Hans Journal of Computational Biology 计算生物学, 2022, 12(3), 40-48
Published Online September 2022 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2022.123006

Hans汉斯

基于马铃薯转录组数据的病毒组装软件比较

涂　振[1*]，钟子旸[1*]，孟繁烨[1]，邹　莹[2]，李佳炜[1]，余　涛[1]，郑经涛[1]，夏军辉[1]，张　舒[3#]，聂碧华[1#]

[1]华中农业大学园艺林学学院，湖北 武汉
[2]恩施州农业科学院，湖北 恩施
[3]湖北省农业科学院植保土肥研究所，湖北 武汉

收稿日期：2022年7月29日；录用日期：2022年8月29日；发布日期：2022年9月6日

摘　要

随着高通量测序技术的成熟和成本的降低，转录组数据呈现爆发式增长。转录组数据中除了包含寄主马铃薯自身的转录本以外，还可能包含寄主受到RNA病毒侵染而带来的病毒序列信息，因此可以低成本地从转录组数据中进行病毒基因组挖掘。本研究通过比较SOAPdenovo、IDBA-UD、Trinity 三种主流软件对RNA-seq数据的组装效果，发现Trinity软件组装得到的结果中序列信息最丰富，且长序列最多，但组装过程耗时较长；相对而言，SOAPdenovo和IDBA-UD耗时较短，但组装结果中序列信息较少且长序列较少，所以推荐使用Trinity软件进行基于转录组数据的病毒基因组组装。

关键词

高通量测序，转录组，病毒，从头组装

Comparing of Three Softwares for Virus Genome Assembly Based on Potato Transcriptome Data

Zhen Tu[1*], Ziyang Zhong[1*], Fanye Meng[1], Ying Zou[2], Jiawei Li[1], Tao Yu[1], Jingtao Zheng[1], Junhui Xia[1], Shu Zhang[3#], Bihua Nie[1#]

[1]College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan Hubei
[2]Enshi Academy of Agricultural Sciences, Enshi Hubei
[3]Institute for Plant Protection & Soil Fertilizer, Hubei Academy of Agricultural Sciences, Wuhan Hubei

Received: Jul. 29th, 2022; accepted: Aug. 29th, 2022; published: Sep. 6th, 2022

[*]第一作者。
[#]通讯作者。

文章引用: 涂振, 钟子旸, 孟繁烨, 邹莹, 李佳炜, 余涛, 郑经涛, 夏军辉, 张舒, 聂碧华. 基于马铃薯转录组数据的病毒组装软件比较[J]. 计算生物学, 2022, 12(3): 40-48. DOI: 10.12677/hjcb.2022.123006

Abstract

With the continuous maturity of high-throughput sequencing technology and the reduction of cost, transcriptome data show explosive growth. The potato transcriptome data not only contains the transcripts of the potato itself, but also contains the viral sequence information caused by the infection of viruses in the sample, so the virus genome mining can be carried out from the transcriptome data. In this study, the assembly results of three mainstream software (SOAPdenovo, IDBA-UD and Trinity3) were compared based on the same RNA-seq data, it was found that Trinity software resulted the most abundant sequence information and the longest sequences, but the assembly process took a long time; meanwhile, SOAPdenovo and IDBA-UD cost a relatively short time, but generated less sequence information and shorter sequences in the assembly results. Thus, it is recommended to use Trinity software to assemble virus genome based on transcriptome data.

Keywords

High-Throughput Sequencing Technology, Transcriptome Data, Virus, De Novo Assembly

1. 引言

马铃薯(*Solanum tuberosum* L.)属茄科茄属作物，具有良好的适应性和均衡的营养，已经在世界约 160 个国家和地区广泛种植，是世界第四大粮食作物[1]。我国马铃薯种植面积和产量均居世界第一，但单产量远不及世界发达国家水平，而病毒侵染引起的种薯退化是导致减产的重要原因之一[2]。马铃薯病毒病会导致其块茎产量和品质的下降，长期以来给广大马铃薯产业工作者造成了巨大损失，也制约了我国产业发展。

