Method S1. Plant material and RNA extraction.

Veratrum californicum plant material was obtained from wild populations in northern Utah. Tissue culture was initiated from wild collected seed and grown in the dark at 24°C on a combination of Linsmaier and Skoog vitamins (Linsmaier E.M. 1965) and Murashige and Skoog media (Murashige 1962) supplemented with 0.5 mg/l 1-naphthaleneacetic acid (Sigma). Refer to Table S9 for detailed media components. RNA extraction for each tissue (bulb, flower, leaf, fall rhizome, spring rhizome, fall root, green shoot, white shoot, and tissue culture samples) was performed as previously described (Johnson *et al.* 2012) (protocol 13). RNA quantity and integrity were evaluated with a NanoDrop 2000 (Thermo Scientific) and a Bioanalyzer 2100 (Agilent Technologies) prior to cDNA library preparation.

References

Johnson, M.T., Carpenter, E.J., Tian, Z., Bruskiewich, R., Burris, J.N., Carrigan, C.T., Chase, M.W., Clarke, N.D., Covshoff, S., Depamphilis, C.W., Edger, P.P., Goh, F., Graham, S., Greiner, S., Hibberd, J.M., Jordon-Thaden, I., Kutchan, T.M., Leebens-Mack, J., Melkonian, M., Miles, N., Myburg, H., Patterson, J., Pires, J.C., Ralph, P., Rolf, M., Sage, R.F., Soltis, D., Soltis, P., Stevenson, D., Stewart, C.N., Surek, B., Thomsen, C.J., Villarreal, J.C., Wu, X., Zhang, Y., Deyholos, M.K. and Wong, G.K. (2012) Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS One*, **7**, e50226.

Linsmaier E.M., S.F. (1965) Organic growth factor requirements of tobacco tissue cultures. *Physiologia Plantarum*, **18**, 100-127.

Murashige, T.S., F. (1962) A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiologia Plantarum*, **15**, 473–497.

Method S2. *Veratrum californicum* metabolite extraction for quantitation by LC-MS/MS.

Extracts were prepared by grinding frozen plant tissue in liquid nitrogen followed by 5 minutes of vortexing in 70% ethanol added in a 200 µl to 100 mg w/v ratio. Samples were subject to centrifugation for 10 minutes (14,000 X g) at room temperature and the supernatant filtered through a 0.2 µm PTFE membrane (Millipore) prior to injection. Extracts were diluted 10-10,000 fold with 70% ethanol, depending on alkaloid concentration, prior to LC-MS/MS analysis (refer to LC-MS/MS protocol in Experimental Procedures).

Method S3. Transcriptome assembly and determination of relative contig expression.

cDNA library construction, Illumina paired-end sequencing, and *de novo* transcriptome assembly for *Veratrum californicum* were performed at the National Center for Genome Resources (Santa Fe, New Mexico). For the transcriptome assembly, 50 bp paired-end Illumina reads for each tissue were first examined for gross abnormalities and poor sequence quality and trimmed with the FASTX Toolkit. (**http://hannonlab.cshl.edu/fastx_toolkit**). Quality control was as previously described (Kilgore *et al.* 2014).

Iseli, C., Jongeneel, C.V. and Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 138-148.

Kilgore, M.B., Augustin, M.M., Starks, C.M., O'Neil-Johnson, M., May, G.D., Crow, J.A. and Kutchan, T.M. (2014) Cloning and characterization of a norbelladine 4'-O-methyltransferase involved in the biosynthesis of the Alzheimer's drug galanthamine in Narcissus sp. aff. pseudonarcissus. *PLoS One*, **9**, e103223.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J. and Wang, J. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, **20**, 265-272.

Lottaz, C., Iseli, C., Jongeneel, C.V. and Bucher, P. (2003) Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, **19 Suppl 2**, ii103-112.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A. and Finn, R.D. (2012) The Pfam protein families database. *Nucleic acids research*, **40**, D290-301.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res*, **19**, 1117-1123.

