Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

Chromosome-level reference genome of the Soursop (Annona muricata), a new resource for Magnoliid research and tropical pomology

Joeri Strijk[1], Damien Hinsinger[1,2], Mareike Roeder[3], Lars Chatrou[4], Thomas Couvreur[5,6], Roy Erkens[7], Hervé Sauquet[8], Michael Pirie[9], Daniel Thomas[10], and Kunfang Cao[1]

[1]Guangxi University
[2]Alliance for Conservation Tree Genomics, Pha Tad Ke Botanical Garden
[3]Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences
[4]Ghent University
[5]Institut de recherche pour le développement
[6]Institut de Recherche pour le Développement
[7]Maastricht University
[8]Royal Botanic Gardens and Domain Trust
[9]Johannes Gutenberg University of Mainz
[10]Singapore Botanic Gardens

June 1, 2020

Abstract

Annonaceae contain important commercially grown tropical crops, but development of other promising species is hindered by a lack of genomic resources to build breeding programs. In addition, Annonaceae are part of the Magnoliids, an ancient lineage of angiosperms for which evolutionary relationships with other major clades have remained unclear. To provide resources to both breeders and evolutionary researchers, we report the chromosome-level genome assembly of the soursop (Annona muricata). We assembled the soursop genome using a total of 444.32 Gb of DNA sequences that were generated using PacBio and Illumina short-reads, in combination with 10XGenomics, Bionano data and Hi-C sequencing. 949 scaffolds were assembled to a final size of 656.77Mb, with a scaffold N50 of 3.43 Mb. Repeat sequences accounted for 54.87% of the genome, and 23,375 protein-coding genes with an average of 4.79 exons per gene were annotated using de novo, RNA-seq and homology-based approaches. Reconstruction of the historical population size of A. muricata showed a slow but regular contraction of the population, likely related to Cenozoic climate changes. The soursop is the first genome assembled in Annonaceae, supporting further studies of floral evolution in Magnoliids, and providing an essential resource for delineating relationships of major lineages at the base of the angiosperms. Both genome-assisted improvement and conservation efforts will be strengthened by the availability of the soursop genome. The genome assembly as a community resource will further strengthen the role of Annonaceae as model species for research on the ecology, evolution and domestication potential of tropical species in pomology and agroforestry.

Keywords

magnoliids, Annonaceae, crop improvement, basal angiosperms, pomology, high quality draft genome

Introduction

Since the publication of the first plant genome (*Arabidopsis thaliana;* Arabidopsis genome Initiative 2000), there has been a steady increase in the number of sequenced eudicot and monocot genomes. However, with

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

the exception of the iconic *Amborella trichopoda* , angiosperm diversity represented by the ancient lineages of Nymphaeales, Austrobaileyales, Chloranthales, and magnoliids has largely been overlooked. After eudicots and monocots, Magnoliidae are the most diverse clade of angiosperms (Massoni et al. 2014) with 9,000-10,000 species in four orders (Canellales, Piperales, Laurales and Magnoliales). Despite this diversity and economic value (e.g. avocado, black pepper, cinnamon, soursop), only four genomes in three families have been published to date (Chaw et al. 2019; Chen et al. 2019; Hu et al. 2019; Rendón-Anaya et al. 2019). Analysis of such genomic data was expected to resolve the still unclear relationships of magnoliids with the rest of angiosperms (Soltis and Soltis 2019). However, recently published results strongly disagree on the position of magnoliids, supporting either a sister relationship to eudicots and monocots (Chen et al. 2019; Hu et al. 2019; Rendón-Anaya et al. 2019), or to eudicots alone Chaw et al. (2019).

Here we report the genome sequence of *Annona muricata* (the soursop) which is one of the c. 2450 species of the custard apple family (Annonaceae) (Rainer and Chatrou 2014), the second most species-rich family of magnoliids (Chatrou et al. 2012). Its species are frequent components of tropical rain forests worldwide (Gentry 1993; Tchouto 2006; Punyasena 2008; Sonké 2014). Widely known examples include ylang-ylang (*Cananga odorata*), used for its essential oils, and species of the Neotropical genus *Annona* , cultivated for their edible fruits, medicinal and pharmaceutical properties.

Annona muricata originated in the Caribbean and Central America, but is now widely cultivated in tropical and subtropical regions around the world. It is a small (up to 9 m), evergreen tree, typically with hairy branches when young. Leaves are oblong to oval, with a glossy green surface, while flowers are simple, with green sepals and thick yellowish petals (Figure 1a-b). The fruits are ovoid, dark green and tuberculate (Figure 1c-d), and can be up to 30 cm long, with a moderately firm texture. Their flesh is juicy, acidic, whitish and aromatic. The fruit contains significant amounts of vitamins (e.g. vitamin C, vitamin B1 and B2), but also the neurotoxic annocianin. Both fruits and leaves as well as seeds have been long used to treat a wide range of ailments owing to its pharmacological activities (e.g. anti-microbial, -leishmanial, -hyperglycemic, -parasitic, -inflammatory, -neuralgic, -rheumatic a.o.). More recently, research has focussed increasingly on the potential of using compounds extracted from*Annona muricata* parts to treat various lines of carcingeous cell lines.

