# A draft genome and transcriptome of common milkweed (*Asclepias syriaca*) as resources for evolutionary, ecological, and molecular studies in milkweeds and Apocynaceae

Kevin A. Weitemier[1,7]

Shannon C. K. Straub[2]

Mark Fishbein[3]

C. Donovan Bailey[4]

Richard C. Cronn[5]

Aaron Liston[6]

[1]Department of Fisheries and Wildlife, Oregon State University, 104 Nash Hall, Corvallis, OR 97331, USA

[2]Department of Biology, Hobart & William Smith Colleges, 113 Eaton Hall, Geneva, NY 14456, USA

[3]Department of Plant Biology, Ecology & Evolution, Oklahoma State University, 301 Physical Sciences, Stillwater, OK, 74078, USA

[4]Department of Biology, New Mexico State University, PO Box 30001, MSC 3AF, Las Cruces, NM, 88003, USA

[5]Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, OR, 97331, USA

[6]Department of Botany & Plant Pathology, Oregon State University, 2082 Cordley Hall, Corvallis, OR 97331, USA

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26488v2 | CC BY 4.0 Open Access | rec: 18 Apr 2019, publ: 18 Apr 2019

[7]Corresponding author:

Kevin Weitemier

Department of Fisheries and Wildlife, Oregon State University, 104 Nash Hall, Corvallis, OR 97331, USA

Fax: 541-737-3590

kevin.weitemier@oregonstate.edu

Keywords: Genome, *Asclepias*, milkweed, Apocynaceae, cardenolide, chromosome evolution

Running Title: *Asclepias* nuclear genomic resources

## ABSTRACT

Milkweeds (*Asclepias*) are used in wide-ranging studies including floral development, pollination biology, plant-insect interactions and co-evolution, secondary metabolite chemistry, and rapid diversification. We present a transcriptome and draft nuclear genome assembly of the common milkweed, *Asclepias syriaca*. This reconstruction of the nuclear genome is augmented by linkage group information, adding to existing chloroplast and mitochondrial genomic resources for this member of the Apocynaceae subfamily Asclepiadoideae. The genome was sequenced to 80.4× depth and the draft assembly contains 54,266 scaffolds ≥1 kbp, with N50 = 3415 bp, representing 37% (156.6 Mbp) of the estimated 420 Mbp genome. A total of 14,474 protein-coding genes were identified based on transcript evidence, closely related proteins, and ab initio models, and 95% of genes were annotated. A large proportion of gene space is represented in the assembly, with 96.7% of *Asclepias* transcripts, 88.4% of transcripts from the related genus *Calotropis*, and 90.6% of proteins from *Coffea* mapping to the assembly. Scaffolds covering 75 Mbp of the *Asclepias* assembly formed eleven linkage groups. Comparisons of these groups with pseudochromosomes in *Coffea* found that six chromosomes show consistent stability in gene content, while one may have a long history of fragmentation and rearrangement. The progesterone 5β-reductase gene family, a key component of cardenolide production, is likely reduced in *Asclepias* relative to other Apocynaceae. The genome and transcriptome of common milkweed provide a rich resource for future studies of the ecology and evolution of a charismatic plant family.

## INTRODUCTION

The development of genomic resources for an ever-increasing portion of the diversity of life is benefiting every field of biology in myriad ways. The decreasing cost of sequencing and the continual development of bioinformatic tools are allowing even single labs and small collaborations to produce genomic content that is beneficial and accessible to the wider research community. This study presents a draft genome assembly of a species of the milkweed genus *Asclepias* (Apocynaceae), which is the focus of diverse studies including floral development, pollination biology, plant-insect interactions and co-evolution, secondary metabolite chemistry, and rapid diversification.

*Asclepias* sensu stricto is made up of about 130 species in North and South America (Fishbein et al., 2011). The genus in the Americas is found in a wide range of habitats, from deserts to swamps, plains to shaded forests, and may represent a rapid ecological expansion (Fishbein et al., 2018). The common milkweed, *Asclepias syriaca* L., inhabits wide swaths of eastern North America, westward to Kansas, and northward to Canada (Woodson, 1954). It is well known for the milky latex exuded when injured, showy inflorescences, and pods filled with seeds tufted with fine hairs. Like other members of the Apocynaceae, milkweeds produce an array of potent secondary compounds, including cardiac glycosides (specifically cardenolides). Some herbivores possess defenses to avoid or tolerate these compounds, including the monarch butterfly, *Danaus plexippus*. Monarch caterpillars are able to sequester cardenolides from *Asclepias* to use for their own defense, and *Asclepias* are an essential host for monarchs (Brower et al., 1967). The variation within and among *Asclepias* species in types of and investments in defensive compounds and structures has engendered numerous studies of defensive trait evolution

(Agrawal et al., 2012; Agrawal and Fishbein, 2006, 2008; Fishbein et al., 2018; Livshultz et al., 2018; Rasmann et al., 2009, 2011), plant-herbivore ecological interactions (Brower et al., 1967, 1972; Van Zandt and Agrawal, 2004; Vaughan, 1979), and plant-herbivore co-evolution (Agrawal, 2005; Agrawal and Van Zandt, 2003; Labeyrie and Dobler, 2004).

As members of Apocynaceae subfamily Asclepiadoideae, *Asclepias* species possess floral architectures unique among plants, including floral coronas and a central gynostegium composed of the unified stamens and pistil. Most *Asclepias* species are nearly or entirely self-incompatible (Wyatt and Broyles, 1994), and their pollen is packaged into masses, pollinia, which are transferred as a unit from one flower to another. This usually allows a single successful pollination event to fertilize all of the ovules in an ovary, resulting in full-sibling families in each fruit (Sparrow and Pearson, 1948; Wyatt and Broyles, 1990). These features have positioned *Asclepias* as a model in studies of angiosperm reproductive biology (Broyles and Wyatt, 1990; Wyatt and Broyles, 1990, 1994), floral development (Endress, 2006, 2015), selection on floral characters and prezygotic reproductive isolation (La Rosa and Conner, 2017; Morgan and Schoen, 1997), and floral display evolution (Chaplin and Walker, 1982; Fishbein and Venable, 1996; Willson and Rathcke, 1974).

A few genomic resources have been developed for *Asclepias* and other Apocynaceae. The chloroplast and mitochondrial genomes of *Asclepias* have been sequenced (Straub et al., 2011, 2013), and flow cytometry estimates place the nuclear genome size of *A. syriaca* at 420 Mbp (Bai et al., 2012; Bainard et al., 2012). *Asclepias* is not the first member of Apocynaceae to receive nuclear genome sequencing. Genomic sequencing and assembly of *Catharanthus roseus* (Rauvolfioideae) was performed by Kellner et al. (2015) to investigate the production of medicinal compounds (Table 1). Sabir et al. (2016) assembled the genome of *Rhazya stricta* (subfamily Rauvolfioideae) and Hoopes et al. (2017) assembled the *Calotropis gigantea* (Asclepiadoideae) genome, investigating alkaloid diversity and cardenolide production, respectively (Table 1).

The transcriptomes of several species of Apocynaceae have been released as part of broader investigations into medicinally important plants, particularly those producing monoterpene indole alkaloids, including *Tabernaemontana elegans* (Rauvolfioideae), *Rauvolfia serpentina* (Rauvolfioideae), *Rhazya stricta,* and *Catharanthus roseus* (Góngora-Castillo, Childs, et al., 2012; Medicinal Plant Consortium, 2011; Park et al., 2014; Xiao et al., 2013; Yates et al., 2014). The transcriptome of *Calotropis procera* has also been investigated (Hoopes et al., 2017; Kwon et al., 2015; Pandey et al., 2016).

Outside of Apocynaceae the most closely related species to milkweed with a sequenced genome is the diploid ancestor of coffee, *Coffea canephora* (Rubiaceae; Denoeud et al., 2014). *Coffea* is in the same order as *Asclepias*, Gentianales, and *C. canephora* has the same number of chromosomes: $x$=$n$=11, 2$n$=22 (Denoeud et al., 2014). The *Coffea* genome assembly is a high-quality reference, with large scaffolds ordered onto pseudochromosomes (scaffolds that have been ordered based on linkage information, as though on a chromosome; Table 1).

The genomic assembly of *Asclepias syriaca* presented here includes a nearly complete representation of gene space, supported by transcriptome evidence. The heterozygosity present in this obligate outcrossing species is used to develop a panel of SNPs that can be captured via targeted enrichment, and a set of offspring from the sequenced individual is used to cluster assembled scaffolds into linkage groups. A comparison of linkage groups between *Asclepias* and

*Coffea* is presented, providing insights into chromosome organization in *Asclepias*, and chromosomal evolution within Gentianales. Both genome and transcriptome sequences are used to explore gene family evolution, especially as related to cardenolide biosynthesis.

## METHODS

### *Tissue preparation and library construction*

Leaf tissue of *Asclepias syriaca* was sampled from a single individual at the Western Illinois University research farm, raised from seed from a wild population in McDonough County, Illinois (40.29622ºN, 90.89876ºW; Winthrop B. Phippen *s.n.*, OSC 226164, 226165). DNA was extracted from frozen tissue using the FastDNA Spin Kit from MPBiomedicals (Santa Ana, CA, USA) following manufacturer's protocols, modified by the addition of 40 µL 1% polyvinylpyrrolidone and 10 µL β-mercaptoethanol to the lysis solution prior to grinding.

Aliquots of isolated DNA were sheared with a BioRuptor sonicator (Diagenode Inc., Denville, NJ, USA). Two libraries were prepared following the Illumina protocol for paired-end libraries (Solexa, Inc, 2006). Ligated fragments were cut from agarose gels centered around 225 bp and 450 bp, and were amplified through 15 and 14 cycles, respectively, of polymerase chain reaction using Phusion High-Fidelity PCR Master Mix (New England BioLabs, Ipswich, MA, USA) and standard Illumina primers. Cleaned product was submitted for sequencing on an Illumina GAII Sequencer (Illumina Inc., San Diego, CA, USA) at the Center for Genome Research and Biocomputing (CGRB) at Oregon State University (Corvallis, OR, USA). One lane of the 450 bp library was sequenced with 80 bp paired-end reads, and 5 lanes of the 225 bp library were sequenced with 120 bp paired-end reads.

Frozen tissue was sent to GlobalBiologics, LLC (Columbia, MO, USA) for DNA extraction and production of mate-pair libraries using the Illumina Mate Pair Library v2 protocol with average insert sizes of 2750 bp and 3500 bp, and indexed with unique barcode sequences (Bioo Scientific, Austin, TX, USA). Purified DNA was provided to the CGRB for production of a mate-pair library using the Illumina Nextera protocol with an average insert size of 2000 bp. The 2000 bp library was sequenced on an Illumina MiSeq at the CGRB, one of 15 samples pooled on a lane, and sequenced with 76 bp paired-end reads. The 2750 bp library was sequenced on an Illumina HiSeq 2000 sequencer at the CGRB, one of three samples pooled on a lane, and sequenced with 101 bp paired-end reads. The 3500 bp library was sequenced on an Illumina MiSeq at Oregon Health and Science University (Portland, OR, USA) with 33 bp paired-end reads (Table 2).

### *Genomic read processing*

Pairs of reads properly mapping to the *Asclepias* chloroplast or mitochondria, with three or fewer mismatches between the target and query, were filtered out using Bowtie 2 v. 2.1.0 (scoring parameter “--score-min L,-6,0”), samtools v. 0.1.18, and bamtools v. 2.3.0 (Barnett et al., 2013; Langmead and Salzberg, 2012; Li et al., 2009). Portions of reads matching the Illumina adapter sequences were removed with Trimmomatic v. 0.30 and the “ILLUMINACLIP” option (Bolger et al., 2014). Duplicate read pairs from the same library were removed using the custom script fastq_collapse.py (Weitemier, 2014). Paired-end read pairs with sequences that overlapped by ≥7 bp sharing ≥90% identity were merged using the program FLASH v. 1.2.6 (parameters “-m 7 -M 80 -x 0.10”) (Magoč and Salzberg, 2011). The 3' and 5' ends of reads were then trimmed of any bases with a Phred quality score below 30, and any remaining reads less than 30 bp were removed using Trimmomatic.

Summary statistics were calculated using a k-mer distribution plot of reads from the 225 bp insert library after removing chloroplast and mitochondrial reads, but prior to joining with FLASH. K-mers of 17 bp were counted using BBTools script kmercountexact.sh, and estimates of genome size and heterozygosity were calculated using the program gce (Bushnell and Rood, 2015; Liu et al., 2013).

### *RNA-seq library preparation, sequencing, and assembly*

Total RNA was extracted from the individual used for genome sequencing from leaves and buds separately, by homogenizing approximately 200 mg of fresh frozen tissue on dry ice in a Fast-Prep-24 bead mill. Cold extraction buffer (1.5 mL of 3M LiCl/8M urea; 1% PVP K-60; 0.1M dithiothreitol; Tai et al., 2004) was added to the ground tissue. Tissue was then homogenized and cellular debris pelleted at 200×*g* for 10 minutes at 4°C. Supernatant was incubated at 4°C overnight. RNA was pelleted by centrifugation (20,000×*g* for 30 minutes at 4°C) and cleaned using a ZR Plant RNA MiniPrep kit (Zymo Research, Irvine, CA, USA). For each tissue type, an RNA-seq library was prepared using the Illumina RNA-Seq TruSeq kit v. 2.0 with the modifications of Parkhomchuk et al. (2009) to allow strand-specific sequencing by dUTP incorporation.

