DATA NOTE

Improved de novo genome assembly and analysis of the Chinese cucurbit Siraitia grosvenorii, also known as monk fruit or luo-han-guo

Mian Xia1,†, Xue Han2,†, Hang He2,†, Renbo Yu2, Gang Zhen2, Xiping Jia3, Beijiu Cheng1,∗ and Xing Wang Deng2,∗

1Key Laboratory of Crop biology of Anhui Province, Anhui Agricultural University, Hefei, China; 2School of Advanced Agriculture Sciences and School of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking University, Beijing 100871, China; and 3National Demonstration Area of Modern Agriculture in Cangxi, Sichuan Province, China

∗Correspondence address. Xing Wang Deng, E-mail: deng@pku.edu.cn; Beijiu Cheng, E-mail: cbj@ahau.edu.cn http://orcid.org/0000-0003-4518-2381
†These authors contributed equally to this article.

Abstract

Background: Luo-han-guo (Siraitia grosvenorii), also called monk fruit, is a member of the Cucurbitaceae family. Monk fruit has become an important area for research because of the pharmacological and economic potential of its noncaloric, extremely sweet components (mogrosides). It is also commonly used in traditional Chinese medicine for the treatment of lung congestion, sore throat, and constipation. Recently, a single reference genome became available for monk fruit, assembled from 36.9x genome coverage reads via Illumina sequencing platforms. This genome assembly has a relatively short (34.2 kb) contig N50 length and lacks integrated annotations. These drawbacks make it difficult to use as a reference in assembling transcriptomes and discovering novel functional genes. Findings: Here, we offer a new high-quality draft of the S. grosvenorii genome assembled using 31 Gb (∼73.8x) long single molecule real time sequencing reads and polished with ∼50 Gb Illumina paired-end reads. The final genome assembly is approximately 469.5 Mb, with a contig N50 length of 432,384 bp, representing a 12.6-fold improvement. We further annotated 237.3 Mb of repetitive sequence and 30,565 consensus protein coding genes with combined evidence. Phylogenetic analysis showed that S. grosvenorii diverged from members of the Cucurbitaceae family approximately 40.9 million years ago. With comprehensive transcriptomic analysis and differential expression testing, we identified 4,606 up-regulated genes in the early fruit compared to the leaf, a number of which were linked to metabolic pathways regulating fruit development and ripening. Conclusions: The availability of this new monk fruit genome assembly, as well as the annotations, will facilitate the discovery of new functional genes and the genetic improvement of monk fruit.

Keywords: Siraitia grosvenorii; monk fruit; PacBio sequencing; ortholog analysis; RNA-Seq; mogrosides biosynthesis

Data Description

Introduction

Siraitia grosvenorii (luo-han-guo or monk fruit, NCBI Taxonomy ID: 190515) is an herbaceous perennial native to southern China and is a famous specialty in Guilin city, Guangxi Province of
Genome assembly and analysis of *Siraitia grosvenorii*

China (Fig. 1) [1]. In addition to being used as a natural sweetener, *S. grosvenorii* has been used in China as a folk remedy for the treatment of lung congestion, sore throat and constipation for hundreds of years [2]. The ripe fruit of *S. grosvenorii* contains mogrosides, which have become a popular research topic due to their pharmacological characteristics, including putative anti-cancer properties [3]. Additionally, mogrosides are purified and used as a non-caloric, non-sugar sweetener in the United States and Japan, as they are estimated to be approximately 300 times as sweet as sucrose [1, 4]. To date, *S. grosvenorii* fruit was shown to have additional pharmacological effects and contain different types of secondary metabolites [5, 6]. Monk fruit products have been approved as dietary supplements in Japan, the US, New Zealand and Australia [2, 7].