Materials and Methods

Genomic DNA extraction, Illumina sequencing and genome size estimation

High-quality genomic DNA was extracted from freshly frozen leaf tissue of *A. muricata* using the Plant Genomic DNA Kit (Tiangen, Beijing, China), following manufacturer's specification. After purification, a short-insert library (300˜350 bp) was constructed and sequenced on the Illumina HiSeq 2500 platform (Illumina Inc., San Diego, CA, USA), according to manufacturer's specifications. A total of ˜65.47 Gb of raw data were generated. Sequencing adapters were then removed from the raw reads and reads from non-nuclear origin (chloroplast, mitochondrial, bacterial and viral sequences, etc.), screened by aligning them to the nr database (NCBI,*http://www.ncbi.nlm.nih.gov*, accessed on 12/07/2017) using megablast v2.2.26 with the parameters '*-v 1 -b 1 -e 1e-5 -m 8 -a 13* '; The script*duplication_rm.v2* (Strijk et al. 2014) was used to remove the duplicated read pairs; low-quality reads were filtered as follows: 1) reads with [?]10% unidentified nucleotides (N) were removed; 2) reads with adapters were removed; and finally, 3) reads with >20% bases having Phred quality <5 were removed. After the removal of low-quality and duplicated reads, ˜65 Gb of clean data (Table 1) were used for the genome size estimation, based on the 17-mer frequency of Illumina short reads. The formula - '*genome size = (total number of 17-mer)/(position of peak depth)* ' - was used to obtain an estimate of 799.11 Mb. An additional library was built (250 bp), sequenced as above and combined with the 350bp library to generate approximately 900 millions reads to provide a first estimation of the GC content, heterozygosity rate and repeat content.

PacBio, 10X Genomics and Bionano library preparation and sequencing

A 20 kb insert size PacBio library was built as previously described (Strijk et al. 2014). This library was sequenced on the PacBio RS II platform (Pacific Biosciences, Menlo Park, CA, USA), yielding about 37 Gb of data (read quality [?] 0.80, mean read length [?] 7 Kb). 10X Genomics DNA sample preparation, indexing,

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

and barcoding were done using the GemCode Instrument (10X Genomics, Pleasanton, CA, USA). About 0.7 ng of very high molecular weight DNA (>50 kb) was used for GEM reaction procedure during PCR, and 16 bp barcodes were introduced into droplets. Then, the droplets were fractured following the purifying of the intermediate DNA library. Next, we sheared the DNA into 500 bp for fragments constructing libraries, which were finally sequenced on the Illumina HiSeq X platform (Illumina Inc., San Diego, CA, USA), according to the manufacturer. A Bionano optical map was also constructed from Irys platform (BioNano Genomics, San Diego, CA, USA) from the same DNA, of which 95.9 Gb data were generated.

De novo Genome assembly, 10X and optical scaffolding

We used ALLPATHS-LG (Gnerre et al. 2011) and obtained a preliminary assembly of *A. muricata* with a scaffold N50 size of 19,908 kb and corresponding contig size of 8.26 Kb. We used PBjelly (English et al. 2012) to fill gaps with PacBio data. The options were set to "*<blasr>-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore -500 -nproc 10 -noSplitSubreads</blasr>* ". Then, we used Pilon (Walker et al. 2014) with default settings to correct assembled errors. For the input BAM file, we used BWA (Li and Durbin 2009) to align all the Illumina short reads to the assembly and SAMtools to sort and index the BAM file. This second assembly reached a contig N50 of approximately 700 kb. We used fragScaff (Adey et al. 2014) to generate scaffolds from this assembly using the optical map (95.9 Gb – 120.01x, Table 1) and 10X Genomics data (180.04 Gb – 225.30x, Table 1) with default parameters. We assessed the quality of the soursop genome assembly by mapping the Illumina reads back against the assembly.

Hi-C scaffolding

We constructed two Hi-C libraries from flash-frozen soursop leaves by cross-linking HMW gDNA in a 4% formaldehyde solution at room temperature in a vacuum for 30 mins. 2.5 M glycine was added to stop the crosslinking reaction for 5 min, then the sample was kept on ice for 15 min. The sample was then centrifuged at 2500 rpm at 4degC for 10 mins, and the pellet was washed with 500 µl PBS, then centrifuged for 5 min at 2500 rpm. The pellet was resuspended with 20 µl of lysis buffer (1 M Tris-HCl, pH 8, 1 M NaCl, 10% CA-630, and 13 units protease inhibitor), then the supernatant was centrifuged at 5000 rpm at room temperature for 10 min. The pellet was washed twice in 100 µl ice cold 1x NEB buffer and then centrifuged for 5 min at 5000 rpm. The nuclei were re-suspended by 100 µl NEB buffer and solubilized with dilute SDS followed by incubation at 65°C for 10 min. The SDS was neutralized by Triton X-100, then an overnight digestion was applied to the samples with a 4-cutter restriction enzyme *Mbo* I (400 units) at 37°C on a rocking platform.

This was followed by marking the DNA ends with biotin-14-dCTP and blunt-end ligation of the cross-linked fragments. The proximal chromatin DNA was re-ligated by ligation enzymes. The nuclear complexes were reverse cross-linked by incubation with proteinase K at 65°C. DNA was purified using a standard phenol-chloroform extraction protocol (Sambrook and Russell 2006). Biotin was removed from non-ligated fragment ends using T4 DNA polymerase. Sonication-sheared fragment ends (200-600 bp) were repaired using a mixture of T4 DNA polymerase, T4 polynucleotide kinase and Klenow DNA polymerase. Biotin-labeled Hi-C samples were specifically enriched using streptavidin C1 magnetic beads. After adding A-tails to the fragment ends and following ligation by the illumina paired-end (PE) sequencing adapters, Hi-C sequencing libraries were amplified by PCR (12-14 cycles) and sequenced on an Illumina NovaSeq platform (PE 150 bp).

After quality assessment, 66.1 Gb of Illumina reads were retained and assessed for Hi-C cross-linking efficiency using HiCUP (included in Juicer tools 1.5). The Hi-C clean data were aligned against the scaffold assembly using BWA (Camacho et al. 20009). Only the read pairs with both reads aligned to contigs were considered for scaffolding. According to the physical coverage of the resulting alignment (defined as the total bp number spanned by the sequence of reads and the gap between the two reads when mapping to contigs), any misassembly (identified by a drop in the physical coverage of reads along the contig) was split and further considered as two contigs. We used LACHESIS (Lowe and Eddy 1996) to assemble, order and orientate the scaffolds of our draft genome into the 7 Chromosomes of the soursop (Supplementary Table S4).