UniProt, C. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research*, **41**, D43-47.

Method S4. Transcriptome dataset interrogation using Haystack and PlantTribes.

Haystack (http://haystack.mocklerlab.org/) input parameters included a value of 20 for fall root, fall rhizome, spring rhizome and bulb and a value of 1 for leaf, flower, white shoot, green shoot, and tissue culture 1- and 2 weeks after transfer to new media. A correlation cutoff value of 0.7 was used instead of the default of 0.8 to avoid missing true positives.

Gene family circumscriptions in PlantTribes were calculated using the similarity-based clustering of gene models from *Arabidopsis thaliana* (v. 7), *Carica papaya* (v.1), *Populus trichocarpa* (v. 1), *Medicago truncatula* (v. 1), *Oryza sativa* (v. 5), *Sorghum bicolor* (v. 1), *Selaginella moellendorffii* (v. 1), *Physcomitrella patens* (v. 1), and *Chlamydomonas reinhardtii* (v. 1) genome annotations using TribeMCL (Enright *et al.* 2003, Enright *et al.* 2002). Translated transcript assemblies for *Veratrum californicum, Narcissus* sp*. aff. pseudonarcissus* (daffodil), and *Colchicum autumnale* (autumn crocus) were sorted into the resulting gene family clusters using BLAST followed by multiple sequence alignment and gene tree estimation. In addition to this MCL (Markov CLuster algorithm) clustering approach, we developed a complete minimal representative dataset from all available plant species of cytochrome P450 genes relevant to alkaloid biosynthesis to identify candidate cytochromes P450.

Multiple sequence alignment and phylogenetic tree estimation were done on these relevant tribes and gene families using the MAFFT (Multiple Alignment using Fast Fourier Transform) alignment software (Katoh *et al.* 2009) and RAxML (Randomized Axelerated Maximum Likelihood) for maximum likelihood tree generation (Stamatakis 2006).

The RNA-seq transcriptome assembly sequences from *C. autumnale* and *Narcissus* were included in the tribe clustering steps (Supporting Data S1). These two species are also lillioid monocots but do not produce cyclopamine (but instead make the unrelated alkaloids colchicine and galanthamine, respectively), *C. autumnale* and *Narcissus* sequences were included in order to facilitate identification of tribe clusters that only contain *Veratrum californicum* genes, which were then assigned a higher priority for biochemical validation.

Selection criteria were established to score and sort the resulting clades, positive criteria were given a +1 each while penalizations were scored as a -1 each. A given clade was scored on the percentage of clade members that significantly co-localized with cyclopamine (e.g. present in the Haystack output dataset). Clades that did not contain any genes that fit the Haystack model were penalized. Clades containing genes that were *not* significantly co-localized with the alkaloid were penalized. Lastly, clades that contain genes from species that *do not* produce cyclopamine incurred a score penalty and were not chosen for initial biochemical characterization. These criteria were combined to score and rank the clades that

contain Haystack output gene members to identify the clade(s) with the highest likelihood of containing genes that function in the steroid alkaloid biosynthesis pathway.

References

Enright, A.J., Kunin, V. and Ouzounis, C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic acids research*, **31**, 4632-4638.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, **30**, 1575-1584.

Katoh, K., Asimenos, G. and Toh, H. (2009) Multiple alignment of DNA sequences with MAFFT. *Methods in molecular biology*, **537**, 39-64.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688-2690.

Method S5. Construction of viral expression vectors.

Candidate contigs obtained from Haystack analysis were subjected to BLAST searches (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and global alignments to homologous, experimentally characterized gene sequences with the CLC Main Workbench 6.8, for prediction of the open reading frame. Where the reading frame appeared incomplete, Rapid Amplification of cDNA Ends (RACE) was used to obtain the complete coding sequence. *Veratrum californicum* cDNA was prepared from root RNA extracts using M-MLV Reverse Transcriptase (Invitrogen) according to manufacturer's instructions. All primer sequences and PCR programs can be found in Tables S11 and S12, respectively.

