没有充分考虑到 exRNA 数据的特征，因此进行矩阵处理的方法还有待改

进。比如对表达矩阵进行样本库大小的归一化，由于 exRNA 数据的稀疏性和少数基因占据主导的特点，归一化方法可能不够稳定，即使使用外部加入的含量固定的 spikein RNA，也有可能因为实验技术的问题导致不同样本的 spikein 并不完全一致。另外我们还遇到了批次信息和样本类别信息完全重合的情况，如何对这样的数据进行批次效应的处理，以及更好地衡量批次效应的去除还值得探索，2018 年的一篇文章探讨了衡量批次效应去除的指标 kBET^[45]，我们也提出了 mKNN 分数来衡量批次效应去除效果，比 kBET 在小样本上更加稳定，但是一个公认的方便使用的批次效应衡量指标依然有待研究。

机器学习模型的小样本和过拟合 机器学习问题中，为了避免过拟合，测定模型的泛化性能，一般使用交叉验证的方式。在特征选择部分，针对 exRNA-seq 样本量少，特征维度高，容易过拟合的问题，我们使用了 50 轮随机抽样进行交叉验证，测试模型的泛化能力。另外，研究中使用的模型也都具有一定的避免过拟合的能力，如随机森林模型在处理测序数据时具有一定的避免过拟合的能力^[28]，而逻辑斯谛回归也可以通过加入 L_1 正则化等方式对特征挑选进行约束，一定程度上进行正则化，避免过拟合^[46]。在未来获得更大量样本的情况下，过拟合的问题可能会得到一定程度的缓解，机器学习模型的结果将会更加可靠。伴随小样本问题的另一问题是同一套数据内类别不均衡的问题^[47]，常规的处理方法一般采取对样本量少的类别升采样，或者对样本量多的类别降采样，由于本身数据量很小，降采样的方式并不可取，然而升采样时也必须注意根据数据的分布进行采样，复杂的升采样方法常采用生成模型，如自编码器（VAE），生成对抗网络（GAN）等，但是对于样本量本身比较少的数据也并不容易实现。因此如何均衡数据的类别也需要更仔细的思考。

机器学习的另一个限制是生成的模型通常不易解释。即使机器学习算法非常有效，我们通常也无法理解其中的算法结构与基础生物学之间的对应关系。通过机器学习算法得到的具有区分不同表型的生物标志物不一定与疾病的生物学发生或发展有着显着的直接关系。例如，生物标志物可以由感兴趣的疾病过程下游的免疫应答产生。由于我们难以破译驱动机器学习算法的具体机制，因此在尝试将算法应用于与训练对象非常不同的新群组时应更加谨慎，例如从动物模型转换为人类临床样本。

模型的可解释性以及与其他数据结合的可能性 机器学习模型的一个问题和局限是其本身的黑箱特性，即内部的原理难以解释，不容易与实际问题的原理对应。我们挑选出的特征可能难以直接与癌症的生物学机理对应，挑选出的特征的组合也难以被诠释出具体的意义。为了解决这个问题，研究人员从不同角度对模型进行修改，如使用变分自编码器的中间层隐含变量对生物学机制进行诠释，构建基于图的模型对调控网络进行建模等，以增强模型的可解释性。另一方面，研究人员也在探索组合使用多种类型数据进行癌症的预测，如结合甲基化，RNA 编辑和剪接数据以及图像数据等^{[48][49][50]}，对于不同数据的组合，一般可以在建模的不同阶段将其融合，如在数据准备阶段将不同的数据拼接成一个高维张量（tensor），或者对不同类型的数据建立不同的模型，在预测阶段将模型融合等。利用更多来源的数据可以互相补充，加入更多的信息，有助于提高预测的准确率和灵敏性，特异性等。

总之我们认为，在未来随着 exRNA-seq 测序技术的不断发展和样本量的积累，以及 exRNA-seq 分析技术以及特征选择算法的发展，人们将可以挑选出表现更好，泛化能力更强，稳定性更好而且具有更强解释力的基因作为生物标志物，并且广泛应用于癌症尤其是早期癌症的检测与预后治疗中。

插图索引

图 1.1 exSEEK 流程图	5
图 2.1 reads 处理流程图	8
图 2.2 L-exoRNA-seq 数据 mapping 基本情况统计。	9
图 2.3 S-cfRNA-seq 数据 mapping 基本情况统计	11
图 2.4 RNA 长度分布	12
图 2.5 不同 exRNA 和组织数据结构域分布和大小分布	13
图 2.6 小 exRNA-seq 结构域和 miRNA 特征的丰度与多样性	14
图 2.7 小 exRNA-seq 差异表达 RNA 比例	14
图 2.8 表达矩阵处理流程图	15
图 2.9 scImpute 原理	15
图 2.10 样本库大小归一化效果	17
图 2.11 去除批次效应效果	19
图 2.12 综合使用 UCA 和 mKNN 评估矩阵处理效果	20
图 2.13 S-cfRNA-seq 数据的差异表达分析	22
图 2.14 S-exoRNA-seq 数据的差异表达分析	23
图 2.15 挑选不同数量 feature 时 AUC 的变化	27
图 2.16 肝癌和结肠癌 AUC 总结	28
图 2.17 每个样本多轮交叉验证的预测概率箱线图	29
图 2.18 每个样本多轮交叉验证的 AUC 汇总	29
图 2.19 早期肝癌和结肠癌分类的 ROC 曲线	30
图 2.20 综合评估挑选的生物标志物	31
图 2.21 exSEEK 软件模块及使用命令	32
图 2.22 exSEEK 绘图模块	32

