DOI:10.3969/j.issn.1672-5565.201709003

Unix 文本比对分析高通量 RNA-Seq 测序基因表达

宋东光,卢博彬,陈柳婷
(佛山科学技术学院 园艺系,广东 佛山,528231)

摘 要:从 RNA-Seq 高通量测序短序列进行比对及拼接获得较长转录本并确定基因表达量的方法随着转录组测序的广泛开展仍在不断改进,本文利用类 Unix 系统的文本处理命令组合对山茶花开花期叶片及花瓣的转录组序列进行比对、序列拼接及其表达量分析。首先对测序序列进行每 1 万条准随机排序,选取 10 万条序列分别与 100 万条序列进行比对,从每个查询序列随机选取 9 组20 mer比对 100 万条序列去重后获得该序列的转录数量。利用查询序列首尾20 mer从匹配的比对重叠群进行拼接,初次拼接最长为410 mer,超过两个及以上拼接序列的再次进行相互比对及再拼接,最长1 174 mer。用查询序列的比对匹配数表示其拼接前后的表达量,与用互补链进行比对得到的负链表达量相当。用拼接序列进行 NCBI 联网 blast 比对获得了其基因注释。本文得到的结果表明,利用类 Unix 系统文本比对可以有效用于高通量测序基因表达量及进行序列从头组装等分析。

关键词:RNA-Seq;文本比对;基因表达量;重叠群拼接;类 Unix 系统
中图分类号:Q344+.13 **文献标志码**:A **文章编号**:1672-5565(2018)02-119-11

Gene expression analysis from high-throughput RNa-Seq sequencing by Unix Text-aligning

SONG Dongguang, LU Bobin, CHEN Liuting
(*Department of Horticulture*, *Foshan University*, *Foshan* 528231, *Guangdong*, *China*)

Abstract: With the rapid development in transcriptome sequencing nowadays, further improvement in methods for estimating gene expression is underway in aligning and assembling short high-throughput RNA-Seq sequences into longer transcripts. Preliminary sequence alignment, assembly and expression of blade and petal transcriptome of Camellia at flowering stage were reported in this study by the combinations of text-filtering commands in Unix-like operating system. Firstly, near-random sorting of every 10 000 sequences were completed, then 100 000 sequences were aligned to 1 million sequences. 9 randomly selected groups of 20 mers selected from each query sequence were aligned to 1 million sequences, and transcripts were counted after removing duplicated sequences. By first- and-last 20 mers of query sequences, assembly was conducted in matching contigs of each aligned group. The longest sequence in first assembly was 410 mers. The longest sequence was 1174 mers in re-aligning and reassembly of two or more joint sequence. Matched aligning counts of each query sequence were used as its expression before and after assembling, which was approximately equal to the minus strand's expression after comparing with that of complementary strand. Gene connotations were obtained by aligning joint sequences to remote NCBI blast server. The results show that gene expression and de novo assembly could be effectively analyzed by text-aligning in Unix-like system.

Keywords: RNA-Seq; Text-aligning; Gene expression; Contig-joining; Unix-Like system

植物生长发育过程基因表达调控研究近年来随着二代测序的迅速发展得到广泛深入开展,RNA-Seq 高通量测序给获得植物组织器官甚至单细胞全基因组转录信息带来了革命性突破[1-2],随着测序成本的下降,测序数据量剧增,截至 2017 年 8 月,NCBI 收录的 SRA/RNA-Seq 测序数量达到了 65 万

收稿日期:2017-09-18;**修回日期**:2018-01-23.
* **通信作者**:宋东光,男,博士,副教授,研究方向:遗传学;E-mail:3dsong@163.com.

条以上，其中人（*Homo sapiens*）与小鼠（*Mus musculus*）均已接近 20 万条，拟南芥 1.35 万条，玉米 6.7 千条，水稻 3.8 千条，很多数据目前可能并没有递交到 NCBI SRA 数据库。从高通量 RNA 测序结果中进行基因表达等分析近年来开发了很多工具[3-6]，包括没有参考基因组的序列从头组装工具[7-9]。很多工具算法复杂，完成多项分析任务，运行时间较长，研究者可能无法完全得知返回数据的运算细节，这对于从结果分析中发现基因表达的更多信息有一定难度[10-13]。

高通量测序序列较短而量大，利用序列本身直接从测序序列群体进行匹配会更有效找到比对序列，由于测序错误率引起匹配误差则可考虑不使用全长查询序列而使用局部序列进行匹配。为了加快 RNA 高通量测序获得基因表达量及进行从头组装，本文探讨利用开放源代码操作系统的文本过滤命令组合脚本进行 RNA-Seq 测序序列比对及拼接分析，分析过程清晰、简捷，不需要复杂的代码，可广泛用于进行快速 RNA-Seq 基因从头组装及表达量等分析。

1 材料与方法

1.1 植物材料

山茶花品种克瑞默大牡丹和银凯旋的开花期花枝成熟全展叶片和未打开花苞的花瓣，采于佛山市植物园，由佛山市林业科学研究所提供。样品采集后马上液氮速冻并通过干冰运输。

1.2 RNA-Seq 测序

由北京奥维森基因科技有限公司提供测序服务，包括总 RNA 提取及建库，双端测序（Paired-End，Illumina HiSeq 4 000）。序列格式 fastq，提交 6G 高质量数据（Clean data）及 7G 未处理原始测序数据（Raw data），每条序列长度150 mer，合并双端得到每个样品测序序列约5 000 万条。

1.3 操作系统

FreeBSD 11.0（amd64）（http://www.freebsd.org），计算机内存 8G，硬盘 1T，CPU 为 Intel（R）Xeon（R）CPU E3-1230 v5 @ 3.40GHz。

1.4 Unix 文本处理命令

包括 awk，sed，cat，cut，comm，split，uniq，sort，tr 等。所有序列处理与分析由 FreeBSD 操作系统的文本过滤命令及运行命令组合脚本完成。

1.5 序列处理

提取序列（选用 6G clean data），仅保留序列，每条序列加上编号，然后逐级按照每 10→100→5 万条（“split”命令）进行随机排序并按随机方式合并序列。其中，每个 100 万分成若干个目录单独进行每 1 万条的随机排序，再随机组合合并所有序列。

1.6 表达量分析

随机选取上述准随机排序后的 100 万条进行比对后统计表达量。1）比对：按照每 10 万条分割，选取其中任何一个 10 万条序列，每隔 5 个核苷酸取 20 个连续的核苷酸短序列（20 mer，共 27 个），随机选取 9 个20 mer短序列进行与 100 万条的序列匹配比对，统计序列编号去除重复编号获得每条序列与 100 万条匹配后的数量。同时，获得 9 个20 mer序列的互补链亦进行与 100 万条序列匹配比对。2）计算表达量：分别计算两次（正链与互补链）比对的统计数即可初步获得每个查询序列与 100 万条序列比对的匹配序列，即匹配该条序列的正链与互补链的表达量（每 100 万条序列转录物数量）。