病毒在细胞水平侵染马铃薯，因此一旦感病，很难用药物进行治疗，只能主要采取预防的办法，而预防病毒病的关键是对各种马铃薯病毒的快速检测。然而，侵染马铃薯的病毒多达 40 种以上，而且均为 RNA 病毒，用传统的酶联免疫(ELISA)和反转录聚合酶链式反应(RT-PCR)等方法需要投入大量的人力和时间，而且上述方法只能检测数十种主要病毒。

近年来，高通量测序发展迅速，已经被广泛应用于包括植物病毒生物学研究的众多领域[3] [4] [5]。2011 年，马铃薯基因组测序已基本完成，给许多马铃薯科研工作者带来了新的机遇。转录组(transcriptome)是指某种生理条件下细胞中产生的全部转录物(包括 mRNA 和非编码 RNA)的总和[6]。转录水平调控是生物体最重要的调控方式之一。近几年，转录组学研究是科学家研究的热门领域。通过高通量测序对样品的基因表达水平进行分析被称为 RNA-seq，而在所测得的数据中不仅包含着样品的转录本，也含有那些感染寄主的 RNA 病毒序列信息，因此利用植物转录组测序数据，过滤掉寄主的序列，可以对病毒资源进行挖掘[7]。本实验室在进行马铃薯低温糖化以及结薯机制的相关研究中已经进行了大量的转录组测序[8] [9] [10]，利用这些现有的数据，可以快速而低成本的获得潜在的马铃薯病毒序列信息，从而“变废为宝”，为这些马铃薯病毒病的检测和防治工作奠定了重要的基础。

RNA-seq 测序得到非常多很短的读段(reads)，想要获得病毒基因组序列，最核心的问题是将这些短的 reads 拼接组装成长的转录本，因此需要对 RNA-seq 数据进行从头组装(de novo assembly)。从头组装，

是指没有参考基因组的帮助下，将 reads 拼接组装成真实 DNA 序列的过程。目前，基因组的从头组装主要基于以下两种算法：Overlap-Layout-Consensun (OLC)算法[11]和 De Bruijn 图算法[12]。且后者更加适合 reads 较短、数据量较大的二代测序数据进行拼接组装[13]。基于此算法开发的软件有：Trinity、SOAPdenovo、IDBA-UD 等。因此，本研究利用一系列实验室相关研究获得的测序数据，在通过比对过滤掉马铃薯基因组数据后，用上述 3 种软件进行了病毒序列数据的组装，通过比较组装效果，确定从 RNA-seq 测序数据中发掘病毒序列的最佳分析软件。

2. 材料与方法

2.1. 数据收集

研究用到的相关测序数据来自本实验室其它研究积累的转录组测序数据，分别随机选择了 6 个 PE125 数据样本(chang-1, chang-2, CW2-1, CW2-2, duan-1, duan-2)和 6 个 PE150 数据样本(AC142, PVS, E20, E108, Ri-1, Ri-2)。

本研究使用的马铃薯参考基因组序列由国际马铃薯基因组测序协会(International Potato Genome Sequencing Consortium, PGSC)公布的对 DM1-3 (*S. phureja*)进行全基因组测序的基因组序列及其注释文件，版本号为 6.1 (http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml)。相关病毒序列来自 NCBI 公共数据库下载的病毒标准序列数据库(ftp://ftp.ncbi.nih.gov/refseq/release/viral/) [14]。

2.2. 生物信息数据分析方法

2.2.1. 组装前数据预处理

在进行数据组装前，首先使用 FastQC (v0.11.3)对数据进行质量评估，使用 Trimmomatic [15]进行质控，得到 clean_data；接着，将 clean_data 数据比对到马铃薯参考基因组，以区分来源于马铃薯转录本的 reads 和非来源于马铃薯转录本的 reads，即 unmapped reads，作为进行组装方法比较的核心数据。

2.2.2. 使用 Trinity 软件进行组装

对于上一步得到的 unmapped reads，即非来源于马铃薯转录本的数据，用 Trinity (v2.1.0)软件进行重头组装：首先，对于未匹配到马铃薯基因组的数据，即上步得到的 unmapped.bam 文件，使用 samtools (v0.1.18)根据 bam 文件的 flag 标签，提取双端 reads，得到 unmapped_reads1.fastq 和 unmapped_reads2.fastq。本研究所使用的命令如下：

```
samtools view unmapped.bam|awk '{if ($2~/69/) print ">"$1"\n"$10"\n+\n"$NF}' > unmapped_reads1.fastq
samtools view unmapped.bam|awk '{if ($2~/133/) print ">"$1"\n"$10"\n+\n"$NF}' > unmapped_reads2.fastq
```

