DOI:10.12113/202409009

基于 Linux 系统构建 lncRNA-seq 生物信息学本地分析平台

黄平平,王佳容,单亚男,周百灵*
山东省生物物理重点实验室(德州学院生物物理研究院),山东 德州 253023

摘　要:长非编码 RNA(Long noncoding RNA, lncRNA)是一类转录本长度大于 200 nt、缺乏蛋白质编码能力的 RNA 分子,可通过与 mRNA、蛋白质和 miRNA 等分子相互作用从而以多种形式在不同水平调控基因表达。近年来,lncRNA 在人类疾病发生发展、动物生长发育和植物应对环境胁迫等方面的研究一直是国内外研究热点和持续关注的焦点。目前,关于人类、动植物 lncRNA 的研究主要采用 lncRNA 测序(lncRNA-sequencing, lncRNA-seq)的方法。lncRNA-seq 数据的分析依赖众多软件,且大部分是在 Linux 系统通过命令行的方式运行,对于大多数研究者开展数据分析可能存在困难。因此,文中在调研现有的 lncRNA-seq 数据分析流程的基础上,基于 Linux 系统构建 lncRNA-seq 本地分析平台,平台主要包括以下几个步骤:首先采用 trimmomatic 对原始测序数据进行质量控制,使用软件 Hisat2 和 StringTie 进行序列的比对、组装,通过 Salmon 软件进行转录本的定量,结合 CPC2、CNCI、Pfam 和 PLEK 对新的转录本进行编码能力预测。另外,以 2 组马铃薯测序数据为例详细介绍数据分析过程,以便同领域研究者在服务器上构建本地化分析平台、灵活地开展数据分析。

关键词:长非编码 RNA;分析平台;Linux;生物信息学
中图分类号:Q52　　**文献标志码**:A　　**文章编号**:1672-5565(2025)04-323-07

Construction of localized lncRNA-seq bioinformatics analysis platform based on the Linux system

HUANG Pingping, WANG Jiarong, SHAN Yanan, ZHOU Bailing*
(*Shandong Provincial Key Laboratory of Biophysics*, *Institute of Biophysics*, *Dezhou University*, *Dezhou* 253023, *Shandong*, *China*)

Abstract: Long noncoding RNA (lncRNA) is a type of transcript with length longer than 200 nt and lacks protein coding ability. It can regulate the gene expression in various forms at different levels by interacting with mRNA, protein, and miRNA. In recent years, the studies of lncRNA in the occurrence and development of human diseases, animal growth and development, and plant response to environmental stress have attracted considerable attention. At present, the researches on the human, animal and plants lncRNAs are mainly performed based on the lncRNA-sequencing (lncRNA-seq) method. However, the analysis of lncRNA-seq data depends on many software and most of them are run in the Linux system, which might be difficult for most of researchers. Therefore, the paper investigates the existing lncRNA-seq data analysis process and constructs the localized lncRNA-seq analysis platform based on the Linux system. The platform mainly includes the following steps: first, the quality of the original sequencing data is controlled by the trimmomatic software and obtain the clean data; then, the clean data are mapped to the reference genome by hisat2 and the transcripts are assembled by stringtie; further, the quantification of the transcript is performed by salmon; the coding ability of the new transcript is predicted by four software,

收稿日期:2024-09-09;**修回日期**:2024-10-23; **网络首发日期**:2024-11-26.
网络首发地址:https://link.cnki.net/urlid/23.1513.Q.20241126.1430.002
基金项目:山东省自然科学基金青年项目(No. ZR2021QF143);德州学院人才引进项目(No. 2020xjrc216);德州学院横向课题(No. HXKT2023211);山东省教育厅高等学校人才引育计划.
*通信作者:周百灵,女,副教授,研究方向:生物信息学和机器学习.E-mail: blzhou@126.com.
引用格式:黄平平,王佳容,单亚男,等. 基于 Linux 系统构建 lncRNA-seq 生物信息学本地分析平台[J].生物信息学,2025,23(4):323-329.
HUANG Pingping, WANG Jiarong, SHAN Yanan, et al. Construction of localized lncRNA-seq bioinformatics analysis platform based on the Linux system [J]. Chinese Journal of Bioinformatics, 2025, 23(4): 323-329.

including CPC2, CNCI, Pfam and PLEK. In addition, two groups of potato sequencing data are taken as an example to introduce the data analysis process in detail, which can help the related researchers build a localized analysis platform on the server and carry out data analysis flexibly.

