基于转录组测序数据分析及高通量 GO 注释理论的研究

刘粉香[1,2],杨文国[3*],孙勤红[2]

(1. 南京工业职业技术学院,江苏南京 210023;2. 三江学院,江苏南京 210012;3. 南京中医药大学,江苏南京 210023)

摘要 随着二代测序技术的快速发展,转录组测序在越来越多的动植物中完成,人们获得了大批量的转录组数据序列。如何从这些海量的序列数据中挖掘具有生物意义的信息已成为很多研究的关键所在,对未知基因的功能进行预测和注释就是其中一个重要的问题。转录组序列的功能注释是功能基因组学研究的一项重要内容,基因本体论(gene ontology,GO)注释目前是一种最重要的功能注释方式。介绍了利用生物信息学软件进行转录组测序数据分析过程,包括数据质量控制和过滤、从头拼接(*De novo* assembly)、同源比对以及大规模 GO 注释,为从事转录组测序特别是非模式植物转录组测序研究者在数据分析方面提供参考。

关键词 二代测序;转录组;从头拼接;GO 注释

中图分类号 Q-3 **文献标识码** A **文章编号** 0517-6611(2018)31-0088-04

Transcriptome Sequencing Data Analysis and High-throughput GO Annotation

LIU Fen-xiang[1,2], YANG Wen-guo[3], SUN Qin-hong[2] (1. Nanjing Institute of Industry Technology, Nanjing, Jiangsu 210023; 2. Sanjiang University, Nanjing, Jiangsu 210012; 3. Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210023)

Abstract With the development of sequencing technology, the transcriptome sequencing has been completed in more and more plants. A large number of transcriptome sequence data were obtained. How to mine biologically meaningful information from these massive serial data has become the key point of many researches. Predicting and annotating the function of unknown genes is an important issue. Functional annotation of transcriptome sequences is an important part of functional genomics. Gene Ontology (GO) annotation is currently one of the most important functional annotation methods. We introduced the analysis of transcriptome sequencing data using bioinformatics software, including data quality control and filtering, *De novo* assembly, homology comparison and large-scale GO annotation, which provided a reference for researchers engaged in transcriptome sequencing, especially non-model plant transcriptome sequencing in data analysis.

Key words Next-generation sequencing; Transcriptome; *De novo* assembly; GO annotation

广义上的转录组是指生物体细胞或组织在特定状态下所转录出来的所有 RNA 的总和,包括 RNA(即 mRNA)编码蛋白质和 RNA(ncRNA,如 rRNA、tRNA、microRNA 等)非编码蛋白质;狭义上的转录组通常指所有 mRNA 的总和[1]。转录基因组学研究被转录的基因,是挖掘转录基因的功能基因极其重要的途径,功能基因组学研究在基因进化、遗传育种等研究中具有非常重要的意义[2]。转录组研究的技术手段大体上有 EST 序列构建、芯片技术和二代测序技术等。随着二代测序(next generation sequencing)技术的发展和应用,许多物种已经完成了转录组测序。早在 2008 年,Nagalakshmi 等[3]利用 RNA-Seq 技术进行了酵母转录组测序。近年来,越来越多的无参考基因组物种先后完成了转录组测序。2012 年,Zhang 等[4]对不同发育阶段的 6 个麻竹花器官的转录组进行测序,并分析基因的差异表达,最后预测了 81 个转录因子家族在麻竹花组织发育过程中的差异表达。Mudalkar 等[5]于 2014 年对亚麻转录组进行测序,并且在拼接得到的 53 854 个转录本序列数据中发现了 19 379 个 SSR 标记位点。同年,Upadhyay 等[6]通过比较天冬根组织和叶组织转录组拼接结果,发现在根组织中特异表达的基因,从而推测其在体甾皂苷元合成中表达的基因。从目前公布的这些无参考基因组的物种转录组测序数据的研究成果[4-7]来看,转录组测序生物信息学分析的主要内容有:①功能注释、分类及代谢途径分析;②预测编码序列框(CDS);③样品间基因差异表达(2 个及 2 个以上样品);④分子标记(SNPs、SSR)的研究进展。同时,这些研究也反映出转录组测序技术的几个突出优点:①任何物种都可以进行完整的转录组分析(无需了解物种的基因或基因组的信息,可以直接在任何物种中进行最全面的转录组分析);②更准确的基因注释;③不仅可以检测已知的转录本,还可以识别新的基因、鉴定变异体。转录组测序作为一种更为精确的测定方法,在转录组学的应用中具有革命性的意义,开辟了转录组学研究的新纪元[8]。