Libraries were sequenced on an Illumina HiSeq 2000 at the CGRB to yield 101 bp single-end reads. Before further analysis, reads that did not pass the Illumina chastity and purity filters were removed. Trimmomatic 0.20 (Bolger et al., 2014) was used to trim the final base of each read, leading and trailing bases with quality scores below Q20, and all following bases if a sliding window of 5 bp did not have an average quality of at least Q30. Reads shorter than 36 bp after trimming were excluded.

Transcripts were assembled de novo using Trinity (Release 2013-08-14) (Grabherr et al., 2011) for bud and leaf reads separately, as well as combined into a single data set using default settings, except for using a minimum contig length of 101 bp. The same settings were also used to assemble RNA-seq data from leaf tissue of the same *A. syriaca* individual from a library made using ribosomal RNA subtraction (Straub et al., 2013). Best scoring open reading frames (ORFs) were determined for each library using the TransDecoder utility provided with Trinity (Haas et al., 2013). Transcripts were annotated using Mercator (Lohse et al., 2014) and TRAPID (Van Bel et al., 2013).

### *Comparative transcriptome and gene family evolution analyses in Apocynaceae*

For a comparative analysis, transcriptomes were obtained for five other species of Apocynaceae. *Catharanthus roseus* and *Rauvolfia serpentina* transcriptomes were downloaded from the Medicinal Plant Genomics Resource project database (http://medicinalplantgenomics.msu.edu; Góngora-Castillo, Fedewa, et al., 2012), the *Rhazya stricta* (Yates et al., 2014) and *Calotropis procera* (Kwon et al., 2015) transcriptomes were downloaded from NCBI, and the *Tabernaemontana elegans* transcriptome was downloaded from the PhytoMetaSyn Project database (www.phytometasyn.ca; Xiao et al., 2013). All transcriptomes, including that of *A. syriaca*, were checked for duplicate transcripts, and the duplicates removed using the Dedupe tool in BBMap (Bushnell and Rood, 2015). Transcriptomes were checked for completeness using BUSCO v. 1.22 (Simão et al., 2015). Transcripts of all species were assigned to reference gene families using TRAPID. Reference gene family assignments were obtained from two high quality genomes, *Coffea canephora* (Denoeud et al., 2014) and *Vitis vinifera* (PLAZA v. 2.5; Proost et al., 2009).

A phylogenetic framework for comparative analysis was produced using published evolutionary relationships and divergence times in Apocynaceae (Fishbein et al., 2018). The timings of the *Coffea* split from Apocynaceae and the *Vitis* split from Gentianales were based on the estimates of Wikström et al. (2015). In order to examine changes in gene family sizes across Apocynaceae transcriptomes, BadiRate v. 1.35 (Librado et al., 2012) was run using the BDI (birth-death-innovation) stochastic model with a free rate (FR) branch model where each branch can have a different gene turn-over rate. Gains and losses were inferred using Wagner (ordered) parsimony (Kluge and Farris, 1969).

## *Genomic sequence assembly*

Processed read-pairs were assembled into contigs using Platanus v. 1.2.1 (Kajitani et al., 2014). Platanus is designed to assemble highly heterozygous diploid genomes, and initially uses several k-mer sizes during assembly. *Asclepias* reads were assembled with an initial k-mer size of 25 bp with a k-mer step increase of 10 bp up to a maximum k-mer of 110 bp. As part of the expectation for heterozygous assembly, Platanus can merge contigs sharing high identity. We allowed contigs sharing 85% identity to be merged (assembly parameters “-k 25 -u 0.15”).

Scaffolding was performed with Platanus, setting the paired-end reads as “inward pointing” reads and the mate-pair reads as “outward pointing” reads. Reads were mapped to scaffolds using an initial seed size of 21 bp, one link between contigs was sufficient to align them into a scaffold, and scaffolds sharing 85% identity could be merged (scaffolding parameters “-s 21 -l 1 -u 0.15”).

Gaps between scaffolds were closed via local alignment and assembly of reads around the gaps using Platanus. An initial seed size of 21 bp was used to include reads in the mapping around a gap, and a minimum overlap of 21 bp between the newly assembled filler contig and the edges of the scaffold was required to use that contig to fill the gap (gap close parameters “-s 21 -k 21 -vd 21 -vo 21”).

Transcripts were mapped to *Asclepias* scaffolds ≥1 kbp using BLAT v. 32x1; one or more transcripts spanning multiple scaffolds were used to merge those scaffolds (Kent, 2002). This was performed with the program Scubat (<https://github.com/elswob/SCUBAT> accessed 12/17/2015) modified so that scaffolds would not be clipped when joined by cap3 v. 02/10/15 (Elsworth, 2012; Huang and Madan, 1999; Tange, 2011).

### *Contaminant removal*

Merged scaffolds were compared against a genomic database of potentially contaminating organisms with the program DeconSeq standalone v. 0.4.3 (Schmieder and Edwards, 2011). Contaminant databases were downloaded from the DeconSeq website representing bacteria, archaea, viruses, 18S rRNA, zebrafish, mouse, and several human genomes (<http://deconseq.sourceforge.net> accessed January 20, 2016). Fungal genomes were obtained from the National Center for Biotechnology Information (NCBI) including *Alternaria arborescens* accession AIIC01, *Aspergillus fumigatus* AAHF01, *Bipolaris maydis* AIHU01, *Botrytis cinerea* assembly GCA_000832945.1, *Cladosprium sphaerospermum* AIIA02, *Fomitopsis pinicola* AEHC02, *Fusarium oxysporum* AAXH01, *Galerina marginata* AYUM01, *Hypoxylon sp.* JYCQ01, *Penicillium expansum* AYHP01, *Rhodotorula graminis* JTAO01, *Saccharomyces cerevisiae* assembly GCA_000146045.2, and *Trichoderma reesei* AAIL02

(Amselem et al., 2011; Firrincieli et al., 2015; Floudas et al., 2012; Goffeau et al., 1997; Hu et al., 2012; Li et al., 2015; Ma et al., 2010; Martinez et al., 2008; Ng et al., 2012; Nierman et al., 2005; Ohm et al., 2012; Riley et al., 2014; Shaw et al., 2015). The genome of *Solanum lycopersicum* (ITAG 2.4) was downloaded from the Sol Genomics Network (The Tomato Genome Consortium, 2012). The fungal and *Solanum* genomes were prepared as DeconSeq databases following the DeconSeq website, including filtering of repeated Ns, removal of duplicate sequences, and indexing with a custom version of BWA released with DeconSeq (Li and Durbin, 2010; <http://deconseq.sourceforge.net> accessed January 20, 2016).

Genomes obtained from the DeconSeq website and the fungal genomes were used as contaminant databases, the *Solanum* genome was used as a retain database. Scaffolds matching one of the contaminant genomes with ≥80% identity along ≥80% of the scaffold length were excluded as contaminants. Those scaffolds matching both a contaminating genome and the *Solanum* genome were retained.

## *Gene prediction and annotation*

A library of *Asclepias* repetitive elements was created following guidelines in the MAKER Genome Annotation Pipeline online documentation (Jiang, 2015). The program RepeatModeler v. open-1.0.8 was used to integrate the programs RepeatMasker v. open-4.0.5, rmblastn v. 2.2.28, RECON v. 1.08, Tandem Repeats Finder v. 4.07b, and RepeatScout v. 1.0.5 (Bao and Eddy, 2002; Benson, 1999; Price et al., 2005; Smit et al., 2015). Repeat models initially missing a repeat annotation were compared, using BLAT, against a library of class I and class II transposable elements acquired from the TESeeker website (Kennedy et al., 2010, 2011), and matching sequences provided an annotation. Remaining unannotated models were submitted to the online repeat analysis tool, CENSOR, and provided annotations with a score ≥400 and ≥50% sequence similarity (Kohany et al., 2006). A set of proteins from *Arabidopsis thaliana* was filtered to remove proteins from transposable elements, then compared using BLASTX against the *Asclepias* repeat models. The program ProtExcluder.pl v. 1.1 then used the BLASTX output to remove repeat models and flanking regions matching *Arabidopsis* proteins (Altschul et al., 1990; Jiang, 2015).

The set of scaffolds ≥1 kbp were annotated via the online annotation and curation tool GenSAS v. 4.0 (Humann et al., 2016; Lee et al., 2011), which was used to implement the following tools for repeat masking, transcript and protein mapping, ab initio gene prediction, gene consensus creation, and mapping of *Asclepias* predicted proteins:

1) Repeats in the assembled sequence were masked via RepeatMasker v. open-4.0.1 using the *Asclepias* repeat models and using models developed from dicots more broadly (Smit et al., 2015).

2) Multiple datasets were mapped onto *Asclepias* scaffolds in order to assist with gene prediction. The best ORFs from assembled *Asclepias* transcripts were mapped using both BLAT and BLAST (expect < 1e-50, 99% identity). Assembled transcripts from *Calotropis procera* were mapped with BLAT (Kwon et al., 2015). Proteins from *Coffea canephora* were mapped with BLASTX (e<0.0001; Denoeud et al., 2014). While additional high-quality genomes within Apocynaceae were later released (Hoopes et al., 2017; Sabir et al., 2016), they were not available at the time this work was performed.

3) Genes were predicted using the ab initio tools Augustus v. 3.1.0, SNAP, and PASA (Haas et al., 2003; Korf, 2004; Stanke et al., 2008). Augustus was run using gene models from

*Solanum*, finding genes on both strands, and allowing partial models; SNAP was run using models from *Arabidopsis thaliana*. PASA was informed by the best ORFs from assembled *Asclepias* transcripts.

4) Multiple lines of evidence were integrated into a gene consensus using EVidenceModeler (Haas et al., 2008) with the following weights: Augustus, 1; SNAP, 1; *Coffea* proteins, 5; *Asclepias* transcripts (BLAST), 7; *Asclepias* transcripts (BLAT), 7; *Calotropis* transcripts, 5; PASA, 7. Consensus gene models were then refined using PASA, again informed by *Asclepias* transcripts.

5) Predicted proteins were compared to the NCBI plant RefSeq database using BLASTP (expect < 1e-4, BLOSUM62 matrix; Pruitt et al., 2002), as well as being mapped against protein sequences from *Coffea* and *Catharanthus roseus* (expect < 1e-4; Denoeud et al., 2014; Kellner et al., 2015). Protein families were classified using the InterPro database and InterProScan v. 5.8-49.0 (Jones et al., 2014; Mitchell et al., 2015). Transfer RNAs were identified using tRNAscan-SE v. 1.3.1 (Lowe and Eddy, 1997). Additional open reading frames were found using the getorf tool from the EMBOSS suite, accepting a minimum of 30 bp (Rice et al., 2000).

Some predicted proteins were missing one or more exons, either because they were fragmented on the ends of scaffolds or, rarely, transcript evidence predicted exons with non-canonical splice sites. The predicted coding sequence produced by GenSas for some of these proteins was out of frame. In these cases the coding sequence was translated under all reading frames and a translation lacking internal stop codons was selected, if available.

An estimate of the completeness of the assembled gene space was calculated using the program BUSCO v. 1.22 and a set of 956 conserved single copy plant genes (Simão et al., 2015).

BUSCO was run independently on the set of coding sequences returned following gene prediction as well as on the assembled scaffolds ≥1 kbp using Augustus gene prediction with *Solanum* models. Predicted genes from *Asclepias*, *Catharanthus*, *Coffea*, and *Vitis* (obtained from the PLAZA 3.0 database) were clustered into orthogroups using OrthoFinder v. 0.7.1 (Emms and Kelly, 2015; Proost et al., 2015; The French-Italian Public Consortium for Grapevine Genome Characterization, 2007).

### *Gene analyses*

The P5βR region (PLAZA v. 2.5 gene family HOM000752) was identified in assembled scaffolds with BLAT (Kent, 2002), using the P5βR sequences from *Asclepias curassavica* (ADG56538; Bauer et al., 2010) and *Catharanthus roseus* (KJ873882-KJ873887; Munkert et al., 2015) as references. A maximum likelihood tree was constructed from peptide sequences of two *A. syriaca* regions with high identity to P5βR; six *Catharanthus* P5βR sequences; the *A. curassavica* sequence; P5βR sequences from *Calotropis procera* (Kwon et al., 2015), *C. gigantea* (Hoopes et al., 2017), and *Rhazya stricta* (Sabir et al., 2016); sequences from *Digitalis purpurea* and *D. lantata* (ACZ66261, AAS76634), representing P5βR2 and P5βR paralogs, respectively; and a sequence from *Picea sitchensis* (ABK24388). P5βR sequence alignments were performed using MUSCLE 3.8.425, as implemented in Geneious v. 11.1.5, with a maximum of 10 iterations (Edgar, 2004; Kearse et al., 2012). The optimal models of amino acid substitution, rate variation among sites, and equilibrium frequencies were inferred using the Akaike and Bayesian information criteria, as implemented in the online tool PhyML 3.0, which was also used to infer trees under those models and calculate aBayes support values (Anisimova et al., 2011; Guindon et al., 2010; Guindon and Gascuel, 2003).

### *SNP finding and targeted enrichment probe development*

The Platanus genome assembler uses a de Bruijn graph approach for contig assembly (Kajitani et al., 2014). Certain types of branches in this graph, known as "bubbles," may be caused by heterozygosity and are saved by the program for use in later assembly stages. Here, saved bubbles were filtered to identify those likely to represent heterozygous sites in low-copy regions of the genome.