1.7 重叠群（基因）序列拼接

将 10 万条序列与 100 万条序列比对得到的序列进行重叠群拼接，利用150 mer查询序列的 5′和 3′两端20 mer序列进行匹配，切除原始序列后找出剩下最长的序列进行连接。然后，将近 10 万条重叠群（部分可能没有获得）相互比对即可以去除重复的重叠群，即将重叠群序列再次进行9x20 mer随机片段选取并与其它重叠群序列比对，按照第一个序列号合并重叠群。并再次进行首尾 20 mer 拼接。

1.8 重叠群序列注释

将组装拼接得到的序列与 NCBI 远程服务器核酸（nr/nt）数据库进行比对（https://blast.ncbi.nlm.nih.gov/Blast.cgi? PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome）获得该序列的比对及注释信息。每次上传 500 条进行比对，保存比对结果进行注释提取。

2 结 果

2.1 RNA-Seq 测序基因表达分析流程

RNA-Seq 测序数据量大，选取较合适的序列数据大小来反映基因表达情况很必要，由于计算机运行内存的局限大多数程序都将序列数据进行切割，利用 100 万条进行比对及组装，获得重叠群拼接序列的表达量。RNA-Seq 测序分析流程如图 1。

2.2 RNA-Seq 测序 10 万条及其序列比对分析

山茶花叶片与花瓣 RNA 测序返回数据每个样品近5 000万条（双端测序合并，6G 高质量数据，即 clean data），进行每 1 万条准随机排序（见方法介绍）后如何选择适合的数据大小进行表达分析是很必要的，为此，分析过程进行了多个 100 万条及一个

300 万条的比对比较分析。另外,如果利用全部选取的 100 万条与 100 万条比对将耗尽计算机资源运行会非常慢甚至停止。因此,分析中从其中的 100 万条序列中选取了 10 万条叶片(银凯旋)、10 万条花瓣(克瑞默)分别与多个不同的 100 万条序列进行比对,每个比对得到 10 万条比对结果,每条找到与其序列匹配的序列首先合并其序列号,后面分析再提取其序列进行拼接得到重叠序列。考虑到测序过程是 cDNA 正负链都可能被测序,所以也将 10 万条序列的互补链与 100 万条序列进行了比对。两次比对结果可以合并用于计算每个重叠序列的表达量。由于本文主要介绍文本过滤分析得到基因表达量,测序的质控分析结果在此不列出,下面分别介绍各个比对结果。

图 1　开源系统文本比对分析高通量 RNA-Seq 测序流程图

Fig.1　Flowchart for high-throughput RNA-Seq sequencing by text-filtering in open-source system

2.2.1　10 万条序列与多个 100 万条序列比对

选取多长的序列进行比对合适也是很关键的,如果选取全长150 mer序列进行比对计算机运行会很艰难,因此,考虑用20 mer的长度,要尽可能从150 mer长度的序列找到其他匹配序列,故考虑将150 mer隔5 mer分割将产生 27 个20 mer短序列,下面比较了 24 条序列(银凯旋叶片测序)用其全部的 27 个20 mer及分开的 3 个 9 个20 mer、随机的 9 个 20 mer的比对结果,见表 1。

从表 1 可以看出,用 9 个随机的20 mer短序列与 27 个所有的 20 mer 序列比对得到的比对匹配数量很一致(A, E~I),而不进行随机选取的 3 个 9 个 20 mer的匹配数量与 27 个20 mer的差异较明显(A, B~D)。因此,为了加快比对进程,后面的比对结果全部选取了150 mer序列中随机选取 9 个20 mer进行比对。

表 2 用同样的银凯旋叶片测序的 24 条序列与银凯旋叶片测序的 3 个 100 万条(A/B/C)、1 个 300 万条(D)和克瑞默的 1 个叶片 100 万条(G), 及 2 个银凯旋花瓣的 100 万条(F-双端读 1/G-读 2)进行比对,每个 100 万条是随机选取的,300 万条序列里包含了 A 和 B 及另外的 100 万条(不在列表)。从表 2 可以看到,3 个银凯旋叶片的匹配数量也是很一致的,较高表达的 24 条序列与 300 万条的匹配数量也成倍增加,也说明基因表达并不是简单的线性增加。初步也可以看到两个品种的叶片表达数量是相似的,有些序列与花瓣则有明显的差异。从表 1 和表 2 结果可以看出,用 100 万条序列进行比对是可以反映高通量测序基因表达水平的。

2.2.2　10 万条序列重叠群拼接分析

利用 10 万条序列与多个 100 万条序列比对得到的匹配结果能否反映基因的表达情况,需要进一步对其进行拼接后才能确定其表达量。另外,10 万条序列是否包含了所有表达的基因序列亦需要进一步分析。利用银凯旋叶片 10 万条序列分别与多个 100 万条(两个品种及其叶片、花瓣)序列进行了比对,下面主要列出了其中部分结果。

表 1 银凯旋叶片 24 条序列的20 mer与 1 个叶片 100 万条序列比对的匹配序列数量

Table 1 Matchig sequence amount of 20 mers from 24 sequences of Yangkaixuan blades aligned to 1 million sequences from 1 blade

A	B	C	D	E	F	G	H	I
100	54	68	83	100	100	90	90	90
9	9	7	6	8	8	8	8	8
5	3	3	3	5	5	5	5	5
20	12	7	14	12	18	20	13	20
6	3	6	6	6	6	6	6	6
22	17	17	14	22	22	22	21	22
12	5	6	11	8	10	10	7	7
53	33	38	25	47	35	43	41	45
16	13	7	6	15	7	15	7	6
39	24	29	27	30	30	39	30	36
14	9	8	9	13	14	13	14	11
9	6	8	6	8	8	9	9	9
69	28	38	45	69	66	63	54	64
29	27	27	27	29	29	28	28	29
2 333	1 824	1 655	909	2 331	2 322	2 313	2 326	2 328
3	3	2	2	3	3	3	3	3
23	10	14	14	19	22	22	13	21
3	1	2	3	3	3	3	3	3
50	31	21	20	26	36	28	45	26
10	9	7	4	10	10	10	10	10
13	8	7	11	10	11	12	12	12
293	173	240	241	244	281	240	271	285
6	1	6	1	1	6	6	1	6
3	2	2	3	3	3	3	3	3

注:A: 27 个20 mer;B~D: 27 个20 mer序列从左至右的 9 个20 mer;E~I: 27 个20 mer分 5 次随机选取 9 个20 mer