Keywords: Long noncoding RNA; Analysis platform; Linux; Bioinformatics

长非编码 RNA(Long noncoding RNA, lncRNA)是一类长度为 200~10 000 nt 之间的 RNA,在调控基因表达和多种生物学过程中发挥重要作用。鉴于 lncRNA 的功能重要性,lncRNA 相关研究为国内外持续研究热点领域[1]。普通转录组测序(RNA-sequencing, RNA-seq)方法基于 oligo(dT)捕获技术,主要利用 mRNA 通常都带有 poly(A)尾巴的特性,将 mRNA 和部分 lncRNA 与总 RNA 分离。因此基于公共数据库中海量的普通转录组测序数据,越来越多的动植物 lncRNA 被鉴定出来。近年来,关于人类、动植物 lncRNA 的研究主要采用 lncRNA 测序(lncRNA-sequencing, lncRNA-seq)的方法,该测序方法采用去核糖体链特异性建库技术。与普通 RNA-seq 技术相比,lncRNA-seq 方法检获的 lncRNA 更为全面,可同时揭示大量无 poly(A)结构的 lncRNA。但从 lncRNA/RNA-seq 数据中挖掘 lncRNA 需依赖众多软件、转件安装复杂,且大部分是在 Linux 系统通过命令行的方式运行,而大多数实验生物学家可能对 Linux 系统并不熟悉。

目前,国内外存在部分针对 RNA-seq 数据分析的本地和可视化分析平台,如 Galaxy 和 BioJupies[2-3]。Galaxy 作为生物医学领域最受欢迎的在线生物信息分析工具之一包含多种分析软件,如在 tools 下选择 RNA-seq 中的 Hisat2 和 Stringtie 可分别进行序列比对和组装。但是该网站仅含单个 lncRNA 编码能力预测软件,而为了保证数据的可靠性,对 lncRNA 编码能力的预测往往需要结合多个软件同时分析。BioJupies 分析平台可使研究者通过直观的界面自动化挖掘 GEO 测序数据,也可提交原始测序数据进行分析。但是,BioJupies 分析平台要求原始数据文件的大小在 5 GB 以下,同时不涉及 lncRNA 相关分析。

目前也存在部分流程可从 RNA-seq 数据中鉴定 lncRNA,如 UClncR 采用 Tophat/Hisat2 比对序列、Stringtie 进行转录本组装,进一步预测候选 lncRNA,但该流程仅针对人类数据[4]。LncPipe 流程可完成原始转录组数据到 lncRNA 鉴定的分析,该流程仅适用于人类、小鼠、果蝇和斑马鱼[5]。可见,目前已存在的从 lncRNA/RNA-seq 数据中鉴定 lncRNA 的流程均存在一定的局限性,如仅可应用于特定物种。因此,本论文拟结合应用案例,基于 Linux 系统构建从 lncRNA/RNA-seq 数据中挖掘 lncRNA 的本地化分析平台,能够从根本上解决实验生物学家生物信息学数据分析的难题,帮助同领域研究者在服务器构建本地化分析平台,灵活地开展数据分析。

1 系统设计

1.1 数据质量控制

测序原始数据可能包含低质量序列,影响后续数据分析,因此需采用质量控制软件删除低质量序列,同时删除测序过程中人为添加的接头和引物序列。目前 lncRNA/RNA-seq 原始数据的质量控制广泛采用 Trimmomatic 和 fastp 软件处理[6-7]。其中 Trimmomatic 软件的引用次数高达 25 000 余次,本论文采用该软件进行原始数据质控,如图 1。