表格索引

表 2.1	exRNA 数据收集总结	7
表 2.2	混淆矩阵	26

公式索引

公式 2-1	20
公式 2-2	21
公式 2-3	24
公式 2-4	25
公式 2-5a	25
公式 2-5b	25
公式 2-6	26
公式 2-7	27

参考文献

- [1] wikipidea. what is cancer[EB/OL]. 2019. <https://www.cancer.gov/about-cancer/understanding/what-is-cancer?redirect=true>.
- [2] wikipidea. cancer[EB/OL]. 2019. <https://www.who.int/en/news-room/fact-sheets/detail/cancer>.
- [3] Aravanis A M, DeBusschere B D, Chruscinski A J, et al. A genetically engineered cell-based biosensor for functional classification of agents[J]. Biosensors and Bioelectronics, 2001, 16(7-8): 571–577.
- [4] Cohen J D, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test[J]. Science, 2018, 359(6378): 926–930.
- [5] Xi X, Li T, Huang Y, et al. Rna biomarkers: frontier of precision medicine for cancer[J]. Non-coding RNA, 2017, 3(1): 9.
- [6] Schwarzenbach H, Hoon D S, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients [J]. Nature Reviews Cancer, 2011, 11(6): 426.
- [7] Wei Z, Batagov A O, Schinelli S, et al. Coding and noncoding landscape of extracellular rna released by human glioma stem cells[J]. Nature communications, 2017, 8(1): 1145.
- [8] Ngo TT, Moufarrej MN, Rasmussen ML H, et al. Noninvasive blood tests for fetal development predict gestational age and preterm delivery[J]. Science, 2018, 360(6393): 1133–1136.
- [9] Zhou J, Yu L, Gao X, et al. Plasma microrna panel to diagnose hepatitis b virus-related hepatocellular carcinoma[J]. J Clin Oncol, 2011, 29(36): 4781–4788.
- [10] Kaczor-Urbanowicz K E, Kim Y, Li F, et al. Novel approaches for bioinformatic analysis of salivary rna sequencing data for development[J]. Bioinformatics, 2017, 34(1): 1–8.
- [11] Allen R M, Zhao S, Ramirez Solano M A, et al. Bioinformatic analysis of endogenous and exogenous small rnas on lipoproteins[J]. Journal of extracellular vesicles, 2018, 7(1): 1506198.
- [12] Vitsios D M, Enright A J. Chimira: analysis of small rna sequencing data and microrna modifications[J]. Bioinformatics, 2015, 31(20): 3365–3367.
- [13] Baras A S, Mitchell C J, Myers J R, et al. mirge-a multiplexed method of processing small rna-seq data to determine microrna entropy[J]. PloS one, 2015, 10(11): e0143066.
- [14] Capece V, Garcia Vizcaino J C, Vidal R, et al. Oasis: online analysis of small rna deep sequencing data[J]. Bioinformatics, 2015, 31(13): 2205–2207.
- [15] Li W V, Li J J. An accurate and robust imputation method scimpute for single-cell rna-seq data[J]. Nature communications, 2018, 9(1): 997.

- [16] Bacher R, Chu L F, Leng N, et al. Scnorm: robust normalization of single-cell rna-seq data [J]. *Nature methods*, 2017, 14(6): 584.
- [17] Chen C, Grennan K, Badner J, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods[J]. *PloS one*, 2011, 6(2): e17238.
- [18] wikipedia. Feature selection[EB/OL]. 2019. https://en.wikipedia.org/wiki/Feature_selection.
- [19] Liu S, Dissanayake S, Patel S, et al. Learning accurate and interpretable models based on regularized random forests regression[J]. *BMC systems biology*, 2014, 8(3): S5.
- [20] Yuan T, Huang X, Woodcock M, et al. Plasma extracellular rna profiles in healthy and cancer patients[J]. *Scientific reports*, 2016, 6: 19413.
- [21] Tan C, Cao J, Chen L, et al. Noncoding rnas serve as diagnosis and prognosis biomarkers for hepatocellular carcinoma[J]. *Clinical chemistry*, 2019: clinchem–2018.
- [22] Max K E, Bertram K, Akat K M, et al. Human plasma and serum extracellular small rna reference profiles and their clinical utility[J]. *Proceedings of the National Academy of Sciences*, 2018, 115(23): E5334–E5343.
- [23] Giraldez M D, Spengler R M, Etheridge A, et al. Comprehensive multi-center assessment of small rna-seq methods for quantitative mirna profiling[J]. *Nature biotechnology*, 2018.
- [24] Akat K M, Moore-McGriff D, Morozov P, et al. Comparative rna-sequencing analysis of myocardial and circulating small rnas in human heart failure and their utility as biomarkers[J]. *Proceedings of the National Academy of Sciences*, 2014, 111(30): 11151–11156.
- [25] Li S, Li Y, Chen B, et al. exorbase: a database of circrna, lncrna and mrna in human blood exosomes[J]. *Nucleic acids research*, 2017, 46(D1): D106–D112.
- [26] Hall A E, Turnbull C, Dalmay T. Y rnas: recent developments[J]. *Biomolecular concepts*, 2013, 4(2): 103–110.
- [27] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2[J]. *Genome biology*, 2014, 15(12): 550.
- [28] Díaz-Uriarte R, De Andres S A. Gene selection and classification of microarray data using random forest[J]. *BMC bioinformatics*, 2006, 7(1): 3.
- [29] Anders S, Reyes A, Huber W. Detecting differential usage of exons from rna-seq data[J]. *Genome research*, 2012, 22(10): 2008–2017.
- [30] Risso D, Ngai J, Speed T P, et al. Normalization of rna-seq data using factor analysis of control genes or samples[J]. *Nature biotechnology*, 2014, 32(9): 896.
- [31] Ritchie M E, Phipson B, Wu D, et al. limma powers differential expression analyses for rna-sequencing and microarray studies[J]. *Nucleic acids research*, 2015, 43(7): e47–e47.
- [32] Lopez R, Regier J, Cole M B, et al. Deep generative modeling for single-cell transcriptomics [J]. *Nature methods*, 2018, 15(12): 1053.
- [33] Jain A K, Dubes R C, et al. Algorithms for clustering data: volume 6[M]. [S.l.]: Prentice hall Englewood Cliffs, 1988