The biosynthesis pathway of mogrosides has been extensively studied, and several genes have been identified [8-11]. Squalene is thought to be the initial substrate and precursor for triterpenoid and sterol biosynthesis. Squalene epoxidases (SQE) perform epoxidation, which creates squalene or oxidosqualene, and cucurbitadienol synthase (CDS) cyclizes oxidosqualene to form the cucurbitadienol triterpenoid skeleton, which is a distinct step in phytosterol biosynthesis [12]. Epoxide hydrolases (EPH) and cytochrome P450s (CYP450) further oxidize cucurbitadienols to produce mogrol, which is glycosylated by UDP-glycosyl-transferases (UGT) to form mogroside V (Fig. 2).

The genome of *S. grosvenorii* was first published in 2016 and served the purpose of identifying the genomic organization of the gene families of interest but did not act as the reference in the transcriptome assembly and gene families identification [8]. Although the first draft genome assembly was a useful resource, some improvements remain necessary, including improving the continuity and completeness, genome assembly assessment, annotation of genes and repetitive regions, and analysis of other genomic features. With an average read length now exceeding 10 kb, SMRT sequencing technology from Pacific Biosciences (PacBio) has the potential to significantly improve genome assembly quality [13]. Therefore, we de novo assembled a high-quality genome draft of *S. grosvenorii* using high-coverage PacBio long reads and applied extensive genomic and transcriptomic analyses. This new assembly, annotations, and other genomic features discussed below will serve as valuable resources for investigating the economic and pharmacological characteristics of monk fruit and will also assist in the molecular breeding of monk fruit.

DNA libraries construction and sequencing

A total of 20 μg of genomic DNA was extracted from seedlings of *S. grosvenorii* (variety Qingpiguo) using a modified CTAB method [14] to construct two libraries with an insert size of 20 kb. The plants were introduced from the Yongfu Dis-
Candidate genes involved in the mogrosides biosynthesis pathway. Candidate functional genes were annotated as SQEs, EPHs, CDSs, CYP450s, and UGTs and assigned to the pathway.

Figure 2: Candidate genes involved in the mogrosides biosynthesis pathway. Candidate functional genes were annotated as SQEs, EPHs, CDSs, CYP450s, and UGTs and assigned to the pathway.

Candidate genes Numbers

| Candidate genes | Numbers |
|-----------------|---------|
| SQEs            | 1       |
| EPHs            | 9       |
| CDSs            | 1       |
| CYP 450s        | 78      |
| UGTs            | 38      |

A total of 300 ng of genomic DNA was extracted as described above, and the library was constructed using DNA sequence fragments of ~470 bp, with an approximate insert size of 350 bp. Sequencing was performed using a 2 × 150 paired-end (PE) configuration, and base calling was conducted using the HiSeq Control Software + Off-Line Base Caller (OLB) + GAPipeline-1.6.
Genome assembly and analysis of *Siraitia grosvenorii*

Table 1: SMRT reads used for genome assembly

| Statistics                  | Length (bp) |
|-----------------------------|-------------|
| Total raw data              | 31 G        |
| Mean length of raw reads    | 11 k        |
| N50 of raw reads            | 15,754      |
| Mean length of subreads     | 7.7 k       |
| N50 of subreads             | 11,898      |

Subreads: reads without adapters and low-quality bases.

Table 2: Metrics of *de novo* *S. grosvenorii* genome assembly

| Statistics                  | Contig | Contig (polished) |
|-----------------------------|--------|-------------------|
| Total number                | 4128   | 4128              |
| Total length (bp)           | 467,072,951 | 469,518,713      |
| N50 length (bp)             | 433,684 | 432,384           |
| N90 length (bp)             | 36,820  | 36,953            |
| Max length (bp)             | 7657,852 | 7683,850          |
| GC content (%)              | 33.57   | 33.49             |

Table 3: Summarized benchmarks of the BUSCO assessment

|                      | Monk fruit (%) |
|----------------------|----------------|
| Complete BUSCOs      | 89.2           |
| Complete and single-copy | 59.0       |
| Complete and duplicated | 30.2       |
| Partial              | 2.7            |
| Missing              | 8.1            |

