计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):222−228 [doi: 10.15888/j.cnki.csa.007024]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

六倍体小麦基因组注释流程构建与优化①

祝海栋, 李瑞琳, 何小雨, 赵　丹, 韩鑫胤, 牛北方

(中国科学院 计算机网络信息中心, 北京 100190)
(中国科学院大学 计算机与控制学院, 北京 100190)
通讯作者: 牛北方, E-mail: bniu@sccas.cn

摘　要: 野生小麦是异源六倍体, 基因组规模较大 (约 14 GB), 且包含大量重复序列. 为了培育具有优良性状的新品种, 首先要定位控制目标性状的基因, 因此建立一个完整准确的基因组注释软件流程至关重要. 传统的基因组注释方法基于数据库比对, 具有三个明显的缺点: 一是比对速度慢; 二是难以发现新基因; 三是软件选择没有统一标准. 本文提出了一种新的生物信息学注释流程, 结合了基因数据库比对、转录组高通量测序数据分析、全长转录组单分子测序数据分析等多种技术手段, 实现了六倍体小麦科农 9204 基因组完整准确的注释, 为揭示小麦生长发育规律和培育新品种提供了重要参考和软件技术支撑.

关键词: 基因组; 基因注释; 高通量测序; 生物信息学; 全长转录组

引用格式: 祝海栋,李瑞琳,何小雨,赵丹,韩鑫胤,牛北方.六倍体小麦基因组注释流程构建与优化.计算机系统应用,2019,28(8):222–228. http://www.c-s-a.org.cn/1003-3254/7024.html

Construction and Optimization of Hexaploid Wheat Genome Annotation Process

ZHU Hai-Dong, LI Rui-Lin, HE Xiao-Yu, ZHAO Dan, HAN Xin-Yin, NIU Bei-Fang

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
(School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Wild wheat is a heterologous hexaploid with a large genome size (about 14 GB) and a lot of repetitive sequences. In order to breed new varieties with good traits, we must first locate the genes that control the target traits. Therefore, it is important to establish a complete and accurate genome annotation process. Traditional genomic annotation method based on database alignment has three obvious disadvantages: first, the alignment runs slowly; second, it is difficult to discover new genes; third, there is no uniform standard for software selection. We propose a new analysis process that combines genetic database alignment, transcriptome high-throughput sequencing, and full-length transcriptome single-molecule sequencing data analysis to annotate hexaploid wheat KN9204 completely and accurately. The annotation of the genome provides an important reference and technical support for revealing the growth of wheat and cultivating new varieties.

Key words: genome; genome annotation; high-throughput sequencing; bioinformatics; full length transcriptome

① 基金项目: 国家重点研发计划 (2018YFB0203903, 2016YFC0503607); 中国科学院信息化专项 (XXH13504-08); 国家自然科学基金 (31771466); 青海省科技成果转化专项 (2016-SF-127); 中国科学院“百人计划”海外引进杰出人才择优支持 (牛北方)
Foundation item: National Key Research and Development Program of China (2018YFB0203903, 2016YFC0503607); CAS Special Fund for Informatization Construction (XXH13504-08); National Natural Science Foundation of China (31771466); Special Fund for Transformation of Scientific And Technological Achievements of Qinghai Province (2016-SF-127); Special Support for Introducing Overseas Outstanding Talents of CAS “Hundred Talents Program”(NIU Bei-Fang)
收稿时间: 2019-02-22; 修改时间: 2019-03-08; 采用时间: 2019-03-11; csa 在线出版时间: 2019-08-08

小麦生产对保证粮食安全和农业可持续发展具有重要的现实意义, 促进小麦的增产和品质改良成为当前小麦育种研究的前沿热点. 为了培育具有优良性状的新品种, 首先要定位控制目标性状的基因, 因此建立一套完整准确的大尺度基因组注释流程成为培育新品种过程中的难点之一. 基因组注释主要包括基因识别和基因功能标注两个方面[1], 本文的主要研究方向是基因识别, 主要目标是准确定位基因位置及发现物种特异性基因.

