De novo genome assembly of *Camptotheca acuminata*, a natural source of the anti-cancer compound camptothecin

Dongyang Zhao¹, John P. Hamilton¹, Gina M. Pham¹, Emily Crisovan¹, Krystle Wiegert-Rininger¹,
Brieanne Vaillancourt¹, Dean DellaPenna², and C. Robin Buell¹*

¹Department of Plant Biology, Michigan State University, East Lansing, MI 48824 USA
²Department of Biochemistry & Molecular Biology, Michigan State University, East Lansing, MI 48824 USA

Email addresses: Dongyan Zhao <zhaodon4@msu.edu>, John P. Hamilton <jham@msu.edu>, Gina M. Pham <phamgina@msu.edu>, Emily Crisovan <pankeyem@msu.edu>, Krystle Wiegert-Rininger <wiegertk@msu.edu>, Brieanne Vaillancourt <vaillan6@msu.edu>, Dean DellaPenna <dellapen@msu.edu>, C Robin Buell <buell@msu.edu>

*Correspondence should be addressed to: C. Robin Buell, buell@msu.edu

Manuscript type: Data note
Abstract

Background: *Camptotheca acuminata* is one of a limited number of species that produce camptothecin, a pentacyclic quinoline alkaloid with anti-cancer activity due to its ability to inhibit DNA topoisomerase. While transcriptome studies have been performed previously with various camptothecin-producing species, no genome sequence for a camptothecin-producing species is available to date.

Findings: We generated a high quality de novo genome assembly for *C. acuminata* representing 403,174,860 bp on 1,394 scaffolds with an N50 scaffold size of 1,752 kbp. Quality assessments of the assembly revealed robust representation of the genome sequence including genic regions. Using a novel genome annotation method, we annotated 31,825 genes encoding 40,332 gene models. Based on sequence identity and orthology with validated genes from *Catharanthus roseus* as well as Pfam searches, we identified candidate orthologs for genes potentially involved in camptothecin biosynthesis. Extensive gene duplication including tandem duplication was widespread in the *C. acuminata* genome with 2,571 genes belonging to 997 tandem duplicated gene clusters.

Conclusions: To our knowledge, this is the first genome sequence for a camptothecin-producing species, and access to the *C. acuminata* genome will permit not only discovery of genes encoding the camptothecin biosynthetic pathway but also reagents that can be used for heterologous expression of camptothecin and camptothecin analogs with novel pharmaceutical applications.
Keywords: *Camptotheca acuminata*, camptothecin, genome assembly, genome annotation, tandem duplications

Data Description

Background information on camptothecin, a key anti-cancer natural product

*Camptotheca acuminata* Decne, also known as the Chinese Happy Tree (Figure 1), is an eudicot asterid Cornales tropical tree species within the Nyssaceae family [1] that also contains *Nyssa* spp (tupelo) and *Davidia involucrate* (dove tree); no genome sequence is available for any member of this family. *C. acuminata* is one of a limited number of plant species that produce camptothecin, a pentacyclic quinoline alkaloid (Figure 2A) with anti-cancer activity due to its ability to inhibit DNA topoisomerase [2]. Due to poor solubility, derivatives such as irinotecan and topotecan, rather than camptothecin are currently in use as approved cancer drugs. The significance of these derivatives as therapeutics is highlighted by the listing of irinotecan on the World Health Organization Model List of Essential Medicines [3]. While transcriptome studies have been performed previously with various camptothecin-producing species including *C. acuminata* and *Ophiopogon pumila* (e.g., [4-6]), no genome sequence for a camptothecin-producing species is available to date. We report on the assembly and annotation of the *C. acuminata* genome, the characterization of genes implicated in camptothecin biosynthesis, and highlight the extent of gene duplication that provides new templates for gene diversification.
RNA isolation, library construction, sequencing, and transcriptome assembly

Transcriptome assemblies were constructed using nine developmental RNA-seq datasets described in a previous study [4] that included immature bark, cotyledons, immature flower, immature fruit, mature fruit, mature leaf, root, upper stem, and lower stem. Adapters and low-quality nucleotides were removed from the RNA-seq reads using Cutadapt v1.8 (Cutadapt, RRID:SCR_011841) [7] and contaminating ribosomal RNA reads were removed. Cleaned reads from all nine libraries were assembled using Trinity v20140717 (Trinity, RRID:SCR_013048) [8] with a normalization factor of 50x using default parameters. Contaminant transcripts (5,669 total) were identified by searching the de novo transcriptome assembly against the National Center for Biotechnology Information (NCBI) non-redundant nucleotide database using BLAST+ (v2.2.30) [9, 10] with an E-value cutoff of 1e-5; transcripts with their best hits being a non-plant sequence were removed from the transcriptome.

