Applications in Plant Sciences

Botanical Society of America
leading scientists and educators since 1893

GENOMIC RESOURCES ARTICLES

Genomic and transcriptomic resources for candidate gene discovery in the Ranunculids

Tatiana Arias[1,2,3,*], Diego Mauricio Riaño-Pachón[4,*], and Verónica S. Di Stilio[2,5]

Manuscript received 25 August 2020; revision accepted 3 December 2020.

[1] School of Biological Sciences, The University of Hong Kong, Pokfulam, Hong Kong

[2] Department of Biology, University of Washington, Seattle, Washington 98195-1800, USA

[3] Current address: Tecnológico de Antioquia, Calle 78B No. 72A, 220 Medellín, Colombia

[4] Laboratory of Computational, Evolutionary and Systems Biology, Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, São Paulo 13416-000, Brazil

[5] Author for correspondence: distilio@uw.edu

*These authors contributed equally to this work.

Citation: Arias, T., D. M. Riano-Pachon, and V. S. Di Stilio. 2021. Genomic and transcriptomic resources for candidate gene discovery in the Ranunculids. *Applications in Plant Sciences* 9(1): e11407.

doi: 10.1002/aps3.11407

PREMISE: Multiple transitions from insect to wind pollination are associated with polyploidy and unisexual flowers in *Thalictrum* (Ranunculaceae), yet the underlying genetics remains unknown. We generated a draft genome of *Thalictrum thalictroides*, a representative of a clade with ancestral floral traits (diploid, hermaphrodite, and insect pollinated) and a model for functional studies. Floral transcriptomes of *T. thalictroides* and of wind-pollinated, andromonoecious *T. hernandezii* are presented as a resource to facilitate candidate gene discovery in flowers with different sexual and pollination systems.

METHODS: A draft genome of *T. thalictroides* and two floral transcriptomes of *T. thalictroides* and *T. hernandezii* were obtained from HiSeq 2000 Illumina sequencing and de novo assembly.

RESULTS: The *T. thalictroides* de novo draft genome assembly consisted of 44,860 contigs (N50 = 12,761 bp, 243 Mbp total length) and contained 84.5% conserved embryophyte single-copy genes. Floral transcriptomes contained representatives of most eukaryotic core genes, and most of their genes formed orthogroups.

DISCUSSION: To validate the utility of these resources, potential candidate genes were identified for the different floral morphologies using stepwise data set comparisons. Single-copy gene analysis and simple sequence repeat markers were also generated as a resource for population-level and phylogenetic studies.

KEY WORDS draft genome; floral transcriptome; pollination syndrome; Ranunculaceae; sexual system; Thalictrum hernandezii; Thalictrum thalictroides.

Transcriptomic and genomic resources are a valuable tool for studying development within an evolutionary context, as they enable the search for candidate loci and their regulatory regions, contributing to the upgrading of study systems into model lineages. *Thalictrum* L. is an emerging model lineage within the Ranunculales (Damerval and Becker, 2017), the sister group to all other eudicots (Lane et al., 2018). The genus is therefore an extant representative of a privileged phylogenetic node, before a major evolutionary radiation within the flowering plants (Zeng et al., 2017). *Thalictrum* is a temperate genus of approximately 200 species with floral diversity that encompasses unisexual and wind-pollinated flowers, in association with polyploidy (Soza et al., 2012, 2013). Certain species, mainly *T. thalictroides* (L.) A. J. Eames & B. Boivin, are amenable to virus-induced gene silencing (Di Stilio et al., 2010), which has enabled functional studies of ABC model floral MADS-box genes (Galimba et al., 2012, 2018; Di Stilio et al., 2013; Galimba and Di Stilio, 2015; Soza et al., 2016). Two chloroplast genomes have been assembled to date for this genus, for *T. coreanum* H. Lév. (Park et al., 2015) and *T. thalictroides* (from this data set, Morales-Briones et al., 2019).

Here, we contribute genomic resources for *Thalictrum*, including the first de novo draft genome of diploid, hermaphroditic, and insect-pollinated *T. thalictroides*, and two de novo floral transcriptomes for *T. thalictroides* and for tetraploid, andromonoecious, and wind-pollinated *T. hernandezii* Tausch ex J. Presl (Fig. 1). A qualitative comparison between these transcriptomes provided a preliminary list of enriched transcription factor families and potential candidate genes for developmental studies. Microsatellite markers were also identified as a resource for population-level genetic studies. These genetic resources are presented as a tool to facilitate research on the genetic underpinnings of floral diversity in *Thalictrum*, and across Ranunculales.

FIGURE 1. Floral phenotypes of *Thalictrum* study species. Inflorescence of hermaphroditic *T. thalictroides* with hermaphrodite open flowers (A) and andromonoecious *T. hernandezi*i with staminate buds (♂) and hermaphrodite open flowers (⚥) occurring together in an inflorescence (B); inset, detail of young staminate flower. Scale bar = 1 cm.

METHODS

Thalictrum thalictroides draft nuclear genome

Plant materials and DNA extraction—Two live accessions of *T. thalictroides* (Tt), TtWT478 and TtWT964, were grown in the University of Washington greenhouse (Fig. 1A, Appendix 1). Total DNA was extracted separately from healthy young leaves of two individuals using a modified cetyltrimethylammonium bromide (CTAB) method (Shyu and Hu, 2013). DNA quality and concentration were measured in a Qubit 4 Fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and confirmed with agarose gel electrophoresis.

Library preparation—Total genomic DNA was sheared by sonication for 15–24 min using a Bioruptor (Diagenode, Denville, New Jersey, USA). Size-selected samples (200–400 bp) were extracted using x-tracta Gel Extractor tools (USA Scientific, Ocala, Florida, USA) and purified using the Gel DNA Purification Kit (QIAGEN, Germantown, Maryland, USA) for the end-repair, adenylation of 3′ ends, ligation, and enrichment steps. Sheared DNA was visualized on a 2% low-melting-point agarose gel stained with ethidium bromide. Illumina's TruSeq paired-end (2 × 100 bp) libraries (Illumina, San Diego, California, USA) were generated for each individual with three different insert sizes (170, 500, and 800 bp), following the manufacturer's instructions. Libraries were assessed in an Agilent 2100 Bioanalyzer with the DNA 1000 Kit (Agilent, Santa Clara, California, USA) to determine average length, quantitated by real-time quantitative PCR (qPCR) using a TaqMan Probe (Thermo Fisher Scientific) and amplified off-board on a cBot (TruSeq PE Cluster Kit v3–cBot–HS; Illumina) to generate clusters on the sequencing flow cell, and then sequenced on a HiSeq 2000 (TruSeq SBS Kit v3-HS; Illumina).

Read pre-processing and genome assembly—Raw read quality was visually inspected with FastQC (Andrews, 2015) and then treated with Trimmomatic version 0.38 (Bolger et al., 2014) to remove adapter sequences and low-quality bases, keeping only paired-end reads with at least 80 bp for de novo assembly. Assemblies were initially conducted in CLC Genomics Workbench version 7.5.1 (QIAGEN, Hilden, Germany) and SOAP de novo version 2.04 (Luo et al., 2012a); assemblies were then scanned using BlobTools (Laetsch and Blaxter, 2017) to identify contigs originating from contaminants and finalized in MaSuRCA version 3.2.9 (Zimin et al., 2013). Genome sequences of potential contaminants were downloaded from the National Center for Biotechnology Information (NCBI) and used to map the clean reads with BBDuk (BBMap version 35.85; https://sourceforge.net/projects/bbmap/), keeping only unmapped reads for genome assembly. *K*-mer-based statistics were computed for the clean reads with Jellyfish version 2.2.10 (Marçais and Kingsford, 2011), GenomeScope version 1.0 (Vurture et al., 2017), and Smudgeplots version 0.2.3 (Ranallo-Benavidez et al., 2020). One assembly was generated for each of the two *T. thalictroides* accessions, and a third was generated by combining the data from the two accessions. Each assembly was polished using Pilon version 1.23 (Bruce et al., 2014), with two rounds of error correction. Assembly statistics were computed with Quast version 5.0.2 (Gurevich et al., 2013), and sequence repeats were identified with RepeatModeler version 1.0.11 (http://www.repeatmasker.org/RepeatModeler/) and RepeatMasker version 4.0.9.p2 (http://www.repeatmasker.org/RepeatMasker/). Simple sequence repeat (SSR) markers and loci with di- to hexanucleotide repeats were identified in the genome and transcriptomes using the MicroSAtellite Identification tool (MISA; Thiel et al., 2003). The accuracy of the assemblies was assessed by mapping the contaminant-free clean reads back to the assembly using Bowtie2 version 2.4.1 (Langmead et al., 2012) and computing the fraction of reads that map in the correct orientation (forward-reverse) and within the length range of the insert size used to build the library; the insert size distribution was then estimated from the mapped paired-end reads using the CollectInsertSizeMetrics function in Picard version 2.23.8 (http://broadinstitute.github.io/picard/).

