BMC Genomics

RESEARCH

Open Access

A de novo long-read genome assembly of the sacred datura plant (*Datura wrightii*) reveals a role of tandem gene duplications in the evolution of herbivore-defense response

Jay K. Goldberg[1*], Aaron Olcerst[2], Michael McKibben[1], J. Daniel Hare[2], Michael S. Barker[1] and Judith L. Bronstein[1]

Abstract

The sacred datura plant (Solanales: Solanaceae: *Datura wrightii*) has been used to study plant–herbivore interactions for decades. The wealth of information that has resulted leads it to have potential as a model system for studying the ecological and evolutionary genomics of these interactions. We present a de novo *Datura wrightii* genome assembled using PacBio HiFi long-reads. Our assembly is highly complete and contiguous (N50 = 179Mb, BUSCO Complete = 97.6%). We successfully detected a previously documented ancient whole genome duplication using our assembly and have classified the gene duplication history that generated its coding sequence content. We use it as the basis for a genome-guided differential expression analysis to identify the induced responses of this plant to one of its specialized herbivores (Coleoptera: Chrysomelidae: *Lema daturaphila*). We find over 3000 differentially expressed genes associated with herbivory and that elevated expression levels of over 200 genes last for several days. We also combined our analyses to determine the role that different gene duplication categories have played in the evolution of *Datura*-herbivore interactions. We find that tandem duplications have expanded multiple functional groups of herbivore responsive genes with defensive functions, including UGT-glycosyltranserases, oxidoreductase enzymes, and peptidase inhibitors. Overall, our results expand our knowledge of herbivore-induced plant transcriptional responses and the evolutionary history of the underlying herbivore-response genes.

Keywords Datura wrightii, Herbivory, Genomics, Differential expression, Tandem duplications

Introduction

Datura wrightii (Solanales: Solanaceae) has begun to serve as a model system for research into the ecology and evolution of various plant traits, including both physical and chemical defenses [1, 2], tolerance to herbivory [3], floral phenotypes [4, 5], life histories [6, 7], and ontogenetic changes in defense production throughout an individual plant's lifetime [8, 9]. This is due to the wealth of ecological knowledge already gathered regarding this common plant and the specialist insects with which it interacts [10, 11]. Furthermore, field-based studies have already provided some insights into the evolutionary processes maintaining a trichome dimorphism in naturally occurring populations [2, 12]. As such, the publication of a reference genome for this species would accelerate research into both the molecular mechanisms governing the expression of plant traits and their evolutionary trajectories in a rapidly changing environment.

Here, we present a de novo genome assembly for *D. wrightii*, generated using highly accurate long-reads (PacBio HiFi) [13]. Our assembly is highly complete

*Correspondence:
Jay K. Goldberg
jaykgold@arizona.edu
Full list of author information is available at the end of the article

and contiguous when compared to recent assemblies of closely related species [14–16]. To demonstrate the utility of our genome to future research, we conducted two downstream analyses with it: 1) an assessment of the duplication history of coding regions with our assembly and 2) a differential expression study examining herbivore-induced transcriptional changes. We further use the results of these analyses to determine the role that different categories of gene duplications have had on the herbivore-induced responses of this species.

Gene and whole genome duplications (WGDs) have broad implications for evolutionary processes including the generation of novel traits [17, 18]. WGDs have occurred numerous times throughout the diversification of angiosperms [19]; and these 'ancient' WGDs have been shown to underlie the arms-race dynamic occurring between plants and herbivores [20]. There is widespread evidence of an ancient hexaploidy event that occurred early in the evolution of the Solanaceae [19]. As such we used two different analyses to infer and characterize this ancient WGD in our genome to both detect its presence and assess the ancestral ploidy level of the lineage.

We further use our assembly to conduct a differential expression study into the induced responses of *D. wrightii* to one of its specialist herbivores, the three-lined potato beetle (*Lema daturaphila*) [10, 11]. This analysis utilized older data, generated roughly a decade ago [21]. Using both pairwise and time course analyses, we identified thousands of genes that are significantly differentially expressed when plants are under attack by one of their closest natural enemies, and further show that many of them remain upregulated for multiple days. We conclude by using an over-representation analysis approach to determine the molecular functions that various gene duplication categories have expanded.

Methods

Initial sample collection and preparation, and sequencing

Datura wrightii seeds were originally collected in June 2016 from a wild plant growing at the intersection of Portal Rd. and Foothills Rd. in Portal, Arizona, USA (see https://www.inaturalist.org/observations/156947070 for more details). A cohort of seeds from this collection was germinated in May 2021. Leaf tissue collected for genome sequencing originated from a single individual. Tissue samples were flash frozen and ground under liquid nitrogen immediately before storage at -80C. Samples for differential expression analysis were collected from two experiments conducted at UC Riverside in 2012/13 [21]. Both studies were conducted with the MVV6 line, a backcrossed line originating from seeds originally collected from Moreno Valley, CA that induces production of defenses in response to herbivore attack [22].

Nucleic acid extractions and sequencing strategies

DNA extraction, SMRT bell library preparation, and sequencing (PacBio HiFi, Pacific Bioscience, San Francisco, CA, USA) were performed by the Arizona Genomics Institute (University of Arizona, Tucson, AZ, USA). High molecular weight DNA was extracted from young leaves using the protocol of Doyle and Doyle [23] with minor modifications. Flash-frozen young leaves were ground to a fine powder in a frozen mortar with liquid nitrogen followed by very gentle extraction in 2% CTAB buffer (that included proteinase K, PVP-40 and beta-mercaptoethanol) for 30min to 1h at 50 °C. After centrifugation, the supernatant was gently extracted twice with 24:1 chloroform:isoamyl alcohol. The upper (aqueous) phase was then removed and 1/10th volume 3 M NaAc was added, gently mixed, and then had DNA precipitated with iso-propanol. DNA was collected by centrifugation, washed with 70% ethanol, air dried for 20 min and dissolved thoroughly in elution buffer at room temperature followed by RNAse treatment. DNA purity was measured with Nanodrop, DNA concentration measured with Qubit HS kit (Invitrogen) and DNA size was validated by Femto Pulse System (Agilent).

DNA was sheared to an appropriate size range (10–20 kb) using Megaruptor 3 (Diagenode) followed by Ampure bead purification. The sequencing library was constructed following manufacturers protocols using SMRTbell Prep kit 3.0. The final library was size selected on a Pippin HT (Sage Science) using S1 marker with a 10–25 kb size selection. The recovered final library was quantified with Qubit HS kit (Invitrogen) and size checked on Femto Pulse System (Agilent). The final library was prepared for sequencing with PacBio Sequel II Sequencing kit 2.0 for HiFi library, loaded on 8M SMRT cells, and sequenced in CCS mode for 30 h.

RNA was extracted from root and bud tissue using a ZYMO (Irvine, CA, USA) direct-zol miniprep kit (Cat. # R2050) and sequenced using NovaSeq (Illumina, San Diego, CA, USA) paired-end (150bp) sequencing performed by Novogene (Sacramento, CA, USA). RNA for differential expression was extracted and sequenced as described in Olcerst (2017) [21].