得到双端 reads 后，使用 Trinity (v2.1.0)进行重头组装。本研究所使用的命令如下：Trinity --seqType fq --max_memory 10G --left unmapped_reads1.fastq --right unmapped_reads2.fastq --CPU 10

2.2.3. 使用 SOAPdenovo 软件进行组装

使用 SOAPdenovo (v2.04)进行组装，输入文件不能用上一步 samtools 提取的双端数据，因为 samtools 提取的双端 reads 是全部未匹配的 reads，没有考虑 reads 的配对情况，会出现只有某一端，没有另一端而无法配对的情况。因此使用 bam2fastq (v1.1.0)软件提取双端 reads，该软件默认会去掉无法配对的 reads；另外，SOAPdenovo 组装需要建立一个配置文件，配置文件指定了输入文件的路径、数据特征和程序运行的参数。此处由于会设置 average insert size 参数，所以需要先用 Picard (v1.119)计算 average insert size，命令如下：java-jar/home/software/picard-tools-1.119/CollectInsertSizeMetrics.jar I = accepted_hits.bam O = insertsize.txt H = insertsize.pdf。最后，建好配置文件以后，可以运行 SOAPdenovo 的主程序。本研究使用命令如下(以 k-mer

取 105 为例)：SOAPdenovo-127mer all-s config.txt-o out_K105-K 105-p 10-R & > soapdenovo.log。

2.2.4. 使用 IDBA-UD 软件进行组装

IDBA-UD 与 SOAPdenovo 都需要使用 bam2fastq 软件提取的双端数据；组装前，要先对双端数据进行合并，合并成一个文件：fq2fa--merge--filter unmapped_reads1.fastq unmapped_reads2.fastq unmapped_reads.fa；本研究使用命令如下(以 125 bp reads 为例)：idba_ud -r unmapped_reads.fa--mink 69--maxk 107--step 2 -o./denovo--num_threads 10。

2.2.5. 本地 blast 进行同源性比对

使用本地化 blast+ (v2.2.31)将组装得到的 contig 与所构建的数据库 III 进行比对，研究组装结果中的病毒多样性，这些 contig 中可能会有已报道的对马铃薯有侵染能力的病毒，也可能会发现未报道的对马铃薯有侵染能力的病毒；再使用本地化 blast+将组装得到的 contig 与所构建的数据库 II 进行比对，研究组装结果中 contig 具体与马铃薯病毒的哪个亚型或分离物同源性最高。

3. 结果与分析

3.1. 不同 k-mer 下软件的组装效果比较

本研究选择了 contig N50 和最长 contig 两个指标，作为衡量组装效果的指标，同时也作为选择最适 k-mer 的评判标准。N50 反映的是一个组装结果整体的序列长度，N50 越大，说明组装得到长序列越多，组装效果越好。另外，由于本研究进行的是病毒基因组组装，希望得到最长 contig 越长越好，因此本研究还选择最长 contig 作为最适 k-mer 选择的另一个指标。

本研究对 SOAPdenovo 和 IDBA-UD 两种软件在不同 k-mer 大小下，N50 和 max contig 进行统计。如图 1A 和 B 所示，对于 PE125 的数据(Chang-1 和 CW2-1)，随 k-mer 从 69 增加到 87-91 左右，contig N50 逐渐增大，k-mer 值在 91-95 左右急剧下降，随后逐渐增大，因此 Chang-1 和 CW2-1 分别选择 87 和 91 为最佳的 k-mer。而对于 PE150 的数据(AC142 和 PVS) (图 1C 和图 1D)，随着 k-mer 逐渐增大，contig N50 逐渐增大，在 k-mer 取 127 时达到最大值，因此最佳 k-mer 为 127。但是，对于 max contig 这一指标，无论是 PE125 和 PE150 数据，其变化规律不明显(图 1A-D)。