Table 4: Quality evaluation of the draft genome with the overall alignment rate

| Sample       | Overall alignment rate |
|--------------|------------------------|
| FL-1         | 89.93%                 |
| FL-2         | 87.75%                 |
| FL-3         | 85.83%                 |
| ML-1         | 89.70%                 |
| ML-2         | 89.73%                 |
| ML-3         | 85.07%                 |
| L-1          | 85.95%                 |
| L-2          | 87.39%                 |
| R-1          | 81.50%                 |
| R-2          | 84.36%                 |
| R-3          | 84.57%                 |
| F1-1         | 84.35%                 |
| F1-2         | 91.58%                 |
| F2-1         | 86.83%                 |
| F2-2         | 87.37%                 |

RNA isolation and sequencing

Fresh roots, leaves, and early fruit of *S. grosvenorii* were sampled in our garden in Cangxi County. All samples were stored at -80°C after immediate treatment with liquid nitrogen. Total RNA was isolated from (1) leaves of female plants (FL), (2) leaves of male plants (ML), (3) leaves beside fruits (L), (4) roots (R), (5) fruit of 3 DAA (F1), and (6) fruit of 20 DAA (F2) using the Qiagen RNeasy Plant Mini Kits (Qiagen, CA, USA). PE libraries (PE150 = 5,000) were constructed and subsequently sequenced via the Illumina HiSeq X-Ten platform (Illumina, CA, USA).

Genome assembly

Initial correction of long reads was performed using FALCON (Falcon, RRID:SCR_016089) [15] with cuto- off length = 5,000 according to the distribution of read lengths and -B15, -s400 to cut reads into blocks of 400 Mb and align 15 blocks to another block at the same time. The 25x coverage of the longest corrected reads was extracted with Perl scripts and assembled by mecat2canu command of MECAT [15] with GenomeSize = 420,000,000 estimated in the previous study [8]. This led to a new genome assembly of 467 Mb with a contig N50 size of 434,684 bp (Table 2). This genome size was slightly larger than the estimated 420 Mb [8], which was likely due to the high genome heterozygosity. We used the consensus algorithm Quiver [16] and further polished the assembly with PE reads using Pilon (Pilon, RRID:SCR_014731) [17]. The final assembly produced 4,128 contigs, 614 of which were >100 kb long, with a contig N50 length of 432,384 bp (Table 2). Compared to the preliminary draft of the published *Siraitia* genome, the contiguity was improved more than ∼12.6 times.

Genome assessment

We estimated the completeness of the assembly using Benchmarking Universal Single-Copy Orthologues (BUSCO v2, RRID:SCR_015006) [18] analysis. Of the 1,440 orthologues identified in plants, 1,284 were found in the genome assembly, including 849 in single-copy and 435 in multi-copy (Table 3). In addition, we used RNA-Seq data from different organs to assess the sequence quality. All 15 RNA-Seq libraries were mapped to the assembly using HISAT2 (HISAT2, RRID:SCR_015530) [19], and the overall alignment rate for each data was used as a rough estimation of sequence quality. We also estimated the base error rate of the assembly with both DNA paired-end reads and published DNA short reads [8]. We used BWA-mem [20] to align both short reads to the genome assembly and filtered out low-quality (mapping quality < 30) alignments with SAMtools [21]. Then, we used the Genome Analysis Toolkit VariantFiltering program was used to filter out low-quality variants with the following expression: QD < 2.0 || ReadPosRankSum < -8.0 || FS > 60.0 || QUAL < 50 || DP < 10. Coverage of each alignment file was scanned using Qualimap 2 [23], and the error rate was calculated as the average number of short variants that appear at both alleles (labeled as 1/1 and 1/2 in Table 5) per base. The overall alignment rates of reads in all samples were over 80% (Table 4), and the average base error rate was estimated as less than 1E-3, which suggests a high-quality assembly (Table 5).

