UNIVERSITÉ FRANÇOIS – RABELAIS DE TOURS

ÉCOLE DOCTORALE SSBCV

E.A. 2106 « Biomolécules et Biotechnologies Végétales »

THÈSE présentée par :

Emilien FOUREAU

Soutenue le : **13 juin 2016**

pour obtenir le grade de : **Docteur de l'Université François – Rabelais de Tours**

Discipline/ Spécialité : Sciences de la Vie et de la Santé

Elucidation de la voie de biosynthèse des alcaloïdes de *Catharanthus roseus* et ingénierie métabolique dans la levure

THÈSE dirigée par :

CLASTRE Marc	Maître de Conférences (HDR), Université de Tours

co-encadrée par :

COURDAVAULT Vincent	Maître de Conférences, Université de Tours

RAPPORTEURS :

BOURGAUD Frédéric	Professeur des Universités, Université de Lorraine
SOTTOMAYOR Mariana	Professeur des Universités, Université de Porto

---

JURY :

BOURGAUD Frédéric	Professeur des Universités, Université de Lorraine
CLASTRE Marc	Maître de Conférences (HDR), Université de Tours
COURDAVAULT Vincent	Maître de Conférences, Université de Tours
GIGLIOLI-GUIVARC'H Nathalie	Professeur des Universités, Université de Tours
SOTTOMAYOR Mariana	Professeur des Universités, Université de Porto
TRUAN Gilles	Directeur de Recherches CNRS – INSA de Toulouse

Remerciements

Je désire exprimer ici ma profonde reconnaissance à mon Directeur de thèse, Monsieur Marc Clastre, Maître de conférences à l'EA2106 « Biomolécules et Biotechnologies Végétales », pour son intérêt et son aide précieux portés à ce travail, pour ses conseils avisés et les connaissances qu'il m'a transmises. Que le chercheur en soit très vivement remercié. Merci également au philosophe que tu as su être dans nos nombreuses discussions scientifiques comme lors de ces moments passés en Crête. Tes paroles raisonneront dans ma tête pendant un petit moment encore, même après la thèse.

J'adresse de chaleureux remerciements à mes co-encadrants de thèse, M. Vincent Courdavault, Maîtres de conférences à l'EA2106, et au Professeur Nicolas Papon, qui ont contribué au bon déroulement de ce travail. Merci à toi Vincent pour ton énergie, ta confiance, tes connaissances et ton savoir-faire. J'ai pris un grand plaisir à travailler avec toi, je sais désormais que seuls l'acharnement et la ténacité peuvent palier aux nombreuses déceptions auxquelles la réalité de la science nous confronte. Merci Nicolas, pour m'avoir fait confiance à la suite du Master 2 et avoir contribué à la poursuite de mes études au sein du laboratoire dans d'excellentes conditions. Je te remercie également pour ton expertise dans le domaine des levures et aux multiples conseils que tu m'as donnés dans ce travail sans oublier bien-sûr ta bonne humeur continuelle.

Je souhaite remercier Mme Nathalie Guivarc'h, Professeur des universités et directrice du laboratoire EA2106 « Biomolécules et Biotechnologies Végétales » à Tours pour m'avoir accueilli au sein de son équipe et avoir veillé au bon déroulement de ma thèse.

Merci également au Professeur Sarah O'connor, directrice du laboratoire John Innes de Norwich, ainsi qu'à son équipe pour les nombreux travaux collaboratifs.

Je voudrais remercier les différentes personnes qui m'ont entouré à un moment donné de ma thèse. Le Dr. Arnaud Lanoue pour son expertise sur la métabolomique des AIM, ses conseils et le temps qu'il a consacré à me former sur l'UPLC-MS, tout ne se réduit pas à l'observation de pics… Le Dr. Sébastien Besseau, pour son savoir-faire sur le VIGS, le Dr.

Thomas Dugé de Bernonville pour son expertise en bioinformatique. Mais aussi le Dr. Olivier Pichon, pour la confiance qu'il a su me donner dans mes expériences d'enseignement.

Mes remerciements les plus sincères vont également au Professeur Bougaud Frédéric et au Professeur Sottomayor Mariana qui ont accepté d'être rapporteurs de ce travail, ainsi que monsieur Truan Gilles qui me fait l'honneur de participer au jury.

Parce que passer plus de temps avec nos collègues que notre famille créé des liens, je souhaite adresser mes remerciements les plus sincères à tous les membres du laboratoire « Biomolécules et Biotechnologies Végétales ». Merci à vous tous pour votre accueil, vos conseils, vos rires et sourires qui chaque fois égaient la vie du thésard. Pour tous ces bons moments passés en votre compagnie durant ces années et qui resteront dans ma mémoire une fois parti. A mes chères secrétaires, Catherine, Nathalie, Alexandra, merci pour votre réconfort dans les moments les plus difficiles, merci aussi pour nos petits moments passés lors des pauses café. Les Marie (Marie-Antoinette et Marie-Françoise), un grand remerciement pour votre aide indispensable dans l'entretien des lignées cellulaires. Merci à toi Nath « white » pour ton soutien tout au long de cette thèse, ta sympathie et pour ses moments passés en dehors du laboratoire. Parce que je ne peux pas oublier de remercier Céline et Emeline qui ont grandement contribué à ce travail de thèse. C'est avec un sourire malicieux au coin des lèvres que j'écris ces quelques lignes en repensant à toutes les fois où je suis venu vous taquiner ! Merci à toi aussi Francis, ex-collègue, pour les moments passés en dehors du labo.

Je n'oublie pas non plus tous les membres qui ont pu séjourner ou même juste transiter par ce bon vieux bureau C1180 allias le « bureau des doctorants ». Une chose est sûre c'est qu'après y avoir passé trois ans, on n'en referme pas la porte sans une certaine émotion. Un grand merci aux grands frères, Anthony, Mouadh et Benjamin, dont les souvenirs imprègnent encore les murs de ce bureau. Merci pour tous les bons moments passés ensemble ainsi que votre expérience d'ancien. Elsa, merci pour ta gaieté quotidienne et ton dynamisme. Dimitri, un grand merci pour ton soutien, tes encouragements. « Keep the smile » tu possèdes déjà beaucoup de talent pour ce métier et puis si l'envie te prend de partager à la manière d'un Robin des bois les nombreuses récompenses que tu gagnes dans les congrès, penses à nous. Je te souhaite une très bonne fin de thèse. Ma collègue portos, Ines, merci pour tes conseils, ton expérience et ton soutien dans les moments de doute. Enfin, je souhaite un bon courage aux prochains sur la liste: Tatiana, Franzyska, Florent et Kevin.

Je remercie aussi toutes les personnes que j'aurais pu oublier et l'ensemble du laboratoire EA2106 « Biomolécules et Biotechnologies végétales », l'équipe de chercheurs, les Biatss, les doctorants et les stagiaires pour la bonne ambiance générale du laboratoire qui m'a permis de réaliser la thèse dans de bonnes conditions.

Je souhaiterais enfin exprimer toute ma gratitude et ma reconnaissance à mes proches (ils se reconnaitront). Merci de votre soutien sans faille. A mes parents sans qui rien n'aurait été possible, merci d'avoir fait naitre en moi le goût de l'effort et de la persévérance. Merci de votre amour inconditionnel, ce travail est le fruit de vos nombreux sacrifices…

Résumé

Catharanthus roseus est une plante médicinale produisant divers types d'alcaloïdes indoliques monoterpéniques (AIM) d'intérêt en santé humaine. Ainsi, les AIM dimères comme la vinblastine et la vincristine sont utilisés en chimiothérapie anticancéreuse et les alcaloïdes monomères de type hétéroyohimbine présentent diverses activités pharmacologiques. La fabrication de ces molécules dans la plante est fort complexe. Elle requiert un haut niveau de compartimentation tissulaire et subcellulaire et met en jeu plus d'une trentaine d'étapes enzymatiques, dont certaines sont encore très mal connues. Dans ce contexte, l'objectif de la thèse a consisté à élucider plusieurs étapes enzymatiques de la voie de biosynthèse des AIM. Nos travaux ont permis de caractériser de nouvelles isoformes enzymatiques de la famille des cytochromes P450 ainsi que les réductases qui leur sont associées. Ils ont abouti à l'identification de nouvelles déshydrogénases et mis en évidence, *in planta*, leurs interactions avec la strictosidine synthase suggérant une biosynthèse orientée vers les divers alcaloïdes de type hétéroyohimbine. Enfin, en ayant recours à l'ingénierie métabolique, un segment de la voie de biosynthèse a été transféré dans la levure *Saccharomyces cerevisiae*, lui conférant la capacité de bio-transformer la tabersonine en vindoline, l'un des deux précurseurs finaux des alcaloïdes dimères.

Mots-clefs: *Catharanthus roseus*, alcaloïdes indoliques monoterpéniques, biosynthèse, compartimentation subcellulaire, ingénierie métabolique.

Abstract

Catharanthus roseus is a medicinal plant producing various types of monoterpene indole alkaloids (MIA) with a great interest in human health. MIA dimers such as vinblastine and vincristine are used in cancer chemotherapy and heteroyohimbine monomers alkaloids exhibit various pharmacological activities. The production of these molecules in the plant is very complex. It requires a high level of tissular and subcellular compartmentalization and involves more than thirty enzymatic steps, some of which are largely unknown. In this context, the aim of this thesis was to elucidate several enzymatic steps of the MIA biosynthetic pathway. Our work allowed us to characterize new enzyme isoforms that belong to the family of cytochrome P450 reductase and their associated reductases. They also resulted in the identification of new dehydrogenases highlighted *in planta*, and their interactions with the strictosidine synthase suggest a directed biosynthesis towards various heteroyohimbine alkaloids. Finally, by using metabolic engineering, a segment of the biosynthetic pathway was transferred into the yeast *Saccharomyces cerevisiae*, giving it the ability to bio-convert tabersonine in vindoline, one of the two final precursors of the dimeric alkaloids.

Keywords: *Catharanthus roseus*, monoterpene indole alkaloid, biosynthesis, subcellular compartimentalizaton, metabolic engineering.

Abbréviations

AA : acide aminé

AACT : acétoacétyl-CoA

ADH : alcool déshydrogénase

ADN : acide désoxyribonucléique

ADNc : acide désoxyribonucléique complémentaire

AIM : alcaloïde indolique monoterpénique

ARN : acide ribonucléique

ARNm : acide ribonucléique messager

AS : anthranilate synthase

BiFC : bimolecular fluorescence complementation

CDP-ME : 4-diphospho-cytidyl-2C-méthyl-D-érythritol

CDP-MEP : 4-diphosphocytidyl-2C-méthyl-D-érythritol 2 phosphate

C4H : cinnamate 4-hydroxylase

CMK : 4-diphosphocytidyl-2C-méthyl-D-érythritol kinase

CMS : 4-diphospho-cytidyl-2C-méthyl-D-érythritol synthase

CO_2 : dioxyde de carbone

CPR : NADPH cytochrome P450 réductase

CPR1 : NADPH cytochrome P450 réductase 1

CPR2 : NADPH cytochrome P450 réductase 2

CYP, **P450** ou **CYP450**: cytochrome P450

CYP450-réductase : cytochrome P450 réductase

CS : chorismate synthase

DAT : déacétylvindoline 4-O-acétyl transférase

DFR : diflavine réductase

D4H : désacétoxyvindoline 4-hydroxylase

DHAP : 3-désoxy-D-arabinoheptulosonate-7-phosphate

DMAPP : diméthylallyl diphosphate

7DLH : 7-déoxyloganique acide 7-hydroxylase

7DLGT : 7-déoxyloganétique-acide-glycosyltransférase

DXP : 1-désoxy-D-xylulose-5-phosphate

DXR : 1-désoxy-D-xylulose 5-phosphate réductoisomérase

DXS : 1-désoxy-D-xylulose 5-phosphate synthase

DXS1 : 1-désoxy-D-xylulose 5-phosphate synthase isomère 1

DXS2 : 1-désoxy-D-xylulose 5-phosphate synthase isomère 2

FAD : flavine adénine dinucléotide

FMN : flavine mononucléotide

GES : géraniol synthase

GFP : green fluorescent protein (protéine de fluorescence verte)

G10H : géraniol 10-hydroxylase

G3P : glycéraldéhyde-3-phosphate

GPP : géranyl diphosphate

GPPS : géranyl diphosphate synthase

HAIRY ROOTS : chevelus racinaires

HDR : 4-hydroxy-3-méthylbut-2-ényl diphosphate réductase

HDS : 4-hydroxy-3-méthylbut-2-ényl diphosphate synthase

10HGO : 10-hydroxygéraniol-oxydoréductase

HMBPP : 4-hydroxy-3-méthylbut-2-ényl diphosphate

HMGS : hydroxy-3-méthyl-glutaryl-CoA synthase

HMGR : hydroxy-3-méthl-glutaryl -CoA réductase

H_2O : eau

HYS : hétéroyohimbine synthase

IO : iridoïde oxidase

IS : iridoïde synthase

IDI1 : isopentényl diphosphate isomérase 1

IPP : isopentényl diphosphate

LAMT: acide loganique méthyltransférase

MDR : déshydrogenase/réductase contenant un ion zinc à moyenne chaine

MECS : 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate synthase

MECPP : 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate

MEP : 2-C-méthyl-D-érythritol-4-phosphate

MTSI : monoterpènes séco-iridoïdes

MVA : acide mévalonique

MVD : mévalonate 5-diphosphate décarboxylase

MVK : mévalonate kinase

NAD : nicotinamide adénine dinucléotide

NADP : nicotinamide adénine dinucléotide phosphate

NADPH : nicotinamide adénine dinucléotide phosphate

NAD(P)H:nicotinamide adénine dinucléotide phosphate et nicotinamide adénine dinucléotide

NMT : 2,3-dihydro-3-hydroxytabersonine-N-méthyl transférase

16 OMT : 16 hydroxytabersonine 16-O-méthyltransférase

PAPI : parenchyme associé au phloème interne

PMK : mévalonate phosphate kinase

POD : péroxydase

PPM : parti par million

PRX1 : péroxydase 1

PTS1 : signal d'adressage aux peroxysomes 1 situé en C-terminal

5'RACE : amplification rapide de la partie 5' d'un ADNc

RE : réticulum endoplasmique

RNA-seq : données issues du séquençage des ARN

SDR : déshydrogenase/réductase à courte chaine

SGD : strictosidine β-glucosidase

SLS : sécologanine synthase

SLS1 : sécologanine synthase 1

SLS2 : sécologanine synthase 2

SPIL : séquence d'adressage à la vacuole

STR : strictosidine synthase

TDC : tryptophane décarboxylase

T16H : tabersonine 16-hydroxylase

T16H1 : tabersonine 16-hydroxylase 1

T16H2 : tabersonine 16-hydroxylase 2

THAS : tétrahydroalstonine synthase

TMV : virus de la mosaïque du tabac

T3O : tabersonine-3-oxydase

TPT2 : transporteur de la catharanthine dépendant de l'ATP

T3R : tabersonine-3-réductase

TRV : virus du tabac hochet

TS : tryptophane synthase

UV : ultra-violet

VIGS : extinction de l’expression de gènes induite par un virus

Table des matières

Liste des figures

Introduction

Chapitre I : Introduction Générale

Tous les organismes vivants interagissent avec leurs milieux en prélevant des éléments comme l'eau, les sels minéraux, l'oxygène et/ou le dioxyde de carbone pour produire des métabolites participant aux fonctions cellulaires. Parmi ces organismes, les végétaux interagissent spécifiquement avec leur environnement en synthétisant des métabolites primaires issus directement de la photosynthèse (figure 1). Ces métabolites, dits primaires, sont communs à tous les organismes et leur rôle est d'assurer les fonctions cellulaires de base pour la plante.

Contraintes par leur immobilité, les plantes ont su s'adapter aux caractéristiques du milieu dans lequel elles se développent en synthétisant à partir de leur métabolisme primaire une multitude de molécules spécialisées de structures très diverses des plus simples aux plus complexes. A leur découverte, au milieu du XX^{ème} siècle, elles ont été regroupées sous le terme de « métabolites secondaires » par opposition aux métabolites primaires commun à l'ensemble des organismes vivants (figure 1). Cette dénomination reflète le fait que la plupart de ces composés ne semble pas essentiels à la croissance et au développement des plantes et que pour une certaine partie d'entre eux leur synthèse ne s'effectue qu'en de faibles quantités et sous certaines conditions. Il apparaît aujourd'hui que ces molécules présentent une fonction centrale dans les interactions entre la plante et son environnement (attraction d'insectes pollinisateurs, défense contre des organismes pathogènes, interaction plante-plante, protection contre les rayonnements UV) et que de plus, bien que majoritairement isolées à partir des végétaux, elles ne sont en aucune façon restreintes au règne végétale. Malgré l'énorme variété de métabolites secondaires, le nombre de voies de biosynthèse correspondantes reste limité. Les précurseurs sont essentiellement issus de voies métaboliques de base, comme la glycolyse, le cycle de Krebs ou la voie des pentoses phosphate (figure 1). Il est à noter que ces dernières années, le terme de métabolisme secondaire a été peu à peu supplanté dans la littérature scientifique par le terme de métabolisme spécialisé qui signifie la même chose.

A ce jour, les métabolites secondaires peuvent être regroupés en trois grandes familles en fonction de leur structure chimique (figure 1) : **les composés phénoliques**, qui se caractérisent par la présence d'un noyau benzénique portant au moins un groupement hydroxyle. Cette classe, se compose notamment des phénylpropènes, des coumarines, des

naphtoquinones, des flavonoïdes, des lignanes, des lignines ou encore des tanins qui sont synthétisés à partir de la voie du shikimate. **Les isoprènoïdes ou terpènes**, issus de l'assemblage d'unités isopréniques à 5 atomes de carbone (C_5) constituent probablement la plus large classe de métabolites secondaires incluant les mono- (C_{10}), sesqui-(C_{15}), di-(C_{20}), tri-(C_{30}), tétra-(C_{40}), polyterpènes (> C_{40}),et tous leurs dérivés. Enfin il y a **les composés azotés**, parmi lesquels figurent les glucosinolates et les alcaloïdes.

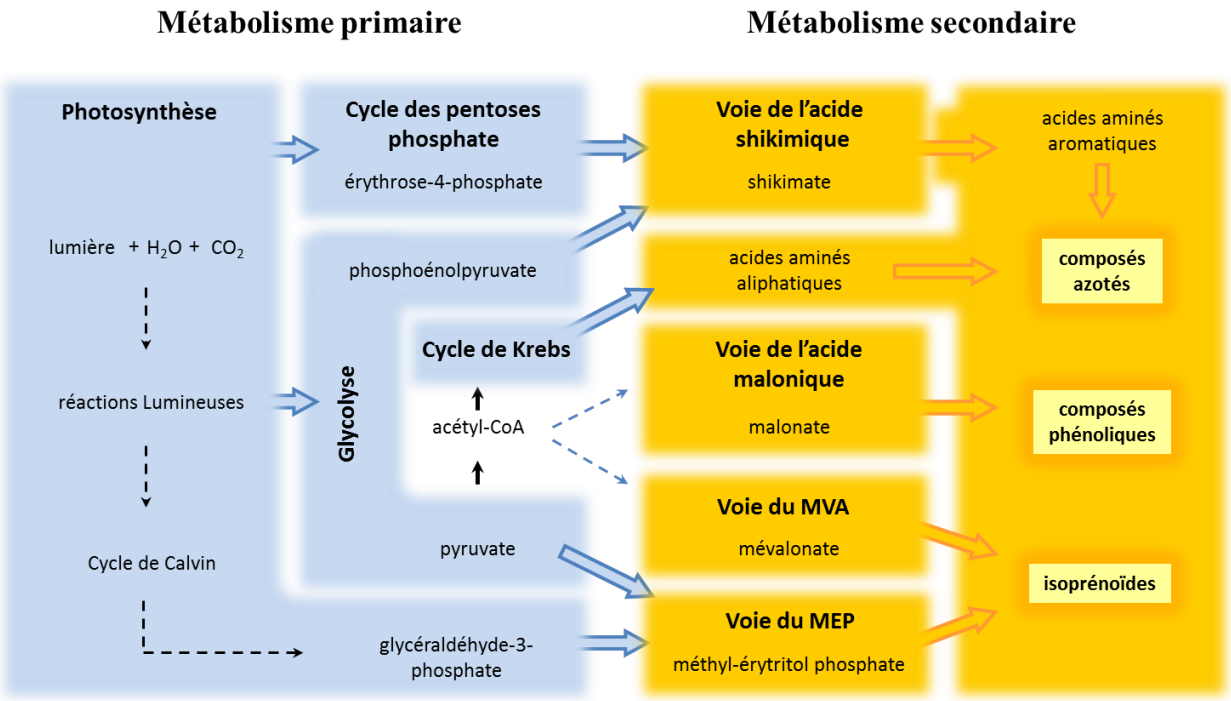


Figure 1 : Représentation simplifiée des voies de biosynthèse du métabolisme primaire et du métabolisme secondaire chez les végétaux et leurs interactions. CO_2 : dioxyde de carbone, H_2O : eau, MVA : mévalonate ou acide mévalonique, MEP : méthyl-érythritol-phosphate.

A la plupart de ces composés, a pu être associée une fonction biologique. Les flavonoïdes et les anthocyanines jouent un rôle dans la pigmentation des fleurs attirant ainsi les pollinisateurs (Conn., 1981) et sont donc directement impliqués dans le processus de reproduction. Les caroténoïdes protègent la chlorophylle des réactions de dégradation photochimique induites par des irradiations trop intenses. Les alcaloïdes quant à eux, du fait de leur toxicité directe ou par une action répulsive, assurent très souvent des fonctions de défenses chimiques contre les pathogènes, les insectes ravageurs ou encore les herbivores.

La nature est la source d'une grande diversité de molécules possédant parfois des propriétés thérapeutiques. L'Homme trouva chez les végétaux des remèdes à ses maux, bénéficiant sans le savoir de la très grande variété de métabolites secondaires contenue dans des extraits de plantes. Des passages d'auteurs classiques attestent que l'opium extrait du pavot (*Papaver somniferum*, Papaveracée) était déjà connu dans les pays de la Méditerranée orientale (Grèce, Crète, Chypre, Egypte) pendant l'antiquité. Au IV^{e} siècle av. J.-C. les différentes parties du pavot entraient dans la confection de divers mélanges, collyres, cataplasmes et pastilles à effet antalgique et narcotique. De nombreux autres exemples jalonnent encore l'histoire, comme l'exécution du philosophe Socrate en 399 avant J.C., par absorption de ciguë (Grande ciguë, *Conium maculatum*), l'utilisation par la reine Cléopâtre d'extrait de Jusquiame noire (*Hyoscyanum niger*), dans un but cosmétique, pour son action dilatatrice de la pupille sous l'effet de l'atropine. Aujourd'hui, près de 80% des habitants de la planète ont principalement recours aux 13000 plantes médicinales répertoriées dans le monde (Farnsworth et *al*., 1985; Tyler, 1994). Même si les avancées scientifiques ont permis d'isoler puis de synthétiser bon nombre de métabolites secondaires, les composés végétaux restent à l'origine de plus de 25% des prescriptions pharmaceutiques. Par ailleurs on assiste aujourd'hui à une recrudescence de l'emploi traditionnel des produits végétaux naturels avec le développement des phytomédicaments (phytothérapies, naturothérapies).