使用 IDBA-UD 软件进行从头组装，支持的最大 k-mer 值是 123。无论是 PE125 还是 PE150 的数据，conitg N50、max contig 与 k-mer 之间存在着相似的规律：对于 PE125 数据(Chang-1 和 CW2-1)，Contig N50 在 k-mer 达到 71 时有一个明显上升，随后保存缓慢线性增长，k-mer 达到 105 后达到最大值，随后有所回落。而 max contig 的长度也随 k-mer 值增大呈现一定的增长态势，以 CW2-1 为例(图 1F)，k-mer 值从 93 增加到 95 时，max contig 长度增加 300 bp 左右，而 Chang-1 (图 1E)在 k-mer 值处于 75~107 之间，max contig 保存不变。因此，为了获得最佳组装效果，该软件对 PE125 数据，k-mer 选择 105 为宜。对于 PE150 数据(AC142 和 PVS) (图 1G 和图 1H)，contig N50 随 k-mer 的变化规律与 PE125 数据相似，只是一般在 k-mer 值 121 左右达到最大值。而 max contig 长度在 contig N50 极值范围内变不大。因此，对于 PE150 的 reads 推荐使用 115~121 作为最优 k-mer 的选择范围。

综上所述，对于 SOAPdenovo 软件，组装效果随着 K-mer 变化无明显变化规律，需要视实际组装情况选择合适 K-mer。而对于 IDBA-UD 软件组装效果与 k-mer 取值的变化规律非常明显，即 K-mer 越大，组装效果越好。因此，IDBA-UD 软件因此更容易选择最合适的 k-mer 值，而且组装效果更加容易让人信服。