Repeat annotation

We scanned the genome using RepeatMasker (RRID:SCR_012954) [24] with Repbase [25] and a *de novo* repeat database constructed with RepeatModeler (RRID:SCR_015027) [26]. Sequences 240 Mb (51.14% of the assembled genome) in length were identified as repetitive elements, which was slightly larger than the 42.8% of *Momordica charantia* [27] and much larger than the 28.2% of *Cu-
Gene annotation

To generate gene models, the S. grossenorii genome was annotated using three gene prediction pipelines including homology-based, de novo, and RNA-Seq data-based prediction. First, we aligned the three cucurbitaceous proteomes downloaded from the cucubit database ([29], cucumber, Chinese Long_v2, melon_v3, and watermelon_97_103_v1) to the genome assembly using TBLASTN with an E-value of 1e-5 and filtering out bad hits (identity <50% and length <50%). The best hit of each retained protein was extracted and further used to predict protein coding gene structures with GeneWise ([30, 31]). Second, we de novo predicted protein coding genes using AUGUSTUS ([32]) with the repeat masked genome. Third, we used StringTie [33] to assemble 15 RNA-Seq alignment files (described above) generated from HISAT2 using the assembly as the reference and TransDecoder [34] to generate an annotation file based on transcripts. Finally, the three respective annotation files were combined using EVidenceModeler ([35]). After combining these gene structure predictions, we obtained 30,565 consensus protein-coding genes (Table 7). We annotated the genes using BLASTp searching against the NCBI nonredundant protein database (nt) and found that 78.3% of the predicted genes had at least one significant homologue (E-value < 1E-3), indicating that the gene structures were credible. We found that the majority of homologous proteins belonged to cucurbits, such as cucumber and muskmelon (Fig. 3). Protein domain and gene ontology term annotations were performed using InterProScan 5 (RRID:SCR_005829, Table 7) [36]. In addition, genes annotated as SQEs, EPHs, CDSs, EPFs, CYP450s, and UGTs were compared with those in other cucurbitaceous genomes, and we found that gene abundance in the five mogroside-related gene families were not significantly different among S. grossenorii, Cucumis sativus, Cucurbita moschata, and Cucurbita maxima ([29], Table 8).

Ortholog analysis

Gene family clustering analysis was accomplished using OrthoMCL ([37]) on protein sequences of S. grossenorii, C. sativus (cucumber_ChineseLong_v2, [29]), Cucumis melo (CM3.5.1, [29]), Citrullus lanatus (watermelon_97_103_v1, [29]), Prunus persica (Prunus.persica.prune1.0, [40]), Solanum lycopersicum (Solanum_lycopersicum.SL2.50, [40]), Arabidopsis thaliana (Tair10, [43]), and Oryza sativa (Oryza_sativa.IRGCSP-1.0.40) [45]. A total of 23,246 S. grossenorii genes were clustered into 26,190 gene families, including 1,471 unique S. grossenorii gene families (Fig. 4A). Compared to other cucurbitaceous plants, S. grossenorii shares fewer gene families with relative species (Fig. 4B), indicating an earlier divergence time than C. lanatus. A total of 834 single-copy gene families were identified and selected to construct the phylogenetic tree using RAxML ([38]) and the alignment was treated with Gblocks [49] with parameters of -t = p -b5 = h -b4 = 5 -b3 = 15 -d = e -n = y. The divergence time was estimated by MCMCtree [50]. Phylogenetic analysis showed that S. grossenorii diverged from the Cucurbitaceae family approximately 40.95 million years ago (Fig. 4C).
Table 7: Gene prediction and annotation

| Weight | Number of predicted genes | RNA-Seq data-based | Ab initio | Homology-based | Integration | Annotation |
|--------|---------------------------|--------------------|-----------|---------------|-------------|------------|
| 10     | 27,304                    | 0.1                | 5         | -             | nr          | -          |

| Tools | HISAT2 | RepeatMasker | AUGUSTUS | BLAST | EVM | BLAST | InterProScan |
|-------|--------|--------------|----------|-------|-----|-------|--------------|
|       | StringTie | TransDecoder |          |       |     |       |              |