For additional transcript support for use in a genome-guided transcriptome assembly to support genome annotation, strand-specific RNA-seq reads were generated by isolating RNA from root tissues and sequencing of Kappa TruSeq Stranded libraries on an Illumina HiSeq 2500 platform generating 150 nt paired-end reads (BioSample ID: SAMN06229771). Root RNA-seq reads were assessed for quality using FASTQC v0.11.2 (FASTQC, RRID:SCR_014583) [11] using default parameters and cleaned as described above.

DNA isolation, library construction, and sequencing

The genome size of C. acuminata was estimated at 516 Mb using flow cytometry, suitable for de novo assembly using the Illumina platform. DNA was extracted from young leaves of C.
acuminata at the vegetative growth stage using CTAB [12]. Multiple Illumina-compatible paired-end libraries (Table 1) with insert sizes ranging from 180-609 bp were constructed as described previously [13] and sequenced to 150 nt in paired-end mode on an Illumina HiSeq2000. Mate-pair libraries (Table 1) with size ranges of 1.3-8.9 kb were made using the Nextera Kit (Illumina, San Diego CA) as per manufacturer’s instructions and sequenced to 150 nt in paired-end mode on an Illumina HiSeq2000.

**Genome assembly**

Paired-end reads (Table 1) were assessed for quality using FASTQC v0.11.2 (FASTQC, RRID:SCR_014583) [11] using default parameters, cleaned for adapters and low quality sequences using Cutadapt v1.8 (Cutadapt, RRID:SCR_011841) [7] and only reads in pairs with each read ≥25 nt were retained for genome assembly. Mate pair libraries (Table 1) were processed using NextClip v1.3.1 (NextClip, RRID:SCR_005465) [14] and only reads from Categories A, B, C were used for the assembly. Using ALLPATHS-LG v44837 (ALLPATHS-LG, RRID:SCR_010742) [15] with default parameters, two paired-end read libraries (180 and 268 bp insert libraries) and all five mate pair libraries (Table 1) were used to generate an initial assembly of 403.2 Mb with an N50 contig size of 108 kbp and an N50 scaffold size of 1,752 kbp (Tables 1 and 2). Gaps (5,076) in this initial assembly were filled using SOAP GapCloser v1.12r6 (GapCloser, RRID:SCR_015026) [16] with four independent paired-end libraries (352, 429, 585, and 609 bp inserts, Table 1); 12,468,362 bp of the estimated 16,471,841 bp of gaps was filled leaving a total of 3,825 gaps (3,772,191 Ns). The assembly was checked for contaminant sequences based on alignments to the NCBI non-redundant nucleotide database using BLASTN (E-value = 1e-5) [10];
a single scaffold of 5,156 bp that matched a bacterium sequence with 100% coverage and 100% identity was removed. Subsequently, five scaffolds less than 1 kbp were removed resulting in the final assembly of 403,174,860 bp comprised of 1,394 scaffolds with an N50 scaffold size of 1,752 kbp (Tables 1 and 2) and 0.9% Ns.

Quality assessments revealed a robust high quality assembly with 98% of the paired-end genomic sequencing reads aligning to the assembly, of which, 99.97% aligned concordantly. With respect to genic representation, 95.3% of RNA-seq-derived transcript assemblies [4] and 74,119 of 74,682 (99%) pyrosequencing transcript reads from a separate study [5] aligned to the genome assembly. A total of 93.6% of conserved Embryophyta BUSCO (BUSCO , RRID:SCR_015008) proteins were present in the assembly as full-length sequences with an additional 2.5% of the Embryophyta proteins fragmented [17].

