武乐,李益楠,孔德信,等.基于Snakemake的RNA-seq数据自动化分析流程RNApipe[J].华中农业大学学报,2022,41(6):143-151.
DOI:10.13300/j.cnki.hnlkxb.2022.06.016

基于Snakemake的RNA-seq数据自动化分析流程RNApipe

武乐[1],李益楠[2],孔德信[1],周志鹏[2]

1.华中农业大学信息学院,武汉 430070; 2.华中农业大学生命科学技术学院,武汉 430070

摘要 为使科研工作者简单高效地分析RNA-seq数据,本研究基于Snakemake工作流程管理系统和Conda环境管理器构建了一个自动化和模块化的工作流程:RNApipe(Github:https://github.com/ywu019/RNApipe.git),其可对来自任何有参物种的RNA-seq数据自动执行质控、比对、定量、鉴定差异基因,以及GO、KEGG、GSEA等功能注释分析;其中,每一步骤的分析结果均以高质量的可视化图片或报告展示,并保留重要的输出文件。使用RNApipe在多个模式物种中的测试与评估结果表明:RNApipe可以平稳运行,且注释结果准确。与现有的自动化分析流程相比,RNApipe的主要特点包括:工作流程较为完整、默认工具消耗时间与资源较少、适用于任何有参物种、全面的可视化、以及用户友好性(易安装、易使用、易扩展)。研究表明,RNApipe便于研究人员快速地从大型RNA-seq测序数据中获取基本信息。

关键词 转录组测序; 差异表达分析; 功能注释; Snakemake; Conda; 自动化数据分析; 可视化

中图分类号 TP311.13 **文献标识码** A **文章编号** 1000-2421(2022)06-0143-09

转录组测序(RNA-seq)是一项对特定组织或细胞中所有的RNA进行高通量测序的技术。近十年来,随着测序技术的发展、测序成本的降低以及分析工具的快速开发,RNA-seq已广泛应用于研究基因表达调控、可变剪切甚至RNA结构等方向,成为生命科学基础研究领域中非常重要的技术[1]。RNA-seq最常应用于识别特定条件下差异表达的基因(differential gene expression,DGE)并鉴定其分子功能[2]。DGE的分析流程主要包括5个步骤:对高通量测序平台产生的原始读段(reads)进行质量检测和控制、将预处理后的reads比对到参考基因组、在基因组或转录组水平上进行定量、鉴定样本间差异表达的基因,以及表征差异表达基因的生物学功能[3]。

目前,RNA-seq的各个分析步骤都已有许多成熟的计算方法。然而,这种分段分析的方法通常会涉及到不同的操作系统和编程语言,这对于科研人员的编程能力提出较高的要求。之前已有多项研究将各个分析步骤集成到自动化分析工作流程中,从而降低对用户编程能力的要求并有效地提高其工作效率。然而,目前已开发的RNA-seq自动化分析工作流程中存在着一些不足之处。例如,VIPER[4]与hppRNA[5]等工具只支持少数几个特定物种的数据分析,IRIS-EDA[6]与ideal[7]等工具仅适用于差异表达基因的下游分析和可视化,BioJupies[8]与RNA-Cocktail[9]等工具不支持原始数据的预处理和质量控制等步骤,而ARRMOR[10]与RASflow[11]等工具缺乏对差异基因的功能注释。

本研究通过Conda部署环境以确保整个工作流程中的软件快速、平稳地安装,并通过Snakemake工作流程管理系统集成各个分析步骤,最终构建了1个流程更完整且更易使用的RNA-seq自动化分析工具——RNApipe,旨在为研究人员快速、简便地从大型RNA-seq测序数据中获取基本信息提供技术支持。

1 材料与方法

1.1 Conda环境与Snakemake工作流程管理系统

对于编程经验有限的用户,安装并维护RNA-seq数据分析中必需的工具及其依赖项通常具有挑

收稿日期: 2022-01-13
基金项目:国家自然科学基金项目(31970552)
武乐,E-mail:wuler005@163.com

战性。Conda作为软件包和环境管理系统，可以自动安装、运行和更新软件包及其依赖项，并且支持虚拟环境的创建、切换与移植，是确保软件管理可持续和可重现的理想工具[12]。因此，RNApipe使用Conda安装和运行软件，并创建虚拟环境。这不仅能够保证RNApipe的顺利安装与平稳运行，还确保了独立于操作系统和机器的数据可重复性分析。

为减少数据分析过程中手动执行的步骤，传统的工作管道通常使用自定义的脚本将多个工具进行链接，然而这种管道通常与本地计算基础架构高度相关，难以共享与维护，产生的结果也较难重复[13]。为此，RNApipe使用Snakemake工作流程管理系统将RNA-seq各个分析步骤进行组合。Snakemake[14]是一个基于python的工作流程管理系统，它具有以下优势：通过集成Conda包管理器自动安装和配置软件，以确保不同平台的可移植性；通过支持高性能并行计算和云计算环境，以实现超越本地基础架构的可扩展性；通过支持工作流间缓存，避免在管道之间重新运行程序而浪费时间和计算资源；并通过将各个步骤分解为模块化规则(rules)，从而降低了代码的复杂性，使代码易读易维护易扩展。最终，RNApipe基于Snakemake管理系统，可以实现通过1条命令执行整个自动化分析过程，极大地简化了程序的运行方式(图1)。