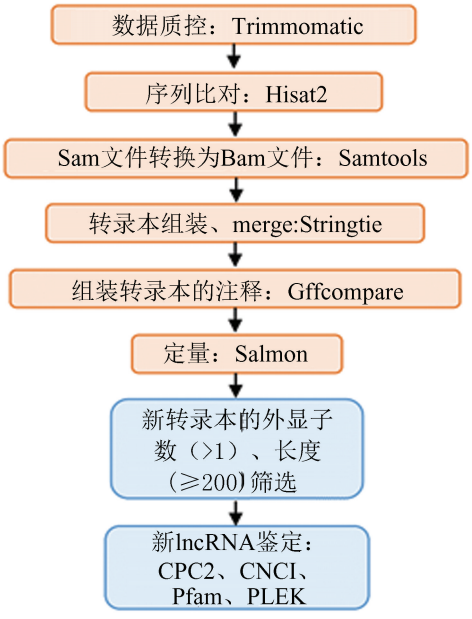

图 1 lncRNA 数据分析流程图

Fig.1 Workflow chart of lncRNA analysis

1.2 序列比对和组装

筛选快速准确的比对、组装软件。lncRNA-seq 针对有参考基因组的样本进行,因此从 lncRNA/RNA-seq 数据中挖掘 lncRNA 需将序列比对到参考基因组,从而确定序列起始位点。目前序列比对主流软件包括 TopHat2、STAR 和 Hisat2。TopHat2 软

件比较耗时,可能需要花费几天的时间来处理单个测序样本[8]。STAR 软件使用后缀数组,处理速度比 TopHat2 快。然而,与 Burrows-Wheeler 算法相比,后缀数组方法对内存要求较高[9]。较之以上两款软件,Hisat2 运行速度最快、需内存低,精确度也较好[10]。就组装软件而言,目前 StringTie 和 Cufflinks 使用比较多,与 Cufflinks 相比,StringTie 运行速度快、具有很高的准确性和新型异构体的发现能力,对于有多重异构体的物种更擅长[11]。鉴于以上原因,本文选择 Hisat2 进行序列比对,StringTie 进行转录本组装。

1.3　转录本定量和编码能力预测

转录本定量软件包括 kallisto、eXpress、Sailfish 和 Salmon 等[12]。鉴于 Salmon 软件具有准确度高、运行速度快、所需内存低等优点,该软件广泛应用于 lncRNA/RNA-seq 数据转录本的定量。

随后需要筛选 lncRNA 鉴定/编码能力预测软件。近年来研究者们开发出多种 lncRNA 鉴定软件,但是不同软件所获得的结果不一致。因此,需结合多种软件提高 lncRNA 鉴定的准确性。目前,广泛使用的转录本编码能力预测软件包括 CPC2[13]、CNCI[14]、Pfam[15]和 PLEK[16]等。本文结合该四种软件来预测转录本的编码能力。

2　应用

2.1　分析系统及案例数据来源

分析平台基于 Linux 系统 CentOS 7(Version:3.10.0-1160.el7.x86_64)操作系统构建。

应用案例采用 NCBI 数据库中野生型马铃薯和转基因抗病马铃薯分别感染致病疫霉(*Phytophthora infestans*)24 h 后的转录组测序数据(PRJNA318049),数据采用的测序方法为双末端测序,测序平台为 Illumina HiSeq 2000。其中,数据 SRR3371859、SRR3371860 和 SRR3371861 为野生型马铃薯叶片感染致病疫霉 24 h 后的转录组测序数据,数据 SRR3371864、SRR3371865 和 SRR3371866 为转基因抗病马铃薯叶片进行相同处理后的数据[17]。马铃薯基因组文件下载链接:http://spuddb.uga.edu/data/DM_1-3_516_R44_potato_genome_assembly.v6.1.fa.gz;马铃薯基因注释文件下载链接:http://spuddb.uga.edu/data/DM_1-3_516_R44_potato.v6.1.hc_gene_models.gff3.gz。