基因注释是基于“同源基因,功能相似”假设的基础[9-10],利用生物信息学方法来搜索未知基因序列与公共数据库中序列的相似性,并通过与数据库中已注释的基因的的同源性来预测未知基因的功能。核酸数据库主要有 GenBank(NCBI)、EMBL 和 DDBJ,蛋白质数据库主要有 UniProt 和 PDB 等,搜索比对软件主要有 Blast 系列软件等。目前基因功能分类主要有 2 种方法:KEGG 功能分类和 Gene Ontology(简称 GO)分类。GO 是国际标准的基因功能分类体系,它提供了一套动态更新的标准词汇表(controlled vocabulary)来全面描述生物体基因和基因产物的性质[11]。GO 共有 3 个本体(ontology),分别描述的是分子功能(molecular function)、细胞组分(cellular component)和生物过程(biological process)[12]。GO 的基本单位是 term[13](节点),每个 term 都对应一个属性。GO 功能分析,一方面给出了基因 GO 功能的分类注释,另一方面给出了基因 GO 功能的显著性富集分析。GO 功能分类注释给出了具有某个 GO 功能的基因数目统计量的基因列表。GO 功能显著性富集分析给出了与基因组背景相比显著富集基因的 GO 功能条目,因而给出了显著相关的基因的生物学功能。该分析首先将所有基因映射到

基金项目 江苏省青年基金项目(BK2015100)。

作者简介 刘粉香(1975—),女,江苏高邮人,讲师,博士,从事生物信息学研究。*通讯作者,副教授,硕士,从事生物信息学研究。

收稿日期 2018-08-14;**修回日期** 2018-09-10

Gene Ontology 数据库的各个 term,计算每个 term 的基因数,然后使用超几何测试来识别 GO 条目,与整个基因组背景相比,显著富集的 GO 条目。转录组测序技术的应用和发展,将大大推动功能基因组学的发展。

尽管转录组测序已成为获得大量植物功能基因组数据的重要技术,但是非模式植物转录组研究也面临许多挑战。首先,从转录组测序中获得大量的短序列,数据分析时对计算机运算速度和内存有较高的要求。其次,由于缺乏参考基因组信息,非模式植物转录组的构建和量化必须依靠从头拼接(*De novo* assembly),错误拼接、不完整拼接、拼接得到的冗余数据都将影响下游分析的质量。另外,非模式植物转录组分析过程包括使用多个在线或本地化数据库、安装和使用 Linux 平台应用程序,以及选择和评估大规模计算参数等。所有这些都将给研究者带来不少困难。笔者以单端测序数据为例,详细介绍非模式植物转录组测序数据的分析过程,包括原始测序数据质量控制和从头开始拼接序列获得转录本序列(transcripts)、Blast 同源比对、Blast2go 进行大规模 GO 注释和基因功能预测等。这套非模式植物转录组分析流程为研究者在相关软件安装、使用方法以及注意事项等方面提供参考。

1　转录组测序数据分析

1.1　测序数据质量控制　笔者以鹰嘴豆(chickpea)的根及芽组织转录组测序数据为例介绍转录组测序数据分析过程、软件使用和结果说明。该数据包含 31 028 774 条长度为 51 bp 的原始序列,可根据数据号 SRR063784 直接从 NCBI 网站的 SRA 数据库下载[14]。

从 SRA 上下载的鹰嘴豆转录组数据为 sra 格式文件,这种文件不能直接使用软件进行分析,需要转化为 fasta 或 fastq[15]格式文件才可以使用。所以,首先使用 sratoolkit(http://www.ncbi.nlm.nih.gov/Traces/sra)中的一个可执行程序 fastq-dump,将下载的 sra 格式的序列文件(SRR063784.sra)转化为 fastq 格式的文件(SRR063784.fastq)。

获得原始数据后,需进行序列的从头拼接,这是后续研究的基础。原始数据中具有大量的测序接头序列、低质量碱基盒未检测碱基(用 N 表示)将严重影响后续组装的质量。所以,首先需要对测序数据做一些预处理,经过质控后得到的数据即为有效数据,也称为 clean data。一般使用 FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)查看 raw data 的质量,为此可执行如下命令:./fastqc -o./ -f fastq SRR063784.fastq,其中,-o 指定文件输出路径,-f 给出输入序列文件格式。FastQC 输出的结果为一个压缩文件,解压后,打开文件夹中 html 格式文件可看到序列文件一些统计信息。统计信息包括每个碱基位点的平均质量值(per base sequence quality)、每条序列平均质量值的分布(per sequence quality scores)、序列 GC 含量(per sequence GC content)、序列是否含有接头(adapter content)等 12 项内容。通过结果报告概要(summary)就可以对数据的情况有一个初步的了解,每一项统计分析前都有一个标志,这种标志共有 3 种颜色:绿色、黄色和红色。绿色代表“通过”(pass),黄色代表“警告”(warn),红色代表“不合格”(fail),FastQC 以此向用户指出需要注意序列数据哪些方面。