3.2. 不同组装软件效果比较

本研究分别使用 Trinity、SOAPdenovo、IDBA-UD 软件对随机选择的 PE125、PE150 各 6 组数据进

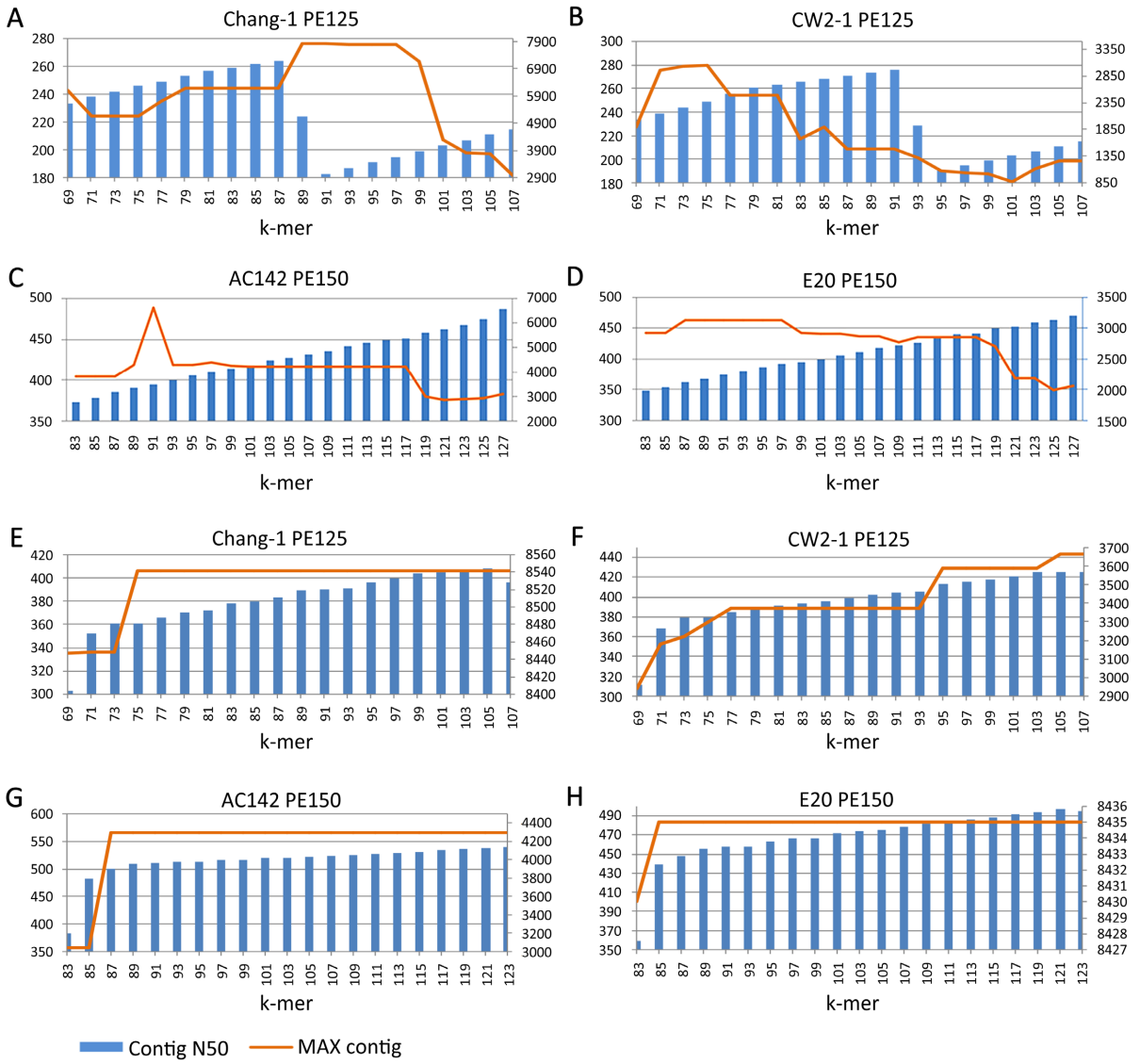

Figure 1. Effect of the k-mer on the de novo assembly quality using SOAPdenovo and IDBA-UD
图 1. SOAPdenovo 和 IDBA-UD 软件 k-mer 值对组装质量的影响

行组装，比较组装效果，选择最适合作为病毒基因组组装的软件。为了衡量软件组装效果，选择了 contig N50、contig 平均长度、contig 中位数长度、最长 contig、大于 1000 的 contig 的数量、组装结果总 contig 数目以及大于 1000 bp 的 contig 数目占总 contig 数目的百分比作为指标，其中前 3 个指标都反映的是组装得到的序列的整体长度，这 3 个指标数值越大，表示组装结果中长序列越多，组装效果越好。对三个组装软件的组装结果如表 1 所示。

以前 3 个指标进行考查，对于 PE125 的数据，这 3 个指标都是 IDBA-UD 的最大，Trinity 次之，SOAPdenovo 最小，以 CW2-1 数据样本为例，IDBA-UD 软件的 3 个指标的值分别是 425 bp、422 bp、387 bp，Trinity 对应的分别是 321 bp、325 bp、277 bp，SOAPdenovo 对应的分别是 276 bp、179 bp、138 bp。对于 PE150 的数据来说，3 个软件对应的 3 个指标的值相差不大，以 AC142 数据样本为例，IDBA-UD 的 3 个指标分别是 540 bp、543 bp、470 bp，Trinity 对应的分别是 444 bp、424 bp、341 bp，SOAPdenovo 对应的分别是 454 bp、459 bp、431 bp。对于最长 contig 这一指标，3 个软件组装结果相差比较大，但是

Trinity 软件的该指标都是最优的，以 CW2-1 数据样本为例，使用 Trinity 软件进行组装最长 contig 达到 8469 bp，而 IDBA-UD 和 SOAPdenovo 软件分别只有 3661 bp 和 1484 bp，远远小于 Trinity 软件。对于大于 1000 的 contig 的数量、组装结果总 contig 数目以及大于 1000 bp 的 contig 数目占总 contig 数目的百分比这 3 个指标进行比较，Trinity 软件优势非常明显，使用该软件组装得到的大于 1000 bp 的 contig 数量以及组装得到的总 contig 数量都是最多的，且远远高于另外 2 个软件，以 CW2-1 数据样本为例，Trinity 软件组装结果中共有 23,571 条序列，其中大于 1000 bp 的有 198 条，对于 IDBA-UD 这 2 个值分别是 1737 和 26，SOAPdenovo 这 2 个值分别是 2882 和 2。这反映了 Trinity 软件的组装结果中包含了丰富的序列信息且长序列最多，这为后续进行序列信息挖掘提供了基础。