The cDNAs encoding CYP90B27 (accession numbers KJ869252, KJ869253), CYP90G1 (accession numbers KJ869258-KJ869261), GABAT1 (accession numbers KJ869262-

RACE was required to determine the 5' sequence of CYP94N1 gene (accession numbers KJ869254-KJ869257). RACE ready cDNA was prepared using the GeneRacer Kit (Invitrogen) according to manufacturer's instructions using *V. californicum* root RNA. Primers 13 and 15 were used for PCR (round 1), followed by amplification using primers 14 and 16 (round 2). Resulting RACE fragments were cloned into PCR-Blunt II-TOPO. The full-length gene was directly amplified from *V. californicum* root cDNA with primers 17 and 18, incorporating BglII/EcoRI restriction sites at the 5' and 3' end of the open reading frame. The amplified product was digested with Bglll/EcoRI and ligated into pVL1392 digested with the same enzymes. Each characterized *V. californicum* contig and subsequent enzyme designation can be found in Table S6.

The cDNA encoding GABA transaminase isozyme 2 from *Solanum lycopersicum* (tomato GABAT2) implicated in steroid alkaloid biosynthesis (accession number AY240230) was isolated from *S. lycopersicum* using the Qiagen RNA-easy kit for RNA extraction followed by cDNA synthesis as described above. Tomato GABAT2 was amplified by PCR using Primers 19 and 20, incorporating PstI/XbaI sites at the 5' and 3' end of the open reading frame. The amplified product and pVL1392 were subject to restriction digest with PstI/XbaI and ligated together, preceding transformation.

Method S6. Virus co-transfection, amplification, and protein production.

Each pVL1392 expression construct was independently co-transfected with the Baculogold Linearized Baculovirus (BD Biosciences) into *S. frugiperda* Sf9 cells according to manufacturer's instructions. Sf9 cells were maintained as previously described (Gesell *et al.* 2011). Virus amplification and protein production proceeded as previously described (Gesell *et al.* 2009). Each cytochrome P450 virus construct was co-expressed with virus containing *E. californica* cytochrome P450 reductase (CPR) while the GABAT1 was produced by single infection. Sf9 cell cultures were also infected with several constructs in parallel. Combinations of each virus used in multiple infections can be found in Table S13. Equal volumes for each virus were used in the multiple infections and adjusted to a total viral volume of 2.5 ml.

References

Gesell, A., Chavez, M.L., Kramell, R., Piotrowski, M., Macheroux, P. and Kutchan, T.M. (2011) Heterologous expression of two FAD-dependent oxidases with (S)-tetrahydroprotoberberine oxidase activity from Arge mone mexicana and Berberis wilsoniae in insect cells. *Planta*, **233**, 1185-1197.

Gesell, A., Rolf, M., Ziegler, J., Diaz Chavez, M.L., Huang, F.C. and Kutchan, T.M. (2009) CYP719B1 is salutaridine synthase, the C-C phenol-coupling enzyme of morphine biosynthesis in opium poppy. *The Journal of biological chemistry*, **284**, 24432-24442.

Method S7. Extraction of multiple infections for Sf9 *in vivo* product production.

Baculovirus infections were carried out for production of each enzymatic product in *S. frugiperda* Sf9 cells and collected as stated in Method S6. 1 ml each of Sf9 cells expressing the various combinations of virus were extracted with 2 volumes of ethyl acetate by vortexing (1 min), centrifugation (16,000 x g; 2 min), and were taken to dryness under N_2. Samples were either derivatized with 40 µl of Sylon HTP and injected onto the GC-MS using the protocol stated in Experimental Procedures or were re-suspended in 50 µl of 80% methanol and analyzed by LC/MS-MS according to the protocol in Experimental Procedures.

Method S8. Assays to clarify order of enzymatic transformations.