**Genome annotation**

We used a novel genome annotation method to generate high quality annotation of the *C. acuminata* genome in which we repeat masked the genome, trained an *ab initio* gene finder with a genome-guided transcript assembly, and then refined the gene models using additional genome-guided transcript assembly evidence to generate a high quality gene model set. We first created a *C. acuminata* specific custom repeat library (CRL) using MITE-Hunter v2011 [18] and RepeatModeler v1.0.8 (RepeatModeler , RRID:SCR_015027) [19]. Protein coding genes were removed from each repeat library using ProtExcluder.pl v1.1 [20] and combined into a single CRL, which hard-masked 143.6 Mb (35.6%) of the assembly as repetitive sequence using RepeatMasker v4.0.6 (RepeatMasker , RRID:SCR_012954) [21]. Cleaned root RNA-seq reads
Table S1, BioSample ID: SAMN06229771) were aligned to the genome assembly using TopHat2 v2.0.13 (TopHat, RRID:SCR_013035) [22] in strand-specific mode with a minimum intron length of 20 bp and a maximum intron length of 20 kb; the alignments were then used to create a genome-guided transcriptome assembly using Trinity v2.2.0 (Trinity, RRID:SCR_013048) [23]. The RNA-seq alignments were used to train AUGUSTUS v3.1 (Augustus: Gene Prediction, RRID:SCR_008417) [24] and gene predictions were generated with AUGUSTUS [25] using the hard-masked assembly. Gene model structures were refined by incorporating evidence from the genome-guided transcriptome assembly using PASA2 v2.0.2 (PASA, RRID:SCR_014656) [26, 27]; with the parameters: MIN_PERCENT_ALIGNED=90, MIN_AVG_PER_ID=99. After annotation comparison, models that PASA identified as being merged and a subset of candidate camptothecin biosynthetic pathways genes identified as mis-annotated were manually curated. The final high-confidence gene model set consists of 31,825 genes encoding 40,332 gene models. Functional annotation was assigned using a custom pipeline using WU-BLASTP [28] searches against the Arabidopsis thaliana annotation (TAIR10; [29]) and Swiss-Prot plant proteins (downloaded on 08-17-2015), and a search against Pfam (v29) using HMMER v3.1b2 (Hmmer, RRID:SCR_005305) [30]. This resulted in 34,143 gene models assigned a putative function, 2,011 annotated as conserved hypothetical, and 4,178 annotated as hypothetical.

C. acuminata is insensitive to camptothecin due to mutations within its own DNA topoisomerase [31] and we identified two topoisomerase genes in our annotated gene set, one of which matches the published C. acuminata topoisomerase (99.78% identity, 100% coverage) and includes the two mutations that confer resistance to camptothecin (Figure 2B), one mutation is specific in C. acuminata and the other is present in both C. acuminata and two
camptothecin-producing *Ophiorrhiza* species. Further quality assessments of our annotation with 35 nuclear-encoded *C. acuminata* genes available from GenBank revealed an average identity of 99.5% with 100% coverage in our annotated proteome while a single gene encoding 1-deoxy-D-xylulose 5-phosphate reductoisomerase (ABC86579.1) had 88.2% identity with 100% coverage that may be attributable to differences in genotypes. One mRNA reported to encode a putative strictosidine beta-D-glucosidase (AES93119.1) was found to have a retained intron that when removed, aligned with 99.3% identity yet reduced coverage (66%) as it was located at the end of a short scaffold. Collectively, the concordant alignment of whole genome shotgun sequence reads to the assembly, the high representation of genic regions as assessed by independent transcriptome datasets (RNA-seq and pyrosequencing) as well as the core Embryophyta BUSCO proteins, when coupled with the high quality gene models as revealed through alignments with cloned *C. acuminata* genes indicate that we have not only generated a high quality genome assembly for *C. acuminata* but also a robust set of annotated gene models.

