The genome and transcriptome of common milkweed (*Asclepias syriaca*): resources for evolutionary, ecological, and molecular studies in milkweeds and Apocynaceae

Kevin A. Weitemier[1,6]

Shannon C. K. Straub[2]

Mark Fishbein[3]

C. Donovan Bailey[4]

Richard C. Cronn[5]

Aaron Liston[1]

[1]Department of Botany & Plant Pathology, Oregon State University, 2082 Cordley 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]Corresponding author:

Kevin Weitemier

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

Fax: 541-737-3573

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 the first nuclear genome and transcriptome assemblies of the common milkweed, *Asclepias syriaca*. This is the first species in Apocynaceae subfamily Asclepiadoideae with reconstructions of the nuclear, chloroplast, and mitochondrial genomes, and the first in the Apocynaceae to have linkage group information incorporated into the nuclear assembly. 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. The progesterone 5β-reductase gene family, a key component of cardenolide production, is likely reduced in *Asclepias* relative to other Apocynaceae. 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 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 the first genome assembly of a species of milkweed, plants that are the focus of diverse studies including floral development, pollination biology, plant-insect interactions and co-evolution, secondary metabolite chemistry, and rapid diversification.

The genus *Asclepias* sensu stricto is made up of about 130 species in North and South America (Fishbein et al., 2011). *Asclepias* 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. 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 & Fishbein, 2006, 2008, Rasmann et al., 2009, 2011), plant-herbivore

ecological interactions (Brower et al., 1967, 1972; Van Zandt & Agrawal, 2004; Vaughan, 1979), and plant-herbivore co-evolution (Agrawal, 2005; Agrawal & Van Zandt, 2003; Labeyrie & Dobler, 2004).

As members of Apocynaceae subfamily Asclepiadoideae, *Asclepias* species possess floral architectures unique among plants, including floral coronas and a central column composed of the unified stamens and pistil. Most *Asclepias* species are nearly or entirely self-incompatible (Wyatt & 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 & Pearson, 1948; Wyatt & Broyles, 1990). These features have positioned *Asclepias* as a model in studies of angiosperm reproductive biology (Broyles & Wyatt, 1990; Wyatt & Broyles, 1990, 1994), floral development (Endress, 2006, 2015), selection on floral characters and prezygotic reproductive isolation (La Rosa & Conner, 2017; Morgan & Schoen, 1997), and floral display evolution, (Chaplin & Walker, 1982; Fishbein & Venable, 1996; Willson & Rathcke, 1974).

Although this is the first genomic assembly of *Asclepias*, it is not the first genomic resource for the genus. The chloroplast and mitochondrial genomes of *Asclepias* have been sequenced (Straub et al., 2011, 2013), and flow cytometry estimates place the nuclear genome size at 420 Mbp (Bai et al., 2012; Bainard et al., 2012). *Asclepias* is not the first member of Apocynaceae to receive nuclear genome sequencing. 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. Genomic sequencing and assembly of *Catharanthus roseus* (Rauvolfioideae) was

performed by Kellner et al. (2015) to investigate the production of medicinal compounds (Table 1), and includes nearly all of the *Catharanthus* gene space on unlinked scaffolds.

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 *Rauvolfia serpentina* (Rauvolfioideae) and *Catharanthus roseus* (Góngora-Castillo et al., 2012; Medicinal Plant Consortium, 2011). 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 to have been sequenced 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: $x=n=11$, $2n=22$ (Denoeud et al., 2014). The *Coffea* genome assembly is a high-quality reference, with large scaffolds ordered onto pseudochromosomes (Table 1). Additionally, both the *Catharanthus* and *Coffea* gene models contain functional annotations.

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, the first such resource in Apocynaceae. 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 related to cardenolide biosynthesis.

METHODS

A summary of experimental methods is presented here; details for all methods are provided in the Supplementary Materials.

Tissue preparation, library construction, sequencing, and assembly

DNA was extracted from frozen *Asclepias syriaca* leaf tissue from an individual raised from seed from a wild population. Paired-end and mate-pair libraries with multiple insertion sizes were prepared and sequenced on Illumina platforms (Table 2). Following read filtering and cleanup, a distribution of 17 bp k-mers was used to calculate summary statistics, genome size, and heterozygosity values.

Total RNA was extracted from the sequenced individual from leaves and buds separately, and a strand-specific RNA-seq library prepared for each. Following Illumina sequencing, filtered reads were assembled de novo with Trinity (Grabherr et al., 2011) for combined bud and leaf reads. Best scoring open reading frames (ORFs) were determined from this assembly.

Genome assembly was performed using Platanus v. 1.2.1, a program designed for highly heterozygous diploid genomes (Kajitani et al., 2014). Platanus uses several k-mers, and because of the expectation for substantial heterozygosity, merges highly similar contigs and scaffolds. Transcripts were mapped to *Asclepias* scaffolds ≥1 kbp, and those scaffolds were merged where they were linked by one or more transcripts. Scaffolds matching a database of potentially contaminating organisms were removed.

Gene prediction and annotation

The set of scaffolds ≥1 kbp were annotated via the annotation and curation tool GenSAS v. 4.0 (Humann et al., 2016), which was used to: mask repeats from a custom *Asclepias* repeat library; map *Asclepias* ORFs, *Calotropis procera* transcripts, and *Coffea canephora* proteins onto assembled scaffolds; perform ab initio gene prediction and integrate this with transcript and protein mapping evidence to form a gene consensus; and map *Asclepias* predicted proteins against protein sequences from *Coffea*, *Catharanthus roseus*, and the NCBI plant RefSeq database (Pruitt et al., 2002). Completeness of the assembled gene space (predicted CDS and assembled scaffolds) was estimated using the program BUSCO v. 1.22 and a set of 956 conserved single-copy plant genes (Simão et al., 2015). Predicted genes from *Asclepias*, *Catharanthus*, *Coffea*, and grape (*Vitis vinifera*, Vitaceae) were clustered into orthogroups.