表 2 24 条序列 9 个随机20 mer与多个 100 万条及一个 300 万条序列比对的匹配序列数量

Table 2 Matching sequence amount of 9 sets of randomly-selected 20 mers from 24 sequences aligned to multiple sets of 1 million sequences and 1 set of 3 million sequences

A	B	C	D	E	F	G
92	74	92	268	83	33	29
7	2	3	13	4	0	0
5	0	0	6	1	0	0
10	11	10	32	6	2	1
6	3	3	12	2	1	0
22	24	19	58	7	0	1
8	5	3	18	6	0	1
46	55	44	136	44	1	0
14	25	25	58	6	0	0
39	49	40	141	52	52	28
13	15	13	50	7	7	10
9	8	6	18	2	2	2
61	67	55	171	52	52	55
27	20	20	60	25	39	46
2 326	2 280	2 174	6 805	1 706	0	3
3	1	0	4	0	1	0
15	17	11	40	8	17	15
3	3	0	9	4	0	0
26	23	23	70	29	44	33
10	5	4	17	5	1	1
11	9	14	35	9	0	1
246	237	247	730	237	179	192
1	0	1	2	2	1	6
3	4	3	7	2	2	1

注:RNA-Seq 序列,A~C: 100 万条银凯旋叶片;D~300 万条银凯旋叶片(包含了 A 和 B);E:克瑞默叶片 100 万条; F、G:银凯旋花瓣 100 万条(F-双端读 1/G-读 2)。

银凯旋叶片 10 万条序列与包含自身序列的叶片 100 万条进行比对获得的 9 个比对结果如图 2,比对过程大约一周(每天约 1.5 万条,可同时运行 6 个相同任务计算机运行其他程序不受影响),对比对结果进行排序去重复(剩下92 025条),及去除没有匹配的(9 701条),得到 90 298 条匹配(见图 3),最高匹配序列6 615 个,大于1 000 个的2 371 个,小于1 000的87 927个。

13790001@10030283#10031079#10227640#10784934#11585699#120054#12256675#12257140#1244510#1248813#125058#12842830#13653240#13790001#13792297#14002071#14002499#14720862#14721150#15359604#15359981#15512955#15518852#15717508#15719171#15861148#15867274#16405050#16405266#1721230#17213698#17413389#1746935#17687883#17717276#18403645#18407014#18420049#18420371#18420566#18421230#18759319#18759409#18898412#18898988#18899421#20064224#20064549#20064714#20065155#20130071#20911840#21082533#21452618#21772332#21777613#22160445#22164155#22165959#22357494#22576124#23482870#23486738#23488277#23816585#23817641#23891026#23897892#24209180#24649130#24649943#3058420#3674018#3811242#3814089#3817951#5137625#5507266#5507434#5653681#5682317#5867541#5868785#5987411#6344761#6909494#7093446#8633286#8634940#8636314#8637382#9519931# 92

13790002@121492#1247601#13790002#3671166#547577#7404346#7404958# 7

13790003@11589968#13790003#15859197#5828058#8199936# 5

13790004@12690180#13790004#22090983#22091263#23488863#5132969#547461#5863179#786904#8632096# 10

13790005@10039718#10039969#13790005#18756587#22124791#22125634# 6

13790006@10039011#12636975#12636976#13790006#15008575#15352937#20917993#21774501#21895593#22049458#22165103#23818346#2499272#3053875#5164905#5252541#5862991#5865723#5982728#8001380#9522892#9522939# 22

13790007@13790007#17219706#21892924#21893015#3819076#3819766#5824094#6907384# 8

13790008@12270811#12275027#1243016#12632272#13580017#13580488#13790008#14005086#15350537#15350892#15350973#15355580#15714146#15852663#15864145#15868284#16409672#16893578#17689844#18407575#18422260#18461714#19645216#20067011#22040855#22092446#22166228#22167811#22572742#23502091#23502400#23815845#2602323#5134924#5166515#542570#5485814#6901310#6903837#6904446#6906228#6909410#8008366#8637922#8638348#9991588# 46

13790009@10221372#11580806#13790009#15005592#15352116#17212249#17718326#20062587#24204232#2608523#5259584#5820631#7093108#7401625# 14

图 2　银凯旋叶片 10 万条序列与 1 个 100 万条序列(包含自身 10 万条)部分序列比对匹配数

Fig.2　Partial matching amount of 100 000 sequences aligned to 1 set of 1 million sequences(including its own 100 000 sequences) of Yinkaixuan blade

10000000#10030780#10786854#12256352#12257434#12278490#1249442#12638796#13589627#13650493#13656646#13794319#14006122#14720115#15355852#15356140#15356347#15515500#15517692#15519647#15519648#15541714#15541817#15717377#16408909#1728313#17415704#1741657#18406923#18460931#18469710#18774656#18776602#20060374#20916641#21082473#21412826#2142180#22040470#22041921#22044155#22166437#22167871#22168761#22351490#23506484#23816299#23834994#23891412#23897997#23899786#24202425#24206609#2437256#2437425#2437579#24643945#2495876#3053732#3056165#3057077#3675092#3676103#3810041#6340678#6341301#6344426#6344511#6348981#6349074#6349348#650083#6906412#7096275#7098634#8147710#8148512#8149930#8215437#9511238#9523102#9525340#9998849# 83

10000000#10031824#10031914#10032083#10051814#10054264#10054266#10057931#10796381#10797230#10798370#124944 2#12639136#12639224#12767235#13589627#13654262#14004166#15359872#15519562#15545789#15717048#15861081#1640 3125#16408909#16772604#16777065#16777331#16893615#16894157#17411107#1747171#17686963#18404751#18407588#18 460931#18774055#18774327#18779557#18891510#19649923#20067910#20919652#21082708#21415206#21418381#2142163# 2142300#2144882#21776432#21776531#21776813#21779316#22041921#22095079#22123206#22166437#22355349#22576314 #22576770#23484828#23504032#23816587#23818335#23832830#23892289#23897997#23899453#24205638#2431138#243142 5#2438799#2495506#2498101#3676103#3810041#3812314#3813005#5137973#5162268#5163213#5163718#5167526#5168589 #5502919#5821609#5822336#5822408#5823605#5863251#5864721#5980005#5981316#5982503#5982714#5988630#5989482# 6344426#6344511#6345595#6345858#652868#6907007#7096275#7097650#751572#759618#8000173#8004741#8141366#8149 822#8198583#8199141#8215437#8219318#9523384#9529114#9529461#9992065# 119