Repeat sequences in the soursop genome

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

Transposable elements in the genome assembly were identified both at the DNA and protein level. We used RepeatModeler (Smit and Hubley 2008) to develop a de novo transposable element library. RepeatMasker (Smit and Hubley 2017) was applied for DNA-level identification using Repbase and the de novo transposable element library. At the protein level, RepeatProteinMask was used to conduct WU-BLASTX (Camacho et al. 20009) searches against the transposable element protein database. Overlapping transposable elements belonging to the same type of repeats were merged.

The tRNA genes were identified by tRNAscan-SE (Lowe and Eddy 1996) with eukaryote parameters. The rRNA fragments were predicted by aligning them with *Arabidopsis thaliana* and *Oryza sativa* template rRNA sequences using BlastN (Camacho et al. 20009) at E-value of 1E-10. The miRNA and snRNA genes were predicted using INFERNAL (Nawrocky and Eddy 2013) by searching against the Rfam database (Nawrocky et al. 2014).

Gene annotation

RNA preparation sequencing and transcriptome assembly

Total RNA was extracted from leaves, flowers, bark and both young and ripe fruits (Table 1) using the RNAprep Pure Plant Kit, and genomic DNA contamination was removed using RNase-Free DNase I (both from Tiangen, Beijing, China). The integrity of RNA was evaluated on a 1% agarose gel, and its quality and quantity were assessed using a NanoPhotometer spectrophotometer (IMPLEN, Munich, Germany) and an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbroon, Germany). RNA sequencing (RNA-Seq) libraries were constructed using the NEBNext mRNA Library Prep Master Mix Set for Illumina (New England Biolabs, Beverly, MA, USA) following the manufacturer's instructions. The PCR products obtained were purified (AMPure XP system, Beckman Coulter Inc., Indianapolis, IN, USA) and library quality was assessed on the Agilent Bioanalyzer 2100 system. Library preparations were sequenced on an Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA), generating 100-bp paired-end reads. Raw reads were filtered by removing those containing undetermined bases ('N') or excessive numbers of low-quality positions (>10 positions with quality scores <10). Then the high-quality reads were mapped to the A. muricata genome using Tophat (v2.0.9) (Kim et al. 2013) with the parameters of '*-p 10 -N 3 –read-edit-dist 3 -m 1 -r 0 –coverage-search –microexon-search* '.

Annotation

Protein coding genes were predicted through a combination of homology-based prediction, de novo predictions and transcriptome based predictions, using the repeat-masked genome sequence: 1) Structural annotation of protein coding genes and protein domains was performed by aligning the protein sequences of the soursop against a representative set of angiosperms (*Amaranthus hypochondriacus* , *Amborella trichopoda* , *Aquilegia coerulea* , *Arabidopsis thaliana* ,*Coffea canephora* , *Musa acuminata* , *Nelumbo nucifera* ,*Oryza sativa* , *Vitis vinifera*) using TblastN (Camacho et al. 20009) with an E-value cutoff of 1E-5. The blast hits were conjoined by Solar (Yu et al. 2006) and for each, Genewise (Birney et al. 2004) was used to predict the exact gene structure in the corresponding genomic regions; 2) Five ab initio gene prediction programs, including Augustus (Stanke et al. 2006), Genscan (Burge and Karlin 1997), GlimmerHMM (Majoros et al. 2004), Geneid (Blanco and Abril 2009) and SNAP (Korf 2004) were used to predict coding genes on the repeat-masked genomes; 3) Finally, RNA-seq data were mapped to the genome using Tophat (Kim et al. 2013), and then cufflinks (Trapnell et al. 2012) was used to assemble transcripts to gene models.

All gene models predicted from the above three approaches were combined into a non-redundant set of gene structures with EVidenceModeler (EVM) (Haas et al. 2008). Then we filtered out low quality gene models based on 2 criteria : (1) coding region lengths of [?]150 bp and (2) those supported only by ab initio methods and with FPKM<1.

Functional annotation of protein coding genes was evaluated by BLASTP (Camacho et al. 20009) (e-value 1E-05) against two integrated protein sequence databases – SwissProt and TrEMBL (Boeckmann et al. 2003). The annotation information of the best BLAST hit derived from the database, was transferred to our

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

gene set. Protein domains were annotated by searching InterPro (Hunter et al. 2008) and Pfam (El-Gebali 2018) databases, using *InterProScan* (Quevillon et al. 2005) and*Hmmer* (Finn et al. 2011), respectively. Gene Ontology (GO) terms for each gene were obtained from the corresponding InterPro or Pfam entry. The pathways in which the gene might be involved were assigned by blasting against the KEGG database (https://www.genome.jp/kegg/), with an E-value cut-off of 1E-05.

Positive selection

To detect positive selection on protein-coding sequences, we calculated the number of synonymous substitutions per site (Ks) and nonsynonymous substitutions per site (Ka) for a set of angiosperms (*Amborella trichopoda* , *Arabidopsis thaliana* , *Helianthus annuus* ,*Nelumbo nucifera* and *Oryza sativa*) in addition to *A. muricata* . A ratio Ka/Ks > 1 is an indication of positive selection. We used MUSCLE (Edgar 2004) to generate MSA for the protein and nucleotide sequences, and Gblocks (Castresana 2002) with default parameters (-b3 8;-b4 10;-b5 n) to remove poorly aligned positions of alignments. The maximum likelihood-based branch test implemented in the PAML package (Yang 2007) was used to produce an estimate of the genic Ka/Ks ratio, calculated from the entire length of the protein sequences.