1.2 RNApipe工作流程框架

整个RNApipe工作流程(图1)包括：(1)RNApipe启动时，首先使用fastp[15](默认)或Cutadapt[16]对每个样本进行检测并修剪接头、去除低质量碱基和reads；随后使用FastQC[17]检测reads质量，质检结果由MultiQC[18]汇总为一个可视化报告。(2)质量达到用户要求的reads会被HISAT2[19](默认)或STAR[20]比对到参考基因组，比对结果由Qualimap2[21]进行质量评估，比对文件以BAM或Bigwig格式(可选)保存。(3)使用featureCounts[22](默认)或HTSeq-count[23]对BAM文件进行定量(count)。(4)其中表达矩阵(count)用于DESeq2[24]、edgeR[25](可选)或limma[26](可选)鉴定差异表达的基因。TPM标准化的表达矩阵用于样本间基因的表达量分析。(5)使用KOBAS-i[27]对差异基因进行GO功能注释与KEGG pathway通路分析，对所有表达的基因做GSEA基因集富集分析。以上每个步骤都可以输出高质量的可视化图片或报告，并保留重要的输出文件和表格数据。

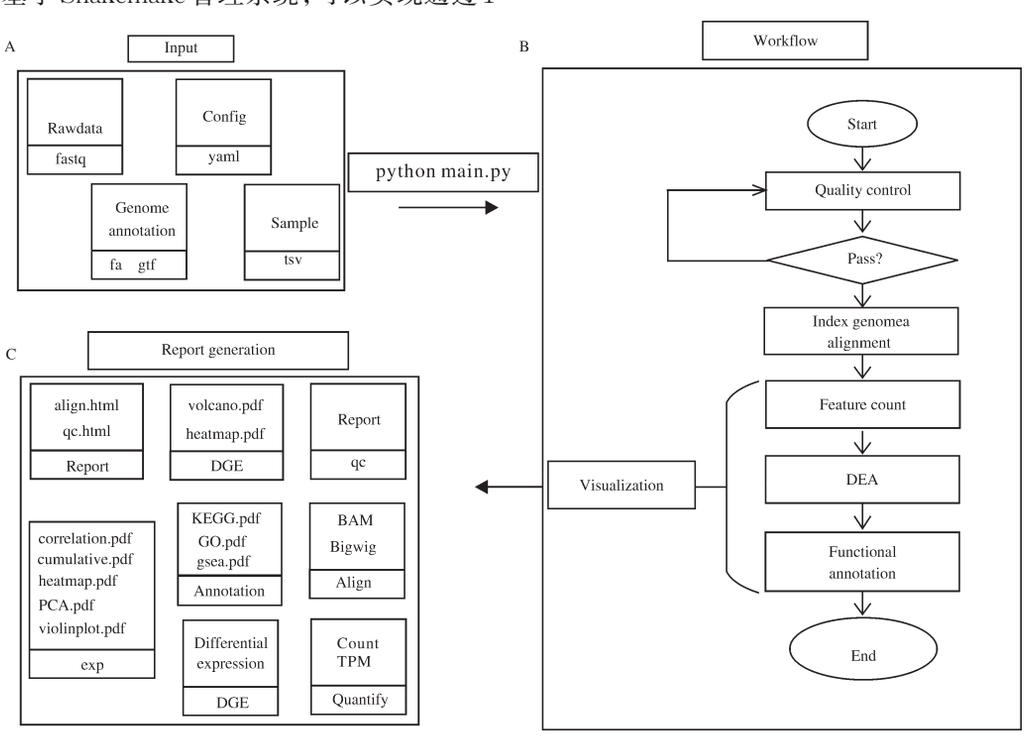


图1 RNApipe数据分析流程图

Fig.1 The workflow chart of RNApipe analysis

1.3 RNApipe使用方法

RNApipe的源代码已经上传至Github供用户下载和使用：https://github.com/ywu019/RNApipe.git。以下简单介绍RNApipe的安装和使用方法，具

体使用方法可见 Github 中的说明文档(Documentation.pdf)。

1)下载 RNApipe 工具与安装环境。首先通过 Python3.7 在 Linux 系统中安装 miniConda3 环境，随后通过以下 3 条命令完成 RNApipe 工具的安装与环境的创建：① 从 GitHub 下载 RNApipe：it clonehttps://github.com/ywu019/RNApipe.git；②创建环境：conda env create -n RNApipe -f envs/envs.yaml；③激活环境：conda activate RNApipe。

2)用户指定输入文件。在运行 RNApipe 前需要指定：基因组文件(.fa)、注释文件(.gtf)与压缩的测序文件(.fastq.gz)(图 1A)。其中基因组文件和注释文件建议从 NCBI 下载，测序文件支持单端和双端测序，其命名应符合“条件_重复”(或“条件_重复_双端”)的形式，例如 root_1_R1.fastq.gz。

随后修改 configs/目录下的样本表(samples.tsv)和配置文件(config.yaml)。用户可以根据实验条件指定样本表中的条件列“condition”与重复列“replicate”的信息，并修改配置文件中的基本设置，配置文件主要包含以下变量：

PROJECT：#项目名称；

READPATH:#测序文件 fastq.gz 的路径；

END:#测序数据类型是 SE(single-end)还是 PE(paired-end)；

OUTPUTPATH：#输出文件的路径；

GENOME:#基因组文件路径；

ANNOTATION:#注释文件路径；

CONTROL：#对照组名称，与 samples.tsv 中 condition 一致；

TREAT:#处理组名称；

SPEICES_ABBREVIATION：物种简称，在 speices_abbreviations 中查询；

TRIM：fastp #[fastp，cutadapt] 选择工具执行数据预处理；

ADAPTER：AGATCGGAAGAGC #接头序列；

ALIGN：hisat2 #[hisat2，STAR] #比对软件；

BIGWIG：["no"] #[yes，no] #是否将 bam 格式的文件转化为 bigwig 格式；

DEA：DESeq2 #[DESeq2，edgeR，limma] #差异基因分析软件；

QUANTIFY：featureCounts #[featureCounts，htseq] #定量软件；

THREAD："10" #线程数目；

TIME："5" #程序延迟时间。

3)运行工作流程。准备好所有的输入文件后，用户只需要通过一条命令(python main.py)即可在本地或集群环境中运行 RNApipe 整个工作流程。运行结果保存在用户指定的输出目录中，logs/目录下生成各个分析步骤的日志文件，其中记录了程序的运行时间、标准输出和标准错误等信息。