Asclepias transcripts, as well as transcripts from publicly available transcriptomes of other Apocynaceae (*Catharanthus roseus, Rauvolfia serpentina, Rhazya stricta*, and *Tabernaemontana elegans*) and two outgroups (*Coffea canephora* and *Vitis vinifera*) were assigned to gene family. Changes in gene family sizes, including the progesterone 5β-reductase family (below), were estimated using a birth-death-innovation stochastic model based on phylogenetic information from whole plastomes.

Progesterone 5β-reductase gene family

One of the key genes involved early in the cardenolide synthesis pathway, progesterone 5β-reductase (P5βR), was identified in assembled scaffolds based on published sequences from *Asclepias curassavica* and *Catharanthus roseus* (Bauer et al., 2010; Munkert et al., 2015). A maximum likelihood tree was constructed from peptide sequences of two *A. syriaca* regions with

high identity to P5βR, as well as paralogs from *A. curassavica*, *Catharanthus*, *Digitalis* (Lamiaceae), and *Picea* (Pinaceae). This sampling represents a subset of the analysis performed by Bauer et al. (2010), with the addition of *Catharanthus* sequences to test the orthology of *Asclepias* P5βR homologs.

SNP finding, targeted enrichment, and linkage analyses

Low-copy heterozygous sites were identified from the genome assembly, and 20,000 RNA oligos targeting 17,684 scaffolds were produced for target enrichment.

Seeds from six fruits were collected and germinated from the open pollinated plant that was the subject of genome sequencing. Due to the pollination system of *Asclepias*, seeds in a fruit are likely fertilized by a single pollen donor (Sparrow & Pearson, 1948; Wyatt & Broyles, 1990), meaning up to six paternal parents are represented among the 96 mapping offspring. Following DNA extraction, barcoded libraries were prepared, enriched for targeted SNP regions, pooled, and sequenced on an Illumina HiSeq 3000.

Processed reads from 90 offspring (excluding 6 with low sequencing depth) and reads from the sequenced individual were mapped onto assembled scaffolds, and SNPs called for each individual. Two subsets of SNPs were retained. The first retained SNPs where the maternal parent was heterozygous and the paternal parents for all offspring were homozygous for the same allele. 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.

SNPs from the full set of individuals were used to cluster matching scaffolds into 11 core linkage groups. The SNPs from the full-sibling set, mapping to many more scaffolds, were

clustered into linkage groups and assigned to the core linkage groups via shared SNPs between the subsets.

Scaffolds on the core linkage groups were matched to *Coffea* coding sequences 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* pseudochromosome, and recombination fractions were measured among these loci.

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 draft assembly of *Asclepias syriaca* contains 54,266 scaffolds ≥1 kbp, with N50 = 3415 bp, representing 37% (156.6 Mbp) of the estimated genome (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 similar to lengths 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 just coding sequences identified 742 complete genes (302 of which were duplicated), and an additional 84 fragmented genes, from the set of 956 plant genes, representing 86.4%. When applied to the entire genome assembly BUSCO identified 818 complete genes (209 duplicated) and 77 fragmented genes, representing 93.6% 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 (165 duplicated, 24 fragmented), representing 85.1%. Presence of these genes in other transcriptomes ranged from 80.2% in *Tabernaemontana* to 86.7% in *Rauvolfia*, indicating that the *Asclepias* transcriptome assembly was of similar completeness to the Apocynaceae transcriptomes publically available at the time of analysis. All Apocynaceae transcriptomes except *Catharanthus* showed increased duplication of the 429 genes with 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 five Apocynaceae transcriptomes showed 5302 gene families were common to all. The *Asclepias* transcriptome contained 5762 gene families also present in the *Coffea* genome. *Asclepias* had the highest number of gene gains among all lineages with 31,374, nearly double the 16,907 gene gains observed in the lineage with the second highest number of gains, the *Rauvolfia* plus *Catharanthus* lineage. Although *Asclepias* had the largest number of gene gains, it did not have the highest gene birth rate over time (0.01314 events per gene per million years; Fig. 3), which was observed for the *Rauvolfia* plus *Catharanthus* plus *Tabernaemontana* lineage (0.08652 events per gene per million years). *Asclepias* had close to the median value for gene death rate, and one of the lowest innovation rates compared to other lineages (Fig. 3).

Linkage mapping and synteny within Gentianales

Filtering SNPs from the set of all 96 offspring retained over 16,000 SNPs where 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 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 held 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 high 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 it's 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*, *Digitalis*, *Catharanthus*, and *Picea*. The optimal model of sequence evolution selected by AIC was the LG+G+F model of peptide substitution, rate variation among sites, and estimation of equilibrium frequencies (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 *Catharanthus* paralog P5βR6 (Fig. 6). The sequence from the unlinked scaffold and *Catharanthus* P5βR6 together are sister to all other P5βR sequences analyzed, except *Picea* (when the *Picea* P5βR is placed at the root). The P5βR sequence from linkage group 11 groups strongly with the sequence from *A. curassavica*, within a clade including P5βR1 sequences from both *Digitalis* and *Catharanthus*.

Analysis of the P5βR gene family across Apocynaceae showed that this gene family is much more diverse in *Rauvolfia*, *Catharanthus*, and *Tabernaemontana*, with most of the expansion of the gene family occurring in the common ancestor of these three (Fig. 3). However, this interpretation may change as more Apocynaceae 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 more distantly related organisms 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. 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). 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, and the heterozygosity rate in the sequenced 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 *Catharanthus*, the closest relative with an assembled genome, contains 33,829 called genes (Kellner et al., 2015). 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* (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 collapsed into a single contig during the assembly stage that was meant to only 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 assembly's fragmented nature.