**Gene duplication and orthology analyses**

During our annotation efforts, it was readily apparent that there was substantial gene duplication including tandem gene duplication in the *C. acuminata* genome. Paralogous clustering of the *C. acuminata* proteome revealed 5,516 paralogous groups containing 15,806 genes. We identified tandem gene duplications in the *C. acuminata* genome based on if: 1) two or more *C. acuminata* genes were present within an orthologous/paralogous group; 2) there were no more than 10 genes in between on a single scaffold; and 3) the pairwise gene distance was less than 100 kbp [32]. Under these criteria, a total of 2,571 genes belonging to 997
tandem duplicated gene clusters were identified. Gene ontology analysis showed that tandem duplicated genes are significantly enriched in “response to stress” ($p < 0.0001, \chi^2$ test) while under-represented in most other processes, especially “other cellular processes” and “cell organization and biogenesis” ($p < 0.0001, \chi^2$ test).

To our knowledge, *C. acuminata* is the first species within the Nyssaceae family with a genome sequence. To better understand the evolutionary relationship of *C. acuminata* with other asterids and angiosperms, we identified orthologous and paralogous groups using our annotated *C. acuminata* proteome and the proteomes of three other key species (*Arabidopsis thaliana*, *Amborella trichopoda*, and *Catharanthus roseus*) using OrthoFinder (v0.7.1) [33] with default parameters. A total of 12,667 orthologous groups containing at least a single *C. acuminata* protein were identified with 9,659 orthologous groups common to all four species (Figure 3; Table S2). Interestingly, *C. acuminata* contains less singleton genes (8,868) than *A. trichopoda* and *C. roseus*, and gene ontology analysis demonstrated that these genes were highly enriched in “transport”, “response to stress”, and “other metabolic and biological processes” ($p < 0.0001, \chi^2$ test) while dramatically under-represented in “unknown biological processes” ($p < 0.0001, \chi^2$ test), suggesting these genes may be involved in stress responses and other processes specific to *C. acuminata*.

**Uses for the *C. acuminata* genome sequence and annotation**

Generation of a high-quality genome sequence and annotation dataset for *C. acuminata* will facilitate discovery of genes encoding camptothecin biosynthesis as physical clustering can be combined with sequence similarity and co-expression data to identify candidate genes, an
approach that has been extremely useful in identifying genes in specialized metabolism in a
number of plant species (see [34-36]). In C. acuminata, geranylgeranyl diphosphate from the 2-
C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate (MEP) pathway is used to
generate secologanic acid via the iridoid pathway and tryptamine from tryptophan
decarboxylase are condensed by strictosidinic acid synthase to generate strictosidinic acid that
is then converted into camptothecin in the alkaloid pathway via a set of unknown steps [37]
(Figure 4A). Catharanthus roseus, Madagascar periwinkle, produces vinblastine and vincristine
via the MEP and iridoid pathways for which all genes leading to the biosynthesis of the iridoid
secologanin have been characterized [35]. Using sequence identity and coverage with
characterized C. roseus genes from the MEP and iridoid pathway (Figure 4A), we were able to
identify candidate genes for all steps in the MEP and iridoid pathway in C. acuminata (Table 3).
The downstream steps in camptothecin biosynthesis subsequent to formation of strictosidinic
acid involve a broad set of enzymes responsible for reduction and oxidation [37] and a total of
343 cytochrome P450s (56 paralogous gene clusters and 120 singletons; Table S3) were
identified which can serve as candidates for the later steps in camptothecin biosynthesis.

Though not absolute, physical clustering of genes involved in specialized metabolism has been
observed in a number of species across a number of classes of specialized metabolites [34, 38].
With an N50 scaffold size of 1,752 kbp, we observed several instances of physical clustering of
genes with homology to genes involved in monoterpane indole alkaloid biosynthesis which may
produce related compounds in C. acuminata. Using characterized genes involved in the
biosynthesis of vinblastine and vincristine from C. roseus as queries [35] (Figure 4A, Table 3), we
identified a single C. acuminata scaffold (907 kbp, 86 genes; Figure 4B) that encoded genes with
sequence identity to isopentenyl diphosphate isomerase II within the MEP pathway, 8-
hydroxygeraniol oxidoreductase (GOR, three complete and one partial paralogs), 7-
deoxyloganic acid 7-hydroxylase (7DLH) within the iridoid pathway, and a protein with
homology to C. roseus 16-hydroxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase
(NMT) within the alkaloid pathway suggesting that access to a high contiguity genome assembly
may facilitate discovery of genes involved in specialized metabolism in C. acuminata. Tandem
duplications of genes involved in specialized metabolism have been reported previously [39, 40]
and via divergence either in the coding region or promoter sequence which lead to neo- and
sub-functionalization at the enzymatic or expression level, respectively, have been shown to
contribute to the extensive chemical diversity within a species [40, 41].
The C. acuminata genome can also be used to facilitate our understanding of the mechanisms
by which camptothecin production evolved independently in distinct taxa such as C. acuminata
(Nyssaceae) and O. pumila (Rubiaceae). For example, a comparative analysis of C. acuminata
and O. pumila may be highly informative in not only delineating genes involved in camptothecin
biosynthesis but also in revealing key evolutionary events that led to biosynthesis of this critical
natural product across a wide phylogenetic distance. As noted above, camptothecin is
cytotoxic and as a consequence, derivatives of camptothecin are used as anti-cancer drugs.
Perhaps most exciting, the ability to decipher the full camptothecin biosynthetic pathway will
yield molecular reagents that can be used to not only synthesize camptothecin in heterologous
systems such as yeast, but also produce less toxic analogs with novel pharmaceutical
applications.
Availability of Supporting Information