Gene predictions and orthogroup identification—RNA-seq data sets generated from *T. thalictroides* flowers (see below), together with a 1KP project transcriptome (Matasci et al., 2014) (http://www.onekp.com) and predicted protein sequences from the *Aquilegia coerulea* E. James genome (Filiault et al., 2018), were used as extrinsic evidence for protein-coding gene prediction using Braker2 version 2.1.2 (Hoff et al., 2019; Bruna et al., 2020). RNA-seq data were also employed for protein-coding gene prediction with StringTie version 1.3.6 (Pertea et al., 2015) and TransDecoder version 5.5.0 (https://transdecoder.github.io); ab initio protein-coding gene prediction was generated with SNAP daf76ba (Korf, 2004) using *Arabidopsis thaliana* (L.) Heynh. as a model. Final models for protein-coding genes were produced with EVidenceModeler version 1.1.1 (Haas et al., 2008). tRNAs were predicted with tRNAscan-SE version 2.0 (Chan and Lowe, 2019) and ribosomal RNA genes with RNAmmer version 1.2 (Lagesen et al., 2007). Other non-protein-coding RNA (ncRNA) were predicted with Infernal version 1.1.2 (Nawrocki and Eddy, 2013) and Rfam version 14.1 (Kalvari et al., 2018). Predicted protein-coding genes were evaluated with TransRate version 1.0.3 (Smith-Unna et al., 2016) against *Papaver somniferum* L. and *A. coerulea* (Ranunculales), and *A. thaliana* and *Solanum lycopersicum* L., representing the two main eudicot orders.

Genome annotation—Functional annotation was conducted with BLASTP (against the TAIR, SwissProt, and TrEMBL protein databases). Protein functional descriptions and Gene Ontology (GO) terms were added with the "Automated Assignment of Human Readable Descriptions" tool (AHRD version 3.3.3) (The Tomato Genome Consortium et al., 2012). Conserved regions, domains, and sequence features were identified with InterProScan version 5.35-74.0 (Jones et al., 2014). Transcription-associated proteins (TAPs) were identified using the approach described for the construction of PlnTFDB version 3.0 (http://plntfdb.bio.uni-potsdam.de) (Pérez-Rodríguez et al., 2010). Classification rules and software to assign protein sequences into TAP families based on their domain architecture were obtained from https://bitbucket.org/diriano/mytfdb. Benchmarking Universal Single-Copy Orthologs (BUSCO) version 3.0 (Waterhouse et al., 2017) was used with near-universal single-copy orthologs selected from OrthoDB version 9.0 for Embryophyta (Zdobnov et al., 2017). Clusters of orthologous genes (orthogroups) were identified with OrthoFinder version 2.3.3 (Emms and Kelly, 2019). Sequence similarity was computed with Diamond version 0.9.24 (Buchfink et al., 2015) and gene clusters were generated with MCL version 14-137, with 1.5 inflation (Enright et al., 2002).

Thalictrum thalictroides and *T. hernandezii* floral transcriptomes

Plant materials and RNA extractions—Fresh open flowers were collected from an individual of *T. thalictroides* (Fig. 1A) that was also used for genome sequencing (TtWT478, hermaphroditic flowers) and from *T. hernandezii* (Fig. 1B, Th_HWT441 hermaphroditic and Th_SWT441 staminate flowers) (Fig. 1, Appendix 1). Flowers of *T. thalictroides* and *T. hernandezii* were flash-frozen in liquid nitrogen and total RNA was extracted with TRIzol (Invitrogen, Carlsbad, California, USA), following the manufacturer's instructions. RNA quality (RIN ≥ 6.5) and concentration were determined in an Agilent 2100 Bioanalyzer and with agarose gel electrophoresis.

RNA library preparation—Library preparation and sequencing were carried out by the Beijing Genomics Institute (BGI Group, Hong Kong). Flower RNA-seq libraries were prepared for *T. thalictroides* hermaphrodite (Tt_H), *T. hernandezii* hermaphrodite (Th_H), and *T. hernandezii* staminate (Th_S) flowers using Illumina's TruSeq mRNA (unstranded) paired-end (2 × 100 bp) kit and sequenced in a HiSeq 2000 Illumina sequencer. A mixed floral stage library of *T. thalictroides* (1KP, library name: GBVZ; Matasci et al., 2014) was downloaded from NCBI's Sequence Read Archive and also included in the analysis.

Read pre-processing, sequence assembly, and annotation—Contaminant reads were removed using BBDuk (BBMap version 35.85) by mapping against the same contaminants detected during genome assembly (see above); de novo assembly was then conducted in Trinity version 2.8.5 (Grabherr et al., 2011) and assessed for completeness with BUSCO version 3.0 (as for the genome assembly). Two de novo transcriptome assemblies were generated, one for *T. thalictroides* and another for *T. hernandezii* (combining libraries for the two flower types). Only contigs larger than 200 bp were used in further analyses. Read mapping metrics were computed against the assemblies using Bowtie2 version 2.4.1 (Langmead et al., 2012) with the parameters *--maxins 1000 --very-sensitive* and Salmon version 0.14.0 (Patro et al., 2017) with the parameters *quant --validateMappings --seqBias*, as a measure of transcriptome accuracy. Assembled transcripts were compared against NCBI's non-redundant protein database using Diamond version 0.9.23 and assessed in MEGAN-LR version 6.14.2 (Huson et al., 2018). Polypeptides encoded by the assembled transcripts were identified with TransDecoder version 5.5.0, including BLASTP hits against the SwissProt database and profile hidden Markov model hits against the Pfam database (El-Gebali et al., 2019). Functional annotation, including identification of TAPs and clustering of shared (orthologous) genes, or "orthogroups," was carried out as described above.

Identification of candidate genes—First, we performed a three-way comparison to identify orthogroups among the *T. thalictroides* genome and the *T. thalictroides* and *T. hernandezii* de novo transcriptomes. We considered that orthogroups present in the *T. thalictroides* genome and transcriptome, but not found in the *T. hernandezii* transcriptome, are more likely to contain high-confidence genes involved in the development of floral traits that characterize insect-pollinated flowers (Fig. 1A). Conversely, orthogroups expressed exclusively in *T. hernandezii* (i.e., not found in the *T. thalictroides* transcriptome) that map to the *T. thalictroides* draft genome were considered more likely to contain high-confidence genes involved in the development of floral traits that characterize wind-pollinated flowers (Fig. 1B). Second, we analyzed *T. hernandezii*-specific orthogroups to identify transcripts associated with the different floral sexes, i.e., staminate (male) vs. hermaphrodite. To that end, we computed the expression level of the de novo–assembled *T. hernandezii* transcripts in the two *T. hernandezii* data sets (Ther_S and Ther_H) using Salmon version 0.14.0 with options *--seqBias --validateMappings --recoverOrphans --libType A* and considered a transcript as expressed when it had at least a single mapped read. For the data intersections of interest, we performed an enrichment analysis of the families of TAPs using a Fisher's exact test, correcting *P* values for the false discovery rate using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995). GO term enrichment analysis was conducted using topGO (Alexa

TABLE 1. Summary statistics for de novo genome assemblies of *Thalictrum thalictroides*.[a]

Metric	Consensus	WT964[b]	WT478[c]
No. of contigs	44,860	41,892	41,384
Total length (bp)	243,091,176	205,112,749	167,308,766
Largest contig (bp)	119,890	80,531	57,549
GC (%)	36.23	36.29	36.61
N50 (bp)	12,761	10,189	8017
L50	4281	4739	4875

[a]Assembly statistics were computed using Quast version 5.0.2 (Gurevich et al., 2013).
[b,c]Sequenced accessions from two different individuals.

and Rahnenfuhrer, 2010), with the weight0 method, correcting *P* values for the false discovery rate using the Benjamini–Hochberg method.