Table 8: Abundance analysis of the mogrosides synthesis related gene families

| S. grosvenorii | C. sativus | C. moschata | C. maxima |
|----------------|------------|-------------|-----------|
| SQE            | 5 (5)      | 1           | 2         | 1         |
| EPH            | 30 (8)     | 23          | 29        | 22        |
| CYP450         | 276 (191)  | 213         | 289       | 234       |
| UGT            | 156 (131)  | 124         | 137       | 121       |
| CDS            | 1 (1)      | 1           | 2         | 3         |

The numbers quoted are the number of genes belonging to each gene family annotated in monk fruit genome version 1.

Transcriptomic analysis

Mogrosides are produced during fruit development in S. grosvenorii and are not found in vegetative tissues [8]. Thus, we performed an extensive transcriptomic analysis of early fruit at two stages (stage 1 sampled at 3 days after anthesis and stage 2 sampled at 20 days after anthesis) and of leaves to identify transcripts involved in mogroside synthesis in early fruit. Using the genome-wide annotation, RNA-Seq reads were mapped to the genome assembly, and read count tables were generated using HISAT2 and StringTie [33] for the next step of differential expression analysis. DESeq2 [RRID:SCR_000154] [51] was used to detect differential gene expression among L, F1, and F2 with the criteria of padj < 0.01 and |log2FoldChange| > 1. Genes that were up-regulated with fruit development were merged and used for KEGG pathway enrichment analysis with KOBAS (RRID:SCR_006350) [52]. Thirteen pathways were significantly enriched (corrected P < 0.01), and the most enriched pathways were related to metabolic pathways. In particular, the sesquiterpenoid and triterpenoid biosynthesis pathways were significantly enriched, indicating that genes involved in the biosynthesis of secondary metabolites, including mogrosides, perform their functions in the very early fruit (Fig. 5). Genes possibly related to mogrosides biosynthesis in early fruit according to the gene annotation were assigned to the mogrosides synthesis pathway (Fig. 2).

Discussion

S. grosvenorii is an important herbal crop with multiple economic and pharmacological values. Mogrosides, the main effective components of S. grosvenorii fruit, are partial substitutes of sucrose because of its extremely sweet and noncaloric charac-
teristics as more progress is made on molecular breeding and purification processes. Additionally, monk fruit could serve in contrast to other cucurbitaceous plant because of its earlier divergence from the common ancestor than some other well-studied cucurbits (cucumber, muskmelon), and it may be a new system for the investigation of plant sex determination. In the present study, we sequenced and assembled the second version of the monk fruit genome. With a great improvement in completeness and accuracy, the genome as well as the annotations will provide valuable resources and reference information for transcriptome assembly and novel gene discovery. These resources and further transcriptomics analysis of ripe fruit and young fruit will facilitate studies of the secondary metabolite synthesis pathways and monk fruit breeding.

**Availability of supporting data**

The genomic and transcriptomic sequencing reads were deposited in the Genome Sequence Archive under accession number CRA000522 and ENA (European Nucleotide Archive) under accession numbers PRJEB23465, PRJEB23466, and PRJEB25737. Supporting data are also available in the GigaScience database, GigaDB [53].

**Abbreviations**
Genome assembly and analysis of Siraitia grosvenorii

Carbon fixation in photosynthetic organisms
Sesquiterpenoid and triterpenoid biosynthesis
Phenylpropanoid biosynthesis
Carbon metabolism
Biosynthesis of amino acids
Photosynthesis
Starch and sucrose metabolism
Glycolysis / Gluconeogenesis
Tyrosine metabolism
Plant hormone signal transduction
Amino sugar and nucleotide sugar metabolism
Biosynthesis of secondary metabolites

Figure 5: KEGG pathway enrichment analysis of candidate functional genes

CDS: cucurbitadienol synthase; CYP450: cytochrome P450; EPH: epoxide hydrolase; F1: fruit of 3 DAA; F2: fruit of 20 DAA; FL: female plants; L: leaves beside fruits; ML: male plants; PacBio: Pacific Biosciences; PE: paired-end; R: root; SMRT: single molecule real time sequencing; SQE: squalene epoxidase; UGT: UDP-glycosyl-transferase.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
X.W.D., B.C., H.H., and M.X. planned and coordinated the project. M.X. collected and grew the plant material. R.Y. and G.Z. collected the samples and performed experiments. Genome assembly, annotation, phylogenetic analysis, and manuscript writing were completed by X.H., M.X., H.H., and X.W.D.