Parmi les métabolites secondaires, une famille de molécules est particulièrement riche en composés actifs. Il s'agit de la famille des alcaloïdes avec près de 12000 molécules identifiées à ce jour. On retrouve les alcaloïdes dans approximativement 20% des espèces végétales (Ziegler et Facchini, 2008) et chez quelques animaux (Harborne et *al*., 1999). Chez les plantes, quatre classes d'alcaloïdes ont été décrites. Les **alcaloïdes indoliques monoterpéniques** (3000 molécules) (Almagro et *al*., 2015) sont issus de la condensation d'un acide aminé, le tryptophane avec un dérivé monoterpénique la sécologanine. Les **alcaloïdes benzyliquinoline** (2500 molécules) dérivent de la tyrosine (Hagel et Facchini, 2013), les **alcaloïdes tropaniques** sont issus de l'arginine et de la phénylalanine tandis que les **alcaloïdes puriniques** sont fabriqués à partir de la purine (Ashihara et *al.*, 2008). L'un des alcaloïdes les plus controversé de l'histoire est sans doute la cocaïne extrait de la feuille de coca (*Erythroxylum coca*). Utilisé tout d'abord comme énergisant et stimulant dans des boissons comme le Coca-cola, il fut rapidement retiré de la consommation. En effet, suite à ses effets indésirables comme l'accoutumance, les comportements psychotiques et les hallucinations, la cocaïne a été classifiée comme stupéfiant. Par ailleurs, ce sont les effets

nuisibles de la consommation courante de cocaïne qui sont à l'origine du terme « toxicomanie ». Malgré le caractère toxique ou hallucinogène de certains alcaloïdes, beaucoup d'entre eux entrent dans la composition de nombreux médicaments en tant que principe actif. Parmi les plus connus, on peut citer la morphine, premier alcaloïde identifié et isolé du latex du pavot (*Papaver somniferum*) et utilisée comme analgésique, la quinine issue de *Cinchona succirubra* administrée en tant qu'antipaludéen ou encore les anticancéreux vincristine et vinblastine issus de la Pervenche de Madagascar (*Catharanthus roseus*) (figure 2).

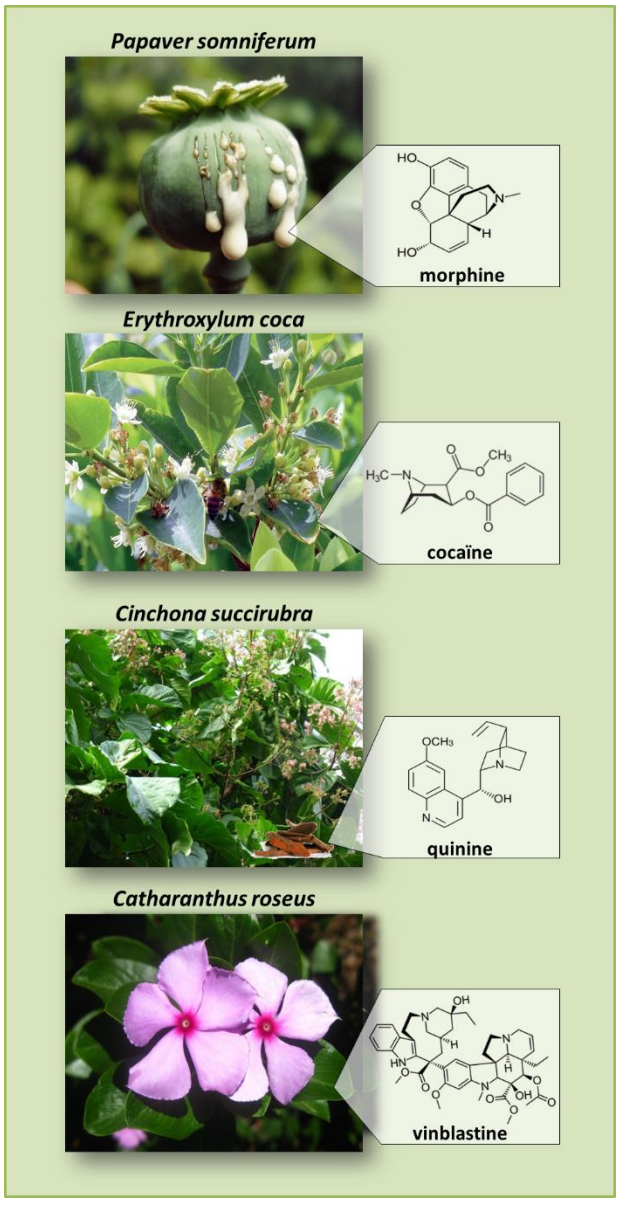


Figure 2 : Structure d'alcaloïdes extraits des végétaux et possédant une activité thérapeutique. Parmi les centaines de milliers de produits naturels existants, les alcaloïdes

contiennent le plus grand nombre de composés actifs. Les alcaloïdes sont des molécules azotées, composées d'un cycle aromatique contenant un atome d'azote.

La culture de plantes représente aujourd'hui encore l'une des étapes du processus d'obtention de molécules d'intérêt pharmacologique. De nombreuses plantes ont ainsi été identifiées et exploitées pour leur capacité à fournir ces composés utilisés par l'homme sans que leurs mécanismes de biosynthèse ainsi que leur rôle pour la plante soient connus. Cette ignorance a longtemps été considérée comme sans conséquence. Cependant, le fait que les molécules les plus intéressantes soient le plus souvent celles qui sont en plus faibles concentrations dans la plante se répercute directement sur les coûts de production. D'autre part, dans de nombreux cas, la synthèse chimique de ces molécules s'avère difficile et donc peu rentable économiquement.

Aussi, s'est dessinée peu à peu l'idée qu'une production optimisée de ces molécules devait s'appuyer sur la compréhension des mécanismes biochimiques mis en œuvre pour leur biosynthèse. Ces recherches ont notamment pris leur essor dans les années 1980-90 et ont conduit à l'élaboration de cultures *in vitro* de cellules, tissus et organes végétaux en vue de produire des métabolites secondaires d'intérêt. Bien que les résultats n'aient pas été à la hauteur des espérances car les taux de production furent insuffisant, l'ensemble de ces travaux a jeté les bases de l'élucidation des voie métaboliques et de leur régulation (Zenk, 1991). De nos jours, ces recherches se poursuivent en s'appuyant sur les données « omiques » et s'incarnent dans le champ de la génomique phytochimique qui vise à élucider les bases génomiques de la synthèse des composés en intégrant les approches métabolomiques (Saito, 2013). C'est ainsi qu'ont pu être réalisées des avancées majeures ces dernières années permettant aujourd'hui d'envisager d'implémenter des segments de voies métaboliques végétales dans des microorganismes hôtes en vue de produire les métabolites d'intérêt. Cette ingénierie métabolique, qui s'inscrit dans le domaine de la biologie synthétique (Keasling, 2012) est en plein essor dans de nombreux domaines comme ceux des biocarburants, des nouveaux matériaux et des métabolites d'intérêt pharmaceutique et/ou cosmétique.

C'est dans le contexte de la génomique phytochimique et de l'ingénierie métabolique que s'inscrit mon travail de thèse. Ainsi, il a eu pour objet d'élucider plusieurs étapes **de la voie de biosynthèse des alcaloïdes indoliques monoterpéniques de la plante médicinale**

Catharanthus roseus **et de reconstituer un segment de cette voie dans la levure** ***Saccharomyces cerevisiae*****, en vue de produire certains alcaloïdes d'intérêts.**

Chapitre II : *Catharanthus roseus* et les alcaloïdes indoliques monoterpéniques

II.1 Taxonomie et origine de *Catharanthus roseus*

Catharanthus roseus, plus communément appelée la pervenche de Madagascar, est une plante herbacée pérenne endémique de Madagascar appartenant à la famille des Apocynacées. Elle fut découverte par le naturaliste suédois Carl von Linné en 1759, sous le nom initial de *Vinca rosea* et révisée par le botaniste britannique George Don en 1837 qui l'a classée dans le genre *Catharanthus*. Actuellement huit espèces composent le genre *Catharanthus* dont sept originaires de l'île de Madagascar avec *Catharanthus coriaceus* Markgr, *C. lanceus* (Bojer ex A.DC.) Pichon, *C. longifolius* (Pichon) Pichon, *C. ovalis* Markgr, *Catharanthus roseus* (L.) G.Don. Madagascar (Pervenche de Madagascar), *C. scitulus* (Pichon) Pichon, *C. trichophyllus* (Baker) Pichon et une originaire du subcontinent indien *C. pusillus* (Murray) G.Don (Van der heijden et *al*., 2004).

De nos jours, le genre *Catharanthus* s'est bien exporté à l'ensemble des continents puisqu'on le retrouve maintenant en Afrique, Amérique, Asie, Australie et Europe avec une colonisation des îles du Pacifique et des tropiques. Cette vaste répartition est liée à d'importants transports de cette plante utilisée comme coupe faim et tonifiant pour éliminer la fatigue des marins lors de longues traversées. Bien que la dispersion de *C. roseus* ait eu lieu il y a plusieurs siècles, les plantes sauvages sont toujours récoltées et utilisées aujourd'hui en pharmacopées traditionnelles.

II.2 Une plante aux multiples vertus

Cette plante fait partie intégrante d'un bon nombre de pharmacopées traditionnelles et les vertus médicinales qu'on lui prête divergent selon les régions du globe. De façon générale, deux types d'usages sont encore utilisés aujourd'hui: en **usage externe**, comme en Amérique du Sud, où les extraits de feuilles sont utilisés comme antiseptique pour guérir les blessures, prévenir les hémorragies. La décoction est employée pour le soin de dermatoses ou de candidoses buccales, en bains de bouche. A titre d'exemple, la décoction des fleurs sert aussi

à soigner les conjonctivites et les yeux infectés. Les feuilles sont utilisées broyées ou en décoction pour soulager des piqûres de guêpe en Inde, pour traiter les dartres et les ulcères à Madagascar ou pour désinfecter les plaies au Brésil. En **usage interne**, la décoction de l'ensemble de la plante est considérée en Afrique comme agent hypoglycémique par voie orale. La décoction des feuilles et des fleurs est utilisée pour soigner les problèmes de tension, le diabète et les affections du foie, les indigestions et les ulcères de l'estomac. Dans l'Océan Indien, elle est utilisée traditionnellement en tant que purgatif, vermifuge ou encore dans le traitement de ménorragies. En infusion, elle est réputée au nord du Vietnam pour jouer un rôle antipaludéen et diurétique. Enfin, dans toute la région Indopacifique, les feuilles sont utilisées comme coupe-faim (Nammi et *al*., 2003 ; Mostofa et *al*., 2007 ; Aslam et *al*., 2010).

Malgré ces effets bienfaisants, beaucoup de cas d'intoxications ont lieu chaque année suite à l'emploi de la plante sous forme de tisane. En dépit de son potentiel thérapeutique, la plante est largement connue pour sa toxicité. En effet, celle-ci renferme plus d'une centaine d'alcaloïdes tous plus ou moins toxiques. Traditionnellement utilisée pour ses propriétés hypoglycémiantes, la pervenche de Madagascar a tout d'abord fait l'objet de recherches approfondies pour la mise au point de nouveaux médicaments antidiabétiques. De récents travaux ont montré qu'une administration d'extraits bruts de *C. roseus* entraîne une augmentation de la sécrétion d'insuline dans le sang de lapin et de rats diabétiques (Nammi et *al*, 2003 ; Mostofa et *al*, 2007). Ces propriétés pseudoantidiabétiques motivèrent, dès la fin des années 1950, les premiers travaux de recherche sur de nombreux alcaloïdes produits par la plante. L'intérêt thérapeutique de certains AIM produits par cette plante accéléra les recherches pour sélectionner et isoler des cultivars présentant une plus forte concentration de ces composés (Dutta et *al.*, 2005), allant même aujourd'hui jusqu'à développer des outils de biotechnologies (culture de calles et de cellules) pour améliorer les rendements de production (Mujib et *al.*, 2012).

II.3 Nature et distribution des alcaloïdes chez *Catharanthus roseus*

A ce jour, 130 alcaloïdes ont pu être identifiés ou isolés à partir des différents organes de la plante (Duffin, 2000). La plupart d'entre eux appartiennent à la classe des alcaloïdes indoliques monoterpéniques (AIM) (Duffin, 2000 ; Van der Heijden *et al*., 2004). La classe des AIM peut se subdiviser en deux sous-catégories. D'un côté les AIM de type monomère

comme l'ajmalicine, la serpentine, la catharanthine, ou encore la vindoline. De l'autre, les AIM de type dimère telle que la vinblastine et la vincristine premièrement identifiées par Ralph Noble et Charles Beer en 1958 (Van der Heijden et *al.*, 1989). Les AIM monomères sont regroupés en trois sous-catégories distinctes : les monomères de type corynanthe comme l'ajmalicine, les monomères de types iboga dérivant de la stemmadénine dont fait partie la catharanthine (El Sayed et Verpoorte, 2007) et celle des monomères de type aspidosperma dont le représentant principal est la vindoline (figure 3 et figure 4).

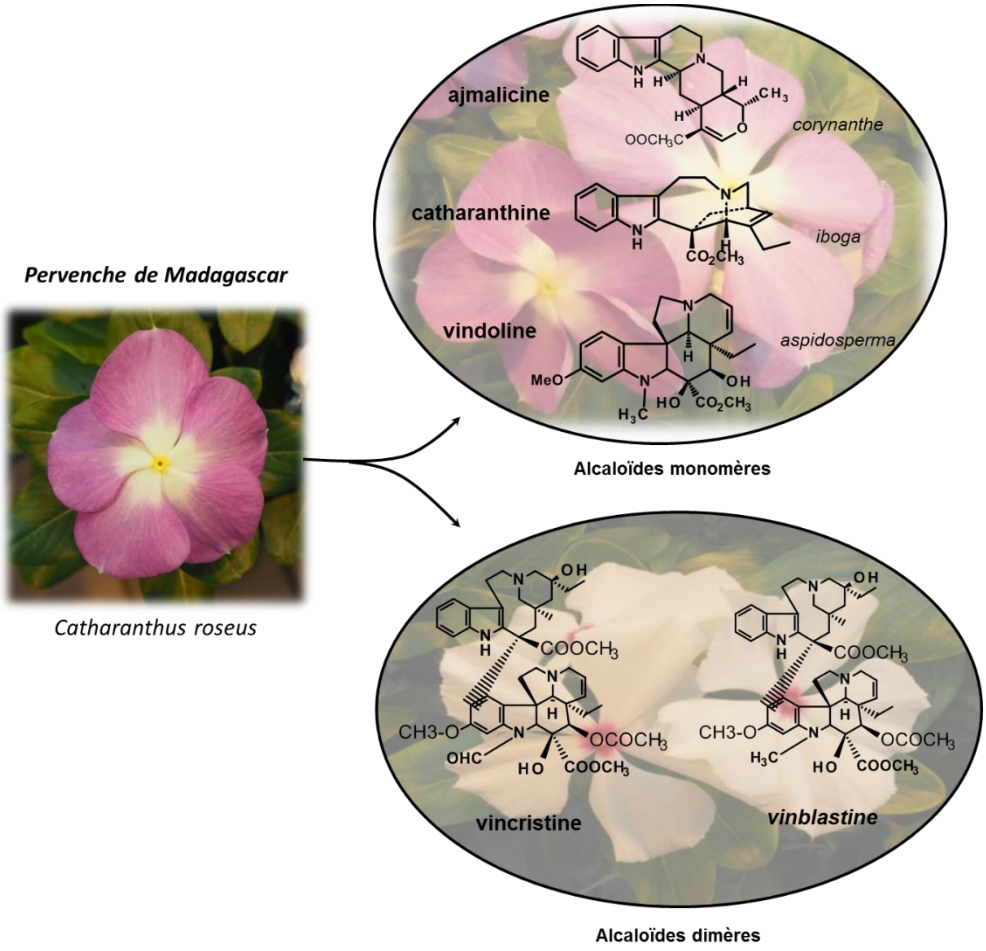


Figure 3 : Structure chimique des AIM de la pervenche de Madagascar. *C. roseus* présente deux types d'AIM : les AIM monomères et les AIM dimères.

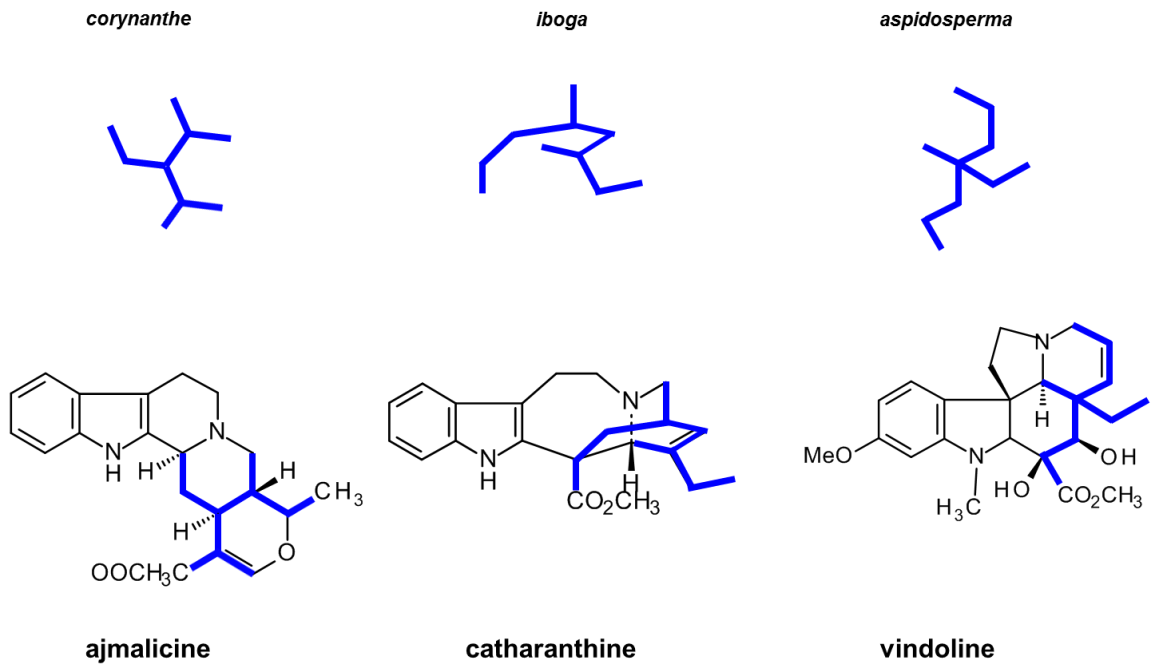


Figure 4 : Représentation des trois classes majeures d'AIM et de leurs représentants chez *C. roseus*. Les AIM dérivent tous de la structure carbonnée de la sécologanine et de la tryptamine. Ils appartiennent à trois classes distinctes, *corynanthe*, *iboga* et *aspidosperma*. La partie dérivée de la sécologanine de leurs squelettes carbonnés est dessinée en bleue (figure 4).

Les AIM ne sont pas synthétisés dans les graines mais deux à trois semaines après germination. Leur concentration fluctue selon les cycles de vie de la plante et leur distribution n'est pas homogène dans la plante. On a pu remarquer des concentrations élevées dans les tissus en développement de jeunes feuilles, qui décroissent en fonction de l'état de maturité des feuilles plus âgées (Balsevich et Bishop, 1989 ; Naaranlahti *et al.*, 1991). Chez *C. roseus*, la majorité des AIM sont présents dans tous les tissus de la plante mais de façon inégale. Seuls les alcaloïdes dimères apparaissent exclusivement dans la partie aérienne de la plante ainsi que la vindoline considérée comme l'alcaloïde majoritaire des feuilles de *C. roseus* (Westekemper *et al.*, 1980). Certains alcaloïdes de types *corynanthe*, comme la serpentine et l'ajmalicine sont principalement présents dans les racines (figure 5).

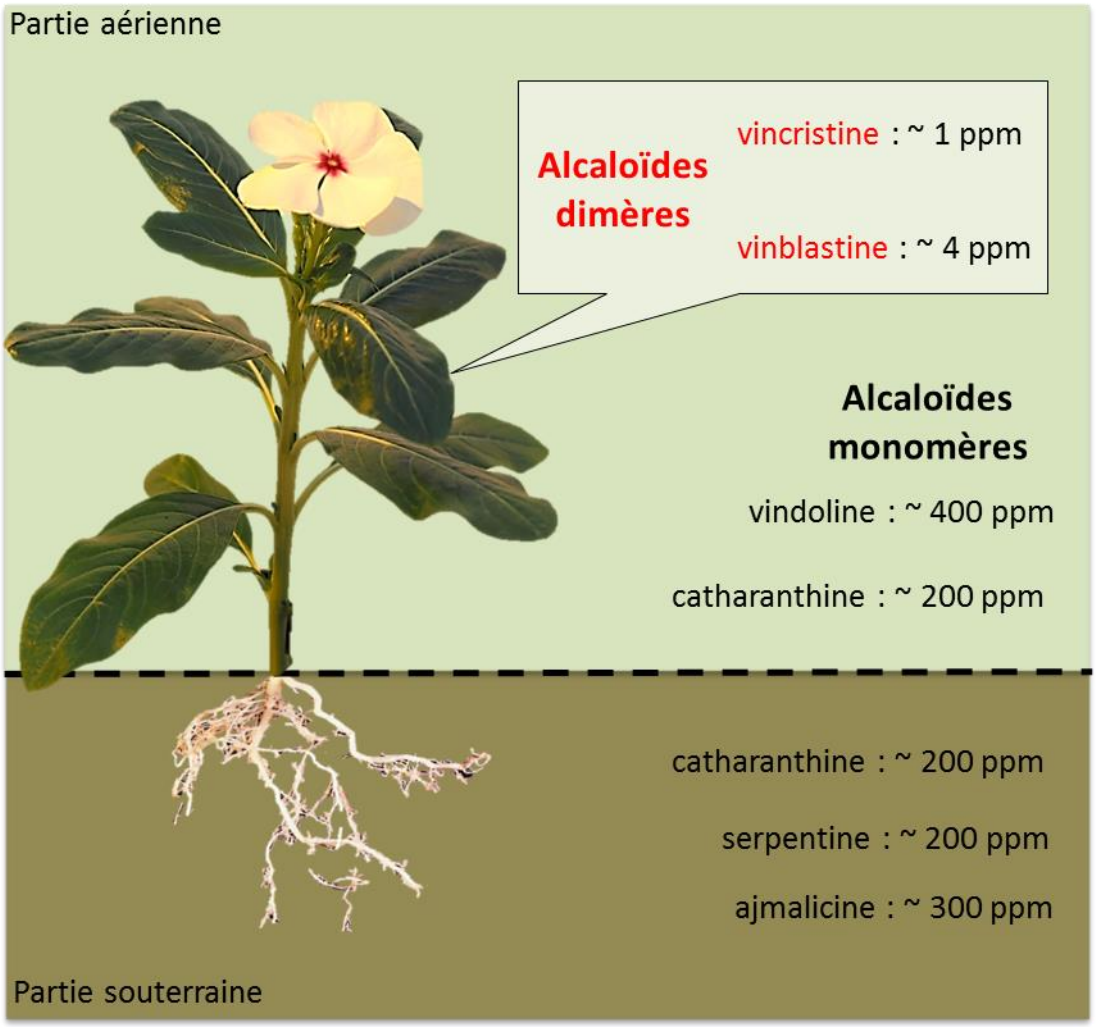


Figure 5 : Distribution et concentration des AIM de *C. roseus.* ppm : partie par million.