RESULTS

A draft nuclear genome for *Thalictrum thalictroides*

Genome sequencing and de novo assembly—Paired-end sequencing of the six libraries from two live accessions of *T. thalictroides* resulted in 49,105,897 sequenced fragments or 8.8 Gbp, after quality-trimming and contaminant removal (Appendix 2). The genome sequences of contaminants identified by BlobTools version 1.0.1 (Laetsch and Blaxter, 2017; Appendix S1), as well as mitochondria and plastids (Appendix S2), were used to map the clean reads with bbduk2, keeping only unmapped reads for further processing. Short-read data from TtWT478 had contamination from an aphid (GCF_000142985.2) and its bacterial endosymbiont (GCF_000009605.1; Appendix S1), which was removed. Metrics for the *T. thalictroides* draft genome assemblies were generated with MaSuRCA version 3.2.9 (Table 1; Appendices S3, S4). The best assembly, as measured by mapped reads (Appendix S4), assembly contiguity metrics (Table 1), and gene content (BUSCOs; Fig. 2), was generated by combining both accessions. This "consensus" assembly consisted of 44,860 contigs, with N50 = 12,761 bp (Table 1, Appendix S2) and 83.8% complete conserved embryophyte single-copy genes (88.5% when considering complete and fragmented BUSCOs; Fig. 2). The BUSCO estimate increased to 84.5% after gene prediction (90.9% when considering complete and fragmented BUSCOs). We confirmed with Smudgeplots (Ranallo-Benavidez et al., 2020) that the joint data set behaves like a diploid genome (Appendix S5), and the low level of duplicated BUSCOs in the consensus supports this (46/1440, or 3.2%; Fig. 2). We were able to map, on average, 5.75% more high-quality reads to the consensus assemblies than to either individual genome across the different next-generation sequencing libraries (Appendix S4). Mean contig coverage for the consensus assembly (from the two combined accessions) was 55× (median = 31.9×). Our genome size estimate from the consensus was 243.1 Mbp, comparable to the 286.4 Mbp estimate from *k*-mer frequency statistics in GenomeScope version 1.0; the heterozygosity estimate was 1.23% (Appendix S6).

More than one third of the genome (37.6%) could be assigned to different classes of repeat elements, with the most abundant class being the autonomous long terminal repeat retroelements, particularly from the Copia (7.7%) and Gypsy (6.8%) superfamilies (Appendix S7). SSRs were also a common occurrence in the genome; 65,651 were identified and the most common SSR was dinucleotide (Table 2; Arias et al., 2020a).

We predicted 33,624 protein-coding genes and 1936 non-coding RNA loci (Appendix S8). Orthogroup analysis resulted in 67.4% of predicted protein-coding genes forming clusters with at least one of the four reference genomes included in the analysis (Appendix S9); most were shared by all five species, providing support for our gene predictions. TransRate analysis also showed that a large number of predicted proteins in *T. thalictroides* had best bi-directional BLAST hits against the reference genomes of *A. thaliana*, *S. lycopersicum*, *A. coerulea*, and *P. somniferum* (Appendix S10), and 88.3% of predicted protein-coding genes had hits against InterPro member databases (Arias et al., 2020b). Functional descriptions were added for 22,603 protein-coding genes. We identified 1569 TAPs; 1258 of these were transcription factors (TFs) that can be grouped into 68 TF families, and the remaining 311 belonged to 29 families of other transcriptional regulators (oTRs) (Fig. 3, Appendix S11).

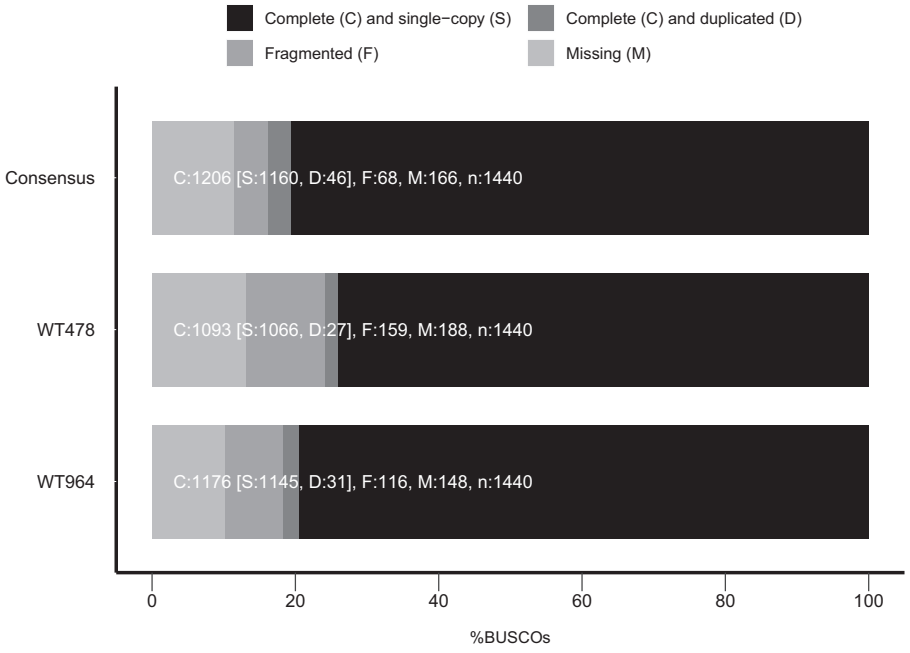

FIGURE 2. Proportion of Benchmarking Universal Single-Copy Orthologs (BUSCOs) for the *Thalictrum thalictroides* genome (total number of BUSCOs = 1440 genes in Embryophyta from OrthoDB version 9.0). Data are shown for the two sequenced accessions (WT964 and WT478) and the consensus.

Thalictrum thalictroides and *T. hernandezii* floral transcriptome assembly

De novo transcriptome assemblies of *T. thalictroides* (Tt; GHXU00000000) consisted of 54,104 contigs (N50 = 1817 bp), while *T. hernandezii* (Th; GHXT00000000) had 124,707 contigs (N50 = 1703 bp), with 80.1% and 82.9% identified complete BUSCOs, respectively (Table 3). For *T. thalictroides*, the total

TABLE 2. Simple sequence repeat (SSR) motif distribution in the *Thalictrum thalictroides* genome and the *T. thalictroides* and *T. hernandezii* (hermaphrodite and staminate combined) floral transcriptomes.

SSR properties	*T. thalictroides* genome[b]	*T. thalictroides* transcriptome	*T. hernandezii* transcriptome
Total no. of identified SSRs[a]	65,651	11,826	18,631
No. of SSRs present in tandem	3582	381	538
No. of contigs with SSR	22,867	9844	16,512
No. of contigs with >1 SSR	12,596	1634	1898

[a]SSR sequences are available at https://doi.org/10.6084/m9.figshare.11984370.v8.
[b]With abundance ≥5% of total SSRs in the genome.

percentage of identified complete BUSCOs increased to 94.2% when including both the transcripts predicted from the genome sequence and those from the de novo transcriptome assembly (96.3% when considering complete plus fragmented BUSCOs). TransDecoder identified 26,407 ORFs per sample for *T. thalictroides* and 52,313 for *T. hernandezii* (Arias et al., 2020c). There were approximately twice as many conditional reciprocal best BLAST hits in *T. hernandezii* compared to *T. thalictroides* with either of the reference transcriptomes, which is consistent with the former being a tetraploid (Table 4). A search within the *Thalictrum* floral transcriptomes detected 30,457 SSR markers, with approximately half of the analyzed contigs containing SSRs. The most common SSR was trinucleotide, followed by dinucleotide (Table 2; Arias et al., 2020a).

High representation of transcription factors in floral transcriptomes

The *T. thalictroides* and *T. hernandezii* floral transcriptomes contained 3541 transcripts that could be assigned into 94 TAP families; of these, *MYB-related*, *AP2-EREBP*, *bHLH*, and *bZIP* were among the top families. MADS-box genes (which include the floral organ identity genes) were found in all floral transcriptomes at approximately 2% relative abundance, and several TAP families were represented at significantly different levels in the three transcriptomes (Fig. 3).

Functional annotation of floral transcriptomes

To optimize orthogroup detection, *Thalictrum* transcriptomes were compared against multiple reference genomes: (1) *T. thalictroides* draft genome (this work); (2) *A. thaliana* (Brassicaceae); and (3) the two phylogenetically most closely related genomes available, *A. coerulea* (Ranunculaceae) and *P. somniferum* (Papaveraceae, Ranunculales). As a result of these comparisons, we identified 14 *T. thalictroides*- and 75 *T. hernandezii*-specific orthogroups (Appendix S12). Stepwise comparisons were conducted to (a) validate transcripts against the *T. thalictroides* draft genome and (b) conduct inter- and intraspecies qualitative comparisons between wind- vs. insect-pollinated and male vs. hermaphrodite floral morphologies (Fig. 1).