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), since the divergence between *Coffea* (Rubiaceae) and *Asclepias* (Apocynaceae) occurred at the base of the Gentianales (Backlund et al., 2000). 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 & 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 M to 8 M 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 M and 7 M 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 when fractionation occurred within 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, which also is made up of a single exon (Kellner et al., 2015). These two sequences together are sister to all *Asclepias*, *Catharanthus*, and *Digitalis* P5βR sequences analyzed (Fig. 6).

While at least two P5βR paralogs have been identified in a wide range of plants, and *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* does not (Abere et al., 2014; Agrawal et al., 2012; Sivagnanam & Kumar, 2014). *Calotropis* is known to produce β-cardenolides (Bauer et al., 2010; Pandey et al., 2016), and different transcripts from *Calotropis* map to all P5βR regions in *Asclepias*. 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

This study represents the first draft genome assembly with linkage information in Apocynaceae, and the second among the >20,000 species of Gentianales, 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 & 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, available at https://ir.library.oregonstate.edu/concern/datasets/vd66w525h. A genome browser is available at www.milkweedgenome.org.

REFERENCES

Abere, T. A., Ojogwu, O. K., Agoreyo, F. O., & Eze, G. I. (2014). Antisickling and toxicological evaluation of the leaves of *Rauwolfia vomitoria* Afzel (Apocynaceae). *Journal of Science and Practice of Pharmacy*, *1*, 11–15.

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

Agrawal, A. A., & Fishbein, M. (2006). Plant defense syndromes. *Ecology*, *87*, S132–S149.

Agrawal, A. A., & 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*, 10057–10060.

Agrawal, A. A., Petschenka, G., Bingham, R. A., Weber, M. G., & Rasmann, S. (2012). Toxic cardenolides: chemical ecology and coevolution of specialized plant–herbivore interactions. *New Phytologist*, *194*, 28–45. https://doi.org/10.1111/j.1469-8137.2011.04049.x

Agrawal, A. A., & Van Zandt, P. A. (2003). Ecological play in the coevolutionary theatre: genetic and environmental determinants of attack by a specialist weevil on milkweed. *Journal of Ecology*, *91*, 1049–1059. https://doi.org/10.1046/j.1365-2745.2003.00831.x

Backlund, M., Oxelman, B., & Bremer, B. (2000). Phylogenetic relationships within the Gentianales based on *ndhF* and *rbcL* sequences, with particular reference to the Loganiaceae. *American Journal of Botany*, *87*, 1029–1043.

Bai, C., Alverson, W. S., Follansbee, A., & Waller, D. M. (2012). New reports of nuclear DNA content for 407 vascular plant taxa from the United States. *Annals of Botany*, *110*, 1623–1629. https://doi.org/10.1093/aob/mcs222

Bainard, J. D., Bainard, L. D., Henry, T. A., Fazekas, A. J., & Newmaster, S. G. (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*, 1240–1250. https://doi.org/10.1111/j.1469-8137.2012.04370.x

Bauer, P., Munkert, J., Brydziun, M., Burda, E., Müller-Uri, F., Gröger, H., … Kreis, W. (2010). Highly conserved progesterone 5β-reductase genes (P5βR) from 5β-cardenolide-free and 5β-cardenolide-producing angiosperms. *Phytochemistry*, *71*, 1495–1505. https://doi.org/10.1016/j.phytochem.2010.06.004

Brower, L. P., Brower, J. van, & Corvino, J. M. (1967). Plant poisons in a terrestrial food chain. *Proceedings of the National Academy of Sciences of the United States of America*, *57*, 893–898.

Brower, L. P., McEvoy, P. B., Williamson, K. L., & Flannery, M. A. (1972). Variation in cardiac glycoside content of Monarch butterflies from natural populations in eastern North America. *Science*, *177*, 426. https://doi.org/10.1126/science.177.4047.426

Broyles, S. B., & Wyatt, R. (1990). Paternity analysis in a natural population of *Asclepias exaltata*: multiple paternity, functional gender, and the “pollen-donation hypothesis.” *Evolution*, *44*, 1454–1468. https://doi.org/10.2307/2409329

Chaplin, S. J., & Walker, J. L. (1982). Energetic constraints and adaptive significance of the floral display of a forest milkweed. *Ecology*, *63*, 1857–1870. https://doi.org/10.2307/1940126

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., … Gabaldón, T. (2016). Genome sequence of the olive tree, *Olea europaea*. *GigaScience*, *5*, 1–12. https://doi.org/10.1186/s13742-016-0134-5

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., … Lashermes, P. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, *345*, 1181–1184. https://doi.org/10.1126/science.1255274

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Computational Biology*, *10*, e1003998. https://doi.org/10.1371/journal.pcbi.1003998

Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., … Wheat, C. W. (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*, 8362–8366. https://doi.org/10.1073/pnas.1503926112

Endress, P. K. (2006). Angiosperm floral evolution: morphological developmental framework. In *Advances in Botanical Research* (Vol. Volume 44, pp. 1–61). Academic Press.

Endress, P. K. (2015). Development and evolution of extreme synorganization in angiosperm flowers and diversity: a comparison of Apocynaceae and Orchidaceae. *Annals of Botany*, mcv119. https://doi.org/10.1093/aob/mcv119

Fishbein, M., Chuba, D., Ellison, C., Mason-Gamer, R. J., & Lynch, S. P. (2011). Phylogenetic relationships of *Asclepias* (Apocynaceae) inferred from non-coding chloroplast DNA sequences. *Systematic Botany*, *36*, 1008–1023. https://doi.org/doi:10.1600/036364411X605010

Fishbein, M., & Venable, D. L. (1996). Evolution of inflorescence design: Theory and data. *Evolution*, *50*, 2165–2177.