Assay for GC-MS: Cytochrome P450 enzyme assay conditions were identical to those stated above using *S. frugiperda* Sf9 cell suspensions with the following modifications. First, 12 assays each containing CYP90G1 + CPR, CYP94N1 + CPR, or control cytochrome P450 + CPR and each with pure 22(*R*)-hydroxycholesterol were allowed to incubate overnight at 30°C. Like assays were pooled, extracted 3 times with 2 volumes ethyl acetate, dried under N_2, and re-suspended in 180 µl of 25% DMSO. Extracts containing the enzymatic product of the 22-hydroxycholesterol 26-hydroxylase/ oxidase + CPR and 22(*R*)-hydroxycholesterol were divided equally and used as substrate for 6 assays containing CYP90G1 + CPR and 6 assays containing control cytochrome P450 + CPR. Extracts containing the enzymatic product of CYP90G1 + CPR and 22(*R*)-hydroxycholesterol were divided and used as substrate in 6 assays containing 22-hydroxycholesterol 26-hydroxylase/ oxidase + CPR and 6 assays containing control cytochrome P450 + CPR. Control P450 + CPR assay was run in parallel, treated identically and added to another control P450 assay. Assays were allowed to incubate for 20 min at 30°C then stopped by addition of 20 µl of 20% TCA with vortexing. Like assays were pooled, extracted, derivatized, and analyzed by GC-MS using the protocol stated in Experimental Procedures. Refer to Figure S11 a for an overview of the experiment.

Assay for LC-MS/MS: All assays used crude Sf9 cell suspensions. Enzyme assays started with a combination of CYP90B27 + CPR and CYP94N1 + CPR (8 individual reactions) in parallel to CYP90B27 + CPR and CYP90G1 + CPR (8 reactions). Assays were extracted, and added to CYP94N1 + CPR, CYP90G1 + CPR, or GABAT1 for several possible enzyme combinations (4 reactions each). Like samples were pooled, extracted, and added to 2 reactions each with enzyme not yet utilized previously. Refer to Figure S11 b for an overview of the experiment. Samples were taken at each step post extraction for LC-MS/MS analysis and run by the protocol stated in Experimental Procedures.

Method S9. Enzymatic product purification for NMR and high resolution MS for structure elucidation.

Large-scale 750 ml *S. frugiperda* Sf9 cultures were grown expressing viral combinations 5-7 (Table S13) of the *Veratrum californicum* enzymes as previously described (Gesell *et al.* 2009). Cells were collected after three days and re-suspended in 10 ml of 100 mM tricine pH 7.4/ 5 mM thioglycolic acid; then extracted 3 times with 2 volumes of hexane or ethyl acetate. The remaining aqueous supernatant was extracted once with 1 volume of hexane or ethyl acetate. Extracts for each infection were then pooled, dried under N_2, and re-suspended in 5 ml of absolute methanol.

The extracts were purified on a Waters HPLC system equipped with a 2707 autosampler, 1525 binary pump, 2998 photodiode array detector, and Waters Fraction Collector III. In some cases, samples were cleaned up by Solid Phase Extraction (SPE), before HPLC purification. For HPLC, extracts were concentrated to 500 µl and then injected in 50 µl portions onto a Phenomenex Gemini C-18 NX column (150 X 2.00 mm, 5 µm) with the same solvents used for LC-MS/MS as described in Experimental Procedures with the following binary gradient: Solvent B was held at 20% for 2 min, then 2-11 min 20-30% B, 11-18 min 30-100% B, 18-30 min 100% B, 30-31 min 100-20% B, and held at 20% B for an additional 5 minutes. The flow rate was 0.5 ml/min; 0.5 ml fractions were collected. The resulting fractions were then analyzed by GC-MS or LC-MS/MS as described in Experimental Procedures, and selected samples were analyzed by NMR or by high resolution MS. NMR spectra were acquired in MeOD at 600 MHz on a BrukerAvance 600 MHz spectrometer equipped with a BrukerBioSpin TCI 1.7 mm MicroCryoProbe. Proton, gCOSY, ROESY, gHSQC, and gHMBC spectra were acquired; ^{13}C chemical shifts were obtained from the HSQC and HMBC spectra. Chemical shifts are reported with respect to the residual non-deuterated MeOD signal. Refer to Data S2 for NMR designations for 22-keto-cholesterol and 22-keto-26-hydroxycholesterol. For high resolution MS, samples were diluted 1:10 in 80% acetonitrile:water (LC-MS grade) containing 0.1% formic acid and infused into an LTQ-Orbitrap Velos Pro (Thermo-Fisher Scientific, San Jose, CA) using a Triversa Nanomate (Advion, Ithaca, NY). Data were collected in positive ion mode, detected in the Orbitrap at a nominal resolution setting of 60,000 at m/z 400. Precursors were determined with a wide SIM scan (m/z 385-430). Precursors were isolated in the