II.4 Rôle des alcaloïdes chez *Catharanthus roseus*

Le rôle physiologique de nombreux métabolites secondaires reste encore très énigmatique aujourd'hui. Leur fonction resta longtemps une interrogation jusqu'à la découverte des propriétés insecticides de la Nicotine dans les années 1940. Puis d'autres travaux confirmèrent le caractère toxique de certains alcaloïdes, ceux-ci pouvant constituer un véritable arsenal chimique de défense pour la plante contre l'attaque d'herbivores et de pathogènes (Wink, 1999). Chez C. *roseus*, le peu d'informations disponibles sur les fonctions attribuées aux AIM suggèrent toutefois leur implication dans des mécanismes de défense. Cette plante serait capable de modifier son profil alcaloïdique en réponse à une blessure (Van dam et *al.*, 1993), en accumulant préférentiellement des alcaloïdes au niveau des zones lésées modifiant ainsi les flux de synthèse. Ainsi, en cas de blessure, la vindoline s'accumule

beaucoup plus rapidement grâce à une accélération de la transformation de la tabersonine en vindoline (Vazquez-Flota *et al.*, 2004). On note aussi une accumulation d'alcaloïdes finaux (synthétisés dans des étapes finales de voies de biosynthèse) comme la vindoline, la vincristine et la vinblastine dans les racines infectées par des agents pathogènes comme les phytoplasmes (Favali et *al*., 2004). Plus récemment, deux modèles complémentaires, fondés sur la compartimentation de la voie de biosynthèse des AIM au niveau cellulaire et subcellulaire, ont attribué aux AIM un rôle physiologique dans la mise en place d'un système de défense contre les insectes herbivores et/ou les pathogènes nécrophages (Roepke *et al*., 2010 ; Guirimand *et al*., 2010). Des travaux précurseurs, (Luijendijk *et al.*, 1996) avaient suggéré que la strictosidine β-glucosidase (SGD), enzyme qui catalyse la déglucosylation de la strictosidine (premier AIM et précurseur de tous les autres chez *C. roseus*) en strictosidine aglycone, jouerait un rôle direct dans la défense contre plusieurs microorganismes au travers des effets toxiques de la forme aglycone. Des travaux plus récents ont proposé un modèle basé sur le concept d'une « bombe à retardement (figure 6). Ce modèle est basé sur une compartimentation distincte de la SGD, présente dans le noyau et de son substrat, la strictosidine qui s'accumule préférentiellement dans la vacuole. En condition normale, la synthèse de strictosidine puis son apport dans le noyau et sa déglucosylation par la SGD serait régulée, assurant ainsi un niveau basal de production d'AIM. Lors d'une attaque par un herbivore, la synthèse de strictosidine augmenterait et la rupture des membranes tonoplastiques et nucléaires provoquerait une mise en contact massive de strictosidine et de SGD provoquant la formation d'importantes quantités de strictosidine aglycone, produit hautement toxique pour l'agresseur car il possède un pouvoir important de réticulation des protéines (Barleben *et al.*, 2007 ; Guirimand *et al.*, 2010).

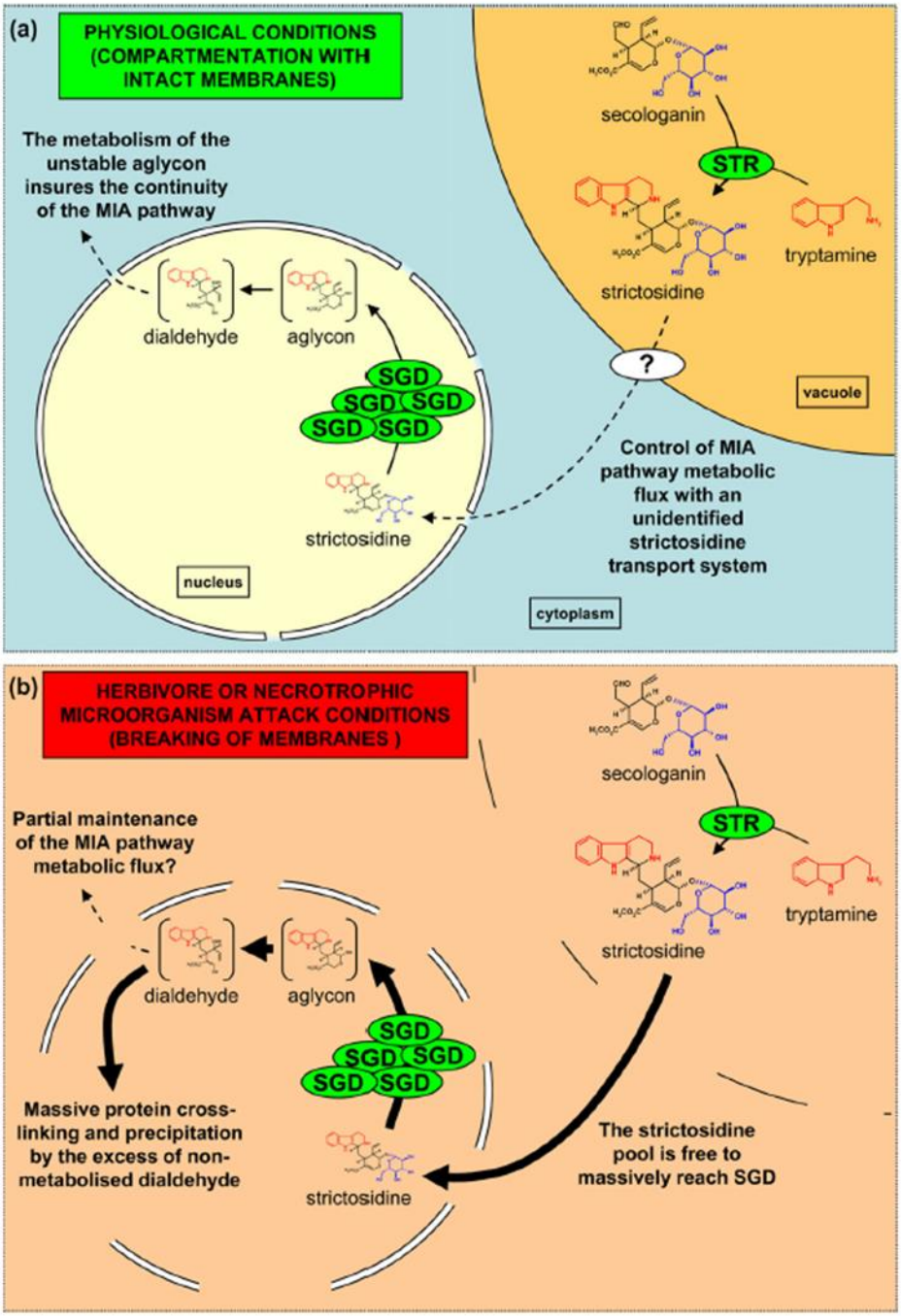


Figure 6 : Illustration de la séquestration de deux enzymes de la voie de biosynthèse des AIM impliquées dans la défense de la plante (adaptée d'après Guirimand et al., 2010). En condition normale (de non attaque) a) la SGD est séquestrée dans le noyau de la cellule et la STR dans la vacuole. Le flux métabolique de la strictosidine au noyau est alors contrôlé. Lors d'une attaque b) la rupture du tonoplaste et des membranes nucléaires, entraîne une mise en contact de la strictosidine et de la SGD pour produire l'aglycone hautement réactif. strictosidine synthase (STR); strictosidine β-glucosidase (SGD).

II.5 Utilisation des alcaloïdes de *Catharanthus roseus* en santé humaine

L'activité biologique la plus recherchée depuis la fin des années 1950 est l'effet anti-tumeur de certains alcaloïdes dimères comme la vincristine et la vinblastine. La propriété antinéoplasique de ces AIM réside principalement dans la capacité de pouvoir perturber la polymérisation des microtubules entraînant ainsi l'arrêt de la division cellulaire en métaphase et induisant l'apoptose (Jordan et *al*., 1998). Ces métabolites interfèrent avec les contacts entre tubulines et empêchent ainsi leur assemblage en microtubule (Gigant et *al*., 2005).

Les AIM dimères sont encore aujourd'hui utilisés dans des traitements chimio-thérapeutiques : la vinblastine est utilisée préférentiellement dans le traitement de la maladie de Hodgkin's tandis que la vincristine (la forme oxydée de la vinblastine) est très active pour traiter les leucémies lymphoblastiques. Toutefois un traitement prolongé avec ces AIM entraîne de multiples effets secondaires indésirables comme des vomissements et une perte des réflexes (Avila et *al*., 1994).

La recherche de médicaments ayant moins d'effets secondaires a conduit à l'élaboration d'AIM semi-synthétiques de plus faible toxicité. La vindesine et la vinorelbine par exemple, sont obtenus par semi-synthèse chimique à partir de la vinblastine, et la vinflunine est un dérivé fluoré de la vinorelbine. Ces alcaloïdes semi-synthétiques possèdent des profils pharmacologiques différents. La vindésine est utilisée dans le traitement de certains lymphomes, la vinorelbine dans le traitement du cancer du poumon (Gregory et *al*., 2000) et la vinflunine dans le traitement de carcinomes mammaires (Bellmunt *al*., 2013). Par ailleurs, outre leurs activités anticancéreuses, la vinblastine et la vincristine possèdent une activité antimicrobienne et sont aussi utilisées comme traitements antiparasitaires sur *Trypanosoma cruzi* responsable de la trypanosomiase humaine (Grellier et *al*., 1999).

Chapitre III : Biosynthèse des AIM chez *Catharanthus roseus* et architecture des voies métaboliques

III.1 Biosynthèse des AIM : contexte

Les AIM sont des molécules organiques constituées d'un hétérocycle azoté. Chez *Catharanthus roseus*, un grand nombre de ces AIM dérive de la strictosidine considéré comme le premier AIM formé à partir duquel sont synthétisés tous les autres (Van der Heijden et *al*., 2004). Leurs faibles taux de biosynthèse *in planta*, ont conduit à la recherche de stratégies alternatives de production et à l'étude de leurs mécanismes de régulation afin d'optimiser la production de ces AIM. A ce titre, de nombreux travaux de recherche ont été menés dans les années 1980 avec notamment la sélection de culture cellulaire de *C. roseus* hautement productrice d'alcaloïdes (Deus et *al*., 1982). Les conditions de cultures (nutriments, facteurs de croissance, pH, température, lumière) ainsi que la sélection d'explants appropriés pour la mise en place de lignées cellulaires ont été très étudiés (Van der Heijden et *al*., 1989). Cependant l'utilisation de ce type de stratégie à un niveau industriel s'avère être limité car la vitesse de synthèse des métabolites d'intérêts reste trop lente (Verpoorte *et al.,* 1998) avec des taux de production encore insuffisants voire identiques à ceux retrouvés *in planta*. De plus, beaucoup d'alcaloïdes ne sont pas synthétisés dans des cultures cellulaires de *C. roseus,* puisque seuls les AIM monomériques comme l'ajmalicine, la serpentine ou la tabersonine ont été isolés (Hallard, 2000). Pour d'autres AIM, leur production s'avère compliquée. L'une des explications repose sur le fait que la voie de biosynthèse est extrêmement compartimentée dans la plante entière et requière des cellules hautement différenciées (El-sayed et Verpoorte, 2007). La culture d'hairy roots (chevelus racinaires) s'est avérée être une alternative prometteuse avec une croissance rapide des racines en milieu liquide et la mise en place de cellules différenciées capables de produire un large spectre d'AIM. Cependant même si les rendements de production sont meilleurs qu'en culture des cellules indifférenciées, la production des métabolites reste restreinte aux AIM monomères.

Toutefois, au cours du temps, à partir des années 1970-80, les cultures cellulaires de *C. roseus* se sont révélées être un matériel de choix en vue de caractériser la voie de biosynthèse des AIM et de comprendre les mécanismes régulant leur production. Ces travaux se sont

appuyés sur des études biochimiques visant à caractériser certaines étapes catalytiques et à purifier les enzymes natives (Zenk, 1991). Puis, à partir de la fin des années 1990, des approches de biologie moléculaire sont venues renforcer l'approche biochimique et ont permis d'isoler les premiers gènes codant les enzymes de la voie de biosynthèse des AIM.

A partir de 2010, on peut dire que *C .roseus* est devenue la référence en matière d'étude du métabolisme des AIM. Ce statut de plante modèle est en grande partie lié à l'avancée des connaissances sur son métabolisme alcaloïdique décrit comme l'un des plus complexes mettant en jeu plus de 30 étapes enzymatiques (Geu-Flores et *al*., 2012 ; Stavrinides et *al*., 2015). Cette avancée, impulsée par l'évolution récente des méthodes d'analyses transcriptomiques basées sur d'importants programmes de recherche comme « The Medicinal Plant Genomics Consortium» (Gòngora-castillo et *al*., 2012), « PhytoMetaSyn » (Xiao et *al*., 2013) ou encore « CathaCyc » (Van Moerkercke et *al*., 2013) ont permis d'accélérer la recherche de gènes impliqués dans le métabolisme des AIM. Couplées avec de nouvelles méthodes de caractérisations fonctionnelles des gènes, comme la technique VIGS (Virus-Induced Gene Silencing) (Carqueijeiro et *al*., 2015), ces études ont ouvert de réelles perspectives quant à la découverte et à la caractérisation complète de la voie de biosynthèse des AIM. Elles ont également permis de montrer la complexité de ce métabolisme, notamment au niveau de certaines étapes catalytiques, en dévoilant la présence de plusieurs isoformes enzymatiques.

Ces méthodes d'analyses transcriptomiques couplées aux techniques de caractérisation fonctionnelles de gènes par silencing (VIGS) misent en place au cours de ma thèse ont abouti à l'identification de nouveaux gènes présentés dans la partie résultat.

III.2 Voie de biosynthèse des AIM

Les recherches menées sur la caractérisation de la voie de biosynthèse des AIM ont conduit à l'élucidation de nombreuses enzymes et intermédiaires métaboliques ainsi qu'à une meilleure connaissance de l'architecture globale de la voie. Celle-ci peut se scinder en 3 blocs distincts (figure 7) : 1) la synthèse du précurseur indolique, la tryptamine, à partir de la voie du shikimate; 2) la synthèse du précurseur monoterpénique, la sécologanine dans la voie des

monoterpènes sécoiridoïdes (MTSI) à partir de la condensation des précurseurs terpéniques issus de la voie MEP ; 3) la synthèse de la strictosidine et les étapes post-strictosidine menant entre autre, jusqu'à la catharanthine et la vindoline qui, une fois condensés, aboutissent aux deux AIM dimères que sont la vinblastine et la vincristine (figure 7) .

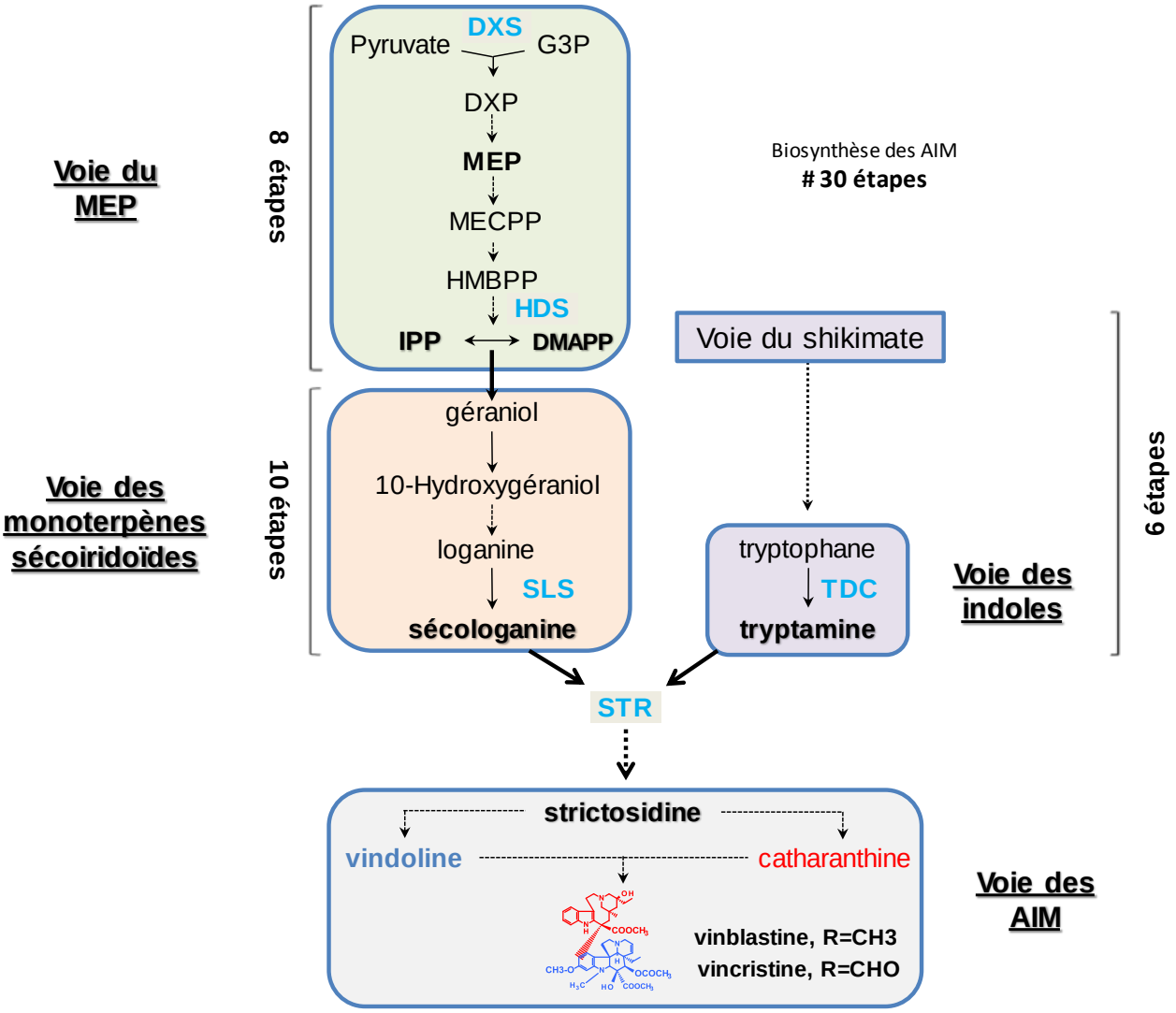


Figure 7 : Voie de biosynthèse simplifiée des alcaloïdes indoliques monoterpéniques chez *Catharanthus roseus.* DXS : 1-désoxy-D-xylulose 5-phosphate synthase (DXP synthase); **G3P** : glycéraldéhyde-3-phosphate ; **DXP** : 1-déoxy-D-xylulose 5-phosphate ; **MEP** : 2-C-méthyl-D-érythritol-4-phosphate ; **MECPP** : 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate ; **HMBPP** : 4-hydroxy-3-méthylbut-2-ényl diphosphate; **HDS** : 4-hydroxy-3-méthylbut-2-ényl

diphosphate synthase (HMBPP synthase); **IPP** : isopentényl diphosphate ; **DMAPP** : diméthyllallyl diphosphate ; **SLS** : sécologanine synthase ; **TDC** : tryptophane décarboxylase ; **STR** : strictosidine synthase. Une flèche continue indique une seule étape enzymatique alors qu'une flèche pointillée indique une succession d'étapes enzymatiques.

III.2.1 Biosynthèse du précurseur indolique des AIM : la tryptamine

La tryptamine est le précurseur indolique des AIM. Elle est issue de la décarboxylation du tryptophane (un acide aminé synthétisé dans la voie du shikimate) par la tryptophane décarboxylase (TDC) (De Luca et *al*., 1989 ; Penning et *al*., 1989). Cette enzyme est à l'interface entre le métabolisme secondaire (voie des indoles) et le métabolisme primaire de la voie du shikimate (figure 8). Celle-ci débute par la condensation de l'érythrose-4-phosphate dérivant du cycle de Calvin (photosynthèse) avec le phosphoénolpyruvate (voie des pentoses phosphates) (Maeda et Dudareva, 2012). La condensation de ces deux composés entraîne la formation du 3-désoxy-D-arabinoheptulosonate-7-phosphate (DAHP) qui subit ensuite une cyclisation pour former l'acide shikimique (shikimate). Ce dernier subit plusieurs réactions enzymatiques pour ensuite être transformé en chorismate par la chorismate synthase (CS). Le chorismate est de façon générale le précurseur de bon nombre d'acides aminés aromatiques (phénylalanine, tryptamine, tyrosine) qui alimentent plusieurs voies de biosynthèse et notamment celle des composés azotés comme les AIM. Le chorismate est ensuite transformé en anthranilate par l'anthranilate synthase (AS) en présence de glutamine. L'anthranilate est ensuite transformé en tryptophane par la tryptophane synthase (TS) puis ce dernier est décarboxylé en tryptamine par la TDC.

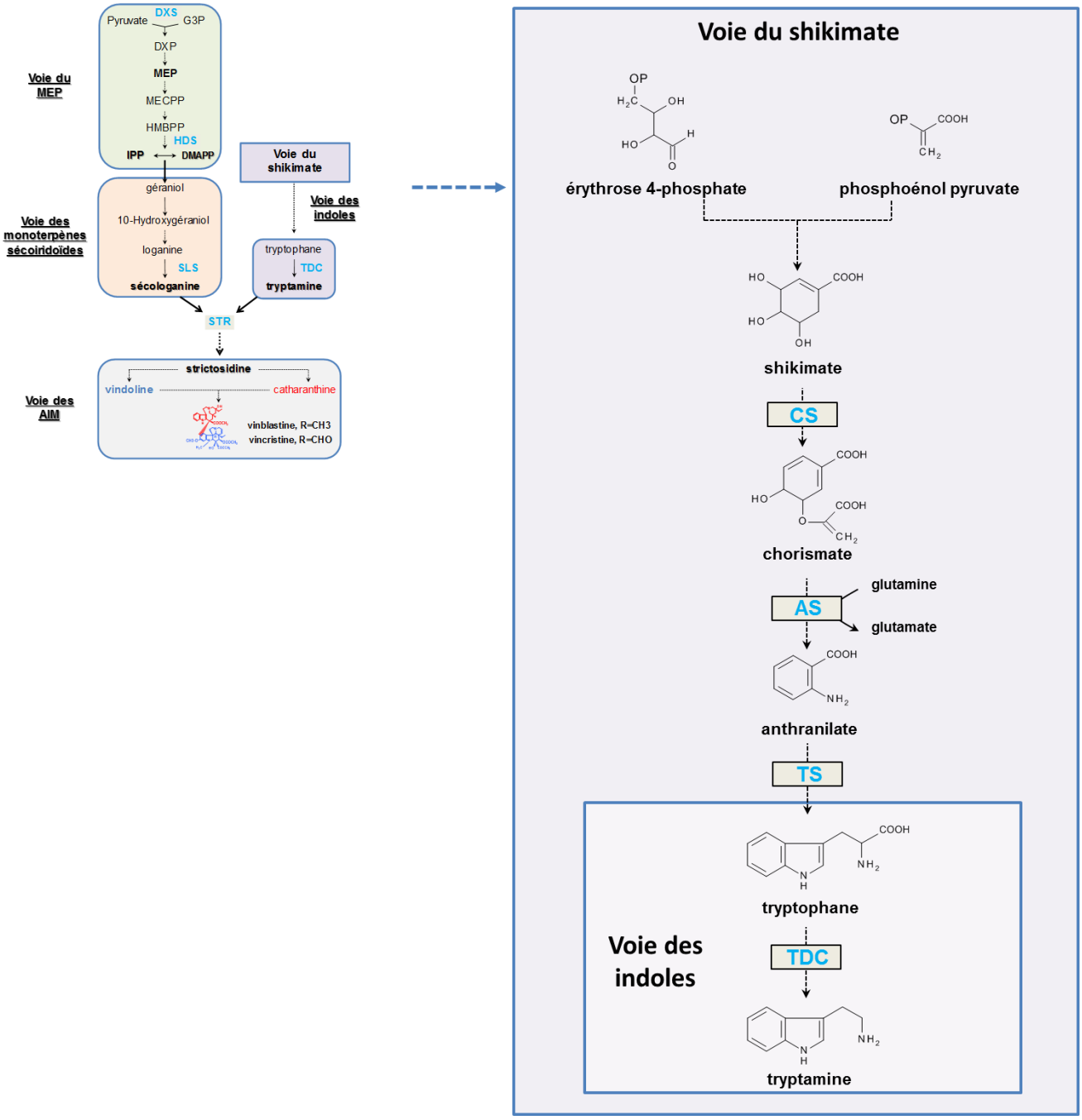


Figure 8 : Voie de biosynthèse de la tryptamine chez *Catharanthus roseus*, précurseur indolique des AIM. CS : chorismate synthase ; AS : anthranilate synthase ; TS : tryptophane synthase ; TDC : tryptophane décarboxylase. Les traits discontinus indiquent une succession de plusieurs étapes enzymatiques.