2　结果与分析

为了让用户了解 RNApipe 的工作原理和基本功能，本研究利用 RNApipe 对来自 4 日龄野生型拟南芥根部(root)和芽部(shoot)的单端 RNA-seq 完整数据集(PRJNA321304；GEO：GSE81332)进行自动化分析，其中根部设为对照组，芽部设为试验组。

2.1　RNApipe 概要

图 2A 展示了用户从 Github 下载的 RNApipe 所包含的完整数据集。example_data/目录代表 1 个真实的测试数据子集(GSE81332)，便于用户在运行实际项目前对系统和软件进行测试。Documentation.pdf 文档对 RNApipe 的安装、使用方法以及注意事项进行了详细的介绍。

用户在运行项目时应在 project/目录中设置类似于 example_data/的项目结构，并将测序数据、参考基因组和注释文件分别置于相应的子目录中(图 2A)，再根据实验条件修改 config.yaml 文件(图 2B)和 sample.tsv 文件(图 2C)。当输入文件设置完毕，只需要 1 条简单的命令(图 1 命令)，RNApipe 将根据 config.yaml 文件中的设置，通过 Snakemake 对 main.py 中的规则自动执行整个分析过程。输出结果存于 trim/、align/、quantify/、DEA/、annotation/、visualize/等目录，分别对应质控、比对、定量、差异分析、功能注释和图形可视化结果(图 1C)。其余文件则不需要修改，除非用户需要自定义工作流程。

2.2　数据质量可视化

为了保证分析结果的准确性，RNApipe 分别对质控后的数据和比对文件进行质量检测，并将每个样本文件生成的质控报告(html)进行汇总(图 3)。通过质控报告检查当前 reads 是否仍需要修剪，当数据质量较低时，例如：碱基质量差、GC 含量不稳定、存在接头等，则建议继续修剪以提高数据质量。如图 3A 所示，6 个样本在质控后的碱基质量均在 Q30 以上，表明碱基质量较好，可用于后续数据分析。图 3B 展示了 6 个样本的详细比对情况。

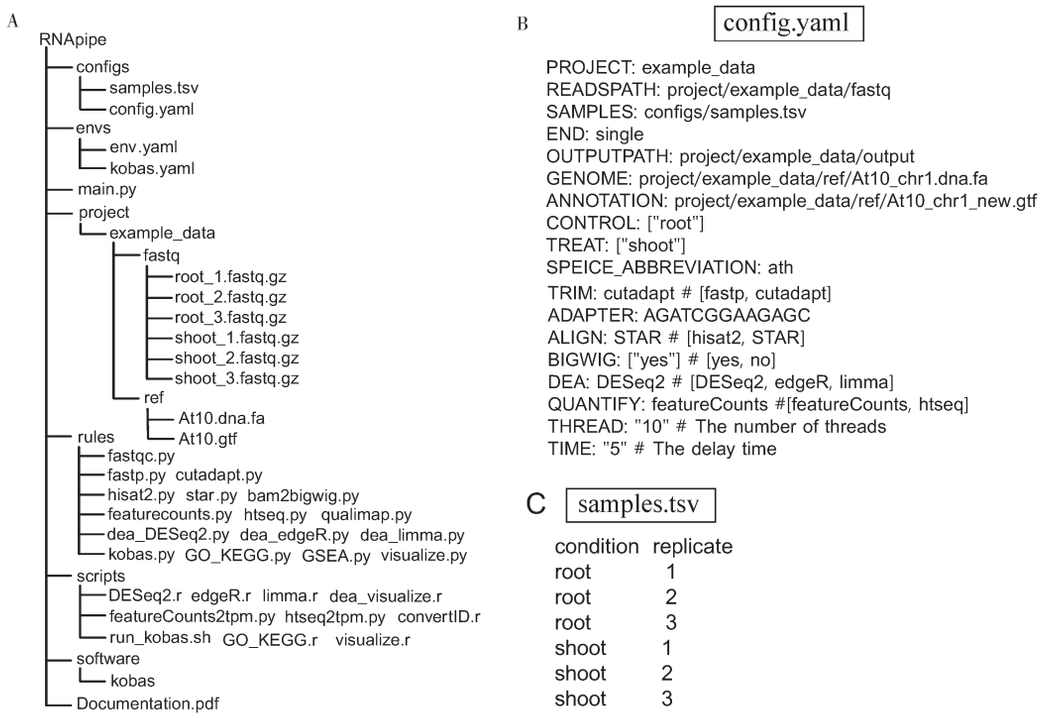
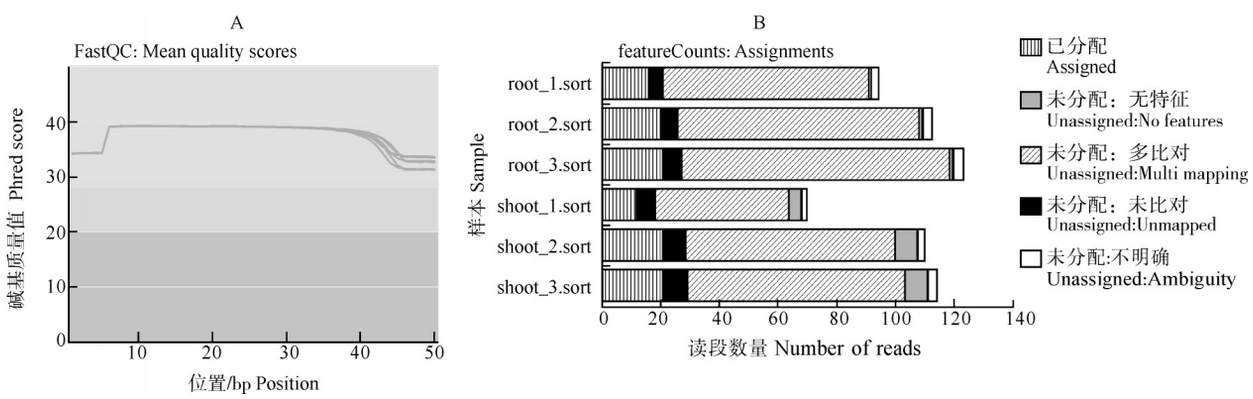