The program CD-HIT-EST v. 4.5.4 was used to cluster any bubbles sharing ≥90% identity, which were removed, leaving only unique bubbles (Li and Godzik, 2006). Unique bubbles were mapped against the set of *Asclepias* scaffolds ≥1 kbp using BLAT at minimum identity thresholds of 90% and 95% (Kent, 2002). A set of 4000 SNP probes developed from a preliminary study using a similar approach, but from a different genome assembly, were mapped against the assembly presented here with a 90% identity threshold (Weitemier et al., 2014). One appropriate bubble from each scaffold <10 kbp, and up to two bubbles from scaffolds ≥10 kbp, were selected, up to a total of 20,000 bubbles. Bubbles mapping only once within the ≥90% identity mapping analysis were selected first, progressively adding bubbles that either mapped to ≤4 locations in the ≥90% identity mapping or mapped to ≤3 locations in the ≥95% identity mapping. Bubble sequences were trimmed to 80 bp, and centered around the SNP site where possible. Potential SNP probes were further analyzed by MYcroarray (now Arbor BioSciences, Ann Arbor, MI, USA) and excluded if they were predicted to anneal in a solution hybridization reaction to >10 locations within the *Asclepias* genome at 62.5-65°C or >2 locations above 65°C. Twenty thousand RNA oligos suitable for targeted enrichment, matching 17,684 scaffolds, were produced by MYcroarray.

### *Linkage mapping population*

Mature follicles were collected from the open pollinated plant that was the subject of genome sequencing. Approximately 100 seeds from six follicles collected from four stems of this plant (1, 3, 1, and 1 follicle per stem) were germinated and grown at Oklahoma State University. Due to the pollination system of *Asclepias*, seeds in a fruit are likely to be fertilized by a single pollen donor (Sparrow and Pearson, 1948; Wyatt and Broyles, 1990), meaning up to six paternal parents are represented among the 96 mapping offspring.

Seeds were surface sterilized in 5% bleach and soaked for 24 hr in distilled water. The testa was nicked opposite from the micropylar end and the seeds germinated on moist filter paper, in petri dishes, in the dark, at room temperature. Germination occurred within 4-7 days, and seedlings were planted into MetroMix 902 media in plug trays when radicles attained a length of 2-3 cm. Seedlings were again transplanted to 3-inch deep pots following the expansion to two sets of true leaves. Seedlings were grown under high intensity fluorescent lights in a controlled environment chamber at 14 hr daylength at approximately 27˚C. Plants were grown for approximately 90 days, harvested, and rinsed in distilled water, and frozen at -80˚C. DNA was extracted from roots, shoots, or a combination of roots and shoots using the FastDNA® kit (MP Biomedicals, Santa Ana, California) and Thermo Savant FastPrep® FP120 Cell Disrupter (Thermo Scientific, Waltham, MA, USA). DNA quantity and quality were visualized using agarose gel electrophoresis and quantified with a Qubit® fluorometer (Invitrogen, Carlsbad, CA, USA) and Quant-iT™ DNA-BR Assay Kit.

Ninety-six genomic DNA samples were diluted as necessary with ultrapure water to obtain approximately 3 μg in 100 ul and sheared on a Bioruptor UCD-200 (Diagenode) at low power for 12 cycles of 30 s on/30 s off. Several samples required sonication for 5-10 additional cycles to achieve a high concentration of fragments at the target size of 300-400 bp. Illumina-compatible, dual-indexed libraries were produced with the TruSeq® HT kit (Illumina), each with a unique barcode.

Barcoded libraries were pooled by equal DNA mass in three groups of 32 samples. These were enriched for targeted SNP regions using RNA oligos and following MYcroarray MYbaits protocol v. 3.00. Enriched pools were then themselves evenly pooled and sequenced with 150 bp paired-end reads on an Illumina HiSeq 3000 at the CGRB, producing 120.3 Mbp of sequence data (NCBI short read archive: SRX2163716-SRX2163811).

### *Linkage analyses*

Reads were processed using Trimmomatic v. 0.33 to remove adapter sequences, bases on the ends of reads with a Phred quality score below three, and clipping once a sliding window of 4 bp fell below an average quality score of 17 (Bolger et al., 2014). Processed reads for 90 samples (excluding 6 with low sequencing depth) were mapped onto the assembled scaffolds using bowtie2 with "sensitive" settings and a maximum fragment length of 600 bp (Langmead and Salzberg, 2012). Reads from the 225 bp insert library of the sequenced individual were also mapped back onto assembled scaffolds using the same settings. Mappings for all individuals and the parent were combined using samtools v. 0.1.16 and SNPs called using the bcftools "view" command (Li et al., 2009).

Two subsets of SNPs were retained. The first was a subset of SNPs where the maternal parent was heterozygous and the paternal parents for all offspring were homozygous for the same allele. The file containing all variants was converted to a format suitable for the R package

OneMap, using a custom perl script (Tennessen, 2015), retaining only sites heterozygous in one parent, the maternal sequenced individual. In this filtering the minor genotype abundance (either heterozygote or homozygote) needed to be at least 24 across 90 samples, loci could have up to 30% missing individuals, and alternative genotypes within individuals were ignored if their Phred probability score was 15 or above (i.e., of the three possible genotypes AA, Aa, aa, one should be most probable with a low Phred score and the other two less probable with Phred scores above 15).

The second subset retained SNPs from 22 full siblings (from the fruit producing the most offspring) for loci in which either the maternal or paternal parent, but not both, were heterozygous. Filtering in this set required a minor genotype abundance of at least five, loci could have up to four missing individuals, and genotypes with Phred probabilities of 20 or above were ignored (i.e., the final genotype calls are more certain because alternative genotypes are less likely).

SNP sets were clustered into linkage groups in R v. 3.2.2 using the package OneMap v. 2.0-4 (Margarido et al., 2007; R Core Team, 2014). One SNP from each scaffold was selected from SNPs among the full set of individuals, and were grouped using a logarithm of odds (LOD) threshold of 8.4. This clustered SNP loci into eleven clear groups, referred to here as the core linkage groups.

From the full-sibling set of SNPs, those held on the same scaffold and with identical genotypes across individuals (i.e., in perfect linkage) were grouped, and SNPs on different scaffolds in perfect linkage with no missing data were grouped. This was performed separately for loci where either the maternal or paternal parent was heterozygous. These loci were clustered into groups using LOD scores 6.1, 6.0, and 5.5. Each of these groupings produced hundreds of groups, but each contained about 22 groups that were substantially larger than the others.

A custom R script was used to combine the linkage group identity of scaffolds in the core linkage groups with scaffolds and groups in the sibling sets (Weitemier, 2017). For example, scaffold A could be assigned to a linkage group if it was in perfect linkage in the sibling set with scaffold B, and scaffold B was also present in the core linkage groups. If multiple scaffolds were perfectly linked, but associated with different core linkage groups, no unknown scaffolds would be assigned unless the most common core linkage group was three times as common as the next core group.

Linkage groupings in the sibling sets could be assigned to core linkage groups based on the membership of the scaffolds they contained. If the markers indicating that a sibling group should belong to a certain core linkage group were ten times as common as markers supporting a second most common assignment, then the sibling group was assigned to the core group, and all unknown scaffolds it contained also assigned to that group. (For example, sibling group A contains ten scaffolds known to be on core linkage group 1, one scaffold known to be on core linkage group 2, and one unknown scaffold; sibling group A is assigned to core linkage group 1 and the unknown scaffold is similarly assigned.)

This process was performed iteratively, progressively assigning scaffolds to core linkage groups. It was performed first with the sibling set grouped with LOD 6.1, then the grouping with LOD 6.0, finally the grouping with LOD 5.5.

### Synteny within Gentianales

Scaffolds found within the core linkage groups were mapped to *Coffea* coding sequences (BLASTN, expect < 1, best hit chosen) and mapped to their location on *Coffea* pseudochromosomes. Six *Asclepias* linkage groups had a roughly one-to-one correspondence with a *Coffea* pseudochromosome (e.g., most of the scaffolds from that linkage group, and few from other linkage groups, mapped to the pseudochromosome). From these six linkage groups one marker was selected for every 1 Mbp segment of the *Coffea* chromosome. Recombination fractions were measured among these loci using OneMap (retaining “safe” markers with THRES=5) and converted to cM using the Kosambi mapping function.

## RESULTS

### Sequencing and read processing

Paired-end sequencing produced 215.6 million pairs of reads representing 50.0 Gbp of sequence data, and mate-pair sequencing produced 52.8 million pairs of reads for 9.9 Gbp of sequence data. After read filtering and processing, 30.7 Gbp of paired-end sequence data remained along with 3.0 Gbp of mate-pair data. This represents total average sequence coverage of 80.4× on the 420 Mbp *Asclepias syriaca* genome (Table 2).

The distribution of 17 bp k-mers from the largest set of paired-end reads demonstrates a clear bi-modal distribution, with peaks at 43× and 84× depth (Fig. 1), corresponding to the sequencing depth of heterozygous and homozygous portions of the genome, respectively. This k-mer distribution provides a genome size estimate of 406 Mbp, and a site heterozygosity rate estimate of 0.056.

### Sequence assembly and gene annotation

The assembly of *Asclepias syriaca* contains 54,266 scaffolds ≥1 kbp, with N50 = 3415 bp, representing 37% of the estimated genome (156.6 Mbp of sequence plus 5.8 Mbp of gaps, Table 3). When including scaffolds ≥200 bp the assembly sums to 229.7 Mbp, with N50 = 1904 bp. The largest scaffold is 100 kbp, and 10% of the *Asclepias* genome, 42.82 Mbp, is held on 2343 scaffolds ≥10 kbp.

Within the 156.6 Mbp of scaffolds ≥1 kbp, 1.25 million putative open reading frames were identified, along with 193 transfer RNA loci. Assembled repeat elements made up about 75.7 Mbp. A total of 14,474 protein-coding genes were identified based on transcript evidence, closely related proteins, and ab initio models. These are predicted to produce 15,628 unique mRNAs, and are made up of a total of 87,496 exons with an average length of 225.3 bp. The median length of predicted proteins is 303 amino acids (mean = 402), which is shorter than lengths predicted in *Calotropis* (median = 367, mean = 448), similar to those predicted in *Coffea* (median = 334, mean = 402), but longer than those predicted in *Catharanthus* (median = 251, mean = 340; Fig. 2). Of the 14,474 predicted genes, 13,749 (95.0%) mapped to either *Coffea* or *Catharanthus* proteins, and 9811 mapped to RefSeq proteins.

Of 32,728 assembled *Asclepias* transcripts representing the best scoring ORFs, 31,654 (96.7%) mapped onto scaffolds ≥1 kbp. For *Calotropis*, 92,115 (88.4%) transcripts were mapped to *Asclepias* scaffolds, while 23,182 (90.6%) proteins from *Coffea* mapped to the assembly. BUSCO analysis of the genome assembly identified 818 of the 956 plant genes in its set, of which 209 were identified as duplicates. An additional 77 BUSCO genes were fragmented, meaning they were found in the genome assembly, but with a length outside two standard deviations of the mean BUSCO length for that gene. This represents a total of 895 BUSCO genes, or 93.6%. When applied to just the set of coding sequences BUSCO identified 742 complete genes (302 duplicated) and 84 fragmented genes, representing 86.4% of the conserved plant gene set. Apocynaceae transcriptomes were compared using the BUSCO set of 429 genes common to eukaryotes. The *Asclepias* transcriptome contained 365 of the genes (117 duplicated, 21 fragmented), representing 85.1%. Presence of these genes in other transcriptomes (*Catharanthus, Rauvolfia*, *Rhazya*, *Tabernaemontana*, *Calotropis*) ranged from 83.7% in *Calotropis* to 86.5% in *Tabernaemontana*, indicating that the *Asclepias* transcriptome assembly was of similar completeness to Apocynaceae transcriptomes publically available at the time of analysis. All Apocynaceae transcriptomes showed increased duplication of the 429 genes with approximately 2× the number of duplicates on average compared to the *Coffea*, *Catharanthus,* and *Vitis* genomes.

Among 100,114 predicted genes from *Asclepias*, *Catharanthus*, *Coffea*, and *Vitis*, 69.9% were clustered into 13,906 orthogroups. *Asclepias* had the highest percentage of genes placed in orthogroups, 81.6%, but those genes only represent 9837 orthogroups, the lowest of the four genomes. *Asclepias* shared the fewest orthogroups with other species (Table S1).

Comparison of all six Apocynaceae transcriptomes showed 5195 gene families were common to all. The *Asclepias* transcriptome contained 5762 gene families also present in the *Coffea* genome. There were 58 gene families with 1-3 gene copies in *Asclepias* that were not present in other Apocynaceae. Among Apocynaceae lineages, *Asclepias* was not unusual in its number of gene gains or losses. *Asclepias* had close to the median number of gene gains among all lineages with 5697 (median = 5791.5), much less than the 15,831 gene gains inferred in the lineage with the highest number of gains, *Rauvolfia*. Similarly, the number of gene losses in *Asclepias* at 905 was just below the median number of losses (median = 1136), and much less than the 7619 losses inferred for *Catharanthus*. While *Asclepias* had one of the highest gene birth rates over time (0.01082 events per gene per million years; Fig. 3), it was lower than that of close relative *Calotropis* (0.01463 events per gene per million years), and the rate inferred for the *Rauvolfia* plus *Catharanthus* plus *Tabernaemontana* lineage (0.14406 events per gene per million years) was an order of magnitude greater. *Asclepias* had close to the median value for gene death rate (0.00177 events per gene per million years). However, *Asclepias* had the second highest gene innovation rate (0.00069 events per gene per million years) compared to other lineages (Fig. 3). As with gene birth rate, the gene innovation rate of the *Rauvolfia* plus *Catharanthus* plus *Tabernaemontana* lineage (0.00146 events per gene per million years) was an order of magnitude higher.