- [34] Cover T M, Hart P E, et al. Nearest neighbor pattern classification[J]. IEEE transactions on information theory, 1967, 13(1): 21–27.
- [35] wikipidea. Assignment problem[EB/OL]. 2019. https://en.wikipedia.org/wiki/Assignment_problem.
- [36] Butler A, Hoffman P, Smibert P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species[J]. Nature biotechnology, 2018, 36(5): 411.
- [37] Buttner M, Miao Z, Wolf A, et al. Assessment of batch-correction methods for scRNA-seq data with a new test metric[J]. BioRxiv, 2017: 200345.
- [38] Domingos P M. A few useful things to know about machine learning.[J]. Commun. acm, 2012, 55(10): 78–87.
- [39] Kleinbaum D G, Dietz K, Gail M, et al. Logistic regression[M]. [S.l.]: Springer, 2002
- [40] Scholkopf B, Smola A J. Learning with kernels: support vector machines, regularization, optimization, and beyond[M]. [S.l.]: MIT press, 2001
- [41] Liaw A, Wiener M, et al. Classification and regression by randomforest[J]. R news, 2002, 2 (3): 18–22.
- [42] wikipidea. Roc curve[EB/OL]. 2019. https://en.wikipedia.org/wiki/Receiver_operating_characteristic.
- [43] Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine[J]. Bioinformatics, 2012, 28(19): 2520–2522.
- [44] Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter notebooks-a publishing format for reproducible computational workflows.[C]//ELPUB. [S.l.: s.n.], 2016: 87–90.
- [45] Büttner M, Miao Z, Wolf F A, et al. A test metric for assessing single-cell RNA-seq batch correction[J]. Nature methods, 2019, 16(1): 43.
- [46] Ng A Y. Feature selection, l1 vs. l2 regularization, and rotational invariance[C]//Proceedings of the twenty-first international conference on Machine learning. [S.l.]: ACM, 2004: 78.
- [47] Chawla N V, Japkowicz N, Kotcz A. Special issue on learning from imbalanced data sets[J]. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 1–6.
- [48] Cappelli E, Felici G, Weitschek E. Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction[J]. BioData mining, 2018, 11(1): 22.
- [49] Cheng J, Zhang J, Han Y, et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis[J]. Cancer research, 2017, 77(21): e91–e100.
- [50] Wang Q, Armenia J, Zhang C, et al. Unifying cancer and normal RNA sequencing data from different sources[J]. Scientific data, 2018, 5: 180061.
- [51] 薛瑞尼. THUThESIS: 清华大学学位论文模板[EB/OL]. 2017. <https://github.com/xueruini/thuthesis>.

致 谢

感谢鲁志教授对我的悉心培养和指导，两年的时光短暂又充实，让我获益匪浅。感谢鲁志教授实验室的史斌斌，谢宇峰，金云帆对我的课题和帮助，我们共同在本课题投入了诸多的心血和精力，共同完成了这份美妙的研究。感谢实验室王思琦，邢少贞，徐港同学对我的课题给出的启发性的建议。我要专门再次感谢史斌斌同学，从你的身上我学到了让学业和人生变得不一样的可能，谢谢你教会了我这么多。我还要感谢实验室的所有同学，实验室的愉快生活给我的本科生活增加了无数的光彩。感谢 THUTHESIS^[51]，你们十几年来维护的清华毕业论文 LATEX 模板让我深切地体会了对毕业论文应有的敬畏之心，也帮我节约了大量的排版时间，得以专注论文本身。

感谢清华学堂生命科学实验班项目为我营造了优越的科研环境，感谢清华大学生物科学项目四年对我的培养。最后，感谢我的父母和家人对我科研和生活的支持，感谢我的女朋友孟舒晨陪伴我度过了美好的本科生活，你是我的生命之光。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名: _____ 日 期: _____

附录 A 外文资料的调研阅读报告

Literature review for the project "Developing bioinformatics methods on exRNA-seq analysis for cancer"

A.1 Research significance and scientific basis of the project

Compared to DNA and protein biomarkers, RNA has higher tissue specificity, sensitivity, and lower testing costs. These advantages of RNA have made it promising to serve as an alternative, or even more advantageous biomarker in cancer diagnosis^{[1],[2]}. An RNA molecule that can be secreted extracellularly is called exRNA (extracellular RNA) and is a major component of RNA that can be detected in body fluids. Currently, large-scale international research teams (such as the Extracellular RNA Communication Consortium)^[3] and commercial organizations (such as the Gates Foundation, etc to invest 2 billion company GRAIL) has exRNA as the object of focus on exploration, it can be used as desired biomarkers in body fluid biopsies. Although exRNA contains a variety of RNA types, such as mRNA fragments and a variety of non-coding RNAs such as miRNA, piRNA, snRNA, lncRNA and circular RNA, the mainstream research of exRNA at this stage is mostly concentrated on star molecules such as miRNA. More than that, these RNAs have unusually complex processing engineering, such as alternative splicing, editing, etc., which generate a large number of RNA variants (isoforms) in body fluids, increasing the difficulty of these exRNA studies. On the other hand, if we can use the diversity of messenger RNA, we most likely to break through this stage of clinical testing means, promote the development of early screening for cancer^[1]. The extraction of these complex, high-throughput diversity information relies heavily on the development of specialized bioinformatics methods and the development of microsequencing technologies.