图2 RNApipe的文件与目录结构

Fig.2 The file and directory structures of RNApipe



A:读段中碱基质量的平均值 The mean quality value across each base position in the read; B:样本比对情况 A brief mapping summary of all six samples.

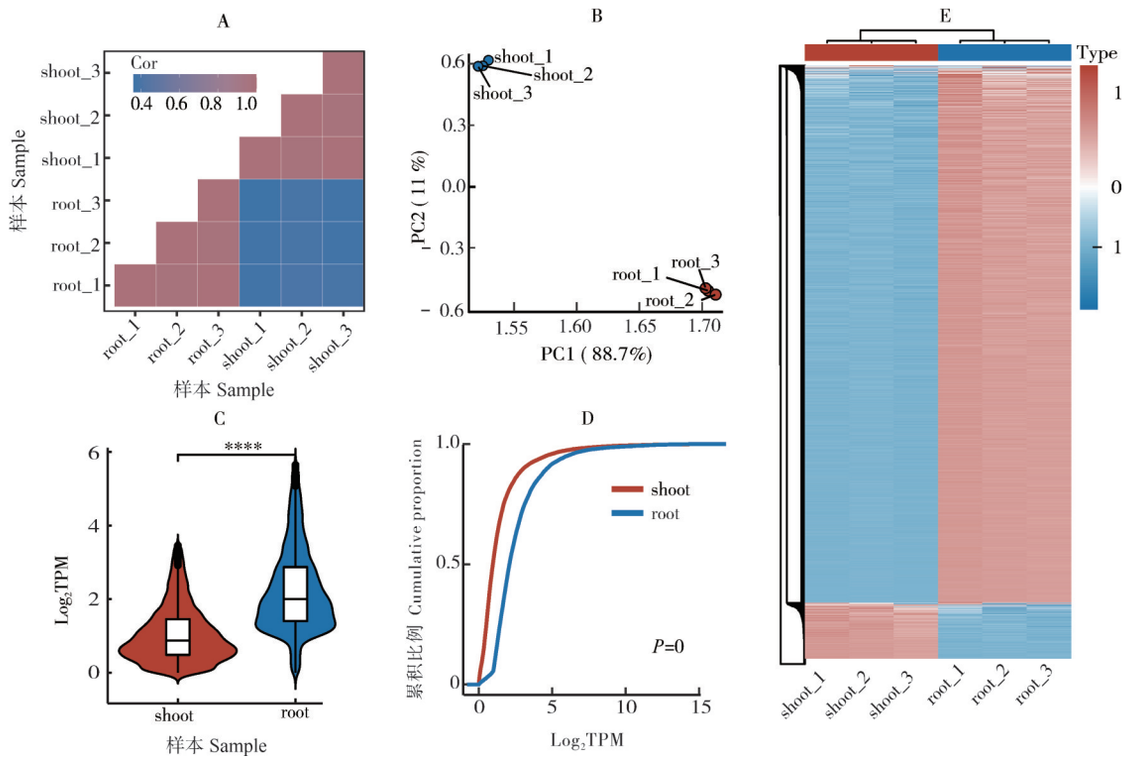
图3 数据质量检测报告

Fig.3 Report on data quality inspection

2.3 基因表达量可视化

在完成质量控制和比对后,RNApipe将表达矩阵(count)进行TPM标准化,以消除测序深度和基因长度对表达量的影响。设置阈值(TPM<1)过滤低表达的基因,得到的表达量矩阵用于可视化图表展示(图4)。

