Title: Mining of the Uncharacterized Cytochrome P450 Genes Involved in Alkaloid Biosynthesis in California Poppy Using a Draft Genome Sequence

Author(s): Hori, Kentaro; Yamada, Yasuyuki; Purwanto, Ratmoyo; Minakuchi, Yohei; Toyoda, Atsushi; Hirakawa, Hideki; Sato, Fumihiko

Citation: Plant and Cell Physiology (2018), 59(2): 222-233

Issue Date: 2018-02-01

URL: http://hdl.handle.net/2433/229010

Right: © The Author(s) 2017. Published by Oxford University Press on behalf of Japanese Society of Plant Physiologists. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Type: Journal Article

Textversion: publisher

Kyoto University
Mining of the Uncharacterized Cytochrome P450 Genes Involved in Alkaloid Biosynthesis in California Poppy Using a Draft Genome Sequence

Kentaro Hori¹, Yasuyuki Yamada¹, Ratmoyo Purwanto¹, Yohei Minakuchi², Atsushi Toyoda², Hideki Hirakawa³ and Fumihiko Sato¹,∗

¹Graduate School of Biostudies, Kyoto University, Kitashirakawa, Sakyo, Kyoto, 606-8502 Japan
²National Institute for Genetics, 1111 Yata, Mishima, Shizuoka, 411-8540 Japan
³Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba, 292-0818 Japan

*Corresponding author: E-mail, fsato@lif.kyoto-u.ac.jp
(Received October 20, 2017; Accepted December 26, 2017)

Land plants produce specialized low molecular weight metabolites to adapt to various environmental stresses, such as UV radiation, pathogen infection, wounding and animal feeding damage. Due to the large variety of stresses, plants produce various chemicals, particularly plant species-specific alkaloids, through specialized biosynthetic pathways. In this study, using a draft genome sequence and querying known biosynthetic cytochrome P450 (P450) enzyme-encoding genes, we characterized the P450 genes involved in benzylisoquinoline alkaloid (BIA) biosynthesis in California poppy (Eschscholzia californica), as P450s are key enzymes involved in the diversification of specialized metabolism. Our in silico studies showed that all identified enzyme-encoding genes involved in BIA biosynthesis were found in the draft genome sequence of approximately 489 Mb, which covered approximately 97% of the whole genome (502 Mb). Further analyses showed that some P450 families involved in BIA biosynthesis, i.e., the CYP80, CYP82 and CYP719 families, were more enriched in the genome of E. californica than in the genome of Arabidopsis thaliana, a plant that does not produce BIAs. CYP82 family genes were highly abundant, so we measured the expression of CYP82 genes with respect to alkaloid accumulation in different plant tissues and two cell lines whose BIA production differs to estimate the functions of the genes. Further characterization revealed two highly homologous P450s (CYP82P2 and CYP82P3) that exhibited 10-hydroxylase activities with different substrate specificities. Here, we discuss the evolution of the P450 genes and the potential for further genome mining of the genes encoding the enzymes involved in BIA biosynthesis.

Keywords: Benzylisoquinoline alkaloid • Cytochrome P450 • Draft genome sequence • Eschscholzia californica • Genome mining • Macarpine.

Abbreviations: BBE, berberine bridge enzyme; BIA, benzylisoquinoline alkaloid; Cj, Coptis japonica; CNMT, coacinarine N-methyltransferase; CYP80B1, N-methylcoclaurine hydroxylase; CYP82N2v2, protopine 6-hydroxylase; CYP82N4, N-methylstylopane 14-hydroxylase; CYP719A2/A3, styloline synthase/canadine synthase; CYP719AS, cheilanthiforine synthase; DB10H, dihydrobenzophenanthridine alkaloid 10-hydroxylase; DB12H, dihydrobenzophenanthridine alkaloid 12-hydroxylase; DBOX, dihydrobenzophenanthridine alkaloid oxidase; Ec, Eschscholzia californica; ESI, electrospray ionization; EST, expressed sequence tag; IQA, isoquinoline alkaloid; LC, liquid chromatography; ML, high-macarpine-producing; MAL, low-macarpine-producing; MP, mate-pair; MS, mass spectrometry; MSH, N-methyllystilpine 14-hydroxylase; m/z, mass-to-charge ratio; NGS, next-generation sequencing; NR, non-redundant; OMT, O-methyltransferase; P6H, protopine 6-hydroxylase; Ps, Papaver somniferum; RNA-Seq, RNA sequencing; SOMT, scoulerine 9-O-methyltransferase; SRA, Sequence Read Archive; UTR, untranslated region.

Footnotes: The draft genome sequences were deposited in the DDBJ/GenBank/EMBL databases (accession numbers: BEHA01000001-BEHA0105253, 53,253 entries), and the annotated gene information is available at the Eschscholzia Genom Data Base (http://eschscholzia.kazusa.or.jp). The raw PE and MP read data are available from the NCBI SRA database under the BioProject accession number PRJDB6318. The nucleotide sequences of Eschscholzia californica for CYP80B5, CYP82B2, CYP82D170–172, CYP82N1, CYP82N2v3, CYP82N2v4, CYP82N5–7, CYP82P1v1, CYP82P1v2, CYP82P2–6, CYP82AT1–3, CYP82AU1, CYP82AV1 and CYP719C1 were submitted to DDBJ/GenBank/EMBL under the accession numbers LC316736–LC316762.