2.2　软件安装

分析平台主要涉及的软件如表 1 所示,软件的安装主要采用 Anaconda,基于 Anaconda 的 bioconda 通道可方便快捷地一键安装绝大多数 lncRNA-seq 生物信息学分析软件,包括 hisat2、stringtie、gffcompare、samtools、salmon、gffread、faSomeRecords、PLEK 和 trimmomatic 等,可极大程度地简化安装过程中软件的下载和环境配置。Anaconda 安装的命令在 CSDN 博客上面有很多详细教程可供学习。Anaconda 安装完成后,可先创建 RNAseq 环境,将上述软件安装在相同环境下,命令如下(括号内为注释内容,$ 符号代表一条命令的起始):

表 1　所用分析软件及其功能和来源

Table 1　Analysis software used and its functions and sources

软件名称	功能描述	软件地址	参考文献
Anaconda	软件安装	https://www.anaconda.com/download	-
sratoolkit	格式转换	https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit	-
gffread	格式转换	https://github.com/gpertea/gffread	[18]
samtools	格式转换	https://github.com/samtools/samtools	[19]
faSomeRecords	序列提取	https://github.com/santiagosnchez/faSomeRecords	-
Trimmomatic	质量控制	http://www.usadellab.org/cms/index.php? page=trimmomatic	[6]
Hisat2	序列比对	https://www.ccb.jhu.edu/software/hisat/index.shtml	[10]
StringTie	转录本组装	https://ccb.jhu.edu/software/stringtie	[11]
gffcompare	组装转录本评估	https://github.com/gpertea/gffcompare	[20]
Salmon	转录本定量	https://github.com/COMBINE-lab/Salmon	[12]
CPC2	转录本编码能力预测	http://cpc2.cbi.pku.edu.cn/; https://github.com/gao-lab/CPC2_standalone	[13]
CNCI	转录本编码能力预测	https://github.com/www-bioinfo-org/CNCI	[14]
Pfam	转录本编码能力预测	http://pfam.xfam.org	[15]
PLEK	转录本编码能力预测	https://sourceforge.net/projects/plek/files	[16]

$ conda create -n RNAseq(创建名为 RNAseq 的环境)

$ conda activate RNAseq(激活 RNAseq 环境)

$ conda install -c bioconda hisat2(安装 hisat2 软件)

在使用软件之前,需要先激活相应的环境,如使用 hisat2 软件之前,需要先激活 RNAseq 环境(命令为:conda activate RNAseq),针对某种软件也可单独创建环境,如:

$ conda create -n stringtie

$ conda activate stringtie

$ conda install -c bioconda stringtie

少部分软件无法通过 conda 安装,可通过表 1 中的链接进行下载,然后在服务器进行安装。例如 sratoolkit 软件的下载与安装如下:

$ mkdir sratoolk(创建名为 sratoolk 的文件夹)

$ cd sratoolk(进入 sratoolk 目录)

$ curl -O https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-centos_linux64.tar.gz(下载当前最新 Linux 版本的 sratoolkit 软件)

$ tar xvf sratoolkit.current-centos_linux64.tar.gz(文件解压缩)

$ cd sratoolkit.3.0.1-centos_linux64/bin(进入 bin 目录)

$ echo export PATH = $ PATH:/../sratoolk/sratoolkit.3.0.1-centos_linux64/bin >> ~/.bashrc(配置环境,".."为 sratoolk 上级详细路径)

$ cd ..(返回上一目录)

$ source ~/.bashrc(配置环境)

2.3 数据分析流程

为方便数据分析,可建立一个专用文件夹(如 localplatform)用于整个分析流程所产生的数据的存储。

第一步,原始测序数据的准备,基因组和注释文件的下载:

$ wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos5/sra-pub-zq-14/SRR003/371/SRR3371859.sralite.1(下载原始测序数据)

$ wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos5/sra-pub-zq-14/SRR003/371/SRR3371860.sralite.1

$ wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos5/sra-pub-zq-11/SRR003/371/SRR3371861/SRR3371861.sralite.1

$ wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos5/sra-pub-zq-11/SRR003/371/SRR3371864/SRR3371864.sralite.1

$ wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos5/sra-pub-zq-11/SRR003/371/SRR3371865/SRR3371865.sralite.1

$ wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos5/sra-pub-zq-14/SRR003/371/SRR3371866.sralite.1

$ wget http://spuddb.uga.edu/data/DM_1-3_516_R44_potato_genome_assembly.v6.1.fa.gz(下载基因组文件)

$ wget http://spuddb.uga.edu/data/DM_1-3_516_R44_potato.v6.1.hc_gene_models.gff3.gz(下载注释文件)

$ gzip -d DM_1-3_516_R44_potato_genome_assembly.v6.1.fa.gz(基因组文件解压缩)

$ mv DM_1-3_516_R44_potato_genome_assembly.v6.1.fa PotatoGenome.fa(将基因组文件重命名为 PotatoGenome.fa)

$ gzip -d DM_1-3_516_R44_potato.v6.1.hc_gene_models.gff3.gz(注释文件解压缩)

$ mv DM_1-3_516_R44_potato.v6.1.hc_gene_models.gff3 Potatogenome.gff3(注释文件重命名)

$ gffread Potatogenome.gff3 -T -o Potatogenome.gtf(注释文件格式转换)

第二步,分别采用 sratoolkit 软件将下载的原始测序数据转化为 fastq 格式、trimmomatic 软件对转化后的双端测序数据进行质量控制,六个数据分别运行以下两条命令:

$ /../sratoolk/sratoolkit.3.0.1-centos_linux64/bin/fastq-dump /localplatform/ SRR3371859.sralite.1 --split-3

$ trimmomatic PE -phred33 -threads 10 /localplatform/SRR3371859.sralite.1_1.fastq /localplatform/SRR3371859.sralite.1_2.fastq /localplatform/SRR3371859.sralite.1_1.clean.fastq /localplatform/SRR3371859.sralite.1_1.unpaired.fastq /localplatform/SRR3371859.sralite.1_2.clean.fastq /localplatform/SRR3371859.sralite.1_2.unpaired.fastq AVGQUAL:30 1>sample1.QC.log 2>&1

第三步,构建 hisat2 索引,采用 hisat2 软件分别将 6 个质控后的数据比对到参考基因组;通过 samtools 软件将文件格式由 sam 转化为 bam:

$ hisat2_extract_splice_sites.py Potatogenome.gtf > Potatogenome.ss(构建索引)

$ hisat2_extract_exons.py Potatogenome.gtf > Potatogenome.exon(构建索引)

$ hisat2-build -p 16 --exon Potatogenome.exon --ss Potatogenome.ss PotatoGenome.fa genome_tran (构建索引)

$ hisat2 --dta -x /localplatform/genome_tran -1 /localplatform/SRR3371859.sralite.1_1.clean.fastq -2 /localplatform/SRR3371859.sralite.1_2.clean.fastq -S /localplatform/SRR3371859.sam(序列比对)

$ samtools sort -@ 8 -o /localplatform/SRR3371859.bam /localplatform/SRR3371859.sam (格式转化)

第四步,采用 stringtie 软件基于以下命令分别对 6 个样本进行转录本的组装;然后对所有样本的转录本进行合并,其中 mergelist.txt 文件为六个样本组装后的 gtf 文件名称:

$ stringtie -p 8 -G /localplatform/Potatogenome.gtf -o /localplatform/SRR3371859.gtf -l SRR3371859 /localplatform/SRR3371859.bam(转录本组装)

$ stringtie --merge -p 16 -G /localplatform/Potatogenome.gtf -o /localplatform/localplatformstringtie_merged.gtf /localplatform/mergelist.txt(转录本合并)

第五步,采用 gffcompare 软件评估组装后的转录本与注释文件的匹配情况:

$ gffcompare -r/localplatform/Potatogenome.gtf-G-o localplatformmerged/localplatform/localplatformstringtie_merged.gtf

第六步,采用 salmon 软件对六个样本分别进行转录本的定量,通过该步获得的每个转录本在六个样本中的表达量文件(localplatform_Transcript.count)可用于转录本的差异表达分析:

$ gffread -w /localplatform/localplatform_total_ense.fa -g /localplatform/PotatoGenome.fa/localplatform/localplatformmerged.annotated.gtf(获取所有转录本的序列)

$ salmon index -t /localplatform/localplatform_total_ense.fa -i /localplatform/localplatform_transcripts_index(构建索引)

$ salmon quant -i /localplatform/localplatform_transcripts_index --validateMappings -l A -1 /localplatform/SRR3371859.sralite.1_1.clean.fastq -2 /localplatform/SRR3371859.sralite.1_2.clean.fastq -o /localplatform/SRR3371859.quant(定量)

$ salmon quantmerge --quants /localplatform/*.quant --column NumReads -o /localplatform/localplatform_Transcript.count(将 6 个样本的定量结果合并在一个文件)

第七步,筛选新的转录本,获取其外显子和长度信息:

$ awk'$ 2 == 0 && $ 3 == 0 && $ 4 == 0 && $ 5 == 0 && $ 6 == 0&& $ 7 == 0 {next} 1' localplatform_Transcript.count > localplatform_Transcript_count2(删除在六个样本中表达量均为 0 的转录本)

$ awk'$ 1 ! ~ /Soltu/' localplatform_Transcript_count2 > temp.txt && mv temp.txt localplatform_Transcript_count3(筛选新的转录本)

$ awk'NR>1 {print $ 1}' localplatform_Transcript_count3 > first_column_localplatform_Transcript_count3(获取新的转录本的 ID)

$ awk'NR>1 {print $ 5, $ 6, $ 10}' localplatformmerged.localplatformstringtie_merged.gtf.tmap>localplatformstringtie_merged.gtf.tmap2(获取所有转录本的 ID、外显子和长度信息)

结合文件 first_column_localplatform_Transcript_count3 和 localplatformstringtie_merged.gtf.tmap2,获取新转录本的外显子、长度信息,保存在文件 Novel_transcript_exon_length 中,脚本如下:

$ awk'{print $ 1","}' first_column_localplatform_Transcript_count3 > first_column_localplatform_Transcript_count3_2

$ awk'{print $ 1",", $ 2, $ 3}' localplatformstringtie_merged.gtf.tmap2 > localplatformstringtie_merged.gtf.tmap2_2

$ grep -Ff first_column_localplatform_Transcript_count3_2 localplatformstringtie_merged.gtf.tmap2_2 | awk '{print $ 0}' > test.txt

$ awk'{$ 1=$ 1; gsub(/,$ /, "", $ 1); print}' test.txt > Novel_transcript_exon_length

第八步,筛选外显子数大于 1、长度大于等于 200 的新转录本(Novel_transcript),并且获取其序列信息(Novel_transcript.fa),进行后续蛋白质编码能力预测:

$ awk'($ 2>1 && $ 3>=200) {print $ 0}' Novel_transcript_exon_length > Novel_transcript_exon_length_2

$ awk'{print $ 1}' Novel_transcript_exon_length_2 > Novel_transcript

$ faSomeRecords localplatform_total_ense.fa Novel_transcript Novel_transcript.fa

第九步,基于不同软件对外显子数和长度满足要求的新转录本进行蛋白质编码能力预测:

(1)采用 CPC2 软件进行蛋白质编码能力预测:

$ mkdir cpc2result

```
$ CPC2.py -i /localplatform/Novel_transcript.fa -o /localplatform/cpc2result/Novel_transcript_cpc2
```

根据结果文件 Novel_transcript_cpc2.txt 中的“coding_probability”列可查看每个转录本的编码能力。

(2)采用 CNCI 软件进行蛋白质编码能力预测:

```
$ mkdir CNCIresult
$ python CNCI.py -f /localplatform/Novel_transcript.fa -o /localplatform/CNCIresult/CNCI_novel_transcript_result -m pl -p 50
```