First, we performed an inter-species transcriptome comparison, using the draft genome for validation (Fig. 4A). The three-way intersection in the Venn diagram (5204 orthogroups) represents orthologs that can be mapped to the *T. thalictroides*

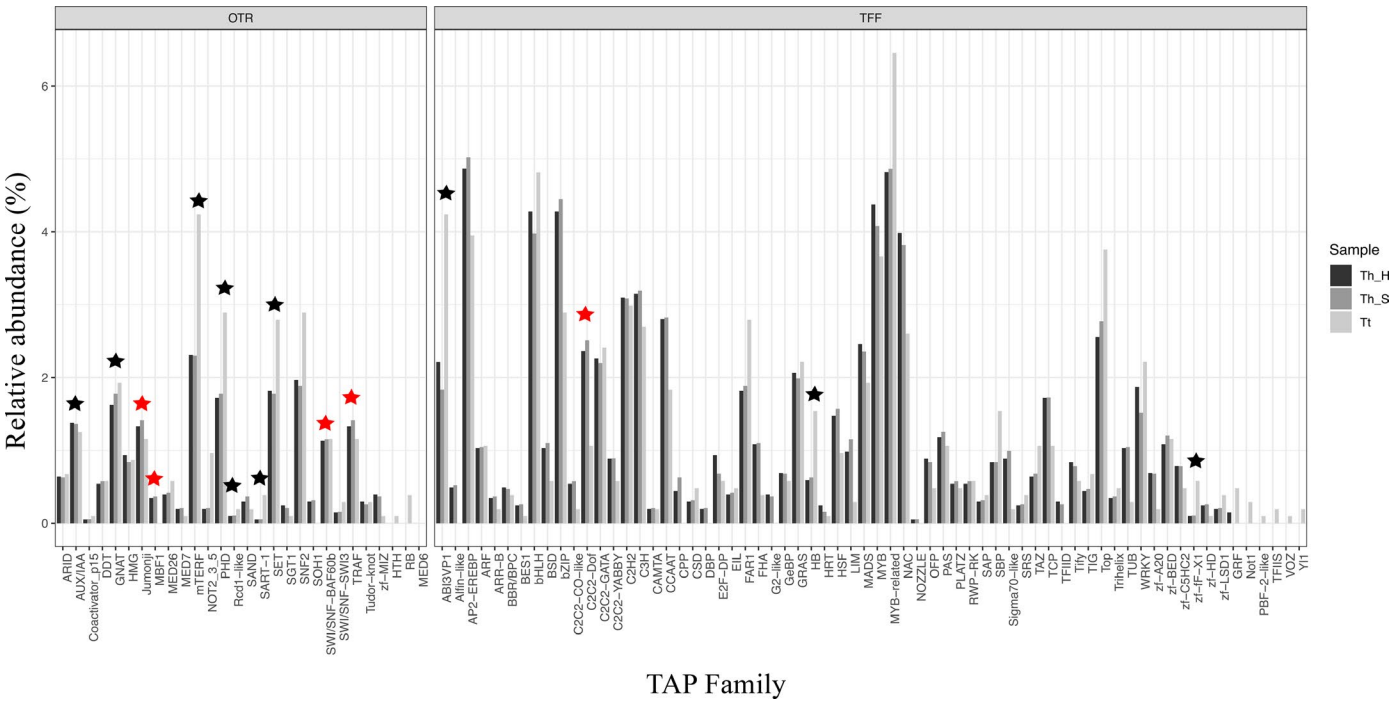

FIGURE 3. The relative abundance of different families of transcription-associated proteins (TAPs) in the floral transcriptomes of *Thalictrum thalictroides* (Tt) and *T. hernandezii* (Th_H, hermaphrodite and Th_S, staminate). Transcripts >0.1 TPM (transcripts per million) are included. Statistically significant differences between samples (chi-square test) are marked with black ($P < 0.01$) or red ($P < 0.05$) asterisks. TFF: transcription factor family, OTR: other transcriptional regulator family.

TABLE 3. Properties of de novo floral transcriptome assemblies of *Thalictrum thalictroides* and *T. hernandezii*.

Metrics	Parameter	*T. thalictroides*	*T. hernandezii*
Contigs	No. of sequences (transcripts)	54,104	124,707
	Largest sequence (bp)	14,923	11,578
	No. of bases	72,336,777	150,432,265
	Mean sequence length	1337.0	1206.3
	No. of sequences >1 Kbp	29,920	60,919
	N50 (bp)	1817	1703
	% GC	40	40
BUSCO	Complete	80.1%	82.9%
	Single copy	42.3%	16.5%
	Duplicated	37.8%	66.4%
	Fragmented	7.8%	7.6%
	Missing	12.1%	9.5%
Read mapping	No. of paired-ends (fragments)	40,504,757	65,712,208
	Proportion of paired-end fragments mapping back to the assembly	87.92% (Salmon)	89.64% (Salmon)
		88.85% overall mapping	90.72% overall mapping
		84.25% concordantly mapping (Bowtie2)	87.03% concordantly mapping (Bowtie2)

draft genome and that are expressed in floral transcriptomes of both species. A GO term enrichment analysis for *biological process* provided a broad characterization of gene functions at this triple intersection, examples of relevant categories include *gene silencing by miRNA*, *maintenance of floral organ identity*, *meristem initiation*, *adaxial/abaxial pattern specification*, and *negative regulation of growth* (Appendix S13; Arias et al., 2020d). A "core" of 9556 orthogroups common to both species is represented by three of the intersecting areas (5204 + 1298 + 3054). Intersection area "a" (3477 orthogroups) comprised 6451 transcripts uniquely expressed in *T. thalictroides* that also mapped to the reference genome (therefore considered of high confidence). Intersection area "b" (3054 orthogroups) comprised 11,251 transcripts found exclusively in *T. hernandezii* and similarly validated by the reference genome. Two sets of species-specific orthogroups not found in the genome (274 and 473 orthogroups each) could represent lineage-specific expansions and/or losses, or artifacts arising from incomplete sequencing. Finally, orthogroups found in both transcriptomes but not in the genome (1298) point to limitations due to fragmentation, as many can be found in other reference genomes (Appendix S12). Second, we performed an intraspecific comparison within *T. hernandezii*-specific orthogroups validated by our draft genome (Fig. 4A area "b"; 3054 orthogroups). Transcripts from male (Ther_S) and hermaphrodite (Ther_H) floral transcriptomes were compared, yielding 447 male-specific and 765 hermaphrodite-specific transcripts (Fig. 4B, areas "c" and "d", respectively).

DISCUSSION

The genome completeness results obtained here are within range of other draft genomes of non-traditional species, such as *Calotropis gigantea* (77–88%; Hoopes et al., 2017) and certain Brassicaceae (61–95%; Lopez et al., 2017), yet lower than the reference genomes used in our orthogroup analysis (94.5–99.6%; Appendix S9). Our genome size estimated range of 243.1–286.4 Mbp was comparable to that of *A. coerulea*, another diploid member of the Ranunculaceae (291.7 Mbp; Filiault et al., 2018). This value is slightly lower than our previous estimate for this species from flow cytometry (1C value = 0.366 pg or 356 Mbp; Soza et al., 2013), which we attribute to the relatively low sequencing depth and to the presence of complex or long repetitive regions (longer than the short reads) that were not recovered in our sequencing and/or assembly. Both of these limitations can be overcome in the near future with additional deep sequencing from third-generation sequencing technologies. Among the most abundant repetitive elements found, long terminal repeat transposable elements have been previously found to underlie homeotic flower mutants of *T. thalictroides* (Galimba et al., 2012). Our heterozygosity estimate of 1.23% is higher than that of *A. coerulea* (0.2–0.35%; Filiault et al., 2018) but still within the range of obligate outcrossers (Leffler et al., 2012). The number of TF families detected in the *T. thalictroides* draft genome is comparable to that found in the genomes of *A. thaliana*, *S. lycopersicum*, *A. coerulea*, and *P. somniferum*. The ratio of TF to oTRs was similar among the three

TABLE 4. Comparative metrics of de novo floral transcriptome assemblies of *Thalictrum thalictroides* and *T. hernandezii*.

	T. thalictroides		*T. hernandezii*	
Parameter	Reference *Aquilegia coerulea*	Reference *T. thalictroides*	Reference *Aquilegia coerulea*	Reference *T. thalictroides*
No. of CRBB[a] contigs	40,771	46,744	97,550	90,591
Proportion of CRBB contigs	0.75357	0.86397	0.78223	0.72643
No. of reciprocal best hits per reference	0.99289	1.31451	2.74325	2.20615
No. of CRBB reference transcripts	18,877	19,387	18,323	21,224
Proportion of CRBB reference transcripts	0.45971	0.54519	0.51527	0.51686

[a]Conditional reciprocal best BLAST algorithm, implemented in TransRate (Smith-Unna et al., 2016).