To be specific, liquid biopsies comprising the noninvasive analysis of circulating tumor-derived material (the "tumor circulome"), represent an innovative tool in precision oncology to overcome current limitations associated with tissue biopsies. The "tumor circulome", dened as the subset of circulating components, is derived from

cancer tissue and can be directly or indirectly used as a source of cancer biomarkers in liquid biopsies^[4]. These components include circulating tumor proteins, circulating tumor nucleic acids (ctDNA and ctRNA), CTCs, Extracellular vesicles (EVs), and tumor-educated platelets (TEPs). EVs, ctRNA, and TEPs are relatively new tumor circulome constituents with promising potential at each stage of cancer management. EVs are membranous particles released from all cell types under physiological and pathological conditions, as well as following different types of stimuli, including proteases, ADP, thrombin, inflammatory cytokines, growth factors, biomechanical shear, and stress inducers, and apoptotic signals^[5]. They can be found in almost every bodily fluid, especially blood^[6]. EVs have been recognized as fundamental mediators of intercellular communication, regulating and participating in a plethora of physiological and pathological processes, including cancer^[6]. Based on their biogenesis, content, and secretory pathways, EVs can be divided into two broad categories: exosomes and microvesicles^[6].

The suitability of EVs as cancer biomarkers lies in the fact that the molecular cargoes they carry can be considered a molecular fingerprint of the cell of origin^[7]. Similar to ctDNA and CTCs, EVs can be a source of quantitative and qualitative information. Quantitative information comprising EV numbers can inform the presence of malignant disease and tumor burden. For example, circulating exosome levels are increased in breast and pancreatic cancer^[8] and the number of circulating microparticles (MPs) is higher in patients with multiple myeloma (MM) compared with healthy individuals^[9].

RNA markers are characterized by diverse species and complex forms, and have important scientific research value With the development and application of various RNA sequencing methods, it has been found that there are a large number of novel non-coding RNAs^[10] different from miRNAs in the genome. In the past, research on the use of RNA sequencing for disease diagnosis mainly focused on changes in transcript expression levels caused by certain diseases^[11], gene fusion^[2] and so on, but with the development and integration of RNA high-throughput sequencing technology and bioinformatics technology, people found much more information than that, and detected clinically relevant RNA species^[1] complex, comprising: different types of RNA (mRNA, miRNA, tRNA, Y RNA, snoRNA, circRNA, lncRNA, etc.), different isoforms,

modified species, intracellular and extracellular distribution, etc. The discovery and understanding of RNA diversity, has a very important scientific value; at the same time, the use of RNA biodiversity information is also likely to break at this stage of clinical testing methods to promote early cancer screening.

The detection of clinically relevant RNA species is complex, including different RNA types (mRNA, tRNA, miRNA, Y-RNA, snoRNA, circRNA, lncRNA, etc.), different splicing forms, modified species, intracellular and extracellular distribution, etc. The RNA content of EVs, including both coding and noncoding (nc)RNAs, has been widely studied.

A.2 Current research status internationally

The research progress related to the RNA marker research of this subject is briefly introduced as follows:

Internationally, some large-scale research teams and commercial organizations have begun to conduct exploratory research on exRNA (extra-cellular RNA) as a biomarker. Recently, the National Center for Translational Science (NCATS) under the NIH of the United States launched the exRNA research project ERCC (Extracellular RNA Communication Consortium) 3, which includes several research directions, including 1) exRNA treatment protocols; 2) molecular markers; 3) The mechanism of action and other aspects have also funded multiple research groups to conduct research. Mainly to detect EML4-ALK gene fusion and EGFR T790M gene mutation in plasma samples^[14]. The first representative product is ExoDx lung (ALK), EML4-ALK is used to detect non-small cell lung cancer patient's plasma exosomes transcripts. In addition, many teams began using RNA circulating in the mother's body fluids of information to reflect the health of the fetus.

In previous studies of body fluid exRNA, miRNA was a star molecule and received extensive attention. miRNAs can be endogenously expressed in a variety of cells and secreted into a variety of body fluids (blood, saliva, and urine). Based on these features, miRNA can be used as a non-invasive biomarker to become one of the ideal

candidate biomarkers of human diseases including cancer^[16]. A recent study of non-small cell lung cancer (NSCLC, accounting for 85% of all cases of lung cancer) in patients with early plasma samples were genome-wide miRNA expression profiling, miRNA found 24 kinds of circulation, has a high diagnostic value^[17]; another study, serum of miRNA expression has undergone radical prostatectomy for prostate cancer patients were analyzed, the results found 43 kinds of miRNA, can distinguish between different disease stages 14 prostate cell lines and patient samples^[18]. China has also made significant achievements in the research and application of miRNA biomarkers. For example, in a study of liver cancer, Chinese scientists using miRNA microarray patients with hepatocellular carcinoma and the normal population, patients with chronic hepatitis, liver cirrhosis patient plasma samples, and screened seven most significant effect of miRNA, the establishment of a multi-index Logistic regression analysis model to distinguish between liver cancer patients and other control populations^[19]. The kit based on this result has completed multi-center clinical validation and recently passed the certification of the State Food and Drug Administration. Many of these detection methods based on miRNA has reached a very high (about 80% -90%) sensitivity (sensitivity), and specificity (specificity) is also very good, up to about 70-80% and more; but it also shows There are still about $\frac{1}{4}$ to $\frac{1}{5}$ misdiagnosis rates and room for improvement.