Raw genomic sequence reads and transcriptome reads derived from root tissues are available in the NCBI Sequence Read Archive under project number PRJNA361128. All other RNA-seq transcriptome reads were from Bioproject PRJNA80029 [4]. The genome assembly and annotation are available in the Dryad Digital Repository [42] and through the Medicinal Plant Genomics Resource [43] via a genome browser and search and analysis tools.

Abbreviations

2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate (MEP), 7-deoxyloganic acid 7-hydroxylase (7DLH), 8-hydroxygeraniol oxidoreductase (GOR), 16-hydroxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase (NMT), custom repeat library (CRL), National Center for Biotechnology Information (NCBI), RNA-sequencing (RNA-seq)

Competing Interests

The authors have declared that no competing interests exists.

Author Contributions

CRB oversaw the project. DZ performed the genome assembly, assisted in genome annotation and analyzed data. JH annotated the genome and analyzed data. EC, GP, and KWR constructed libraries and analyzed data. BV analyzed data. DDP provided intellectual oversight. DZ, JH, and CRB wrote the manuscript.
Acknowledgements

Funding for this work was provided in part by a grant to CRB and DDP from the National Institute of General Medical Sciences (1RC2GM092521) and funds to CRB and DDP from Michigan State University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Figure Legends

Figure 1. *Camptotheca acuminata* Decne, the Chinese Happy Tree, is a member in the Nyssaceae family that produces the anticancer compound camptothecin.

Figure 2. Genome aspects of *Camptotheca acuminata*. (A) Structure of camptothecin. (B) Key amino acid mutations (red rectangles) in DNA topoisomerase I in camptothecin-producing and non-producing species and their phylogenetic relationship.

Figure 3. Venn diagram showing orthologous and paralogous groups between *Amborella trichopoda*, *Arabidopsis thaliana*, *Camptotheca acuminata*, and *Catharanthus roseus*.

Figure 4. Key portions of the proposed camptothecin biosynthetic pathway and an example of physical clustering of candidate genes in *Camptotheca acuminata*. (A) The methylerythritol phosphate (MEP) pathway (green), iridoid pathway (blue), and condensation of secologanic acid with tryptamine via strictosidinic acid synthase (STRAS) to form strictosidinic acid prior to downstream dehydration, reduction, and oxidation steps yielding camptothecin. DXS, 1-deoxy-D-xylulose 5-phosphate synthase 2; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; CMS, 4-diphosphocytidyl-methylerythritol 2-phosphate synthase; CMK, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MCS, 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, GCPE protein; HDR, 1-hydroxy-2-methyl-butenyl 4-diphosphate reductase; IPI, plastid isopentenyl pyrophosphate, dimethylallyl pyrophosphate isomerase; GPPS, geranyl pyrophosphate synthase; GES, plastid geraniol synthase; G8H, geraniol 8-hydroxylase; GOR, 8-hydroxygeraniol oxidoreductase; CYC1, iridoid cyclase 1; 7-DLS, 7-deoxyloganetic acid synthase; 7-DLGT, 7-deoxyloganetic acid glucosyltransferase; 7-DLH, 7-
deoxyloganic acid hydroxylase; SLAS, secologanic acid synthase; TDC, tryptophan decarboxylase.