III.2.2 Biosynthèse du précurseur monoterpénique des AIM : la sécologanine

a) Production de l'isopentényl diphosphate et du diméthyllallyl diphosphate

La biosynthèse du précurseur monoterpénique (la sécologanine) repose au préalable sur un enchaînement de réactions enzymatiques permettant la synthèse de précurseurs isoprénoïdes que sont l'isopentényl diphosphate (IPP) et son isomère, le diméthyllallyl diphosphate (DMAPP) (figure 9 et figure 10). Chez les végétaux, l'IPP peut être synthétisé selon deux voies distinctes. La voie du MVA (acide mévalonique (figure 9)) caractérisée dans les années 1950. Longtemps considérée comme étant essentiellement cytosolique, des études récentes ont permis de montrer que les étapes finales de la voie du MVA étaient localisées au sein des péroxysomes (Simkin et *al*., 2011).Cette voie se compose d'une dizaine d'étapes enzymatiques (Bouvier et *al*., 2005) (figure 9). La première, correspond à la condensation de trois unités d'acétyl-CoA en 3-hydroxyméthyl-glutaryl-coA (HMG-CoA) par l'acétoacétyl-CoA (AACT) d'une part et l'hydroxy-3-méthyl-glutaryl-CoA synthase (HMGS) d'autre part. L'HMG-CoA est ensuite réduit par l'hydroxy-3-méthyl-glutaryl -CoA réductase (HMGR) en mévalonate, puis celui-ci est phosphorylé en mévalonate 5-diphosphate par la phosphomévalonate kinase (MVK) (Riou et al., 1994). Il est ensuite transformé par la mévalonate 5-phosphate kinase (PMK) en mévalonate 5-diphosphate (Tsay et Robinson, 1991). Ce dernier est finalement décarboxylé par la mévalonate-diphospho décarboxylase (MVD) pour former l'IPP (Cordier et *al*., 1999). L'isomérisation de l'IPP en DMAPP est ensuite assurée par l'isopentényl diphosphate isomérase (IDI).

L'IPP est également produit à partir de la voie du MEP, découverte initialement chez certaines bactéries (Rohmer et *al*., 1993) (figure 10), puis chez les plantes (Lichtenthaler et *al*., 1997). Localisée au niveau des plastes et des stromules, cette voie est considérée comme la source principale d'IPP et de DMAPP qui alimente la voie de biosynthèse des monoterpènes sécoiridoïdes. La voie MEP mène aussi à la production de caroténoïdes, des chlorophylles ou encore des quinones (Bouvier et *al*., 2005). La première étape de la voie MEP débute par la condensation du glycéraldéhyde-3-phosphate (G3P) et du pyruvate, en 1-déoxy-D-xylulose-5-phosphate (DXP), réaction catalysée par la 1-déoxy-D-xylulose-5-phosphate synthase (DXS). Chez *C. roseus* la DXS est issue d'une famille multigénique et existe sous trois isoformes : CrDXS, CrDXS1 et CrDXS2 codées respectivement par les gènes

Cr*dxs*, Cr*dxs1* et Cr*dxs2*. (Chahed et *al*., 2000 ; Han et *al*., 2013). Le DXP est ensuite réduit en 2–C-méthyl-D-érythritol-4-phosphate (MEP) par la 1-désoxy-D-xylulose 5-phosphate réductoisomérase (DXR). Ce dernier est alors condensé par la 4-cytidil-diphospho-2-C-méthyl-D-érythritol synthase (CMS) afin de donner le 4-diphospho-cytidyl-2C-méthyl-D-érythritol (CDP-ME). Celui-ci est ensuite phosphorylé par la 4-diphospho-cytidyl-2C-méthyl-D-érythritol kinase (CMK) pour donner le 4-diphosphocytidyl-2C-méthyl-D-érythritol 2 phosphate (CDP-MEP) puis le groupement cytidine de la CDP-MEP est supprimé par la 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate synthase (MECS) pour donner le 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate (MECPP). Ce métabolite est alors transformé en 4-hydroxy-3-méthylbut-2-ényl diphosphate (HMBPP) par l'action de la 4-hydroxy-3-méthylbut-2-ényl diphosphate synthase (HDS) et réduit par la 4-hydroxy-3-méthylbut-2-ényl diphosphate réductase (HDR) en IPP et en DMAPP. Au sein de la voie MEP, le pool d'IPP et de DMAPP est régulé par une réaction d'isomérisation de l'IPP en DMAPP effectuée par une enzyme IPP isomérase 1 (IDI1).

Des études utilisant des précurseurs radiomarqués ont montré que c'est la voie MEP qui participe principalement à l'élaboration des sécoiridoïdes et des AIM chez *C. roseus* (Contin et *al*., 1998 ; Hong et *al*, 2003).

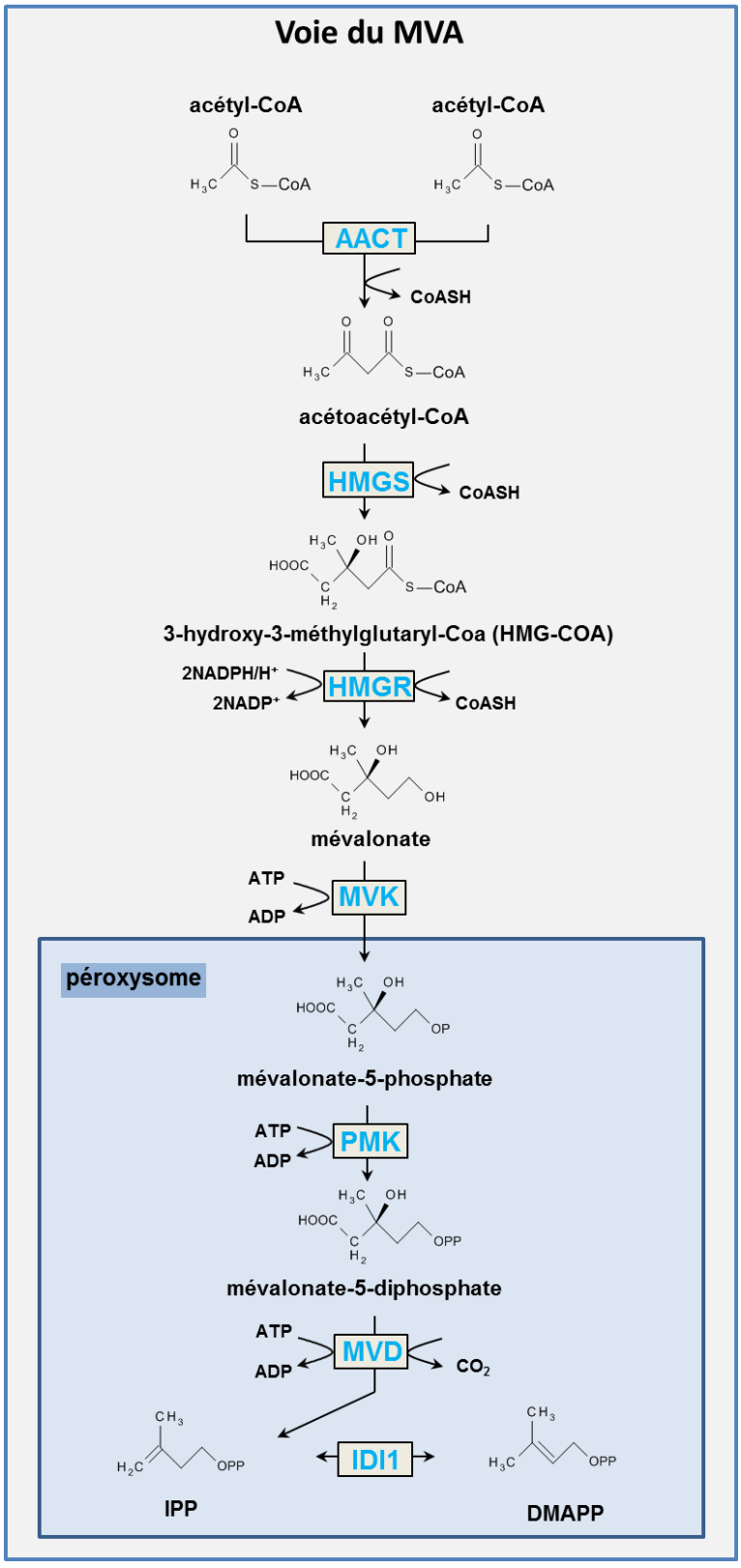
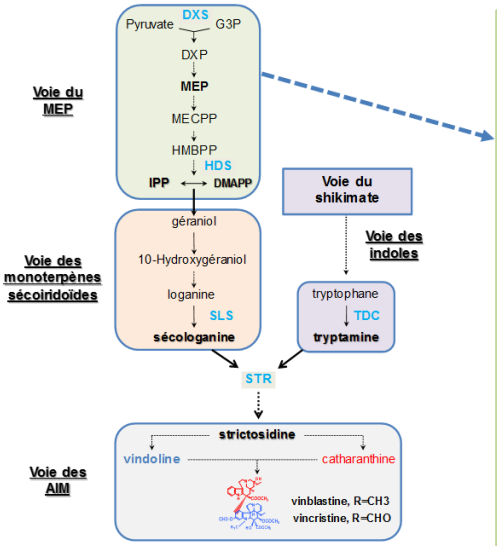


Figure 9 : Voie de biosynthèse de l'IPP à partir de la voie du mévalonate. AACT : acétoacétyl-CoA ; **HMGS** : hydroxy-3-méthyl-glutaryl-CoA synthase ; **HMGR** : hydroxy-3-méthl-glutaryl -CoA réductase ; **MVK** : mévalonate kinase ; **PMK** : mévalonate phosphate kinase ; **MVD** : mévalonate 5-diphosphate décarboxylase ; **IDI1** : IPP isomérase 1 ; **IPP** : isopentényl diphosphate ; **DMAPP** : diméthyllallyl diphosphate.


Voie du MEP
Pyruvate
DXS
G3P
DXP
MEP
MECPP
HMBPP
HDS
IPP
DMAPP
Voie du shikimate
Voie des indoles
Voie des monoterpènes sécoiridoïdes
géraniol
10-Hydroxygéraniol
loganine
SLS
sécologanine
tryptophane
TDC
tryptamine
STR
strictosidine
vindoline
catharanthine
vinblastine, R=CH3
vincristine, R=CHO
Voie des AIM

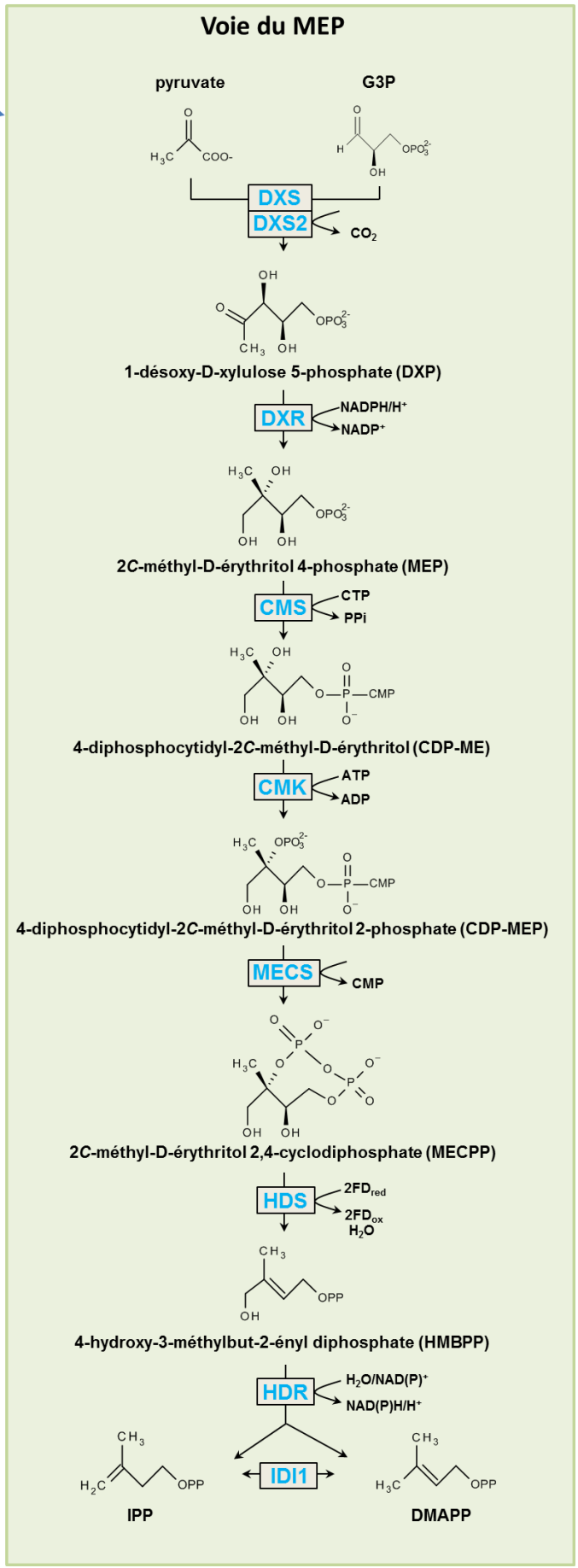
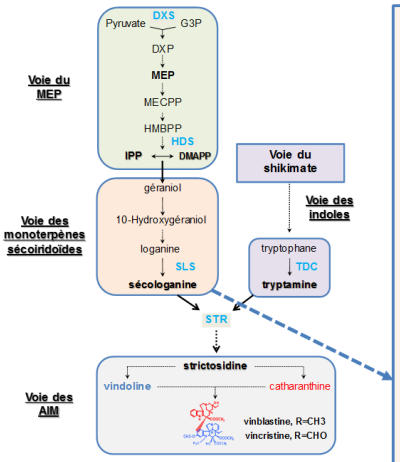

Voie du MEP
pyruvate
G3P
DXS
DXS2
CO2
1-désoxy-D-xylulose 5-phosphate (DXP)
DXR
NADPH/H+
NADP+
2C-méthyl-D-érythritol 4-phosphate (MEP)
CMS
CTP
PPi
4-diphosphocytidyl-2C-méthyl-D-érythritol (CDP-ME)
CMK
ATP
ADP
4-diphosphocytidyl-2C-méthyl-D-érythritol 2-phosphate (CDP-MEP)
MECS
CMP
2C-méthyl-D-érythritol 2,4-cyclodiphosphate (MECPP)
HDS
2FDred
2FDox
H2O
4-hydroxy-3-méthylbut-2-ényl diphosphate (HMBPP)
HDR
H2O/NAD(P)+
NAD(P)H/H+
IDI1
IPP
DMAPP

Figure 10 : Voie de biosynthèse de l'IPP et du DMAPP à partir de la voie du MEP. DXS et son isomère **DXS2** : 1-désoxy-D-xylulose 5-phosphate synthase ; **DXP** : 1-déoxy-D-xylulose 5-phosphate ; **DXR** : 1–désoxy-D-xylulose 5-phosphate réductoisomérase ; **MEP** : 2-C-méthyl-D-érythritol-4-phosphate ; **CMS** : 4-diphospho-cytidyl-2C-méthyl-D-érythritol synthase ; **CDP-ME** : 4-diphospho-cytidyl-2C-méthyl-D-érythritol ; **CMK** : 4-diphosphocytidyl-2C-méthyl-D-érythritol kinase ; **CDP-MEP**: 4-diphosphocytidyl-2C-méthyl-D-érythritol 2 phosphate ; **MECS** : 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate synthase ; **MECPP** : 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate ; **HDS** : 4-hydroxy-3-méthylbut-2-énYl diphosphate synthase; **HMBPP** : 4-hydroxy-3-méthylbut-2-énYl diphosphate; **HDR** : 4-hydroxy-3-méthylbut-2énYl diphosphate réductase; **IDI1** : isopentényl diphosphate isomérase 1 ; **DMAPP** : diméthyllallyl diphosphate ; **IPP** : isopentényl diphosphate.

b) Production de la sécologanine

La sécologanine est synthétisée au cours d'une succession d'une dizaine d'étapes enzymatiques (Miettinen et *al*., 2014) à partir des précurseurs (IPP et DMAPP) issus de la voie MEP (figure 11). L'IPP et le DMAPP sont condensés sous l'action de la géranyl diphosphate synthase GPPS pour former le géranyl diphosphate (GPP) : précurseur des composés monoterpéniques. Les GPPS sont des isoformes d'enzymes se présentant sous forme d'homodimères et d'hétérodimères. Les GPPS hétérodimères sont des complexes enzymatiques composés d'une petite sous-unité non catalytique (SSU) et d'une large sous-unité catalytique (LSU). Les GPPS homodimères quant à elles sont exclusivement constituées de deux larges sous-unités (LSU) qui diffèrent de celles des GPPS hétérodimères (Tholl et *al*., 2004). Chez *C. roseus*, deux formes de GPPS ont été découvertes : une GPPS homodimère ainsi qu'une forme de GPPS hétérodimère (Rai et *al*., 2013). Le GPP est ensuite hydroxylé en géraniol par la géraniol synthase GES (Simkin et *al*., 2013), puis hydroxylé en 10-hydroxygéraniol, par la géraniol 10-hydroxylase (G10H) (Collu et *al*., 2001). La G10H appartient à la famille des enzymes cytochromes P450. Ces enzymes sont associées à des NADPH-cytochrome P450 réductases (CPR) pour pouvoir effectuer la réaction enzymatique (Facchini, 2001 ; Meijer et *al*., 1993b). Au sein de la voie de biosynthèse des monoterpènes sécoiridoïdes, plusieurs étapes enzymatiques sont catalysées par des cytochromes P450. Ce type d'enzyme est très présent au sein du règne végétal et semble très important dans la conversion des précurseurs de la voie des MTSI (Schröder et *al*., 1999). Le G10H est ensuite

oxydé en dialdéhyde 10-hydroxygéranial par la 10-hydroxygéraniol-oxydoréductase (10HGO) (Miettenen et *al*., 2014) puis transformé en cis-trans-iridodial par l'iridoïde synthase (IS) (Geu-flores et *al*., 2012). Le cis-trans-iridodial est ensuite oxydé par un second complexe associant l'iridoide oxydase (IO, un autre cytochrome P450) avec une CPR P450 réductase pour donner l'acide 7-déoxyloganétique. Celui-ci est glycosylé par la 7-déoxyloganétique-acide-glycosyltransférase (7DLGT) (Asada et *al*., 2013) pour donner l'acide 7-déoxyloganique. Ce dernier est ensuite converti en acide loganique par un nouveau cytochrome P450, la 7-déoxyloganétique-acide 7-hydroxylase (7DLH) (Salim et *al*., 2013) puis l'acide loganique est méthylé par l'acide loganique méthyltransférase (LAMT) (Murata et *al*., 2008) pour former la loganine. Enfin, la loganine est transformée en sécologanine dans une réaction catalysée par la sécologanine synthase (SLS), un cytochrome P450 qui clive le motif cyclopentane de la loganine (Irmler et *al*., 2000). **Cette dernière joue un rôle clé dans la synthèse du précurseur monoterpénique des AIM. Un paragraphe dédié à cette enzyme sera développé dans la partie Résultats car elle fait l'objet d'une partie de mes de travaux de thèse.**


Voie du MEP
Pyruvate
G3P
DXS
DXP
MEP
MECPP
HMBPP
HDS
IPP
DMAPP
Voie du shikimate
Voie des monoterpènes sécoiridoïdes
géraniol
10-Hydroxygéraniol
loganine
SLS
sécologanine
Voie des indoles
tryptophane
TDC
tryptamine
STR
strictosidine
vindoline
catharanthine
vinblastine, R=CH3
vincristine, R=CHO
Voie des AIM

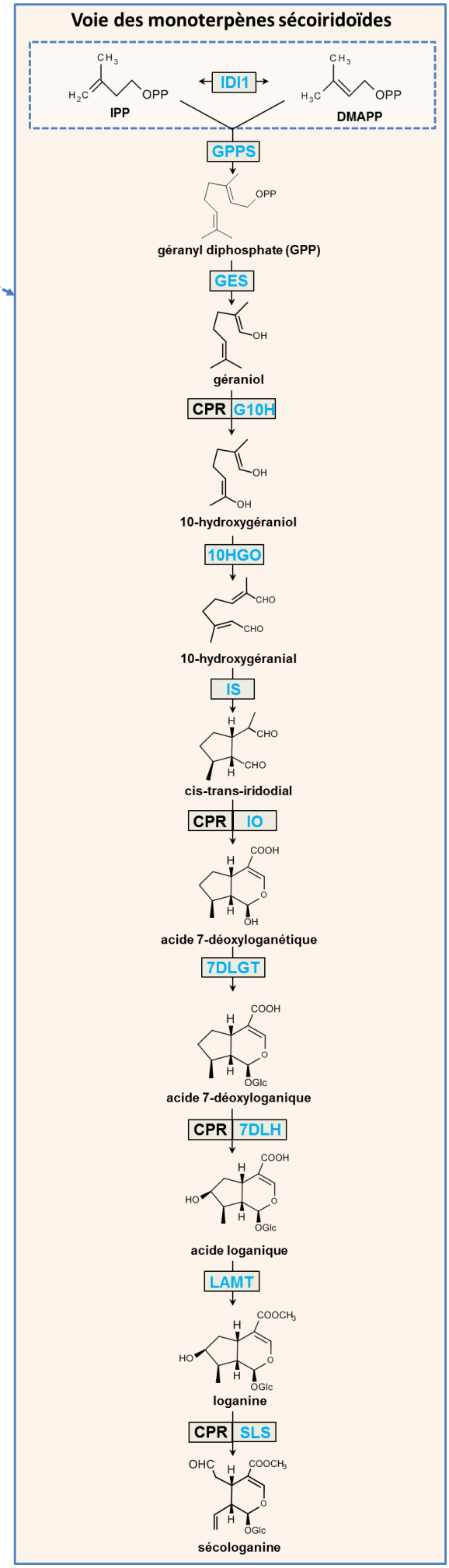

Voie des monoterpènes sécoiridoïdes
IDI1
IPP
DMAPP
GPPS
géranyl diphosphate (GPP)
GES
géraniol
CPR G10H
10-hydroxygéraniol
10HGO
10-hydroxygéranial
IS
cis-trans-iridodial
CPR IO
acide 7-déoxyloganétique
7DLGT
acide 7-déoxyloganique
CPR 7DLH
acide loganique
LAMT
loganine
CPR SLS
sécologanine

Figure 11 : Voie de biosynthèse des monoterpènes sécoiridoïdes conduisant à la synthèse du précurseur terpénique, la sécologanine. IPP : isopentényl diphosphate ; DMAPP : diméthylallyl diphosphate ; IDI1 : IPP isomérase 1 ; GPPS : géranyl diphosphate synthase ; GES : géraniol synthase ; G10H : géraniol 10-hydroxylase ; CPR : cytochrome P450 réductase ; 10HGO : 10-hydroxygéraniol oxydoréductase ; IS : iridoïde synthase ; IO : iridoïde oxydase ; 7DLGT : acide 7-déoxyloganétique glucosyltransférase ; 7DLH : 7-désoxyloganine 7-hydroxylase ; LAMT : acide loganique méthyltransférase ; SLS : sécologanine synthase.