The discoveries and research of new exRNA progress quickly, exRNA types that can be used as biomarkers are far more than miRNA. Nearly all known classes of RNA have been found in systemic circulation and, to a certain extent, each has the potential to serve as a cancer biomarker^[20]. The most important classes of ctRNA potentially suitable as biomarkers are mRNAs, miRNAs, and long ncRNAs (lncRNAs). Their analysis is performed with techniques ranging from qRT-PCR or dPCR-based assessment of single or small panels of RNAs to the comprehensive characterization of RNA (especially miRNAs) signatures via RNASeq20. By definition, a variety of extracellular RNA are collectively referred to as exRNA, they play an important role in the communication between cells, can be transported to tissues adjacent to or distant, it is taken up and transferred they carry genetic regulation of target cells and information. exRNA would normally be wrapped into the exosomes (with exosomes), microvesicles

vesicles (microvesicles, MVs)^[21] and the like, RNA and protein complexes of non-vesicular structures (RNPs)^[22], which have different grain. The size of the diameter can usually be separated by ultracentrifugation and physical sedimentation^[23] and can be distinguished according to different surface markers. Studies have shown that MVs, exosomes, and RNPs have different RNA compositions, such as miRNAs that are abundantly enriched in exosomes; RNPs contain large amounts of tRNA and Y-RNA fragments, making it easy to extract them from MVs^[24].

Circulating exosomal mRNA has been used to investigate the mutational status of KRAS and BRAF in patients with CRC^[25], and exosomal EGFR vIII mRNA has the potential for the diagnosis of EGFRvIII-positive high-grade gliomas^[26]. In another report, the detection of androgen receptor splice variant 7 (AR-V7) in plasmatic exosomes by ddPCR was shown to be a good predictor of resistance to hormonal therapy n prostate cancer^[27]. Numerous lung cancer-related gene fusions are also readily identified in both vesicular and nonvesicular mRNA and have value as biomarkers^[27]. Among the nonvesicular fraction of ctRNAs, circulating human telomerase reverse transcriptase (hTERT, the catalytic subunit of the telomerase complex) mRNA demonstrated greater diagnostic and prognostic accuracy than PSA for prostate cance^[28].

With regards to miRNAs, plasma exosomal miR-196a and miR-1246 levels have the potential for the early diagnosis of pancreatic cancers^[29], and panels of miRNAs have been shown to be reliable biomarkers for the diagnosis^[30] or prognosis^[31] of lung cancer. More recently, a serum exosomal miRNA signature was proven to be an innovative tool for the differential diagnosis of gliomas^[31].

It is also well-known that there is a lncRNA called PCA3 and it has been identified as a molecular marker in the urine for prostate cancer. For example, plasma exosome LINC00152 levels have been linked to gastric cancer^[32], and the combination of two mRNAs and one lncRNA in serum exosomes have diagnostic potential for CRC^[33]. Furthermore, serum exosomal HOTAIR lncRNA has applicability in the diagnosis and prognosis of glioblastoma multiforme^[34]. More recently, a panel of five circulating lncRNAs was studied as promising diagnostic biomarkers for gastric cancer^[35].

A recent study shows that circular RNAs can be detected in urine and has the potential to serve as prostate cancer's biomarker^[36]. Circular RNAs (circRNAs) are single-stranded, covalently closed RNA molecules that are produced from pre-mRNAs

through a process called back splicing and were initially proposed to be splicing-associated noise^[37]. Recent studies have shown that circRNAs may be involved in microRNA (miRNA) inhibition^[38], epithelial-mesenchymal transition^[39], and tumorigenesis^[40]. Further, circRNA expression can be tissue specific^[39], and some evidence supports the translation of some circRNAs^{[41],[42]}. CircRNAs are highly stable and can be found in exosomes, cell-free saliva, and plasma. Therefore, with improved detection and characterization methodologies, circRNAs may be potential biomarkers or therapeutic targets.

Therefore, there are different extracellular components in the body fluid, and the RNA species and content of the different components are very different.

In order to sequence exRNAs in body fluids, high-throughput sequencing technology for low-input RNAs is one of the key issues. The amount of free RNA in body fluids is low^[43], and RNA itself is easily degraded by endogenous and exogenous RNase, so some specialized methods and commercial kits have been established for the extraction and sequencing of trace RNA, for example, in single cell transcription. and genomic experiments, building an RNA library can be a single cell sequencing of RNA picograms (pg) level, the main application is SMARTer^[44] library construction sequencing technology and other technical MALBAC^[45], these two techniques were used to capture trace the template switch full-length RNA and the use of multiple annealing circular loop amplification techniques allow the ends of the amplicon to complement each other to prevent exponential amplification.

In recent years, a variety of different RNA library construction methods have been developed for exRNA sequencing, and the results obtained by different RNA library construction methods have significantly different results. The cell-free RNA-seq obtained by four different RNA-seq methods was significantly different in rRNA ratio, library size, microbiome, and detectable different types of RNA.