Inference of historical changes in population size

We used PSMC (Liu and Hansen 2017) to infer the variation in population size of the soursop based on the observed heterozygosity in the diploid genome. As PSMC was shown to perform reliably for scaffolds >100 kb, we removed shorter scaffolds from the assembly. 312 scaffolds > 100 kb were kept, totalling 646.64 Mb (98.46 percent of the total assembly). We assume a generation time of 15 years (Collevatti et al. 2014) and a per-generation mutation rate of 7x10-9. PSMC was otherwise conducted using default parameters.

Organellar genome reconstruction

The chloroplast of *Annona muricata* was reconstructed using GetOrganelle (Jin et al. 2019), with a subset of the Illumina paired reads (˜18 millions reads) and default parameters. The three Illumina libraries were then mapped against the resulting circular contig to detect any misassemblies. The draft plastome was annotated using CPGAVAS 2 (Shi et al. 2019), using default settings and deposited in GenBank (GB number pending).

Mapping of hybridization capture data

To exemplify usefulness of the v2 assembly, we mapped the data from Couvreur et al. (2019). As *Annona muricata* was not present in this study, we retrieved the raw reads of *Annona glabra* , filtered them by removing any position from both ends with a quality < Q20, and mapped them against our v2 assembly using Bowtie2, with default parameters. Regions with a mapping depth >30 were annotated in Geneious Prime (Biomatters, Ltd., Auckland, New Zealand), and displayed on the chromosomes using circos 0.69-9 (Krzywinski et al. 2009).

Results and Discussion

High quality Annona genome

We estimated genome size and heterozygosity to be 799.11 Mb and 0.08%, respectively, with a repeat content of 59.76%. The GC content ranged from 35.46% (350 bp library) to 37.64% (250bp library). The first genome assembly using only Illumina data and the assembly program SOAPdenovo (Luo et al. 2012) was approximately 595.5 Mb, with a contig N50 of 8,258 bp, a scaffold N50 of 19,908 bp (620.3Mb total length).

A total of 444.32 Gb of data were produced using Illumina, PacBio, 10X Genomics and Bionano technologies, corresponding to 556x coverage of the soursop genome. This sequencing strategy provided sequencing depths of 163X, 46X, 225X and 120X for Illumina, PacBio, 10X Genomics and Bionano libraries sequencing, respectively (Table 1).

The first scaffolding step resulted in a v1 assembly comprising 949 scaffolds, with a scaffold N50 length of 3.43 Mb (Table 2) for a total assembly length of 656.78Mb. The longest scaffold was 20.46 Mb (GC

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

content of 34.35%) and 29 scaffolds were longer than 5Mb. Scaffolds longer than 100kb totalled 646.64Mb (98.45% of the total length). This level of contiguity is similar to that obtained in *Liriodendron chinensis* (N50=3.5 Mb (Chen et al. 2019)) but smaller than obtained in*Cinnamomum kanehirae* (N50=50.4Mb after Hi-C scaffolding (Chaw et al. 2019)) and better than other genomes assembled at scaffold-level (Wei et al. 2018; Arimoto et al. 2019; Zhang et al. 2019). A total of 97.16% reads can be mapped, covering >99.92% of the genome, excluding gaps. 99.81% of the genome was covered with a depth >20x, which guaranteed the high accuracy of the assembly for SNPs detection (Supplementary Table 1). SNP calling on the final assembly yielded a heterozygosity rate of 0.032%, lower than 0.08% as estimated by the K-mer analysis (Supplementary Figure 1). We then Used Hi-C scaffolding to improve the v1 assembly and produce a chromosome-level assembly, hereafter referred as "v2 assembly". Assembly statistics after Hi-C scaffolding are summarized in Table 2. The *Annona muricata* genome information after HI-C scaffolding is summarized in Table 3. Sequencing quality assessment is shown in Supplementary Table S2. Statistics for the final soursop genome assembly are as follows: the total length of contig is 652,885,881 bp, the length of contig N50 reaches 743,350 bp; the total length of scaffold is 656,813,740 bp, and the length of scaffold N50 reaches 93,205,713 bp. 97.38% of the contigs from the v1 assembly were included in the v2 assembly.

Repeat sequences in the Annona genome

Repeats accounted for 54.87% of the genome, an intermediate value between *Cinnamomum* (Lauraceae, 48%) and *Liriodendron*(Magnoliaceae, 63.81%). Long Terminal Repeat (LTR) retrotransposons were the most abundant TE, representing 41.28% of the genome (56.25% in *Liriodendron chinense*), followed by DNA repeats (7.29%) (Supplementary Table 2, Figure 2a). The stout camphor tree genome exhibited a different balance between types, with LTR (25.53%) and DNA transposable elements (12.67%) being less dominant. No significant recent accumulation of LTRs and LINEs was found in the interspersed repeat landscape, but a concordant accumulation around 40 units was detected (Figure 2b). Assuming a substitution rate similar to the one found in *Liriodendron* (1.51x10-9 subst./site/year), we estimate this burst of transposable elements to have occurred 130-150 Ma ago. By far the main contribution to this old expansion of repeat copy-numbers were the LTRs, with an increase of up to approximately 1% at 42 units. We identified 1201 microRNA, 560 transfer RNA (tRNA), 315 ribosomal RNA (rRNA), and 3198 small nuclear RNA (snRNA) genes (Supplementary Table 3).