III.2.3 Biosynthèse de la strictosidine et étapes post-strictosidine

La strictosidine, est formé à partir de la condensation de la sécologanine avec la tryptamine (De luca et *al*., 1987 ; Maresh et *al*., 2007) (figure 12). Cette réaction est catalysée par la strictosidine synthase (STR). Des études ont montré que cette enzyme est codée par un gène unique, identifié pour la première fois chez l'espèce *Rauwolfia serpentina* (Kutchan et *al*., 1988) puis chez *C. roseus* (Pasquali et *al*., 1992) et *Ophiorrhiza pumila* (Yamazaki et *al*., 2003). Des expériences d'immunoblot de protéines avec des anticorps spécifiques de STR, ont permis de montrer que les isoformes retrouvées étaient issues de modifications post-traductionnelles avec un repliement de la protéine qui est différente d'une espèce à l'autre. La présence de sept isoformes de STR (De Waal et *al.*, 1995) atteste de la possibilité de modifications post-traductionnelles.

Les étapes finales de conversion de la strictosidine vers des AIM complexes sont encore très mal connues avec seulement quelques étapes élucidées. Parmi ces étapes, seule la déglucosylation de la strictosidine en strictosidine aglycone catalysée par la SGD est connue (figure 12). Identifiée à l'origine dans des cultures cellulaires indifférenciées de *C. roseus*, cette enzyme possède la capacité de former des complexes multimères de très haut poids moléculaire (Luijendijk et *al*., 1998). Sa caractérisation fonctionnelle chez *C. roseus* et *R. serpentina* (Geerlings et *al*., 2000 ; Gerasimenko et *al*., 2002) a montré que SGD est codée par un seul gène montrant plus de 60% d'homologie avec les glucosidases de plantes. Les étapes enzymatiques suivantes conduisant, d'une part à la tabersonine et d'autre part à la catharanthine, via vraisemblablement la stemmadénine sont inconnues (figure 12).

Les AIM monomères de type « corynanthe », (ou « hétéroyohimbine ») comme l'ajmalicine et la tétrahydroalstonine sont issus de la cathénamine (Hemscheidt et zenk, 1985 ; Blom et *al*., 1991). **Les étapes catalytiques impliquées dans la biosynthèse d'AIM de type corynanthe ont été étudiées au cours cette thèse et ont permis d'identifier de nouvelles déshydrogénases/réductases. La voie de biosynthèse de la vindoline est aujourd'hui très bien caractérisée avec l'élucidation récente de deux enzymes intervenant dans les étapes centrales de conversion des dérivés de la tabersonine vers la vindoline. Elles sont présentées dans cette thèse, puis explicitées dans la partie résultats. Ces travaux ont permis d'amorcer une production de la vindoline dans un système hétérologue de levure**.

La synthèse de vindoline à partir de la tabersonine s'appuie sur 7 étapes enzymatiques : la première étape consiste en une hydroxylation de la tabersonine par la tabersonine 16-hydroxylase (T16H) pour produire la tabersonine 16-hydroxylée. Extraite à partir de feuilles de *C. roseus* cette enzyme est une monooxygénase de type P450 dépendante du NADPH (St-Pierre et De Luca, 1995). Plus récemment, des travaux ont montré l'implication d'une seconde isoforme de T16H (T16H2) impliquée dans la synthèse de 16 hydroxy-tabersonine. L'isoforme T16H2 est préférentiellement impliquée dans la synthèse de vindoline dans les jeunes feuilles alors que T16H1 est restreinte à la synthèse de vindoline dans les fleurs (Besseau et *al*., 2013).

L'étape suivante est la 16-O-méthylation de la 16 hydroxy-tabersonine par la 16 hydroxytabersonine 16-O-méthyltransférase (16 OMT) en 16 méthoxytabersonine (Murata et *al*., 2005 ; Levac et *al*., 2007) qui est convertie ensuite en 16-méthoxy-2,3-dihydroxy-3-hydroxy-tabersonine par une étape d'hydratation. De récentes découvertes effectuées en 2015 ont permis de montrer que cette hydratation de la 16 méthoxytabersonine était réalisée par deux étapes enzymatiques successives. Il s'agit tout d'abord d'une oxydation en époxytabersonine par la tabersonine-3-oxydase (T3O) appartenant également à la famille des cytochromes P450 (Kellner et *al*., 2015). Nous avons participé à ce travail bien qu'il ne soit pas présenté dans ce manuscrit de thèse. La seconde étape dans laquelle l'imine tabersonine (formée à la suite de réactions spontanées de l'époxytabersonine sur elle-même) est réduite en 16-méthoxy-2,3-dihydroxy-3-hydroxy-tabersonine par une déshydrogénase. la tabersonine-3-réductase (T3R) a été caractérisée récemment par un autre groupe (Qu et *al*., 2015) et nous-même et fait l'objet de la Partie 4 des résultats du présent manuscrit. La suite de la voie de

biosynthèse fait intervenir la 2,3-dihydro-3-hydroxytabersonine-N-méthyl transférase (NMT) qui assure la méthylation de la 16-méthoxy-2,3-dihydroxy-3-hydroxy-tabersonine aboutissant à la désacétoxyvindoline (De Luca et *al*., 1987 ; Liscombe et *al*., 2010). Ce composé est ensuite transformé en déacétylavindoline par une dioxygénase, la désacétoxyvindoline-4-hydroxylase (D4H) (De Carolis et De Luca, 1993 ; Vasquez-Flota et *al*., 1997). Enfin, la vindoline est produite par la déacétylvindoline 4-O-acétyltransférase (DAT) à partir de la déacétylvindoline (De Luca et *al*., 1985 ; St-Pierre et *al*., 1998).

Concernant, les AIM dimères d'intérêt pharmacologique tels que la vinblastine et la vincristine, ils sont issus de la condensation de la catharanthine et de la vindoline, réaction catalysée par une péroxydase 1 PRX1 (Sottomayor et *al*., 1998 ; Sottomayor et *al*., 2003). Cette condensation aboutit à la formation d'une molécule d'iminium qui sera transformée en 3,4-anhydrovinblastine. De cette dernière dériveront la vinblastine et sa forme oxydée la vincristine *via* des étapes enzymatiques non encore caractérisées.

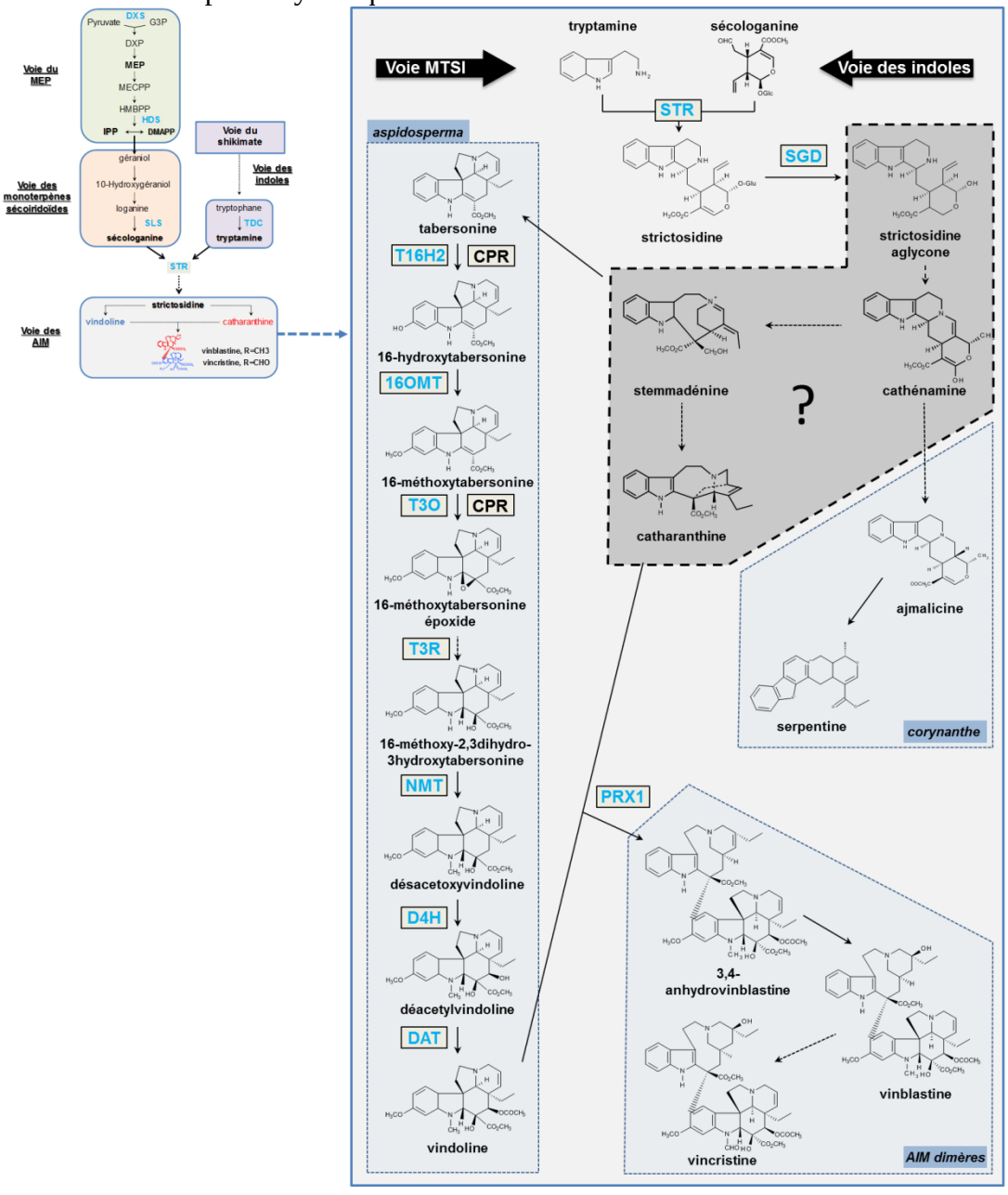


Figure 12 : Etapes finales de la voie de biosynthèse des AIM chez *Catharanthus roseus*. Le premier AIM formé est la strictosidine issue de la condensation de la sécologanine et de la tryptamine. La strictosidine est le précurseur de tous les AIM de *C. roseus.* **STR** : strictosidine synthase ; **SGD** : strictosidine β-D-glucosidase ; **CPR** : cytochrome P450 réductase ; **T16H2** : tabersonine 16-hydroxylase isomère 2 ; **16OMT** : 16-hydroxytabersonine-O-méthyltransférase ; **T3O** : 16-méthoxytabersonine 3-oxygénase ou tabersonine-3-oxydase ; **T3R :** tabersonine-3-réductase **;** **NMT** : 2,3-dihydro-3-hydroxytabersonine-N-méthyltranférase ; **D4H** : déacétoxyvindoline-4-hydroxylase ; **DAT** : déacétylvindoline-4-O-acétyltransférase ; **PRX1** : péroxydase 1.

III.3 Architecture de la voie de biosynthèse des AIM

L'augmentation croissante du nombre de gènes découverts codant des enzymes de la voie de biosynthèse des AIM chez *C. roseus* n'a eu de cesse d'alimenter le questionnement sur l'organisation et l'architecture de cette voie. Une connaissance plus approfondie de cette architecture devrait permettre une meilleure compréhension des flux métaboliques s'exerçant à travers les cellules et les différents tissus de la plante. Ainsi l'étude de la localisation d'enzymes impliquées dans le métabolisme des AIM chez *C. roseus* est devenue un axe de recherche intensif avec le développement de nouvelles techniques plus sensibles et plus spécifiques. Bien que certaines données fassent encore défaut, ces recherches ont permis d'établir pour *C. roseus* un modèle très abouti en terme de distribution tissulaire et cellulaire.

III.3.1 Organisation tissulaire

Chez *C. roseus*, les premières études visant à identifier et caractériser la localisation des enzymes impliquées dans la voie de biosynthèse des AIM ont été menées avec des approches de coloration histochimique dans les années 1980-1990 (De Luca et Cutler, 1987). Des techniques plus récentes d'hybridation *in situ* (St-Pierre et *al*., 1999 ; Burlat et *al*., 2004), de microdissection laser ou d'enrichissement d'épiderme par abrasion au carborundum (Murata et De Luca., 2005) ont été utilisées pour élucider l'architecture tissulaire de la voie de biosynthèse des AIM. Le schéma actuel fait état de trois niveaux tissulaires : (figure13), le parenchyme associé au phloème interne (PAPI), l'épiderme foliaire et les cellules spécialisées

des parenchymes palissadiques et lacuneux que sont les laticifères et les idioblastes (St-Pierre et *al*., 2013 ; Courdavault et *al*., 2014).

a) La synthèse des monoterpènes: du parenchyme associé au phloème interne jusqu'aux épidermes

Des analyses menées au cours de ses dix dernières années ont permis de montrer que les transcrits correspondants à la DXS, DXR, HDS, MECS et à IDI, enzymes assurant les étapes de biosynthèse des précurseurs terpéniques des AIM (Burlat et *al*., 2004 ; Oudin et *al*., 2007), ainsi que des transcrits de la GES, enzyme assurant la conversion du précurseur terpénique dans la voie des monoterpènes sécoiridoïdes (MTSI) (Simkin et *al*., 2013), sont associés au PAPI. De manière complémentaire, il a été établi que les étapes centrales de la voie des MTSI, coordonnées par la G10H, 10HGO, IS, IO, 7-DLGT et 7-DLH impliquent ce même tissu (Burlat et *al*., 2004 ; Geu-flores et *al*., 2012 ; Miettinen et *al*., 2014). Ces résultats montrent donc que la synthèse des précurseurs isoprénoïdes des AIM de la voie MEP et leur transformation en monoterpènes sécoiridoïdes s'effectuent très majoritairement au sein du PAPI. Cependant les deux dernières étapes de la voie MTSI impliquant la LAMT et la SLS sont quant à elles localisées au niveau de l'épiderme (foliaire), (Irmler et *al*., 2000 ; Guirimand et *al*., 2011 ; Geu-Flores et *al*., 2012). Ces résultats suggérent l'existence d'un transporteur, exprimé dans le mésophylle permettant l'acheminement de l'acide loganique depuis le PAPI jusqu'à l'épiderme, où il sera successivement transformé en loganine puis en sécologanine (figure 13).

b) Synthèse du premier AIM dans les épidermes

Plusieurs étapes importantes de la synthèse des AIM s'effectuent dans le tissu épidermique, avec notamment la formation de la strictosidine. En effet, en plus de la LAMT et la SLS, la TDC, assurant la formation du précurseur indolique, la tryptamine, a été localisée au niveau de l'épiderme (St Pierre et *al*., 1999). Les deux étapes suivantes sont aussi associées à ce tissu, puisqu'on y retrouve la STR et la SGD (Guirimand et *al*., 2010 ; Murata et *al*., 2005 ; Mahroug et *al*., 2006). Certaines étapes enzymatiques connues n'ont pas encore été localisées mais il semble toutefois que l'épiderme puisse puisse assurer partiellement la transformation de la tabersonine en vindoline. Ainsi, la T16H1, T16H2 ainsi que la 16OMT ont été localisées dans le tissu épidermique (Levac et *al*., 2008, Murata et *al*., 2008 ; Besseau

et *al*., 2013). La catharanthine quant à elle est synthétisée au niveau des épidermes. De récentes études ont permis de montrer que ce composé, toxique pour les insectes et champignons pathogènes est en grande partie excrété à la surface de la cuticule par un transporteur ATP dépendant nommé TPT2 (Yu et De luca, 2013 ; Roepke et *al*., 2010) (figure 13).

c) Etapes finales de la synthèse des AIM dans les laticifères et idioblastes

Les laticifères ainsi que les idioblastes sont bien connus pour leur capacité de stockage de bon nombre de substances toxiques de défense. Chez *C. roseus*, ces différents types cellulaires voient s'accumuler certains AIM (St Pierre et *al.,* 1999). A ce jour, seules les deux dernières étapes de la voie de biosynthèse de la vindoline catalysées respectivement par la D4H et la DAT ont été localisées au niveau des cellules spécialisées (St Pierre et *al*., 1999 ; Guirimand et *al*., 2011a). L'absence d'information concernant la distribution tissulaire de T3O, T3R et NMT ne permet pas aujourd'hui de prédire la nature du métabolite transitant de l'épiderme vers les cellules spécialisées (figure 13).

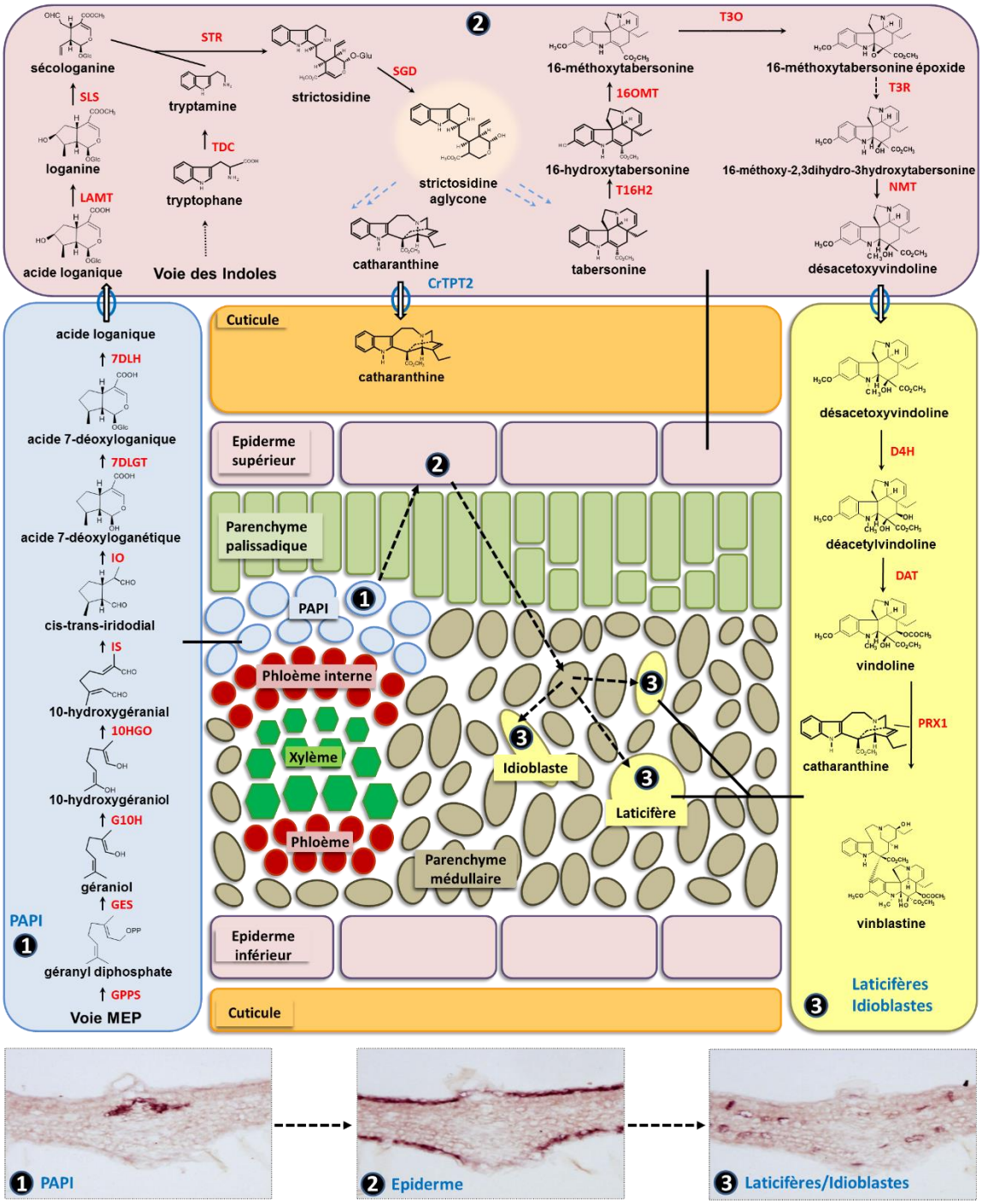


Figure 13 : Organisation tissulaire de la voie de biosynthèse des AIM chez *C. roseus* (adaptée d'après Courdavault et *al.*, 2014). Voie MEP : voie de biosynthèse du méthyl érytritol phosphate ; GPPS : géranyl diphosphate synthase ; GES : géraniol synthase ; G10H : géraniol 10-hydroxylase ; CPR : cytochrome P450 réductase ; 10HGO : 10-hydroxygéraniol oxydo-réductase ; IS : iridoïde synthase ; IO : iridoïde oxydase ; 7DLGT : acide 7-déoxyloganétique glucosyltransférase ; 7DLH : 7-désoxyloganine 7-hydroxylase ; LAMT : acide loganique méthyltransférase ; SLS : sécologanine synthase ; TDC : tryptophane

décarboxylase ; STR : strictosidine synthase ; SGD : strictosidine β-D-glucosidase ; T16H2 : tabersonine 16-hydroxylase ; 16OMT : 16-hydroxytabersonine-O-méthyltransférase ; T3O : tabersonine 3-oxydase ; T3R : tabersonine-3-réductase ; NMT : 2,3-dihydro-3-hydroxytabersonine-N-méthyltranférase ; D4H : déacétoxyvindoline-4-hydroxylase ; DAT : déacétylvindoline-4-O-acétyltransférase ; PRX1 : péroxydase 1 ; TPT2 : transporteur de la catharanthine ATP dépendant . Les flèches discontinues indiquent plusieurs réactions enzymatiques successives ; PAPI : parenchyme associé au phloème interne.