Genome assembly/annotation

CCS output (ie: HiFi reads; 3952061 reads; 65.26Gb total; mean length = 16524) were assembled using hifiasm-0.16.0 [24] with default settings. Jellyfish v2.2.10 [25] was used for kmer counting (kmer size = 101bp) before using the GenomeScope2.0 web portal [26] to estimate genome size (Fig. S1). Genome quality was examined using Bandage v0.8.1 [27], BUSCO v5.1.3 (odb10_solanales; Fig. 1) [28], the blobtools v1.1 pipeline (Fig. S2) [29] employing minimap v2-2.24 [30] for alignment and the nt

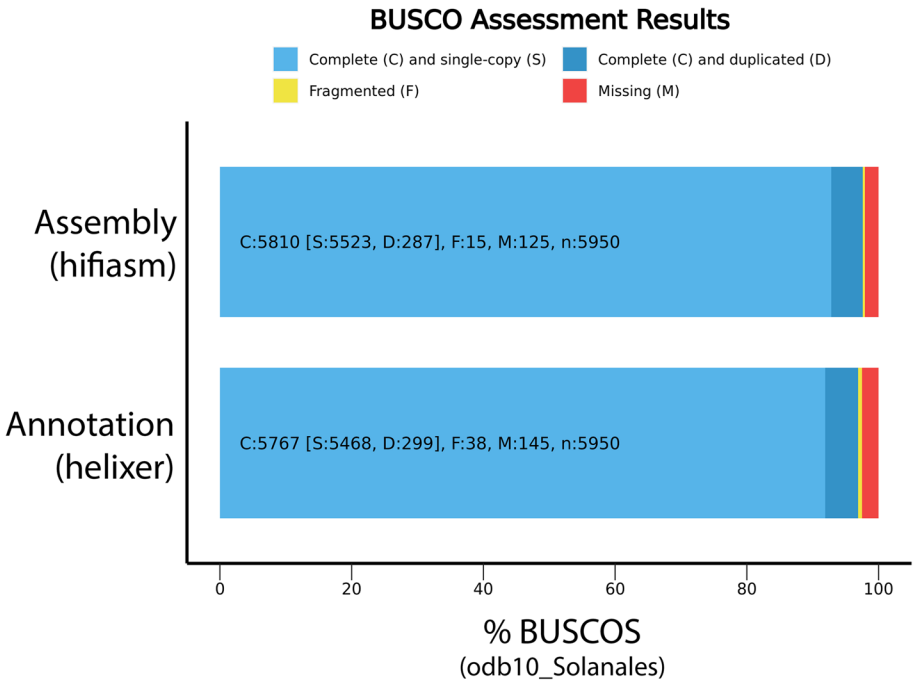

Fig. 1 Bar charts showing the gene content represented in both our assembly (top bar) and our annotation (bottom bar). Annotation assessment was run in proteome mode

sequence database for taxonomic identification (BLAST 2.13.0) [31], and Inspector [32]. Inspector was also used for error correction/polishing (Table S1). Repetitive elements were identified using RepeatModeler v2.0.1 and RepeatMasker v4.1.0 [33] (Table S2). Structural gene annotation was carried out using the Helixer v0.3.1 algorithm pipeline [34, 35] using the pre-made land plant training dataset. Functional annotation was done using InterProScan v5.45–80.0 [36] and blastp (using blast v2.13.0) [31] comparisons to the UniProt-Swissprot database [37]. Annotation files were combined into a single gff using the manage_functional_annotation.pl script in the AGAT v 1.2.0 toolkit [38]. Detailed annotation statistics are shown in Table S3.

RNA-seq read alignments

Raw RNA-seq reads were aligned to the annotated reference genome and counted using STAR [39] with the default parameters. STAR was selected for its flexible alignment parameters (up to 10 mismatches per read, so long as total mismatches do not exceed 30% of their length), and automatic read counting functionality when given an annotation (a.gff3 file in our case). Reads aligned to our reference genome at an average rate of 80% per library, with a multi-mapping rate of less than 10% for all libraries. An average of 0.3% of reads from each library was removed for mapping to over 10 loci. Alignment details can be found in Table S4. Untransformed read counts were passed to DESeq2 [40] in R [41] for statistical analysis of differential gene expression.

Gene duplication categorization

We used two programs, MCScanX [42] and frackify [43], to investigate the history of gene duplication events that generated our observed *D. wrightii* gene content and test for signatures of ancient WGDs. Using MCScanX we made inter- and intraspecific syntenic comparisons of the *D. wrightii* and *Ipomoea purpurea* reference genomes [42]. We visualized the syntenic depth ratio of each collinear loci in the interspecific comparison using matplotlib [44]. Additionally we appended Ka/Ks values for each collinear gene pain in both the inter- and intraspecific syntenic comparison using the add_kaks_to_synteny.pl script available through MCScanX [42]. The distribution of Ks values for each comparison were transformed into density functions with the nparam_density and gaussian_kde functions from the Numpy and Scipy python libraries [45]. Maxima of each WGD and ortholog divergence peak in each density distribution were found with the find_peaks function from the Scipy python library [45]. We then used the syntenic and Ks inferences as input for Frackify to identify paleologs in the D. wrightii reference genome [43]. Finally, we classified other paralogs in the genome as tandem, dispersed, proximal, and segmental duplicates using duplicate_gene_classifier available through MCScanX [42].

Differential expression experiment 1: pairwise comparison & gene set enrichment analysis

Adults and larvae of *L. daturaphila both* feed on the leaves of *D. wrightii.* Both life stages remove leaf tissue in irregular holes between the veins (Hare & Elle 2002). The overall lifecycle of the beetle is about a month, and the duration of its larval period is about a week, depending upon temperature. Many generations of *L. daturaphila* are produced over the nine-month growing season of *D. wrightii.* Our pairwise experiment consisted of three replicates of control/ *L. daturaphila*-induced leaf samples ($N=6$ total RNAseq libraries) collected from greenhouse grown plants in May 2012. Samples were collected after *L. daturaphila* larvae had been allowed to feed for 24h. Additional details regarding growing conditions and sample collection can be found in Olcerst [21]. We used the Wald test in DESeq2 [40] to test for pairwise differences between control and *L. daturaphila*-challenged samples while accounting for individual differences between sample pairs. We used a significance cutoff of $P_{adj}=0.05$ for all gene-wise analyses without any fold-change cutoff for differential expression. We then performed a gene set enrichment analysis (GSEA) on the results of our pairwise experiment using the ClusterProfiler 4.0 package [46]. Each GO ontology (biological processes, molecular functions, and cellular component) was analyzed separately. We further separated each ontology into separate up- and down-regulated gene lists as prior studies have found this approach to be more robust than grouping all DEGs [47].

Differential expression experiment 2: time course

The quantities and composition of volatile compounds induced in *D. wrightii* by *L. daturaphila* both vary with the time after herbivore damage [48]. This time course study asked how the pattern of gene induction might also vary over time since induction. Our experiment consisted of samples taken at 5 timepoints during larval herbivory (0h [before treatment], 12h, 24h, 48h, 96h; 6 per timepoint with a balanced design; except 24h where $N=5$, 3 control, 2 *L. daturaphila*-induced). We used a likelihood ratio test (LRT) approach to analyze these data. Our full model consisted of timepoint, treatment type, and the interaction term. The reduced model lacked the interaction term; thus, this analysis tested the significance of the interaction term specifically (i.e. genes for which expression over time differed by treatment). We then used the DEGreport package [49] to identify co-expressed gene groups within the time course dataset and visualize their expression levels over time.

Duplicated gene over-representation analyses

To determine the role of gene duplication categories in expanding functional gene groups, we used an over-representation analysis (ORA) approach via the ClusterProfiler package in R [46]. This tests for enriched GO terms within a subset of genes compared to a background set of genes. We only used the molecular function ontology for this analysis. We analyzed each gene duplication category as separate subsets against the total set of genes with assigned GO terms ($N=24939$). We also conducted this analysis using only the set of significantly differentially expressed genes with assigned GO-terms ($N=1954$).