10000000#10031824#10031914#10032083#10051814#10054264#10054266#10057931#10796381#10797230#10798370#124944 2#12639136#12639224#12767235#13589627#13654262#14004166#15359872#15519562#15545789#15717048#15861081#1640 3125#16408909#16772604#16777065#16777331#16893615#16894157#17411107#1747171#17686963#18404751#18407588#18 460931#18774055#18774327#18779557#18891510#19649923#20067910#20919652#21082708#21415206#21418381#2142163# 2142300#2144882#21776432#21776531#21776813#21779316#22041921#22095079#22123206#22166437#22355349#22576314 #22576770#23484828#23504032#23816587#23818335#23832830#23892289#23897997#23899453#24205638#2431138#243142 5#2438799#2495506#2498101#3676103#3810041#3812314#3813005#5137973#5162268#5163213#5163718#5167526#5168589 #5502919#5821609#5822336#5822408#5823605#5863251#5864721#5980005#5981316#5982503#5982714#5988630#5989482#6 344426#6344511#6345595#6345858#652868#6907007#7096275#7097650#751572#759618#8000173#8004741#8141366#8149822#8198583#81 99141#8215437#8219318#9523384#9529114#9529461#9992065# 119

图 3　银凯旋叶片 10 万条序列与 1 个 100 万条序列比对结果排序去重复后匹配序列

Fig.3　Matching sequence after sorting and removing duplicated ones of 100 000 sequences aligned to one set of 1 million sequences of Yinkaixuan blade

利用得到的90 298条比对结果的序列号从序列文件找出各自150 mer的序列,利用这些序列将其首尾20 mer序列作为匹配序列与找到的每一行内各自的匹配序列进行拼接,连接后得到加长的拼接序列,得到90 298条拼接序列,拼接过程运行时间较长约10 天(按照每个窗口 1 万条分别进行拼接时间只需要 3 d)。为了进一步合并拼接序列,去除了含有 10 个连续 polyA/C/G/T 的拼接序列,其中没有得到重叠群的序列有7 912条,序列超过150 mer的77 954条(重叠群序列见图 4), 拼接长度最长410 mer。

```
13790001%GTCCCTGTCTGGTACAATATGGTCGATCAAGGTCTTGTTAATGAACCAGTCTTTTCATTTTGGTTTAATCGCAATTC
TGGTGAGAATGAAGGGGGTGAAATTGTTTTTGGTGGGGTTGATTCTAATCATTTCAAGGGTGAGCATACTTTTTTTCTCGTGG
TTCAGAAAGGCTATTGGCAGTTTGATATGGGTGATGTCCTAATTGATGGAGAAACAACTGG TTTTTGTGGTGGTGGTTGTTCA
GCAATTGCCGATTCTGGAACCTCTTTGTTGGCAGGACCTACAACAATAATTACTCAAATTAATCATGCCATTGGAGCCTCTGG
GATTGTAAGCCAGGAATGCAAAGCAGTGGTTGCTCAATATGGGAAAATGATACTGGAAATGCTTTT 392
13790002 %AATTTTCCGGTGTCTGCTTGGTCAGCGTCTTTGTGCTGTAGGGTTTTTGCTTGAGCGCAAATTTGGATGCCCTCAGG
ATGTCGAAGGATGCGTGGTTGGTAAAGACATTTTCATTGTCCAGACACGCCCACAACCTCAATAAAAGGCTTCTTCCCATAAC
TTTTAACAATGTATGTATATCTACATTGTCAAAGGATGAGCTGTTGGTAAAGACATTTTCATTGTCCAGATATGCCTGCAACCT
TAAATAAGGCTTCTTCTCAGAACTTTAACCATGTATGTA 284
13790003%GGAAACCTGGTGAGGTCACGAAGCTAATTAATATGCGTGGTGAACTCCATAGCAGATTCCAAGTTGTGAAGCAGC
GAATGGCCCTTTGGGAAGAGATATCTGCAAACCTGCTGGCTGATGGGATAAACCGAAGTCCTGGCCAGTGCAAATCTCTATG
GACATCTCTGGTCCAGAAACATGAGGAAATCAAGGGTGATGAGAAAAGCAAGAAAAATTGGCCTTACTTTGAGGACATGAAT
AAGATTTTATCCGACTTGGAGGCAACAGCAGCAACAACAAAATGATTGGACTAGTGATATAATAGCTTACCTTCCAGAATTCT
TTTGCAAAGAGCTTTTGTCGAAGGGTAATTATCCAC 358
13790004%GCTGTGCTAATATGTTGTGATAGCTCACTTCCTGGTATAAGCTTCTCTGCTACAGATACATTATAATTTGGAACGAA
AATTACCTTCAAATAGCTGTTGACCTCGGGATCATTGTTGACAACAGCCCCAACATCATTCACTAGCTTGACTATTCTTTTTGC
ATTTGTATATGTCGCAAATGCTTTTCCTCCAAGCATGATGGTGCGTGGTGTAGTATTTTTCCTCTCTTCAGGGCTCATCTCCTTT
AACTTCTTGTACCTATAGACCGCACCCAGAATATTTAGTAGCTGTCTTTTATATTCATGGATCCGCTTGACCTGTATGTCAAAG
AGGCTATTTGGGTCAATGCTTACACCT 357
13790005%TGAGAATCTAAAGTGGCTCCAAATATTTAACAAGTATGACCTATACAGCAAGAGTAAAGTCAGAATCGATATCGA
AAAAGTCAAGCCATACTATCTCTCTCTCATTAAAAAGTACTTCCCTGAGAAGCTGAGATGGTGAAAATTTGGAGCTATGGGTT
GATAATTTTAGCGTGTGAATGGTGATGGAGACTTTTAGATGCATTTCTTTTTTTTAGGGGTTTTTGTTTTGATT 232
13790006%TCTGTATTTATAGAAGCTTTAATTTTCGTGCTTATCTTAATTGTTGGTTTAGTTTATGCATGGCGAAAAGGGGCATT
GGAATGGTCTTAGCTCCTGAATATTCAGACAATAAAAAGAAAAAAAAAAGATTGAGACAGTTATGAATTCCATTGAGTTTAC
CTTACTTGATCGAACAACCCAAAATTCAGTTATTTCAACTACATTCAATGATCTTTCAAATTGGTCAAGACTCTCTAGTTTATG
GCCCCTTCTCTATGGTACTAGTTGTTGCTTCATTGAATTTGCTTCACTAATAGGTTCACGATTCGATTTTGATCGTTATGGACTG
GTACCAAGATCAAGTCCTAGACAAGCAGACCTAATTTTAACAGCTGGAACAGTAACAATGAAAA 354
```

图 4 叶片 10 万条序列与 100 万条序列比对匹配序列第一次拼接部分序列

Fig.4 Partial sequences in first-time assembly of the blade's aligned matching sequences of 100 000 sequences and 1 million sequences