Genes involved in plant defense and disease resistance

We identified 23,375 genes, 21,336 of them supported by at least two of the predictive methods described above (Supplementary Figure 2), with an average coding-region length of 1.1 kb and 4.79 exons per gene, similar to other angiosperms (Supplementary Table 4). We assessed both the quality of our gene predictions and completeness of our assembly using BUSCO (Simao et al. 2015) and CEGMA (Parra et al. 2007). 231 CEGs genes (93.15%) and 899 (94%) of the BUSCO orthologous single copy genes were retrieved from the soursop assembly (Supplementary Table 5).

22,769 (97.4%) genes were annotated through SwissProt and TrEMBL and GO-terms were retrieved for 20,595 (88.1%) genes (Supplementary Table 6). Comparing gene content in *Annona* with that found in the stout camphor tree, we found a striking difference in diversity of resistance genes. Of 387 resistance genes in *Cinnamomum* , 82% were nucleotide-binding site leucine-rich repeat (NBS-LRR) or with a putative coiled-coil domain (CC-NBS-LRR). By contrast, the soursop genome contains a similar number of resistance genes (301 annotations), but only 0.66% (2 genes) of them are NBS-LRR or CC-NBS-LRR genes. These results suggest the presence of different evolutionary strategies within magnoliids with respect to pathogen resistance.

We identified 77 genes putatively under positive selection (p-value<0.01, FDR < 0.05). We identified the 10 most enriched gene families and retrieved their GO-terms. Two families have GO-terms (Supplementary Table 7) and none of these families have a defined KEGG pathway.

Historical fluctuations in population size

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

We determined *Annona muricata* exhibits heterozygous and homozygous SNP ratios of 0.0032% and 0.0001%, respectively. This very low heterozygosity, usually found in cultivated species that experienced strong bottlenecks during domestication (Eyre-Walker et al. 1998; Doebley et al. 2006; Zhu et al. 2007), was not due to an intense, recent decrease in population size, as shown by our PSMC analysis. Instead, the very low heterozygosity observed in soursop was due to a slow and regular reduction of the species population sizes (Figure 3). The slow but regular reduction in population size of *A. muricata* is compatible with the Quaternary contraction of tropical regions in several parts of the world, and suggests that the soursop may have been severely affected by climate changes, as many other tropical taxa (Barlow et al. 2018). Contrary to the situation in most crop plants (Eyre-Walker et al. 1998), this reduction did not result from a genetic bottleneck during domestication. However, the very low heterozygosity in soursop could make future genetic improvement difficult, and will likely require outcrossing with wild relatives (Zamir 2001).

Mapping of genes from hybridization capture

We mapped the loci previously obtained using targeted enrichment of nuclear genes from the study of Couvreur et al. (2019) to the v2 assembly, and superimposed their position and density onto the circular chromosome map (Figure 1) using circos 0.69-9 (Krzywinski et al. 2009). A total of 2328 regions with gene coverage higher than 30 were identified across the genome. Mapping was significantly lower in the regions with high numbers of repeat sequences.

Conclusions

This study presents the first high-quality genome assembled for a plant in the Annonaceae - a large tropical tree family of global ecological and economic importance. The *Annona muricata* genome provides a vital resource for research on floral morphology diversity, on the early evolution of Magnoliids and on the conservation of this tropical tree species. The soursop genome is not only an exceptional resource for the scientific community, but also for breeders of other tropical trees (e.g. avocado, *Annona* species, pepper, *Magnolia*) as it provides novel data on disease resistance and plant defense. Of particular relevance is the positional information inherent in genome data, which is absent from transcriptomes, allowing breeders to use linkage disequilibrium estimation in their programs (Barabaschi et al. 2015).

Acknowledgements

Genome sequencing, assembly and annotation were conducted by Novogene Bioinformatics Institute, Beijing, China; Project No. P2016112416. We kindly acknowledge Ghent University Botanical Garden for granting access to their living collections of *A.muricata* .

Data accessibility

Raw reads were deposited to EBI (project PRJEB30626).

Abbreviations

BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; CDS: coding sequence; CEGMA: Core Eukaryotic Genes Mapping Approach; Gb: gigabase pairs; GC: guanine- cytosine; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; LINE: long interspersed nuclear element; LTR: long terminal repeats; Mb: megabase pairs; PacBio: Pacific Biosciences; PAML: Phylogenetic Analysis by Maximum Likelihood; SNAP: Scalable Nucleotide Alignment Program; TE: transposable element; tRNA: transfer RNA.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported through the Guangxi Province One Hundred Talent program and the Bagui Scholarship team funding under Grant No. C33600992001 to JSS, and the China Postdoctoral Science Foundation

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

(grant number 2015M582481 and 2016T90822) to DDH. The basis for this manuscript was laid down during the 2015 Dialogue seminar on Annonaceae under the Joint Scientific Thematic Research Programme (JSTP) funded by the Netherlands Organisation for Scientific Research and the Chinese Academy of Sciences (grant number 045.011.020). TLPC was supported by the Agence Nationale de la Recherche (grant AFRODYN: ANR-15-CE02-0002-01). MDP was supported by the Heisenberg programme of the Deutsche Forschungsgemeinschaft (PI 1169/3-1).

References

1. Arabidopsis genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–796.

2. Massoni J, Forest F, Sauquet H. Increased sampling of both genes and taxa improves resolution of phylogenetic relationships within Magnoliidae, a large and early-diverging clade of angiosperms. Mol Phylogenet Evol. 2014;70:84–93.

3. Chaw SM, Liu YC, Wu YW, Wang HY, Lin CY, Wu CS, Ke HM, Chang LY, Hsu CY, Yang HT, Sudianto E. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. Nature plants. 2019;5(1):63-73.

4. Chen J, Hao Z, Guang X, Zhao C, Wang P, Xue L, Zhu Q, Yang L, Sheng Y, Zhou Y, Xu H. Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. Nature plants. 2019;5(1):18-25.