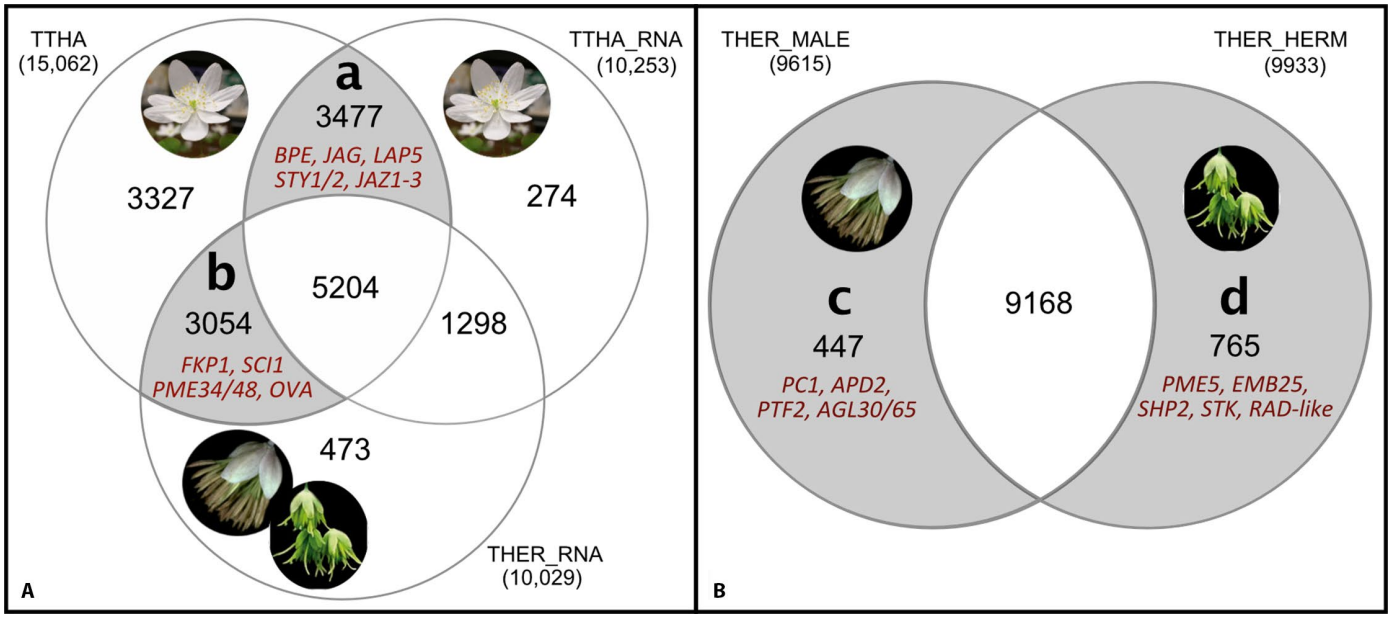

FIGURE 4. The number of shared and unique orthogroups or transcripts among the *Thalictrum* data sets in this study, with examples of candidate genes (in red based on *Arabidopsis*, see text for details). The total number of orthogroups per data set is indicated in parentheses. Representative flower pictures are shown for *T. thalictroides* (TTHA), insect-pollinated with hermaphrodite flowers, and *T. hernandezii* (THER), wind-pollinated with hermaphrodite (THER_HERM) and staminate (THER_MALE) flowers. (A) Interspecies comparison of floral transcriptomes (TTHA_RNA and THER_RNA) and validation against the *T. thalictroides* draft genome (TTHA). Area "a" represents *T. thalictroides*-exclusive orthogroups found in the draft genome; "b" represents *T. hernandezii*-exclusive orthogroups (both floral types combined) found in the draft genome. (B) Intraspecies comparison of transcripts of two floral types in *T. hernandezii* (male and hermaphrodite). Area "c" represents male flower–specific transcripts; "d" represents hermaphrodite flower–specific transcripts.

representatives of Ranunculales (*T. thalictroides*, *A. coerulea*, and *P. somniferum*), with approximately 4.1 TFs per oTR, and lower than estimates for *A. thaliana* and *S. lycopersicum* (5.1 TF per oTR).

The use of multiple reference genomes captured most orthologs in our transcriptomes, aiding in the validation of our gene predictions. Because *T. hernandezii* is a tetraploid (Soza et al., 2013), de novo transcriptome assembly was expected to include up to four expressed alleles per gene, thus explaining the larger number of assembled transcripts and of reciprocal best hits per reference for this species (Tables 3, 4), as well as the larger fraction of duplicated BUSCOs.

Applications: Data-mining examples for candidate genes in flower development

To test for applications of our results, we identified potential candidate genes for the morphological differences between flower types in the two species (Fig. 1). Our goal was to provide a preliminary, qualitative working list of candidate genes for future investigations of the genetic basis of distinct sexual systems (hermaphroditic vs. unisexual) and pollination modes (insect vs. wind). To that end, we first searched for known candidate genes within our comparisons (Fig. 4). Venn diagram areas "a" and "b" (Fig. 4A) represent functional orthogroups for flowers with distinct morphologies due to differing pollination modes: insect-pollinated *T. thalictroides* vs. wind-pollinated *T. hernandezii*. Venn diagram areas "c" and "d" (Fig. 4B) represent examples of transcripts expressed in flowers with distinct sexual systems: staminate flowers, with sepals and stamens, and hermaphrodite flowers, with added carpels. It is possible that a small number of these orthogroups are expressed at low levels in both species, but that due to the lack of replicates they would appear as differentially expressed. *Thalictrum thalictroides* has petaloid sepals that are comparatively bigger and white, upright stamens with smaller anthers, and carpels with short styles and stigmas. *Thalictrum hernandezii* has smaller flowers with smaller, green sepals, pendant stamens with larger anthers on longer filaments, and carpels with longer styles and stigmas (Fig. 1). Based on these phenotypic differences, we predict that the transcriptome comparisons could yield genes involved in processes such as cell elongation or cell division (longer stamen filaments and styles), increased flexibility (in pendant filaments), epidermal cell elongation (extended stigmatic papillae in wind-pollinated flowers; Di Stilio et al., 2009), or increased pollinator grip in petaloid organs, among others. A subset of the genes emerging from our comparisons fit these criteria, thus serving as validation for the usefulness of our data sets as a resource (see below).

First, we searched for previously characterized candidate genes in the *T. thalictroides* draft genome and the three transcriptomes. A homolog of *MIXTA-like2* (*ThtMYBML2*, FJ487606.1) with a role in papillate cells and stigmatic papillae was identified in both species, as a full transcript in the genome assembly (GenBank: KAF5204412.1) and in the *T. hernandezii* transcriptome (90% protein identity, TSA:GHXT01017115), and in two fragments in the *T. thalictroides* transcriptome assembly (TSA:GHXU01051721

and GHXU01036124). A second candidate for differences in morphology between species is the *Thalictrum STYLE2.1* ortholog, which is involved in style length in tomato (Chen et al., 2007). A *Solanum* query (UniProt B6CG44) was used to retrieve sequences from the *T. thalictroides* genome (65% protein identity, GenBank: KAF5189420.1) and the transcriptomes (65% identity, TSA:GHXU01012872; 67% identity, TSA:GHXT01076288). The presence of these candidate genes at the three-way intersection of all data sets (Fig. 4A) suggests that regulatory changes in expression levels, rather than on/off switches, likely underlie the phenotype differences.

We then identified candidate orthogroups or transcripts of interest in our data set via stepwise comparisons (Arias et al., 2020e). Orthogroups exclusive to *T. thalictroides* (Fig. 4A, area "a") and relevant to the insect pollination syndrome included orthologs of *BIG PETAL* (*BPE*; Szécsi et al., 2006) and *JAGGED* (*JAG*; Sauret-Güeto et al., 2013) with a potential role in large, petaloid sepals (Fig. 1A); *LESS ADHESIVE POLLEN 5* (*LAP5*; Dobritsa et al., 2010) for "stickier" pollen (higher in insect-pollinated species); *STYLISH 1/2* (*STY1/2*; Sohlberg et al., 2006) for regulation of style length (short) and stigma size (compact or "capitate"); and jasmonate (JA) hormone pathway *JAZ 1-3* (Figueroa and Browse, 2015) in stamen filament elongation (shorter and flattened). Other "usual suspects" in flower development enriched in this area included *WUSCHEL-related homeobox* (*WOX3-4*) contributing to general aspects of floral architecture and morphology (Costanzo et al., 2014), the YABBY family *INNER-NO-OUTER* (*INO*) in ovule development (Simon et al., 2017), and the TCP family *CYCLOIDEA* (*CYC*) in flower symmetry (Hileman and Cubas, 2009).