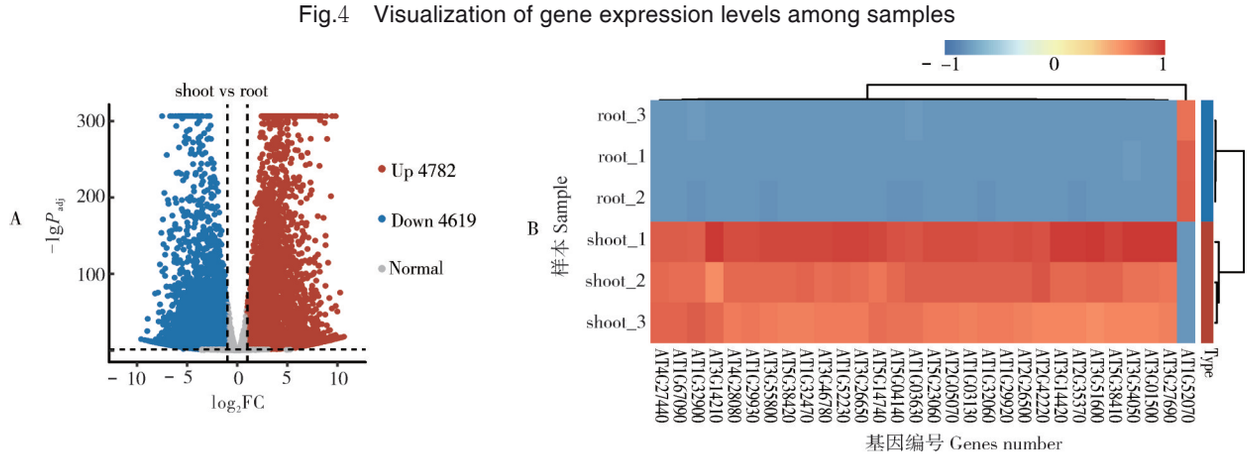
RNApipe使用cor函数与ggplot2包(v3.3.3)将所有样本基因表达谱的成对相关性可视化为热图,相关系数的绝对值越接近1,表明样本之间的相关性越高,这有助于用户进一步识别相关性较低的离群值样本。结果显示拟南芥根部或芽部中各个生物学重复间的相关性较高,而样本间的相关性较差(图4A)。此外,RNApipe使用FactoMineR(v2.4)和factoExtra(v1.0.7)包进行主成分分析(PCA),通过将数据降为二维,再进行K-means聚类和计算欧式距离,距离值越小则表示数据的相似度越高,用户可以据此分离和排除离群值。PCA结果同样显示组内的生物学重复聚类良好,而样本间差异较大(图4B)。RNApipe还支持使用pheatmap包(v1.0.12)展示样本中所有表达基因的聚类图谱,横轴顶部的线图表示



A：热图展示样本间的相关性 The heatmap shows the correlation among samples；B：PCA 主成分分析 PCA principal component analysis；C：箱线图展示样本总体表达量 The boxplot shows the overall expression level of samples；D：累积分布曲线图展示样本总体表达量变化 The cumulative distribution curve shows the changes of expression levels in samples；E：热图根据样本内所有表达基因的相关性进行聚类 The heatmap is clustered according to the correlation of all expressed genes in the samples.

图 4　样本间基因表达量的可视化

Fig.4　Visualization of gene expression levels among samples



A：火山图展示根部和芽部间差异表达的基因 Volcano plot shows differentially expressed genes between root and shoot；B：热图展示样本间变化最大的前 30 个差异基因 The heatmap shows the top 30 differentially expressed genes.

图 5　样本间差异表达基因的可视化

Fig.5　Visualization of differentially expressed genes among samples

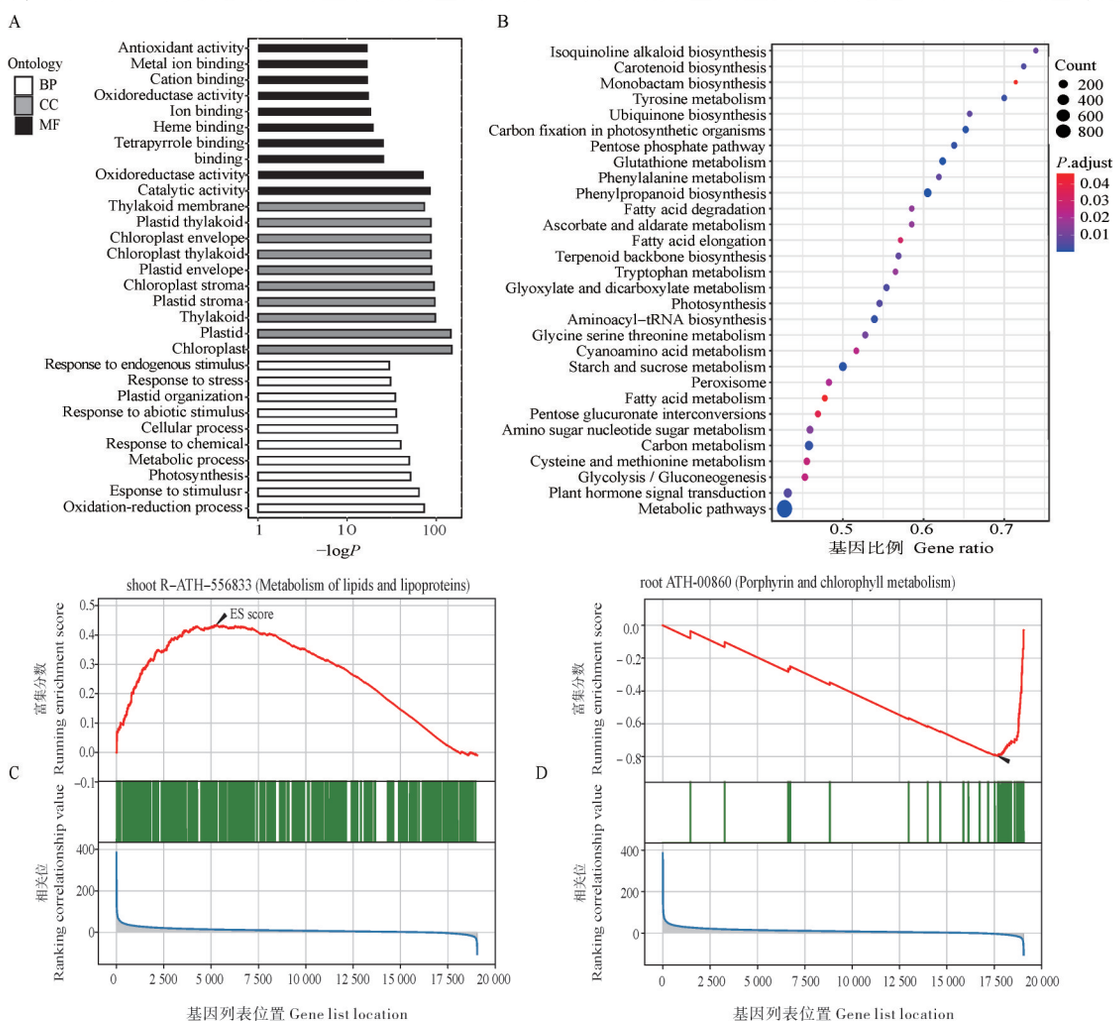
样本的聚类，左侧线图则表示基因的聚类。此外，RNApipe 使用 stat_ecdf 和 geom_boxplot 函数分别绘制累积分布曲线图（图 4C）与箱线图（图 4D）以评估 2 组样本中基因的整体表达水平。同样地，这 2 种分析方法也都表明拟南芥根部的整体基因表达水平显著高于芽部。与图 4A、B 一致，聚类图谱的结果同样表明组内重复的相关性较高，而组间差异较大（图 4E）。同时，该结果还表明拟南芥根部的基因总体表达量高于芽部（图 4E）。