In addition to the abundance of RNA expression, variants and isoforms resulting from post-transcriptional regulation of RNA can also serve as markers for the development of cancer. Studies have found that in cancer cells, compared with normal cells, RNA splicing, poly (A) tailing and editing and other post-transcriptional

regulatory events are misregulated. Shen et al. systematically identified a number of alternative splicing events^[46] associated with the clinical outcome of cancer by analyzing TCGA data. Xia et al. also analyzed the data of TCGA and found that in cancer cells, a large number of cancer-related genes are regulated by alternative polyadenylation, and the 3'UTR shortened genes are prone to be up-regulated, demonstrating alternative polyadenylation is also closely related to the development of cancer^[47]. In recent years, more and more studies have found that RNA editing events are also erroneously regulated in cancer patients. Many RNA editing sites in mRNA and miRNA are identified due to differences in editing levels between cancer patients and normal people. For potential biomarkers^{[48],[49]}.

A.3 Previous analyzing tools for small RNA-seq

There are several existing tools for small RNA-seq analysis: ExceRpt^[50], TIGER^[51], Chimira^[52], miRge^[53], and Oasis^[54]. Some imputation, normalization and batch correction tools including scImpute, SCnorm, and combat may also be useful.

The exRNA-seq dataset has certain properties: fragmented, sparse, and heterogeneity. Some tools consider its fragmented characteristics and there are many scRNA-seq analyzing tools for normalization and batch correction.

Tools for Integrative Genome analysis of Extracellular sRNAs (TIGER) was performed on mouse lipoproteins, bile, urine, and livers. A key advance for the TIGER pipeline is the ability to analyze both host and non-host sRNAs at genomic, parent RNA and individual fragment levels. Moreover, TIGER facilitated the comparison of lipoprotein sRNA signatures to disparate sample types at each level using hierarchical clustering, correlations, beta-dispersions, principal coordinate analysis and permutational multivariate analysis of variance. TIGER analysis was also used to quantify distinct features of exRNAs, including 5 miRNA variants, 3 miRNA non-templated additions and parent RNA positional coverage. The researchers have observed that miRNA explained less than 5% of quality sequencing depth of lipoproteins and only 15% of liver sequencing depth. It seems that non-coding RNAs are processed to smaller fragments creating an enormously diverse pool of sRNAs in cells and extracellular fluids. They found there are several sRNAs are derived from tRNA, snRNA,

and rRNA. TIGER also have more advanced ability to analyze sRNAs since other tools are restricted to miRNAs or endogenous (host) sRNAs, including Chimira, Oasis, and miRge.

Small RNA-seq pipeline exceRpt can be used for processing and analyzing the results of the experimental data. Alignment to exogenous genomes and their quantification results were used for the analyses of small RNAs of exogenous origin.

Some single cell tools for normalization and batch correction are also useful. Methods used to quantify mRNA abundance introduce systematic sources of variation that can obscure signals of interest. Consequently, an essential first step in most mRNA-expression analyses are normalization, whereby systematic variations are adjusted to make expression counts comparable across genes and/ or samples. Within-sample normalization methods adjust for gene-specific features, such as GC content and gene length, to facilitate comparisons of a gene's expression within an individual sample; whereas between-sample normalization methods adjust for sample-specific features, such as sequencing depth, to allow for comparisons of a gene's expression across samples. SCnorm is a method for between-sample normalization which also allows gene-specific features to be adjusted. A number of methods are available for between-sample normalization in bulk RNA-seq experiments. Most of these methods calculate global scale factors (one factor is applied to each sample, and this same factor is applied to all genes in the sample) to adjust for sequencing depth. These methods demonstrate excellent performance in bulk RNA-seq, but they are compromised in the single-cell setting because of an abundance of zero-expression values and increased technical variability. scRNA-seq data show systematic variation in the relationship between transcript-specific expression and sequencing depth (which we refer to as the count-depth relationship) that is not accommodated by a single scale factor common to all genes in a cell. Global scale factors adjust for a count-depth relationship that is assumed to be common across genes. When this relationship is not common across genes, normalization via global scale factors leads to overcorrection for weakly and moderately expressed genes and, in some cases, under normalization of highly expressed genes. SCnorm uses quantile regression to estimate the dependence of transcript expression on sequencing depth for every gene. Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group.

Within-group adjustment for sequencing depth is then performed using the estimated scale factors to provide normalized estimates of expression. Although SCnorm does not require experimental RNA spike-ins, performance may be improved if spike-ins that span the range of expression observed in endogenous genes are available.

scImpute is currently the state-of-art method to do imputation on scRNA-seq data. It is a statistical method to accurately and robustly impute the dropouts in scRNA-seq data. scImpute automatically identifies likely dropouts, and only perform imputation on these values without introducing new biases to the rest data. scImpute also detects outlier cells and excludes them from imputation. Evaluation based on both simulated and real human and mouse scRNA-seq data suggests that scImpute is an effective tool to recover transcriptome dynamics masked by dropouts. scImpute is shown to identify likely dropouts, enhance the clustering of cell subpopulations, improve the accuracy of differential expression analysis, and aid the study of gene expression dynamics. scImpute focuses on imputing the missing expression values of dropout genes, while retaining the expression levels of genes that are largely unaffected by dropout events. Hence, scImpute can reduce technical variation resulted from scRNA-seq and better represent cell-to-cell biological variation, while it also avoids introducing excess biases during its imputation process. To achieve the above goals, scImpute first learns each gene's dropout probability in each cell by fitting a mixture model for each cell type. Next, scImpute imputes the (highly probable) dropout values of genes in a cell by borrowing information of the same gene in other similar cells, which are selected based on the genes not severely affected by dropout events. Comprehensive studies on both simulated and real data suggest that compared with the raw scRNA-seq data, the imputed data by scImpute better present cell type identity and lead to more accurate DE analysis results.

conclusion In the review, we have discussed the potential application of exRNA in diagnosis and prognosis of some complex diseases including cancer. We have discussed some previous progress and their limitations. Some challenges we face when dealing with exRNA data analysis and cancer prediction. It is clear that an integrative and better tool for exRNA data analysis is essential and useful. We aim to develop such kind of tool for mapping, expression matrix construction, matrix processing, and feature selection.