了解数据大致情况后,使用工具包 NGS QC Toolkit (http://59.163.192.90:8080/ngsqctoolkit/)中的 IlluQC.pl 对 raw data 进行进一步过滤,为此可执行如下命令:perl IlluQC.pl -se SRR063784.fastq N A -s 20 -l 70 -o./SRR063784_NGS/,其中,-se 给出输入的 single-end 的序列文件,N 表示不过滤接头接头文库(FastQC 结果显示 reads 不包含接头),A 表示自动识别 fastq 文件的版本(不同版本采用不同的质量标示方案),-s 设置 Phred 值,-l 设置大于设定 Phred 值的 read length 占该序列长度的比例,-o 指定输出文件路径。在执行上述命令时,当 raw data 中的 reads 的 Phred 值≥20(即 base calling 正确率要大于等于 99%)的碱基数≥reads 长度的 70%时,reads 被保留,否则被过滤掉。

程序运行结束后,所有输出的结果文件都保存在文件夹 SRR063784_NGS 中。其中,output_SRR063784.html 中记录了 raw data 质量和数据过滤记录,SRR063784.fastq_filter 是过滤后的序列(clean data)文件。过滤后,31 028 774 条 raw reads 中有 24 735 426 条(79.72%)高质量 reads 保留下来,保留下来的 clean data 将用于从头拼接。

1.2　从头拼接　从头拼接是将 *De novo* 测序得到的序列拼接组装成连续较长的序列[16]。将这些拼接后得到的较长序列与公共数据库中公布的基因或蛋白质序列进行同源比对分析(Blast),最终可以确定基因序列。从头组装是进行无参考序列及短序列组装、快速获得表达基因的一种有效的方法。近年来,研究者们设计了各种适用于 *De novo* assembly 的软件。目前,常用的拼接软件有 Trans-ABySS(http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss)、SOAPdenovo(http://soap.genomics.org.cn/soapdenovo.html)、Trinity(http://trinityrnaseq.sourceforge.net)、Velvet(http://www.ebi.ac.uk/~zerbino/velvet)、Velvet/Oases(http://www.ebi.ac.uk/~zerbino/oases)。

该研究使用 Velvet 结合 Oases 进行转录组序列的 *De novo* 拼接。由于 Velvet 默认的 K-mer 值上限为 31,若要使用的 K-mer 值大于 31,则需要重新编译软件。例如,若将 K-mer 值上限设置为 57,则可执行如下编译命令:make ‘MAXKMERLENGTH=57’,另外,部分 Velvet 算法支持多核计算,对 OPENMP 选项进行编译后,这部分程序即可使用多核运行。如需编译 OPENMP 选项,可执行如下编译命令:make ‘OPENMP=1’,编译好软件后,首先选择 5 个不同的 K-mer 值(27、31、37、41、47)进行单端测序序列的拼接,并执行如下 velveth 命令:./velveth chickpea 27,47,10 -short -fastq SRR063784_hq.fastq、./velveth chickpea 31,41,10 -short -fastq SRR063784_hq.fastq,其中,chickpea 为输出文件名称;27,47,10 表示输入多个 K-mer 值,27≤K≤47(K 为奇数),10 为 K 值步长(步长为偶数);-fastq 指出输入文件格式为 fastq;-short 指出输入数据类型。结果将产生 5 个文件夹,分

别为 chickpea_27、chickpea_31、chickpea_37、chickpea_41、chickpea_47。每个文件夹里包含 2 个文件,分别是 Roadmap 以及 Sequences。