将获得的重叠群序列(77 954 条)进行再次相互比对(按照第一次比对的方法),没有得到匹配的序列有 16 478 条,得到超过 1 条以上重叠群的有37 256条,合并第一个序列号后为16 754条,将这些序列选取第一个序列头尾20 mer进行拼接得到新的重叠群,长度最长为1 174 mer(部分序列截图见图 5), 长度超过500 mer的有9 729条。因为是按照第一个序列号组合的,新的16 574条重叠群仍然有匹配的序列群,所以需要进一步拼接,按照初次拼接的方法用第一个序列的150 mer进行20 mer分割随机选取 9 个去找到匹配的重叠群,再次拼接后得到6 125个具有一个以上的匹配群序列,10 629个不再出现匹配,而且最多匹配序列号只有 218 个。进一步可以对 6 125 条序列进行 clustalw 多重比对后可以获得单一的重叠群序列(本文没有列出)。

2.2.3 叶片 10 万条序列拼接重叠群的表达量分析

上面的拼接中,得到的单一拼接序列包括再次拼接得到的16 478条及将含有 2 个以上的16 754条合并第一个序列号再比对匹配拼接后得到的10 629条,仍有 6 125 条序列含有 2 条及以上的匹配序列,如果进一步合并可以去除重复的匹配序列。因此,从 10 万条序列与 100 万条序列的比对匹配后拼接可以大致获得 3 万条左右的拼接序列。利用这些拼接序列的匹配序列按照拼接过程合并其表达数量,就得到它们的表达量计数。由于测序不排除正负链,所以计算表达量应该将正负链的表达进行计数。下面介绍得到的所有拼接序列的表达量。

第二次拼接得到的16 478条单一拼接序列正负链的表达量相当,最高3 369条,大于 100 的有 53 条,而大于1 000的只有 16 条(即千分之一),小于等于 10 的最多有 14 762 条。限于篇幅,图 6 列出了大于 100 的所有序列正负链表达量。

第二次拼接得到的含有 2 条以上匹配序列的16 754条序列作为拼接序列合并其表达量,正负链也是相当的,而表达量普遍较高,大于1 000的有 547 条,大于等于10 000有 20 条,大于 100 的有 3 040,小于等于 100 的有13 739条,小于等于 10 的4 204 条。图 7 也列出了 50 条序列的表达量。表达量较高是因为合并了多个匹配序列组,如图 7 的第一个序列(正链表达量为25 520)拼接合并了 86 组首序列号相同的序列组,其中单组表达序列数大于 500 的就有 58 组(最高为 758 个)。

13790001%GGTGCCCTCTGCAAAGTGTTATTTCTCTGTTGCGTGCTATTTACATTCCAAGTATAAGTCAAGCTACT
CTAGCACCTATAAGAAAAATGGGAAATCTGCTGAAATTCATTATGGAACCGGAGCTATTGCAGGTTTCTTTAGCC
AGGACCATGTGAAAATGGGTGATCTTGTTGTAAAGGATCAGGATTTTATTGAGGCAACTAGAGAGCCTGGCATC
ACATTCTTGGCTGCCAAGTTTGATGGTATACTTGGACTTGGTTTTCAAGAGATATCAGTTGGGAATGCTGTCCCT
GTCTGGTACAATATGGTCGATCAAGGTCTTGTTAATGAACCAGTCTTTTCATTTTGGTTTAATCGCAATTCTGGTG
AGAATGAAGGGGGTGAAATTGTTTTTGGTGGGGTTGATTCTAATCATTTCAAGGGTGAGCATACTTTTTTTCTCG
TGGTTCAGAAAGGCTATTGGCAGTTTGATATGGGTGATGTCCTAATTGATGGAGAAACAACTGGTTTTTGTGGT
GGTGGTTGTTCAGCAATTGCCGATTCTGGAACCTCTTTGTTGGCAGGACCTACAACAATAATTACTCAAATTAAT
CATGCCATTGGAGCCTCTGGGATTGTAAGCCAGGAATGCAAAGCAGTGGTTGCTCAATATGGGAAAATGATACT
GGAAATGCTTTTGTTCCCGTGGTTCAGAAAGGCTATTGGCAGTTTGATATGGGTGATGTCCTAATTGATGGAGAA
ACAACTGGTTTTTGTGGTGGTGGTTGTTCAGCAATTGCCGATTCTGGAACCTCTTTGTTGGCAGGACCTACAAC
AATAATTACTCAAATTAATCATGCCATTGGAGCCTCTGGGATTGTAAGCCAGGAATGCAAAGCAGTGGTTGCTC
AATATGGGAAAATGATACTGGAAATGCTTTTAGCACAGGCTGAGCCCCAAAAAATCTGCTCT 951
13790002%CAATTTTCCGGTGTCTGCTTGGTCAGCGTCTTTGTGCTGTAGGGTTTTTGCTTGAGCGCAAATTTGGA
TGCCCTCAGGATGTCGAAGGATGCGTGGTTGGTAAAGACATTTTCATTGTCCAGACACGCCCACAACCTCAATA
AAAGGCTTCTTCCCATAACTTTTAACAATGTATGTATATCTACATTGTCAAAGGATGAGCTGTTGGTAAAGACATT
TTCATTGTCCAGATATGCCTGCAACCTTAAATAAGGCTTCTTCTCAGAACTTTAACCATGTATGTACTTCCCATAA
CTTTTAACAATGTATGTATATCTACATT GTCAAAGGATGAGCTGTTGGTAAAGACATTTTCATTGTCCAGATATGC
CTGCAACCTTAAATAAGGCTTCTTCTCAGAACTTTAACCATGTATGTA 418
13790004%CGCACCACAGAGAAAGAAATTCTCCTCCCCAATTTCCTCCCTGATTTCCACATTAGCCCCATCTAATG
TTCCTATGATGAGACAGCCATTGAGTGCAAACTTCATGTTGCTTGTGCCACTTGCCTCCATACCCGCTGTGCTAA
TATGTTGTGATAGCTCACTTCCTGGTATAAGCTTCTCTGCTACAGATACATTATAATTTGGAACGAAAATTACCTT
CAAATAGCTGTTGACCTCGGGATCATTGTTGACAACAGCCCCAACATCATTCACTAGCTTGACTATTCTTTTTGC
ATTTGTATATGTCGCAAATGCTTTTCCTCCAAGCATGATGGTGCGTGGTGTAGTATTTTTCCTCTCTTCAGGGCTC
ATCTCCTTTAACTTCTTGTACCTATAGACCGCACCCAGAATATTTAGTAGCTGTCTTTTATATTCATGGATCCGCT
TGACCTGTATGTCAAAGAGGCTATTTGGGTCAATGCTTACACCTATTCTTTTTGCATTTGTATATGTCGCAAATGC
TTTTCCTCCAAGCATGATGGTGCGTGGTGTAGTATTTTTCCTCTCTTCAGGGCTCATCTCCTTTAACTTCTTGTACC
TATAGACCGCACCCAGAATATTTAGTAGCTGTCTTTTATATTCATGGATCCGCTTGACCTGTATGTCAAAGAGGC
TATTTGGGTCAATGCTTACACCT 697