Results

Genome assembly and annotation

Our de novo assembly is highly complete and contiguous (N50 = 179Mb; Longest contig = 202Mb; 1144 contigs; BUSCO odb10_Solanales Complete = 97.7% [Single-copy = 92.9%, Duplicated = 4.8%], Fragmented = 0.2%, Missing = 2.1%; Fig. 1; Table 1). The total length of our assembly is 2.2Gb, larger than other *Datura* assemblies [14, 15] and close to the prediction obtained from GenomeScope (2.086 Gb; Fig. S1). When compared to previous *Datura* assemblies, we find that our assembly contains far fewer contigs that are orders of magnitude longer (see Table 1). This improvement is due entirely to our use of PacBio HiFi reads, rather than previous generation sequencing technologies. Inspector analysis found several structural errors ($N=50$) and small-scale errors ($N=37211$; 16.63 per Mb) in our assembly. Polishing our assembly did not substantially reduce the number of structural errors ($N_{polished}=49$), but did reduce the number of small-scale errors in our assembly ($N_{polished}=2808$; 1.255 per Mb). Detailed output from inspector, before and after polishing, are found in Table S1. Blobtools analysis determined that no contamination was present in our assembly and that all contigs/reads were identified

Table 1 Summary of assembly statistics compared to those of *Datura stramonium* genomes previously published by De-La-Cruz et al. (2021) [15]

	D. wrightii **(Portal, AZ, USA)**	***D. stramonium*** **(Ticumán, MX)**	***D. stramonium*** **(Teotihuacán, MX)**
Total Size (Gb)	2.2	1.48	1.28
Contig number	1144	27915	30392
Largest contig (Mb)	202	3.13	2.11
N50	179 Mb	84.1 kb	58.2 kb
N75	121 Mb	44.5 kb	32.6 kb
L50	5	4557	5713
L75	10	10641	13166
BUSCO complete (%)	97.70	91	81.70
BUSCO db (odb10)	Solanales	Solanaceae	Solanaceae

as belonging to the Solanaceae (Fig. S2). 86.11% of the assembly is repetitive elements, primarily retroelements such as long terminal repeats (Table S2). Structural gene annotation using the helixer pipeline was able to preserve the majority of gene content represented in our assembly (BUSCO odb10_Solanales: Complete = 96.9%, [Single-copy = 91.9%, Duplicated = 5.0%], Fragmented = 0.6%, Missing = 2.5.8%; N_{genes} = 45500; Fig. 1). Of the structurally annotated genes, 42745 were functionally annotated; 37040 of which have inferred gene names from blastp against the Uniprot/Swissprot database [37]. Detailed annotation statistics are found in Table S3.

Gene duplication analyses

We used MCScanX [42] to test for the presence of the ancient Solanaceae WGD [16, 19] using a Ks analysis of the colinear gene blocks within our assembly. This analysis identified a large peak (Ks = 0.70), indicative of a rapid burst of gene duplications and consistent with the predictions of ancient WGD (Fig. 2A). We used a neighboring lineage of the Solanaceae that does not share the ancient WGD (Solanales: Convolvulaceae: *Ipomoea purpurea*) to calculate the syntenic depth of loci collineary to the *D. wrightii* genome assembly (Fig. 2B). A significant number of collinear loci (N = 972) had a syntenic depth ratio of 1:3, consistent with paleohexaploid ancestry. These inferences were further validated using Frackify to identify multi-copy paleologs retained from the Ks peak at 0.70. Among the paleologs retained in duplicate, 17% were retained in the triple copy state. We further used MCScanX to classify all the remaining paralogs within the largest 50 contigs of the genome assembly (Table 2; Table S8). We found that dispersed duplications are most common (N = 20523; Table 2), 75% of which were identified as paleologs by Frackify. Ancient WGDs are likely to have generated a significant amount of the gene content observed in the *D. wrightii* genome as well (N = 30714). All of these results are consistent with prior analyses that the ancient Solanaceae WGD was an ancient hexaploidization [16, 19].

Differential expression experiment 1: pairwise comparison & gene set enrichment analysis

Our pairwise analysis found that control and *L. datu-raphila*-induced plants had distinct gene expression profiles (Fig. 3A) driven by 3555 significantly differentially expression genes (DEGs; Fig. 3B; Table S5). Most

Table 2 Results of MCScanX gene duplication classifier function and GO-term over-representation analysis by duplication type. Gene counts refer to the total number of genes (ie: with or without assigned GO-terms)

Duplication type	Gene count		Enriched GO-terms (Molecular function)	
	Whole genome	DEGs only	Whole Genome	DEGs only
Singleton	4268	431	NA	NA
Dispersed	20523	1329	10	1
Proximal	2140	154	28	2
Tandem	4064	487	45	8
Segmental/WGD	14505	1196	13	1

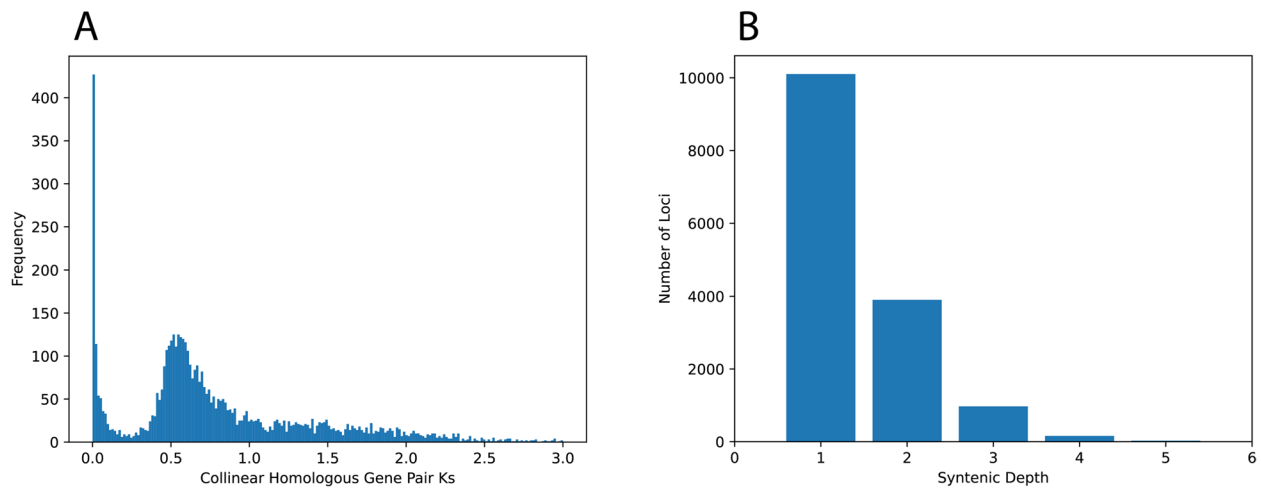

Fig. 2 **A** Results of MCSxanX showing the frequency of non-synonymous substitutions in homologous syntenic gene blocks. The large peak at ~ 0.5Ks indicates a large burst of gene duplications, consistent with the presence of an ancient whole-genome duplication. **B** Frackify results showing the syntenic depth of gene blocks in the *Datura wrightii* genome compared to *Ipopmoea purpurea*. The elevated number of triplicated syntenic blocks indicates an ancient hexaploidy state in *D. wrightii* that is not present in *I. purpurea*