其次运行 velvetg,由于这里使用 Velvet 结合 Oases 进行转录组测序序列组装,所以运行 velvetg 时只设置 1 个参数。具体执行如下命令:./velvetg chickpea_27 -read_trkg yes,该命令中的-read_trkg 参数要求结果给出更细致的拼接描述(yes 表示打开该选项)。当程序运行结束时,屏幕上会显示 nodes 数 n_{50} 的值、最长 contig 的长度(bp)以及总的组装序列的大小。同时,文件夹 chickpea_27 中将产生 8 个文件,分别是 contigs. fa、LastGraph、Pregraph、Sequences、Graph2、Log、Roadmaps 和 stats. txt。contigs. fa 即为拼接得到的 contigs 文件,Log 文件记录 Velvet 运行情况(包括开始时间、软件版本、执行命令、运行结果),stats. txt 文件则记录对拼接得到的每一条 contig 的描述。对 velveth 产生的其他 4 个文件夹进行同样的操作(分别运行 velvetg),最终产生 5 个组装结果。

比较这 5 个拼接结果的 n_{50} 长度、contigs 的数目(nodes)和 contigs 的平均长度这 3 个参数,选择最好的拼接结果。如图 1 所示,当 K-mer 为 37 时,拼接得到的 n_{50} 长度最长(620 bp)、最大的 contig 长度最长(7 339 bp)、contigs 的平均长度较长(202 bp),所以最终选择 K-mer 值为 37 时的拼接结果进行后续分析。

最后运行 oases 对 Velvet 拼接得到的 contigs 进行进一步的拼接,最终获得转录本(transcripts)。运行 oases 的前提是安装并运行了 Velvet,并且需要将 Velvet 所在的文件夹命名为"velvet"或者指明 Velvet 的路径,为此可执行如下命令:make 'VELVET_DIR = ~/software/velvet',值得注意的是 oases 默认的 K-mer 值上限为 31,若使用的 K-mer 值大于 31,则在使用软件前需重新编译 K-mer 的值。若将 K-mer 值上限设置为 75,可执行如下命令:make 'MAXKMERLENGTH = 75',运行 oases 时执行如下命令:oases chickpea_37,运行结束后文件夹 chickpea_37 中产生 2 个文件,分别是 transcripts. fa 和 contig-ordering. txt。transcripts. fa 为包含组装得到的 transcripts 文件,而 contig-ordering. txt 记录了每一个 transcripts 中 contigs 的组成情况(图 1)。

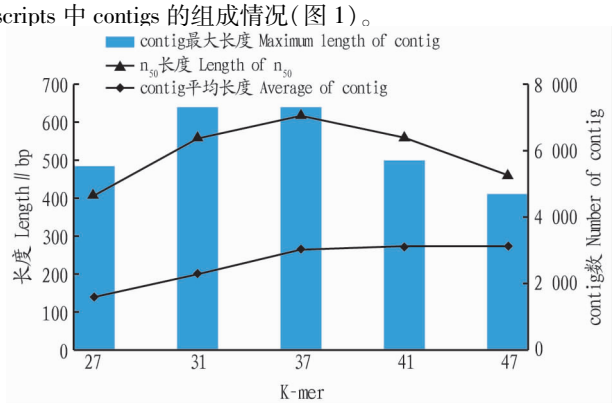

图 1 使用不同 K-mer 值进行从头拼接的结果

Fig. 1 The use of different K-mer values from the head assembly

将拼接得到的 contig 或 scaffold 从大到小排序,累加其长度,当累加长度达总 contig 或 scaffold 长度 50%的时候,最后一个 contig 或 scaffold 的长度即为 n_{50} 的值。

1.3 基因注释与功能分类 基因注释是通过比对已知数据库中已被注释的同源基因的信息推断未知基因的功能。Blast+(ftp://ftp. ncbi. nlm. nih. gov/ blast/executables/blast+/LATEST/)中 Blastx 的功能是将输入核苷酸序列翻译成蛋白,并将其与蛋白质数据库比对,最后输出几个相似度高的结果。该研究使用 Blastx 将拼接得到的 transcripts 比对到 nr 数据库(NCBI 非冗余蛋白质数据库)。

要进行本地 Blast 搜索,首先需要从 NCBI 的 ftp 站点下载并格式化数据库 nr. gz。将下载的 nr. gz 放在目录 ncbi-blast-2. 2. 25+/bin/中,解压后,利用文件夹 bin/中的可执行文件 makeblastdb 格式化数据库,为此可执行如下命令:makeblastdb - in nr - dbtype prot -parse_seqids -out nrdb,其中,-in(nr)输入待格式化的文件(nr),-dbtype(prot)给出数据库类型(蛋白质数据库),-parse_seqids 启动序列 ID 解析,-out(nrdb)指定输出文件名。