ion-trap and transferred to the HCD cell for fragmentation at 35 NCE (m/z 418) and 50 NCE (m/z 398). Data were analyzed manually using the Qualbrowser application of Xcalibur (Thermo-Fisher Scientific, San Jose, CA).

References

Gesell, A., Rolf, M., Ziegler, J., Diaz Chavez, M.L., Huang, F.C. and Kutchan, T.M. (2009) CYP719B1 is salutaridine synthase, the C-C phenol-coupling enzyme of morphine biosynthesis in opium poppy. *The Journal of biological chemistry*, **284**, 24432-24442.

Method S10. Phylogenetic analysis of cytochrome P450 enzymes across species using deep transcriptome sequence data from 1KP and MonAToL projects.

Sequences for P450 genes were identified and pulled from various sources. Transcriptome data for monocotyledon taxa were taken from the MonAToL project (DEB-0830020) and the 1KP project (https://sites.google.com/a/ualberta.ca/onekp/home). Taxa were selected to represent major clades but were not exhaustive of the sampling available. We also included genome sequences from the 22 sequenced land plant genomes used by the *Amborella* genome project to represent eudicots and other angiosperms (Amborella Genome 2013). The final taxa used in reconstructing the P450 gene tree can be found in Table S8. Additionally, a P450 sequence from *S. cerevisiae* was used as the outgroup (GenBank Accession: U34636.1).

Transcriptomic data was assembled using Trinity r2013-02-25 and resulting assemblies were filtered using FPKM (fragments per kilobase of exon per million fragments mapped). Sequences where an isoform represented less than 1% of all reads mapping to a gene were discarded. Assemblies were translated using the ORF estimation Transdecoder r20131110 software packaged with Trinity. Blastp (Camacho *et al.* 2009) was used to blast genomic and transcriptomic amino acid sequences using the *S. lycopersicum* GAME4, GAME6, GAME7, and GAME8 (Refer to Table S7 for accession numbers) and *Veratrum californicum* CYP90B27, CYP90G1, and CYP94N1 (Refer to Table S7 for accession numbers) P450 amino acid sequence. Best blast hits were identified using an initial e-value threshold of 1e-10 followed by a minimum overlapping length of 85% between the query and subject sequences. These filtering criteria resulted in putative P450 genes of at least 1017 basepairs in length for all taxa.

Amino acid sequences for all putative P450 sequences were aligned using MAFFT v.7.029b (Katoh and Standley 2014) under default settings. Trees were estimated using RAxML v.8.0.22 (Stamatakis 2006) under the GTR + • • • • • • • • • • with 500 bootstrap replicates and *S. cerevisiae* as the outgroup.

References

Amborella Genome, P. (2013) The Amborella genome and the evolution of flowering plants. *Science*, **342**, 1241089.

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

Katoh, K. and Standley, D.M. (2014) MAFFT: iterative refinement and additional methods. *Methods in molecular biology*, **1079**, 131-146.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688-2690.