Table 1. Comparison of the assembly quality for SOAPdenovo, IDBA-UD and Trinity
表 1. SOAPdenovo、IDBA-UD 和 Trinity 3 个软件组装结果比较

samples	software	contig N50 (bp)	mean of contig length (bp)	median of contig length (bp)	max contig length (bp)	contig number of over 1000 bp	total contig number	ratio of contig over 1000 bp to total contig
CW2-1 (PE125)	SOAPdenovo	276	179	138	1484	2	2882	0.07%
	Trinity	321	325	277	8469	198	23,571	0.84%
	IDBA-UD	425	422	387	3661	26	1737	1.50%
CW2-2 (PE125)	SOAPdenovo	277	178	135	1702	3	3082	0.10%
	Trinity	322	324	278	8160	1138	47,198	2.41%
	IDBA-UD	426	425	389	3101	23	1811	1.27%
Chang-1 (PE125)	SOAPdenovo	264	166	118	6169	7	6128	0.11%
	Trinity	321	325	279	9695	173	29,916	0.58%
	IDBA-UD	408	391	364	8541	28	2366	1.18%
Chang-2 (PE125)	SOAPdenovo	263	166	116	8469	6	5829	0.10%
	Trinity	338	339	289	9699	306	31,574	0.97%
	IDBA-UD	456	423	389	8446	46	2584	1.78%
duan-1 (PE125)	SOAPdenovo	281	189	135	4966	11	3688	0.30%
	Trinity	329	331	284	9694	260	34,340	0.76%
	IDBA-UD	431	418	379	8533	49	2722	1.80%
duan-2 (PE125)	SOAPdenovo	285	194	141	7173	11	3660	0.30%
	Trinity	323	326	280	9696	220	33,831	0.65%
	IDBA-UD	407	399	363	8489	32	2748	1.16%
AC142 (PE150)	SOAPdenovo	454	459	431	1114	4	262	1.53%
	Trinity	444	424	341	4534	2501	61,707	4.05%
	IDBA-UD	540	543	470	4297	277	5237	5.29%
PVS (PE150)	SOAPdenovo	427	367	392	3034	19	2071	0.92%
	Trinity	509	466	369	9727	6561	112,067	5.85%
	IDBA-UD	543	515	460	5618	733	15,540	4.72%

Continued

Ri-1 (PE150)	**SOAPdenovo**	446	408	423	910	0	285	0.00%
	Trinity	446	425	343	4945	2341	60,717	3.86%
	IDBA-UD	523	530	464	4319	231	5325	4.34%
Ri-2 (PE150)	**SOAPdenovo**	440	401	410	1108	3	370	0.81%
	Trinity	445	424	341	4481	2599	65,451	3.97%
	IDBA-UD	548	551	476	3897	309	5601	5.52%
E108 (PE150)	**SOAPdenovo**	453	458	426	2708	9	999	0.90%
	Trinity	454	430	345	8496	3551	84,627	4.20%
	IDBA-UD	434	342	245	4651	315	16,785	1.88%
E20 (PE150)	**SOAPdenovo**	452	467	432	2062	12	706	1.70%
	Trinity	463	435	349	8471	3997	93,936	4.26%
	IDBA-UD	497	491	440	8435	324	10,235	3.17%

3.3. 组装得到病毒序列

使用 Trinity 软件对处理后的数据进行组装，比对病毒数据库后对结果进行整理分析。在收集的数据中共找到了 10 种植物病毒，9 种 RNA 病毒和 1 种 DNA 病毒，总计 2129 条植物病毒序列，其中 314 条比对长度超过 1000 bp，大部分序列比对结果的一致性都在 95%以上，其中 114 条覆盖 90%以上的基因组，接近数据库中的病毒全长序列。

4. 讨论

4.1. 不同软件的 k-mer 选择

使用 SOAPdenovo 对序列进行组装，k-mer 的取值对基因组组装结果影响较大。本研究参考了其他生物信息领域科研人员的建议，k-mer 设置范围取 reads 长度的 55%~85%，且为奇数。因此，对于 PE125 的数据，k-mer 选择 69~107，对于 PE150 的数据，k-mer 选择 83~127。

使用 IDBA-UD (v1.2.0)进行组装不需要设置不同大小的 k-mer 分步运行，该软件可以从小的 k-mer 开始到大的 k-mer 进行递增计算，多个 k-mer 一次就可运行完成。合并后的数据再进行组装，设置最大、最小的 k-mer 参数，与上一步选择相同的 k-mer，但是由于 IDBA-UD 软件支持的最大 k-mer 是 123，因此对于 150 bp 的 reads 最大 k-mer 只能选择到 123。