格式化数据库后,即可运行 Blastx 将拼接得到的 transcipts 比对到本地 nr 数据库,为此执行如下命令:./blast+/bin/blastx -query transcripts. fa -out transcripts. xml -db ~/software/blast+/bin /nrdb -outfmt 5 -evalue 1. 0E-6 -max_target _seqs 10 -num_threads 20,上述命令中,-query 给出输入待比对数据文件路径及数据文件名(transcripts. fa),-out 指定输出文件名(transcripts. xml),-db 指定用于比对的数据库名称(nrdb),-outfmt 指定 输入数据格式(xml 格式),-evalue 设置输出结果的 E-value 值,-num_threads:使用多线程运算。

拼接的结果中有 42 203 条 transcripts 参与比对,其中 38 622 条(91. 5%)transcripts 获得相似性搜索结果(基因注释)。此次比对获得的 hits 在大豆中的分布最多(47 520),其次是鹰嘴豆(33 898)。这样的结果表明,一方面参与比对的序列与豆科植物基因表现出显著的相似性,另一方面表明公共数据库中可获得的鹰嘴豆的基因组资源依然较少[13]。

Blast+只是一种预测新基因功能的基本工具,仅通过 Blast 的结果无法得到新基因的 GO 注释信息。可以将 Blast 搜索结果文件(xml 文件)作为 Blast2Go[17] 的输入数据,使用 Blast2Go 软件进行 GO 注释,最终得到与输入序列相关的 GO 注释信息,并将 GO 注释信息分为 molecular function、cellular component 和 biological process 3 类及其子类。

2 高通量 GO 注释工具 Blast2Go

目前,能进行基因产物功能注释的生物信息学软件或生物信息学方法有很多[18],但是对非模式物种测序序列进行大规模功能注释的软件不多。在获得 Blast 结果后,如果再到基因本体论网站查询相关的 GO 注释信息,将会浪费大量的时间[19]。Blast2Go 是一款用于大规模 GO 注释的工具,Blast2Go 是一套在植物基因组研究中对未知基因功能分析的综合软件,其主要特点是:①综合多种注释策略,输出格式多样,支持多种注释数据库,包括 GO、Enzyme Codes、InterPro

以及 KEGG;②直观的图形化界面,可输出多种结果统计图;③综合处理数据,除对序列做 GO 注释,还可以进行 KEGG Pathway 分析等,并能根据用户的设置进行分析;④可进行大规模数据的本地自动化注释,可一次性处理 20 000 条序列的分析。Blast2Go 的注释进程包括 3 个步骤:Blast、Mapping 和 Annotation。

2.1　启动 Blast2Go　进入 Blast2Go 主页(http://www.blast2go.com/),下载适合计算机内存容量的版本,下载后得到图形化界面程序 blast2go *.jnlp。运行 Blast2Go 有 3 个必要条件:①网络连接;②JAVA 运行环境(JRE);③配置本地数据库(本地数据库包含了执行 Mapping 步骤的必要信息)。若使用 Blast2Go Pro(Blast2Go 的付费版本),则可以使用 Blast2Go 提供的在线数据库,无需再配置本地数据库。在 lunix 下打开 Blast2Go 运行界面,可执行命令:Javaws -Xnosplash blast2go *.jnlp,打开 Blast2Go 运行界面后,在运行 Blast2Go 之前,需设置数据库。可供选择的数据库有 3 类:①公共数据库;②本地数据库(事先本地化的数据库);③Pro Server(Blast2Go Pro 用户可选)。

2.2　Blast 步骤　启动 Blast2Go 后,可直接输入 Blast 的结果文件(xml 格式),也可以直接输入拼接后的结果文件进行 Blast 比对。用户可选择的 Blast 方式有 3 种:①在 NCBI 运行 Blast;②使用本地 Blast(Blast+,需本地化数据库);③使用 CloudBLAS 进行 Blast。

使用 NCBI 的 Blast+进行本地 Blast 比对时,可以选择的 Blast 程序有 4 种,即 Blastx、Blastp、Blastn 和 tBlastx[13]。用户可以根据自己的需要设置 E-value 值,同时 Blast2Go 提供数目众多的数据库供用户选择,如 nr、nt、swissprot、refseq_protein、est 等。选择合适的 Blast 程序及比对数据库,设置 E-value 值和最大 hits 数后,点击"start"便开始 Blast 比对步骤:①直接输入 Blast 结果(xml 文件);②输入序列文件;③选择 Blast 运行方式;④Blast 设置,包括选择 Blast 运行程序、选择比对数据库、设置 e 值、设置 Blast hits 数以及输出文件格式;⑤查看 Blast 结果统计图。