Thalictrum hernandezii–specific orthogroups relevant to the wind-pollination syndrome (Fig. 4A, area "b") included orthologs of *Arabidopsis FLAKY POLLEN 1* (*FKP1*) affecting pollen coat qualities (Ishiguro et al., 2010) relevant to pollen adaptations to wind pollination (less "sticky"); *PECTIN METHYLESTERASE 34* (*PME34*), which is highly expressed in stamen filaments and relevant to long, flexible filaments and styles (Gou et al., 2012), and *PME48*, a promoter of pollen tube elongation and thus relevant for successful fertilization through long styles (Leroux et al., 2015); and *STIGMA/STYLE CELL-CYCLE INHIBITOR 1* (*SCI1*; DePaoli et al., 2014), relevant to the extended styles and stigmas. Most members of the *OVULE ABORTION* (*OVA*) family (Berg et al., 2005), relevant to sex determination, were also specific to this andromonoecious species.

Comparisons between *T. hernandezii* hermaphrodite and staminate flower transcriptomes (intra-individual; Fig. 4B) were best suited to identify carpel-specific candidate genes (only present in hermaphrodite flowers) and, to a lesser extent, genes related to sexually dimorphic features of stamens and sepals, which are found in both flower types. Out of 765 transcripts uniquely expressed in *T. hernandezii* hermaphrodite flowers (Fig. 4B, area "d"), relevant candidates included orthologs of *PECTIN METHYL ESTERASE 5* (*PME5*) expressed in carpels and during fruit development (Louvet et al., 2006), *SHATTERPROOF 2* (*SHP2*) and *SEEDSTICK* (*STK*) in ovule development (Favaro et al., 2003), *EMBRYO DEFECTIVE 25* (*EMB25*) in embryo development (Meinke, 2020), and *RADIALIS-like* (*RAD-like*) in ovule and embryo development (Baxter et al., 2007). Among the 447 male-specific transcripts, MIKC* MADS-box genes *AG-like30/65* are main contributors to pollen and pollen-tube function (Adamczyk and Fernandez, 2009) and orthologs with pollen development function, such as *POLLEN CALCIUM BINDING PROTEIN 1* (*PC1*; Wang et al., 2008), *ABERRANT POLLEN DEVELOPMENT 2* (*APD2*; Luo et al., 2012b), and *POLLEN EXPRESSED TRANSCRIPTION FACTOR 2*, affecting pollen germination (*PTF2*; Niu et al., 2013).

Conclusions

This study provides genomic and transcriptomic resources for *Thalictrum*, a representative of an early-diverging lineage of eudicots with distinct floral morphologies representing diversity in sexual and pollination systems. Genomic resources for *T. thalictroides* and transcriptomes for *T. thalictroides* and *T. hernandezii* (Ranunculaceae) generated here increased the known set of protein-coding genes for this genus to 33,624 (predicted from genome sequence, BioProject: PRJNA439007), from approximately 132 nuclear genes, 10,461 expressed sequence tags, and 130 population sets available in NCBI databases to date. The value of these resources has been exemplified in the identification of transposable elements, molecular markers, and putative candidate genes. Future potential uses of these resources include the identification of other genes of interest and their regulatory regions (draft genome), as well as primer design to contribute to ongoing phylogenetic and population-level studies in *Thalictrum* (e.g., Humphrey and Ossip-Drahos, 2018; Timerman and Barrett, 2018) and in other Ranunculids.

ACKNOWLEDGMENTS

Funding was provided by the Research and Conference Grants Administration System (RCGAS) of The University of Hong Kong's Small Project Funding (to T.A.); the National Science Foundation's Opportunities for Promoting Understanding through Synthesis–Mid-Career Synthesis (OPUS-MCS; DEB 1911539); and The Fred C. Gloeckner Foundation, Inc. (to V.S.D.). Computational resources were provided by the University of Washington, the Earlham Institute, Institut national de la recherche agronomique (INRA), and Center for Nuclear Energy in Agriculture of the University of São Paulo. D.M.R.P. is a level 2 Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) research fellow (Brazil; 310080/2018-5). The authors thank Gemy George Kaithakottil (Earlham Institute, United Kingdom), Eric Montaudon (National Institute for Agricultural, Food and Environmental Research, France), and Jorge Mario Muñoz and Juliana Arcila (Corporación para Investigaciones Biológicas, Colombia) for help with scripts.

AUTHOR CONTRIBUTION

T.A. co-developed questions and framework, obtained funding, made collections, performed lab work, performed analyses, and wrote the initial draft of the manuscript. D.M.R.P. co-developed the framework, performed analyses, and edited the manuscript. V.D.S. co-developed questions and framework, performed analyses, mentored T. A., and wrote and edited the manuscript. All authors approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Raw reads and assemblies have been deposited in the National Center for Biotechnology Information (NCBI; BioProject PRJNA439007

and Sequence Read Archive [SRA] SRP136081) (Appendices 1, 2). Additional data are available in Figshare: GO enrichment analysis: https://doi.org/10.6084/m9.figshare.12465254.v1;MISA:https://doi.org/10.6084/m9.figshare.11984370.v8; orthogroups: https://doi.org/10.6084/m9.figshare.11984358.v3; TAPs: https://doi.org/10.6084/m9.figshare.11984313.v1; and commands and their main arguments used for the main steps of analyses: https://doi.org/10.6084/m9.figshare.13224371.v1.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Phylum-level taxonomic assignment of contaminants in the genome of *Thalictrum thalictroides* (from BlobPlots). The top plot summarizes sequence coverage and GC content for each contig in the assembly (circles), with *T. thalictroides* contigs shown in yellow. The size of the circles is proportional to contig size; the color key for taxa is provided in the inset legend. The bottom plot shows the frequency distribution of the contaminating taxa for each of the two accessions used.

APPENDIX S2. Genome accession numbers of species, mitochondria and plastids used for contaminant removal using BBduk.

APPENDIX S3. Frequency of contig length in de novo genome assemblies of *Thalictrum thalictroides*.

APPENDIX S4. Metrics for clean read mapping to de novo genome assemblies of *Thalictrum thalictroides* (after quality-trimming and contaminant removal).

APPENDIX S5. Estimation of ploidy from clean reads using Smudgeplots version 0.2.3 (*k*-mer = 21). Color intensity reflects the approximate number of *k*-mers per bin, from purple (low) to yellow (high). The most abundant *k*-mer pair is the diploid heterozygous (AB).

APPENDIX S6. Genome size estimation with GenomeScope version 1.0.

APPENDIX S7. Transposable and other repeat elements in the *Thalictrum thalictroides* genome.

APPENDIX S8. Number of predicted RNA types in the *Thalictrum thalictroides* genome.

APPENDIX S9. Identification of orthogroups in the *Thalictrum thalictroides* genome using columbine (*Aquilegia coerulea*), tomato (*Solanum lycopersicum*), opium poppy (*Papaver somniferum*), and *Arabidopsis thaliana*. Protein identifiers for each orthogroup and species are available at https://doi.org/10.6084/m9.figshare.11984358.v1.

APPENDIX S10. Evaluation of the predicted transcriptome from the *Thalictrum thalictroides* genome assembly (Tt) against four reference genomes (see Appendix S9). Metrics are from TransRate (Smith-Unna et al., 2016).

APPENDIX S11. Number and type of transcription-associated proteins (TAPs) and proportion of complete BUSCO in the genomes of *Thalictrum thalictroides* (TTHA), *Aquilegia coerulea* (ACOE), *Papaver somniferum* (PSOM), *Solanum lycopersicum* (SLYC), and *Arabidopsis thaliana* (ATHA).

APPENDIX S12. The number of shared and unique orthologous genes (orthogroups) at the intersection between five species and six data sets. Genome comparisons between *Thalictrum thalictroides* (TTHA, this study) and the reference genomes of *Arabidopsis thaliana* (ATHA), *Aquilegia coerulea* (ACOE), and *Papaver somniferum* (PSOM); and two *Thalictrum* floral transcriptomes, *T. thalictroides* (TTHA_RNA, this study) and *T. hernandezii* (THER_RNA; this study).

APPENDIX S13. Enriched Gene Ontology categories for comparisons between floral transcriptomes of *Thalictrum thalictroides* (Tt) and *T. hernandezii* (Th).

LITERATURE CITED

Adamczyk, B. J., and D. E. Fernandez. 2009. MIKC* MADS domain heterodimers are required for pollen maturation and tube growth in *Arabidopsis*. *Plant Physiology* 149: 1713–1723.

Alexa, A., and J. Rahnenfuhrer. 2010. topGO: Enrichment analysis for gene ontology. *R Package Version* 2: 2010.