Acknowledgments
This research was supported by the National Key R&D Program of China (2017YFA0503800) to X.W.D. and in part by the National Demonstration Area of Modern Agriculture in Cangxi, Sichuan Province, China.

References
1. Zhang JS, Dai LH, Yang JG et al. Oxidation of cucurbitadienol catalyzed by CYP87D18 in the biosynthesis of mogrosides from Siraitia grosvenorii. Plant Cell Physiol 2016;57:1000–7.
2. Liu C, Dai LH, Dou DQ et al. A natural food sweetener with anti-pancreatic cancer properties. Oncogenesis 2016;5:1–7.
3. Cao C, Lin LM, Su F et al. Chemistry and pharmacology of Siraitia grosvenorii: A review. Chinese J Nat Med 2014;12:89–102.
4. Nie RL. The decadal progress of triterpene saponins from Cucurbitaceae (1980–1992). Acta Bot Yunnan 1994;16:201–8.
5. Wang Q, Qin HH, Wang W et al. The pharmacological research progress of Siraitia grosvenorii. J Guangxi Tradit Chin Med Univ 2010;13:75–76.
6. Zhang H, Li XH. Research progress on chemical compositions of Fructus Momordicae. J Anhui Agri Sci 2011;39:4555–56, 4559.
7. Pawar RS, Krynitsky AJ, Rader JI. Sweeteners from plants–with emphasis on Stevia rebaudiana (Bertoni) and Siraitia grosvenorii (Swingle). Anal Bioanal Chem 2013;405:4397–407.
8. Ikkin M, Davidovich-Rikanati R, Cohen S et al. The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from Siraitia grosvenorii. Proc Natl Acad Sci USA 2016;113:E7619–28.
9. Dai LH, Liu C, Zhu YM et al. Functional characterization of cucurbitadienol synthase and triterpene glycosyltransferase involved in biosynthesis of mogrosides from Siraitia grosvenorii. Plant Cell Physiol 2015;56:1172–82.
10. Zhang JS, Dai LH, Yang JG et al. Oxidation of cucurbitadienol catalyzed by CYP87D18 in the biosynthesis of mogrosides from Siraitia grosvenorii. Plant Cell Physiol 2016;57:1000–7.
11. Tang Q, Ma XJ, Mo CM et al. An efficient approach to finding Siraitia grosvenorii triterpene biosynthetic genes by RNA-seq and digital gene expression analysis. BMC Genomics 2011;12:343.
12. Shibuya M, Adachi S, Ebizuka Y. Cucurbitadienol synthase, the first committed enzyme for cucurbitacin biosynthesis, is a distinct enzyme from cycloartenol synthase for phytosterol biosynthesis. Tetrahedron 2004;60:6995–7003.
13. Zimin AV, Stevens KA, Crepeau MW et al. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience 2017;6:1–4.
14. Porubsky S, Bailey LG, Baum BR. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Rep 1997;15:8–15.
15. Xiao CL, Chen Y, Xie SQ et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods 2017;14:1072–4.
16. Chien CS, Alexander DH, Marks P et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 2013;10:563–9.
17. Walker BJ, Abeel T, Shea T et al. Pilon: an integrated tool for
comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9:e112963.

18. Simão FA, Waterhouse RM, Ioannidis P et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31:3210–2.

19. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357–60.

20. Li H, Handsaker B, Wysoker A et al. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 2009;25:2078–79.

21. Li H, Handsaker B, Wysoker A et al. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 2009;25:2078–79.

22. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303.

23. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 2016;32:292–94.

24. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 2009;3:4–14.

25. Visser M, Van der Walt AP, Maree HJ et al. Extending the sRNAome of apple by next-generation sequencing. PLoS One 2014;9:e95782.

26. Smit A, Hubley R. RepeatModeler Open-1.0.8. 2008; http://www.repeatmasker.org/RepeatModeler.html.

27. Urasaki N, Takagi H, Natsume S et al. Draft genome sequence of bitter gourd (Momordica charantia), a vegetable and medicinal plant in tropical and subtropical regions. DNA Res 2016;24:51–58.

28. Huang S, Li R, Zhang Z et al. The genome of the cucumber, Cucumis sativus L. Nat Genet 2009;41:1275–81.

29. Cucurbit Genomics Database: http://cucurbitgenomics.org. Accessed 12 Jun 2018.

30. Wise2: https://www.ebi.ac.uk/~birney/wise2/, Accessed 12 Jun 2018.

31. Gupt V, Estrada AD, Blakley I et al. RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing. Gigascience 2015;4:5.

32. Stanke M, Tzvetkova A, Morgenstern B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol 2006;7:S11.1–8.

33. Pertea M, Pertea GM, Antonescu CM et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33:290–5.

34. TransDecoder GitHub: https://github.com/TransDecoder/TransDecoder, Accessed 12 Jun 2018.

35. Haas BJ, Salzberg SL, Zhu W et al. Automated eukaryotic gene structure annotation using EVIDencEModeler and the Program to Assemble Spliced Alignments. Genome Biol 2008;9:R7.

36. Quevillon E, Silventoinen V, Pillai S et al. InterProScan: protein domain identifier. Nucleic Acids Res 2005;33:W116–20.

37. Li L, Stoeckert CJr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003;13:2178–89.

38. Garcia-Mas J, Benjak A, Sanseverino W et al. The genome of melon (Cucumis melo L.). Proc Natl Acad Sci USA 2012;109:11872–77.

39. Guo S, Zhang J, Sun H et al. The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. Nat Genet 2013;45:51–8.

40. Ensembl Plants: https://plants.ensembl.org/, Accessed 12 Jun 2018.

41. International Peach Genome Initiative, Verde I, Abbott AG et al. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 2013;45:487–94.

42. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 2012;485:635–41.

43. TAIR: http://Arabidopsis.org/, Accessed Jun 2018.

44. Lamesch P, Berardini TZ, Li D et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 2012;40:D1202–10.

45. International Rice Genome Sequencing Project. The mapped sequence of the rice genome. Nature 2005;436:793–800.

46. Stamatakis A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–13.

47. MUSCLE: https://www.ebi.ac.uk/Tools/msa/muscle/, Accessed 12 Jun 2018.

48. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–97.

49. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56:564–77.

50. Battistuzzi FU, Billing-Ross P, Palivai A et al. Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. Mol Biol Evol 2011;28:2439–42.

51. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

52. Xie C, Mao X, Huang J et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 2011;39:W316–22.

53. Xia et al. | 9

54. Cucumis melo L.). Proc Natl Acad Sci USA 2012;109:11872–77.

55. Guo S, Zhang J, Sun H et al. The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. Nat Genet 2013;45:51–8.

56. Stamatakis A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–13.

57. MUSCLE: https://www.ebi.ac.uk/Tools/msa/muscle/, Accessed 12 Jun 2018.

58. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–97.

59. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56:564–77.

60. Battistuzzi FU, Billing-Ross P, Palivai A et al. Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. Mol Biol Evol 2011;28:2439–42.

61. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

62. Xie C, Mao X, Huang J et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 2011;39:W316–22.

63. Xia et al. | 9

58. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–97.

59. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56:564–77.

60. Battistuzzi FU, Billing-Ross P, Palivai A et al. Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. Mol Biol Evol 2011;28:2439–42.

61. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.