(B) Physical clustering of homologs of genes involved in the methylerthritol phosphate, iridoid, and alkaloid biosynthetic pathways of Catharanthus roseus on scaffold 151 of C. acuminata. GOR: 8-hydroxygeraniol oxidoreductase; NMT: 16-hydroxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase; 7DLH: 7-deoxyloganic acid 7-hydroxylase; IPP2: isopentenyl diphosphate isomerase II. Gene IDs are below the arrows.
Table 1. Input libraries and sequences for *de novo* assembly of the *Camptotheca acuminata* genome.

<table>
<thead>
<tr>
<th>BioProject ID</th>
<th>BioSample ID</th>
<th>Fragment size (bp)</th>
<th>No. of cleaned read pairs</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220985</td>
<td>180</td>
<td>96,955,546</td>
<td>ALLPATHS-LG assembly</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220986</td>
<td>268</td>
<td>89,381,055</td>
<td>ALLPATHS-LG assembly</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220987</td>
<td>352</td>
<td>61,207,691</td>
<td>GapCloser</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220988</td>
<td>429</td>
<td>50,688,562</td>
<td>GapCloser</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220989</td>
<td>585</td>
<td>21,856,610</td>
<td>GapCloser</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220990</td>
<td>609</td>
<td>22,217,954</td>
<td>GapCloser</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220991</td>
<td>8,111</td>
<td>9,923,643</td>
<td>ALLPATHS-LG assembly</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220992</td>
<td>7,911</td>
<td>7,652,519</td>
<td>ALLPATHS-LG assembly</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220993</td>
<td>1,377</td>
<td>12,800,554</td>
<td>ALLPATHS-LG assembly</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220994</td>
<td>3,179</td>
<td>13,138,503</td>
<td>ALLPATHS-LG assembly</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06220995</td>
<td>8,879</td>
<td>13,599,241</td>
<td>ALLPATHS-LG assembly</td>
</tr>
</tbody>
</table>

All libraries were sequenced in paired end mode generating 150 nt reads.
Table 2. Metrics of the final assembly of *Camptotheca acuminata* genome.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total scaffold length (bp)</td>
<td>403,174,860</td>
</tr>
<tr>
<td>Total no. of scaffolds (bp)</td>
<td>1,394</td>
</tr>
<tr>
<td>Maximum scaffold length (bp)</td>
<td>8,423,530</td>
</tr>
<tr>
<td>Minimum scaffold length (bp)</td>
<td>1,002</td>
</tr>
<tr>
<td>N50 scaffold size (bp)</td>
<td>1,751,747</td>
</tr>
<tr>
<td>N50 contig size (bp)</td>
<td>107,594</td>
</tr>
<tr>
<td>No. Ns</td>
<td>3,772,191 (0.9%)</td>
</tr>
<tr>
<td>No. gaps</td>
<td>3,825</td>
</tr>
</tbody>
</table>
Table 3. Identification of candidate camptothecin biosynthetic pathway genes in the *Camptotheca acuminata* genome as revealed by sequence identity and coverage with characterized genes from the 2-C-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate and iridoid biosynthetic pathways from *Catharanthus roseus*.