对于 Trinity 软件，由于该软件不需要手动设置 k-mer 大小，该软件可以自动计算最优 k-mer，因此本研究并未对 Trinity 软件进行上述研究。

4.2. 病毒基因组装

在本研究中利用 SOAPdenovo、IDBA-UD、Trinity 三种软件对测序数据进行组装，通过对组装效果的衡量对比，我们发现 Trinity 软件运行病毒序列组装，信息最为丰富，长序列最多，但运行时间长；而 SOAPdenovo 和 IDBA-UD 这两款软件耗时都比较短，但是组装效果不佳，序列信息少且长序列少，所以在追求组装质量的前提下，本研究建议选择 Trinity 进行全部 RNA-seq 数据的组装。但从在耗时以及消耗资源的角度看，SOAPdenovo 和 IDBA-UD 有非常明显的优势，相同线程、内存资源条件下，组装同一个样本，Trinity 所消耗的时间是 SOAPdenovo 和 IDBA-UD 的几十倍。综上所述，Trinity 软件组装效果最

好，Smith 等人和 Sparks 等人分别对艺神袖蝶和茶翅蝽进行转录组测序，使用 Trinity 进行数据拼接组装，都发现和鉴定了新病毒[16] [17]。因此推荐使用 Trinity 软件进行基于转录组数据的病毒基因组组装。本研究中，利用 Trinity 软件对未匹配到马铃薯参考基因组的数据进行组装，组装结果比对病毒数据库，共发现 9 种 RNA 病毒和 1 种 DNA 病毒。其中有常见的 6 种马铃薯病毒，所占比例最大且为 90.94%。此外还发现了 TuMV 等几种目前尚未报道对马铃薯有侵染能力的病毒，表明了使用 Trinity 组装马铃薯病毒序列的可靠性。但是，如果考虑到时间成本的话推荐使用 IDBA-UD 软件。而 SOAPdenovo 组装“碎片”较多，不适合病毒基因组组装。

利用生物学分析软件构建基于 RNA-seq 数据的马铃薯病毒基因组挖掘平台，分析 RNA-seq 数据中病毒多样性，研究病毒变异、进化，发现和鉴定可能存在的新病毒或者尚未报道的对马铃薯有侵染能力的病毒，为马铃薯与病毒互作研究、马铃薯抗性材料筛选奠定了基础，对马铃薯抗病毒育种具有重要意义。

4.3. 测序数据深度挖掘

高通量测序技术由于其快速、准确、低成本的特点，已经被广泛应用于生物学研究中。近年来，越来越多的植物基因组测序完成，与马铃薯相关的转录组数据快速增长。同时测序技术也在不断创新，二代测序技术由于其通量高、精度高、信息丰富、成本低的特点已经广泛应用于生命科学研究的各个领域，产生了大量的转录组测序数据。但是，大量的测序数据在满足了特定的生物意义分析之后，就被闲置起来，十分可惜。但是，我们可以从其它角度进行深度挖掘，从这些闲置的数据中发现潜藏的有价值信息。

基于以上理论和技术基础，在本研究中收集了实验室数据库中的转录组数据，进行深度挖掘，挖掘其中的病毒序列，分析能够侵染马铃薯的病毒类型、病毒多样性。但由于测序最初的目的是针对宿主马铃薯进行研究，所以大部分数据是宿主序列，病毒序列丰度不足，因此导致组装到完整病毒序列的数量有限。本研究针对转录组数据构建了基于 RNA-seq 数据的病毒组装平台，该平台能够用于分析 RNA-seq 数据中病毒序列多样性，能发现一些新的、有研究价值的病毒序列，具有良好的应用前景。

致　谢

作者感谢陈汝豪、李春燕等同学为本研究提供的相关测序数据，感谢罗鸣博士在分析方法上提出的意见和建议。感谢农业农村部马铃薯生物学与生物技术重点实验室为本研究提供了良好的科研平台。

基金项目

国家自然科学基金项目(31971989)；农业部华中作物有害生物综合治理重点实验室/农作物重大病虫草害防控湖北省重点实验室开放基金课题(2019ZTSJJ4)；湖北省农业科技创新中心项目(2019-620-003-001)；国家大学生创新创业基金(202010504041)。

参考文献

[1] 白人朴. 关于我国马铃薯产业发展振兴的思考[J]. 农机科技推广, 2017(3): 4-6.