近些年, 基因组测序技术突飞猛进, 其发展过程包含三个阶段: 1975 年由桑格和考尔森开创的链终止法标志着第一代 DNA 测序技术的诞生, 但测序成本高、通量低等缺点严重影响了其大规模的应用; 第二代测序建立在聚合酶链式反应扩增的基础上, 主要特点是为边合成边测序, 测序结果读长短、测序速度快、吞吐量大[2]; 第三代测序技术的核心是以单分子为目标, 旨在解决第二代测序在准确性和组装困难方面的问题. 测序技术的高速发展, 大大满足了测序深度、重测序等大规模基因组的研究需求, 改变了生命科学诸多领域的研究面貌, 也给小麦等大尺度基因组的注释研究奠定了重要基础.

1 小麦基因组注释流程研究现状

传统的基因注释方法主要为数据库比对, 通过把基因组片段与已有的亲缘物种基因数据库比对, 得到目标基因. 这种方法较为简便, 但具有三个明显的缺点: 一是对比速度慢, 原因是该方法中需要与较多的数据库进行比对分析, 因此耗时长, 尤其是用于小麦等较大基因组时该缺点更为明显; 二是难以发现新的基因, 由于依赖数据库比对得到的基因都是目前相近物种中广泛存在的基因, 物种特有的基因不会被识别, 造成注释的不完整. 转录组测序可以全面快速的获取物种在某一时期和特定组织中所有表达的基因序列, 常被用于研究物种基因结构和基因功能[3]. 但是转录组分析软件繁多, 缺乏统一的选择标准, 且分析过程中涉及多个软件配合完成, 分析流程中不可避免地会存在软件间衔接困难、格式转换和大量数据重复读写等问题. 另外, 由于各种软件在内存、CPU 等资源利用方面存在较大差异, 且多数情况下生物信息学中的分析过程依赖于脚本生成的流程, 没有并行优化, 因此资源利用率和分析效率较低. 针对上述问题, 本文提出了整合基因组和转录组数据进行基因注释的分析流程, 以提高注释的完整性和准确性.

2 实验数据与测试环境

本次研究中使用的测试数据包括: 科农 9204 小麦基因组组装数据, 数据大小 14.24 GB; 二代转录组测序样本 77 个, 单样本大小约为 17 GB; 三代全长转录组测序样本 2 个, 单样本大小约为 40 GB. 测试环境为超级计算系统“元”. 其包含 270 台计算节点, 每节点采用 2 个 Intel Xeon E5-2680V3 处理器 (2.5 GHz、12 核), 单节点 CPU 计算能力 0.96Tflops, 配备 256 GB 内存. 操作系统为 Linux version 2.6.32-358.el6.x86_64, CentOS release 6.4 (Final). 系统中配置 Python、Perl、C++等基本编译和运行环境.

3 基因组注释分析软件流程

本节分为 3 个部分, 建立小麦基因组注释分析流程, 并对部分环节实现优化.

3.1 数据库比对注释

数据库比对注释是最传统和最常用的注释方式. 其主要方法是把待注释的基因组逐一与各个近亲物种已有基因比对, 获取注释结果. TriAnnot[4]是为解读小麦基因组而开发的一个流程, 集合了 Blast、Repeat Masker 等开源软件, 比对了 NCBI[5]、TAIR10[6]等开放数据库, 对转座子、编码基因、非编码序列、分子标记进行了多步的处理分析, 可以得到比较完整的注释结果. 因此, 本文对 TriAnnot 注释软件进行优化并对科农 9204 小麦基因组进行初步注释.

3.1.1 优化方法

为了提高注释效率, 本研究的主要贡献是实现 TriAnnot 注释软件的优化, 重点分为 3 个方面: 单任务多实例并行优化、多核计算并行优化, 多数据库查找并行优化, 下面给出具体的方法与实现.

首先, 对 TriAnnot 注释软件的单任务多实例并行优化. 六倍体小麦基因组较大, 每条染色体的平均长度接近 700 MB, 给序列比对带来许多困难. 为了便于比对分析, 在注释过程中, 必须把染色体切分成小的片段, 本研究中选择的切分大小为 1 MB, 切分时的保留的重复长度为 50 KB. 切分后的每个片段即为每个实例, 实例之间相对独立. 为了提高注释速度, 本研究采用了多实例并行, 即在每个时刻都有多个实例同时执行. 因为

每个步骤的 CPU 和内存使用率各不相同, 该优化策略可以实现资源的充分利用.