图 5　叶片 10 万条序列与 100 万条序列比对匹配序列第二次拼接部分序列

Fig.5　Partial sequences in second-time assembly of the blade's aligned matching sequences of 100 000 sequences and 1 million sequences

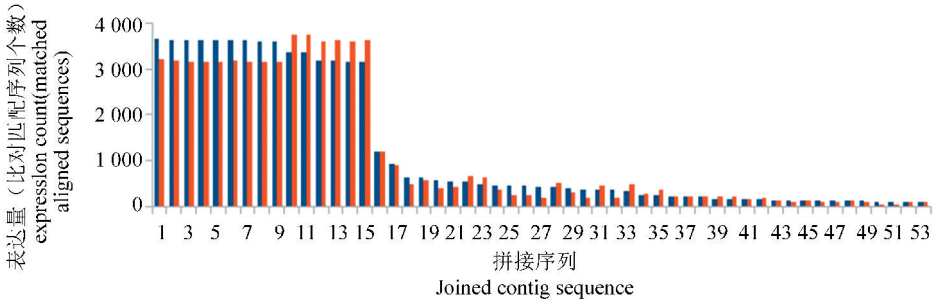

图 6　第二次拼接得到的单一序列表达量大于等于 100 的正负链表达量柱状图

Fig.6　Histogram of plus and minus strands with single sequence expression greater than or equal to 100 in second-time assembly

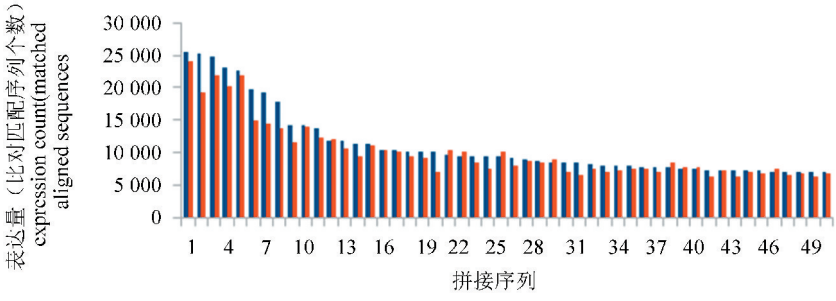

图 7　第二次拼接得到含有 2 个几以上匹配序列组合并后最高表达的 50 个序列的正负链表达量柱状图

Fig.7　Histogram of plus and minus strands expression of 50 highest expression sequences with more than 2 matching sequences after combination in second time assembly

RNA-Seq 测序过程首先对 RNA 进行了片段化，所以测序得到的序列是随机选取任何 RNA 片段获得的。由于一些序列比对没有找到匹配，部分序列进行第二次拼接没有找到匹配，因此，表达量的表示直接用第一次比对的 150mer 查询序列得到的匹配比对序列数作为该序列的表达量。

2.3 重叠群拼接序列注释初步分析

单一拼接序列与含有 2 个以上拼接序列组的拼接序列其是否真实反映了某个基因，需要进一步利用该序列与 NCBI 核酸数据库进行 blast 比对，从返回比对序列及其注释可以初步推知。

每次用 500 条序列（fasta 格式）与 NCBI 的 blastn 服务器进行比对，保存 text 格式的比对结果，提取得到了所有拼接序列的 blast 比对及其注释，完成所有 3 万条拼接序列的远程 blalstn 比对只需要半天。Blastn 程序只是返回比对打分大于 70%以上的比对，前面的两组拼接序列中再次拼接得到的16 478条单一序列 blastn 比对返回8 889条相似打分（Identities）高于 70%的结果，而再次拼接大于多于两条拼接序列的 16 754条序列进行 blastn 比对返回12 255条，说明较长的序列可以得到更多匹配的 blastn 比对。对从 16 478 条单一拼接序列进行 blastn 比对返回的8 889条注释进行第一条注释提取去重复后得到7 511条不同注释，其中一组 53 条为叶绿体全基因组序列注释，其他的均小于 10 条，出现 2 条及以上相同注释的有 1 034条，只有 1 条注释的6 476条，注释中出现山茶属（*Camellia*）植物的有 906 条。多于两条拼接序列的 16 754条序列进行 blastn 比对返回12 255条进行注释提取去除重复后得到8 418条注释，其中两组（162+36）为两个不同植物叶绿体全基因组序列注释，只有 1 条注释的6 462条，2 条及以上的1 954条，出现山茶属（*Camellia* ）植物的有1 962条。因此，两组拼接序列合并的有用基因注释只有15 678条，即是用 10 万条叶片测序得到的序列与 100 万条序列比对后拼接得到的 3.3 万条拼接序列得到的基因注释，没有得到有效注释的拼接序列为 17 554 条。表达量超过10 000的有 20 条，其中 16 条得到有效注释，结果见表 3，注释中仍然有相同基因名的。本文的注释分析是初步的（只是对 blastn 返回注释的第一个进行了提取），而且其中出现 *Camellia* 的注释仍有重复，需要进一步分析。

3 讨 论

3.1 100 万条序列作为分析 RNA 测序基因表达量的可行性

本文利用开源系统文本过滤命令组合脚本对高通量 RNA 测序（RNA-Seq）的单次测序结果（6G 高质量序列，约 2 500 万条）进行基因序列拼接及表达量分析，用 10 万条对任意 100 万条（按照每 1 万条进行了随机重排序）序列进行比对（利用随机 9 组20 mer 长度寻找匹配序列）得到的匹配序列中拼接得到了重叠拼接序列，并获得了正负链的表达量。序列拼接长度最长达到1 174 bp，10 万条序列只有 8 千条没有得到匹配序列，拼接后得到单一拼接序列为16 478条，多于 1 条的有16 754条，为此这些序列再次进行了拼接，去除重复后得到10 629条单一拼接序列，而仍然有6 125条出现多于 1 条的匹配序列。