Some data issues include sparsity, fragmentation, heterogeneity and batch effect. We plan to design and apply certain methods to deal with these problems in our project.

References

- [1] Byron, S. A. et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* 17, 257-271, doi:10.1038/nrg.2016.10 (2016).
- [2] Xi, X. et al. RNA Biomarkers: Frontier of Precision Medicine for Cancer. *Non-Coding RNA* 3, doi:10.3390/ncrna3010009 (2017).
- [3] Ainsztein, A. M. et al. The NIH Extracellular RNA Communication Consortium. *J Extracell Vesicles* 4, doi:UNSP 27493 10.3402/jev.v4.27493 (2015).
- [4] De Rubis, G. e. a. Circulating tumor DNA - current state of play and future perspectives. *Pharmacol. Res.* (2018).
- [5] Taylor, J. a. B., M. Proteins regulating microvesicle biogenesis and multidrug resistance in cancer., doi: <http://dx.doi.org/10.1002/pmic.201800165> (2018).
- [6] van Niel, G. e. a. Shedding light on the cell biology of extracellular vesicles. *Nat. Rev. Mol. Cell Biol.* (2018).
- [7] Torrano, V. e. a. Vesicle-MaNiA: extracellular vesicles in liquid biopsy and cancer. *Curr. Opin. Pharmacol.* (2018).
- [8] Melo, S. A. e. a. Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature* 523, 177–182 (2015).
- [9] Krishnan, S. R. e. a. Isolation of human CD138(+) microparticles from the plasma of patients with multiple myeloma. *Neoplasia* 18, 25-32 (2016).
- [10] Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74, doi:10.1038/nature11247 (2012).
- [11] Sparano, J. A. et al. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *New Engl J Med* 373, 2005-2014, doi:10.1056/NEJMoa1510764 (2015).
- [12] Vardiman, J. W. et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* 114, 937-951, doi:10.1182/blood-2009-03-209262 (2009).
- [13] Garcia-Romero, N. e. a. Extracellularvesiclescompartmentin liquid biopsies: clinical application. *Mol. Aspects Med.* 60, 27-37 (2018).
- [14] Brock, G. et al. Liquid biopsy for cancer screening, patient stratification and monitoring. *Transl Cancer Res* 4, 280-290, doi:10.3978/j.issn.2218-676X.2015.06.05 (2015).
- [15] Tsui, N. B. Y. et al. Maternal Plasma RNA Sequencing for Genome-Wide Transcriptomic Profiling and Identification of Pregnancy-Associated Transcripts. *Clin Chem* 60, 954-962, doi:10.1373/clinchem.2014.221648 (2014).
- [16] Schwarzenbach, H. et al. Clinical relevance of circulating cell-free microRNAs in cancer.

Nat Rev Clin Oncol 11, 145-156, doi:10.1038/nrclinonc.2014.5 (2014).

- [17] Wozniak, M. B. et al. Circulating MicroRNAs as Non-Invasive Biomarkers for Early Detection of Non-Small-Cell Lung Cancer. Plos One 10, doi:ARTN e0125026 10.1371/journal.pone.0125026 (2015).
- [18] Singh, P. K. et al. Serum microRNA expression patterns that predict early treatment failure in prostate cancer patients. Oncotarget 5, 824-840, doi:DOI 10.18632/oncotarget.1776 (2014).
- [19] Zhou, J. et al. Plasma MicroRNA Panel to Diagnose Hepatitis B Virus-Related Hepatocellular Carcinoma. J Clin Oncol 29, 4781-4788, doi:10.1200/Jco.2011.38.2697 (2011).
- [20] Zaporozhchenko, I. A. e. a. The potential of circulating cell-free RNA as a cancer biomarker: challenges and opportunities. Expert Rev. Mol. Diagn 18, 133-145 (2018).
- [21] Valadi, H. et al. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9, 654-U672, doi:10.1038/ncb1596 (2007).
- [22] Vickers, K. C. et al. MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins (vol 13, pg 423, 2011). Nat Cell Biol 17, 104-104, doi:10.1038/ncb3074 (2015).
- [23] Nakano, I. et al. Extracellular vesicles in the biology of brain tumour stem cells—Implications for inter-cellular communication, therapy and biomarker development. Semin Cell Dev Biol 40, 17-26, doi:10.1016/j.semcd.2015.02.011 (2015).
- [24] Wei, Z. et al. Coding and noncoding landscape of extracellular RNA released by human glioma stem cells. Nat Commun 8, 1145, doi:10.1038/s41467-017-01196-x (2017).
- [25] Hao, Y. X. e. a. KRAS and BRAF mutations in serum exosomes from patients with colorectal cancer in a Chinese population. Oncol. Lett. 13, 3608–3616 (2017).
- [26] Manda, S. V. e. a. Exosomes as a biomarker platform for detecting epidermal growth factor receptor-positive high-grade gliomas. J. Neurosurg. 128, 1091–1101 (2018).
- [27] Aguado, C. e. a. Fusion gene and splice variant analyses in liquid biopsies of lung cancer patients. Transl. Lung Cancer Res 5, 525–531 (2016).
- [28] March-Villalba, J. A. e. a. Cell-free circulating plasma hTERT mRNA is a useful marker for prostate cancer diagnosis and is associated with poor prognosis tumor characteristics. Plos One (2012).
- [29] Xu, Y. F. e. a. Plasma exosome miR-196a and miR-1246 are potential indicators of localized pancreatic cancer. Oncotarget 8, 77028–77040.
- [30] Jin, X. e. a. Evaluation of tumor-derived exosomal miRNA as potential diagnostic biomarkers for early-stage non-small cell lung cancer using next-generation sequencing. Clin. Cancer Res. 23, 5311–5319.
- [31] Liu, Q. e. a. Circulating exosomal microRNAs as prognostic biomarkers for non-small-cell lung cancer. Oncotarget 8, 13048–13058.
- [32] Li, Q. e. a. Plasma long noncoding RNA protected by exosomes as a potential stable biomarker for gastric cancer. Tumour Biol. 36, 2007–2012 (2015).
- [33] Dong, L. e. a. Circulating long RNAs in serum extracellular vesicles: their characterization