2.4 差异基因的鉴定

RNA-seq下游分析的第一步是鉴定样本间差异表达的基因。RNApipe集成了当下主流的差异表达工具：DESeq2、edgeR、limma。这些工具具有不同的模型和优势（见Documentation.pdf），用户可以根据实验需求在config.yaml中选择其中任一工具进行分析。RNApipe默认使用DESeq2软件在基因水平上执行差异基因的鉴定，DESeq2使用基因表达的原始丰度而不是标准化丰度进行统计检验[28]。DESeq2分析的结果包括：对数变化（logFC）、错误发现率（FDR）以及调整后的P值。图5展示了拟南芥芽部相较于根部显著差异基因（$|\log FC| \geqslant 1$且$P_{adj} < 0.05$）的数目，以及前30个变化最大的显著差异基因的表达量热图。除了可视化图片，输出结果还包括基因表达矩阵和显著差异基因列表，以便科研人员进一步分析和后续验证。

2.5 功能注释

差异基因的功能注释对于了解特定条件所影响的生物学过程至关重要。基因本体论（GO）分析[29]与京都基因和基因组百科全书（KEGG）途径分析[30]是对差异表达基因进行分类的重要工具。RNApipe通过KOBAS-i数据库对差异基因的生物学功能和参与的通路进行注释（图6）。GO分析从分子功能（MF）、生物过程（BP）和细胞成分（CC）3个方面展示，拟南芥根部与芽部间差异基因的GO分类注释，表明差异基因主要与氧化还原和胁迫相应等过程相



A：样本间差异基因的GO功能注释 GO functional annotation of differentially expressed genes that among samples；B：样本间差异基因的KEGG通路分析 KEGG pathway analysis of differentially expressed genes among samples；C：shoot R-ATH-556833的GSEA富集分析 GSEA enrichment analysis of shoot R-ATH-556833；.D：root ATH-00860的GSEA富集分析 GSEA enrichment analysis of shoot root ATH-00860.

图6 样本间基因功能注释的可视化

Fig.6 Visualization of gene functional annotations among samples

关，并显著富集于叶绿体、质体等部位（图 6A）。KEGG 通路分析是探索差异基因的功能和相关通路的另一个重要工具。KEGG 气泡图展示了根部与芽部间差异基因显著富集的前 30 条 KEGG 通路，气泡的大小代表富集到该通路中基因数目的多少，气泡的颜色从红到蓝代表富集程度越来越显著。结果表明差异基因主要富集在代谢和植物激素信号转导等通路（图 6B）。

GO 和 KEGG 富集分析往往侧重于识别两组间差异表达显著的基因，这可能导致遗漏部分差异表达不显著却有重要功能和意义的基因，另外，GO 和 KEGG 分析无法判断差异基因的变化方向是上调还是下调。基因集富集分析（GSEA）不需要指定明确的差异阈值，算法会根据实际数据的整体趋势对基因的表达差异进行排序，随后检验数据库中预先设定的基因集富集在该排序列表的顶端或底端，因此，GSEA 检测到的是基因集而不是单个基因的表达变化。GSEA 图展示了脂质和脂蛋白通路中的基因集的表达在拟南芥的芽中呈现上调的趋势，而叶啉与叶绿素代谢通路中的基因集的表达在根中呈现下调的趋势（图 6C）。

2.6　RNApipe 与其他自动化分析工具的比较

与现有的自动化分析工具对比（表 1），RNApipe 通过 Conda 虚拟环境保证了所有软件的自动安装与运行，通过 Snakemake 工作流程管理系统整合各个分析步骤，保证了数据的可重复性与功能的可扩展性，并通过用户指定输入文件，保证了 RNApipe 适用于所有有参物种的分析。先前的工作流程大多到鉴定差异基因就截止，然而差异基因的功能注释对于生物学问题的研究更具有指导意义，因此 RNApipe 通过 KOBAS-i 数据库进行功能注释。该数据库目前整合了最新的 KEGG 数据库，支持 5 944 个物种的 KEGG 注释和 71 个物种的 GO 注释信息。另外，RNApipe 在各个分析步骤中集成了多款当下主流的软件，并在 log 日志文件中记录同类型软件的运行情况。通过比较运行时间和所耗硬件资源，我们筛选出一套最优软件组合作为 RNApipe 的默认流程（FastQC-fastp-HISAT2-featureCounts-DESeq2-KOBAS-i）。同时，用户也可以根据自己的实验数据，通过修改 config.yaml 中的参数选择非默认软件进行分析。RNApipe 另一重要的特点是在每一分析阶段都提供详细的高质量图表供用户直接使用。