[2] 白艳菊, 李学湛, 文景芝, 杨明秀. 中国与荷兰马铃薯种薯标准化程度比较分析[J]. 中国马铃薯, 2006, 20(6): 357-359.

[3] Adams, I.P., Glover, R.H., Monger, W.A., *et al*. (2009) Next-Generation Sequencing and Metagenomic Analysis: A Universal Diagnostic Tool in Plant Virology. *Molecular Plant Pathology*, **10**, 537-545. https://doi.org/10.1111/j.1364-3703.2009.00545.x

[4] Rwahnih, M.A., Daubert, S., Golino, D., *et al*. (2009) Deep Sequencing Analysis of RNAs from a Grapevine Showing Syrah Decline Symptoms Reveals a Multiple Virus Infection that Includes a Novel Virus. *Virology*, **387**, 395-401. https://doi.org/10.1016/j.virol.2009.02.028

[5] Kreuze, J.F., Perez, A., Untiveros, M., *et al*. (2009) Complete Viral Genome Sequence and Discovery of Novel Viruses

by Deep Sequencing of Small RNAs: A Generic Method for Diagnosis, Discovery and Sequencing of Viruses. *Virology*, **388**, 1-7. https://doi.org/10.1016/j.virol.2009.03.024

[6] Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics*, **10**, 57-63. https://doi.org/10.1038/nrg2484

[7] Batty, E.M., Nicholas, W.T.H., Amy, T., *et al*. (2013) A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples. *PLOS ONE*, **8**, e66129. https://doi.org/10.1371/journal.pone.0066129

[8] Shan, J., Song, W., Zhou, J., *et al*. (2013) Transcriptome Analysis Reveals Novel Genes Potentially Involved in Photoperiodic Tuberization in Potato. *Genomics*, **102**, 388-396. https://doi.org/10.1016/j.ygeno.2013.07.001

[9] Ai, Y., Jing, S., Cheng, Z., *et al*. (2021) DNA Methylation Affects Photoperiodic Tuberization in Potato (*Solanum tuberosum* L.) by Mediating the Expression of Genes Related to the Photoperiod and GA Pathways. *Horticulture Research*, **8**, Article No. 181. https://doi.org/10.1038/s41438-021-00619-7

[10] Liu, X., Chen, L., Shi, W., *et al*. (2021) Comparative Transcriptome Reveals Distinct Starch-Sugar Interconversion Patterns in Potato Genotypes Contrasting for Cold-Induced Sweetening Capacity. *Food Chemistry*, **334**, Article ID: 127550. https://doi.org/10.1016/j.foodchem.2020.127550

[11] Pevzner, P.A, Tang, H. and Waterman, M.S. (2001) An Eulerian Path Approach to DNA Fragment Assembly. *Proceedings of the National Academy of Sciences*, **98**, 9748-9753. https://doi.org/10.1073/pnas.171285098

[12] Idury, R.M. and Waterman, M.S. (1995) A New Algorithm for DNA Sequence Assembly. *Journal of Computational Biology*, **2**, 291-306. https://doi.org/10.1089/cmb.1995.2.291

[13] Zhang, W., Chen, J., Yang, Y., *et al*. (2012) A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLOS ONE*, **6**, e17915. https://doi.org/10.1371/journal.pone.0017915

[14] Wang, B., Ma, Y., Zhang, Z., *et al*. (2011) Potato Viruses in China. *Crop Protection*, **30**, 117-1123. https://doi.org/10.1016/j.cropro.2011.04.001

[15] Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics*, **30**, 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

[16] Sparks, M.E., Gundersenrindal, D.E. and Harrison, R.L. (2013) Complete Genome Sequence of a Novel Iflavirus from the Transcriptome of *Halyomorpha Halys*, the *Brown Marmorated Stink Bug*. *Genome Announcements*, **1**, e00910-13. https://doi.org/10.1128/genomeA.00910-13

[17] Smith, G., Macias-Muñoz, A. and Briscoe, A.D. (2014) Genome Sequence of a Novel Iflavirus from mRNA Sequencing of the Butterfly *Heliconius erato*. *Genome Announcements*, **2**, e00398-14. https://doi.org/10.1128/genomeA.00398-14