其次, 实现 TriAnnot 注释软件的单任务多实例并行优化. 在实验中, 针对每个软件的特点, 本研究采用相应的调度方式优化. RepeatMasker 通过相似性比对来识别重复序列, 可以屏蔽序列中转座子重复序列和低复杂度序列[7]. 本研究在流程中加入了 RepeatMasker 的多核心并行, 可以根据机器硬件情况指定 4 至 24 核心实现并行运行, 并且可以通过使用-qq 指令加快比对效率.

最后, 实现对 TriAnnot 注释软件的多数据库查找并行优化. SIMSearch 软件通过使用多个同源数据库进行序列比对找到亲缘关系较近的基因序列. 为了加快同源基因的查找速度, 研究采用了多数据库并行的方案, 把多个同源数据库同时读取到内存中, 将每个基因片段在多个核心上与不同的数据库进行对比.

3.1.2 软件与数据库

TriAnnot 依赖的软件及其下载地址如表 1 所示, 数据库及其下载地址如表 2 所示.

表 1 TriAnnot 主要依赖软件

软件	版本	地址
BLAST	2.2.21	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST
cross-match	0.990329	http://www.incogen.com/public_documents/vibe/details/crossmatch.html
augustus	2.4	http://augustus.gobics.de/
EuGene	4.0	http://eugene.toulouse.inra.fr/
HMMER	3.0	http://eugene.toulouse.inra.fr/
InterProScan	4.6	http://www.ebi.ac.uk/interpro/
TEannot	1.4	http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET
RepeatMasker	3.2.6	http://www.repeatmasker.org/
TRF	4	http://tandem.bu.edu/
tRNAscan-SE	1.23	http://lowelab.ucsc.edu/tRNAscan-SE/
exonerate	2.2.0	http://www.ebi.ac.uk/~guy/exonerate/

表 2 TriAnnot 主要数据库

数据库	版本	地址
ALL_Repbase	16.03	http://www.girinst.org/server/RepBase/index.php
AT_unigene	70	ftp://ftp.ncbi.nih.gov/repository/UniGene/Arabidopsis_thaliana/
HV_unigene	56	ftp://ftp.ncbi.nih.gov/repository/UniGene/Hordeum_vulgare/
OS_unigene	80	ftp://ftp.ncbi.nih.gov/repository/UniGene/Oryza_sativa/
TA_unigene	56	ftp://ftp.ncbi.nih.gov/repository/UniGene/Triticum_aestivum/
ZM_unigene	77	ftp://ftp.ncbi.nih.gov/repository/UniGene/Zea_mays/
SB_unigene	29	ftp://ftp.ncbi.nih.gov/repository/UniGene/Sorghum_bicolor/
SO_unigene	14	ftp://ftp.ncbi.nih.gov/repository/UniGene/Saccharum_officinarum/
HS_unigene	224	ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/
WH_embl_mito	AP008982	http://www.ncbi.nlm.nih.gov/nuccore/78675232

3.1.3 分析流程

该步骤的输入为基因组装得到的 KN9204 小麦基因序列文件. 六倍体小麦有 21 条染色体, 此外还有少量未有效定位到染色体上的基因片段, 在其中加入 100 个未知碱基标识“N”, 构成未分组染色体, 共 22 条 fasta 序列, 每条序列单独输入.

TriAnnot 软件运行前需要下载完整的基因数据库. 主要参数包括: -W 指定工作目录, -s 指定输入的 fasta 文件, -t 指定注释流程 xml 文件, --type 设置输入为核酸, --maxlength 设置最大序列长度, --splitseq 设置超过最大长度的序列自动切分, --overlap 设置切分时冗余长度.

软件的输出为 gff 文件, 包含了详细的内含子、外显子、编码区、转座子等注释.

3.2 转录组高通量测序

为了准确注释物种特异性基因, 本研究结合了转录组高通量测序数据, 选取了苗期、孕穗期、7 天、14 天等不同时期的根、叶、穗等不同组织的样本, 测序深度约为 30 X. 常用的转录组分析工具有 HISAT、SATR、StringTie、Cufflinks 等. 使用不同的分析工具和方法对分析结果的准确度和耗时影响较大, 需要根据特定的数据集及特定的研究目标选择合适的分析工

具和方法. HISAT 解决了转录组中仅有不连续的外显子难以比对的问题, 对比上代主流转录组比对工具 Tophat 效率高 50 倍, 且内存需求更少[8]. StringTie 继承于 Cufflinks, 在准确性方面有了较大提升, 且可以通过输入数据库比对注释结果提高在已知基因区域的准确性, 在组装的过程中会计算每个基因及可变剪切的表达水平. 综合以上优点, 对于复杂的小麦基因组, 本文使用 HISAT[9]和 StringTie[10]工具进行转录组组装. 主要分为以下四个步骤

(1) 建立 HISAT2 基因组索引. 转录组数据分析过程遇到的第一个问题就是, 小麦上亿条 reads 如何在保证错误率在可接受的范围内, 高效率地比对到基因组上. 针对上述问题, 需要根据基因组序列使用 hisat2-build 命令建立索引.