Introduction

Land plants produce specialized low molecular weight metabolites to adapt to various environmental stresses, such as UV radiation, pathogen infection, wounding and damage caused by animal feeding (Sato 2014, Wurtzel and Kutchan 2016). Due to the variations in...
stresses, plants produce diverse chemicals, especially plant species-specific alkaloids, through specialized biosynthesis pathways (Facchini 2001, Facchini et al. 2004, Ziegler and Facchini 2008). For example, isoquinoline alkaloids (IQAs) are specifically found in the Ranunculaceae, Papaveraceae, Berberidaceae and Rutaceae families, among others, and often consist of chemically diverse structures that are specific to plant species, including morphinans (e.g. narcotic morphine in opium poppy), emetines (e.g. anti-moebic emetine in ipecac) and galantamines (e.g. acetylcholineesterase inhibiting galantamine in snowdrop), as well as more generally distributed protoberberines (e.g. antibiotic berberine in the Berberidaceae, Papaveraceae, Ranunculaceae and Rutaceae families), benzophenanthridines (e.g. antibiotic sanguinarine in Papaveraceae), aporphines (e.g. the potent anti-HIV agent magnoflorine in Ranunculaceae) and pavines (e.g. eschscholizine in Papaveraceae), among others (Hagel and Facchini 2013). Biosynthetic pathways involved in specialized metabolism have thus far been characterized using metabolite identification and tracer experiments. Biosynthetic enzymes have been characterized and purified from metabolite-producing plant organs and/or in vitro cultured cells to confirm the pathways, and cDNAs for biosynthetic enzymes have been isolated (Sato 2013, Sato 2014). For example, in the 1970s, Zenk’s group characterized many IQA biosynthetic pathways, after which several groups purified their main enzymes and isolated the cDNA of those enzymes using the sequences of purified proteins involved in IQA biosynthesis (Tanahashi and Zenk 1990, Studler and Zenk 1993, Kraus and Kutchan 1995, Takeshita et al. 1995, Morishige et al. 2000, Choi et al. 2002, Liscombe et al. 2009). Due to advances in molecular technologies, especially DNA sequencing, expressed sequence tag (EST) libraries [later involving transcriptome data produced by using next-generation sequencing (NGS)] were used to select several candidate enzyme-encoding genes based on the signature sequences of O-methyltransferase (OMT) and cytochrome P450 (P450) (Morishige et al. 2002, Ikezawa et al. 2007, Ikezawa et al. 2009), the relative abundances of transcripts or the results of correlation analyses between the accumulation of metabolites and transcripts (Liscombe et al. 2009, Takemura et al. 2013). Integrated transcriptome and metabolome analyses have been especially powerful tools for selecting candidates, particularly when plant clones and/or cell culture lines with different levels of metabolite productivity are used (Liscombe et al. 2009, Takemura et al. 2013); however, preparing adequate cell and plant collections is difficult. Therefore, we attempted to isolate candidate genes directly from the genome sequence because tomato and potato genome information has been successfully used both to analyze the steroidal glycoalkaloid biosynthesis enzyme-encoding gene cluster and to isolate candidate genes in those plant species (Itkin et al. 2013).

In this study, we used *Eschscholzia californica* cv. Hitozeaki (California poppy), a member of the Papaveraceae family that produces various types of benzylisoquinoline alkaloids (BIAs), i.e. aporphine-, pavine-, protoberberine-, protopine- and benzophenanthridine-type alkaloids (Fabre et al. 2000). The major alkaloid biosynthesis pathways and biosynthetic enzymes of these BIAs have been characterized at the molecular level (Hagel and Facchini 2013, Sato 2013). In fact, benzophenanthridine alkaloid biosynthesis from (S)-reticuline to dihydrosanguinarine has been characterized at the genetic level (Dittrich and Kutchan 1991, Ikezawa et al. 2007, Liscombe and Facchini 2007, Ikezawa et al. 2009, Beaudoin and Facchini 2013, Takemura et al. 2013), and this information has been used to reconstruct its biosynthetic pathway (Fossati et al. 2014, Hori et al. 2016). However, the genes involved in the biosynthesis of macarpine have not been characterized, although two P450s and two OMTs (Fig. 1), i.e. dihydrobenzophenanthridine alkaloid 10-hydroxylase (DB10H), dihydrobenzophenanthridine alkaloid 12-hydroxylase (DB12H), 10-hydroxydihydrobenzophenanthridine alkaloid 10-O-methyltransferase (10OMT) and 12-hydroxydihydrobenzophenanthridine alkaloid 12-O-methyltransferase (12OMT), involved in macarpine biosynthesis have been identified (De-Eknukul et al. 1992, Kammerer et al. 1994). Macarpine biosynthesis is especially interesting because its biosynthetic pathway from reticuline involves six P450s: cheilanthiflorine synthase (CYP719A5), stylophone synthase/canadine synthase (CYP719A2/A3), N-methylstylophone hydroxylase (CYP82N5), protopine 6-hydroxylase (CYP82N2v2), DB10H and DB12H. Additionally, the anti-proliferative effects of macarpine on several cancer cell lines have received increased attention (Slaninová et al. 2007).

Because the evolution of specialized metabolism is enhanced by the diversification of key enzymes such as P450 (Mizutani and Sato 2011, Nelson and Werpck-Reichhart 2011), we also focused on the diversification of the P450 family in our draft genome sequence research. Our results suggest that draft genome sequencing can be a useful resource both for mining the biosynthetic enzyme genes involved in specialized metabolism and for developing tools for bioengineering. In particular, the identification of many silenced genes suggests that plant cells have the potential to expand their specialized metabolism, as reported in microbial cell systems (Ozaki et al. 2013, Rutledge and Challis 2015).

The full and/or comparative sequencing of related genomes is usually used in these types of studies. By using Illumina sequencing and genome assembly, we obtained an *E. californica* draft genome sequence that covered >95% of the whole genome (502 Mb). After the assembly, we searched the genome sequence and its gene clusters for sequences related to BIA biosynthesis using known biosynthetic enzyme genes. Our search results indicated that all characterized biosynthetic enzyme genes were found in the draft genome, indicating that the draft sequence sufficiently covered the genome; however, a gene cluster for BIA biosynthesis was not found. Further characterization of the identified P450 genes indicated that the California poppy genome was enriched with genes belonging to the CYP80, CYP82 and CYP719 families; these genes are involved in BIA biosynthesis in *E. californica* but not in *Arabidopsis thaliana*. Notably, CYP82 genes were clustered in some scaffolds, and the results of additional expression analyses suggested that these genes functioned differently in plant tissues and cultured cells with unique alkaloid profiles. A functional analysis using transgenic yeast revealed that CYP82P2 and CYP82P3 function as 10-hydroxylases in dihydr benzophenanthridine alkaloid biosynthesis and exhibit different substrate specificities and expression profiles. The molecular evolution of P450 and metabolic diversification of BIA biosynthesis are also discussed.
Results
Draft genome sequencing and assembly