Gärtner, D. E., Keilholz, W., & Seitz, H. U. (1994). Purification, characterization and partial peptide microsequencing of progesterone 5β-reductase from shoot cultures of *Digitalis purpurea*. *European Journal of Biochemistry*, *225*, 1125–1132. https://doi.org/10.1111/j.1432-1033.1994.1125b.x

Gärtner, D. E., Wendroth, S., & Seitz, H. U. (1990). A stereospecific enzyme of the putative biosynthetic pathway of cardenolides. *FEBS Letters*, *271*, 239–242. https://doi.org/10.1016/0014-5793(90)80415-F

Góngora-Castillo, E., Childs, K. L., Fedewa, G., Hamilton, J. P., Liscombe, D. K., Magallanes-Lundback, M., … Buell, C. R. (2012). Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PLoS ONE*, *7*, e52506. https://doi.org/10.1371/journal.pone.0052506

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*, 644–652. https://doi.org/10.1038/nbt.1883

Hoopes, G. M., Hamilton, J. P., Kim, J., Zhao, D., Wiegert-Rininger, K., Crisovan, E., & Buell, C. R. (2017). Genome Assembly and Annotation of the Medicinal Plant Calotropis gigantea, a Producer of Anti-Cancer and Anti-Malarial Cardenolides. *G3: Genes|Genomes|Genetics*. https://doi.org/10.1534/g3.117.300331

Humann, J., Ficklin, S. P., Lee, T., Cheng, C.-H., Jung, S., Wegrzyn, J., … Main, D. (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.

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., … Simon, P. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics*, *48*, 657.

Jiao, Y., & Paterson, A. H. (2014). Polyploidy-associated genome modifications during land plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130355. https://doi.org/10.1098/rstb.2013.0355

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., … Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*, 1384–1395. https://doi.org/10.1101/gr.170720.113

Kellner, F., Kim, J., Clavijo, B. J., Hamilton, J. P., Childs, K. L., Vaillancourt, B., … O'Connor, S. E. (2015). Genome-guided investigation of plant natural product biosynthesis. *The Plant Journal*, *82*, 680–692. https://doi.org/10.1111/tpj.12827

Kwon, C. W., Park, K.-M., Kang, B.-C., Kweon, D.-H., Kim, M.-D., Shin, S. W., … Chang, P.-S. (2015). Cysteine protease profiles of the medicinal plant *Calotropis procera* R. Br. revealed by de novo transcriptome analysis. *PLoS ONE*, *10*, e0119328. https://doi.org/10.1371/journal.pone.0119328

La Rosa, R. J., & Conner, J. K. (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*, 150–160. https://doi.org/10.3732/ajb.1600328

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

Medicinal Plant Consortium. (2011, October). Release of the medicinal plant consortium transcriptome resources. Retrieved August 18, 2016, from http://medicinalplantgenomics.msu.edu/final_version_release_info.shtml

Morgan, M. T., & Schoen, D. J. (1997). Selection on reproductive characters: floral morphology in *Asclepias syriaca*. *Heredity*, *79*, 433.

Munkert, J., Pollier, J., Miettinen, K., Van Moerkercke, A., Payne, R., Müller-Uri, F., … Goossens, A. (2015). Iridoid synthase activity is common among the plant progesterone 5β-reductase family. *Molecular Plant*, *8*, 136–152. https://doi.org/10.1016/j.molp.2014.11.005

Pandey, A., Swarnkar, V., Pandey, T., Srivastava, P., Kanojiya, S., Mishra, D. K., & Tripathi, V. (2016). Transcriptome and Metabolite analysis reveal candidate genes of the cardiac glycoside biosynthetic pathway from *Calotropis procera*. *Scientific Reports*, *6*, 34464.

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

Pruitt, K., Brown, G., Tatusova, T., & Maglott, D. (2002). The Reference Sequence (RefSeq) Database. In J. McEntyre & J. Ostell (Eds.), *The NCBI Handbook* (p. Chapter 18).

Bethesda, MD: National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK21091/

Rahman, A. Y. A., Usharraj, A. O., Misra, B. B., Thottathil, G. P., Jayasekaran, K., Feng, Y., … Alam, M. (2013). Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genomics*, *14*, 75. https://doi.org/10.1186/1471-2164-14-75

Rasmann, S., Agrawal, A. A., Cook, S. C., & Erwin, A. C. (2009). Cardenolides, induced responses, and interactions between above- and belowground herbivores of milkweed (*Asclepias* spp.). *Ecology*, *90*, 2393–2404.

Rasmann, S., Erwin, A. C., Halitschke, R., & Agrawal, A. A. (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*, 16–25. https://doi.org/10.1111/j.1365-2745.2010.01713.x

Sabir, J. S. M., Jansen, R. K., Arasappan, D., Calderon, V., Noutahi, E., Zheng, C., … Ruhlman, T. A. (2016). The nuclear genome of *Rhazya stricta* and the evolution of alkaloid diversity in a medically relevant clade of Apocynaceae. *Scientific Reports*, *6*, 33782.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

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

Sparrow, F. K., & Pearson, N. L. (1948). Pollen compatibility in *Asclepias syriaca*. *Journal of Agricultural Research*, *77*, 187–199.