2.3　Mapping 步骤　Blast 步骤完成后,接着可以进行 Mapping 步骤。Mapping 是一个检索与 Blast 得到的 hits 相关的 GO terms 的进程。Blast2Go 进行 3 种不同的 Mapping 方式:①Blast 结果中的基因序列号(accession number)用来检索基因名称,检索会用到 2 个由 NCBI 提供的 Mapping 文件(gene-infor、gene2accession);②Blast 结果中的 GI identifiers 用于重新检索在 UniProt ID 号,检索使用来自 PIR(the protein information resource,蛋白质信息资源数据库)[20]非冗余参考蛋白质数据库的 Mapping 文件,这个非冗余参考蛋白质数据库搜罗了来自 PSD、UniProt、Swiss-Prot、TrEMBL、RefSeq、GenPept 以及 PDB 数据库的蛋白质信息;③Blast 结果中的基因序列号(accession number)直接在 GO 数据库中的 DBXRef Table 中进行搜索。

2.4　Annotation 步骤　Mapping 步骤结束后,进入 Annotation 注释步骤。通过 Annotation 步骤,将 Mapping 步骤中获得的 GO terms 分配到各个输入序列,得到与输入序列相关的 GO 注释信息,并将 GO 注释信息分为 molecular function,cellular component 和 biological process 这 3 类及其子类。利用大量的序列数目和 GO terms 的结果数目,通过 GO slim(GO 联合会提供的简化本体论术语)将得到的 GO terms 归类到更高层次的 terms,从而可以在更高的层次上研究基因的功能。

2.5　利用 Blast2Go 在 GO 注释结果中挖掘信息　利用 Blast2Go 还可以进行 KEGG Pathway 分析。KEGG(kyoto encyclopedia of genes and genomes)是系统分析基因功能、基因组信息数据库,KEGG 可以查询整合代谢途径(pathway),这样有利于研究者将基因及表达信息作为一个整体网络进行研究。在 Blast2Go 注释的过程中,会给出相关 unigene 的 EC(enzyme code)号。在代谢通路中,EC 号是节点(酶)的识别符,即通过 EC 号,可以找到 unigene 参与的生物学通路(pathway),因此能推断出对应的 unigene 如何参与生命活动及其在生命活动中发挥的作用(图 2)。

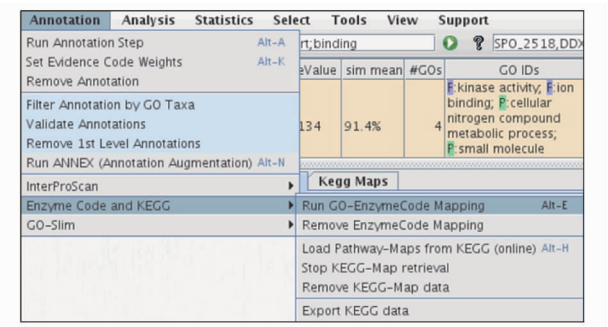

图 2　KEGG pathway 分析

Fig. 2　KEGG pathway analysis

3　讨论

目前,绝大多数已报道的转录组研究资料仅介绍了某个物种的转录组研究成果,很少有资料介绍转录组分析中使用的软件及软件的详细使用方法。该研究以 NCBI 网站 SRA 数据库下载的 Illumina 测序平台产生的数据(sra 文件)为例,使用工具包 NGS QC Toolkit 中的 IlluQC.pl 对 raw data(31 028 774 条 raw reads)进行过滤得到 clean data(24 735 426 条 clean reads)。随后使用 Velvet/Oases 进行转录组拼接,最后进行基因注释和功能分类。最终,拼接得到 42 203 条 transcripts 中,有 38 622 条(91.5%)transcripts 获得相似性搜索结果,这表明转录组测序技术是功能基因组学研究的有利手段。

该研究详细介绍了转录组测序数据(single-end)分析的流程,但研究者在具体的数据分析过程中,可能还会遇到各种各样的问题。如测序中出现的错误会影响到从头拼接的质量,所以在质量控制时,会根据数据质量情况对 reads 末端碱基进行适当的剪切(trimming)。其次,该研究使用的是 Single-end reads,所以在进行拼接时,可以直接运行 velvet。在组装 Paired-end reads 时,由于 velvet 软件只能采用两端序列混合在一起的 fasta 或 fastq 文件,因此需先使 shuffleSe-

(下转第 100 页)

5 结语

下沉式绿地能够实现绿地多功能化、就地消纳雨水径流、减少外排雨水量、雨水资源化利用、改善生态环境等多种目标[26],除了要满足科学性、适宜性、生态性三大原则进行合理选址外,还要有科学的竖向设计、合理的土壤结构、耐涝的乡土植物,尽量避免次生危害。这些选址应当因地制宜,依据不同城市的环境特点进行调整和变化,应针对不同类型城市的选址原则及要求进一步开展研究。

参考文献

[1] 俞孔坚.打造"海绵城市"别忽视民间水利工程[J].中州建设,2016(15):54,55.