(2) 将所有二代测序 reads 比对到基因组. 使用 HISAT2 利用基因组索引将高通量测序 reads 比对到基因组上. 参数-p 指定并行核心数, -x 指定索引位置, --dta 为组装提供锚点. 使用 samtools 将比对结果按染色体和起始位点排序.

(3) 使用 StringTie 对排序完成的 reads 进行组装. 不同组织中表达数据差异相对较大, 比对到基因组的 reads 也各有不同, 这些因素都会影响组装的效率.

(4) 将所有转录本的组装结果使用 StringTie 的 merge 模块合并. 由于不同组织和不同时期表达的基因各不相同, 为了获取更加完整的注释, 需要对多个测序样本合并. merge 步骤可以跨多个测序样本生成统一的转录本. 首先要创建一个文本文件, 该文件包含所有转录本组装结果路径, 文本的每行是单个样本组装结果文件路径. 参数设置为: --merge 指定使用合并模块, -p 指定并行核心数, 输入上述文本文件, 即可得到最终的二代转录组组装结果.

3.3 全长转录组单分子测序数据处理

二代测序可以准确地进行基因定量分析研究, 但是受读长限制, 不能得到全转录本的信息. 全长转录组采用单分子实时测序技术, 通过构建哑铃型文库, 以环形方式循环测序[11]. 因此, 通过全长转录组单分子测序可以不经过组装, 准确、直接地获取整个转录本. 三代测序存在单碱基错误率较高的问题[12], 本研究使用 PacBio 公司发布的 SMRTLINK Pipeline[13], 对三代测序得到的数据进行过滤与质量控制. 由于全长转录组测序成本相对较高, 本次研究采取了常用的组织混合测序方式. 选取了叶、穗、幼叶、幼根四种组织混合, 设置两个生物学重复, 共得到两组测序数据. 数据处理过程主要分为以下 3 个步骤:

(1) 使用 SMRTLINK 进行三代测序数据的清洗. 主要分为三个步骤, 首先召回环形一致性序列, 包括单碱基纠错和序列过滤; 然后对序列分类, 包括去除接头、polyA 尾部和串联子; 最后进行迭代的聚类纠错, 主要是合并相似的序列, 形成全长转录本. 该软件提供了用户可视化接口, 安装后使用浏览器访问服务器地址的对应端口即可进入管理界面. 在管理界面中, 使用“数据管理”选项导入原始测序结果文件, 然后使用“SMRT 分析”选项, 选择分析流程为“Iso-Seq”, 设置相关参数, 选取对应的样本即可开始全长转录组的纠错. 在本次研究中, 我们设置的参数主要有以下几个: By Strand CCS: OFF; Maximum Dropped Fraction: 0.8; Maximum Subread Length: 15000; Minimum Predicted Accuracy: 0.75; Minimum SNR: 3.75; Polish CCS: ON; 其余参数均为默认值.

(2) 使用 GMAP[14]比对全长转录本到基因组. GMAP 具有一次对多条 reads 同时进行比对的优点, 比对结果较为可靠, 因此, 本文采用 GMAP 将全长转录本比对到基因组上. 为了提高运算速度, GMAP 比对阶段对全长转录本序列进行数据分割, 将分割后的多个数据进行并行处理. 首先使用 gmap-build 建立索引, 由于小麦基因组较大, 会自动使用长索引. 使用-D 参数指定索引存储位置, -d 参数指定索引前缀, 输入基因组 fasta 文件即可开始建立索引, 然后使用 gmapl 命令开始比对. 指定的索引存储位置和前缀需与上述过程中对应参数相同, -B 指定批处理个数, -t 指定并行核心数, -f 指定输出格式, -O 指定顺序输出. 使用 samtools 将 bam 文件按染色体和起始位点排序.