Straub, S. C. K., Cronn, R. C., Edwards, C., Fishbein, M., & Liston, A. (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biology and Evolution*, *5*, 1872–1885. https://doi.org/10.1093/gbe/evt140

Straub, S. C. K., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., … Liston, A. (2011). Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics*, *12*, 211. https://doi.org/10.1186/1471-2164-12-211

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

Thorn, A., Egerer-Sieber, C., Jäger, C. M., Herl, V., Müller-Uri, F., Kreis, W., & Muller, Y. A. (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*, 17260–17269. https://doi.org/10.1074/jbc.M706185200

Van Zandt, P. A., & Agrawal, A. A. (2004). Community-wide impacts of herbivore-induced plant responses in milkweed (*Asclepias syriaca*). *Ecology*, *85*, 2616–2629. https://doi.org/10.1890/03-0622

Vaughan, F. A. (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*, 89–100. https://doi.org/10.1007/BF00987690

Wang, N., Thomson, M., Bodles, W. J. A., Crawford, R. M. M., Hunt, H. V., Featherstone, A. W., … Buggs, R. J. A. (2012). Genome sequence of dwarf birch (*Betula nana*) and cross-

species RAD markers. *Molecular Ecology*, *22*, 3098–3111. https://doi.org/10.1111/mec.12131

Wikström, N., Kainulainen, K., Razafimandimbison, S. G., Smedmark, J. E. E., & Bremer, B. (2015). A Revised Time Tree of the Asterids: Establishing a Temporal Framework For Evolutionary Studies of the Coffee Family (Rubiaceae). *PLOS ONE*, *10*, e0126690. https://doi.org/10.1371/journal.pone.0126690

Willson, M. F., & Rathcke, B. J. (1974). Adaptive design of the floral display in *Asclepias syriaca* L. *The American Midland Naturalist*, *92*, 47–57. https://doi.org/10.2307/2424201

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

Wyatt, R., & Broyles, S. B. (1990). Reproductive biology of milkweeds (*Asclepias*): Recent advances. In S. Kawano (Ed.), *Biological approaches and evolutionary trends in plants* (pp. 255–272). San Diego, California: Academic Press, Inc.

Wyatt, R., & Broyles, S. B. (1994). Ecology and evolution of reproduction in milkweeds. *Annual Review of Ecology and Systematics*, *25*, 423–441.

Zhan, S., Merlin, C., Boore, J. L., & Reppert, S. M. (2011). The Monarch butterfly genome yields insights into long-distance migration. *Cell*, *147*, 1171–1185. https://doi.org/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.

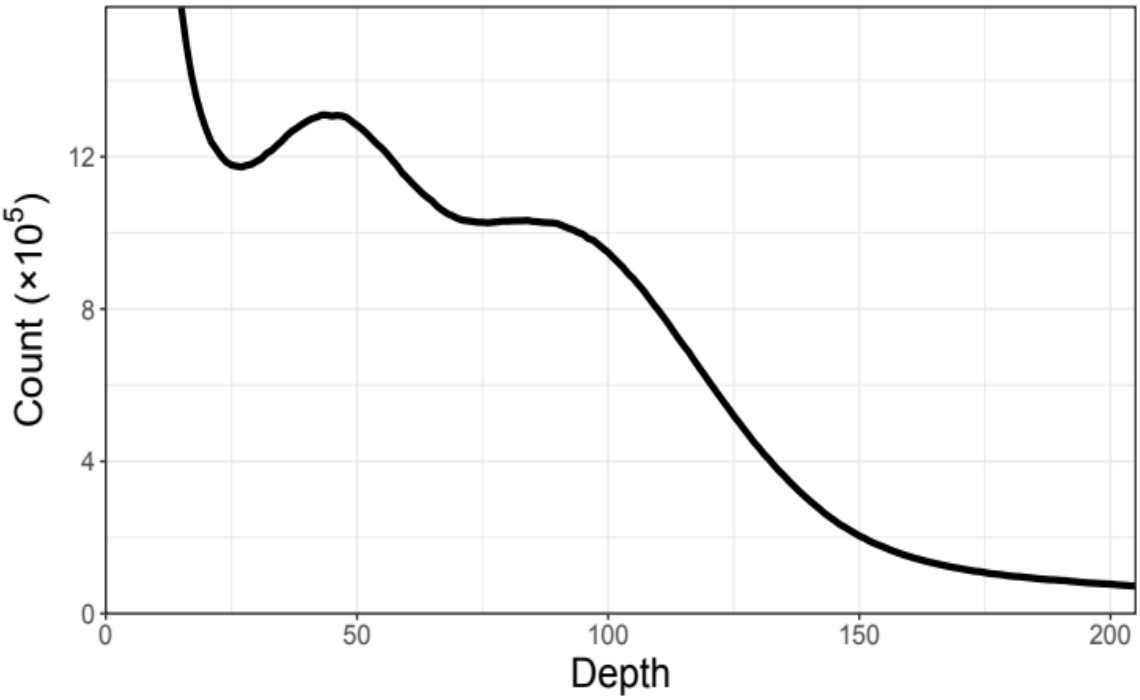

Figure 2: Peptide length histograms of *Asclepias*, *Coffea*, and *Catharanthus*. Mean and median peptide lengths are provided in the legend.