[2] 章林伟.海绵城市建设概论[J].给水排水,2015,41(6):1-7.

[3] 中华人民共和国住房和城乡建设部.关于印发《海绵城市建设技术指南——低影响开发雨水系统构建(试行)》的通知:建城函〔2014〕275号[A].2014-10-22.

[4] 张铁锁,刘九川.下沉式绿地的应用浅析[C]//河南省科学技术协会.科技、工程与经济社会协调发展——河南省第四届青年学术年会论文集(下册).郑州:河南省科学技术协会,2004:3.

[5] 邱巧玲."下沉式绿地"的概念、理念与实事求是原则[J].中国园林,2014,30(6):51-54.

[6] 李俊奇,车伍,池莲,等.住区低势绿地设计的关键参数及其影响因素分析[J].给水排水,2004,30(9):41-46.

[7] 邵洪波.下沉式绿地的设计和对城市排水的影响分析[J].城市建设理论研究:电子版,2013(9):1-4.

[8] 程江,徐启新,杨凯,等.下凹式绿地雨水渗蓄效应及其影响因素[J].给水排水,2007,33(5):45-49.

[9] 叶睿超,王秀英.下沉式绿地在自贡市推广的价值初探[J].安徽农学通报,2014,20(19):83-84.

[10] 游漪凡,丁若莹,万明磊,等.城市下沉式绿地雨水调蓄技术探讨及优化[J].市政技术,2017,35(2):110-112,116.

[11] 龙春英,葛嘉浩.基于海绵城市对下沉式绿地雨水景观的探讨[J].安徽建筑,2016,23(3):13-14,18.

[12] 车伍,赵杨,李俊奇.海绵城市建设热潮下的冷思考[J].南方建筑,2015(4):104-107.

[13] 邹宇,许乙青,邱灿红.南方多雨地区海绵城市建设研究:以湖南省宁乡县为例[J].经济地理,2015,35(9):65-71,78.

[14] 苏义敬,王思思,车伍,等.基于"海绵城市"理念的下沉式绿地优化设计[J].南方建筑,2014(3):39-43.

[15] 黎意如.浅谈海绵城市理论在下沉式绿地中的融合运用[J].绿色环保建材,2017(1):209.

[16] 牛志强,崔飞飞.基于"海绵城市"概念的郑州市下沉式绿地优化设计[J].科技展望,2017,27(28):26-27.

[17] 王艳丽,王园园.海绵城市理念下的下沉式绿地优化设计分析[J].住宅与房地产,2016(30):84.

[18] 危薇.基于"海绵城市"理念的下沉式绿地优化设计探讨[J].城市建筑,2016(29):48.

[19] 林苡.海绵城市理念下的下沉式绿地研究与优化:以西咸新区沣西新城为例[D].西安:长安大学,2017.

[20] 沈杨霞,张建林.海绵城市中植物景观的品种选择[J].现代园艺,2016(21):90-91.

[21] 许铭宇,卢艺菲.基于海绵城市视角的下沉式绿地应用[J].浙江农业科学,2018,59(8):1394-1395,1398.

[22] 孟颖斌,李忠民.青岛在海绵城市建设中的植物选择与配置[J].广东园林,2018,40(1):65-68.

[23] Ministerium fuer Umwelt und Naturschutz, Land wirtschaft und Verbraucherschutz des Landes Nordrhei-Westfalen. Naturnahe Regenwasserbewirtschaftung[M]. Duisburg:WAZ-DRUCK,2001.

[24] GEIGER W,DREISEITL H. Neue Wege fuer das Regenwasser:Handbuch zum Rueckhalt und zur Versickerung von Regenwasser in Baugebieten [M]. Muenchen:Muenchen Oldenbourg Industrieverlag GmbH,2001.

[25] 暴雨检视迁安海绵建设效果雨水随下随渗 试点区域积水问题基本消除[EB/OL].[2018-07-25].http://ts.hebnews.cn/2018-07/25/content_6966588.htm.