After the genomic DNA was digested, 150 bp of several paired-end (PE) and mate-pair (MP) libraries (PE300, PE500, MP3k, MP5k, MP8k, MP10k, MP15k and MP20k) containing average insert sizes of 342–15,125 bp were sequenced via Illumina sequencing 610 million, 510 million, 220 million, 250 million, 120 million, 360 million, 130 million and 170 million times, respectively, producing a coverage value 790 times the genome size (Supplementary Table S1). The sequences were then assembled by Platanus 1.2.1.1 (Kajitani et al. 2014) after the removal of adaptor sequences and an error correction by SOAPec v2.01 using a k-mer size of 27, a low-frequency k-mer cut-off of 19 and manual correction methods. The sequences were ultimately assembled into 53,253 scaffolds that covered approximately 489 Mb of the California poppy genome (approximately 97% of the 502 Mb genome) (Cui et al. 2006) (Table 1). The N50 of the scaffolds consisted of approximately 753 kb, and the longest was 4 Mb. To validate the completeness of the sequences, the EST and RNA sequencing data of E. californica from NCBI (http://www.ncbi.nlm.nih.gov/) and PhytoMetaSyn (http://www.phytometasyn.ca/) databases were mapped onto the draft genome sequence. The mapping results indicated that >90% of the RNA sequences (approximately 80,000 in total) were mapped onto the draft genome and covered >50% of the known RNA sequences (Supplementary Table S2). Gene prediction was also performed with the RNA sequence data obtained from the NCBI Sequence
We searched for orthologs of the biosynthetic enzyme genes involved in the synthesis of morphinan alkaloids (e.g. CYP719B1; Gesell et al. 2009) and gene clusters associated with noscapine alkaloid biosynthesis (Winzer et al. 2012); such gene clusters are specific to the opium poppy. Our search did not identify any of these genes, indicating that morphinan and noscapine biosyntheses are specific to opium poppy at the genetic level. Additional searches for clusters of the biosynthetic enzyme genes involved in BIA biosynthesis also yielded no matches in the scaffolds (Supplementary Table S3), except for the identified CYP82 gene clusters, which are discussed later (Supplementary Table S5; Supplementary Fig. S2).

### Analysis of P450 families

In BIA biosynthesis and many other specialized metabolite biosynthesis pathways, P450s are important for determining chemical diversity in metabolism (Nelson and Werck-Reichhart 2011). Plants are highly enriched with P450s; thus far, 135 families (http://drnelson.uthsc.edu/biblioD.html) have been reported. Therefore, using tblastn searches, we estimated the amount of P450 family genes in the *E. californica* draft genome based on sequence similarity. We also compared the number of P450 family genes with that of *A. thaliana*, as only three P450 families, i.e. CYP80, CYP82 and CYP719, have been identified as components of BIA biosynthesis (CYP80B1, Pauli and Kutchan 1998; CYP719A1, Ikezawa et al. 2003; CYP719A5, Ikezawa et al. 2007; CYP80G2, Ikezawa et al. 2008; CYP719B1, Gesell et al. 2009; CYP719A2/A3, Ikezawa et al. 2009; CYP82N4 (M5H), Beaudoin and Facchini 2013; CYP82N2v2 (P6H), Takemura et al. 2013; CYP82Y1, Dang and Facchini 2014). As shown in Supplementary Table S4, the numbers of some of the P450 families differed between the plant species. Specifically, CYP80, CYP82 and CYP719, which are involved in BIA biosynthesis, were more enriched in *E. californica* than in *A. thaliana*; these families were named by

### Table 1 Eschscholzia californica draft genome assembly statistics

<table>
<thead>
<tr>
<th>Draft genome sequence (ECA_r1.0)</th>
<th>Predicted genes (ECA_r1.0_cds_braker1)</th>
<th>ECA_r1.0_cds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total No. of sequences</td>
<td>53,253</td>
<td>93,339</td>
</tr>
<tr>
<td>Average (bases)</td>
<td>489,064,912</td>
<td>86,538,158</td>
</tr>
<tr>
<td>Maximum (bases)</td>
<td>4,249,866</td>
<td>927</td>
</tr>
<tr>
<td>Minimum (bases)</td>
<td>30</td>
<td>16,281</td>
</tr>
<tr>
<td>N50 (bases)</td>
<td>752,971</td>
<td>300</td>
</tr>
<tr>
<td>G + C%</td>
<td>36.5</td>
<td>1,263</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;500 bases No. of sequences</td>
<td>37,795</td>
<td>59,645</td>
</tr>
<tr>
<td>Total length (bases)</td>
<td>482,984,812</td>
<td>74,697,139</td>
</tr>
<tr>
<td>Average (bases)</td>
<td>12,779</td>
<td>1,252</td>
</tr>
<tr>
<td>Maximum (bases)</td>
<td>29,978</td>
<td>1,798</td>
</tr>
<tr>
<td>Minimum (bases)</td>
<td>1,718</td>
<td>1,718</td>
</tr>
<tr>
<td>&gt;1 kb No. of sequences</td>
<td>15,580</td>
<td>1,798</td>
</tr>
<tr>
<td>Total length (bases)</td>
<td>467,061,688</td>
<td>28,552,063</td>
</tr>
<tr>
<td>Average (bases)</td>
<td>29,978</td>
<td>1,718</td>
</tr>
<tr>
<td>Maximum (bases)</td>
<td>28,552,063</td>
<td>1,718</td>
</tr>
<tr>
<td>&gt;5 kb No. of sequences</td>
<td>2,223</td>
<td>199</td>
</tr>
<tr>
<td>Total length (bases)</td>
<td>444,303,297</td>
<td>1,266,048</td>
</tr>
<tr>
<td>Average (bases)</td>
<td>199,867</td>
<td>1,266,048</td>
</tr>
<tr>
<td>Maximum (bases)</td>
<td>4,249,866</td>
<td>927</td>
</tr>
</tbody>
</table>

Read Archive (SRA) database (accession: SRX096037) and by Augustus v3.0.3, as described in the Materials and Methods. The genes predicted by Augustus were further annotated by similarity searches against the NCBI non-redundant (NR) database. The gene set that included the genes predicted by Augustus was named CSE_r1.0_cds_braker1. Alternatively, the gene set excluding the genes without similarity matches against the NR database but with similarity matches against transposable elements and pseudogenes was named CSE_r1.0_cds. The numbers of the genes in the CSE_r1.0_cds_braker1 and CSE_r1.0_cds groups are shown in Table 1. The draft genome sequences were deposited in the DDBJ/GenBank/EMBL databases (accession Nos. BEHA01000001–BEHA01053253; 53,253 entries), and the annotated gene information is available at the Eschscholzia Genome DataBase (http://eschscholzia.kazusa.or.jp).