5. Rendon-Anaya M, Ibarra-Laclette E, Mendez-Bravo A, Lan T, Zheng C, Carretero-Paulet L, Perez-Torres CA, Chacon-Lopez A, Hernandez-Guzman G, Chang TH, Farr KM. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. Proceedings of the National Academy of Sciences. 2019 Aug 20;116(34):17081-9.

6. Hu L, Xu Z, Wang M, Fan R, Yuan D, Wu B, Wu H, Qin X, Yan L, Tan L, Sim S. The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. Nature communications. 2019;10(1):1-1.

7. Soltis DE, Soltis PS. Nuclear genomes of two magnoliids. Nat Plants. 2019;5:6–6.

8. Rainer H, Chatrou LW. AnnonBase: World species list of Annonaceae [Internet]. 2014 [cited 2018 Dec 13]. Available from: https://www.catalogueoflife.org/col/details/database/id/40

9. Chatrou LW, Erkens RHJ, Richardson JE, Saunders RMK, Fay MF. The natural history of Annonaceae. Bot J Linn Soc. 2012;169:1–4.

10. Gentry AH. Four neotropical rainforests. New Haven: Yale University Press; 1993.

11. Tchouto MGP, Yemefack M, De Boer WF, De Wilde JJFE, Van Der Maesen LJG, Cleef AM. Biodiversity hotspots and conservation priorities in the Campo-Ma'an rain forests, Cameroon. Biodivers Conserv. 2006;15:1219–52.

12. Punyasena SW, Eshel G, McElwain JC. The influence of climate on the spatial patterning of Neotropical plant families. J Biogeogr. 2008;35:117–30.

13. Sonke B, Couvreur T. Tree diversity of the Dja Faunal Reserve, southeastern Cameroon. Biodivers Data J. 2014;2.

14. Strijk JS, Hinsinger DD, Zhang F, Cao K. Trochodendron aralioides, the first chromosome-level draft genome in Trochodendrales and a valuable resource for basal eudicot research. GigaScience. 2019 Nov;8(11):giz136.

15. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM. High-quality draft assemblies of mammalian genomes from massively parallel sequence

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

data. Proceedings of the National Academy of Sciences. 2011;108(4):1513-8.

16. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE. 2012;7:e47768–e47768.

17. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one. 2014;9(11).

18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

19. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, Ronaghi M, Amini S, Gunderson KL, Steemers FJ, Shendure J. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. Genome research. 2014 Dec 1;24(12):2041-9.

20. Sambrook J, Russell DW. Purification of nucleic acids by extraction with phenol: chloroform. Cold Spring Harbor Protocols. 2006;(1):pdb-rot4455.

21. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC bioinformatics. 2009;10(1):421.

22. Lowe TM, Eddy SR. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1996;25:955–64.

23. Smit AF, Hubley R. RepeatModeler Open-1.0 [Internet]. 2008 [cited 2018 Jul 6]. Available from: http://www.repeatmasker.org

24. Smit A, Hubley R, Green P. RepeatMasker Open-4.0.6 [Internet]. 2017 [cited 2018 Jul 6]. Available from: http://www. repeatmasker. org

25. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–2935.

26. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD. Rfam 12.0: updates to the RNA families database. Nucleic acids research. 2015;43(D1):D130-7.

27. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36–R36.

28. Yu X-J, Zheng H-K, Wang J, Wang W, Su B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics. 2006;88:745–751.

29. Birney E, Clamp M, Durbin R. GeneWise and genomewise. Genome Res. 2004;14:988–995.

30. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–W439.

31. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268:78–94.

32. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20:2878–2879.

33. Blanco E, Abril JF. Computational gene annotation in new genome assemblies using GeneID. Bioinforma DNA Seq Anal. Totowa: Humana Press; 2009. p. 243–261.

34. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

35. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012;7(3):562-78.

36. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology. 2008;9(1):R7.

37. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research. 2003;31(1):365-70.

38. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD. InterPro: the integrative protein signature database. Nucleic acids research. 2009;37(suppl_1):D211-5.

39. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer EL. The Pfam protein families database in 2019. Nucleic acids research. 2019;47(D1):D427-32.

40. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. Nucleic acids research. 2005;33(suppl_2):W116-20.

41. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39:W29–W37.

42. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5(1):113.

43. Castresana J. Gblocks, v. 0.91 b. online version available at http://molevol. cmima. csic. es/castresana. Gblocks_server. html (accessed 2 February 2010). 2002.

44. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 2007;24(8):1586-91.

45. Liu S, Hansen MM. PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. Mol Ecol Resour. 2017;17:631–41.

46. Collevatti RG, Telles MPC, Lima JS, Gouveia FO, Soares TN. Contrasting spatial genetic structure in Annona crassiflora populations from fragmented and pristine savannas. Plant Syst Evol. 2014;300:1719–27.

47. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. BioRxiv. 2019:256479.

48. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic acids research. 2019;47(W1):W65-73.

49. Couvreur TL, Helmstetter AJ, Koenen EJ, Bethune K, Brandao RD, Little SA, Sauquet H, Erkens RH. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. Frontiers in plant science. 2019;9:1941.

50. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome research. 2009;19(9):1639-45.

51. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1(1):2047-17X.

52. Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, Xia E, Lu Y, Tai Y, She G, Sun J. Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. Proceedings of the National Academy of Sciences. 2018;115(18):E4151-8.

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

53. Arimoto A, Nishitsuji K, Higa Y, Arakaki N, Hisata K, Shinzato C, Satoh N, Shoguchi E. A siphonous macroalgal genome suggests convergent functions of homeobox genes in algae and land plants. DNA Research. 2019;26(2):183-92.