(3) 合并多个样本的全长转录组结果. 合并时使用 TAMA 软件, 共分为两个步骤. 首先根据比对到基因组上的位置情况合并可变剪切, 然后合并多个测序样本的转录本.

3.4 合并注释结果

为了得到高质量的注释结果, 需对上述结果进行合并和过滤, 在本次研究中我们开发了一个自动化合并注释的软件 Annotator, 该软件包含的功能模块有格式转换, 结果合并, 去除重复序列, 过滤可变剪切, 根据证据支持评价可信度, 编码区预测, 蛋白翻译等多个步

骤, 最终生成 gff 注释文件和转录本序列、编码区序列、编码蛋白序列. Annotator 详细流程如图 1 所示. 合并过程分为以下 5 个步骤:

(1) 转换结果文件为 bed12 格式. 本步骤调用了 cufflinks[15]软件的 gffread 模块和 bedops[16]软件, 将数据库比对注释得到 gff 文件和二代转录组组装得到的 gtf 文件转换为 bed 文件. bed 格式使用单行定义单个基因, 具有简单易读的特点.

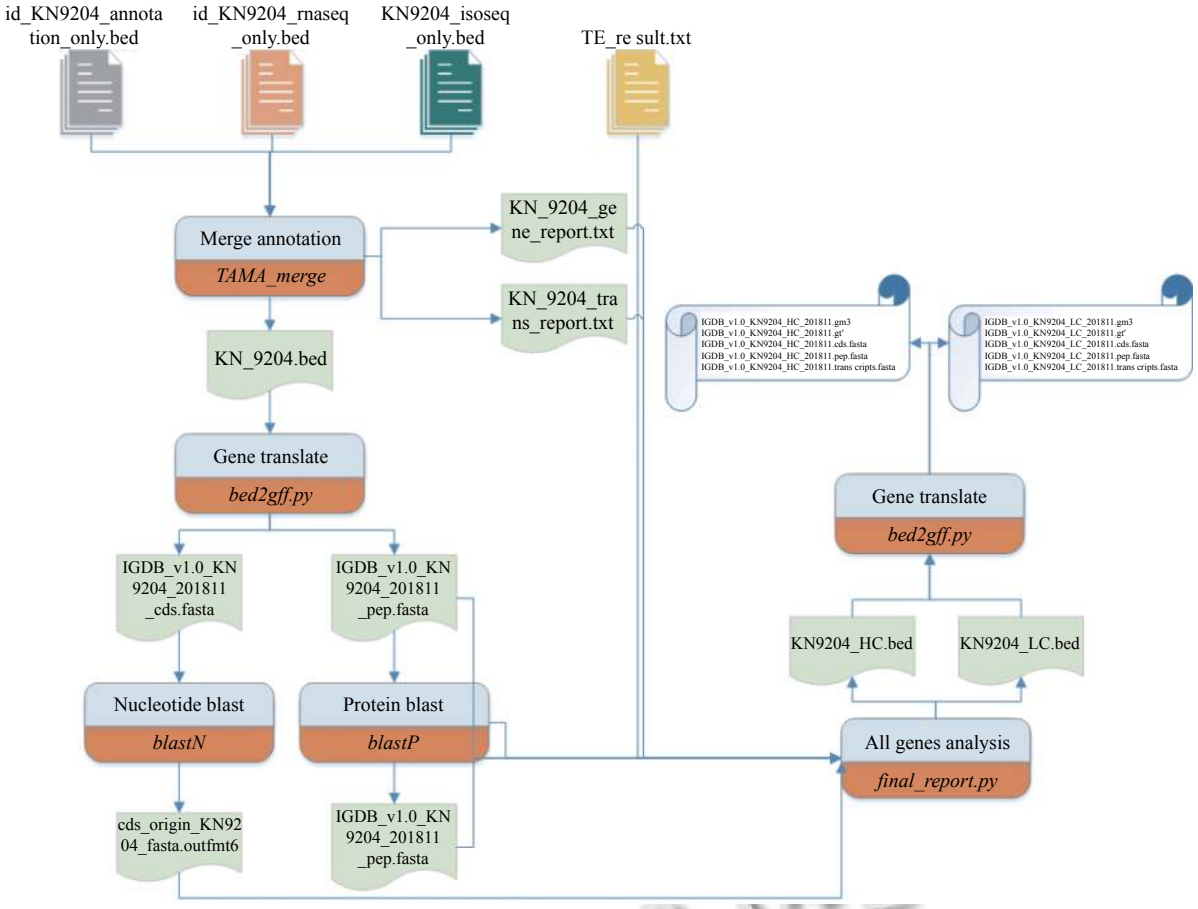

图 1 注释合并流程

(2) 合并数据库比对注释、二代转录组组装、三代全长转录组结果. 本步骤使用了 TAMA 的 merge 模块, 生成含有全部基因的 bed 文件. 根据每个基因的支持证据的不同, 分为高可信度基因和低可信度基因.

(3) 过滤重复的可变剪切. 由于测序误差或 reads 组装错误的不可避免, 测序结果中可变剪切会出现许多冗余, 因此需要对重复的可变剪切进行过滤. 过滤过程中, 保留的优先级依次为全长转录组得到的可变剪切结果、数据库比对注释结果中的可变剪切, 由于二代转录组组装有更多的错误可能, 其优先级最低.