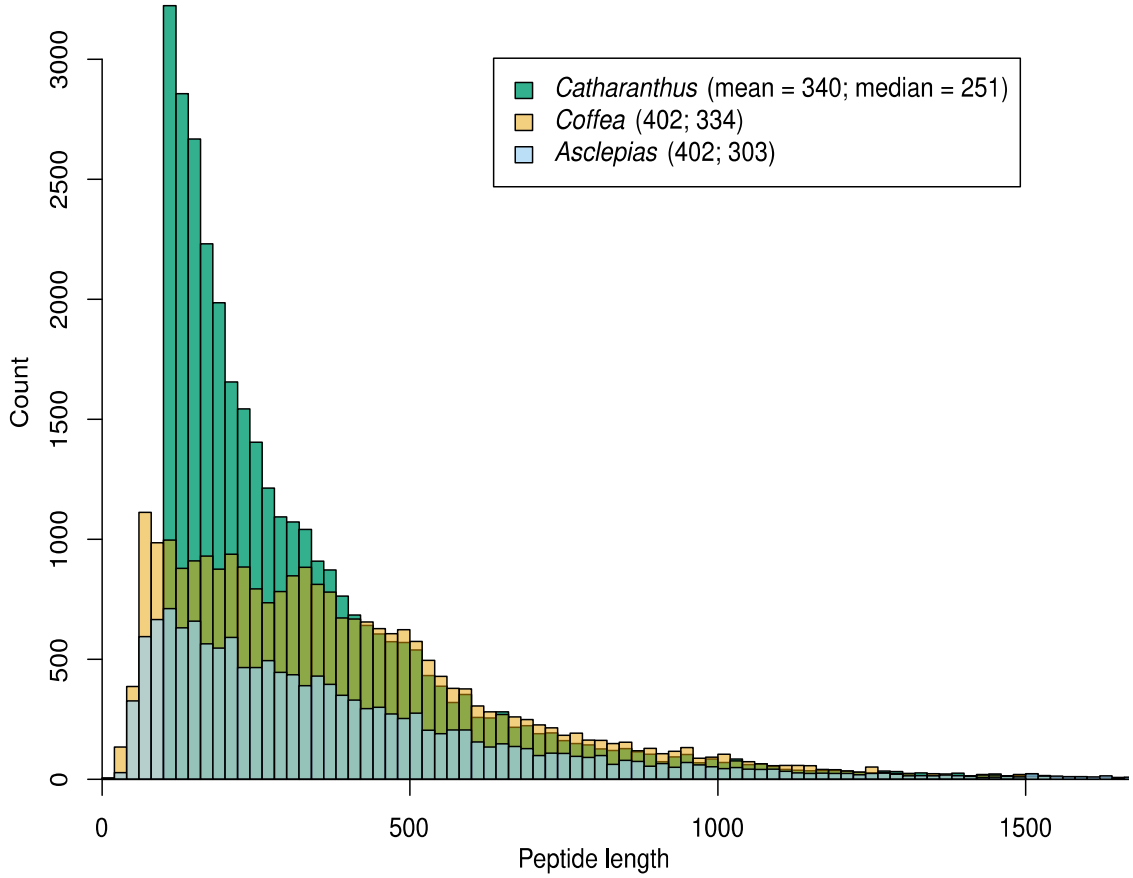

Figure 3: Gene family evolution in Apocynaceae inferred from transcriptomes.
The ultrametric tree depicts the phylogenetic relationships and estimated divergence times of sampled Apocynaceae and outgroups (*Coffea*, *Vitis*). All nodes had 100% bootstrap support, except one denoted by an asterisk, which had 97% bootstrap support. 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.

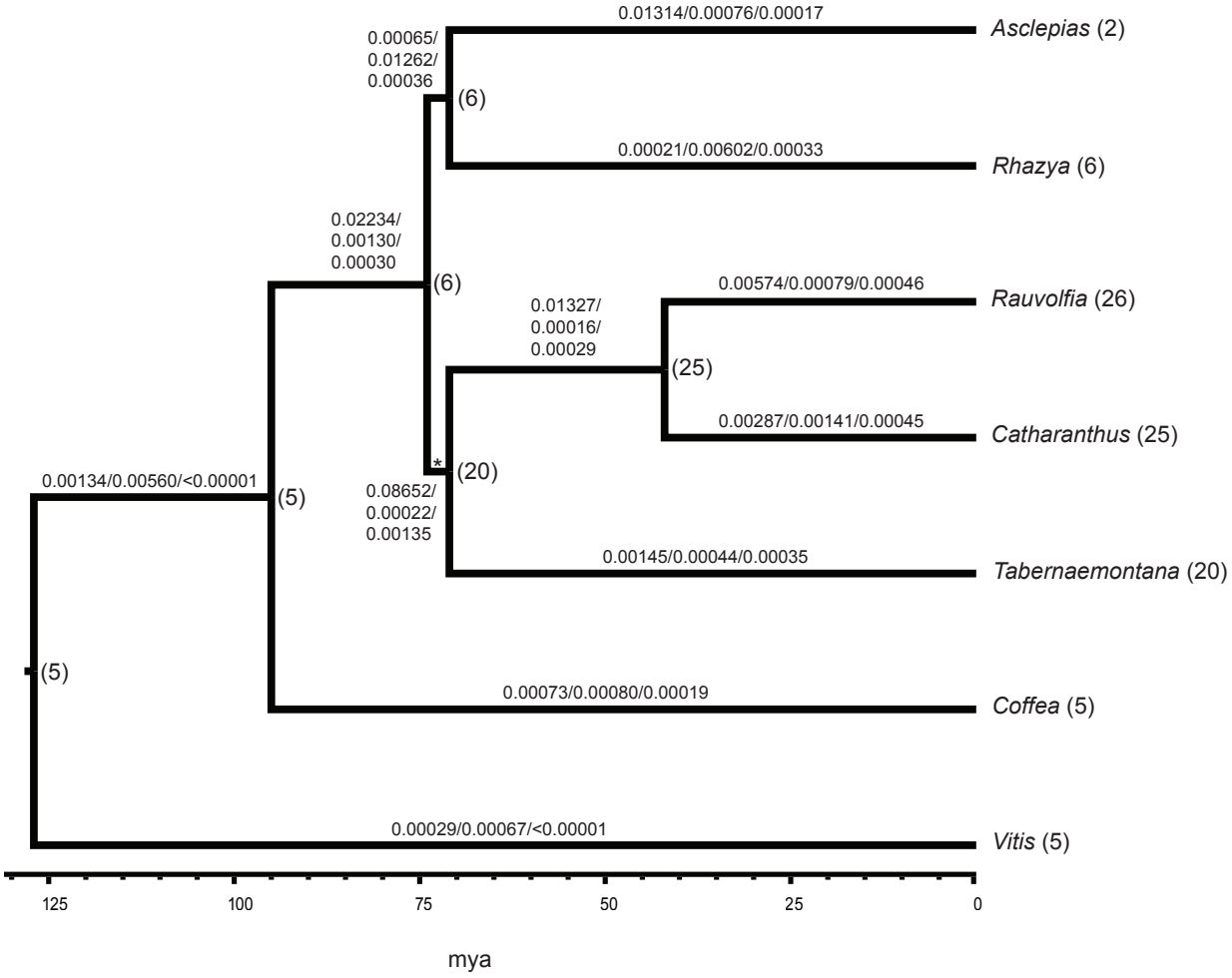

Figure 4: Counts of *Asclepias* linkage group scaffolds mapping to *Coffea* pseudochromosomes. Each column includes scaffolds from a single *Asclepias* linkage group, each row includes scaffolds mapping to a *Coffea canephora* pseudochromosome. *Coffea* chromosome 0 represents unassigned *Coffea* regions.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