We further searched the genome for previously known genes that are involved in BIA biosynthesis, including reticuline and sanguinarine biosynthesis. Queries with tblastn (Altschul et al. 1990) revealed that all the BIA biosynthetic enzyme genes except CYP719A3 were in the draft genome (Supplementary Table S3). However, a CYP719A3 gene with an unfixed N-sequence was also found in the draft genome, and the presence of this gene was confirmed by genomic PCR and sequencing. This finding indicates that the draft genome is a sufficient platform for further sequence analyses. In fact, when the coclaurine finding indicates that the draft genome is a sufficient platform for further sequence analyses. In fact, when the coclaurine
the P450 nomenclature committee (Dr. D.R. Nelson, University of Tennessee, Memphis, TN, USA; Supplementary Table S5). Among these P450s, the CYP82 family was the most diversified.

The CYP80 family contained CYP80B1 (N-methylnorcoclaurine hydroxylase, Pauli and Kutchan 1998) and an uncharacterized CYP80 gene (named CYP80B5) that had a 70% amino acid sequence identity to CYP80B1. The CYP719 family consisted of seven genes, one of which exhibited 95% homology with CYP719A5. On the other hand, the CYP82 family consisted of 28 genes, of which only three (CYP82N2v2, CYP82NS and CYP82Y1) have been biochemically characterized; another four genes could be potential homologs of previously identified genes. These results indicate that the CYP82 family is rich in uncharacterized biosynthetic enzyme genes that are involved in BIA biosynthesis. Many uncharacterized CYP82 genes also formed gene clusters (Supplementary Fig. S2). Although CYP82P3 contained an unfixed N-sequence in the draft genome, its full sequence was determined by genomic PCR/reverse transcription–PCR and nucleotide sequencing. PseudoCYP82 was determined to be a pseudogene due to the presence of a stop codon in the coding frame.

When the genome structure, including the presence of introns, was analyzed, all genes in the CYP719 family and some CYP82 genes, such as protopine 6-hydroxylase (CYP82N2v2 and its homologs), lacked introns (Supplementary Table S6). On the other hand, CYP80 and many CYP82 family genes contained an intron in the region of their fourth substrate recognition site.

The phylogenetic analysis of the CYP82 genes showed that they were quite diverse (Supplementary Fig. S3), and some CYP82 subfamily genes, e.g. CYP82P (P2, P4–P6), CYP82B (B2–B4) and CYP82D (D170, D171), clustered in scaffolds, including Eca_sc000081.1, Eca_sc007583.1 and Eca_sc006699.2 (Supplementary Fig. S2). To characterize the functions of these uncharacterized CYP82 genes, we analyzed their expression in both plant tissues and cell lines that exhibited different BIA production abilities.

CYP82 family genes show diverse expression and function

Using specific gene primers, we analyzed the expression of uncharacterized CYP80 and CYP82 genes, except for CYP80B1, CYP82N2v2 and CYP82NS and their homologs (i.e. CYP80B5, CYP82N1, CYP82N2v3 and CYP82N2v4) (Supplementary Table S7). Due to their high sequence similarity, CYP82P1v1 and CYP82P1v2 were analyzed as a set of genes. First, we examined their expression in E. californica plant tissues that exhibited different alkaloid accumulation profiles (Supplementary Fig. S4). Our previous analysis indicated that roots, which accumulate benzophaenanthridine alkaloids, showed high expression levels of sanguinarine biosynthetic genes such as CYP80B1, N-methylcoclaurine 4’-O-methyltransferase (4’OMT), berberine bridge enzyme (BBE), CYP719A5, CYP719A2 and CYP719A3 (Ikezawa et al. 2007, Ikezawa et al. 2009). Similarly, genes involved in macarpine biosynthesis, such as DB10H and DB12H, were expected to be expressed in the roots in which macarpine accumulated. Aerial parts, such as leaves and petioles, are known to accumulate pavine-type alkaloids such as escholtzine (Fabre et al. 2000), so pavine-type alkaloid biosynthetic enzyme-encoding genes were expected to be expressed in those tissues.

As shown in Fig. 2, each P450 gene showed a distinct expression profile. CYP82N7 was specifically expressed in the roots and cultured cells. CYP82D170, CYP82P1v1/CYP82P1v2 and CYP82P2 were expressed in the petioles, roots and cultured cells. In contrast, CYP82B2, CYP82B3, CYP82B4 and CYP82P6 were not expressed in the roots but were highly expressed in the aerial parts. CYP82D171 and CYP82P3 were preferentially expressed in the wild-type cultured cells. There was no accumulation of the CYP82AT2, CYP82AU1, CYP82AV1, CYP82D172, CYP82N6, CYP82P4 and pseudoCYP82 transcripts. The data for CYP80B5, CYP82AT1 and CYP82P5 transcripts were not obtained due to large replication errors.

Next, we analyzed the expression of CYP82 genes in two cultured E. californica cell lines with differences in macarpine productivity: a low-macarpine-producing (ML) cell line and a high-macarpine-producing (MH) cell line (Supplementary Fig. S5). The MH cells showed exclusive or higher expression of CYP82D170, CYP82P1v1/CYP82P1v2, CYP82P2 and CYP82P5 than the ML cells, whereas the ML cells showed higher expression of CYP82AT3, CYP82B2, CYP82D171, CYP82N7, CYP82P3 and CYP82N7 than the MH cells (Fig. 3). No expression of CYP82AT1, CYP82AT2, CYP82AU1, CYP82AV1, CYP82B3, CYP82B4, CYP82P6 or pseudoCYP82 was observed in either cell line. The data for CYP82D172 and CYP82N6 were not obtained due to large replication errors.