54. Zhang T, Qiao Q, Novikova PY, Wang Q, Yue J, Guan Y, Ming S, Liu T, De J, Liu Y, Al-Shehbaz IA. Genome of Crucihimalaya himalaica, a close relative of Arabidopsis, shows ecological adaptation to high altitude. Proceedings of the National Academy of Sciences. 2019;116(14):7137-46.

55. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

56. Parra G, Bradnam K, Korf I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.

57. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS. Investigation of the bottleneck leading to the domestication of maize. Proc Natl Acad Sci U S A. 1998;95:4441–6.

58. Doebley JF, Gaut BS, Smith BD. The Molecular Genetics of Crop Domestication. Cell. 2006;127:1309–21.

59. Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: Severe bottleneck during domestication of rice. Mol Biol Evol. 2007;24:875–88.

60. Barlow J, Franca F, Gardner TA, Hicks CC, Lennox GD, Berenguer E, Castello L, Economo EP, Ferreira J, Guenard B, Leal CG. The future of hyperdiverse tropical ecosystems. Nature. 2018;559(7715):517-26.

61. Zamir D. Improving plant breeding with exotic genetic libraries. Nat Rev Genet. 2001;2:983–983.

62. Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Vale G, Cattivelli L. Next generation breeding. Plant Science. 2016;242:3-13.

Tables, Figures and Supplementary Materials

Table 1. Sequencing data used for the *A. muricata* genome assembly and annotation.

Step	**Technology**	**Tissue**	**insert size**	**Bases generated (Gb**
Genome assembly	Illumina reads	leaves	250 bp	65.96
			350 bp	65.47
	PacBio reads	leaves	20kb	36.95
	10X	leaves		180.04
	Bionano	leaves		95.9
	Total			444.32
Chromosomes scaffolding	Hi-C	leaves	N.A.	66.17
Genome annotation	Illumina reads	flowers (several developmental stages)	350 bp	5.52
		young fruit	350 bp	9.93
		ripening fruit	350 bp	5.73
		bark	350 bp	4.80
		leaves	350 bp	5.04
	Total			25.51

Table 2. Assembly statistics

	length	**length**
	Contig(bp)**	**Scaffold(**
Assembly v1 (Illumina+ PacBio + 10X + BioNano)	**Assembly v1** (Illumina+ PacBio + 10X + BioNano)	**Assembly**

	length	length
Total	652885881	656774640
Max	4254538	20459086
Number>=2000	-	-
N50	784561	3429555
N60	632116	2673626
N70	483912	2112119
N80	346983	1573287
N90	207456	964101
Assembly v2 (Assembly v1 + Hi-C)	**Assembly v2** (Assembly v1 + Hi-C)	**Assembly**
Total	652885881	656813740
Max	4254538	122620176
Number>=2000	-	-
N50	743350	93205713
N60	578736	89409058
N70	451341	85026703
N80	320782	69840041
N90	184498	60483854

** Contig after scaffolding

Table 3. Chromosome characteristics of the v2 assembly

	Chromosome name	Cluster Number	Sequences Length
Hic_asm_0	Amur4	49	89409058
Hic_asm_1	Amur1	68	122620176
Hic_asm_2	Amur3	57	93205713
Hic_asm_3	Amur2	75	118991926
Hic_asm_4	Amur7	34	60483854
Hic_asm_5	Amur5	62	85026703
Hic_asm_6	Amur6	53	69840041

Figure legends

Figure 1. ***Annona muricata* description and genomic landscape.** Top : a) leaves; b) mature flower; c) mature fruit. Bottom: Circular view of the chromosome organization of *Annona muricata* , with genomic features indicated from outer to inner layers in sequence windows of 200 kb; d) Structural organisation of the chromosomes arranged by size, indicated in Mb; e) loci density from Couvreur*et al.* 2019; f) GC deviation; g) GC content (percentage); h) Gene breadth (i.e. the percentage of the sequence window occupied by coding regions) heatmap; i) Gene density (i.e. the number of genes found in one sequence window) histogram; j) TE protein breadth heatmap; k) TE protein density histogram; l) Transposon breadth heatmap; m) Transposon density histogram. In i), k) and m), values above and below the mean are indicated in green and red, respectively.

Figure 2. TE characteristics in the soursop genome. a) Distribution of repeat classes in the soursop genome; b) Divergence distribution of transposable elements in the genome of *Annona muricata* . Both Kimura substitution level (CpG adjusted) and absolute time are given.

Figure 3. Population size variation in soursop. Effective population size history inferred by the PSMC method (black line), with 100 bootstraps shown (red lines).

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.

Supplementary Figure captions

Supplementary Figure 1. K-mer distribution analysis for genome size and heterozygosity estimation.

Supplementary Figure 2. Gene prediction support. Number of predicted genes supported by RNA-seq transcripts (rna_0.5), homology to known proteins (homolog_0.5) or ab initio inference (denovo_0.5).

Supplementary Table captions

Supplementary Table 1. Mapping rate and genome coverage of the*Annona muricata* assembly.

Supplementary Table 2. Classification of TEs content.

Supplementary Table 3. Non coding RNA content in the soursop genome.