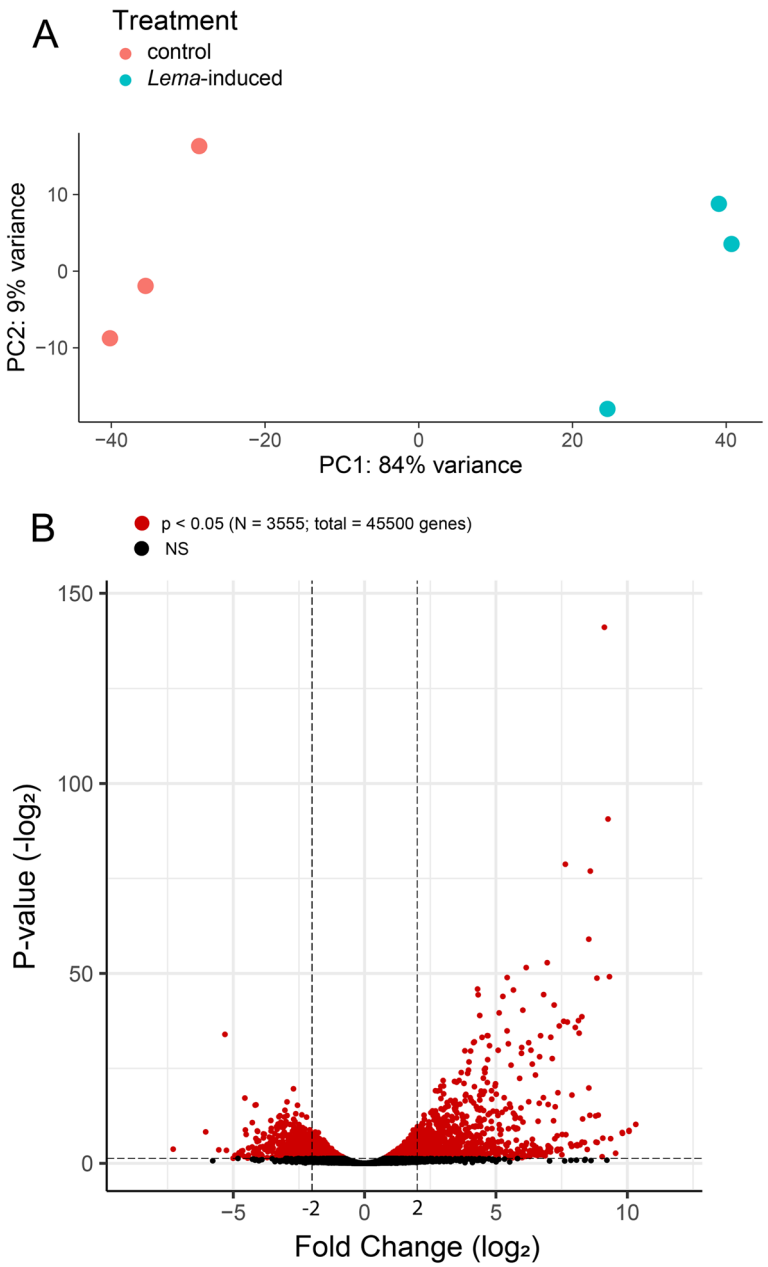

Fig. 3 Results of pairwise differential expression analysis using only the samples collected in May 2012. **A** Results of exploratory PCA to qualitatively screen for differences between treatment groups. There was a clear qualitative difference between the control and *L. daturaphila*-induced samples driven by PC1. **B** Volcano plot showing the relationship between significance and fold-change for each gene in our *D. wrightii* genome annotation. A change of 2/-2 is marked for reference but was not used as a cutoff for any analysis. Significant genes ($p < 0.05$) are shown in red, non-significant genes in black

DEGs ($N = 1985$) were up regulated in response to *L. daturaphila* feeding (Fig. 3B). Up-regulated genes had a far greater range of fold changes (Log_2FC max = 10.3) than down-regulated genes (Log_2FC min = -7.3). Functional annotation identified many genes as being involved with known herbivore-response processes such as jasmonic acid signaling and terpene synthesis (Table S5). Gene set enrichment analysis (GSEA) further confirmed this and demonstrated that several functional groups are differentially expressed in *L. daturaphila* -induced plants (Fig. 4; Table S6). Terpene synthase activity was notably enriched within our set of up-regulated DEGs alongside other known herbivore-response functions such as peptidase inhibitors, oxidoreductase enzymes,