Enzymatic activity of novel P450s

Using the expression profile data, we characterized the enzymatic activity of the macarpine biosynthesis candidate genes, i.e. CYP82D170, CYP82N7, CYP82P1v1, CYP82P2 and CYP82P5, which were clearly expressed in the MH line and roots. Full-length cDNAs were prepared from the cDNA library by reverse transcription–PCR using gene-specific primer pairs (Supplementary Table S8) designed from the draft genome sequence; expression vectors of these cDNAs for Saccharomyces cerevisiae were constructed to prepare recombinant microsomal fractions to measure enzymatic activities.

When the recombinant microsomal fractions reacted with 6-benzophenanthridine alkaloid substrates (dihydrochelerythrine, dihydrochelirubine, dihydrosanguinarine, chelerythrine, chelirubine and sanguinarine; Supplementary Fig. S6), only the microsomal fraction expressing CYP82P2 produced a new hydrophilic peak [mass-to-charge ratio (m/z) increase of 16] from dihydrosanguinarine (m/z 334), suggesting that a hydroxylation reaction occurred (Supplementary Fig. S7). Transgenic Pichia pastoris cells expressing CYP82P2 confirmed this reaction (Fig. 4a–d). The co-culture of transgenic Pichia cells expressing the CYP82P2 and G11 O-methyltransferase isolated from MH cells (Purwanto et al. 2017) produced another new peak that had an identical retention time and mass spectrometry (MS) fragment pattern to the dihydrochelirubine (Fig. 4e–h), and the dihydrosanguinarine 7-hydroxylase function of CYP82P2 was confirmed. However, a recombinant CYP82P2
neither hydroxylated dihydrochelerythrine nor converted dihydrochelirubine, indicating that other P450s, such as 12-hydroxylase and/or dihydrochelerythrine 10-hydroxylase, exist.

Based on the finding that CYP82P2 is a dihydrochelerythrine 10-hydroxylase, we searched for additional hydroxylases involved in macarpine biosynthesis and for dihydrochelerythrine 10-hydroxylases. Because none of the five candidates (CYP82D170, CYP82N7, CYP82P1v1, CYP82P2 and CYP82P5) showed hydroxylation activity for dihydrochelirubine, we searched for dihydrochelerythrine 10-hydroxylase, which produces 10-hydroxychelerythrine, in the ML cells. Although CYP82P4 was a good candidate for these reactions because it was located in the same scaffold as CYP82P2 and exhibits high sequence homology, the full-length CYP82P4 cDNA could not be isolated from the cDNA library. Nevertheless, the highly homologous CYP82P3 gene was amplified from 10-hydroxychelerythrine-producing ML cells, and its enzymatic activity was examined.

When transgenic S. cerevisiae cells expressing CYP82P3 were prepared and reacted with the substrate alkaloids, the cells not only formed 10-hydroxydihydrochelerythrine but also formed 10-hydroxylated dihydrochelerythrine (m/z 366) from the corresponding substrates (Fig. 5). The products were identified by their retention times and MS fragment patterns in comparison with those of authentically prepared 10-hydroxydihydrochelerythrine. CYP82P3 did not show any enzymatic activity against other benzophenanthridine alkaloids, i.e. sanguinarine, chelerythrine, chelirubine or dihydrochelirubine.

**Sequence characteristics of uncharacterized P450s**

Two isolated P450s, CYP82P2 and CYP82P3, shared high sequence similarity (Supplementary Figs. S3, S8) but showed different expression profiles and substrate specificities (Figs. 2–5). Furthermore, some P450 genes showed no expression in the examined plant tissues and cell cultures. Although heterologous expression in microbial cells is useful for examining the biological function of enzymes, recombinant protein production can be difficult. In fact, CYP82D171 was not transformed in the S. cerevisiae cells. Therefore, we examined the sequence characteristics of the uncharacterized P450s.

First, we examined the conservation of P450 signatures in several candidate CYP82 P450s that were found in this study, e.g., CYP82D170, CYP82N7, CYP82N3, CYP82N7, CYP82P1v1, CYP82P2, CYP82P3, CYP82P5 and CYP82P6. These P450s, except for CYP82N3, CYP82P1v1 and CYP82P6, contained the conserved domains of eukaryotic P450s, i.e. the K-helix region, aromatic region and heme-binding domain, and there was a conserved threonine residue in the I-helix region of the monoxygenase-type P450s (Supplementary Fig. S8). In CYP82N3, a conserved threonine was replaced with serine, and, in CYP82P1v1 and CYP82P6, a serine replacement occurred at a conserved glycine/alanine residue found in non-canonical methylenedioxy-forming enzymes (CYP719A1, CYP719A2, CYP719A3 and CYP719A5) and the C–C phenol-coupling enzyme (CYP80G2). These modifications suggest that CYP82N3, CYP82P1v1 and CYP82P6 may have unique functions. However, other P450s such as CYP82D170, CYP82D171, CYP82N7 and CYP82P5 had conserved threonine and glycine/alanine residues, and these enzymes, as well as CYP82P2 and CYP82P3, were expected to have general functions as hydroxylases. More biochemical and physiological characterizations are needed to identify the remaining hydroxylases involved in macarpine and pavine biosynthesis and their physiological roles in BIA biosynthesis.

![Fig. 2 Expression analysis of novel P450 genes in E. californica plant tissues. Transcript expression was measured by quantitative real-time PCR using the cDNAs of the plant tissues and wild-type cultured cells of E. californica. The relative expression levels were calculated against the values of the β-actin gene. Error bars indicate the SD calculated from three technical replicates.](https://academic.oup.com/pcp/advance-article-abstract/doi/10.1093/pcp/pcx210/4781735)
Discussion

Due to the rapid progress of NGS technology, the whole-genome sequencing of non-model organisms is more achievable, although many obstacles still exist. In this study, we used a commercial variety of California poppy (*E. californica* cv. Hitoezaki) to determine the California poppy draft genome sequence. Despite the heterozygosity of the genome of this plant species, the genome sequence obtained by this study still contained many unfixed N-sequences and >53,000 scaffolds, and the draft sequence covered 95% of the genome (approximately 489 Mb of 502 Mb) and transcriptome data. Heterozygosity and unfixed N-sequences hindered the identification of genes, but the current data still sufficiently covered the biosynthetic enzyme genes involved in the BIA pathway (*Table 1*), and the N-sequences found in CYP719A3 and CYP82P3 could be fixed by the amplification of the genomic DNA after nucleotide sequencing. These results indicated that NGS-based draft genome sequencing can provide a sufficient platform for the characterization of specialized metabolism; specialized metabolism is diversified, and its expression is limited to specific cells or tissues and affected by environmental conditions.