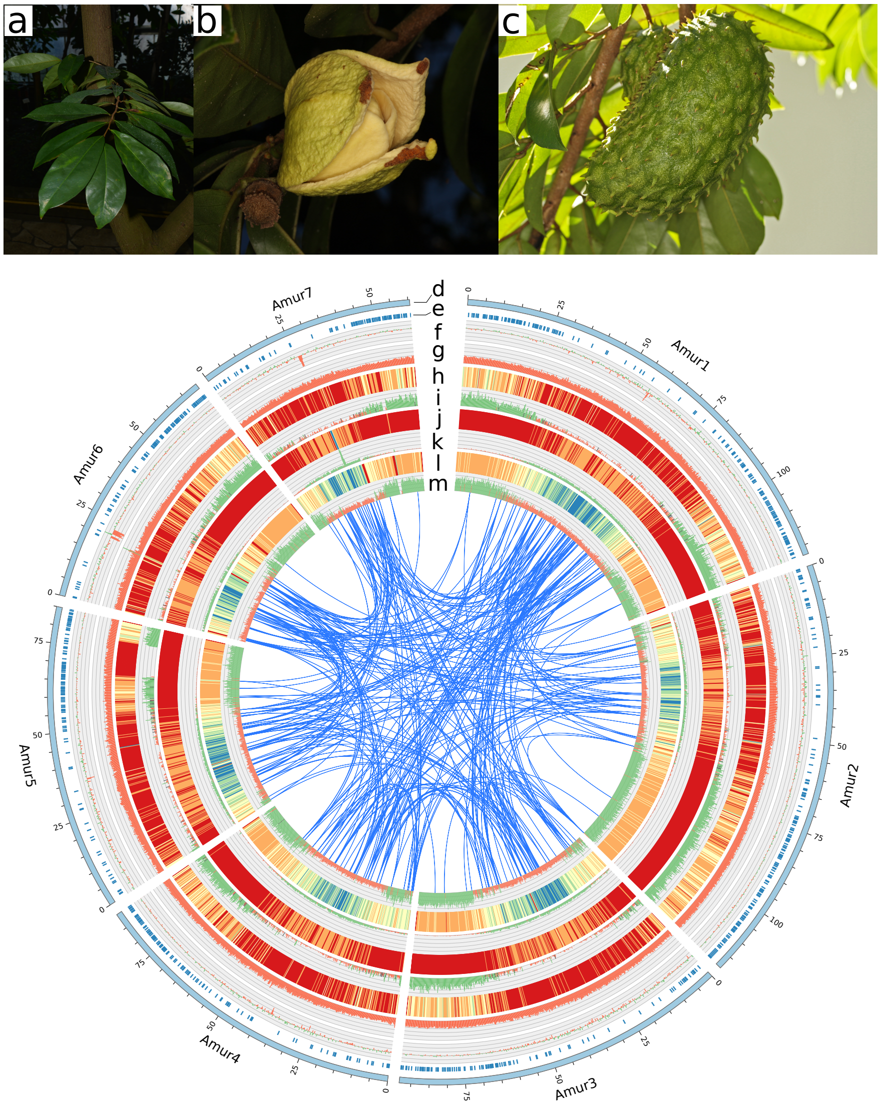
Supplementary Table 4. Characteristics of the annotated genes in 10 angiosperms species.

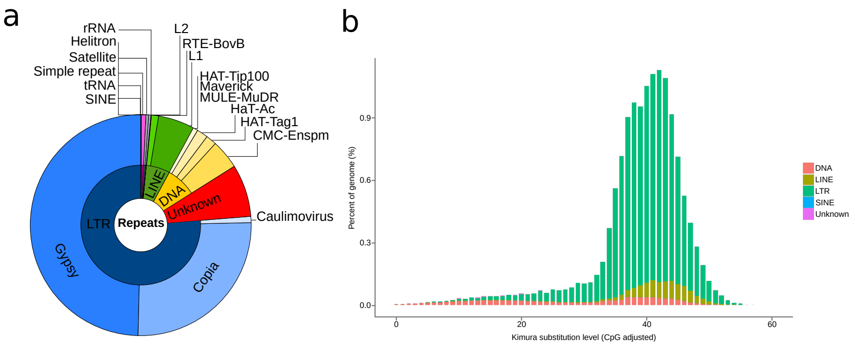
Supplementary Table 5. CEGMA and BUSCO assessments of the gene annotations.

Supplementary Table 6. Overview of annotated genes per database.

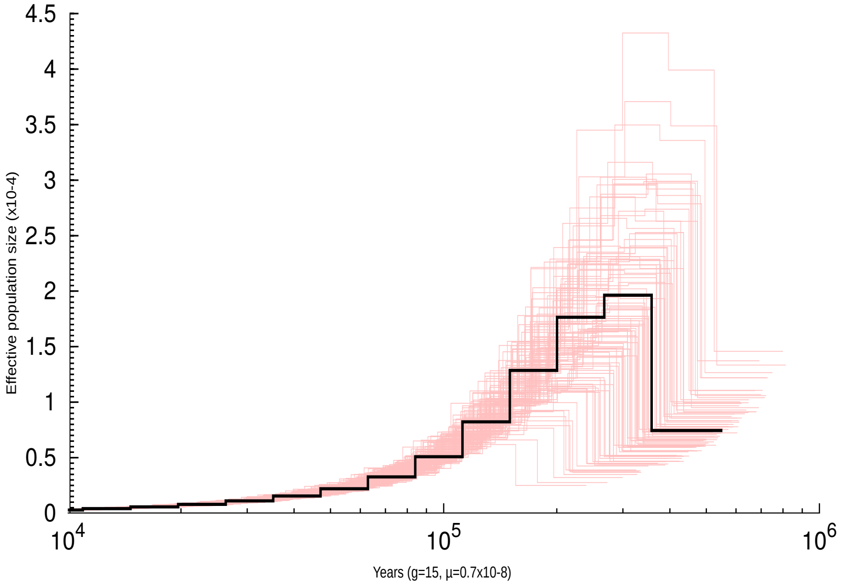
Supplementary Table 7. Retrieved GO-terms for the genes under positive values.

Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.





Posted on Authorea 1 Jun 2020 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/au.159103606.66673541 — This a preprint and has not been peer reviewed. Data may be preliminary.


Effective population size (x10-4)
10^4
10^5
10^6
Years (g=15, μ=0.7x10-8)