<table>
<thead>
<tr>
<th>Description</th>
<th>Abbreviation</th>
<th>Protein</th>
<th>Camptotheca Gene ID</th>
<th>% coverage</th>
<th>% identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-deoxy-D-xylulose 5-phosphate synthase 2</td>
<td>DXS</td>
<td>ABI35993.1</td>
<td>Cac_g024944.t1</td>
<td>98</td>
<td>77.60</td>
</tr>
<tr>
<td>1-deoxy-D-xylulose-5-phosphate reductoisomerase</td>
<td>DXR</td>
<td>AAF65154.1</td>
<td>Cac_g016318.t1</td>
<td>100</td>
<td>88.82</td>
</tr>
<tr>
<td>4-diphosphocytidyl-methylerythritol 2-phosphate synthase</td>
<td>CMS</td>
<td>ACI16377.1</td>
<td>Cac_g018722.t1</td>
<td>88</td>
<td>77.82</td>
</tr>
<tr>
<td>4-diphosphocytidyl-2-C-methyl-D-erythritol kinase</td>
<td>CMK</td>
<td>ABI35992.1</td>
<td>Cac_g021688.t1</td>
<td>99</td>
<td>76.17</td>
</tr>
<tr>
<td>2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase</td>
<td>MCS</td>
<td>AAF65155.1</td>
<td>Cac_g008169.t1</td>
<td>100</td>
<td>73.77</td>
</tr>
<tr>
<td>GCPE protein</td>
<td>HDS</td>
<td>AAO24774.1</td>
<td>Cac_g022763.t1</td>
<td>100</td>
<td>88.65</td>
</tr>
<tr>
<td>1-hydroxy-2-methyl-butenyl 4-diphosphate reductase</td>
<td>HDR</td>
<td>ABI30631.1</td>
<td>Cac_g014659.t1</td>
<td>100</td>
<td>83.77</td>
</tr>
<tr>
<td>plastid isopentenyl pyrophosphate:dimethylallyl pyrophosphate isomerase</td>
<td>IPI</td>
<td>ABW98669.1</td>
<td>Cac_g008847.t1</td>
<td>76</td>
<td>91.06</td>
</tr>
<tr>
<td>Geranyl pyrophosphate synthase</td>
<td>GPPS</td>
<td>ACC77966.1</td>
<td>Cac_g026508.t1</td>
<td>51</td>
<td>76.50</td>
</tr>
<tr>
<td>Iridoid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>geraniol 8-hydroxylase</td>
<td>G8H</td>
<td>CAC80883.1</td>
<td>Cac_g017987.t1</td>
<td>95</td>
<td>76.71</td>
</tr>
<tr>
<td>8-hydroxygeraniol oxidoreductase</td>
<td>GOR</td>
<td>AHK60836.1</td>
<td>Cac_g027560.t1</td>
<td>100</td>
<td>71.69</td>
</tr>
<tr>
<td>iridoid synthase</td>
<td>ISY</td>
<td>AFW98981.1</td>
<td>Cac_g006027.t1</td>
<td>100</td>
<td>65.65</td>
</tr>
<tr>
<td>Name</td>
<td>Accession</td>
<td>Description</td>
<td>Identity</td>
<td>Similarity</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------------</td>
<td>------------------</td>
<td>----------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>iridoid oxidase</td>
<td>IO</td>
<td>AHK60833.1</td>
<td>97</td>
<td>78.44</td>
<td></td>
</tr>
<tr>
<td>UDP-glucose iridoid glucosyltransferase</td>
<td>7DLGT</td>
<td>BAO01109.1</td>
<td>100</td>
<td>77.11</td>
<td></td>
</tr>
<tr>
<td>7-deoxyloganic acid 7-hydroxylase</td>
<td>7DLH</td>
<td>AGX93062.1</td>
<td>96</td>
<td>69.58</td>
<td></td>
</tr>
<tr>
<td>loganic acid methyltransferase</td>
<td>LAMT</td>
<td>ABW38009.1</td>
<td>95</td>
<td>53.91</td>
<td></td>
</tr>
<tr>
<td>secologanin synthase</td>
<td>SLS</td>
<td>AAA33106.1</td>
<td>99</td>
<td>64.94</td>
<td></td>
</tr>
</tbody>
</table>

Note: Only the top hit from the BLAST search is presented.
References


398 Additional files

399 Supplemental tables:

400 Table S1. RNA-sequencing libraries used in this study.

<table>
<thead>
<tr>
<th>BioProject ID</th>
<th>BioSample ID</th>
<th>Tissue</th>
<th>No. cleaned reads</th>
<th>Estimated bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255206</td>
<td>mature leaf</td>
<td>90,862,580</td>
<td>5,451,754,800</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255207</td>
<td>immature bark</td>
<td>84,537,958</td>
<td>5,072,277,480</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255208</td>
<td>root</td>
<td>88,940,668</td>
<td>5,336,440,080</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255215</td>
<td>young flower</td>
<td>71,435,806</td>
<td>4,286,148,360</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255216</td>
<td>immature fruit</td>
<td>84,250,338</td>
<td>5,055,020,280</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255217</td>
<td>mature fruit</td>
<td>47,811,342</td>
<td>2,868,680,520</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255222</td>
<td>cotyledons</td>
<td>74,037,722</td>
<td>4,442,263,320</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255223</td>
<td>upper stem</td>
<td>76,105,786</td>
<td>4,566,347,160</td>
</tr>
<tr>
<td>PRJNA80029</td>
<td>SAMN00255224</td>
<td>lower stem</td>
<td>72,680,940</td>
<td>4,360,856,400</td>
</tr>
<tr>
<td>PRJNA361128</td>
<td>SAMN06229771</td>
<td>root</td>
<td>55,435,804</td>
<td>7,224,198,331</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>771,909,254</td>
<td>49,309,244,481</td>
</tr>
</tbody>
</table>