III.3.2 Organisation subcellulaire

Outre la compartimentation tissulaire originale, la voie de biosynthèse des AIM offre également une organisation subcellulaire des plus complexes. Elle transite par plusieurs compartiments, soulevant des questions quant au transport de métabolites, à leur stockage et à la régulation des flux de métabolites. Les premières études visant à étudier la localisation subcellulaire d'enzymes impliquées dans la voie de biosynthèse des AIM ont été menées par des approches classiques d'immunohistochimie ou par des fractionnements cellulaires associés à des études d'activité enzymatique. Au cours de ces dernières années, le développement récent de techniques basées sur la fusion de protéines fluorescentes telle que la *Green Fluorescent Protein (GFP*) (Guirimand et *al*., 2009) avec les protéines que l'on souhaite localiser a permis d'accélérer l'étude de la localisation subcellulaire d'enzymes impliquées dans la voie des AIM. De façon générale, l'adressage d'une protéine est déterminé par la présence de différents signaux d'adressage et/ou de rétention au sein de sa séquence protéique. Grâce à l'imagerie GFP, il est devenu possible d'entreprendre une étude systématique de la localisation subcellulaire des enzymes du métabolisme des AIM de *C. roseus*. Ainsi, aujourd'hui, il s'avère que l'architecture subcellulaire de la voie des AIM implique les plastes, le cytosol, le réticulum endoplasmique, la vacuole et le noyau. (figure 14).

Dans cette organisation subcellulaire, les plastes hébergent les 10 premières étapes enzymatiques assurant la conversion du G3P et du pyruvate jusqu'au géraniol (Courdavault et *al*., 2014). Plus particulièrement il a été mis en évidence chez *C .roseus* que les isoformes enzymatiques DXS, DXS2 ainsi que DXR sont localisées au niveau des plastoglobules un compartiment spécifique des plastes en relation avec le réseau de thylakoïdes. Les autres enzymes de la voie MEP, de CMS à HDR, présentent, quant à elles, une localisation diffuse dans le stroma des plastes incluant les stromules, qui sont des protubérances formées à partir

de la membrane externe des plastes. Leurs séquences protéiques présentent toutes, dans leur région N-terminal, une séquence d'adressage aux plastes. Des expériences d'imagerie GFP ont permis d'établir que la présence de ces motifs d'adressage N-terminaux était suffisante pour une localisation au sein du stroma (Guirimand et *al*., 2009).

Les enzymes qui font suite à la voie MEP pour aboutir à la formation de géraniol, à savoir l'isoforme longue (qui possède une séquence d'adressage plastidiale) de IDI1, la GPPS hétérodimérique et la GES sont toutes trois plastidiales (Simkin et *al*., 2011 ; Guirimand et *al*.,2012 ; Rai et *al*., 2013). Par ailleurs, il a été montré que la GEs est également présente dans les stromules des plastes.

Les réactions enzymatiques suivantes dans la voie des MTSI s'opèrent essentiellement dans le cytosol ou au niveau du réticulum endoplasmique (RE). On y retrouve notamment les P450 associés aux CPR formant des complexes protéiques ancrés, très vraissemblablement, au niveau de la face cytosolique du RE. Ainsi des expériences d'imagerie GFP ont montré que la G10H était ancrée au RE en accord avec la présence d'hélice transmembranaire sur son extrémité N-terminale (Guirimand et *al*., 2009). Ces mêmes études montré une association étroite entre le cortex du RE et les stromules des plastes. Le rapprochement de ces deux structures pourrait faciliter le passage du géraniol du plaste (synthétisé par la GES) vers le cytosol où il est converti en 10-hydroxygéraniol par la G10H ancrée au RE. Le 10-hydroxygéraniol est ensuite transformé dans le cytosole en 10-hydroxygéranial puis en cis-iridodial et trans-iridodial dans le cytosol par la 10HGO et l'IS (Miettinen et *al*., 2014 ; Geu-Flores et *al*., 2012). L'iridodial formé subit une oxydation catalysée par l'IO ancrée au RE sur pour former l'acide 7-déoxyloganétique. Celui-ci est transformé en acide 7 déoxyloganique dans le cytosol (par la 7DLGT), puis en acide loganiquepar la 7DLH ancrée au RE (Miettinen et *al*., 2014). La LAMT, conduisant à la loganine est cytosolique (Guirimand et *al*., 2011), la SLS, est localisée au niveau du RE (Guirimand et *al*., 2011a). Concernant la biosynthèse de la tryptamine, il a été montré que la TDC est cytosolique. Ces résultats ont donc permis d'établir que la biosynthèse du précurseur terpénique (sécologanine) et du précurseur indolique (tryptamine) des AIM s'achève dans le cytosol. Ces deux composés sont transportés au niveau de la vacuole où ils sont condensés par la STR pour former la strictosidine.

Cette enzyme possède un peptide d'adressage en N-terminal suivi d'une séquence SPIL, adressant la protéine vers la vacuole qui est son lieu d'accumulation (Guirimand et *al*.,

2010). La strictosidine est déglucosylée par la SGD générant l'aglycone de strictosidine. Cette réaction enzymatique s'effectue dans un compartiment cellulaire inattendu puisque des approches d'imagerie GFP ont permis de localiser la SGD dans le noyau (Guirimand et *al*., 2010). La localisation différentielle de la SGD dans le noyau et de son substrat produit dans la vacuole constitue un système de défense vis à vis de pathogènes qui a été précédemment au Chapitre II. Les étapes suivant la conversion de la strictosidine en aglycone sont très mal connues et les enzymes ne sont pas caractérisées. Pour autant il semble que les réactions successives de transformation de l'aglycone en catharantine ou en vindoline soient cytosoliques et impliquent vraisemblablement des P450 ancrés au RE.

Les réactions enzymatiques de la voie de biosynthèse de la tabersonine à la vindoline s'opèrent principalement dans le cytosol. De récentes études utilisant des techniques d'imagerie GFP ont permis d'établir que les deux isoformes de T16H était ancrée à la face cytosolique du RE (Guirimand et *al*., 2011 ; Besseau et *al*., 2013). D'autres travaux ont permis de montrer la localisation cytosolique de la 16 OMT (Guirimand et *al*., 2011a). La T3O est ancrée au RE (Kellner et al., 2015) et la T3R est cytosolique (Qu et *al*., 2015). La NMT a été associée au thylakoïdes des plastes de *C. roseus* à partir d'un gradient de fractionnement cellulaire (Dethier et De Luca, 1993), mais la localisation de cette enzyme doit être confirmée. Les deux dernières étapes de la voie de biosynthèse de la vindoline, catalysées respectivement par la D4H et la DAT sont cytosoliques (Guirimand et *al*., 2011a).

Sur la base de ces résultats il semble très important de souligner que la voie de biosynthèse des AIM possède un haut niveau de complexité, impliquant notamment de nombreux éléments de régulation dont font partie la, compartimentation tissulaire spécifique et la compartimentation inter-organites au sein des cellules. Aussi, la compartimentation elle-même et les mécanismes de transports intercellulaires et inter-organites qu'elle nécessite sont autant de facteurs pouvant limiter le flux métabolique et qu'il est nécessaire d'appréhender en profondeur si l'on souhaite développer des plateformes d'ingénierie métabolique visant à produire les AIM, que ce soit in planta ou dans des microorganismes.

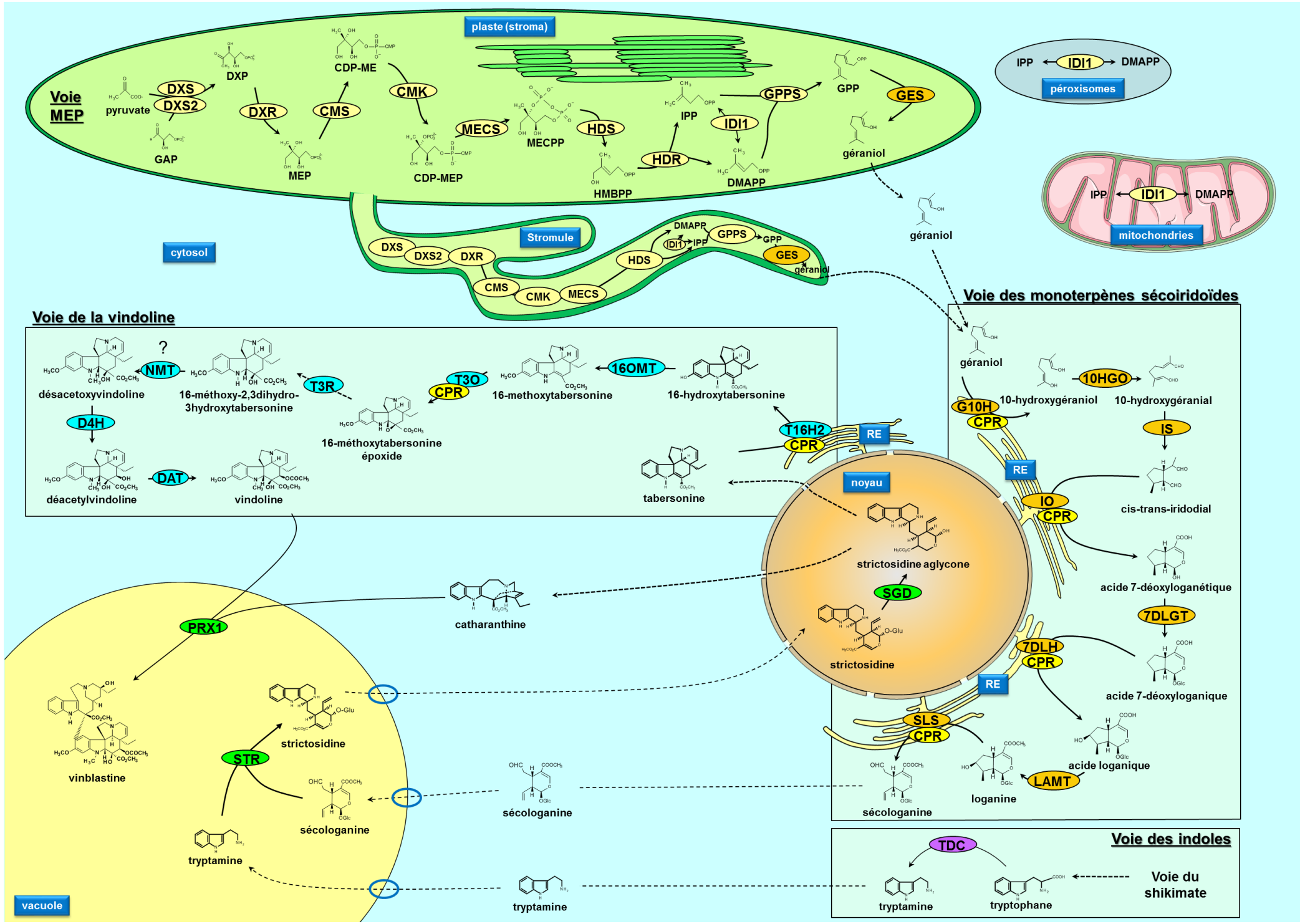

plaste (stroma)
Voie MEP
pyruvate
DXS
DXS2
DXP
DXR
MEP
CMS
CDP-ME
CMK
CDP-MEP
MECS
MECPP
HDS
HMBPP
HDR
IPP
IDI1
DMAPP
GPPS
GPP
GES
géraniol
GAP
cytosol
Stromule
DXS
DXS2
DXR
CMS
CMK
MECS
HDS
DMAPP
IDI1
IPP
GPPS
GPP
GES
géraniol
IPP
IDI1
DMAPP
péroxisomes
IPP
IDI1
DMAPP
mitochondries
Voie de la vindoline
désacetoxyvindoline
NMT
?
16-méthoxy-2,3dihydro-3hydroxytabersonine
T3R
16-méthoxytabersonine époxide
T3O
CPR
16-methoxytabersonine
16OMT
16-hydroxytabersonine
D4H
déacetylvindoline
DAT
vindoline
T16H2
CPR
RE
tabersonine
Voie des monoterpènes sécoiridoïdes
géraniol
10HGO
10-hydroxygéraniol
10-hydroxygéranial
G10H
CPR
IS
RE
cis-trans-iridodial
IO
CPR
acide 7-déoxyloganétique
7DLGT
acide 7-déoxyloganique
7DLH
CPR
RE
SLS
CPR
acide loganique
LAMT
loganine
sécologanine
noyau
strictosidine aglycone
SGD
strictosidine
catharanthine
PRX1
vinblastine
strictosidine
STR
sécologanine
tryptamine
vacuole
sécologanine
tryptamine
Voie des indoles
TDC
tryptamine
tryptophane
Voie du shikimate

Figure 14 : Organisation subcellulaire de la voie de biosynthèse des AIM chez *C. roseus.* **DXS** et son isomère **DXS2** : 1-désoxy-D-xylulose 5-phosphate synthase ; **DXP** : 1-déoxy-D-xylulose 5-phosphate ; **DXR** : 1–désoxy-D-xylulose 5-phosphate réductoisomérase ; **MEP** : 2-C-méthyl-D-érythritol-4-phosphate ; **CMS** : 4-diphospho-cytidyl-2C-méthyl-D-érythritol synthase ; **CDP-ME** : 4-diphospho-cytidyl-2C-méthyl-D-érythritol ; **CMK** : 4-diphosphocytidyl-2C-méthyl-D-érythritol kinase ; **CDP-MEP**: 4-diphosphocytidyl-2C-méthyl-D-érythritol 2 phosphate ; **MECS** : 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate synthase ; **MECPP** : 2-C-méthyl-D-érythritol-2,4-cyclodiphosphate ; **HDS** : 4-hydroxy-3-méthylbut-2-ényl diphosphate synthase; **HMBPP** : 4-hydroxy-3-méthylbut-2-ényl diphosphate; **HDR** : 4-hydroxy-3-méthylbut-2ényl diphosphate réductase; **IDI1** : isopentényl diphosphate isomérase 1 ; **DMAPP** : diméthyllallyl diphosphate ; **IPP** : isopentényl diphosphate ; **GPPS :** géranyl diphosphate synthase **; GES :** géraniol synthase **; G10H :** géraniol 10-hydroxylase **; CPR :** cytochrome P450 réductase **; 10HGO :** 10-hydroxygéraniol oxydo-réductase **; IS :** iridoïde synthase **; IO :** iridoïde oxydase **; 7DLGT :** acide 7-déoxyloganétique glucosyltransférase **; 7DLH :** 7-désoxyloganine 7-hydroxylase **; LAMT :** acide loganique méthyltransférase **; SLS :** sécologanine synthase **; TDC :** tryptophane décarboxylase **; STR :** strictosidine synthase **; SGD :** strictosidine β-D-glucosidase **; T16H2 :** tabersonine 16-hydroxylase **; 16OMT :** 16-hydroxytabersonine-O-méthyltransférase **; T3O** : tabersonine-3-oxydase ; **T3R :** tabersonine-3-réductase **; NMT** : 2,3-dihydro-3-hydroxytabersonine-N-méthyltranférase ; **D4H** : déacétoxyvindoline-4-hydroxylase ; **DAT** : déacétylvindoline-4-O-acétyltransférase ; **PRX1** : péroxydase 1 ; **TPT2 :** transporteur de la catharanthine dépendant de l'ATP**.** Les flèches discontinues indiquent plusieurs réactions enzymatiques successives ainsi que des transports de métabolites.

Chapitre IV : Production des AIM de *Catharanthus roseus* par ingénierie métabolique

Depuis longtemps, les Hommes ont utilisé des substances tirées de leurs environnements naturels pour se soigner, bénéficiant sans le savoir d'une large classe de métabolites secondaires aux propriétés thérapeutiques. Ces métabolites secondaires, représentent aujourd'hui plus d'un tiers des composés thérapeutiques utilisés (Newman et Cragg, 2012). Au 19ème siècle, les chimistes ont commencé à synthétiser des molécules pour élaborer des médicaments. Aujourd'hui, après un peu moins de 200 ans de synthèse chimique, l'industrie pharmaceutique se tourne vers des stratégies alternatives pour élaborer des produits de santé qui visent à substituer aux procédés chimiques des procédés catalytiques biologiques. Ces stratégies reposent sur l'ingénierie métabolique qui, à travers le développement de nouveaux outils moléculaires et la manière d'appréhender les systèmes vivants, est aujourd'hui considérée comme un domaine de la biologie de synthèse

IV.1 La biologie de synthèse

La biologie de synthèse (ou biologie synthétique) est une discipline en émergence. Elle s'appuie sur la biologie des systèmes et le génie génétique tout en y ajoutant une dimension d'ingénierie fondée sur des principes de standardisation et de modélisation. Son dessein est de concevoir des systèmes complexes nouveaux, dotés ou non de fonctions déjà présentes dans la nature.

Ainsi la biologie synthétique se situe est double et se situe dans la compréhension des principes gouvernant la biologie (apprendre en construisant), notamment avec la construction de composés élémentaires d'ADN mais aussi dans la construction d'organismes accomplissant des fonctions biologiques complexes. L'ADN, considéré comme le principal support de l'information à l'origine du maintien de l'homéostasie cellulaire, fait de la biologie moléculaire une base essentielle de la biologie synthétique. Les découvertes fondamentales de la biologie moléculaire ont abouti au développement des techniques du génie génétique dans les années 1970. Les résultats concluant d'expérimentations sur l'ADN recombinant ont ouvert des perspectives industrielles notamment pour la production de protéines

thérapeutiques : la société Genentech s'est ainsi rendue célèbre, en 1978, pour avoir réussi à exprimer l'insuline recombinante humaine dans les bactéries (Crea et *al*., 1978).

Puis, avec l'automatisation des techniques de séquençage de l'ADN, la biologie est rentrée au milieu des années 1990 dans l'ère des sciences omiques générant depuis, de vastes ressources génomiques. Celles-ci, alliées aux développements d'outils de plus en plus performants du génie génétique, ont ouvert la voie à une ingénierie de la biologie qui se veut plus rationnelle et qui s'appuie sur les sciences de l'ingénieur (modélisation informatique, mathématiques…) pour concevoir ou modifier des systèmes biologiques dotés de caractéristiques innovantes (Képès, 2011). De nombreuses approches sont adoptées en biologie de synthèse. Parmi celles les plus couramment décrites, nous trouvons**, les approches top-down** (de haut en bas) et **bottom-up** (de bas en haut).

L'approche top-down est basée sur la déconstruction du vivant. Elle vise à disséquer les systèmes biologiques pour parvenir à les simplifier au maximum en supprimant les fonctions non essentielles à leur survie, leur laissant notamment un génome minimum. L'enjeu biotechnologique est d'obtenir une cellule-châssis facile à cultiver et à manipuler qui apporterait l'horloge, les quantités de matière et l'énergie nécessaire, pour optimiser l'expression des gènes introduits et la fonction que l'on souhaite réaliser (Forster et *al*., 2006). Cette approche s'accorde avec la devise de la biologie synthétique qui conçoit « d'apprendre en construisant ». L'exemple le plus à même d'illustrer cette approche concerne les travaux réalisés sur la bactérie *Mycoplasma genitalium*. L'étude a révélé que sur les 487 gènes codant des protéines, 387 ainsi que 43 gènes ARN sont nécessaires et suffisants à la croissance de la bactérie (Glass et *al*., 2006).

L'approche bottom-up (de bas en haut), quant à elle, est une vision constructive. Les fonctions biologiques sont assemblées en modules ou briques génétiques que l'on assemble hiérarchiquement, selon les caractéristiques recherchées, à la manière de composants de circuits électroniques (Zhang et *al*., 2011). Cette approche est employée pour le design de génome assisté par ordinateur, permettant de tester *in silico* l'assemblage de gènes dans un logiciel lié à une base de données de briques génétiques élémentaires comme les BioBricks® (Rouilly et *al*., 2007). Les BioBricks® ou « légos moléculaires » sont des séquences d'ADN (intégrant des gènes ou des séquences de régulation génique) que l'on peut aisément assembler dans le but de faciliter la bio-ingénierie des systèmes biologiques (Knight, 2003).

IV.2 Production de molécules biosynthétiques par ingénierie métabolique

L'ingénierie métabolique est une méthode qui permet le développement de nouveaux bioprocédés faisant appel à des usines cellulaires pour produire des composés originaux ou des molécules difficiles à obtenir par synthèse chimique. Ainsi, elle consiste à construire, modifier et/ou améliorer des voies métaboliques par l'ajout, le retrait ou la modification des gènes impliqués dans les processus biochimiques (Stephanopoulos et *al*., 1999 ; Yang et *al*., 1998). Elle peut être appliquée à l'organisme source ou à des organismes hôtes facilement manipulables dans lesquels on a reconstruit des voies métaboliques hétérologues.

La biologie de synthèse a un rôle fondamental à jouer dans le développement de l'ingénierie métabolique. En effet, la conception de nouveaux outils moléculaires (comme les BioBricks) et de programmes de modélisation conduiront à une standardisation des procédures et réduiront considérablement le temps et les coûts associés à l'ingénierie métabolique (Keasling ; 2012).

Comme son nom l'indique, l'ingénierie métabolique porte sur le métabolite, molécule de faible poids moléculaire qui se démarque des macromolécules comme les protéines qui sont traditionnellement les produits issus du génie génétique. Assez curieusement, il n'existe pas, à l'heure actuelle, un terme précis permettant de différencier un même métabolite produit par ingénierie métabolique ou par synthèse chimique. Aussi, afin de clarifier ces aspects, nous proposons d'adopter les termes suivants. Le terme de métabolite biosynthétique recouvrera les molécules de faible poids moléculaire issues de la biologie de synthèse par ingénierie métabolique. Par ailleurs, on fera la distinction entre un métabolite semi-biosynthétique qui est produit chimiquement à partir d'un précurseur élaboré par ingénierie métabolique et un métabolite semi-synthétique, élaboré par synthèse chimique à partir d'un précurseur extrait d'une source biologique naturelle.

IV.2.1 Importance du châssis cellulaire : levure optimisée pour la production

L'utilisation d'organisme hétérologue (différent de l'organisme source), comme « usine cellulaire » pour produire des métabolites d'intérêts nécessite toutefois quelques

manipulations afin d'obtenir un châssis cellulaire optimisé pour la production, comprenant un métabolisme endogène *in vivo* et des flux de synthèse tournés vers la production de molécules biosynthétiques, tout en conservant l'intégrité et la viabilité du système. Dans un premier temps, le choix de l'hôte cellulaire est déterminant, et doit montrer de bon prérequis pour la production de métabolites d'intérêts. A ce titre, les organismes procaryotes ont été très utilisés dans le développement de nombreux médicaments et notamment pour la production de protéines recombinantes comme l'insuline. Seulement, la complexité enzymatique des voies de biosynthèses de plantes ainsi que leur architecture subcellulaire ont fait des organismes eucaryotes, des hôtes de choix pour la synthèse de métabolites exprimés en plantes. L'exemple le plus à même d'illustrer cette approche de métabolisme engineering sont les travaux Ro et *al*., 2006, avec la production de l'acide artémisinique dans un système hétérologue de levure. Le choix de *S. cerevisiae* s'est effectué dans un second temps après avoir essayé de produire l'amorphadiene (précurseur de l'artémisinine) en *E. coli*. Pour travailler dans cette souche bactérienne, les auteurs ont dû transférer une partie de la voie du mévalonate de *S. cerevisiae* complémentée avec des gènes de *Staphylococcus aureus* pour obtenir un métabolisme basal permettant une synthèse orientée vers le FPP (précurseurs terpénique essentiel pour la production des précurseurs de l'artémisinine en plante). Chez *S. cerevisiae*, la voie de biosynthèse du mévalonate aboutit sur la production du FPP qui est un précurseur de l'ergostérol, lui-même utilisé pour fabriquer des composants de la paroi des levures. Ce modèle eucaryote présente un métabolisme basal adéquat pour produire à partir du FPP de l'acide artémisinique lorsqu'une partie de la voie de biosynthèse de l'artémisine *d'Artémisia Annua* lui est transférée (Ro et *al*., 2006 ; Paddon et Keasling, 2014).