[26] 王思思,苏义敬,车伍,等.景观雨水系统修复城市水文循环的技术与案例[J].中国园林,2014,30(1):18-22.

(上接第 91 页)

quences_fastq.pl 或 shuffleSequen ces_fasta.pl 将 paired-end 数据结合在一起。大多数拼接软件使用的算法最初都是为基因组测序设计的,但由于可变剪切的存在,一个基因通常都会编码多个转录本,这给真核生物转录组拼接带来巨大的挑战[16]。

另外,由于一般实验室计算机内存限制无法一次性完成所有数据的 GO 注释,可以将拼接后得到的转录本大文件(transcript.fa)分成几个大小合适的 fasta 文件进行基因注释及 GO 分类,在查看 annotation 结果图(Statistics -> Annotation Statistics)时可分别将注释结果以 txt 格式输出(save-> export as text),最终将结果汇总即可。

参考文献

[1] COSTA V,ANGELINI C,DE FIES I,et al. Uncovering the complexity of transcriptomes with RNA-Seq[J]. Journal of biomedicine and biotechnolog,2010,2010:1-19.

[2] 刘红亮,郑丽明,刘青青,等.非模式生物转录组研究[J].遗传,2013,35(8):955-970.

[3] 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.

[4] ZHANG X M,ZHAO L,LARSON-RABIN Z,et al. *De novo* sequencing and characterization of the floral transcriptome of *Dendrocalamus latiflorus* (Poaceae:Bambusoideae)[J]. PLoS One,2012,7(8):1-15.

[5] MUDALKAR S,GOLLA R,GHATTY S,et al. *De novo* Transcriptome analysis of an imminent biofuel crop,*Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers[J]. Plant Mol Biol,2014,84(1/2):159-171.

[6] UPADHYAY S,PHUKAN U J,MISHRA S,et al. *De novo* leaf and root transcriptome analysis identified novel genes involved in Steroidal sapogenin biosynthesis in *Asparagus racemosus* [J]. BMC Genomics,2014,15:1-13.

[7] LOGACHEVA M D,KASIANOV A S,VINOGRADOV D V,et al. *De novo* sequencing and characterization of floral transcriptome in two species of buckwheat(*Fagopyrum*)[J]. BMC Genomics,2011,12:1-17.

[8] 井赵斌,魏琳,俞靓,等.转录组测序及其在牧草基因资源发掘中的应用前景[J].草业科学,2011,28(7):1364-1369.

[9] 周华,张新,刘腾云,等.高通量转录组测序的数据分析与基因发掘[J].江西科学,2012,30(5):607-611.

[10] 黄子夏,柯才焕,陈军.大规模 GO 注释的生物信息学流程[J].厦门大学学报(自然科学版),2012,51(1):139-143.

[11] WANG Z Y,FANG B P,CHEN J Y,et al. *De novo* assembly and characterization of root trascriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato(*Ipomoea batatas*)[J]. BMC Genomics,2010,11(1):726-739.

[12] 郝大程,马培,穆军,等.中药植物虎杖根的高通量转录组测序及转录组特性分析[J].中国科学,2012,42(5):398-412.

[13] HARRIS M A,CLARK J,IRELAND A,et al. The Gene Ontology(GO)database and informatics resource[J]. Nucleic acids research,2004,32:258-261.

[14] GARG R,PATEL R K,TYAGI A K,et al. *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification[J]. DNA Research,2011,18(1):53-63.

[15] COCK P J A,FEILDS C J,GOTO N,et al. The Sanger FASTQ file format for sequences with quality scores,and the Solexa/Illumina FASTQ variants[J]. Nucleic acids research,2010,38(6):1767-1771.

[16] CLARKE K,YANG Y,MARSH R,et al. Comparative analysis of *de novo* transcriptome assembly[J]. Science China life science,2013,56(2):156-162.

[17] CONESA A,GÖTZ S. Blast2Go:A comprehensive suite for functional analysis in plant genomics[J]. International journal of plant genomics,2008,2008:1-12.

[18] KUMAR S,DUDLEY J. Bioinformatics software for biologist in the genomics era[J]. Bioinformatics,2007,23(14):1713-1717.

[19] 王成刚,莫志宏.整合 BLAST 搜索与 GO 注释的软件 GoBlast[J].中国生物化学与分子生物学报,2006,22(12):1003-1006.

[20] 胡绍军.蛋白质组学数据库信息资源的开发与利用[J].图书馆学研究,2006(7):77-82.