401 Table S2. Orthologous groups of genes from *Camptotheca acuminata* and three other plant species.

402 This is available as a separate XLS file
Table S3. P450 paralogous genes in *Camptotheca acuminata*.

This is available as a separate XLS file

Table S4. Expression abundance matrix (fragments per kbp exon model per million mapped reads) from different tissues of *Camptotheca acuminata*.

This is available as a separate XLS file
Figure 2

B

Homo sapiens AAA61207
Camptotheca acuminata BAG31376
Camptotheca acuminata Cac_g012488
Camptotheca acuminata Cac_g021767
Ophiorrhiza pumila BAG31373
Ophiorrhiza liukiuensis BAG31374
Ophiorrhiza japonica BAG31375
Catharanthus roseus BAG31377
Arabidopsis thaliana NP_200341
Saccharomyces cerevisiae AAA35162

camptothecin
present
absent

Direct/indirect camptothecin binding

Homo sapiens AAA61207
Camptotheca acuminata BAG31376
Camptotheca acuminata Cac_g012488
Camptotheca acuminata Cac_g021767
Ophiorrhiza pumila BAG31373
Ophiorrhiza liukiuensis BAG31374
Ophiorrhiza japonica BAG31375
Catharanthus roseus BAG31377
Arabidopsis thaliana NP_200341
Saccharomyces cerevisiae AAA35162

Direct/indirect camptothecin binding

Ophiorrhiza pumila BAG31373
Ophiorrhiza liukiuensis BAG31374.
Ophiorrhiza japonica BAG31375
Catharanthus roseus BAG31377.
Camptotheca acuminata Cac_g021767
Camptotheca acuminata BAG31376.
Camptotheca acuminata Cac_g012488
Arabidopsis thaliana NP_200341.
Homo sapiens AAA61207
Saccharomyces cerevisiae AAA35162.

Ophiorrhiza pumila BAG31373
Ophiorrhiza liukiuensis BAG31374
Ophiorrhiza japonica BAG31375
Catharanthus roseus BAG31377
Camptotheca acuminata Cac_g021767
Camptotheca acuminata BAG31376
Camptotheca acuminata Cac_g012488
Arabidopsis thaliana NP_200341
Homo sapiens AAA61207
Saccharomyces cerevisiae AAA35162.

Figure 2

Click here to download Figure Fig_2.pptx
A. thaliana
singletons: 7,854

A. trichopoda
singletons: 11,353

C. acuminata
singletons: 8,868

C. roseus
singletons: 13,827

Figure 3
Click here to download Figure Fig_3.pptx
**Figure 4**

Click here to download Figure [Fig_4_updated.pptx](#).

**A**

<table>
<thead>
<tr>
<th>MEP</th>
<th>DXS</th>
<th>DXR</th>
<th>CMS</th>
<th>CMK</th>
<th>MCS</th>
<th>HDS</th>
<th>HDR</th>
<th>IPI</th>
<th>GPPS</th>
<th>IRIDOID</th>
</tr>
</thead>
</table>

**B**

- **GOR**
  - Cac_g027558
  - Cac_g027559
  - Cac_g027560
  - Cac_g027561

- **NMT**
  - 6 genes in between

- **7DLH**
  - 8 genes in between

- **IPP2**
  - 13 genes in between

6 genes in between
8 genes in between
13 genes in between

GOR GOR GOR GOR NMT 7DLH IPP2

---

IRIDOID

TDC

STRAS
Click here to access/download

**Supplementary Material**  
Supplemental_Table_2_DZ.xlsx