根据结果文件 CNCI.index 中的“index”列可查看每个转录本的编码能力。注:若分析物种为动物,则指定模型为“ve”,即“-m ve”

(3)采用 Pfam 软件进行蛋白质编码能力预测:

```
$ mkdir Pfamresult
$ transeq/ localplatform/Novel_transcript.fa / localplatform/Pfamresult/Novel_transcript_AA.fa -frame=1
$ pfam_scan.pl -fasta /localplatform/Pfamresult/Novel_transcript_AA.fa -dir /PfamData/ -out /localplatform/Pfamresult/Pfam_novel_transcript_result
```

结果文件 Pfam_novel_transcript_result“seq id”列中存在的转录本具备蛋白质编码能力,因此筛选“seq id”列中不存在的转录本为潜在 lncRNA。

(4)采用 PLEK 软件进行蛋白质编码能力预测:

```
$ mkdir PLEKresult
$ python PLEK.py -fasta /localplatform/Novel_transcript.fa -out /localplatform/PLEKresult/PLEK_novel_transcript_result -thread 30
```

结果文件 PLEK_novel_transcript_result 第一列显示为“Non-coding”则表明对应的转录本为 lncRNA。

综上,通过以上九步分析流程可以获得转录本的表达量信息,以及基于四款软件的 lncRNA 预测结果。基于以上结果,研究者可采用 R 语言进行转录本差异表达分析,筛选不同样本间差异表达的 lncRNA,以及进行 lncRNA 共表达靶基因预测等。基于 R 语言的下游数据分析流程较为简单,不依赖服务器即可完成,因此本文不再进行展示。

3 讨论与展望

lncRNA 作为人、动植物体内重要的调控因子,广泛地参与不同的生物学过程。近年来关于 lncRNA 的研究依然为持续研究热点,通过 lncRNA 相关关键词在 PubMed 数据库中检索共检获 58 433 篇文章(截止到 2024.7),近五年来的年文章数量均超过 6 000 余篇。目前,通过 lncRNA/RNA-seq 方法鉴定、筛选差异表达的 lncRNA 为揭示重要科学问题提供研究思路,但是大多数研究者对 lncRNA 分析流程并不熟悉。

文中整合已有软件,基于 Linux 系统构建了一套可以从 lncRNA/RNA-seq 数据中挖掘 lncRNA 的本地化分析平台。平台采用 CPC2[13]、CNCI[14]、Pfam[15]和 PLEK[16]四种软件进行转录本编码能力预测,提高 lncRNA 鉴定的准确性。该四种软件均适用于动植物 lncRNA 的鉴定,因此本平台可同时应用于动植物 lncRNA 的研究。另外,以马铃薯转录组数据为应用案例,详细展示数据分析流程,包括质量过滤、序列比对、组装、注释、定量和转录本编码能力预测等每一步所采用软件的使用方法以及详细的运行脚本。同时,数据分析过程中涉及到文本的处理,比如提取转录本 ID 等,因此本文在数据分析模块也展示了常见的文本处理命令的使用方法(如 awk 文本处理工具),可帮助相关研究者在本地服务器构建数据分析平台,灵活开展数据分析。

文中所构建的从 lncRNA/RNA-seq 数据鉴定 lncRNA 的本地化分析平台也存在一定的局限性。主要体现在需要研究者具备初级的 Linux 操作基础,能够在服务器进行简单的脚本运行。

参考文献(References)

[1]ZHOU Bailing, JI Baohua, SHEN Congcong, et al. EVlncRNAs 3.0: An updated comprehensive database for manually curated functional long non-coding RNAs validated by low-throughput experiments[J]. Nucleic Acids Research, 2024, 52(D1): D98-D106. DOI: 10.1093/nar/gkad1057.

[2]JALILI V, AFGAN E, GU Qiang, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update[J]. Nucleic Acids Research, 2020, 48(W1): W395-W402. DOI: 10.1093/nar/gkaa434.