### *Linkage mapping and synteny within Gentianales*

Following filtering, the set of all 96 offspring retained over 16,000 SNPs for which the maternal parent was heterozygous and all the paternal parents were homozygous for the same allele. These were located on 8495 scaffolds, covering 43.5 Mbp. Ninety of 96 individuals were sequenced at adequate depth to inform linkage group analyses. At a likelihood of odds (LOD) score of 8.4, 7809 scaffolds were clustered into 11 groups, the core linkage groups, representing 41.9 Mbp.

Filtering for SNPs among just the largest group of full-siblings, in which one parent (but not both) was heterozygous, found 83,854 SNPs on 18,333 scaffolds. These SNPs were consolidated by perfect linkage and then clustered at LOD scores of 6.1, 6.0, and 5.5. Combining scaffolds from the core linkage groups with those clustered among the full-sibling group ultimately provided a combined linkage set, with linkage group assignments to 16,285 scaffolds, representing 75.0 Mbp.

Mapping of scaffolds from just the core linkage groups to *Coffea* pseudochromosomes found several linkage group/pseudochromosome “best hit” pairs (e.g., most *Asclepias* scaffolds from a linkage group mapped to one pseudochromosome, while few scaffolds from other linkage groups mapped to that pseudochromosome). *Asclepias* linkage groups 2, 4, 6, 7, 8 and 9 mapped in this manner to *Coffea* pseudochromosomes 10, 8, 6, 11, 3, and 1, respectively (Figs. 4, 5). From these six linkage groups, SNPs were chosen mapping to every 1 Mbp region (if available) of the corresponding *Coffea* pseudochromosome. Recombination distances were measured among these markers and their relative positions within *Asclepias* plotted against their position in *Coffea* (Figs. S1-S6). Monotonically increasing or decreasing series of points in these plots represent loci in *Asclepias* and *Coffea* that maintain their relative positions. Several such marker clusters are seen in these plots (e.g., Fig. S2), though they tend to cover only short chromosomal regions and are often interrupted by markers from outside the cluster.

## ***Progesterone 5β-reductase gene family***

One region on linkage group 11 had 98.4% identity with peptide sequence from progesterone 5β-reductase from *Asclepias curassavica* (Table S2). This region was supported by *A. syriaca* transcriptome evidence, as well as mapped *Calotropis* transcripts and *Coffea* proteins. Approximately 500 bp downstream from this gene, a second region was identified sharing 52% amino acid identity with the first region, for 70% of its length. The second region lacks transcript evidence from *A. syriaca*, though portions of *Calotropis* transcripts and *Coffea* peptides map to it.

Gene predictions from Augustus and SNAP include potential exons within the region, and the region includes P5βR conserved motifs I, II, and III, and portions of motifs IV, V, and VI described by Thorn et al. (2008). It is interpreted here as a pseudogene of P5βR, ΨP5βR (Table S2).

Paralogs of P5βR have been described in other angiosperms including *Arabidopsis*, *Populus*, *Vitis*, and *Digitalis*, and the P5βR2 paralog occurs on a chromosome separate from that of P5βR1 in *Arabidopsis* and *Populus* (Bauer et al., 2010; Pérez-Bermúdez et al., 2010). Due to frame shifts and ambiguous exon boundaries in ΨP5βR, it is difficult to assess the correct peptide sequence it initially encoded, and therefore difficult to fully align with *Digitalis* P5βR1 and P5βR2 sequences. However, a few motifs, particularly a triple tryptophan at the N-terminal end of the sequence, suggest its origin from P5βR1, a conclusion supported by its position adjacent to the coding P5βR in *Asclepias*.

A third region on an unlinked scaffold exhibited moderate (37%) identity with the peptide sequence from linkage group 11 (Table S2). This region includes an intact reading frame and is matched by transcripts from *Calotropis*, though a lack of *Asclepias* transcripts matching this region indicates that it may not be regularly expressed within leaves or buds. A peptide alignment was made for this sequence, the known coding P5βR in *Asclepias*, and P5βR sequences from *A. curassavica, Calotropis procera, C. gigantea, Rhazya*, *Digitalis, Catharanthus*, and *Picea* to infer the phylogeny of this locus. The optimal model of sequence evolution selected by AIC was the LG+G+I model of peptide substitution, rate variation among sites, and proportion of invariable sites (BIC selected the LG+G model, but tree topologies were identical and are not shown). A maximum-likelihood estimate of the P5βR gene tree grouped the unlinked *Asclepias* sequence with a paralog from *Rhazya* (originating on supercontig 3 from Sabir et al., 2016) and *Catharanthus* paralog P5βR6 (Fig. 6). Together these are sister to all other P5βR sequences analyzed, except *Picea*, which was used to root the gene tree. The P5βR sequence from linkage group 11 is  strongly supported as the most cosely related sequence to the one from *A. curassavica*, within a clade including P5βR1 sequences from *Digitalis* and *Catharanthus*.

Analysis of the P5βR gene family across Apocynaceae showed that this gene family is largest in *Rauvolfia*, *Catharanthus*, and *Tabernaemontana*, with most of the expansion occurring in the common ancestor of these three (Fig. 3). However, this interpretation may change as more Apocynaceae genomes and transcriptomes become available.

## DISCUSSION

The *Asclepias syriaca* nuclear genome assembly presented here represents a large fraction of the protein-coding gene space, despite very high levels of heterozygosity and sequence data restricted to Illumina short reads. Gene space coverage is supported by high proportions of BUSCO plant core genes found within the assembly (93.6%) as well as assembled transcripts mapping to the assembly (96.7%).  A substantial portion of genes from related plant species mapped to the assembly as well, including 88.4% of transcripts from *Calotropis* and 90.6% of amino acid sequences from *Coffea*.

Overall, the *Asclepias* assembly is fragmented when compared to other plant genomes assembled using either long reads or deep sequencing of known contiguous fragments (e.g. BACs or fosmids). Assembly was also hindered by poor quality mate-pair libraries containing low proportions of properly paired fragments (Table 2). However, assembly results are typical for a sequencing project relying entirely on short reads, especially for organisms with high levels of heterozygosity. For example, the *Asclepias* N50 value of 3.4 kbp compares favorably to the assembly of the rubber tree, *Hevea brasiliensis*, genome (N50 = 2972 bp; Rahman et al., 2013), though it is not as contiguous as the dwarf birch, *Betula nana*, genome (N50 = 18.6 kbp; Wang et al., 2012), which incorporated several mate pair libraries. The assembly of the olive tree, *Olea europaea*, genome was also very similar to *Asclepias*, with N50 = 3.8 kbp prior to the inclusion of fosmid libraries (Cruz et al., 2016). See, however, the assembly of *Calotropis gigantea* using paired-end and mate-pair reads (N50 = 805 kbp, Table1; Hoopes et al., 2017) as and example of a less fragmented assembly. The effect of high heterozygosity is clearly seen in the comparison of *Asclepias* and *Catharanthus* assemblies (Kellner et al., 2015). While sequence data and genome assembly methods are similar between the two, *Asclepias* has an estimated heterozygosity rate of >1 SNP per 20 bp, whereas the heterozygosity rate in the inbred *Catharanthus* cultivar is estimated at <1 SNP per 1000 bp. This resulted in a N50 of 27.3 kbp assembled from only a single *Catharanthus* Illumina library (Table 1).

Functional annotations were applied to a high proportion (95.0%) of the 14,474 called genes, which were mapped to proteins from *Catharanthus roseus* and/or to *Coffea canephora*. The number of called genes is well below the typical value for plant genomes: the genome of *Calotropis gigantea*, the closest relative with an assembled genome, contains 19,536 gene loci (Hoopes et al., 2017). The genomes of *Rhazya* and *Catharanthus* contain 21,164 and 33,829 called genes, respectively (Kellner et al., 2015; Sabir et al., 2016). The genome of *Coffea* contains 25,574 protein-coding genes, and the genome of tomato, *Solanum lycopersicum*, from the sister order, Solanales, contains 36,148 (Denoeud et al., 2014; The Tomato Genome Consortium, 2012).

It is likely that the gene count in *Catharanthus* is an overestimate, a possibility in fragmented genome assemblies (Denton et al., 2014), as indicated by the excess of short predicted proteins relative to *Coffea* and *Calotropis* (Fig. 2). By contrast, the 14,474 called genes in *Asclepias* is likely an underestimate of the true number. While the size distribution of predicted *Asclepias* proteins is quite similar to that of *Coffea*, *Asclepias* contains fewer proteins of all sizes, and the dramatic reduction of orthogroups found in *Asclepias* relative to other species argues for deficiency in gene calling. While it's possible that similar genes were mistakenly collapsed into a single contig during the assembly stage meant to collapse alleles at a single locus, this should only occur with genes isolated on small contigs and should not affect the number of orthogroups identified. Nevertheless, the high proportion of matches between the *Asclepias* genome assembly, *Asclepias* transcripts, and gene sets from related organisms, indicates that the assembly likely does contain sequence information for nearly the full complement of genes, but that some of these have not been recognized by gene calling algorithms due to the fragmented nature of the assembly.

### *Synteny within Gentianales*

Six of eleven linkage groups in *Asclepias* show high synteny at a chromosomal scale with the pseudochromosomes of *Coffea* (Figs. 4, 5). This suggests that these chromosomes have remained largely stable and retained the same gene content for over 95 Myr, throughout the evolution of the Gentianales (Wikström et al., 2015). These stable chromosomes may have remained largely intact for a much longer period as well. The stable *Coffea* pseudochromosomes (1, 3, 6, 8, 10, and 11) retain largely the same content as inferred for ancestral core eudicot chromosomes, exhibiting little fractionation, even after an inferred genome triplication at the base of the eudicots, 117-125 Myr ago (see Fig. 1B in Denoeud et al., 2014; Jiao and Paterson, 2014).

Despite the conservation of gene content, gene order within stable chromosomes may be more labile. Plots of recombination distance among markers in *Asclepias* against physical distance in *Coffea* show several sets of markers in *Coffea* that retain their relative order in *Asclepias*, but are frequently interrupted by loci found elsewhere on the same *Coffea* pseudochromosome. For example, within *Asclepias* linkage group 2 there is a set of markers that retain their same relative ordering from positions 3 million to 8 million on *Coffea* pseudochromosome 10 (Fig. S1). However, these markers in *Asclepias* are interrupted by markers mapping to positions closer to the origin on the same *Coffea* pseudochromosome as well as a marker mapping to the far end. The most conserved synteny is between *Asclepias* linkage group 8 and *Coffea* pseudochromosome 3, which show complete synteny except for an apparent transposition of markers at positions 2 million and 7 million on *Coffea* pseudochromosome 3 (Fig. S2).

Contrasting the stability in gene content of six *Coffea* pseudochromosomes, pseudochromosome 2 is inferred to contain portions of at least five ancestral core eudicot chromosomes. This suggests significant fractionation in this chromosome since the eudicot triplication event (Denoeud et al., 2014). Even between *Coffea* and *Asclepias*, pseudochromosome 2 maps to portions of several *Asclepias* linkage groups (Figs. 4, 5). Therefore, the fractionation within this chromosome appears to have either occurred only within the branch leading from the Gentianales ancestor to *Coffea*, or occurred earlier and then continued along the branch leading to *Asclepias*. If the latter is true, then a higher frequency of rearrangement may be a characteristic of this chromosome within the Gentianales, relative to other chromosomes. Analyses of chromosomal rearrangements in *Rhazya* (Figure 1 in Sabir et al., 2016) support this view, suggesting several rearrangements between the core eudicot triplication event and the Gentianales ancestor, and continued rearrangement between that ancestor and *Rhazya*. However, mapped genomic resources within other Asterids outside of Gentianales are scarce, and are only found in taxa that have undergone additional genome duplication events since the eudicot triplication (e.g., *Solanum*, *Daucus*; Iorizzo et al., 2016; The Tomato Genome Consortium, 2012), complicating synteny assessments that might resolve the timing of fractionation of this chromosome.

The production of physical maps of both *Asclepias* and *Coffea* chromosomes will help resolve how frequently synteny has been disturbed between the two taxa. The ordered scaffold maps presented here (Figs. S1-S6) contain only a few dozen markers, and trends apparent now could be altered on maps with much greater resolution. The *Coffea* pseudochromosomes, meanwhile, are still ultimately ordered by recombination frequency, and about half of the scaffolds are placed with unknown orientation (Denoeud et al., 2014), which could manifest here as apparent transpositions among adjacent markers.

## *Progesterone 5β-reductase gene family*

The name *Asclepias* comes from the Greek god of medicine, Asclepius, whose name was applied to this genus for its potent secondary compounds. The cardenolides of *Asclepias* belong to a class of steroidal compounds, cardiac glycosides, used to treat cardiac insufficiency. While the genetic pathway that produces β-cardenolides (the form of cardenolide that includes the medicinal compound digitoxin) is largely unknown, one of the early steps involves the conversion of progesterone to 5β-pregnane-3,20-dione (Gärtner et al., 1990, 1994), catalyzed by the enzyme progesterone 5β-reductase (P5βR). Orthologs of P5βR occur broadly across seed plants, even in taxa that do not produce β-cardenolides, including *Asclepias*, which only produces α-cardenolides (Bauer et al., 2010). The P5βR1 locus has been characterized in *Asclepias curassavica*, but information about its genomic context has remained unknown.