The absence of both a gene cluster involved in noscapine biosynthesis (e.g. CYP82X2 and CYP82Y1, Winzer et al. 2013) and the morphinan alkaloid gene (e.g. CYP719B1, Gesell et al. 2008) in the draft genome of California poppy also indicated that noscapine and morphinan biosynthetic enzyme-encoding genes are specific to the opium poppy, despite the identification of some homologs in the California poppy genome. Our draft genome sequence is a good starting platform for dissecting the molecular evolution of BIA biosynthesis in Papaveraceae and comparing results with the recently published draft genome sequence of *Macleaya cordata*, which is also a member of the Papaveraceae family that produces BIAs (Liu et al. 2017). The *M. cordata* draft genome sequence also showed an absence of noscapine- and morphine-specific genes. The genome sequence of *Aquilegia coerulea*, an IQA-producing member of the Ranunculaceae family (https://phytozome.jgi.doe.gov/pz/portal.html#info?alias=Org_Aceroerulea), was recently reported. These genome sequences enable future analyses of the evolution of BIA biosynthesis in the Papaveraceae and Ranunculaceae families.

As the present study indicates, draft genome sequencing is especially useful for analyzing individual genes, such as the P450s involved in BIA biosynthesis. We estimated all the possible P450 genes in the draft genome and clarified the diversification of P450 family genes, i.e. the CYP80, CYP82 and CYP719 genes involved in BIA biosynthesis in *E. californica*. In particular, the CYP82 genes formed gene clusters within scaffolds (*Supplementary Fig. S2*), and the evolution of these genes via gene duplication was estimated by their high sequence similarity and conservation of exon–intron junction sites (*Supplementary Table S6*). However, the results of the analysis of individual gene expression showed diverse expression profiles among the plant tissues and cultured cells (*Figs. 2, 3*). Furthermore, the functional analysis of P450 homologs showed functional diversification, similar to the case of CYP82P2 and CYP82P3, which function as dihydrosanguinarine 10-hydroxylase and dihydrosanguinarine/dihydrochelerythrine 10-hydroxylase, respectively, in BIA biosynthesis (*Figs. 4, 5*). Interestingly, the expression of the CYP82P2 and CYP82P3 genes was correlated with the alkaloid profiles in each cell; the high expression of CYP82P2 induced macarpine production in the MH cells, and the CYP82P3 expression induced 10-hydroxychelerythrine production in the ML cells. The relative expression levels of CYP82P2 and CYP82P3 in the roots were lower than those in the cultured cells, and the accumulation of macarpine was low in the roots. These data suggest that CYP82P2/P3 (DB10H) proteins are important key regulators in macarpine biosynthesis.

In this study, we characterized P450 genes based on a draft genome sequence and indicated the diversification of both their
expression patterns and functions in plant specialized metabolism. Importantly, many genes were not expressed under the current analytical conditions. Functional analyses of these non-expressed and uncharacterized genes using our alkaloid production platform and microbial systems (Minami et al. 2008, Hori et al. 2016, Nakagawa et al. 2016, Ehrenwoth and Peralta-Yahya 2017) would be useful to understand the future molecular evolution of secondary metabolism and the development of more diversified metabolite production, as shown in microbial genomes by the activation of potential genes (Ozaki et al. 2013).

**Materials and Methods**

**Plant materials**

*Eschscholzia californica* (cultivar Hitoezaki, Takii Seed Co., Ltd.) plants were grown in soil at 23°C under white light (100–200 μmol photons m⁻² s⁻¹) and a 10 h light/14 h dark photoperiod, and were used for the genomic DNA extraction. The total RNA for the expression analysis was prepared from the leaf blades, petioles and roots of plants grown for 5 months. Suspension-cultured wild-type, S-38 line [ML cell line expressing the scoulerine 9-O-methyltransferase (*SMT*) gene of *Coptis japonica* (Takemura et al., 2010b)] and A5-1 line [MH cell line expressing the EcCYP719A5 gene (Takemura et al., 2010a)] *E. californica*...
cells were grown in Linsmaier–Skoog medium (pH 5.7) (Linsmaier and Skoog 1965) containing 10 µM naphthaleneacetic acid, 1 µM benzyladenine and 3% sucrose at 23°C in the dark at 3 week intervals. ML and MH cells grown for 2 d and wild-type cells grown for 2 weeks were used for the total RNA preparation.

**Chemicals**

Sanguinarine was purchased from Sigma-Aldrich Co., LLC. N-Methyltetrahydroberberine, protopine, allocryptopine and chelerythrine were retrieved from the collection of Dr. K. Iwasa. 10-Hydroxychelerythrine and chelirubine were isolated from the ML cells and the cultured *E. californica* cells with the ectopic expression of the *C. japonica* WRKY1 gene, respectively, using a CombiFlash RF+ (Teledyne Isco) purification system, and the identities of these compounds were determined based on their retention times and *m/z* values by liquid chromatography–MS (LC-MS), as previously described (Takemura et al. 2010b). 10-Hydroxydihydrochelerythrine, dihydrochelirubine, dihydroxyguinarine and dihydroxychelerythrine were prepared by a NaBH₄ reduction from 10-hydroxydihydrochelerythrine, chelirubine, sanguinarine and cheler- ythrine, respectively.