and potential application as biomarkers for diagnosis of colorectal cancer. *Biomarkers Prev.* 25, 1158–1166 (2016).

- [34] Tan, S. K. e. a. Serum long noncoding RNA HOTAIR as a novel diagnostic and prognostic biomarker in glioblastoma multiforme. *. Mol. Cancer* (2018).
- [35] Zhang, K. e. a. Genome-wide lncRNA microarray profiling identifies novel circulating lncRNAs for detection of gastric cancer. *Theranostics* 7, 213–227 (2017).
- [36] Josh N. Vo, M. C., Yajia Zhang, ..., Dan R. Robinson, Alexey I. Nesvizhskii, Arul M. Chinnaiyan. The Landscape of Circular RNA in Cancer. *Cell* (2019).
- [37] Capel, B., Swain, A., Nicolis, S., Hacker, A., Walter, M., Koopman, P., Goodfellow, P., and Lovell-Badge, R. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* (1993).
- [38] Hansen, T. B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. Natural RNA circles function as efficient microRNA sponges. *Nature* (2013).
- [39] Conn, S. J., Pillman, K.A., Toubia, J., Conn, V.M., Salmanidis, M., Phillips, C.A., Roslan, S., Schreiber, A.W., Gregory, P.A., and Goodall, G.J. The RNA binding protein quaking regulates formation of circRNAs. *Cell* (2015).
- [40] Guarnerio, J., Bezzi, M., Jeong, J.C., Paffenholz, S.V., Berry, K., Naldini, M.M., Lo-Coco, F., Tay, Y., Beck, A.H., and Pandol, P.P. Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations. *Cell* (2016).
- [41] Legnini, I., Di Timoteo, G., Rossi, F., Morlando, M., Briganti, F., Sthandier, O., Fatica, A., Santini, T., Andronache, A., Wade, M., et al. Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Mol. Cell* (2017).
- [42] Pamudurti, N. R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., Hanan, M., Wyler, E., Perez-Hernandez, D., Ramberger, E., et al. Translation of CircRNAs. *Mol. Cell* (2017).
- [43] Schwarzenbach, H. et al. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* 11, 426–437, doi:10.1038/nrc3066 (2011).
- [44] Yang, H. et al. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* 10, 1556–1566, doi:10.1038/nprot.2015.105 (2015).
- [45] Chapman, A. R. et al. Single cell transcriptome amplification with MALBAC. *Plos One* 10, e0120889, doi:10.1371/journal.pone.0120889 (2015).
- [46] Shen, S. H. et al. SURVIV for survival analysis of mRNA isoform variation. *Nature Communications* 7, doi:ARTN 11548 10.1038/ncomms11548 (2016).
- [47] Xia, Z. et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nature Communications* 5, doi:ARTN 5257 10.1038/ncomms6274 (2014).
- [48] Wang, Y. M. et al. Systematic characterization of A-to-I RNA editing hotspots in microRNAs across human cancers. *Genome Res* 27, 1112–1125, doi:10.1101/gr.219741.116 (2017).
- [49] Liang, H. The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Res* 76, doi:10.1158/1538-7445.Am2016-2661 (2016).

- [50] Kaczor-Urbanowicz, K. E., Kim, Y., Li, F., Galeev, T., Kitchen, R.R., Gerstein, M., Koyano, K., Jeong, S.-H., Wang, X., Elashoff, D., et al. Novel approaches for bioinformatic analysis of salivary RNA sequencing data for development. *Bioinformatics* 34, 1-8 (2018).
- [51] Allen, R. M., Zhao, S., Ramirez Solano, M.A., Zhu, W., Michell, D.L., Wang, Y., Shyr, Y., Sethupathy, P., Linton, M.F., Graf, G.A., et al. Bioinformatic analysis of endogenous and exogenous small RNAs on lipoproteins. *J Extracell Vesicles* 7 (2018).
- [52] Vitsios DM, E. A. Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*. 31, 3365–3367 (2015).
- [53] Baras AS, M. C., Myers JR, et al. miRge - a multiplexed method of processing small RNA-seq data to determine microRNA entropy. *PLoS One*. (2015).
- [54] Capece V, G. V. J., Vidal R, et al. Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics* 31, 2205–2207. (2015).