L'obtention d'un « châssis cellulaire » optimisé pour la production de molécules biosynthétiques nécessite parfois même de modifier l'expression de gènes impliqués dans des processus biochimiques du métabolisme de l'organisme hôte, afin d'augmenter ou de faire converger l'ensemble de son métabolisme vers un métabolite précis. Ce concept du « gene engineering » donne souvent lieu à une amélioration de l'expression des gènes *in vivo* entraînant une augmentation des flux de synthèse. Les travaux de Donald et *al*., 1997 ont montré une augmentation de l'expression de l'HMGR impliquée dans la voie du mévalonate de *S. cerevisiae* lorsque celle-ci était tronquée tHMGR (exprimant que le site catalytique de l'enzyme), entraînant par la suite une meilleure production du mévalonate. Cette isoforme tronquée est d'ailleurs utilisée dans la production d'artémisine et de strictosidine en système hétérologue levure (Paddon et Keasling, 2014 ; Brown et *al*., 2015). Un autre exemple de

« gene engineering » est illustré dans les travaux de Brown et *al*., 2015, avec l'utilisation d'une farnesyl diphosphate synthase mutée mFPS144 de poulet (Stanley Fernandez et *al*., 2000) possédant une activité GPP synthase plus importante que FPP synthase. Cette mutation du site catalytique, entraîne une synthèse préférentielle du GPP d'ordinaire non produit chez *S. cerevisiae*. Cette modification génique est mise à profit dans la levure pour orienter la voie du mévalonate endogène vers la production de GPP, précurseur essentiel dans la voie des monoterpènes sécoiridoïdes pour produire la strictosidine (Brown et *al*., 2015).

Associé à ces modifications géniques, il semble parfois nécessaire d'amplifier ou au contraire de diminuer l'expression de certains gènes intervenant dans le métabolisme de l'hôte de façon à canaliser les flux de métabolites vers la production des molécules recherchées. Dans leur publication, Ro et *al*., 2006, ont utilisé un promoteur répressible (PMET3) pour réprimer l'expression de la squalène synthase (*ERG9*) (qui catalyse la première étape de biosynthèse des stérols après la formation du FPP). Par la même occasion ils ont sur-exprimé le facteur de transcription *upc2-*, qui régule la biosynthèse des stérols afin de favoriser la synthèse de FPP pour la production d'amorphadiène (Ro et *al*., 2006). Par ailleurs, certaines enzymes associées aux voies de biosynthèse, utilisent certains cofacteurs tels que le NADPH qui est produit dans la voie des pentoses phosphates. Une augmentation de l'expression du gène *ZWF1* codant la G6PDH impliquée dans la synthèse du NADPH chez les levures, entraîne une augmentation de l'accumulation du NADPH au profit de la production de xylitol (Kwon et *al*., 2006). Les cytochromes P450 (CYP450) sont des enzymes couramment retrouvées dans les voies métaboliques de plantes, elles forment des complexes moléculaires avec des cytochromes P450 réductase (CPR). Dans ce complexe les CPR sont nécessaires pour transférer des électrons vers les CYP450 pour permettre l'utilisation d'un proton provenant du NADPH, une fois avoir obtenu un châssis cellulaire optimal, avec un métabolisme endogène, convergeant vers la synthèse de précurseurs adaptés. Pour la production de molécules recherchées il convient de transférer des parties, voire des voies de biosynthèse entières pour reconstituer des voies de biosynthèse hétérologues *in vivo*.

IV.2.2 Standardisation des méthodes de transfert de gènes en levure

La synthèse de molécules biosynthétiques par ingénierie métabolique requiert ensuite une méthode efficace de transfert ou de modification des gènes appartenant aux voies métaboliques. Dans les premières études, des plasmides auto-réplicatifs adaptés pour

l'expression de gènes en levure ont été utilisés, posant parfois un problème de stabilité d'expression malgré la présence d'ARS (séquence autonome de réplication). Pour pallier à ce problème majeur de nouveaux vecteurs d'expression ont été construits dans le but d'intégrer des séquences directement dans le génome des levures. Ces stratégies ont permis l'essor des systèmes URA blaster assurant un transfert de gènes par recombinaison homologue. Toutefois ces techniques de recombinaison homologue ne sont pas optimisées pour transférer des parties ou des voies métaboliques entières, d'autant plus que les séquences d'intégration utilisées pour la recombinaison à un locus précis sont de taille importante et peuvent gêner l'intégration d'autres gènes. Le système Cre-loxP est quant à lui connu comme une technologie de recombinaison de l'ADN ayant lieu au niveau de petits sites-spécifiques, largement utilisée pour effectuer des suppressions, des insertions, des translocations et inversions dans l'ADN des cellules. (Deng, 2012). Le système de recombinaison Cre-loxP est emprunté au bactériophage P1 (bactériophage tempéré d'*Escherichia coli*) (Argos et *al*., 1986 ; Sternberg et *al*., 1986 ; Sauer et Henderson, 1988), dans lequel le gène *Cre* code une recombinase à ADN de 38-kDa appelée Cre qui reconnaît des petits sites ADN loxP (locus de crossing over du bactériophage P1) de 34 paires de bases et catalyse une recombinaison intra et intermoléculaire entre deux sites loxP. Le site loxP se compose d'une région centrale de 8 paires de bases encadrées de séquence palindromiques de 13 paires de bases (figure 15). Ce système permet une recombinaison entre deux séquences loxP et une excision de l'ADN se trouvant entre les deux sites loxP en formant un ADN circulaire.

Sur la base de cette nouvelle technologie, de nouveaux vecteurs (plasmides pXP) exploitant le système Cre-loxP ont été designés pour permettre l'intégration des gènes d'une voie métabolique dans le génome des levures (Fang et *al*., 2010). Ces vecteurs possèdent un squelette commun avec la plupart des plasmides utilisés pour exprimer des gènes en levure. Notamment la présence de sites multiples de clonage (dans lesquels sont clonés les transgènes que l'on souhaite exprimer) encadrés par des promoteurs (pPGK et pTEF1) et terminateurs (tCYC1) de levure ainsi que la présence de marqueurs métaboliques de levure (CAN1, HIS3, LEU2, MET15, TRP1, URA3) idéales pour sélectionner les souches transformées. Ils sont également dotés d'outils moléculaires (origine de réplication pMB1, gène de résistance à l'ampicilline) pour leur propagation en bactérie. L'originalité de ces constructions vient de la présence de sites loxP positionnés de part et d'autre de marqueurs métaboliques sur les plasmides pXP, permettant un recyclage de ces derniers par recombinaison homologue au niveau des sites loxP, une fois insérées dans le génome des levures (Deng, 2012). La mise en

place d'un tel système d'intégration de gènes chez les levures nécessite en plus de l'utilisation d'un plasmide pXP (dans lequel est cloné le gène que l'on souhaite exprimer en levure), l'utilisation d'un second plasmide pYES-Cre exprimant la recombinase virale Cre pour exciser le marqueur métabolique sous forme d'ADN circulaire par recombinaison homologue au niveau des sites loxP.

Pour ce faire la souche de levure est transformée selon la méthode de Nielsen et *al*., 2007 dans laquelle le plasmide pXP sert de base pour générer par amplification PCR deux fragments d'ADN chevauchants. Le premier, composé d'un gène (gène X) que l'on souhaite exprimer en levure encadrée par un promoteur (P_{PGK}) et terminateur (T_{CYC1}) et d'une séquence chevauchante à son extrémité 3' (CHE), le tout flanqué par une séquence homologue d'un marqueur métabolique (HIS3) en 5' : HIS3::P_{PGK}-gèneX-T_{CYC1}::CHE. Le second, constitué d'une séquence chevauchante CHE en son extrémité 5', d'un gène codant pour un marqueur métabolique de levure (LEU2) encadré de deux sites de recombinaison loxP ainsi qu'une séquence homologue d'un marqueur métabolique de levure en 3'(HIS3): CHE ::loxP-LEU2-loxP ::HIS3. Ces deux fragments d'ADN sont ensuite utilisés en plus du plasmide pYES-Cre pour transformer une souche de levure. Dans notre exemple, l'intégration du gène X va s'effectuer par recombinaison homologue entre les deux brins d'ADN au niveau des séquences chevauchantes internes ainsi qu'au niveau du marqueur métabolique HIS3. La recombinase Cre quant à elle va exciser le marqueur métabolique LEU2 sous la forme d'un ADN circulaire, permettant de réutiliser ce même marqueur pour l'intégration d'un autre gène.

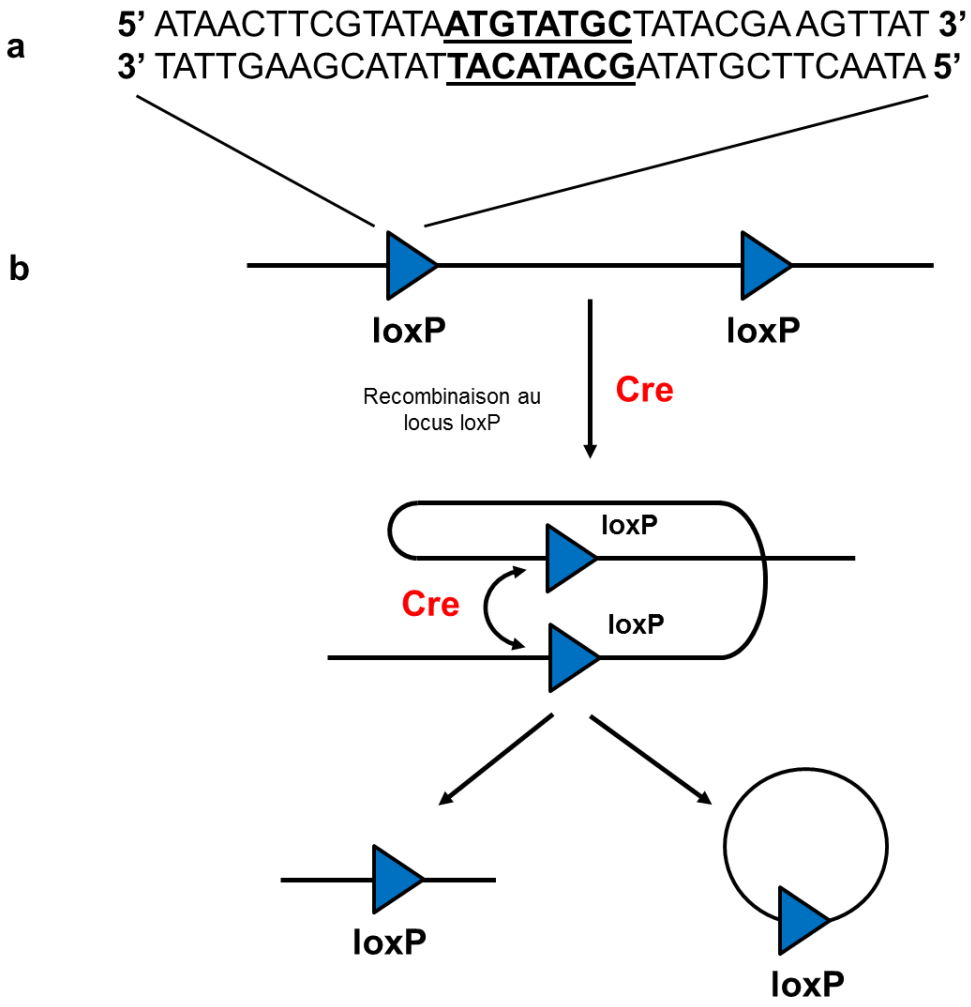


Figure 15 : Représentation schématique du système de recombinaison homologue Cre-loxP. a) Le site loxP se compose d'une séquence de 34 paires de bases constituée d'une séquence centrale de 8 paires de bases (soulignée), encadrée par deux séquences répétées de 13 paires de bases. **b)** le système Cre-loxP entraîne une recombinaison entre deux sites loxP et génère un ADN linéaire ainsi qu'un ADN circulaire qui est excisé contenant chacun site loxP

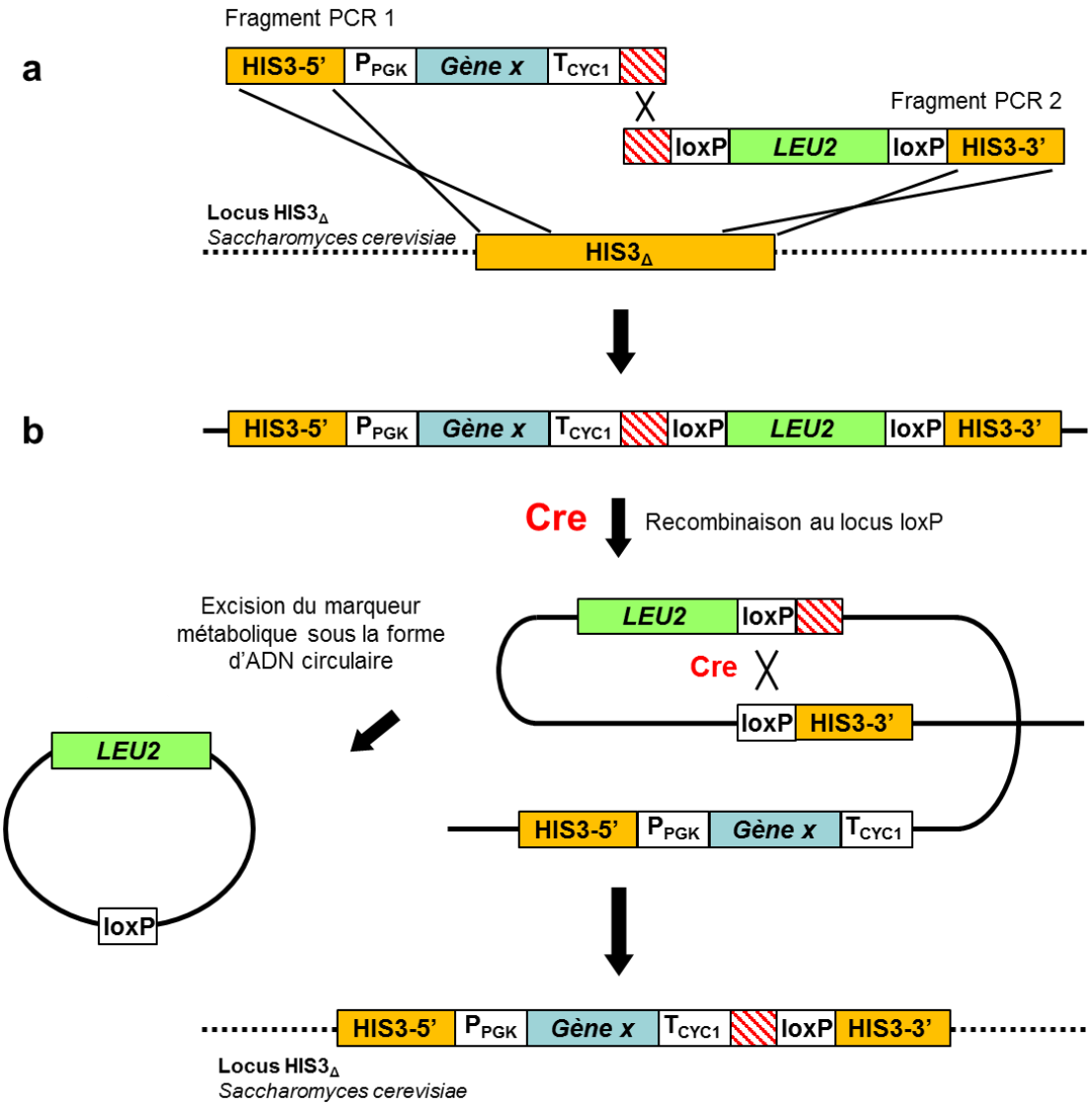


Figure 16 : Modèle d'intégration de gènes appartenant à des voies métaboliques hétérologues dans le génome de *Saccharomyces cerevisiae.* a) intégration du gène X avec un marqueur métabolique de levure au niveau du locus $HIS3_{\Delta}$ dans le génome de *S. cerevisiae.* Deux fragments PCR sont amplifiés à partir d'un plasmide pXP préalablement construit. Ces deux fragments PCR 1 et 2 comportent respectivement le gène X encadré par un promoteur et un terminateur de levure P_{PGK} et T_{CYC1}, ainsi qu'un marqueur métabolique LEU2 utilisé pour identifier les levures transformées. La recombinaison au locus $HIS3_{\Delta}$ est amorcée en positionnant sur l'extrémité 5' et 3' des fragments PCR 1 et 2 des séquences homologues correspondant aux extrémités 5' et 3' du marqueur métabolique HIS3. De la même façon une séquence chevauchante est utilisée pour provoquer une recombinaison entre les deux brins PCR. **b)** l'intégration des deux brins PCR au locus $HIS3_{\Delta}$ est suivi d'une recombinaison entre les deux sites loxP grâce à la recombinase virale Cre. Le marqueur

métabolique LEU2 est recyclé et excisé sous forme d'ADN circulaire, le gène X reste intégré au locus $HIS3_{\Delta}$.

IV.2.3 Les médicaments biosynthétiques issus de la biologie de synthèse

Depuis les années 1970 le génie génétique a fourni des ressources considérables permettant un essor des biotechnologies de l'ADN pour exprimer des gènes d'origines différentes dans une cellule hôte. Toutefois la grande majorité des bio-médicaments produits par ce type d'approche sont des protéines recombinantes. Des approches plus récentes d'ingénierie métabolique, s'inspirant de ce concept, s'intéressent également à la bio-production de molécules d'intérêt pharmaceutique en transférant des voies de biosynthèses dans des organismes hétérologues. Les composés produits par ces approches sont des petites molécules dont l'extraction à partir de l'organisme source (qui produit naturellement la molécule) ou la synthèse chimique totale s'avèrent plus coûteuse et parfois plus complexe que la biosynthèse par un organisme génétiquement modifié.

Parmi ces petites molécules, les terpènes et dérivés terpéniques sont très étudiés car leurs voies de biosynthèse sont relativement bien connues chez l'ensemble des êtres vivants. L'isoprène (C_5H_8) est le précurseur commun pour la synthèse de ces molécules et peut être assemblé de façons différentes selon les voies métaboliques. On leur reconnaît plusieurs fonctions biologiques essentielles: hormones (gibbérellines, stéroïdes), pigments photosynthétiques (phytol, caroténoïdes), phéromones végétales, stabilisateurs des membranes cellulaires (cholestérol, phytostérol)...Par ailleurs, les terpènes représentent la plus grande classe de métabolites secondaires dans la nature. Bon nombre de ces métabolites d'origine végétale ou animale présentent des propriétés pharmacologiques couvrant plusieurs domaines thérapeutiques majeurs tel que l'oncologie (paclitaxel, eleuthorbin, limonène, squalamine) (Elson et Yu,1994), l'inflammation (pseudopterosin, 1-8-cineole, linalool) (Look et *al*., 1986) ou encore l'infectiologie (Balint, 2001 ; Friedman et *al*., 2004). Du fait de leur intérêt en santé humaine et des progrès récents réalisés sur la connaissance de leur biosynthèse, ces métabolites se retrouvent au centre des recherches en ingénierie métabolique visant à les produire de façon biosynthétique. Deux composés se retrouvent particulièrement en première ligne. Il s'agit de l'hydrocortisone et de l'artémisinine.

L'hydrocortisone ou cortisol est une hormone stéroïdienne appartenant à la famille des glucocorticoïdes produite par les glandes surrénales. Cette hormone est impliquée dans le métabolisme glucido-lipidique et dans les processus prévenant l'inflammation. Cette dernière est un précurseur essentiel pour la synthèse d'autres hormones stéroïdiennes, et peut être entièrement synthétisée chimiquement en 40 étapes (Woodward et *al.*, 1952) . Elle est actuellement obtenue par hémi-synthèse incluant une étape de bio-conversion microbienne. Face à une augmentation des demandes, l'industrie pharmaceutique s'est mise à rechercher de nouvelles méthodes de synthèse moins coûteuses, aboutissant en 2003 sur la production de cortisol par voie biotechnologique chez *Saccharomyces cerevisiae* (Szczebara et *al*., 2003). Il a ainsi été créé une voie métabolique « synthétique » du cortisol chez *Saccharomyces cerevisiae*, en ajoutant au génome de la levure une quinzaine de gènes provenant d'organismes différents (boeufs, végétaux, humains) et en modulant l'expression de certains gènes endogènes afin de favoriser l'expression de ceux impliqués dans la biosynthèse de cortisol. En complément, un travail d'ingénierie génique (nombre de copies, place des introns, séquences régulatrices...) a été effectué afin d'optimiser les étapes de la biosynthèse. La chaîne de respiration mitochondriale a même été optimisée pour que la levure dispose d'assez d'énergie pour sa nouvelle fonction « d'usine cellulaire ». Une optimisation des processus de synthèse de l'hydrocortisone chez *Saccharomyces cerevisiae* a permis de produire des quantités remarquables d'hydrocortisone (10 mg/L) à partir de glucose ou d'éthanol, avec très peu de sous-produits polluants. Le passage à l'étape industrielle aura lieu très prochainement, le groupe pharmaceutique Sanofi ayant annoncé vouloir investir 150 millions d'euros pour produire en France plus d'une centaine de tonnes par biologie de synthèse.

L'artémisinine est un composant naturel extrait de l'armoise annuelle (Artemisia annua), une plante native de Chine et connue de longue date dans la pharmacopée chinoise sous le nom de Qinghao. L'artémisinine est utilisée dans le traitement des infections du paludisme et s'avère être efficace sur tous les Plasmodium et soigne même des formes neurologiques graves de paludisme (Tu, 2011). Après la découverte de l'artémisinine, plusieurs dérivés hémi-synthétiques (produits chimiquement à partir de l'artémisinine naturelle extraite d'Artemisia annua) ont été mis au point afin d'améliorer notamment le profil de solubilité et l'efficacité thérapeutique (Meshick et *al*., 1996). Les plus connus sont la dihydroartémisinine, l'artémether et l'artésunate. Selon l'OMS, les associations thérapeutiques à base d'artémisinine représentent les meilleurs traitements antipaludiques disponibles à l'heure actuelle (WHO, 2006). Seulement ils sont bien plus chers que les antipaludiques

classiques extraits ; de plus la production d'artémisinine est largement dépendante du cycle de croissance de la plante, des conditions climatiques durant sa croissance et de sa récolte. Par ailleurs, il existe une variabilité génétique entre les plants ayant une influence sur les quantités de métabolites secondaires biosynthétisés. Le rendement d'extraction moyen est d'environ 5 kg d'artémisinine pour 1 tonne de feuilles séchées d'Armoise correspondant à environ un hectare de culture.

L'artémisinine représente aujourd'hui le premier médicament commercialisé, produit à partir d'une souche de *S. cerevisiae* exprimant une voie hétérologue de l'acide artémisinique modifiée par ingénierie métabolique. Pour produire ce métabolique, tous les gènes de la voie endologue du mévalonate (MVA) ont été sur-exprimés, augmentant considérablement les titres de FPP (Westfall et *al*., 2012). De plus, 4 gènes appartenant à la plante *Artemisia annua* ont été surexprimés à l'aide de plasmides.