表 1　RNApipe 与其他自动化分析工具的比较

Table 1　Comparison of RNApipe with other automated analysis tools

工作流程 Workflow	质量控制 Quality control	物种 Organism	GO/KEGG 分析 GO/KEGG analysis	GSEA 分析 GSEA analysis	安装方式 Installation	硬件需求 Hardware requirement	编程需求 Programming requirement	文献 References
RNApipe	√	All	√	√	简单 Easy	低 Low	低 Low	本研究 This study
RASflow	√	All	×	×	简单 Easy	低 Low	低 Low	[11]
UTAP	√	5	×	×	简单 Easy	高 High	低 Low	[31]
ARMOR	√	All	×	×	简单 Easy	高 High	低 Low	[10]
VIPER	√	2	√	√	简单 Easy	高 High	低 Low	[4]
BioJupies	×	2	√	×	网站 Web	低 Low	低 Low	[8]
hppRNA	√	2	×	×	中等 Medium	低 Low	中等 Medium	[5]
aRNApipe	√	All	×	×	困难 Hard	高 High	高 High	[32]
RNACocktail	×	All	×	×	困难 Hard	低 Low	高 High	[9]

注 Note：√：支持 Support；×：不支持 Unsupported；All：所有的有参物种 All participating species.

3　讨　论

利用 RNA-seq 数据鉴定差异表达基因为研究重要的科学问题提供了思路，若将整个过程的分析步骤整合为一个自动化、模块化的流程，将极大地方便科研工作者的使用。为此，本研究开发了一个适用于任何有参物种的、流程更加完整的 RNA-seq 自动化分析工作流程：RNApipe。RNApipe 工作流程旨在分析来自任何有参物种的 RNA-seq 数据。目前，RNApipe 已在以下 5 个模式物种的完整数据集中完成测试与评估：人（PRJNA390636；GEO：GSE100075）、小鼠（PRJNA630257；GEO：GSE149838）、拟南芥（PRJNA321304；GEO：GSE81332）、粗糙脉孢菌（PRJ-

NA392079；GEO：GSE100539）、酵母（PRJNA318684；GEO：GSE80357）。分析结果表明，RNApipe运行稳定，功能注释结果符合已发表文献的数据。实际数据集的可视化结果存放于https://github.com/ywu019/RNApipe_pj.git.，与现有的自动化分析工具对比，RNApipe具有安装容易、使用方便、易于扩展、功能丰富等特点。RNApipe将帮助科研工作者以一种简单、高效且可重复的方式从RNA-seq数据中获取有价值的信息。

RNApipe目前主要适用于对二代Illumina测序平台产生的短读长（short-read）数据进行差异基因分析。在RNApipe的基础上，用户可以通过修改rules命令来扩展其他功能，例如可变剪切与单核苷酸变异的检测等。另外，由于常规RNA-seq是对所有细胞的RNA进行测序的技术，因此无法辨别细胞的类型和空间信息。近年来兴起的单细胞转录组测序技术（scRNA-seq）常应用于研究不同组织或器官中（例如正常和癌症）差异基因的表达，为研究单细胞水平的基因表达谱提供了机会。空间转录组测序技术通过将成像技术与scRNA-seq相结合，可以绘制出转录本在组织中表达的位置，进一步提高组织分辨率。scRNA-seq技术与空间转录组学技术的普及同时伴随着研究人员在数据分析方面面临的挑战，RNApipe的结构设计则可以为scRNA-seq自动化数据分析工具的构建提供参考。

参考文献 References

[1] STARK R, GRZELAK M, HADFIELD J. RNA sequencing: the teenage years[J]. Nature reviews: genetics, 2019, 20(11): 631-656.

[2] NAGALAKSHMI U, WANG Z, WAERN K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing[J]. Science, 2008, 320(5881): 1344-1349.

[3] CONESA A, MADRIGAL P, TARAZONA S, et al. A survey of best practices for RNA-seq data analysis[J/OL]. Genome biology, 2016, 17: 13[2022-01-13]. https://doi.org/10.1186/s13059-016-0881-8.

[4] CORNWELL M, VANGALA M, TAING L, et al. VIPER: visualization pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis[J/OL]. BMC bioinformatics, 2018, 19(1): 135[2022-01-13]. https://doi.org/10.1186/s12859-018-2139-9.

[5] WANG D P. hppRNA—a Snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples[J]. Briefings in bioinformatics, 2017, 19(4): 622-626.

[6] MONIER B, MCDERMAID A, WANG C, et al. IRIS-EDA: an integrated RNA-Seq interpretation system for gene expression data analysis[J/OL]. PLoS computational biology, 2019, 15(2): e1006792[2022-01-13]. https://doi.org/10.1371/journal.pcbi.1006792.

[7] MARINI F, LINKE J, BINDER H. Ideal: an R/Bioconductor package for interactive differential expression analysis[J/OL]. BMC bioinformatics, 2020, 21(1): 565[2022-01-13]. https://doi.org/10.1186/s12859-020-03819-5.

[8] TORRE D, LACHMANN A, MA'AYAN A. BioJupies: automated generation of interactive notebooks for RNA-seq data analysis in the cloud[J]. Cell systems, 2018, 7(5): 556-561.

[9] SAHRAEIAN S M E, MOHIYUDDIN M, SEBRA R, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis[J/OL]. Nature communications, 2017, 8: 59[2022-01-13]. https://doi.org/10.1038/s41467-017-00050-4.

[10] ORJUELA S, HUANG R Z, HEMBACH K M, et al. ARMOR: an automated reproducible MOdular workflow for preprocessing and differential analysis of RNA-seq data[J]. G3 Genes|Genomes|Genetics, 2019, 9(7): 2089-2096.