A coding P5βR ortholog was located in *Asclepias syriaca* on linkage group 11, sharing 98.4% amino acid identity with P5βR from *A. curassavica*. This gene is supported by transcripts from *Asclepias*, as well as mapped transcripts from *Calotropis* and proteins from *Coffea*. The presence of a novel P5βR pseudogene was also identified closely downstream from the expressed gene (Table S2). Sharing high identity with the expressed P5βR, including several conserved motifs, it clearly originated from a P5βR duplication at some point. However, it is assumed to be non-functional due to its degraded exons interrupted by multiple stop codons and lack of expression evidence from the transcriptome.

A third region in *Asclepias*, on an unlinked scaffold, was matched by multiple P5βR sequences from *Catharanthus* (Table S2). This region is made up of a single open reading frame that shares only moderate identity with the *Asclepias* coding P5βR, and is not supported by *Asclepias* transcript evidence. In a P5βR phylogeny, the unlinked *Asclepias* region is sister to *Catharanthus* P5βR6 and a copy from *Rhazya* (Kellner et al., 2015; Sabir et al., 2016). These sequences together are sister to all other P5βR sequences analyzed except *Picea,* which was used for rooting (Fig. 6).

While at least two P5βR paralogs have been identified in a wide range of plants, and *Rhazya, Rauvolfia, Catharanthus,* and *Tabernaemontana* exhibit expression evidence of multiple paralogs, *Asclepias* is reduced for this group of genes. *Rauvolfia* and *Tabernaemontana* are known to produce cardenolides, but *Catharanthus* and *Rhazya* do not (Abere et al., 2014; Agrawal et al., 2012; Hoopes et al., 2017; Sivagnanam and Kumar, 2014). *Calotropis* is known to produce β-cardenolides (Bauer et al., 2010; Pandey et al., 2016), and contains two P5βR paralogs (Hoopes et al., 2017). It is possible that the fragmented nature of the current assembly precludes identification of all existing P5βR paralogs in *A. syriaca*, however, both genome assembly and transcript evidence point toward one functional P5βR locus. While multiple genes are involved in the production of β-cardenolides, it may be that the reduction in the P5βR family is responsible for the lack of these compounds in *Asclepias*, which only contains α-cardenolides.

## CONCLUSIONS

We present a draft genome assembly with linkage information of *Asclepias syriaca*, assigning nearly half of scaffolds to linkage groups. While the assembly remains fragmented, multiple lines of evidence indicate that nearly all of the gene space of *Asclepias* is represented within the assembly.

Linkage information allowed assessment of synteny across the order Gentianales. Six of eleven chromosomes retain similar gene content across the order, and these chromosomes have likely remained stable since the divergence of eudicots. One chromosome has either experienced dramatic fractionation since the divergence of Rubiaceae from other Gentianales, or experienced earlier fractionation that continued within Gentianales.

*Asclepias syriaca* and its relatives are important systems for a wide range of evolutionary and ecological studies, and are an important component of many ecosystems, serving as prolific nectar producers and as hosts to a range of specially adapted species. The availability of the

*Asclepias* genome, coupled with genomic data from symbiotic organisms, particularly insects, promises to inform important mechanisms of co-evolution (Agrawal and Fishbein, 2008; Edger et al., 2015; Zhan et al., 2011). We expect that the data presented here will advance these studies and aid the discovery of novel insights into the origin and evolution of a charismatic family, the production of important secondary compounds, and the ecological and evolutionary relationships between milkweeds and their communities.

## ACKNOWLEDGMENTS

The authors wish to kindly thank the following for important contributions to this work: Winthrop Phippen for cultivating *Asclepias*, supplying tissue for sequencing, and harvesting fruits. Nicole Nasholm, Matt Parks, LaRinda Holland, Zoe Austin, and Lisa Garrison for DNA extraction and library preparation. The Oregon State University Center for Genome Research and Biocomputing for expert sequencing facilities and computational infrastructure. Access to the TAIR database was provided under the Terms of Use, accessed on August 11, 2016, available at http://www.arabidopsis.org/doc/about/tair_terms_of_use/417.

Funding for this work is provided by the National Science Foundation Division of Environmental Biology awards 0919389 (to M.F.) and 0919583 (to R.C.C. and A.L.) and Integrative Organismal Systems award 1238731 (to C.D.B). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## DATA AVAILABILITY

The whole genome shotgun project and transcriptome shotgun assembly have been deposited at DDBJ/ENA/GenBank under the accessions MSXX01000000 and GFXT01000000, respectively. Additional data has been deposited in the Oregon State University institutional archive (Weitemier, 2017). A genome browser is available at www.milkweedgenome.org.

# REFERENCES

Abere TA, Ojogwu OK, Agoreyo FO, et al. (2014) Antisickling and toxicological evaluation of the leaves of *Rauwolfia vomitoria* Afzel (Apocynaceae). *Journal of Science and Practice of Pharmacy* 1(1): 11–15.

Agrawal AA (2005) Natural selection on common milkweed (*Asclepias syriaca*) by a community of specialized insect herbivores. *Evolutionary Ecology Research* 7: 651–667.

Agrawal AA and Fishbein M (2006) Plant defense syndromes. *Ecology* 87(7): S132–S149.

Agrawal AA and Fishbein M (2008) Phylogenetic escalation and decline of plant defense strategies. *Proceedings of the National Academy of Sciences of the United States of America* 105(29): 10057–10060.

Agrawal AA and Van Zandt PA (2003) Ecological play in the coevolutionary theatre: genetic and environmental determinants of attack by a specialist weevil on milkweed. *Journal of Ecology* 91(6): 1049–1059. DOI: 10.1046/j.1365-2745.2003.00831.x.

Agrawal AA, Petschenka G, Bingham RA, et al. (2012) Toxic cardenolides: chemical ecology and coevolution of specialized plant–herbivore interactions. *New Phytologist* 194(1): 28–45. DOI: 10.1111/j.1469-8137.2011.04049.x.

Altschul SF, Gish W, Miller W, et al. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215(3): 403–410. DOI: doi: 10.1016/S0022-2836(05)80360-2.

Amselem J, Cuomo CA, van Kan JAL, et al. (2011) Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genetics* 7(8): e1002230. DOI: 10.1371/journal.pgen.1002230.

Anisimova M, Gil M, Dufayard J-F, et al. (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology* 60(5): 685–699. DOI: 10.1093/sysbio/syr041.

Bai C, Alverson WS, Follansbee A, et al. (2012) New reports of nuclear DNA content for 407 vascular plant taxa from the United States. *Annals of Botany* 110(8): 1623–1629. DOI: 10.1093/aob/mcs222.

Bainard JD, Bainard LD, Henry TA, et al. (2012) A multivariate analysis of variation in genome size and endoreduplication in angiosperms reveals strong phylogenetic signal and association with phenotypic traits. *New Phytologist* 196(4): 1240–1250. DOI: 10.1111/j.1469-8137.2012.04370.x.

Bao Z and Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* 12(8): 1269–1276. DOI: 10.1101/gr.88502.

Barnett D, Garrison E, Marth G, et al. (2013) *Bamtools*. Available at: https://github.com/pezmaster31/bamtools (accessed 2 November 2013).

Bauer P, Munkert J, Brydziun M, et al. (2010) Highly conserved progesterone 5β-reductase genes (P5βR) from 5β-cardenolide-free and 5β-cardenolide-producing angiosperms. *Phytochemistry* 71(13): 1495–1505. DOI: 10.1016/j.phytochem.2010.06.004.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* 27(2): 573–580.

Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114–2120. DOI: 10.1093/bioinformatics/btu170.

Brower LP, Brower J van and Corvino JM (1967) Plant poisons in a terrestrial food chain. *Proceedings of the National Academy of Sciences of the United States of America* 57(4): 893–898.

Brower LP, McEvoy PB, Williamson KL, et al. (1972) Variation in cardiac glycoside content of Monarch butterflies from natural populations in eastern North America. *Science* 177(4047): 426. DOI: 10.1126/science.177.4047.426.

Broyles SB and Wyatt R (1990) Paternity analysis in a natural population of *Asclepias exaltata*: multiple paternity, functional gender, and the ‘pollen-donation hypothesis’. *Evolution* 44(6): 1454–1468. DOI: 10.2307/2409329.

Bushnell B and Rood J (2015) *BBTools*. Available at: https://sourceforge.net/projects/bbmap/ (accessed 1 December 2015).

Chaplin SJ and Walker JL (1982) Energetic constraints and adaptive significance of the floral display of a forest milkweed. *Ecology* 63(6): 1857–1870. DOI: 10.2307/1940126.

Cruz F, Julca I, Gómez-Garrido J, et al. (2016) Genome sequence of the olive tree, *Olea europaea*. *GigaScience* 5(1): 1–12. DOI: 10.1186/s13742-016-0134-5.

Denoeud F, Carretero-Paulet L, Dereeper A, et al. (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345(6201): 1181–1184. DOI: 10.1126/science.1255274.

Denton JF, Lugo-Martinez J, Tucker AE, et al. (2014) Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Computational Biology* 10(12): e1003998. DOI: 10.1371/journal.pcbi.1003998.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5): 1792–1797. DOI: 10.1093/nar/gkh340.

Edger PP, Heidel-Fischer HM, Bekaert M, et al. (2015) The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences of the United States of America* 112(27): 8362–8366. DOI: 10.1073/pnas.1503926112.

Elsworth B (2012) *Unearthing the genome of* Lumbricus rubellus. Ph.D. dissertation. The University of Edinburgh, Edinburgh, Scotland, U.K. Available at: https://www.era.lib.ed.ac.uk/bitstream/handle/1842/7596/Elsworth2013.pdf.

Emms DM and Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16(1): 1–14. DOI: 10.1186/s13059-015-0721-2.

Endress PK (2006) Angiosperm floral evolution: morphological developmental framework. In: *Advances in Botanical Research*. Academic Press, pp. 1–61.

Endress PK (2015) Development and evolution of extreme synorganization in angiosperm flowers and diversity: a comparison of Apocynaceae and Orchidaceae. *Annals of Botany*: mcv119. DOI: 10.1093/aob/mcv119.

Firrincieli A, Otillar R, Salamov A, et al. (2015) Genome sequence of the plant growth promoting endophytic yeast *Rhodotorula graminis* WP1. *Frontiers in Microbiology* 6: 978. DOI: 10.3389/fmicb.2015.00978.

Fishbein M and Venable DL (1996) Evolution of inflorescence design: Theory and data. *Evolution* 50(6): 2165–2177.

Fishbein M, Chuba D, Ellison C, et al. (2011) Phylogenetic relationships of *Asclepias* (Apocynaceae) inferred from non-coding chloroplast DNA sequences. *Systematic Botany* 36(4): 1008–1023. DOI: doi:10.1600/036364411X605010.

Fishbein M, Straub SCK, Boutte J, et al. (2018) Evolution at the tips: *Asclepias* phylogenomics and new perspectives on leaf surfaces. *American Journal of Botany* 105(3): 514–524. DOI: 10.1002/ajb2.1062.

Floudas D, Binder M, Riley R, et al. (2012) The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* 336(6089): 1715–1719. DOI: 10.1126/science.1221748.

Gärtner DE, Wendroth S and Seitz HU (1990) A stereospecific enzyme of the putative biosynthetic pathway of cardenolides. *FEBS Letters* 271(1–2): 239–242. DOI: 10.1016/0014-5793(90)80415-F.

Gärtner DE, Keilholz W and Seitz HU (1994) Purification, characterization and partial peptide microsequencing of progesterone 5β-reductase from shoot cultures of *Digitalis purpurea*.

*European Journal of Biochemistry* 225(3): 1125–1132. DOI: 10.1111/j.1432-1033.1994.1125b.x.

Goffeau A, Aert R, Agostini-Carbone ML, et al. (1997) The yeast genome directory. *Nature* 387(6632S): 5–5.

Góngora-Castillo E, Childs KL, Fedewa G, et al. (2012) Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PLoS ONE* 7(12): e52506. DOI: 10.1371/journal.pone.0052506.

Góngora-Castillo E, Fedewa G, Yeo Y, et al. (2012) Genomic approaches for interrogating the biochemistry of medicinal plant species. In: David A. Hopwood (ed.) *Methods in Enzymology*. Academic Press, pp. 139–159. Available at: http://www.sciencedirect.com/science/article/pii/B9780124046344000073.

Grabherr MG, Haas BJ, Yassour M, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7): 644–652. DOI: 10.1038/nbt.1883.

Guindon S and Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52(5): 696–704. DOI: 10.1080/10635150390235520.

Guindon S, Dufayard J-F, Lefort V, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59(3): 307–321. DOI: 10.1093/sysbio/syq010.

Haas BJ, Delcher AL, Mount SM, et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* 31(19): 5654–5666. DOI: 10.1093/nar/gkg770.

Haas BJ, Salzberg SL, Zhu W, et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* 9(1): 1–22. DOI: 10.1186/gb-2008-9-1-r7.

Haas BJ, Papanicolaou A, Yassour M, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8(8): 1494–1512.

Hoopes GM, Hamilton JP, Kim J, et al. (2017) Genome Assembly and Annotation of the Medicinal Plant <em>Calotropis gigantea</em>, a Producer of Anti-Cancer and Anti-Malarial Cardenolides. *G3: Genes|Genomes|Genetics*. DOI: 10.1534/g3.117.300331.