**Metabolite analysis**

Alkaloids were extracted in methanol with 0.01 N HCl and analyzed using an LC-MS 2020 system or an LC-MS 8030 system (Shimadzu Corp.) under the following conditions: electrospray ionization (ESI)-MS, 1.5 kV, positive ion mode; column, TSgel 80-TM (4.6×250 mm; TOSOH2); temperature, 40°C; mobile phase, aceto- nitride (A) and water containing 1% (v/v) acetic acid. The elution profile was as follows: 0–15 min, isocratic elution with 40% (v/v) A; 15–18 min, linear gradient from 40% to 80% (v/v) A; 18–40 min, isocratic elution with 80% (v/v) A; 40–43 min, linear gradient from 80% to 40% (v/v) A; and 43–65 min, re-equilibration with 40% (v/v) A. The MS fragment spectra of the alkaloids were detected using the LC-MS 8030 system under the following conditions: ESI-MS, product ion scan mode, *m/z* of 10.00–400.00 and collision energy at −35.0 V. The other conditions were the same as those described above.

**Eschscholzia californica** genome sequencing and assembly

The *E. californica* genomic DNA (50 µg at 100 ng µl⁻¹) was prepared using a NucleoSpin Plant MIDI Kit (TAKARA BIO INC.). The genomic DNA was digested, and 150 bp of several PE and MP libraries (PF300, PE500, MP3k, MP5k, MP8k, MP10, MP15k and MP20k) containing average insert sizes of 342, 545, 2,820, 4,545, 7,926, 8,945, 13,158 and 15,125 bp, respectively, were prepared using a TruSeq DNA Sample Prep Kit and a Nextera Mate Pair Sample Prep Kit (Illumina) according to the manufacturer’s instructions; these libraries were sequenced using an Illumina Hiseq 2500 sequencer 610 million, 510 million, 220 million, 250 million, 120 million, 360 million, 130 million and 170 million times, respectively, to produce a 790 times genome coverage (*Supplementary Table S1*). The raw sequence reads were filtered by both the trimming of adaptor sequences in the reads and error correction using SOAPec v2.01 with a k-mer size of 27 and a low-frequency k-mer cut-off value of 19. The sequence reads were assembled by PLATANUS v1.2.11 (Kajitani et al. 2014) and edited manually. Scaffolds longer than 300 bp were extracted. Scaffolds that matched the genome sequences of bacteria, fungi or humans in NCBI, vector sequences (UniVec in NCBI), PhiX sequences (Illumina) or the chloroplast...
Gene prediction

The RNA sequencing (RNA-Seq) data sampled from the California poppy roots for gene prediction were obtained from the NCBI SRA database (accession: SRX096037). Three nucleotides from the 3' terminus of all of the reads were trimmed because the average quality value (QV) for this region was >20. Fewer than 10 nucleotides at the 3' terminus were further trimmed, and the reads with unknown nucleotides (N) were excluded by using PRINSEQ v0.20.4 (Schmieder and Edwards 2011). In addition, the reads that included >10 bases of nucleotides from the 5' terminus of the adaptor sequence (ACATCGGAAGAAGC) at the 3' terminus were excluded by using the FASTX-toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/). After being trimmed and filtered, the RNA-Seq reads were mapped onto the scaffolds using TopHat v2.0.14 (Trapnell et al. 2009), and the resulting BAM file was processed with BRAKER1 v1.3 (Hoff et al. 2016) to construct a training set for gene prediction. Gene prediction was performed by Augustus v3.0.3 (Stanke et al. 2009) with the training set.

Annotation

The genes predicted by Augustus were subjected to similarity searches against the NCBI NR database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) by BLAST, with an E-value cut-off of 1E-10; domain searches against the InterPro database (Altschul et al. 1997) to construct a training set for gene prediction. Gene prediction was performed by Augustus v3.0.3 (Stanke et al. 2009) with the training set.

Estimation of P450 genes

Arabidopsis thaliana P450 genes were obtained from The Arabidopsis Information Resource (http://www.arabidopsis.org/index.jsp). Eschscholzia californica P450s were estimated by tblastx searches using previously known P450 amino acid sequences as the queries, with the exception of sequences from the CYP95, 713, 737, 740, 733–797 and 800–805 families, which were not available to the public. Partial sequences and N-containing sequences were removed from the results. The introns, start codons and stop codons of CYP80, CYP82 and CYP719 family genes were estimated based on comparisons with cDNA sequences. The CYP80 and CYP82 genes were estimated from the draft genome and were named by the P450 nomenclature committee (Dr. D.R. Nelson, University of Tennessee) (Supplementary Table S5). The nucleotide sequences for CYP80B8, CYP82B2–5, CYP82D170–172, CYP82N1, CYP82N2x1, CYP82N2nx4, CYP82N5–7, CYP82P1v1, CYP82P2v2–6, CYP82A7–3, CYP82A1U, CYP82AV1 and CYP719C1 were submitted to DDBJ/GenBank/EMBL under the accession numbers LC316736–LC316762.

Genomic PCR

To determine the identity of the N-sequences, the E. californica genomic DNAs for CYP719A3, CYP82P2 and CYP82P4 were analyzed by genomic PCR using the primer sequences shown in Supplementary Table S5. CYP82P3 and CYP82P4 were amplified by nested PCR using gene-specific primer pairs.

Alignment analysis of the CYP82 family

Representative plant P450 amino acid sequences were obtained from GenBank for the tree construction, whereas EC-CYP82N5 (N-methylstylopine hydroxylase) was newly identified in this study. The accession numbers for the amino acid sequences were as follows: AAC39454.1, E. californica EcCYP82B1 (wrongly annotated N-methylccoumarine 3'-hydroxylase); BAK20464.1, EcCYP82B2v2 (P6H); AGC92397.1, P. somniferum PsCYP82N3 (P6H); AGC92398.1, PsCYP82N4 (MSH); AFB74614.1, PsCYP82X1 (P450 involved in noscapine biosynthesis); and uncharacterized P450s (AFB74616.1, PsCYP82X2; AFB74617.1, PsCYP82Y1 (N-methylstyrione 1'-hydroxylase); BAF98427.1, C. japonica CjCYP82P1; NP_19492S.1, A. thaliana AtCYP82C2; NP_194923.1, AtCYP82C3; NP_194922.1, AtCYP82C4; NP_180888.1, AtCYP82F1; NP_189154.1 and PsCYP82G1). The CYP82 sequences were aligned by ClustalW (Thompson et al. 1994), and a phylogenetic tree was generated by MEGA6 (using the Neighbor-Joining statistical method, a bootstrap test involving 1,000 replicates and the Jones–Taylor–Thompson model) (Tamura et al. 2013).