Andrews, S. 2015. FastQC: A quality control tool for high throughput sequence data, version 0.11.8. Website http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [accessed 18 December 2020].

Arias, T., D. M. Riaño-Pachón, and V. S. Di Stilio. 2020a. MISA results for transcriptomes (*T. thalictroides* and *T. hernandezii*) and genome (*T. thalictroides*) [published 12 June 2020]. Available at figshare https://doi.org/10.6084/m9.figshare.11984370.v8 [accessed 16 December 2020].

Arias, T., and D. M. Riaño-Pachón. 2020b. Interproscan results [published 14 May 2020]. Available at figshare https://doi.org/10.6084/m9.figshare.12302096.v1 [accessed 16 December 2020].

Arias, T., D. M. Riaño-Pachón, and V. Di Stilio. 2020c. Transdecoder dataset [published 21 June 2020]. Available at figshare https://doi.org/10.6084/m9.figshare.12524036.v1 [accessed 16 December 2020].

Arias, T., D. M. Riaño-Pachón, and V. S. Di Stilio. 2020d. GO enrichment analysis [published 11 June 2020]. Available at figshare https://doi.org/10.6084/m9.figshare.12465254.v1 [accessed 16 December 2020].

Arias, T., D. M. Riaño-Pachón, and V. S. Di Stilio. 2020e. Orthogroups [published 24 June 2020]. Available at figshare https://doi.org/10.6084/m9.figshare.11984358.v3 [accessed 16 December 2020].

Baxter, C. E. L., M. M. R. Costa, and E. S. Coen. 2007. Diversification and co-option of RAD-like genes in the evolution of floral asymmetry. *Plant Journal* 52: 105–113.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57: 289–300.

Berg, M., R. Rogers, R. Muralla, and D. Meinke. 2005. Requirement of aminoacyl-tRNA synthetases for gametogenesis and embryo development in *Arabidopsis*. *Plant Journal* 44: 866–878.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Bruce, J. W., T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9: e112963.

Bruna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky. 2020. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. bioRxiv 2020.08.10.245134 [Preprint] [published 11 August 2020]. Available from https://doi.org/10.1101/2020.08.10.245134 [accessed 22 December 2020].

Buchfink, B., C. Xie, and D. H. Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59–60.

Chan, P. P., and T. M. Lowe. 2019. tRNAscan-SE: Searching for tRNA genes in genomic sequences, 1–14. *In* M. Kollmar [ed.], Gene prediction. Methods in Molecular Biology, vol. 1962. Humana Press, New York, New York, USA.

Chen, K.-Y., B. Cong, R. Wing, J. Vrebalov, and S. D. Tanksley. 2007. Changes in regulation of a transcription factor lead to autogamy in cultivated tomatoes. *Science* 318: 643–645.

Costanzo, E., C. Trehin, and M. Vandenbussche. 2014. The role of *WOX* genes in flower development. *Annals of Botany* 114: 1545–1553.

Damerval, C., and A. Becker. 2017. Genetics of flower development in Ranunculales: A new, basal eudicot model order for studying flower evolution. *New Phytologist* 216: 361–366.

DePaoli, H. C., M. C. Dornelas, and M. H. S. Goldman. 2014. SCI1 is a component of the auxin-dependent control of cell proliferation in *Arabidopsis* upper pistil. *Plant Science* 229: 122–130.

Di Stilio, V. S., C. Martin, A. F. Schulfer, and C. F. Connelly. 2009. An ortholog of MIXTA-like2 controls epidermal cell shape in flowers of *Thalictrum*. *New Phytologist* 183: 718–728.

Di Stilio, V. S., R. A. Kumar, A. M. Oddone, T. R. Tolkin, P. Salles, and K. McCarty. 2010. Virus-induced gene silencing as a tool for comparative functional studies in *Thalictrum*. *PLoS ONE* 5: e12064.

Di Stilio, V. S., N. C. LaRue, and A. M. Sullivan. 2013. Functional recapitulation of transitions in sexual systems by homeosis during the evolution of dioecy in *Thalictrum*. *Frontiers in Plant Science* 4: 487. https://doi.org/10.3389/fpls.2013.00487.

Dobritsa, A. A., Z. Lei, S. Nishikawa, E. Urbanczyk-Wochniak, D. V. Huhman, D. Preuss, and L. W. Sumner. 2010. *LAP5* and *LAP6* encode anther-specific proteins with similarity to chalcone synthase essential for pollen exine development in Arabidopsis. *Plant Physiologist* 153: 937–955.

El-Gebali, S., J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* 47: D427–D432.

Emms, D. M., and S. Kelly. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology* 20: 238.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30: 1575–1584.

Favaro, R., A. Pinyopich, R. Battaglia, M. Kooiker, L. Borghi, G. Ditta, M. F. Yanofsky, et al. 2003. MADS-box protein complexes control carpel and ovule development in *Arabidopsis*. *Plant Cell* 15: 2603–2611.

Figueroa, P., and J. Browse. 2015. Male sterility in *Arabidopsis* induced by overexpression of a MYC5-SRDX chimeric repressor. *Plant Journal* 81: 849–860.

Filiault, D. L., E. S. Ballerini, T. Mandáková, G. Aköz, N. J. Derieg, J. Schmutz, J. Jenkins, et al. 2018. The *Aquilegia* genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* 7: e36426.

Galimba, K. D., and V. S. Di Stilio. 2015. Sub-functionalization to ovule development following duplication of a floral organ identity gene. *Developmental Biology* 405: 158–172.

Galimba, K. D., T. R. Tolkin, A. M. Sullivan, R. Melzer, G. Theissen, and V. S. Di Stilio. 2012. Loss of deeply conserved C-class floral homeotic gene function and C- and E-class protein interaction in a double-flowered ranunculid mutant. *Proceedings of the National Academy of Sciences, USA* 109: E2267–E2275.

Galimba, K. D., J. Martínez-Gómez, and V. S. Di Stilio. 2018. Gene duplication and transference of function in the paleo*AP3* lineage of floral organ identity genes. *Frontiers in Plant Science* 9: 334.

Gou, J.-Y., L. M. Miller, G. Hou, X.-H. Yu, X.-Y. Chen, and C.-J. Liu. 2012. Acetylesterase-mediated deacetylation of pectin impairs cell elongation, pollen germination, and plant reproduction. *Plant Cell* 24: 50–65.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644–652.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072–1075.

Haas, B. J., S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology* 9: R7.

Hileman, L. C., and P. Cubas. 2009. An expanded evolutionary role for flower symmetry genes. *Journal of Biology* 8: 90.

Hoff, K. J., A. Lomsadze, M. Borodovsky, and M. Stanke. 2019. Whole-genome annotation with BRAKER, 65–95. *In* M. Kollmar [ed.], Gene prediction. Methods in Molecular Biology, vol. 1962. Humana Press, New York, New York, USA.

Hoopes, G. M., J. P. Hamilton, J. Kim, D. Zhao, K. Wiegert-Rininger, E. Crisovan, and C. R. Buell. 2017. Genome assembly and annotation of the medicinal plant *Calotropis gigantea*, a producer of anticancer and antimalarial cardenolides. *G3 Genes, Genomes, Genetics* 8: 385–391.

Humphrey, R. P., and A. G. Ossip-Drahos. 2018. Selection imposed by pollination mode minimally influences evolution of pollen morphology in *Thalictrum* (Ranunculaceae). *International Journal of Plant Sciences* 179: 688–696.

Huson, D. H., B. Albrecht, C. Bağcı, I. Bessarab, A. Górska, D. Jolic, and R. B. H. Williams. 2018. MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct* 13: 6.

Ishiguro, S., Y. Nishimori, M. Yamada, H. Saito, T. Suzuki, T. Nakagawa, H. Miyake, et al. 2010. The *Arabidopsis FLAKY POLLEN1* gene encodes a 3-hydroxy-3-methylglutaryl-coenzyme A synthase required for development of tapetum-specific organelles and fertility of pollen grains. *Plant and Cell Physiology* 51: 896–911.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, et al. 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30: 1236–1240.

Kalvari, I., J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, et al. 2018. Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research* 46: D335–D342.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 9: 59.

Laetsch, R., and M. L. Blaxter. 2017. BlobTools: Interrogation of genome assemblies. *F1000 Research* 6: 1287.

Lagesen, K., P. Hallin, E. A. Rødland, H.-H. Stærfeldt, T. Rognes, and D. W. Ussery. 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* 35: 3100–3108.