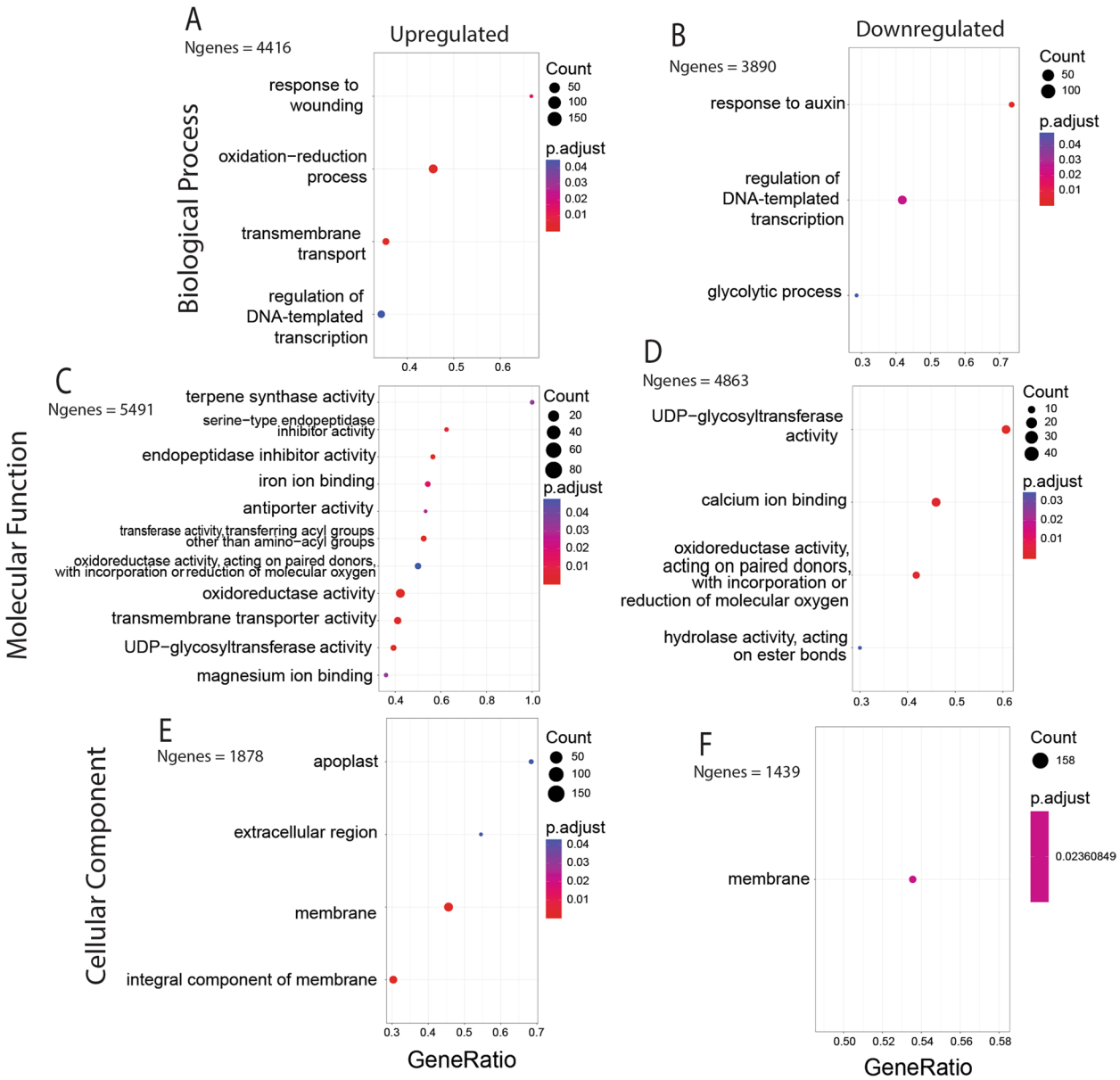

Fig. 4 Results of gene set enrichment analysis of pairwise DEGs. X-axes show the ratio of enriched genes vs the total count of genes sharing that GO term. Dot sizes represent the total number of genes sharing each GO term whereas dot colors represent the *p*-value (adjusted for multiple tests) of each term. Each GO ontology was analyzed separately for up- and down-regulated genes

(See figure on next page.)

Fig. 5 **A** Exploratory PCA of time course data. Although our samples did not form clear clusters, there are definitive patterns. The 24h, 48h, and 96h *L. daturaphila*-induced samples cluster in the top right, whereas the controls for these time points cluster with the 0h samples in the top left. Ranges for 0h and 12h samples do not overlap at all, suggesting that there may also be circadian cycles occurring. **B**, **C** Cluster plots showing the two largest gene clusters with similar expression patterns. These are the only gene clusters that showed a clear separation between treatments throughout the experiment time. Extended gene cluster results are shown in Fig. S3

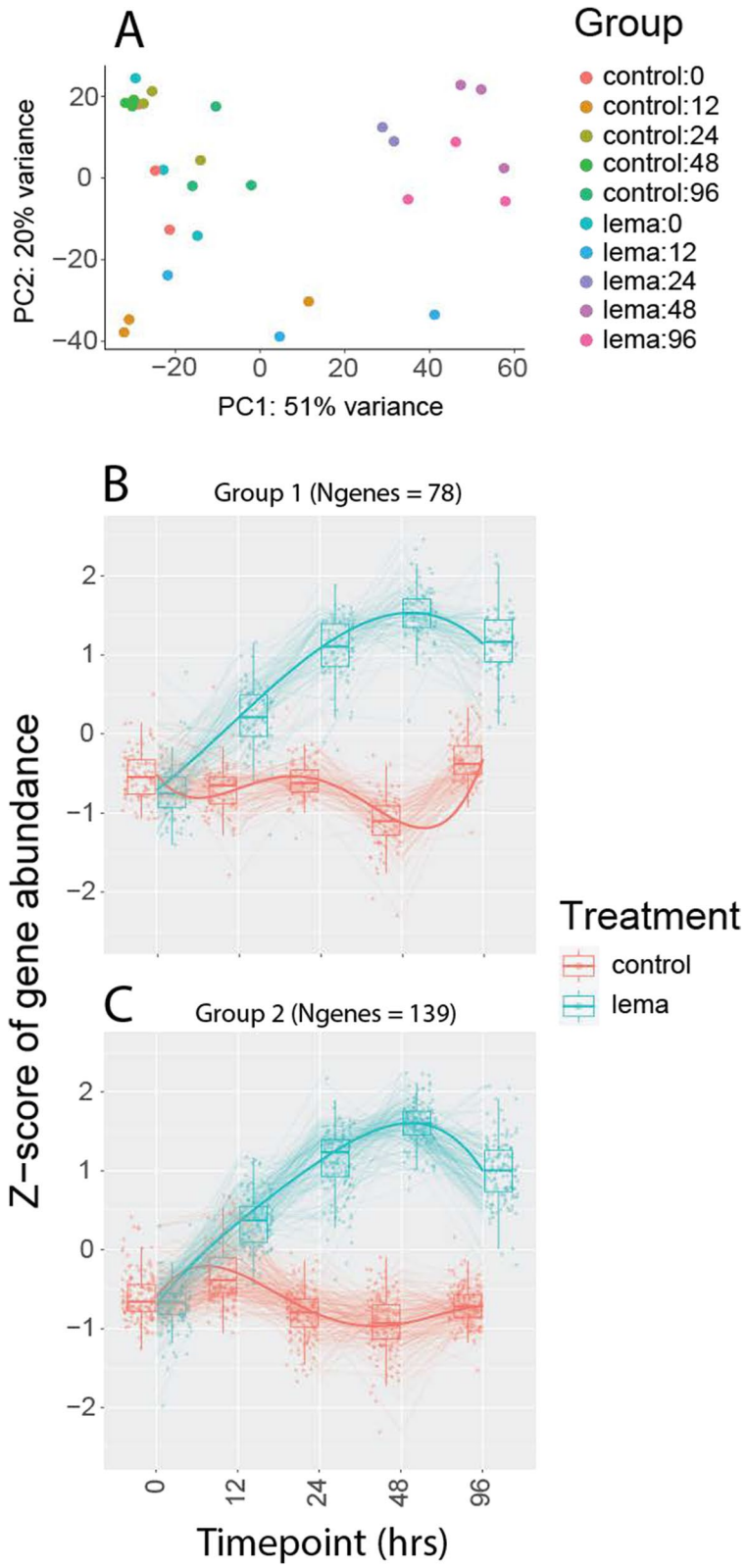

Fig. 5 (See legend on previous page.)

and UDP-glycosyltransferases (Fig. 4C). UDP-glycosyltransferase and oxidoreductase activity were also found to be significantly enriched in our set of down-regulated DEGs (Fig. 4D). Extracellular and apoplastic gene products were found to be enriched in up-regulated DEGs (Fig. 4E), consistent with previous findings in other plant–insect interactions [50]. Gene-wise results of pairwise DEG analysis are found in Table S5 whereas the full GSEA results and summary statistics are shown in Table S6.

Differential expression experiment 2: time course

Our time course analysis generated a shorter list of differentially expressed genes ($N=605$; Fig. 5; Table S7). Principle component analysis (PCA) indicated that *L. daturaphila*-induced samples at 24h, 48h, and 96h formed a distinct cluster (Fig. 5A). We also noticed a strong differentiation between 0h (pre-herbivory) and 12h post-herbivory treatments (Fig. 5A), regardless of treatment type (control or *L. daturaphila*-induced), which may be a sign of background circadian rhythms in gene expression [51]. Most of these genes ($N_{total}=495$) clustered into nine different groups based on their expression profiles over time (Fig. S3), but only the two largest clusters showed a clear difference between control and *L. daturaphila*-induced treatments (Fig. 5B, C). Genes in these clusters were up-regulated in the 24h, 48h, and 96h timepoints, consistent with our PCA results.

Gene duplication over-representation analysis

When assessing the functional groups enriched by different duplications, we found that tandem duplications have expanded the most molecular functions ($N=45$; Table 2; Table S9). Many of these functions – such as chitin binding, hydrolase enzymes, peptidase inhibitors, and terpene synthases – have known roles in defense against herbivores [50, 52]. When we analyzed only the enriched functions within our list of DEGs, we again found tandem duplications to have played the largest role by enriching eight molecular function GO-terms (Fig. 6; Table S10). Many of these functions were also enriched in our genome-wide analysis, including acyltransferase activity, oxidoreductase activity, and endopeptidase inhibitors.