[3]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. DOI: 10.1016/j.cels.2018.10.007.

[4]SUN Zhifu, NAIR A, CHEN Xianfeng, et al. UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq[J]. Scientific Reports, 2017, 7: 14196. DOI: 10.1038/s41598-017-14595-3.

[5]ZHAO Qi, SUN Yu, WANG Dawei, et al. LncPipe: A Nextflow-based pipeline for identification and analysis of long non-coding RNAs from RNA-Seq data[J]. Journal of Genetics and Genomics, 2018, 45(7): 399-401. DOI: 10.1016/j.

jgg.2018.06.005.

[6]BOLGER AM, LOHSE M, USADEL B. Trimmomatic: A flexible trimmer for Illumina sequence data[J]. Bioinformatics, 2014,30(15):2114-2120. DOI: 10.1093/bioinformatics/btu170.

[7]CHEN Shifu,ZHOU Yanqing, CHEN Yaru, et al. Fastp: An ultra-fast all-in-one FASTQ preprocessor[J]. Bioinformatics, 2018,34(17):i884-i890. DOI: 10.1093/bioinformatics/bty560.

[8]KIM D,PERTEA G, TRAPNELL C, et al. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions[J]. Genome Biology, 2013,14(4):R36. DOI: 10.1186/gb-2013-14-4-r36.

[9]DOBIN A,DAVIS C A, SCHLESINGER F, et al. STAR: Ultrafast universal RNA-seq aligner[J]. Bioinformatics, 2013,29(1):15-21. DOI: 10.1093/bioinformatics/bts635.

[10]KIM D, LANGMEAD B, SALZBERG S L. HISAT: A fast spliced aligner with low memory requirements[J]. Nature Methods,2015,12(4):357-360. DOI: 10.1038/nmeth.3317.

[11]PERTEA M, PERTEA G M, ANTONESCU C M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads[J]. Nature Biotechnology, 2015, 33(3):290-295. DOI: 10.1038/nbt.3122.

[12]PATRO R, DUGGAL G, LOVE M I, et al. Salmon provides fast and bias-aware quantification of transcript expression[J]. Nature Methods, 2017, 14(4):417-419. DOI: 10.1038/nmeth.4197.

[13]KANG Yujian, YANG Dechang, KONG Lei, et al. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features[J]. Nucleic Acids Research, 2017,45(W1):W12-W16. DOI: 10.1093/nar/gkx428.

[14]SUN Liang, LUO Haitao, BU Dechao, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts[J]. Nucleic Acids Research, 2013,41(17):e166. DOI: 10.1093/nar/gkt646.

[15]MISTRY J, CHUGURANSKY S, WILLIAMS L, et al. Pfam: The protein families database in 2021[J]. Nucleic Acids Research, 2021, 49(D1):D412-D419. DOI: 10.1093/nar/gkaa913.

[16]LI Aimin, ZHANG Junying, ZHOU Zhongyin. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme[J]. BMC Bioinformatics,2014,15:311. DOI: 10.1186/1471-2105-15-311.

[17]GAO Liangliang, BRADEEN J M. Contrasting potato foliage and tuber defense mechanisms against the late blight pathogen *Phytophthora infestans*[J]. PLoS One, 2016, 11(7): e0159969. DOI: 10.1371/journal.pone.0159969.

[18]PERTEA G, PERTEA M. GFF Utilities: GffRead and gffcompare[J]. F1000Research, 2020, 9: 304. DOI: 10.12688/f1000research.23297.2.

[19]LI Heng, HANDSAKER B, WYSOKER A, et al. The sequence alignment/map format and SAMtools[J]. Bioinformatics,2009,25(16):2078-2079. DOI: 10.1093/bioinformatics/btp352.

[20]PERTEA M,KIM D, PERTEA G M, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown[J]. Nature Protocols,2016,11(9): 1650-1667. DOI: 10.1038/nprot.2016.095.