(4) 预测所有基因的编码区. 该步骤使用三种可能的翻译方式分别将基因翻译为氨基酸序列, 取最长的序列, 得到基因的编码区.

(5) 翻译编码区序列. 根据注释结果中的编码区位置, 将核酸序列翻译为氨基酸序列, 生成序列文件.

4 实验结果与分析

经过优化, 使用 TriAnnot 注释科农 9204 基因组的重复序列时, 速度提升达到 60%, 在 1 号染色体上的测试结果如图 2 所示.

在转录组高通量测序过程中, 建立索引过程输入全基因组大小约为 14 GB, 耗时为 8122 秒, 最大内存使用约为 144 GB. 序列比对时, 输入共 77 个样本, 双端测序的单个 fastaq 文件大小约 17 GB, 比对耗时约 640 秒. StringTie 组装输入 bam 文件大小约为 7 GB, 耗时在 10 小时至 24 小时不等.

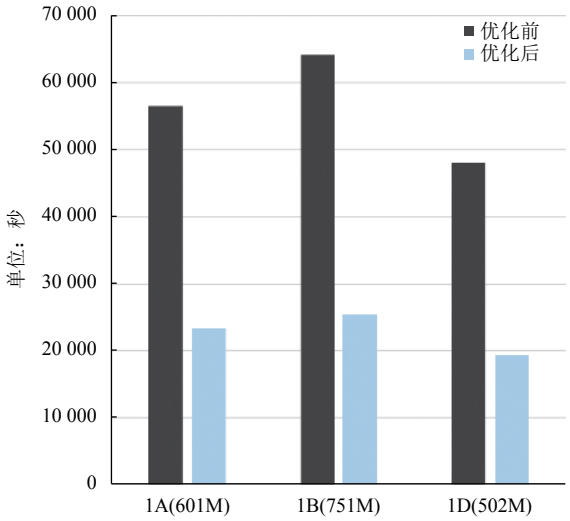

图 2 注释耗时变化

三代全长转录组测序中单样本 bam 文件为 38 GB, 运行时间约为 149 小时. 最终分别输出高质量和低质量的全长转录组 fasta 序列. 最终得到的环形一致性序列质量分布如图 3 所示, 超过 50% 的序列质量均在 0.99 以上, 可信度较高.

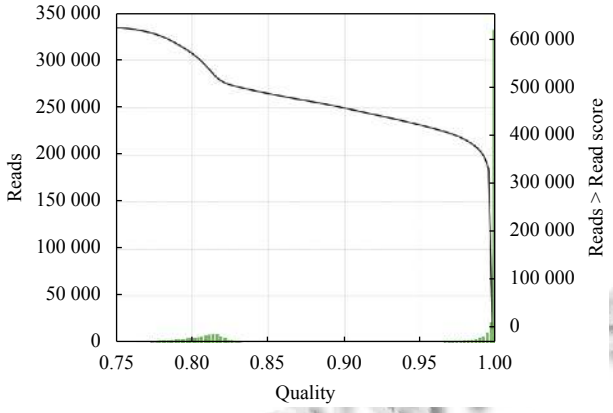

图 3 全长转录组质量分布

本流程在六倍体小麦科农 9204 基因组上完成测试, 共注释出 110 326 个高可信度基因. 对比同源的中国春小麦基因组, 其注释包含 107 891 个高可信度基因[17], 其中有 102 413 个基因匹配, 占中国春基因总数的 94.9%, 占科农 9204 基因总数的 92.8%, 具有高度一致性, 这说明了本流程注释结果具有较高的准确性.