[11] ZHANG X, JONASSEN I. RASflow: an RNA-seq analysis workflow with snakemake[J/OL]. BMC bioinformatics, 2020, 21(1): 110[2022-01-13]. https://doi.org/10.1186/s12859-020-3433-x.

[12] GRÜNING B, DALE R, SJÖDIN A, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences[J]. Nature methods, 2018, 15(7): 475-476.

[13] WRATTEN L, WILM A, GÖKE J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers[J]. Nature methods, 2021, 18(10): 1161-1168.

[14] KÖSTER J, RAHMANN S. Snakemake: a scalable bioinformatics workflow engine[J]. Bioinformatics (Oxford, England), 2012, 28(19): 2520-2522.

[15] CHEN S, ZHOU Y, CHEN Y, et al. Fastp: an ultra-fast all-in-one FASTQ preprocessor[J/OL]. Biochemistry and biophysics reports, 2018, 34(17): i884-i890[2022-01-13]. https://doi.org/10.1093/bioinformatics/bty560.

[16] MARTIN M. Cutadapt removes adapter sequences from high-throughput sequencing reads[J]. EMBnet journal, 2011, 17(1): 10-12.

[17] BROWN J, PIRRUNG M, MCCUE L A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool[J]. Bioinformatics (Oxford, England), 2017, 33(19): 3137-3139.

[18] EWELS P, MAGNUSSON M, LUNDIN S, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report[J]. Bioinformatics, 2016, 32(19): 3047-3048.

[19] KIM D, LANGMEAD B, SALZBERG S L. HISAT: a fast spliced aligner with low memory requirements[J]. Nature methods, 2015, 12(4): 357-360.

[20] DOBIN A, DAVIS C A, SCHLESINGER F, et al. STAR: ultrafast universal RNA-seq aligner[J]. Bioinformatics, 2012, 29(1): 15-21.

[21] OKONECHNIKOV K, CONESA A, GARCÍA-ALCALDE F. Qualimap 2: advanced multi-sample quality control for high-

throughput sequencing data[J]. Bioinformatics, 2015, 32(2): 292-294.

[22] LIAO Y, SMYTH G K, SHI W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features[J]. Bioinformatics, 2013, 30(7): 923-930.

[23] ANDERS S, PYL P T, HUBER W. HTSeq: a Python framework to work with high-throughput sequencing data[J]. Bioinformatics, 2014, 31(2): 166-169.

[24] LOVE M I, HUBER W, ANDERS S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J/OL]. Genome biology, 2014, 15(12): 550[2022-01-13]. https://doi.org/10.1186/s13059-014-0550-8.

[25] ROBINSON M D, MCCARTHY D J, SMYTH G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data[J]. Bioinformatics (Oxford, England), 2010, 26(1): 139-140.

[26] RITCHIE M E, PHIPSON B, WU D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies[J/OL]. Nucleic acids research, 2015, 43(7): e47[2022-01-13]. https://doi.org/10.1093/nar/gkv007.

[27] BU D C, LUO H T, HUO P P, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis[J/OL]. Nucleic acids research, 2021, 49(W1): W317-W325[2022-01-13]. https://doi.org/10.1093/nar/gkab447.

[28] SONESON C, DELORENZI M. A comparison of methods for differential expression analysis of RNA-seq data[J/OL]. eLife, 2013, 14: 91[2022-01-13]. https://doi.org/10.1186/1471-2105-14-91.

[29] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene Ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25-29.

[30] OGATA H, GOTO S, SATO K, et al. KEGG: Kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 1999, 27(1): 29-34.

[31] KOHEN R, BARLEV J, HORNUNG G, et al. UTAP: user-friendly transcriptome analysis pipeline[J]. BMC bioinformatics, 2019, 20(1): 154[2022-01-13]. https://doi.org/10.1186/s12859-019-2728-2.

[32] ALONSO A, LASSEIGNE B N, WILLIAMS K, et al. aRNApipe: a balanced, efficient and distributed pipeline for processing RNA-seq data in high-performance computing environments[J]. Bioinformatics, 2017, 33(11): 1727-1729.

RNApipe: automated analyses of RNA-seq data based on Snakemake

WU Le[1], LI Yi'nan[2], KONG Dexin[1], ZHOU Zhipeng[2]

1. *College of Informatics, Huazhong Agricultural University, Wuhan* 430070, *China*;
2. *College of Life Science and Technology, Huazhong Agricultural University, Wuhan* 430070, *China*

Abstract Transcriptome sequencing technology (RNA-seq) has been widely used in the field of basic scientific studies, but the bioinformatics analysis of RNA-seq data places high requirements on the programming ability of researchers. In order to enable researchers to analyze RNA-seq data simply and efficiently, this article constructed an automated and modular workflow-RNApipe (On Github: https://github.com/ywu019/RNApipe.git) based on the Snakemake workflow management system and Conda environment manager. RNApipe can automatically conduct quality control, alignment, quantification, identification of differential genes, and functional annotation analyses including GO, KEGG, and GSEA with RNA-seq data from any species with a reference genome. The results of analysis in each step are presented in high-quality visualizations or reports, and important output files are preserved. RNApipe has been tested and evaluated in multiple model species. The results showed that RNApipe can run smoothly and the results of annotation are accurate. Compared with the existing pipelines of automated analysis, the main features of RNApipe include (i) the workflow is relatively complete, (ii) the default tools consume less time and resources, (iii) applicable to any parametric species, (iv) comprehensive visualization, and (v) user-friendliness (easy to install, use, and expand). The features of RNApipe mentioned above allow researchers to quickly obtain essential information from large-scale RNA-seq sequencing data.

Keywords RNA-seq; differential expression; functional annotation; Snakemake; Conda; automated data analysis; visualization

（责任编辑：陆文昌）