利用随机 9 组20 mer长度寻找匹配序列（测序长度150 mer，进行每隔 5mer 切割可以得到 27 组 20mer 长度序列），得到的表达量与用 27 组20 mer序列进行比对得到的表达量很一致（见表 1），因而为了加快运行速度只是选用随机 9 组20 mer进行比对就够了。利用叶片 10 万条随机 9 组20 mer对多个叶片或花瓣的 100 万条及 300 万条进行的比对也可以看到表达量很一致（见表 2）。从叶片 10 万条比对结果进行拼接得到的单一（16 478 条）和多于 1 条（16 754 条）的拼接序列进行了表达量分析，正负链的表达量一致（见图 6、图 7）。因此，从以上结果分析可以看到，对高通量测序进行拼接和表达量只随机选取 100 万条及分析正链就可以了。表示表达量的方法很多，如使用 TPM（Tran-scripts per million）来表示高通量测序基因表达量[3]。本文对 RNA-Seq 测序的数据进行了随机打散（如图 1），用每个150 mer查询序列对 100 万条序列进行比对并对比对匹配序列进行了重叠群序列拼接，第一次拼接可以获得最长410 mer序列，其表达量直接用查询序列进行比对获得的匹配序列数表示（见图 6）。我们进行了克瑞默的花瓣和叶片的各自 10 万条序列分别比对花瓣和叶片的 3 组 100 万条序列，对所有序列的匹配数进行了方差分析，差异不显著，比对匹配数高达 5 千的其标准差不超过 125（结果未附）。因此，考虑到测序时 RNA 片段选取随机性及对测序序列进行了随机打散，表达量的表示用每个150 mer查询序列与单个 100 万条序列比对获得的比对匹配数是合适的。

3.2 开放源代码操作系统用于分析高通量基因表达量及从头拼接的优势

开源系统（本文使用 FreeeBSD 系统）的文本过滤功能非常强大，组合命令行脚本可以完成很多分析工作。本文将测序序列作为文本进行处理，从上面结果可以看出使用文本过滤命令用于分析高通量测序序列完全可行的，不少序列分析平台也借助开源系统完成特定任务[6,11]。

表 3　最高表达的 16 条重叠拼接序列 Genbank 注释、blastn 比对打分及表达量(每 100 万)
Table 3　Genbank annotations, blastn aligned scores, and expression (per million) of 16 contig-joining sequences with highest expression

拼接序列号 #长度(bp)	Genbank 号 /注释#长度(bp)	打分,E-value	相似度(%),空格(%)	表达量 (正/负链)
13790063#566	AB623937.1 *Camellia sinensis* DNA, SSR marker, MSE0250#825	586 bits (317), 4×10^{-163}	329/335 (98) 0/335 (0)	25 520/24 086
13791076#555	XM_018969565.1 Juglans regia 30S ribosomal protein S10, chloroplastic #1080	145 bits (78), 2×10^{-30}	158/196 (81), 7/196 (4)	25 165/19 213
13790015#768	EF011075.1 *Camellia sinensis* ribulose-1,5-bisphosphate carboxylase/ oxygenase #769	669 bits (362), 0	378/386 (98), 0/386 (0)	24 894/21 813
13790084#1100	DQ444292.2 *Camellia sinensis* metallothionein-like protein mRNA, complete #531	573 bits (310), 3×10^{-159}	398/440 (90), 8/440 (2)	23 201/20 175
13790022#687	XM_022173664.1 *Helianthus annuus* ubiquitin-conjugating enzyme E2 #790	446 bits (241), 7×10^{-121}	305/337 (91), 0/337 (0)	22 682/21 981
13790737#876	XM_012628700.1 *Gossypium raimondii* stem-specific protein TSJT1-like #986	337 bits (182), 4×10^{-88}	309/371 (83), 6/371 (2)	19 872/14 922
13790753#815	XM_006380342.1 *Populus trichocarpa* hypothetical protein (POPTR_0007s05050g) #823	492 bits (266), 9×10^{-135}	334/368 (91), 0/368 (0)	19 339/14 422
13790521#824	HQ660372.1 *Camellia sinensis* chloroplast ferredoxin I mRNA,complete cds; #583	968 bits (524),0	541/549 (99), 1/549 (0)	17 917/13 717
13790046#1146	XM_010660236.2 *Vitis vinifera* ribulose bisphosphate carboxylase/oxygenase #1685	787 bits (426),0	605/694 (87), 2/694 (0)	14 247/14 064
13790792#1082	KF472133.1 *Camellia sinensis* cultivar Longjin 43 galactinol synthase 1 (GS1) #1435	641 bits (347), 8×10^{-180}	367/376 (98), 3/376 (1)	11 810/12 103
13791924#583	XM_002283173.3 *Vitis vinifera* glutathione S-transferase F9 (LOC100233043), #1040	257 bits (139), 2×10^{-64}	217/255 (85), 4/255 (2)	11 397/9 481
13791162#643	XM_016031456.1 *Ziziphus jujuba* mitochondrial import inner membrane #1384	171 bits (92), 3×10^{-38}	114/125 (91), 0/125 (0)	11 337/11 006
13790620#1089	EF011075.1 *Camellia sinensis* ribulose-1, 5-bisphosphate carboxylase/oxygenase #769	701 bits (379), 0	399/409 (98), 0/409 (0)	10 398/10 323
13791871#435	XM_002263902.4 *Vitis vinifera* 40S ribosomal protein S29 (LOC100241311), #636	231 bits (125), 1×10^{-56}	162/180 (90), 1/180 (1)	10 379/10 037
13791853#931	KF472133.1 *Camellia sinensis* cultivar Longjin 43 galactinol synthase 1 (GS1) #1435	688 bits (372), 0	372/372 (100), 0/372 (0)	10 146/9 335
13791694#324	XM_006360973.2 *Solanum tuberosum* sphingolipid delta(4)-desaturase #1305	156 bits (84), 8×10^{-34}	145/175 (83), 1/175 (1)	10 071/6 920

高通量测序由于数据量大,使用开源系统分析就具有明显优势。本文首先对测序序列按照每 1 万条进行随机排序合并(主要为 split 和 sort 命令),从表达量分析结果可以看到随机排序是很有效的。比对选取全长测序序列(150 mer)显然会减慢比对进程,而较短序列也会得到太多非特异性匹配比对,而简单选取某些区段又会减少比对的匹配数量,为此,本文进行了每隔开 5 个碱基选取20 mer序列作为查询序列,150 mer测序序列可以得到 27 组20 mer序列组,全部用于比对对运算显然也不利,从表 1 中看到,用 27 组20 mer与用随机 9 组20 mer进行比对得到的表达量是一致的,对运算速度影响不大,可以每次同时分析 6 组比对计算机运行速度不受影响,平均速度每天完成 1.5 万条比对(单组),一周就可以同时得到 6 组 10 万条序列与 100 万条的比对。如果将每 10 万条比对 100 万条分成 10 个窗口只需要 2 天就可以。