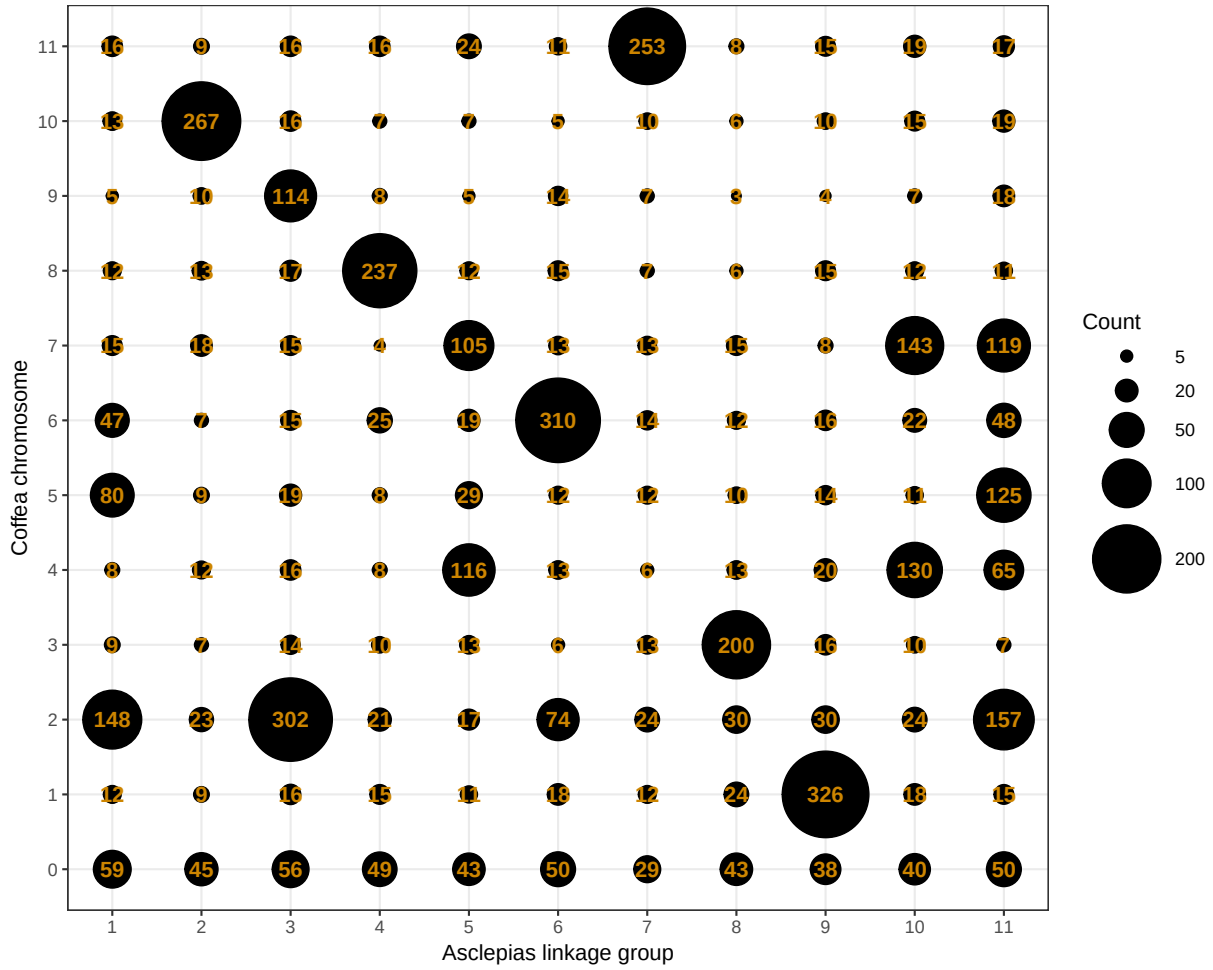

Figure 5: *Asclepias* linkage group scaffolds mapped to *Coffea* pseudochromosomes. *Coffea canephora* pseudochromosomes are shown in rows; the x-axis shows distance along each pseudochromosome. Each vertical bar represents one scaffold from the *Asclepias* core linkage groups, colored by its linkage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