Discussion

Our genome assembly produced using only PacBio HiFi reads is highly complete and contiguous. It is, however, important to note that it is not chromosome-scale or completely error-free. Despite great improvements in genome assembly algorithms and long-read sequencing accuracy, additional datasets (e.g. Hi-C) are generally required to produce chromosome-scale gapless assemblies without substantial structural errors [53]. That said, our largest contigs (>100Mb in length), may represent entire chromosomes and were more than sufficient to detect the presence of an ancient whole-genome duplication (Fig. 2A).

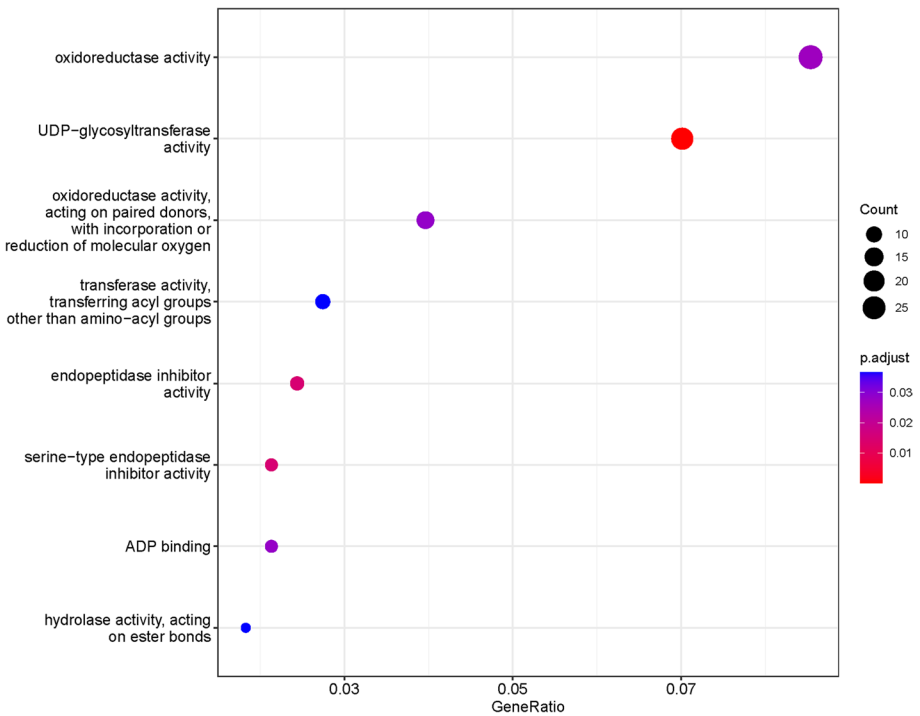

Fig. 6 Results of over-representation analysis examining the molecular function GO-terms of differentially expressed genes enriched by tandem duplications. Eight terms were found to be enriched in our dataset, more than any other duplication type (see Table 2 for details)

Furthermore, most of the gene content was represented, making our assembly suitable for use in mapping-based analysis such as differential expression studies.

Our differential expression analyses identified thousands of genes that are differentially expressed in *D. wrightii* plants that have been attacked by herbivores, hundreds of which remain highly expressed for multiple days. Many of the DEGs that we identified are known homologs of functionally important genes in other plant–insect systems [52], although it is possible that some of these genes may be induced by mechanical wounding and are not specifically involved in responses to herbivory. The lack of wounding controls in our study design is a limitation, although mechanisms to impose mechanical wounding that accurately mimics the small amount of tissue damaged per "bite," and accumulated wounding over time remains challenging for most systems (e.g., Mithofer et al. 2005 [54]). Nonetheless, our functional annotation allows us to confidently identify many DEGs as having functional roles in the production of herbivore-induced plant defenses. Furthermore, the presence of genes with long-term (e.g. multiple-day) changes to expression levels is likely indicative of a role in herbivore defense, as prior studies have found that differential expression of wounding-response genes to be most pronounced on short time scales (e.g. less than 24h) [55]. As such, our list of DEGs provides an excellent starting point for future studies that will lead to further insights into the molecular basis of ecological between plants and herbivorous insects over ecologically relevant time periods.

Gene and genome duplications have been shown to play an important role in the arms race between plants and insects [20], but overall, the molecular underpinnings of plant–insect co-evolution remain poorly resolved. By examining the role of various gene duplication types in expanding the functional repertoire of herbivore-induced genes, we have helped to close this gap in our understanding. The finding that tandem duplications are the most important is consistent with previous research that has found biosynthetic genes involved with the production of secondary metabolites to form distinct clusters generated via tandem duplication bursts [48]. Future studies of the duplication dynamics in both plant and insects may begin to unravel the role of biotic selective pressures in generating chemical defense diversity and the resources presented here will serve to accelerate this line of research.

Concluding remarks

In sum, we present a high quality long-read genome assembly for the sacred datura plant (*Datura wrightii*). We then analyze its gene duplication history and use it as the basis for a genome-guided analysis of herbivore-induced gene expression changes. Multiple tools supported the presence of a well-documented ancient whole-genome duplication event in this species; our analysis identified thousands of differentially expressed genes, some of which have known functional annotations based on comparison to existing reference genomes. We further show, using a GO-term enrichment approach, that tandem duplications have played an important role in the evolution of *D. wrightii*'s herbivore-responsive gene repertoire. Together, these data provide a valuable resource more broadly and will contribute to future studies of angiosperm evolution and the molecular basis of ecological interactions between plants and herbivorous insects.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09894-1.

Additional file 1: Figure S1. Results of kmer analysis to estimate genome size and heterozygosity from raw reads. **Figure S2.** Blobtools coverage plot. (left) the proportion of mapped and unmapped reads. (right) the taxanomic identification of mapped reads. All reads were identified as Solanaceae, indicating no contamination in our assembly. **Figure S3.** Extended results of cluster analysis showing the expression patterns of the nine largest gene clusters identified in our timecourse analysis. Only the two largest gene clusters (also shown in Fig. 5) showed clear differences between treatments over time.

Additional file 2: Table S1. Summary statistics obtained from Inspector before and after assembly polishing.

Additional file 3: Table S2. Repeat content statistics obtained from RepeatMasker.

Additional file 4: Table S3. Detailed summary statistics of genome annotation.

Additional file 5: Table S4. Mapping statistics of RNA-seq reads.

Additional file 6: Table S5. Detailed gene-wise results of pairwise differential expression analysis.

Additional file 7: Table S6. Detailed results of gene set enrichment analysis.

Additional file 8: Table S7. Detailed results of time course analysis.

Additional file 9: Table S8. Results of duplication classification analysis.

Additional file 10: Table S9. Detailed results of duplication type over-representation analysis (total gene content).

Additional file 11: Table S10. Detailed results of duplication type over-representation analysis (differentially expressed genes only).

Acknowledgements

We thank Professor Luciano Matzkin for his guidance throughout the process of working on the project. We would also like to thank the 2016 staff and volunteers of the Southwestern Research Station for assistance with field work and initial collection of seeds.

Authors' contributions

JKG, AO, JDH, and MSB conceived of the project and designed the experiments. JKG collected initial samples for genome assembly. AO and JDH collected samples for differential expression analysis. JKG and MM conducted computational analyses. JKG drafted the initial manuscript. All authors reviewed and contributed to the final version of this manuscript.

Funding
Funding was provided by an NSF postdoctoral research fellowship in biology to JKG (#2010772).