Quantitative real-time PCR

The total RNA was extracted from 2-day-old cultured E. californica cells (ML and MH cells), 2-week-old cultured E. californica cells (wild types) and the tissues (leaf blade, petiole and roots) of the plants that had been growing for 5 months. RNA extraction was performed using an RNasy plant Mini Kit (Qiagen). Single-stranded cDNAs were synthesized from 1 mg of total RNA using a PremixScript RT reagent Kit (TAKARA BIO INC.). Quantitative real-time PCR was performed with gene-specific primer pairs (Supplementary Table S7) using the iQ™ SYBR Green Supermix and a CFX96 Real-Time PCR Detection System (Bio-Rad Laboratories Inc.) under the following PCR conditions: 95°C for 3 min, followed by 40 cycles of 95°C for 10s, 55°C for 20s and 72°C for 20s. The relative expression levels between the samples were calculated by using the standard curve method with the expression level of β-actin as an internal standard.

Isolation of the full-length cDNAs of novel P450s

The full-length cDNAs of novel P450s were amplified by PCR from the total cDNA that was used for the expression analysis using the primer pairs shown in Supplementary Table S7. CYP82D170, CYP82D171, CYP82P1v1, CYP82P2 and CYP82P3 were amplified with gene-specific primers designed from coding sequences of the draft genome, whereas CYP82P5 was amplified using primers corresponding to its putative 5'-untranslated region (UTR)/3'-UTR sequences. The PCR for the CYP82N7 isolation was performed with a forward primer designed for its 5'-UTR and a reverse primer corresponding to an anchor sequence (5'-GCGTTCACGAGGATCCACGATGACAGTG-3') added to the 3' end of the template cDNA. All PCR products were re-amplified with primers to introduce an AvrII, Nhel or XbaI site for the construction of yeast expression vectors (Supplementary Table S8). The sequences of the PCR products were confirmed after they were subcloned into pGEM T-easy or pT7 blue vectors but before their transformation.

Heterologous expression of novel P450s in S. cerevisiae or P. pastoris

The budding yeast expression vectors were constructed in pGYR-SpeI that contained the S. cerevisiae NAPD4-P450 reductase (Ikezawa et al. 2008). The full-length cDNAs of CYP82D170, CYP82D171, CYP82P1v1, CYP82P2, CYP82P5 and CYP82N7 were prepared from their subclones by the use of appropriate restriction enzymes (AvrII, Nhel or XbaI) and ligated into the SpeI site of the pGyr-Spel vectors as previously described (Ikezawa et al. 2007). The expression vectors were introduced into the S. cerevisiae strain AH122, and microsomal fractions prepared from the cultured S. cerevisiae cells were used for the enzyme assay.

A P. pastoris expression vector was constructed by the insertion of the full-length cDNA of CYP82P2 prepared by AvrII into a pPIC3.5K vector (Invitrogen) and introduced into the P. pastoris strain GS115 as described previously (Hori et al. 2016).

Enzyme assays

Six benzophenanthridine alkaloids, i.e. sanguinarine, chelerythrine, chelirubine, dihydroatkinsonine, dihydrochelerythrine and dihydrochelrubine, were used as substrates (Supplementary Fig. S4), and the product formation was analyzed by LC-MS and LC-MS/MS as described in the metabolic analysis section.
In vitro assays were performed in a reaction mixture consisting of 100 mM HEPES/NaOH (pH 7.5), 500 μM NADPH, 100 μM substrate and 10 μg of the microsomal fractions of yeast cells expressing CYP82P1v1, CYP82P2, CYP82P5 or CYP82P7. After an incubation at 30°C for 2 h, the reaction was terminated by the addition of trichloroacetic acid (final concentration of 2%) and methanol (final concentration of 40%). After the proteins precipitated, the reaction products were analyzed using LC-MS and LC-MS/MS systems.

Regarding the in vivo assays, transgenic S. cerevisiae cells that contained the EcCYP82P3 gene and that had been resuspended in 1 ml of HEPES/NaOH buffer (pH 7.5) after a pre-culture period in concentrated SD medium (8% glucose, 1.36% yeast nitrogen base without amino acids and ammonium sulfate, 4% ammonium sulfate and 160 μg ml⁻¹ l-histidine) until OD₆₀₀ = 10–15 were incubated together with the substrate (final concentration of 100 μM) at 30°C for 20 h. In vivo reactions using transgenic Pichia cells expressing CYP82P2 and co-cultures with Pichia cells expressing G11-OMT were performed as previously described, with minor modifications (Hori et al. 2016). The supernatant and methanol cell extracts (containing 0.01 N hydrochloride) of the yeast cell reactions were analyzed by LC-MS and LC-MS/MS as described above.

Database

The sequences of the draft genome (ECA_r1.0), the genes (ECA_r1.0_cds) and the proteins (ECA_r1.0_pep) are available from the Eschscholzia Genome Database (http://eschscholzia.kazusa.or.jp/). The genome sequence is also available from DDBJ/GenBank/EMBL under the accession numbers BEHA01000001–BEHA0103253 (S3253 entries). The raw data of the PE and MP reads are available from the NCBI SRA database under the BioProject accession number PRJDB6318.

Author contributions

K.H. and F.S. conceived this work and designed the experiments. K.H. performed all the experiments. Y.M. and A.T. performed the draft genome sequencing, and Y.M., A.T., K.H., S.T., H.H., Y.Y., and F.S. analyzed the genomic data. R.P. provided G11OMT. K.H. and F.S. conceived this work and designed the experiments. Y.M. and A.T. performed

Supplementary Data

Supplementary data are available online.

Funding

This study was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) [Grant-in-Aid for Scientific Research (S); 26221201 to F.S.]. The draft genome sequencing, and Y.M., A.T., K.H., S.T., H.H., Y.Y., and F.S. analyzed the genomic data. R.P. provided G11OMT. K.H. and F.S. conceived this work and designed the experiments. Y.M. and A.T. performed

Acknowledgments

Computations were partially performed on the NIG supercomputer at the ROIS National Institute of Genetics.

Disclosures

The authors have no conflicts of interests to declare.

References