Lane, A. K., M. M. Augustin, S. Ayyampalayam, A. Plant, S. Gleissberg, V. S. Di Stilio, C. W. Depamphilis, et al. 2018. Phylogenomic analysis of Ranunculales resolves branching events across the order. *Botanical Journal of the Linnean Society* 187: 157–166.

Langmead, B., and S. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.

Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, and M. Przeworski. 2012. Revisiting an old riddle: What determines genetic diversity levels within species? *PLoS Biology* 10: e1001388.

Leroux, C., S. Bouton, M.-C. Kiefer-Meyer, T. N. Fabrice, A. Mareck, S. Guénin, F. Fournet, et al. 2015. PECTIN METHYLESTERASE48 is involved in Arabidopsis pollen grain germination. *Plant Physiology* 167: 367–380.

Lopez, L., E. M. Wolf, J. C. Pires, P. P. Edger, and M. A. Koch. 2017. Molecular resources from transcriptomes in the Brassicaceae family. *Frontiers in Plant Science* 8: 1488.

Louvet, R., E. Cavel, L. Gutierrez, S. Guénin, D. Roger, F. Gillet, F. Guerineau, and J. Pelloux. 2006. Comprehensive expression profiling of the pectin methylesterase gene family during silique development in *Arabidopsis thaliana*. *Planta* 224: 782–791.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, et al. 2012a. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1: 18.

Luo, G., H. Gu, J. Liu, and L.-J. Qu. 2012b. Four closely-related RING-type E3 ligases, APD1–4, are involved in pollen mitosis II regulation in *Arabidopsis*. *Journal of Integrative Plant Biology* 54: 814–827.

Marçais, G., and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27(6): 764–770.

Matasci, N., L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, et al. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3: 17.

Meinke, D. W. 2020. Genome-wide identification of *EMBRYO-DEFECTIVE* (*EMB*) genes required for growth and development in *Arabidopsis*. *New Phytologist* 226: 306–325.

Morales-Briones, D. F., T. Arias, V. S. Di Stilio, and D. C. Tank. 2019. Chloroplast primers for clade-wide phylogenetic studies of *Thalictrum*. *Applications in Plant Sciences* 7(10): 11294.

Nawrocki, E. P., and S. R. Eddy. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29: 2933–2935.

Niu, Q.-K., Y. Liang, J.-J. Zhou, X.-Y. Dou, S.-C. Gao, L.-Q. Chen, X.-Q. Zhang, and D. Ye. 2013. Pollen-expressed transcription factor 2 encodes a novel plant-specific TFIIB-related protein that is required for pollen germination and embryogenesis in *Arabidopsis*. *Molecular Plant* 6: 1091–1108.

Park, S., R. K. Jansen, and S. Park. 2015. Complete plastome sequence of *Thalictrum coreanum* (Ranunculaceae) and transfer of the *rpl32* gene to the nucleus in the ancestor of the subfamily Thalictroideae. *BMC Plant Biology* 15: 40.

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 14: 417–419.

Pérez-Rodríguez, P., D. M. Riaño-Pachón, L. G. G. Corrêa, S. A. Rensing, B. Kersten, and B. Mueller-Roeber. 2010. PlnTFDB: Updated content and new features of the plant transcription factor database. *Nucleic Acids Research* 38: D822–D827.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33: 290–295.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11: 1432.

Sauret-Güeto, S., K. Schiessl, A. Bangham, R. Sablowski, and E. Coen. 2013. *JAGGED* controls *Arabidopsis* petal growth and shape by interacting with a divergent polarity field. *PLoS Biology* 11: e1001550.

Shyu, S.-Y., and J.-M. Hu. 2013. Comparison of six DNA extraction procedures and the application of plastid DNA enrichment methods in selected non-photosynthetic plants. *Taiwania* 58: 268–274.

Simon, M. K., D. J. Skinner, T. L. Gallagher, and C. S. Gasser. 2017. Integument development in *Arabidopsis* depends on interaction of YABBY protein INNER NO OUTER with coactivators and corepressors. *Genetics* 207: 1489–1500.

Smith-Unna, R., C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly. 2016. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research* 26: 1134–1144.

Sohlberg, J. J., M. Myrenås, S. Kuusk, U. Lagercrantz, M. Kowalczyk, G. Sandberg, and E. Sundberg. 2006. *STY1* regulates auxin homeostasis and affects apical–basal patterning of the *Arabidopsis* gynoecium. *Plant Journal* 47: 112–123.

Soza, V. L., J. Brunet, A. Liston, P. Salles Smith, and V. S. Di Stilio. 2012. Phylogenetic insights into the correlates of dioecy in meadow-rues (*Thalictrum*, Ranunculaceae). *Molecular Phylogenetics and Evolution* 63: 180–192.

Soza, V. L., K. L. Haworth, and V. S. Di Stilio. 2013. Timing and consequences of recurrent polyploidy in meadow-rues (*Thalictrum*, Ranunculaceae). *Molecular Biology and Evolution* 30: 1940–1954.

Soza, V. L., C. D. Snelson, K. D. Hewett Hazelton, and V. S. Di Stilio. 2016. Partial redundancy and functional specialization of E-class *SEPALLATA* genes in an early-diverging eudicot. *Developmental Biology* 419: 143–155.

Szécsi, J., C. Joly, K. Bordji, E. Varaud, J. M. Cock, C. Dumas, and M. Bendahmane. 2006. *BIGPETALp*, a *bHLH* transcription factor is involved in the control of *Arabidopsis* petal size. *EMBO Journal* 25: 3912–3920.

The Tomato Genome Consortium, S. Sato, S. Tabata, H. Hirakawa, E. Asamizu, K. Shirasawa, S. Isobe, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635–641.

Thiel, T., W. Michalek, R. Varshney, and A. Graner. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* 106: 411–422.

Timerman, D., and S. C. H. Barrett. 2018. Divergent selection on the biomechanical properties of stamens under wind and insect pollination. *Proceedings of the Royal Society B, Biological Sciences* 285: 20182251.

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, and M. C. Schatz. 2017. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 33(14): 2202–2204.

Wang, Y., W.-Z. Zhang, L.-F. Song, J.-J. Zou, Z. Su, and W.-H. Wu. 2008. Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in *Arabidopsis*. *Plant Physiologist* 148: 1201–1211.

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* 35: 543–548.

Zdobnov, E. M., F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simão, P. Ioannidis, M. Seppey, et al. 2017. OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research* 45: D744–D749.

Zeng, L., N. Zhang, Q. Zhang, P. K. Endress, J. Huang, and H. Ma. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytologist* 214: 1338–1354.

Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29: 2669–2677.

APPENDIX 1. Voucher and source information for *Thalictrum* species in this study.

Species	DNA code	Usage	Voucher accession no. (Herbarium)	Locality	SRA accession no.
T. hernandezii Tausch ex J. Presl	ThWT441	RNA-seq analysis	V. Di Stilio 119 (WTU)	Cultivated from wild seed, Huitzila, Mexico Liston1215	SRR6869420, SRR6869419
T. thalictroides (L.) A. J. Eames & B. Boivin	TtWT478	Whole genome assembly RNA-seq analysis	V. Di Stilio 124 (WTU)	Cultivated from nursery material (MI, USA)	SRR6869426, SRR6869425, SRR6869418 Transcriptome: SRR6869422 GBVZ
T. thalictroides	TtWT964	Whole genome assembly	Unvouchered	Cultivated from nursery material (NC, USA)	SRR6869424, SRR6869423, SRR6869421

Note: MI = Michigan; NC = North Carolina; WTU = University of Washington Herbarium.

APPENDIX 2. Summary statistics and accession numbers for *Thalictrum thalictroides* genome sequences. Data are shown for two genetically distinct individuals (WT964 and WT478), each with three next-generation sequencing libraries of different insert sizes.

		No. of sequenced fragments						
Sample	Library insert size	Raw reads	Trimmed reads	Percent (from raw) after Trimming	Clean reads[a]	Percent (from raw) after cleaning	Clean reads (Gbp)	SRA accession no.
WT964	170	11,429,928	10,146,420	88.8	9,758,045	85.4	1.684	SRX3823744
WT964	500	11,887,778	9,155,669	77.0	8,910,174	75.0	1.497	SRX3823741
WT964	800	11,826,149	9,253,079	78.2	9,130,960	77.2	1.524	SRX3823742
WT478	170	11,484,802	10,191,821	88.7	6,601,641	57.5	1.290	SRX3823739
WT478	500	11,477,882	10,414,934	90.7	7,727,941	67.3	1.502	SRX3823740
WT478	800	11,721,131	9,012,622	76.9	6,977,136	59.5	1.350	SRX3823747

[a]Clean reads are the number of reads remaining after quality-trimming and contaminant removal (see Appendices S1 and S2 for a list of contaminants).