Hu J, Chen C, Peever T, et al. (2012) Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer. *BMC Genomics* 13: 171. DOI: 10.1186/1471-2164-13-171.

Huang X and Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research* 9(9): 868–877. DOI: 10.1101/gr.9.9.868.

Humann J, Ficklin SP, Lee T, et al. (2016) GenSAS v4.0: A web-based platform for structural and functional genome annotation and curation. In: *Plant and Animal Genome XXIV*, San Diego, California, 9 January 2016.

Iorizzo M, Ellison S, Senalik D, et al. (2016) A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics* 48: 657.

Jiang N (2015) Repeat Library Construction-Basic. Available at: http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Basic (accessed 7 October 2015).

Jiao Y and Paterson AH (2014) Polyploidy-associated genome modifications during land plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1648): 20130355. DOI: 10.1098/rstb.2013.0355.

Jones P, Binns D, Chang H-Y, et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9): 1236–1240. DOI: 10.1093/bioinformatics/btu031.

Kajitani R, Toshimoto K, Noguchi H, et al. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* 24(8): 1384–1395. DOI: 10.1101/gr.170720.113.

Kearse M, Moir R, Wilson A, et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12): 1647–1649. DOI: 10.1093/bioinformatics/bts199.

Kellner F, Kim J, Clavijo BJ, et al. (2015) Genome-guided investigation of plant natural product biosynthesis. *The Plant Journal* 82(4): 680–692. DOI: 10.1111/tpj.12827.

Kennedy RC, Unger MF, Christley S, et al. (2010) TESeeker. Available at: https://www3.nd.edu/~teseeker/download.html (accessed 30 September 2015).

Kennedy RC, Unger MF, Christley S, et al. (2011) An automated homology-based approach for identifying transposable elements. *BMC Bioinformatics* 12(1): 1–10. DOI: 10.1186/1471-2105-12-130.

Kent WJ (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Research* 12(4): 656–664. DOI: 10.1101/gr.229202.

Kluge AG and Farris JS (1969) Quantitative phyletics and the evolution of Anurans. *Systematic Biology* 18(1): 1–32. DOI: 10.1093/sysbio/18.1.1.

Kohany O, Gentles AJ, Hankus L, et al. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7(1): 1–7. DOI: 10.1186/1471-2105-7-474.

Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5(1): 1–9. DOI: 10.1186/1471-2105-5-59.

Kwon CW, Park K-M, Kang B-C, et al. (2015) Cysteine protease profiles of the medicinal plant *Calotropis procera* R. Br. revealed by de novo transcriptome analysis. *PLoS ONE* 10(3): e0119328. DOI: 10.1371/journal.pone.0119328.

La Rosa RJ and Conner JK (2017) Floral function: effects of traits on pollinators, male and female pollination success, and female fitness across three species of milkweeds (*Asclepias*). *American Journal of Botany* 104(1): 150–160. DOI: 10.3732/ajb.1600328.

Labeyrie E and Dobler S (2004) Molecular adaptation of *Chrysochus* leaf beetles to toxic compounds in their food plants. *Molecular Biology and Evolution* 21(2): 218–221. DOI: 10.1093/molbev/msg240.

Langmead B and Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4): 357–359. DOI: 10.1038/nmeth.1923.

Lee T, Peace C, Jung S, et al. (2011) GenSAS — An online integrated genome sequence annotation pipeline. *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)* 4: 1967–1973. DOI: 10.1109/BMEI.2011.6098712.

Li B, Zong Y, Du Z, et al. (2015) Genomic characterization reveals insights into patulin biosynthesis and pathogenicity in *Penicillium* species. *Molecular Plant-Microbe Interactions* 28(6): 635–647. DOI: 10.1094/MPMI-12-14-0398-FI.

Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5): 589–595. DOI: 10.1093/bioinformatics/btp698.

Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16): 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li W and Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13): 1658–1659. DOI: 10.1093/bioinformatics/btl158.

Librado P, Vieira FG and Rozas J (2012) BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* 28(2): 279–281. DOI: 10.1093/bioinformatics/btr623.

Liu B, Shi Y, Yuan J, et al. (2013) Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv:1308.2012v1 [q-bio]*. Available at: http://arxiv.org/abs/1308.2012 (accessed 12 August 2016).

Livshultz T, Kaltenegger E, Straub SCK, et al. (2018) Evolution of pyrrolizidine alkaloid biosynthesis in Apocynaceae: revisiting the defence de-escalation hypothesis. *New Phytologist* 218(2): 762–773. DOI: 10.1111/nph.15061.

Lohse M, Nagel A, Herter T, et al. (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell & Environment* 37(5): 1250–1258. DOI: 10.1111/pce.12231.

Lowe TM and Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25(5): 955–964. DOI: 10.1093/nar/25.5.0955.

Ma L-J, van der Does HC, Borkovich KA, et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* 464(7287): 367–373. DOI: 10.1038/nature08850.

Magoč T and Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21): 2957–2963. DOI: 10.1093/bioinformatics/btr507.

Margarido GRA, Souza AP and Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144(3): 78–79. DOI: 10.1111/j.2007.0018-0661.02000.x.

Martinez D, Berka RM, Henrissat B, et al. (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nature Biotechnology* 26(5): 553–560. DOI: 10.1038/nbt1403.

Medicinal Plant Consortium (2011) Release of the medicinal plant consortium transcriptome resources. Available at: http://medicinalplantgenomics.msu.edu/final_version_release_info.shtml (accessed 18 August 2016).

Mitchell A, Chang H-Y, Daugherty L, et al. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* 43(D1): D213–D221. DOI: 10.1093/nar/gku1243.

Morgan MT and Schoen DJ (1997) Selection on reproductive characters: floral morphology in *Asclepias syriaca*. *Heredity* 79(4): 433.

Munkert J, Pollier J, Miettinen K, et al. (2015) Iridoid synthase activity is common among the plant progesterone 5β-reductase family. *Molecular Plant* 8(1). Plant Metabolism and Synthetic Biology: 136–152. DOI: 10.1016/j.molp.2014.11.005.

Ng KP, Yew SM, Chan CL, et al. (2012) Sequencing of *Cladosporium sphaerospermum*, a Dematiaceous fungus isolated from blood culture. *Eukaryotic Cell* 11(5): 705–706. DOI: 10.1128/EC.00081-12.

Nierman WC, Pain A, Anderson MJ, et al. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* 438(7071): 1151–1156. DOI: 10.1038/nature04332.

Ohm RA, Feau N, Henrissat B, et al. (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS pathogens* 8(12): e1003037. DOI: 10.1371/journal.ppat.1003037.

Pandey A, Swarnkar V, Pandey T, et al. (2016) Transcriptome and Metabolite analysis reveal candidate genes of the cardiac glycoside biosynthetic pathway from *Calotropis procera*. *Scientific Reports* 6: 34464.

Park S, Ruhlman TA, Sabir JS, et al. (2014) Complete sequences of organelle genomes from the medicinal plant *Rhazya stricta* (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. *BMC Genomics* 15(1): 405. DOI: 10.1186/1471-2164-15-405.

Parkhomchuk D, Borodina T, Amstislavskiy V, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research* 37(18): e123–e123. DOI: 10.1093/nar/gkp596.

Pérez-Bermúdez P, Moya García AA, Tuñón I, et al. (2010) *Digitalis purpurea* P5βR2, encoding steroid 5β-reductase, is a novel defense-related gene involved in cardenolide biosynthesis. *New Phytologist* 185(3): 687–700. DOI: 10.1111/j.1469-8137.2009.03080.x.

Price AL, Jones NC and Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21(suppl 1): i351–i358. DOI: 10.1093/bioinformatics/bti1018.

Proost S, Bel MV, Sterck L, et al. (2009) PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants. *The Plant Cell Online* 21(12): 3718–3731. DOI: 10.1105/tpc.109.071506.

Proost S, Van Bel M, Vaneechoutte D, et al. (2015) PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research* 43(D1): D974–D981. DOI: 10.1093/nar/gku986.

Pruitt K, Brown G, Tatusova T, et al. (2002) The Reference Sequence (RefSeq) Database. In: McEntyre J and Ostell J (eds) *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information, p. Chapter 18. Available at: https://www.ncbi.nlm.nih.gov/books/NBK21091/.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org/.

Rahman AYA, Usharraj AO, Misra BB, et al. (2013) Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genomics* 14: 75. DOI: 10.1186/1471-2164-14-75.

Rasmann S, Agrawal AA, Cook SC, et al. (2009) Cardenolides, induced responses, and interactions between above- and belowground herbivores of milkweed (*Asclepias* spp.). *Ecology* 90(9): 2393–2404.

Rasmann S, Erwin AC, Halitschke R, et al. (2011) Direct and indirect root defences of milkweed (*Asclepias syriaca*): trophic cascades, trade-offs and novel methods for studying subterranean herbivory. *Journal of Ecology* 99(1): 16–25. DOI: 10.1111/j.1365-2745.2010.01713.x.

Rice P, Longden I and Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6): 276–277. DOI: 10.1016/S0168-9525(00)02024-2.

Riley R, Salamov AA, Brown DW, et al. (2014) Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. *Proceedings of the National Academy of Sciences of the United States of America* 111(27): 9923–9928. DOI: 10.1073/pnas.1400592111.

Sabir JSM, Jansen RK, Arasappan D, et al. (2016) The nuclear genome of *Rhazya stricta* and the evolution of alkaloid diversity in a medically relevant clade of Apocynaceae. *Scientific Reports* 6: 33782.

Schmieder R and Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6(3): e17288. DOI: 10.1371/journal.pone.0017288.

Shaw JJ, Berbasova T, Sasaki T, et al. (2015) Identification of a fungal 1,8-cineole synthase from *Hypoxylon* sp. with specificity determinants in common with the plant synthases. *The Journal of Biological Chemistry* 290(13): 8511–8526. DOI: 10.1074/jbc.M114.636159.

Simão FA, Waterhouse RM, Ioannidis P, et al. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19): 3210–3212. DOI: 10.1093/bioinformatics/btv351.

Sivagnanam S and Kumar A (2014) Preliminary phytochemical analysis of *Tabernaemontana alternifolia*. *International Journal of Pharma and Bio Sciences* 5(2): (B) 283-287.

Smit AFA, Hubley R and Green P (2015) RepeatMasker. Available at: http://www.repeatmasker.org/ (accessed 7 October 2015).

Solexa, Inc (2006) Protocol for whole genome sequencing using Solexa technology. *BioTechniques Protocol Guide 2007*: 291.

Sparrow FK and Pearson NL (1948) Pollen compatibility in *Asclepias syriaca*. *Journal of Agricultural Research* 77(6): 187–199.

Stanke M, Diekhans M, Baertsch R, et al. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24(5): 637–644. DOI: 10.1093/bioinformatics/btn013.

Straub SCK, Fishbein M, Livshultz T, et al. (2011) Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12(1): 211. DOI: 10.1186/1471-2164-12-211.

Straub SCK, Cronn RC, Edwards C, et al. (2013) Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biology and Evolution* 5(10): 1872–1885. DOI: 10.1093/gbe/evt140.

Tai HH, Pelletier C and Beardmore T (2004) Total RNA isolation from *Picea mariana* dry seed. *Plant Molecular Biology Reporter* 22(1): 93–93. DOI: 10.1007/BF02773357.

Tange O (2011) GNU Parallel: The command-line power tool. ;*login: The USENIX Magazine,* February.

Tennessen JA (2015) *MakeOnemapFormatFromVcfNoParentsTwoGenos.Pl*. Perl. Available at: https://github.com/jacobtennessen/MiSCVARS/blob/master/MakeOnemapFormatFromVcfNoParentsTwoGenos.pl (accessed 5 November 2015).

The French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161): 463–467. DOI: 10.1038/nature06148.

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400): 635–641. DOI: 10.1038/nature11119.

Thorn A, Egerer-Sieber C, Jäger CM, et al. (2008) The crystal structure of progesterone 5β-reductase from *Digitalis lanata* defines a novel class of short chain dehydrogenases/reductases. *Journal of Biological Chemistry* 283(25): 17260–17269. DOI: 10.1074/jbc.M706185200.

Van Bel M, Proost S, Van Neste C, et al. (2013) TRAPID: an efficient online tool for the functional and comparative analysis of de novoRNA-Seq transcriptomes. *Genome Biology* 14(12): R134.

Van Zandt PA and Agrawal AA (2004) Community-wide impacts of herbivore-induced plant responses in milkweed (*Asclepias syriaca*). *Ecology* 85(9): 2616–2629. DOI: 10.1890/03-0622.

Vaughan FA (1979) Effect of gross cardiac glycoside content of seeds of common milkweed, *Asclepias syriaca*, on cardiac glycoside uptake by the milkweed bug *Oncopeltus fasciatus*. *Journal of Chemical Ecology* 5(1): 89–100. DOI: 10.1007/BF00987690.

Wang N, Thomson M, Bodles WJA, et al. (2012) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology* 22(11): 3098–3111. DOI: 10.1111/mec.12131.

Weitemier K (2014) *Fastq_collapse.Py*. Python. Available at: https://github.com/kweitemier/fastq_collapse.

Weitemier K (2017) Supplemental data to ‘A draft genome and transcriptome of common milkweed (*Asclepias syriaca*) as resources for evolutionary, ecological, and molecular studies in milkweeds and Apocynaceae’. Oregon State University. DOI: 10.7267/N9M61HDR.

Weitemier K, Straub SCK, Cronn RC, et al. (2014) Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042. DOI: 10.3732/apps.1400042.