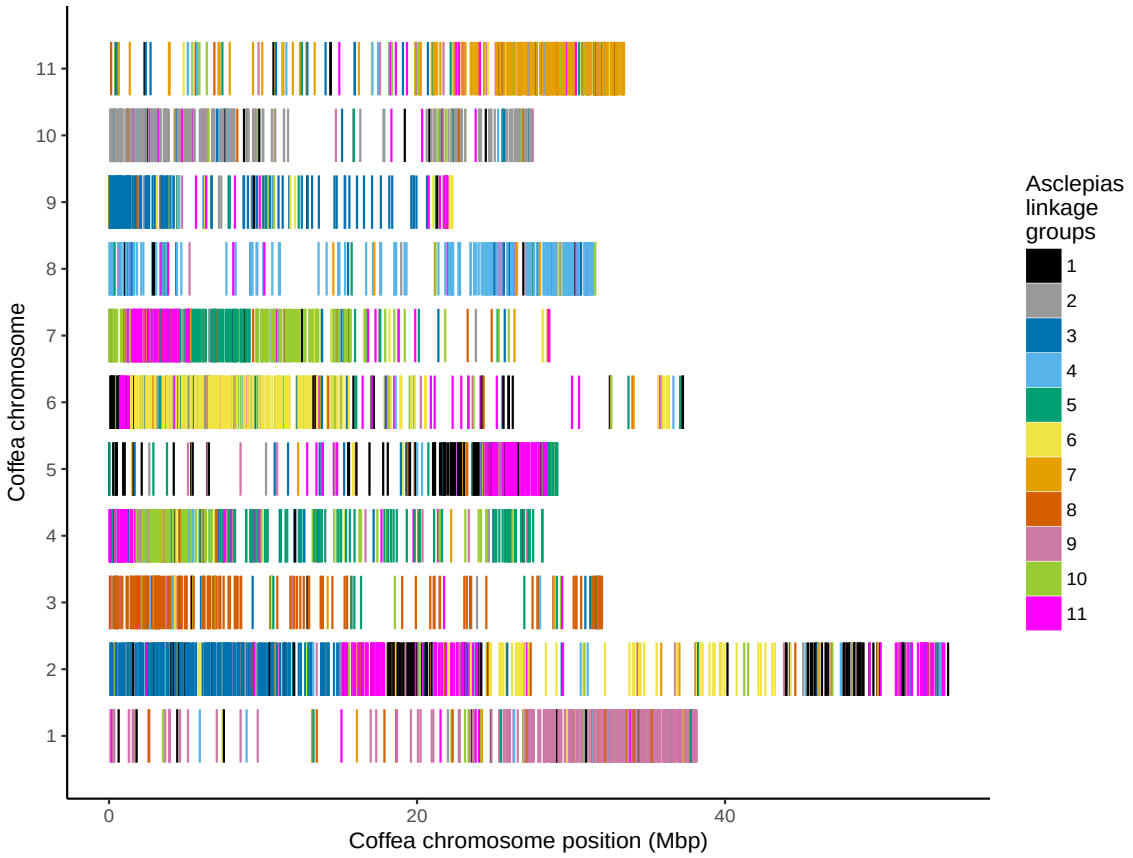

Figure 6: Maximum likelihood phylogeny of progesterone 5β-reductase paralogs. Numbers after labels represent numbered paralogs isolated from that species. *Asclepias syriaca* labels indicate the linkage group from which that sequence originates. Numbers at nodes indicate aBayes support values.

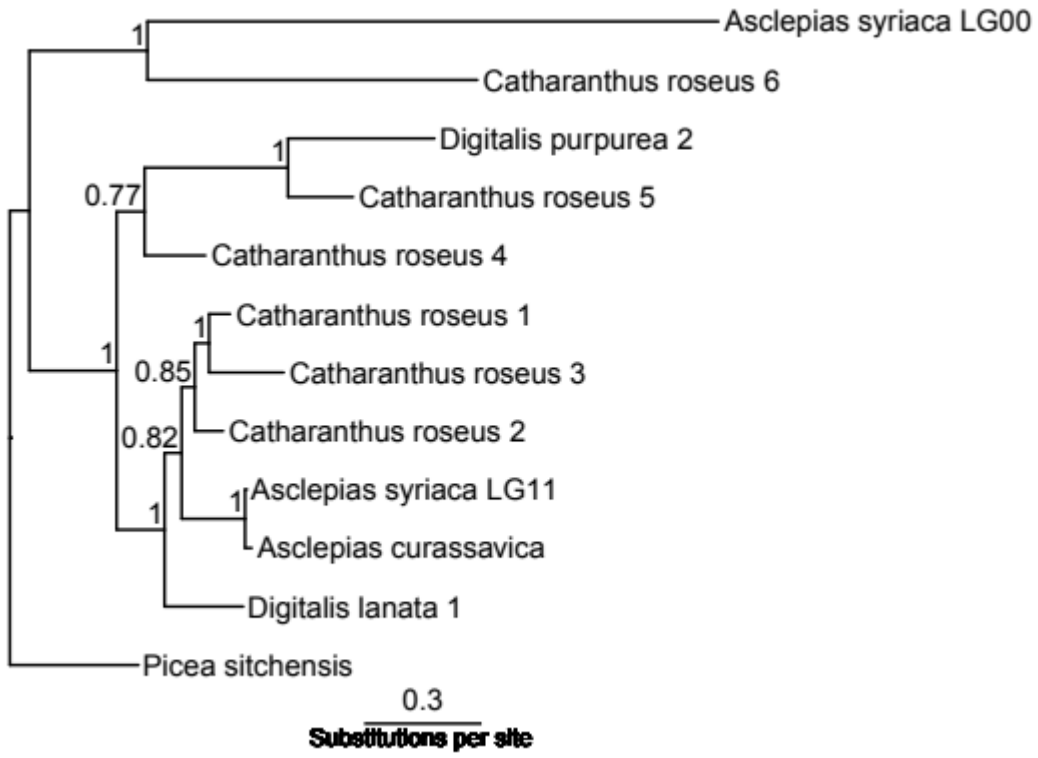

Table 1: Assembly comparison of *Asclepias*, *Catharanthus*, *Rhazya*, and *Coffea*. **Sequencing method** includes technologies and materials used in sequencing; N50 = 50% of the assembly is contained in scaffolds of this length or larger, BAC = bacterial artificial chromosome, SE = single-end, PE = 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
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.26488v1 | CC BY 4.0 Open Access | rec: 6 Feb 2018, publ: 6 Feb 2018

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. **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