Availability of data and materials
Assembly and raw HiFi reads can be found on NCBI (BioProject: PRJNA966699; BioSample: SAMN34546691). Supporting datasets, which includes annotation, can be found at https://github.com/caterpillar-coevolution/Datura-wrightii-genome-project alongside scripts used for computational analyses. Raw transcriptome data will be available on NCBI GenBank (BioProject: PRJNA966699). Seeds from the Portal 2016 cohort can be made available on request. Reviewer link for un-released data: https://dataview.ncbi.nlm.nih.gov/object/PRJNA966699?reviewer=skgd96q1m91kppueafvea7va5p.

Declarations

Ethics approval and consent to participate
No collections were undertaken in protected regions or of protected species, thus no permits were required for fieldwork. In lieu of voucher specimens, observations of the initial plant, alongside the users who identified it, from which seeds were collected in 2016 can be found on iNaturalist (https://www.inaturalist.org/observations/156947070). All methods were carried out in accordance with relevant institutional, national, and international guidelines and legislation.

Consent to publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. [2]Department of Entomology, University of California Riverside, Riverside, CA, USA.

Received: 9 May 2023 Accepted: 11 December 2023
Published online: 02 January 2024

References

1. Hare JD, Walling LL. Constitutive and jasmonate-inducible traits of Datura wrightii. J Chem Ecol. 2006;32:29–47. https://doi.org/10.1007/s10886-006-9349-8.
2. Goldberg JK, Lively CM, Sternlieb SR, Pintel G, Hare JD, Morrissey MB, Delph LF. Herbivore-mediated negative frequency-dependent selection underlies a trichome dimorphism in nature. Evol Lett. 2020;4:83–90. https://doi.org/10.1002/evl3.157.
3. Bronstein JL, Huxman T, Horvath B, Farabee M, Davidowitz G. Reproductive biology of Datura wrightii: the benefits of a herbivorous pollinator. Ann Bot. 2009;103:1435–43. https://doi.org/10.1093/aob/mcp053.
4. Riffell JA, Alarcón R, Abrell L. Floral trait associations in hawkmoth-specialized and mixed pollination systems. Commun Integr Biol. 2008;1:6–8. https://doi.org/10.4161/cib.1.1.6350.
5. Elle E, Hare JD. Environmentally induced variation in floral traits affects the mating system in Datura wrightii. Funct Ecol. 2002;16:79–88.
6. Dupin J, Smith SD. Corrigendum to: "Phylogenetics of Datureae (Solanaceae), including description of the new genus Trompettia and re-circumscription of the tribe" [in Taxon 67: 359–375. 2018]. Taxon. 2019;68:419–419. https://doi.org/10.1002/tax.12056.
7. Dupin J, Smith SD. Phylogenetics of Datureae (Solanaceae), including description of the new genus Trompettia and re–circumscription of the tribe. Taxon. 2018;67:359–75. https://doi.org/10.12705/672.6.
8. van Dam N, Hare J, Elle E. Inheritance and distribution of trichome phenotypes in Datura wrightii. J Hered. 1999;90:220–7. https://doi.org/10.1093/jhered/90.1.220.
9. Kariñho-Betancourt E, Agrawal AA, Halitschke R, Núñez-Farfán J. Phylogenetic correlations among chemical and physical plant defenses change with ontogeny. New Phytol. 2015;206:796–806. https://doi.org/10.1111/nph.13300.
10. Goldberg JK, Sternlieb SR, Pintel G, Delph LF. Observational evidence of herbivore-specific associational effects between neighboring conspecifics in natural, dimorphic populations of Datura wrightii. Ecol Evol. 2021;11:5547–61. https://doi.org/10.1002/ece3.7454.
11. Zhang J, Komail Raza SA, Wei Z, Keesey IW, Parker AL, Feistel F, Chen J, Cassau S, Fandino RA, Grosse-Wilde E, Dong S, Kingsolver J, Gershenzon J, Knaden M, Hansson BS. Competing beetles attract egg laying in a hawkmoth. Curr Biol. 2022;32:861–869.e8. https://doi.org/10.1016/j.cub.2021.12.021.
12. Hare JD, Elle E, van Dam NM. Costs of glandular trichomes in Datura Wrightii: a three-year study. Evolution. 2003;57:793–805. https://doi.org/10.1111/j.0014-3820.2003.tb00291.x.
13. Lang D, Zhang S, Ren P, Liang F, Sun Z, Meng G, Tan Y, Li X, Lai Q, Han L, Wang D, Hu F, Wang W, Liu S. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. GigaScience. 2020;9:giaa123.
14. Rajewski A, Carter-House D, Stajich J, Litt A. Datura genome reveals duplications of psychoactive alkaloid biosynthetic genes and high mutation rate following tissue culture. BMC Genomics. 2021;22:201. https://doi.org/10.1186/s12864-021-07489-2.
15. De-la-Cruz IM, Hallab A, Olivares-Pinto U, Tapia-López R, Velázquez-Márquez S, Piñero D, Oyama K, Usadel B, Núñez-Farfán J. Genomic signatures of the evolution of defence against its natural enemies in the poisonous and medicinal plant Datura stramonium (Solanaceae). Sci Rep. 2021;11:882. https://doi.org/10.1038/s41598-020-79194-1.
16. Xu S, Brockmöller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, Zhou W, Kreitzer C, Stanke M, Tang H, Lyons E, Pandey P, Pandey SP, Timmermann B, Gaquerel E, Baldwin IT. Wild tobacco genomes reveal the evolution of nicotine biosynthesis. Proc Natl Acad Sci. 2017;114:6133–8. https://doi.org/10.1073/pnas.1700073114.
17. Schranz EM, Mohammadin S, Edger PP. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. Curr Opin Plant Biol Genome Stud Mol Genet. 2012;15:147–53. https://doi.org/10.1016/j.pbi.2012.03.011.
18. Selmecki AM, Maruvka YE, Richmond PA, Guillet M, Shoresh N, Sorenson AL, De S, Kishony R, Michor F, Dowell R, Pellman D. Polyploidy can drive rapid adaptation in yeast. Nature. 2015;519:349–52. https://doi.org/10.1038/nature14187.
19. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S, Porsch M, Quint M, Rensing SA, Soltis DE, Soltis PS, Stevenson DW, Ullrich KK, Wickett NJ, DeGironimo L, Edger PP, Jordon-Thaden IE, Joya S, Liu T, Melkonian B, Miles NW, Pokorny L, Quigley C, Thomas P, Villarreal JC, Augustin MM, Barrett MD, Baucom RS, Beerling DJ, Benstein RM, Biffin E, Brockington SF, Burge DO, Burris JN, Burris KP, Burtet-Sarramegna V, Caicedo AL, Cannon SB, Çebi Z, Chang Y, Chater C, Cheeseman JM, Chen T, Clarke ND, Clayton H, Covshoff S, Crandall-Stotler BJ, Cross H, dePamphilis CW, Der JP, Determann R, Dickson RC, Di Stilio VS, Ellis S, Fast E, Feja N, Field KJ, Filatov DA, Finnegan PM, Floyd SK, Fogliani B, García N, Gâteblé G, Godden GT, Goh F, Qi Y, Greiner S, Harkess A, Heaney JM, Helliwell KE, Heyduk K, Hibberd JM, Hodel RGJ, Hollingsworth PM, Johnson MTJ, Jost R, Joyce B, Kapralov MV, Kazamia E, Kellogg EA, Koch MA, Von Konrat M, Könyves K, Kutchan TM, Lam V, Larsson A, Leitch AR, Lentz R, Li F-W, Lowe AJ, Ludwig M, Manos PS, Mavrodiev E, McCormick MK, McKain M, McLellan T, McNeal JR, Miller RE, Nelson MN, Peng Y, Ralph P, Real D, Riggins CW, Ruhsam M, Sage RF, Sakai AK, Scascitella M, Schilling EE, Schlösser E-M, Sederoff H, Servick S, Sessa EB, Shaw AJ, Shaw SW, Sigel EM, Skema C, Smith AG, Smithson A, Stewart CN, Stinchcombe JR, Szövényi P, Tate JA, Tiebel H, Trapnell D, Villegente M, Wang C-N, Weller SG, Wenzel M, Weststrand S, Westwood JH, Whigham DF, Wu S, Wulff AS, Yang Y, Zhu D, Zhuang C, Zuidof J, Chase MW, Pires JC, Rothfels CJ, Yu J, Chen C, Chen L, Cheng S, Li J, Li R, Li X, Lu H, Ou Y, Sun X, Tan X, Tang J, Tian Z, Wang F, Wang J, Wei X, Xu X, Yan Z, Yang F, Zhong X, Zhou F, Zhu Y, Zhang Y, Ayyampalayam S, Barkman TJ, Nguyen N, Matasci N, Nelson DR, Sayyari E, Wafula EK, Walls RL, Warnow T, An H, Arrigo N, Baniaga AE, Galuska S, Jorgensen SA, Kidder TI, Kong H, Lu-Irving P, Marx HE, Qi X, Reardon CR, Sutherland BL, Tiley GP, Welles SR, Yu R, Zhan S, Gramzow L, Theißen G, Wong GK-S, One Thousand Plant

Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574:679–85. https://doi.org/10.1038/s41586-019-1693-2.
20. Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, Hofberger JA, Smithson A, Hall JC, Blanchette M, Bureau TE, Wright SI, dePamphilis CW, Eric Schranz M, Barker MS, Conant GC, Wahlberg N, Vogel H, Pires JC, Wheat CW. The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci. 2015;112:8362–6. https://doi.org/10.1073/pnas.1503926112.
21. Olcerst A. Variation in induced responses of Datura wrightii to herbivore attack: plasticity of volatile organic compound emissions and gene expression across genotypes, ontogeny, and a single attack – doctoral disseration accessed via ProQuest. 2017. https://www.proquest.com/openview/920bf36262446dfa983f0a67e1755bff/1?pq-origsite=gscholar&cbl=18750.
22. Hare JD. Variation in herbivore and methyl jasmonate-induced volatiles among genetic lines of Datura wrightii. J Chem Ecol. 2007;33:2028–43. https://doi.org/10.1007/s10886-007-9375-1.
23. Doyle JJ, Doyle JL. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem Bull. 1987;19:11–5.
24. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5. https://doi.org/10.1038/s41592-020-01056-5.
25. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70. https://doi.org/10.1093/bioinformatics/btr011.
26. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11:1432. https://doi.org/10.1038/s41467-020-14998-3.
27. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31:3350–2. https://doi.org/10.1093/bioinformatics/btv383.
28. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, editor. Gene prediction: methods and protocols, methods in molecular biology. New York: Springer; 2019. p. 227–45. https://doi.org/10.1007/978-1-4939-9173-0_14.
29. Laetsch DR, Blaxter ML. BlobTools: interrogation of genome assemblies. 2017. https://doi.org/10.12688/f1000research.12232.1.
30. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.
31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
32. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with inspector. Genome Biol. 2021;22:1–21. https://doi.org/10.1186/s13059-021-02527-4.
33. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinform. 2009;25:4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.
34. Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, Kiel N, Saadat N, Ebenhöh O, Usadel B, Schwacke R, Bolger M, Weber APM, Denton AK. Helixer–de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. 2023. https://doi.org/10.1101/2023.02.06.527280.
35. Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, Denton AK. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics. 2021;36:5291–8. https://doi.org/10.1093/bioinformatics/btaa1044.
36. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. https://doi.org/10.1093/bioinformatics/btu031.
37. Schneider M, Lane L, Boutet E, Lieberherr D, Tognolli M, Bougueleret L, Bairoch A. The UniProtKB/Swiss-prot knowledgebase and its plant proteome annotation program. J Proteomics Plant Proteomics. 2009;72:567–73. https://doi.org/10.1016/j.jprot.2008.11.010.
38. Dainat J, Hereñú D, Murray DKD, Davis E, Crouch K, LucileSol, Agostinho N. pascal-git, Zollman, Z., tayyrov. NBISweden/AGAT: AGAT-v1.2.0. 2023. https://doi.org/10.5281/zenodo.8178877.
39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.
40. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21. https://doi.org/10.1186/s13059-014-0550-8.
41. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2022. (https://www.R-project.org/).
42. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49. https://doi.org/10.1093/nar/gkr1293.
43. McKibben MTW, Barker MS. Applying machine learning to classify the origins of gene duplications. In: Van de Peer Y, editor. Polyploidy: methods and protocols, methods in molecular biology. Springer, US: New York; 2023. p. 91–119. https://doi.org/10.1007/978-1-0716-2561-3_5.
44. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9:90–5. https://doi.org/10.1109/MCSE.2007.55.
45. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020;17(3):261–72.
46. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation. 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141.
47. Hong G, Zhang W, Li H, Shen X, Guo Z. Separate enrichment analysis of pathways for up- and downregulated genes. J R Soc Interface. 2014;11:20130950. https://doi.org/10.1098/rsif.2013.0950.
48. Hare JD, Sun JJ. Production of induced volatiles by Datura wrightii in response to damage by insects: effect of herbivore species and time. J Chem Ecol. 2011;37:751–64. https://doi.org/10.1007/s10886-011-9985-5.
49. Pantano L. Report of DEG analysis. 2017. https://lpantano.github.io/DEGreport/.
50. Kerchev PI, Fenton B, Foyer CH, Hancock RD. Plant responses to insect herbivory: interactions between photosynthesis, reactive oxygen species and hormonal signalling pathways. Plant Cell Environ. 2012;35:441–53. https://doi.org/10.1111/j.1365-3040.2011.02399.x.
51. Xu X, Yuan L, Xie Q. The circadian clock ticks in plant stress responses. Stress Biology. 2022;2:15. https://doi.org/10.1007/s44154-022-00040-7.
52. Howe GA, Jander G. Plant Immunity to insect herbivores. Annu Rev Plant Biol. 2008;59:41–66. https://doi.org/10.1146/annurev.arplant.59.032607.092825. [55] chitin binding, hydrolase enzymes, peptidase inhibitors, and terpene synthases.
53. Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40:1332–5. https://doi.org/10.1038/s41587-022-01261-x.
54. Mithöfer A, Wanner G, Boland W. Effects of feeding Spodoptera littoralis on lima bean leaves. II. Continuous mechanical wounding resembling insect feeding is sufficient to elicit herbivory-related volatile emission. Plant Physiol. 2005;137:1160–8. https://doi.org/10.1104/pp.104.054460.
55. Xu Z, Pu X, Gao R, Demurtas OC, Fleck SJ, Richter M, He C, Ji A, Sun W, Kong J, Hu K, Ren F, Song J, Wang Z, Gao T, Xiong C, Yu H, Xin T, Albert VA, Giuliano G, Chen S, Song J. Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. BMC Biol. 2020;18:1–14. https://doi.org/10.1186/s12915-020-00795-3.

Publisher's Note