本文也利用文本过滤命令组合脚本进行了从比对匹配序列中分析获得重叠序列并进行拼接,银凯旋叶片 10 万条比对匹配序列约 9 万多条很有多于 1 条匹配序列的,利用头尾20 mer进行对接也实现了重叠群序列的拼接,拼接长度最长1 174 bp(见图 4、图 5)。拼接过程比比对较耗时,约每天每个窗口完成7 000~1 万条,完成 9 万条比对的拼接约需要 10 d 左右,得到77 954条拼接序列(先行去除含有连续 10 个以上的 polyA/C/G/T),7 912 条没有得到拼接。第二次拼接用得到的拼接序列再次进行首尾 20 mer拼接,得到单一的序列16 478条,多于 1 条的按照第一个序列号相同的进行合并后得到 16 754 条,16 754 条序列组的序列利用150 mer测序序列按照随机 9 组20 mer再次进行比对及首尾拼接后得到10 629条单一序列,6 125条仍然含有多于 1 条以上的序列。初步对得到的拼接序列进行 blastn 比对(NCBI)获得了其注释(最高表达的16 条序列的注释见表 4),超过半数拼接序列可以得到有效注释。

本地比对用 blast2(本文使用的系统亦安装了 NCBI 的 blast-2.2.26-ia32-freebsd)也可以进行,利用 RNA-Seq 测序全长150 mer进行其比对速度仍然很快(结果未附)。由于 blast 对双链进行查找比对,不能对正负链的表达量进行分开,与本文利用文本比对得到的匹配转录本数量因而不能直接相比较,也不利于利用本文的拼接方法进行拼接。另外,部分序列利用 blast 无法比对而相同序列用文本比对可以得到匹配序列,且得到的拼接序列进行远程 blast 比对可以获得基因注释(结果未附)。进一步对两者得到的匹配序列是否一致,用银凯旋叶片 10 万条序列查找 100 万条叶片序列中9 701条未得到任何正链匹配的序列(结果 2.2.2),用 blast 去进行比对,发现其中6 247条亦找不到匹配序列,找到一个匹配序列的1 547条,只有 2 个序列多于 100 条(174 和 197),其他匹配介于 2~60 条(结果未附),这些匹配很可能来自同时 blast 查找互补链,符合正常情况下正链匹配少负链也是相对少的情形。上面的比较分析也进一步说明本文的文本比对得到的比对匹配转录数能够确切反映组织内基因的表达水平。

4 结 论

综合以上分析表明,本文建立的利用开源系统文本过滤命令组合脚本对高通量 RNA-Seq 测序进行序列比对及拼接,可以快速获得基因的表达信息。分析方法简捷、高效,用于进行 RNA-Seq 测序分析是完全可行的。

1)本文进行的序列比对前提是序列在 RNA-Seq 测序中被重复测序,直接用序列本身进行文本匹配即可;

2)本文通过将短序列进行每隔 5 个核苷酸取连续 20 个核苷酸序列并进行随机取其中 9 个 20 mer与 100 万条目标序列比对,较好反映匹配目标相似序列数;

3)本文方法可同时进行正链与其互补链分别与目标序列匹配查找,获得两者的表达量相当,表达量用每个查询序列与 100 万条序列比对匹配的序列数表示;

4)本文利用查询序列两端20 mer进行重叠群组内拼接,进行两次拼接可获得大于 1 Kb 重叠序列,因此,进行转录组从头组装是完全可行的;

5)根据从 NCBI 远程 BLAST 比对分析可提取重叠序列的注释,结合重叠群表达量可得到组织转录组信息。

参考文献(References)

[1]GRIFFITH M, WALKER J R, SPIES N C, et al. Informatics for RNA Sequencing: A web resource for analysis on the cloud[J]. PLoS Computational Biology, 2015, 11(8): e1004393. DOI: 10.1371/journal.pcbi.1004393.

[2]TSIRIGOS A, HAIMININ N, BILAL E, et al. GenomicTools: a computational platform for developing high-throughput analytics in genomics[J]. Bioinformatics, 2012, 28(2):282-3. DOI: 10.1093/bioinformatics/btr646.

[3]CONESA A, MADRIGAL P, TARAZONA S, et al.A survey

of best practices for RNA-Seq data analysis[J]. Genome Biology, 2016, 17(1):181. DOI: 10.1186/s13059-016-0881-8.

[4]POPLAWSKI A, MARINI F, HESS M, et al.Systematically evaluating interfaces for RNA-Seq analysis from a life scientist perspective[J]. Briefings in Bioinformatics, 2016, 17(2):213-223. DOI: 10.1093/bib/bbv036.

[5]TRAPNELL C, ROBERTS A, GOFF L, et al. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks[J].Nature Protocol, 2012, 7(3):562-78. DOI: 10.1038/nprot.2012.016.

[6]GRNING BA, FALLMANN J, YUSUF D, et al. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy [J]. Nucleic Acids Research, 2017, 45 (Web Server issue): W560 - W566. DOI: 10.1093/nar/gkx409.

[7]WANG S, GRIBSKOV M. Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis [J]. Bioinformatics, 2017, 33 (3): 327 - 333. DOI: 10.1093/bioinformatics/btw625.

[8]HONAAS L A, WAFULA E K, WICKETT N J, et al. Selecting superior de novo transcriptome assemblies: lessons learned by leveraging the best plant genome[J]. PLoS One, 2016, 11 (1): e0146062. DOI: 10.1371/journal.pone.0146062.

[9]CABAU C, ESCUDIÉ F, DJARI A, et al. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies [J]. PeerJ,2017, 5(7):e2988. DOI: 10.7717/peerj.2988.

[10]DAVIDSON N M, HAWKINS A D K, OSHLACK A, et al. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes [J]. Genome Biology, 2017, 18 (1): 148. DOI:10.1186/s13059-017-1284-1.

[11]SREEDHARAN V T, SCHULTHEISS S J, JEAN G, et al. Oqtans: the RNA-seq workbench in the cloud for complete and reproducible quantitative transcriptome analysis[J]. Bioinformatics, 2014, 30(9):1300-1. DOI: 10.1093/bioinformatics/btt731.

[12]ICAY K, CHEN P, CERVERA A, et al. SePIA: RNA and small RNA sequence processing, integration, and analysis[J]. BioData Mining, 2016, 9(1) 1-18. DOI: 10.1186/s13040-016-0099-z.

[13]PERTEA M, KIM D, PERTEA G M, et al. Transcript-level expression analysis of RNA-Seq experiments with HISAT, StringTie and Ballgown [J]. Nature Protocol, 2016,11(9):1650-67. DOI: 10.1038/nprot.2016.095.