5 结论与展望

本文提出了一种综合运用数据库比对、二代转录组高通量测序、三代全长转录组测序技术获得准确注释的分析流程, 并独立研发了注释软件 Annotator. 随后对流程中用到的部分软件进行了优化, 大大提高了注释效率, 为大尺度多倍体基因组提供了一个较为成熟的注释软件流程.

当前流程也存在一些问题: (1) 注释速度仍然较慢. 数据库比对注释过程是性能提升的主要瓶颈, 仍需优化. (2) 成本较高. 注释的准确性依赖于较高的测序深度, 这会带来成本的大幅提高, 尤其是三代测序更为如此, 这大大限制了该流程的广泛应用.

因此, 在未来的工作中将尝试解决上述问题, 以进一步优化整个流程. 针对注释速度问题, 可以对整个基因组进行更细粒度的并行处理, 提升比对过程中的并行效率; 此外可以使整个比对过程均在内存中进行, 避免中间结果写入硬盘, 减少不必要的时间开销. 针对注释中测序成本问题, 可以在成本较高的三代全长转录组测序中采取多组织混样测序的方案, 选取最关注的组织和时期的样本混合, 通过单次测序降低成本.

参考文献

1 Moriya Y, Itoh M, Okuda S, *et al*. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Research, 2007, 35(Web Server issue): 182–185.

2 Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology, 2008, 26(10): 1135–1145. [doi: 10.1038/nbt1486]

3 Pertea M, Kim D, Pertea GM, *et al*. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. Nature Protocols, 2016, 11(9): 1650–1667. [doi: 10.1038/nprot.2016.095]

4 Leroy P, Guilhot N, Sakai H, *et al*. TriAnnot: A versatile and high performance pipeline for the automated annotation of plant genomes. Frontiers in Plant Science, 2012, 3: 5.

5 Geer LY, Marchler-Bauer A, Geer RC, *et al*. The NCBI biosystems database. Nucleic Acids Research, 2010, 38(suppl_1): D492–D496. [doi: 10.1093/nar/gkp858]

6 Berardini TZ, Reiser L, Li DH, *et al*. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis, 2015, 53(8): 474–485. [doi: 10.1002/dvg.22877]

7 Smit AFA. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current Opinion in Genetics & Development, 1999, 9(6): 657–663.

8 Wen GZ. A simple process of RNA-sequence analyses by

Hisat2, Htseq and DESeq2. Proceedings of the 2017 International Conference on Biomedical Engineering and Bioinformatics. Bangkok, Thailand. 2017. 11–15.

9 Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nature Methods, 2015, 12(4): 357–360. [doi: 10.1038/nmeth.3317]

10 Pertea M, Pertea GM, Antonescu CM, *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology, 2015, 33(3): 290–295. [doi: 10.1038/nbt.3122]

11 Levene MJ, Korlach J, Turner SW, *et al*. Zero-mode waveguides for single-molecule analysis at high concentrations. Science, 2003, 299(5607): 682–686. [doi: 10.1126/science.1079700]

12 Au KF, Underwood JG, Lee L, *et al*. Improving PacBio long read accuracy by short read alignment. PLoS One, 2012, 7(10): e46679. [doi: 10.1371/journal.pone.0046679]

13 Gordon SP, Tseng E, Salamov A, *et al*. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS One, 2015, 10(7): e0132628. [doi: 10.1371/journal.pone.0132628]

14 Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics, 2010, 26(7): 873–881. [doi: 10.1093/bioinformatics/btq057]

15 Roberts A, Trapnell C, Donaghey J, *et al*. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology, 2011, 12(3): R22. [doi: 10.1186/gb-2011-12-3-r22]

16 Neph S, Kuehn MS, Reynolds AP, *et al*. BEDOPS: High-performance genomic feature operations. Bioinformatics, 2012, 28(14): 1919–1920. [doi: 10.1093/bioinformatics/bts277]

17 Appels R, Eversole K, Feuillet C, *et al*. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 2018, 361(6403): eaar7191. [doi: 10.1126/science.aar7191]