Wikström N, Kainulainen K, Razafimandimbison SG, et al. (2015) A Revised Time Tree of the Asterids: Establishing a Temporal Framework For Evolutionary Studies of the Coffee Family (Rubiaceae). *PLOS ONE* 10(5): e0126690. DOI: 10.1371/journal.pone.0126690.

Willson MF and Rathcke BJ (1974) Adaptive design of the floral display in *Asclepias syriaca* L. *The American Midland Naturalist* 92(1): 47–57. DOI: 10.2307/2424201.

Woodson RE (1954) The North American species of *Asclepias* L. *Annals of the Missouri Botanical Garden* 41(1): 1–211.

Wyatt R and Broyles SB (1990) Reproductive biology of milkweeds (*Asclepias*): Recent advances. In: Kawano S (ed.) *Biological Approaches and Evolutionary Trends in Plants*. San Diego, California: Academic Press, Inc., pp. 255–272.

Wyatt R and Broyles SB (1994) Ecology and evolution of reproduction in milkweeds. *Annual Review of Ecology and Systematics* 25: 423–441.

Xiao M, Zhang Ye, Chen X, et al. (2013) Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *Journal of Biotechnology* 166(3): 122–134. DOI: 10.1016/j.jbiotec.2013.04.004.

Yates SA, Chernukhin I, Alvarez-Fernandez R, et al. (2014) The temporal foliar transcriptome of the perennial C3 desert plant *Rhazya stricta* in its natural environment. *BMC Plant Biology* 14(1): 2.

Zhan S, Merlin C, Boore JL, et al. (2011) The Monarch butterfly genome yields insights into long-distance migration. *Cell* 147(5): 1171–1185. DOI: 10.1016/j.cell.2011.09.052.

**Figure 1:** K-mer distribution of *Asclepias syriaca* genomic reads.
Depth is the number of times a certain 17 bp k-mer occurred in the genomic reads, and count is the number of different k-mers at that depth. K-mers with depths below 15 or above 205 are not shown. Within the read set analyzed, 629 million k-mers were unique. Peaks occur at 43× and 84× depth.

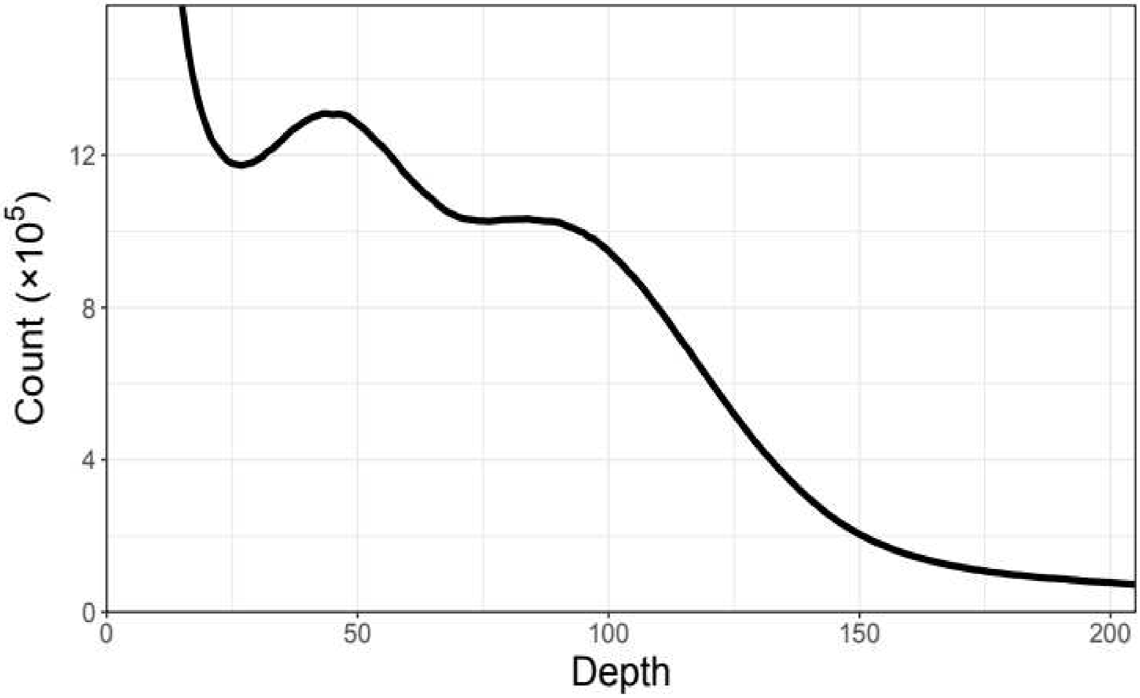

**Figure 2:** Peptide length histograms of *Asclepias*, *Calotropis*, *Coffea*, and *Catharanthus*. Mean and median peptide lengths are provided in the legend.

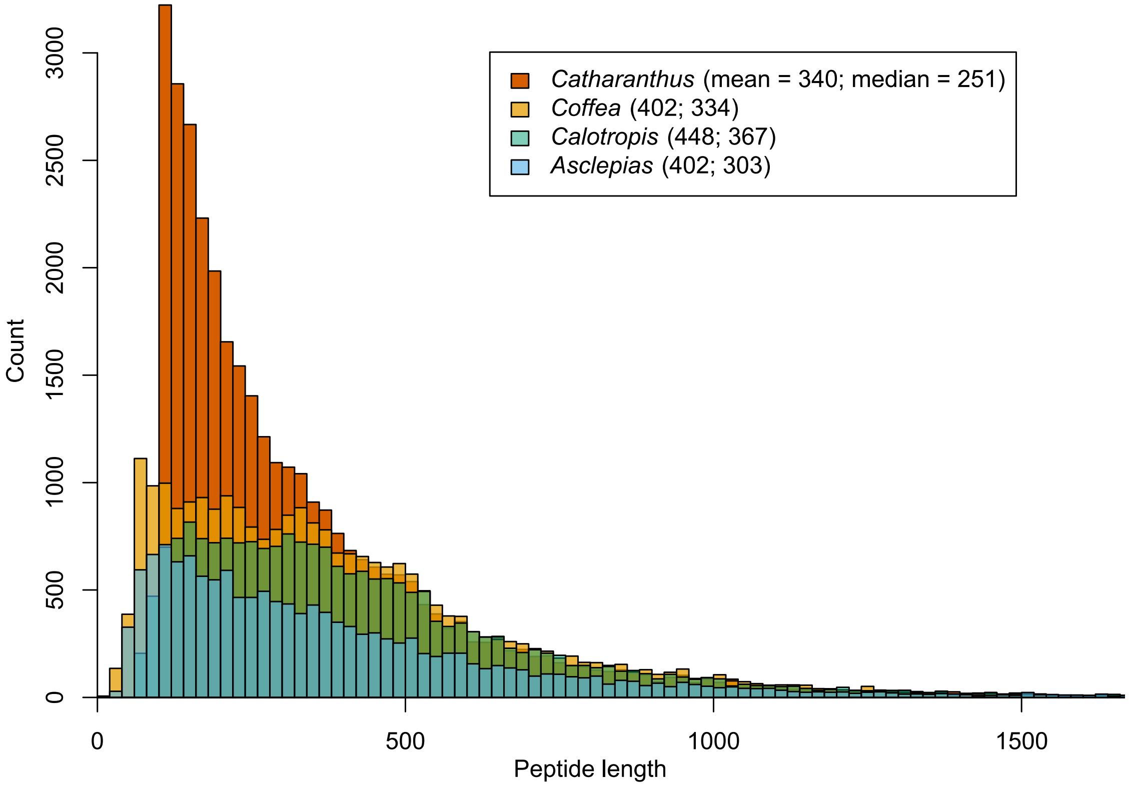

**Figure 3:** Gene family evolution in Apocynaceae inferred from transcriptomes.
The ultrametric tree depicts the phylogenetic relationships and estimated divergence times of sampled Apocynaceae and outgroups (*Coffea*, *Vitis*). The number of gene birth/death/innovation events per gene per million years across all gene families is shown above the branches. Numbers following tip labels represent the observed number of P5βR gene family paralogs, and the inferred number of paralogs present in common ancestors is shown to the right of nodes.

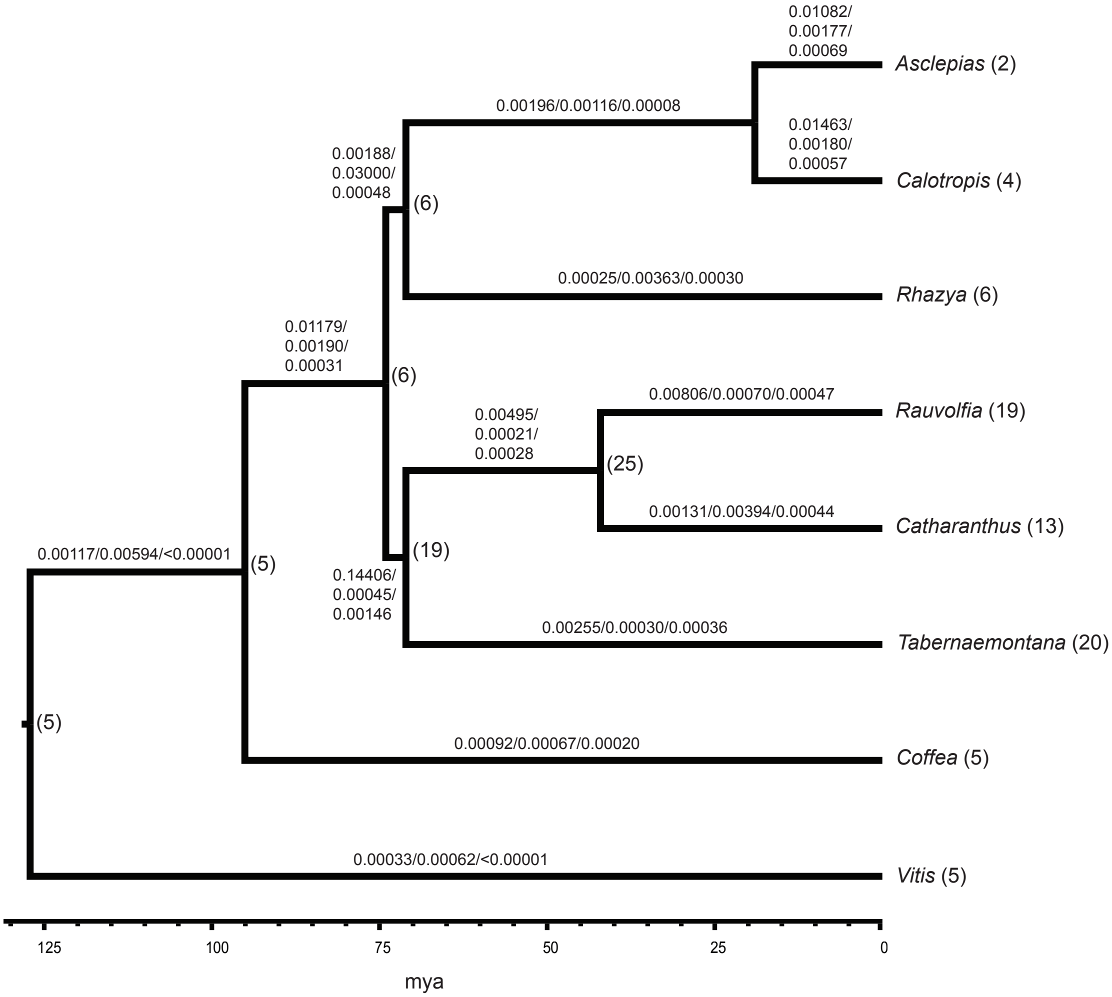

**Figure 4:** Counts of *Asclepias* linkage group scaffolds mapping to *Coffea* pseudochromosomes. Each column includes scaffolds from a single *Asclepias* linkage group, each row includes scaffolds mapping to a *Coffea canephora* pseudochromosome. *Coffea* chromosome 0 represents unassigned *Coffea* regions. Dot size is proportional to the number of mapping scaffolds, which is also provided.

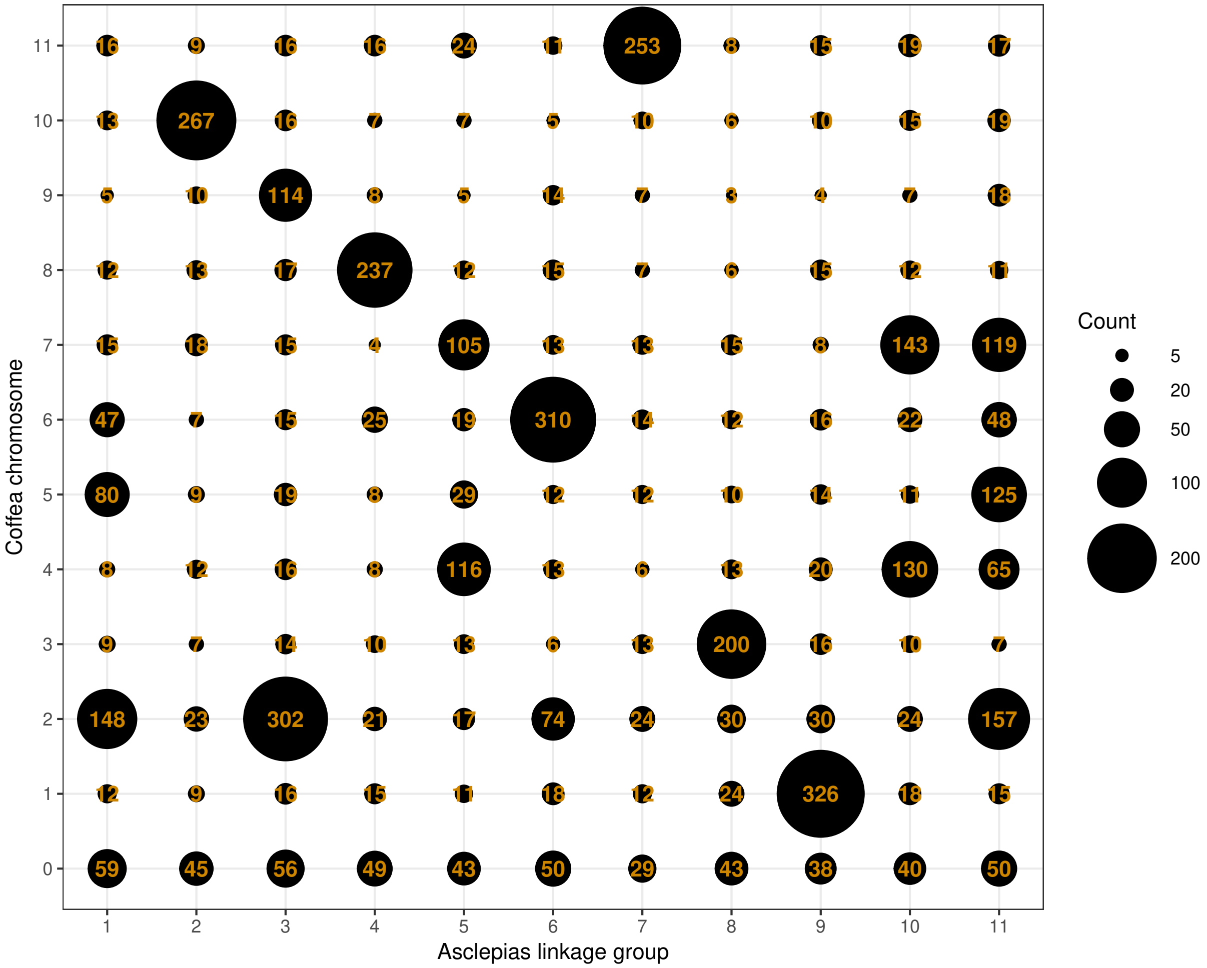

**Figure 5:** *Asclepias* linkage group scaffolds mapped to *Coffea* pseudochromosomes. *Coffea canephora* pseudochromosomes are shown in rows; the x-axis shows distance along each pseudochromosome. Each vertical bar represents one scaffold from the *Asclepias* core linkage groups, colored by its linkage group membership.

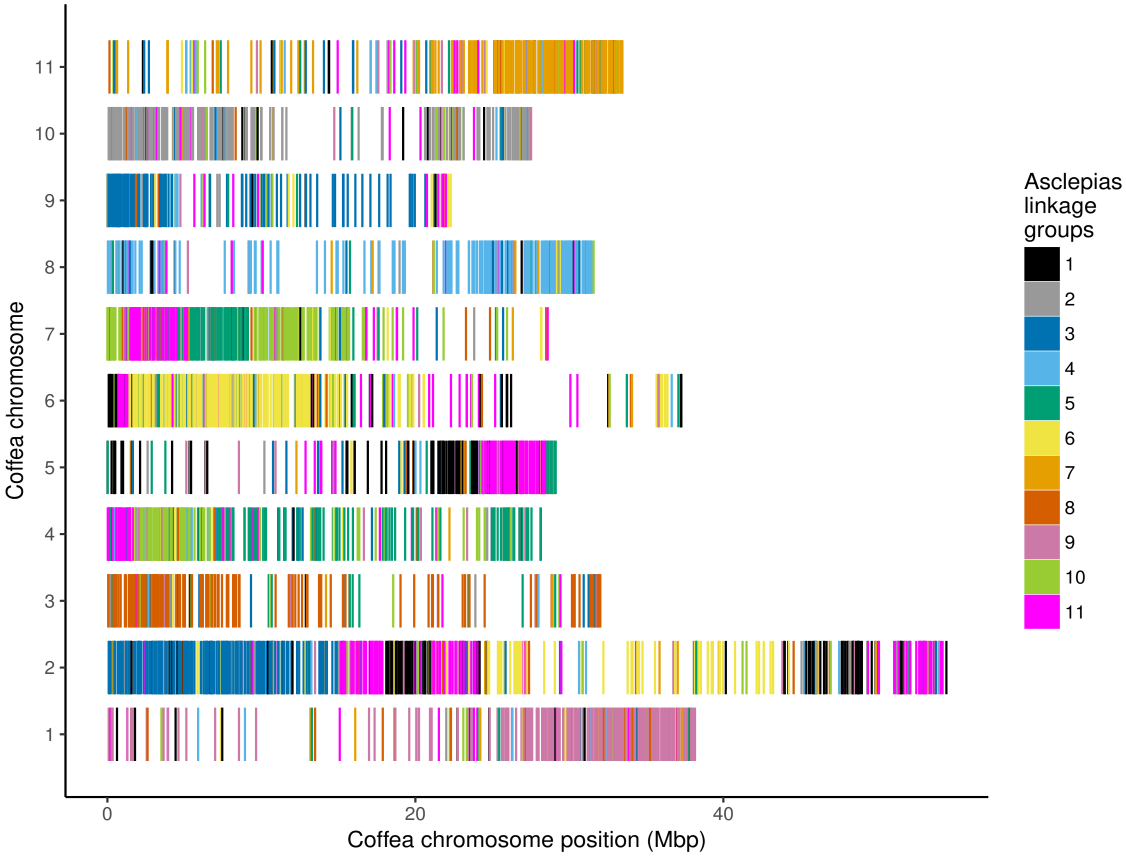

**Figure 6:** Maximum likelihood phylogeny of progesterone 5β-reductase paralogs. *A. syriaca* labels indicate the linkage group from which that sequence originates. *Catharanthus* and *Digitalis* labels indicate numbered paralogs isolated from that species. *Rhazya* labels indicate the originating supercontig from Sabir et al (2016) with two paralogs coming from supercontig 17. *Calotropis procera* labels indicate the originating transcript from Kwon et al (2015). *Calotropis gigantea* labels indicate the originating contig from Hoopes et al (2018). Numbers at nodes indicate aBayes support values. Branch lengths are in substitutions per site.

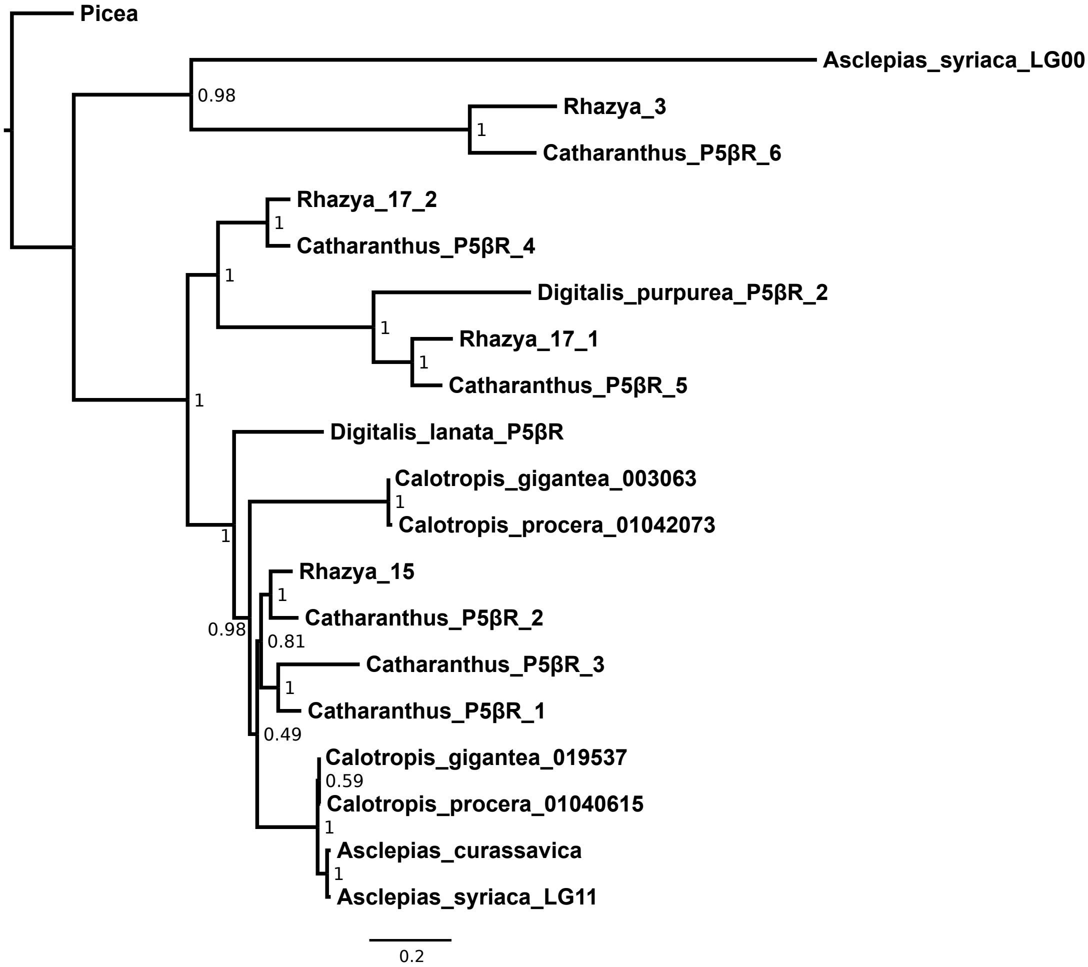

**Table 1:** Assembly comparison of *Asclepias*, *Catharanthus*, *Rhazya*, and *Coffea*. **Sequencing method** includes technologies and materials used in sequencing; N50 = 50% of the assembly is contained in scaffolds of this length or larger, BAC = bacterial artificial chromosome, SE = single-end, PE = paired-end.

| Species | Genome size (Mbp) | Assembly size (Mbp) | N50 (kbp) | # Scaffolds | Sequencing method |
|---|---|---|---|---|---|
| *Coffea canephora* | 710 | 568.6 | 1261 | 13,345 | 454 SE & mate-pair, Illumina SE & PE, BACs, haploid accession |
| *Rhazya stricta* | 200 | 274 | 5500 | 980 | Illumina PE & mate-pair, PacBio, optical mapping |
| *Catharanthus rosea* | 738 | 506 | 27.3 | 41,176 | Illumina PE, inbred accession |
| *Calotropis gigantea* | 225 | 157.3 | 805 | 1,536 | Illumina PE & mate-pair |
| *Asclepias syriaca* | 420 | 156.6 | 3.4 | 54,266 | Illumina PE & mate-pair |

**Table 2:** *Asclepias syriaca* sequencing summary.
**Machine:** Illumina instrument that performed the sequencing; **Raw yield, Processed yield:** Total Mbp of sequence data before and after read processing. **SRA:** NCBI Short Read Archive accession number.

| Library type | Insert size (bp) | Machine | Lanes | Read length (bp) | Clusters | Raw yield (Mbp) | Processed yield (Mbp) | SRA |
|---|---|---|---|---|---|---|---|---|
| Paired-end | 225 | GA II | 5 | 120 | 193,332,028 | 46400 | 29171 | SRX2164079 |
| Paired-end | 450 | GA II | 1 | 80 | 22,244,539 | 3559 | 1530 | SRX322144 |
| Mate-pair | 2000 | MiSeq | 1/15 | 76 | 257,750 | 39 | 34 | SRX2164126 |
| Mate-pair | 2750 | HiSeq 2000 | 1/3 | 101 | 46,704,483 | 9434 | 2819 | SRX322145 |
| Mate-pair | 3500 | MiSeq | 1 | 33 | 5,815,961 | 384 | 195 | SRX322148 |
| RNA-Seq Buds | -- | HiSeq 2000 | 1/4 | 101 | 48,085,747 | 4857 | 2812 | SRX2432900 |
| RNA-Seq Leaf | -- | HiSeq 2000 | 1/4 | 101 | 64,772,831 | 6542 | 3787 | SRX2435668 |
| | | | **Paired-end total** | | 215,576,567 | 49959 | 30701 | |
| | | | **Mate-pair total** | | 52,778,194 | 9857 | 3048 | |
| | | | **RNA-Seq total** | | 112,858,578 | 11399 | 6599 | |

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26488v2 | CC BY 4.0 Open Access | rec: 18 Apr 2019, publ: 18 Apr 2019

**Table 3:** *Asclepias syriaca* assembly statistics.
**Minimum scaffold:** The minimum scaffold size (bp) used for calculations. **Sum:** The sum of the lengths of all included scaffolds, not including gaps. **N80, N50, N20:** The length (bp) of the shortest scaffold in the set of largest scaffolds needed to equal or exceed (N/100)(Sum). **# scaffolds:** Total scaffolds ≥ the minimum size.

| Minimum scaffold | Sum (Mbp) | N80 | N50 | N20 | # scaffolds |
|---|---|---|---|---|---|
| 77 (all) | 265.9 | 317 | 1454 | 7080 | 508851 |
| 200 | 229.7 | 621 | 1904 | 8967 | 221940 |
| 1000 | 156.6 | 1633 | 3415 | 14019 | 54266 |
| 10000 | 42.82 | 12894 | 18998 | 30689 | 2343 |