Deux ans après la première description de la biosynthèse d'acide artémisinique hétérologue dans la levure par (Ro et *al*., 2006), ont obtenu une souche produisant 2,5 grammes d'acide artémisinique (Ara) par litre soit une concentration 25 fois plus élevée que leurs prédécesseurs (Lenihan et *al*., 2008). Afin d'augmenter au maximum les concentrations en Ara les chercheurs ont adopté la souche de *S. cerevisiae* CEN.PK2, plus adaptée aux fermentations à l'échelle industrielle (Paddon et Kiesling, 2014). Les auteurs de ce dernier travail ont également remplacé l'usage du galactose par le glucose comme source de carbone pour la levure, dans le but de faire baisser les coûts de production, le glucose coûtant 100 fois moins cher que le galactose. La levure génétiquement optimisée par Westfall et *al*., 2012 est capable de produire 40 g/L d'amorphadiène mais contrairement aux attentes, la production d'Ara est dix fois plus faible. En effet les chercheurs se sont aperçus que la souche produisait de l'aldéhyde artémisinique (Aral), un intermédiaire d'oxydation hautement réactif et potentiellement toxique (Weathers et *al.*, 2006). La CYP7AV1 n'étant pas capable d'oxyder l'Aral jusqu'à l'Ara, des recherches génomiques ont été effectuées chez A. annua pour trouver l'enzyme catalysantl'oxydation de l'Aral, l'aldéhyde artémisinique déshydrogénase (ALDH1) (Theoh et *al*., 2009). L'expression du gène de l'ALDH1 avec celui de l'alcool artémisinique déshydrogénase(ADH1), également trouvée chez *A. annua*, donne les meilleurs résultats de production d'Ara à ce jour (Paddon et *al*., 2013). Ainsi avec les améliorations du processus de fermentation, la souche de levure intégrant une voie hétérologue de l'Ara est capable de produire la concentration cible de 25 g/L d'Ara fixée au début du projet.

Objectifs et organisation de la thèse:

Les alcaloïdes possèdent de nombreuses activités biologiques. Certains d'entre eux, comme les AIM *de C. roseus,* tels que la vincristine et la vinblastine, présentent une activité thérapeutique de première importance dans les chimiothérapies anticancéreuses. Si les procédés de production actuels, reposent encore sur la culture et l'extraction de ces composés à partir de la pervenche de Madagascar, des alternatives de production utilisant des procédés biotechnologiques n'ont pas permis à ce jour d'amorcer une production industrielle pour ces composés. L'une des raisons premières, réside dans l'extrême complexité de l'architecture cellulaire et subcellulaire de la voie de biosynthèse des AIM chez *C. roseus,* entraînant une faible production de ces composés. L'optimisation de cette production nécessite une parfaite compréhension des mécanismes biochimiques mis en œuvre pour la biosynthèse de ces métabolites. La pervenche de Madagascar, est l'une des plantes médicinales les mieux caractérisées d'un point de vue phytochimique et a fait l'objet de nombreux travaux depuis les années 1970. En effet, nombre d'AIM ont été isolées et identifiées à partir de cette plante. C'est dans ce contexte que s'inscrit mon travail de thèse. Dans un premier temps, il a eu pour objet **d'élucider plusieurs étapes de la voie de biosynthèse des AIM de *C. roseus* (Parties 1,2 et 3)**. La dernière partie (**Partie 4**) relate la **reconstitution d'un segment de cette voie métabolique dans la levure *Saccharomyces cerevisiae*, en vue de produire un alcaloïde d'intérêt**.

La **Partie 1**, expose les méthodes pour identifier les nouvelles étapes enzymatiques de la voie des AIM. Ces méthodes s'appuient (1) sur des analyses des données transcriptomiques de *C. roseus*, pour identifier des gènes candidats codant les enzymes manquantes du métabolisme des AIM ainsi que (2) sur des techniques de caractérisation fonctionnelles de gènes et de la localisation subcellulaire de leurs produits. Ces résultats ont fait l'objet d'une publication dans la revue «Methods in Enzymology».

Les deux parties suivantes (Parties 2 et 3) ont été consacrées à l'utilisation des méthodes d'investigations décrites dans la partie 1 pour identifier de nouvelles enzymes de la voie de biosynthèse des AIM chez *C. roseus.*

La **Partie 2** décrit, d'une part, la caractérisation d'une seconde isoforme du cytochrome P450 sécologanine synthase (SLS). Cette enzyme est impliquée dans le métabolisme des sécoiridoïde (figure 11 de l'introduction) et conduit en particulier à la production de la sécologanine. D'autre part, cette partie expose également les travaux sur la caractérisation des deux cytochromes P450 réductases (CPR) auxquels sont associés les nombreux cytochromes P450. Les travaux de caractérisation de la seconde isoforme de SLS ont été publiés dans la revue «BMC Genomics », tandis que les travaux sur les CPR sont présentés sous forme d'un article en cours de finalisation.

La **Partie 3** rapporte l'identification et la caractérisation de diverses déshydrogénases/réductases impliquées dans la biosynthèse des alcaloïdes de type hétéroyohimbine. Y est exposé la caractérisation de la première enzyme de ce type à avoir été identifiée, à savoir la tétrahydroalstonine synthase (THAS), impliquée dans la formation de la tétrahydroalstonine. Puis, faisant suite à ces travaux, trois autres enzymes de ce type ont également été caractérisées, deux isoformes de THAS ainsi qu'une hétéroyohimbine synthase (HYS). Les travaux sur THAS ont été publiés dans la revue « Chemistry and Biology » et les travaux sur les isoformes de THAS et HYS sont présentés sous forme d'article en cours de finalisation.

Enfin, la **Partie 4** conclue les travaux d'élucidation de la voie de biosynthèse des AIM en caractérisant la tabersonine 3-réductase (T3R), une enzyme des étapes finales de la voie métabolique impliquée dans la formation de la vindoline. L'insertion du gène T3R ainsi que 6 autres gènes dans la levure a été réalisée en vue d'étudier et d'optimiser la production de la vindoline par ingénierie métabolique. Ces travaux sont présentés sous forme d'un article en cours de finalisation.

Résultats

Résultats :

Partie 1: Méthode utilisées pour l'identification et la validation de gènes du métabolisme alcaloïdique

Article 1: Prequels to synthetic biology: from candidate gene identification and validation to enzyme subcellular localization in plant and yeast cells

Ce premier article fait office de "Matériel et Méthodes" du présent manuscrit de thèse. Il présente les principales méthodes mises en œuvre pour élucider les étapes enzymatiques de voies métaboliques partiellement caractérisées en vue d'exploiter ces connaissances dans les approches d'ingénierie métabolique. Il s'appuie sur des méthodes d'investigation portant sur le métabolisme des AIM de C. roseus. Les approches présentées sont transférables à l'étude de diverses voies métaboliques d'autres espèces végétales.

L'article se présente comme un guide de méthodes allant de la recherche de gènes candidats à la validation de leurs produits (enzymes) et leur localisation subcellulaire. Dans la première partie, sont décrites les approches bio-informatiques basées sur l'exploitation des données transcriptomiques de *C. roseus*. L'hypothèse de départ sur laquelle sont basée les analyses transcriptomiques est la suivante : des gènes codant pour des enzymes impliquées dans un même processus physiologique (en l'occurrence ici la biosynthèse des AIM) sont co-exprimés, cela signifie que leurs taux de transcrits (profils d'expression) varient de façon similaire selon les différentes conditions physiologiques. Ainsi, étudier les profils d'expression de l'ensemble des gènes d'une plante peut s'avérer très utile pour associer des fonctions opérant au sein d'une même voie métabolique. En pratique, la comparaison des profils d'expression de l'ensemble des gènes de *C. roseus* (issus des données RNA-seq) obtenus à partir des différentes conditions expérimentales, avec les profils d'expression de gènes connus codant des enzymes caractérisées de la voie métabolique des AIM (ces gènes sont utilisés comme « appât ») permet de sélectionner un panel de gènes candidats potentiellement impliqués dans les étapes enzymatiques inconnues de la voie de biosynthèse des AIM.

Dans la seconde partie de l'article, est détaillée une technique de validation fonctionnelle des gènes candidats identifiés. Le principe repose sur l'extinction du gène candidat et des conséquences de cette extinction sur la biosynthèse des AIM. Ainsi, si l'on observe une diminution de certains AIM lors de l'extinction d'un gène, on pourra supposer que ce gène code une enzyme de la voie de biosynthèse des AIM et on pourra, dès lors, entreprendre des études biochimiques de caractérisation de l'enzyme.

L'extinction du gène s'effectue par méthode d'invalidations induites par un virus, le VIGS (Virus-Induced Gene Silencing). Ce type d'approche a d'abord concerné certaines espèces de *Solanaceae*. Elle a eu recours au virus de la mosaïque du tabac (TMV), au virus de la pomme de terre, et au virus du tabac hochet (TRV) (Burch-Smith et al., 2004 ; Lu et al., 2003 ; Senthil-Kumar et Mysore, 2011). Ce dernier peut désormais être utilisé chez des plantes qui ne sont pas directement les hôtes définitifs du virus comme *Papaver somniferum* et plus récemment chez *C. roseus* (Drea et al., 2007; Hileman et al., 2005; Liu et al., 2002 ; Wijekoon et Facchini, 2011). La méthode VIGS est une technologie qui exploite un mécanisme naturel de défense des plantes contre les virus, à savoir la dégradation des ARNm viraux par les complexes Dicer puis Risc. Chez les plantes infectées par des virus non modifiés, le mécanisme est spécifiquement ciblé sur le génome viral. Cependant, avec des vecteurs viraux portant des fragments d'ADN provenant de gènes de la plante hôte (comme certains gènes de la voie de biosynthèse des AIM chez *C. roseus*), le processus peut être dirigé contre les ARNm correspondants (Burch-Smith et al., 2004 ; Lu et al., 2003 ; Senthil-Kumar et Mysore, 2011) Par ailleurs, la nouveauté dans cette méthode réside dans le fait que l'inoculation du virus TRV modifié dans les plantules, s'effectue avec une méthode de transformation par biolistique alors que la méthode habituelle utilise la transformation *via* la bactérie *Agrobacterium tumefaciens*.

Pour valider la méthode, le choix s'est porté sur l'extinction du gène *pds* codant la phytoène synthase impliquée dans la biosynthèse des caroténoïdes. L'extinction de ce gène provoque un blanchiment des zones foliaires où le virus s'est propagé (figure 17).

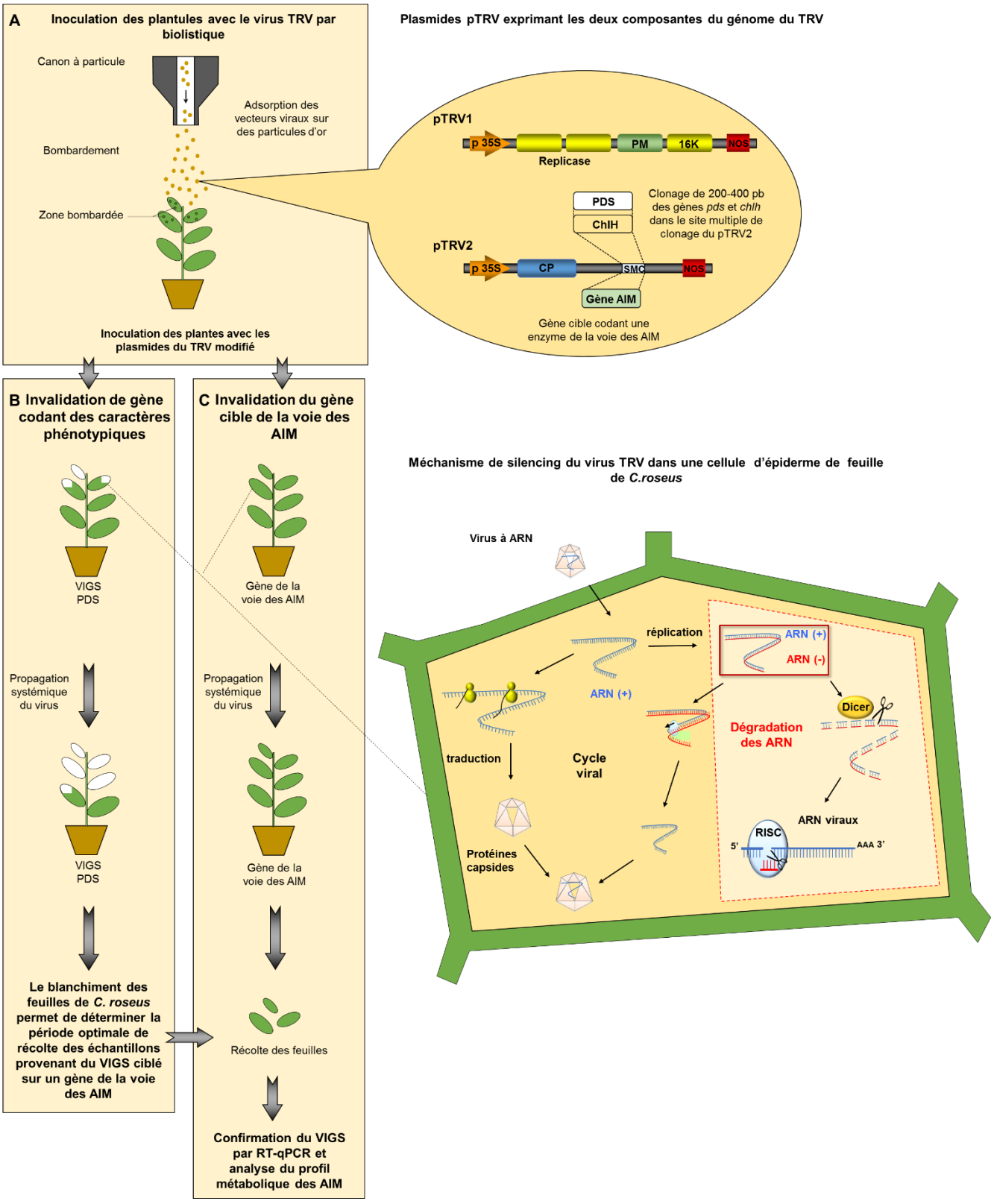


Figure 17 : Schéma du protocole de la méthode VIGS utilisée pour la validation de gènes candidats de la voie de biosynthèse des AIM chez *C. roseus*. (A) Inoculation du TRV modifié par biolistique. Le TRV possède un génome constitué de deux segments d'ARN, ARN1 et ARN2. Les ADNc correspondant à l'ARN1 et l'ARN2 ont été clonés dans des

vecteurs d'expression pour donner les vecteurs pTRV1 et pTRV2. Le pTRV1 possède deux gènes codant des réplicases virales, un gène codant des protéines de mouvement (PM) ainsi qu'une protéine riche en cystéine dont la fonction n'est pas encore connue. Le pTRV2 présente un gène codant pour l'enveloppe du virus (CP) ainsi qu'un site multiple de clonage dans lequel sont clonés les gènes dont on souhaite éteindre l'expression. (B) Le suivi de la cinétique d'apparition du blanchiment des feuilles de *C. roseus* lors de l'extinction du gène *pds*, permet de déterminer la période optimale pour récolter et analyser les échantillons issus du VIGS sur des gènes cibles de la voie des AIM (C).

La troisième partie de l'article traite de la compartimentation subcellulaire des enzymes du métabolisme des AIM. La plupart des voies de biosynthèse de métabolites spécialisés des plantes; possède une compartimentation intracellulaire complexe avec des enzymes adressées dans de multiples compartiments subcellulaires distincts, incluant des transports de métabolites entre ces compartiments. Chez *C. roseus*, six compartiments subcellulaires hébergent la quarantaine d'étapes enzymatiques (en y incluant les étapes non caractérisées au niveau enzymatique) de la voie de biosynthèse des AIM (Courdavault et *al*., 2014). Une telle complexité peut engendrer diverses complications lorsque l'on tente de reconstituer la chaîne métabolique dans les organismes hétérologues, d'autant plus que certains systèmes d'adressage de protéines peuvent être différents d'un organisme à un autre (Geerlings et *al*., 2001). Aussi, s'assurer du bon adressage des protéines dans des organismes hétérologues est donc un prérequis incontournable dans les stratégies d'ingénierie métabolique.

Dans l'article, ces études sont illustrées par les localisations subcellulaires de STR et de SGD chez *C. roseus* et *S. cerevisiae*. Il s'agit de deux enzymes du métabolisme alcaloïdique, catalysant deux réactions enzymatiques successives dans deux compartiments subcellulaires distincts, la vacuole et le noyau (Guirimand et *al*., 2010). Dans ces travaux, nous avons confirmé la localisation vacuolaire de STR chez *C. roseus* et *S. cerevisiae*. En revanche, SGD montre un double adressage au niveau de la vacuole et du noyau chez *S. cerevisiae* alors qu'elle est uniquement nucléaire chez *C. roseus*. De telles différences d'adressage sont à prendre en compte lors de la reconstitution de la voie de biosynthèse AIM dans les levures.

Prequels to Synthetic Biology: From Candidate Gene Identification and Validation to Enzyme Subcellular Localization in Plant and Yeast Cells

E. Foureau*,[1], I. Carqueijeiro*,[1], T. Dugé de Bernonville*,[1], C. Melin*, F. Lafontaine*, S. Besseau*, A. Lanoue*, N. Papon†, A. Oudin*, G. Glévarec*, M. Clastre*, B. St-Pierre*, N. Giglioli-Guivarc'h*, V. Courdavault*,[2]

*Université François-Rabelais de Tours, EA2106 "Biomolécules et Biotechnologies Végétales", Tours, France
†Université d'Angers, Groupe d'Etude des Interactions Hôte-Pathogène, UPRES EA 3142, Angers, France
[2]Corresponding author: e-mail address: vincent.courdavault@univ-tours.fr

Contents

[1] Equal contribution.

Methods in Enzymology
ISSN 0076-6879
http://dx.doi.org/10.1016/bs.mie.2016.02.013

Abstract

Natural compounds extracted from microorganisms or plants constitute an inexhaustible source of valuable molecules whose supply can be potentially challenged by limitations in biological sourcing. The recent progress in synthetic biology combined to the increasing access to extensive transcriptomics and genomics data now provide new alternatives to produce these molecules by transferring their whole biosynthetic pathway in heterologous production platforms such as yeasts or bacteria. While the generation of high titer producing strains remains per se an arduous field of investigation, elucidation of the biosynthetic pathways as well as characterization of their complex subcellular organization are essential prequels to the efficient development of such bioengineering approaches. Using examples from plants and yeasts as a framework, we describe potent methods to rationalize the study of partially characterized pathways, including the basics of computational applications to identify candidate genes in transcriptomics data and the validation of their function by an improved procedure of virus-induced gene silencing mediated by direct DNA transfer to get around possible resistance to Agrobacterium-delivery of viral vectors. To identify potential alterations of biosynthetic fluxes resulting from enzyme mislocalizations in reconstituted pathways, we also detail protocols aiming at characterizing subcellular localizations of protein in plant cells by expression of fluorescent protein fusions through biolistic-mediated transient transformation, and localization of transferred enzymes in yeast using similar fluorescence procedures. Albeit initially developed for the Madagascar periwinkle, these methods may be applied to other plant species or organisms in order to establish synthetic biology platform.

1. INTRODUCTION

For ages, humans have exploited natural compounds, notably those arising from plant specialized metabolisms, as dyes, herbicides, flavors, and scents, or as bioenergy sources, but above of all by taking advantage of their pharmacological properties (Hanson, 2003; Ragauskas et al., 2006). These broad biological activities resulted in the valorization of specialized metabolites in a variety of industries and pharmaceutical applications rendering these molecules inescapable to our life habits. Since the beginning of the 20th century and the progress of synthetic chemistry, chemists began to reproduce these natural products alike both to improve their supply and to get around the drawback of biological sourcing, and subsequently to include rational modifications to generate novel and more potent drugs. However, in spite of decades of efforts and the development of combinatorial chemistry or computer molecular modeling, this task remains challenging for numerous compounds and/or economically unviable given the high complexity of

their structures. As a consequence, it is admitted that more than half of the approved drugs used over the past 30 years are still directly or indirectly extracted from natural sources (Newman & Cragg, 2012). This inextinguishable list of bioactive molecules includes for instance, the prominent quinine, extracted from *Cinchona* tree, which is still one of the major drug used against malaria in combination with artemisinin from *Artemisia annua* (Pollier, Moses, & Goossens, 2011); ajmaline and ajmalicine used, respectively, as antiarrhythmic and in the treatment of hypertension following extraction *Rauwolfia serpentina* (Drewes, George, & Khan, 2003; Pollier et al., 2011); camptothecin and its derivatives topotecan hydrochloride and irinotecan hydrochloride used in chemotherapy and produced from *Camptotheca acuminata* (Thomas, Rahier, & Hecht, 2004); sanguinarine (*Sanguinarea canadensis*)—as antibacterial; or cocaine (*Erythroxylum coca*) as anesthetic (Mano, 2006; Wink, 1999; Zhao & Dixon, 2009); conolidine from *Tabernaemontana* ssp., which was recently proved to have analgesic properties as powerful as the opiates (Tarselli et al., 2011); berberine from *Coptis* with antibiotic and antiinflammatory activities, and more recently shown to have anticancer and antidiabetic activities (Li et al., 2015; Stermitz, Lorenz, Tawara, Zenewicz, & Lewis, 2000); taxol isolated from *Taxus brevifolia* and widely used in chemotherapy cocktails, which was in difficult supply until alternative sources like suspension cell cultures were developed (Miller & Ojima, 2001); ephedrine from *Ephedra sinica* to treat asthma (Lee, 2011) and the antineoplastics vinblastine and vincristine from *Catharanthus roseus* (Madagascar periwinkle) that are still produced via extraction from the periwinkle leaves.

The complexity of the biosynthetic pathways of the specialized metabolites as well as their high degrees of organization in planta has prevented for years, the use of simplified models, such as dedifferentiated plant cell cultures in vitro, to produce these valuable compounds through biotechnological approaches. However, the recent progress in synthetic biology and the dramatic reduction in cost for development of genomic and transcriptomic datasets are now changing this hostile scenario and open new ways to produce these valuable compounds via metabolic engineering of microbial platforms. Over the last 10 years, numerous examples of this type of production in heterologous organisms have been detailed via partial and/or complete biosynthetic pathway reconstitutions through multiple plant gene transfers, starting with the synthesis of artemisinic acid, the artemisinin precursor, in both *Saccharomyces cerevisiae* and *Escherichia coli* (Chang, Eachus, Trieu, Ro, & Keasling, 2007; Paddon et al., 2013; Ro et al., 2006), the production of

benzylisoquinoline alkaloids in yeast transformed with enzymes from three plant sources and humans (Cassels et al., 1995; Hawkins & Smolke, 2008) and the recent synthesis in yeast of the precursor of all monoterpene indole alkaloids (MIAs), strictosidine, achieved by the transfer of 14 MIA biosynthetic genes from *C. roseus* along with seven additional genes and three gene deletions (Brown, Clastre, Courdavault, & O'Connor, 2015).

The development of such strategies relies on a complete characterization of the biosynthetic pathways of specialized metabolites as a commitment for the effective transfer in heterologous organisms of the whole set of genes constituting these pathways. Except for a few number of major drugs, the knowledge of biosynthetic pathways still remains scarce as compared to their immense diversity and complexity. From this point of view, accurate exploitations of transcriptomic data now have the potential to rationalize the identification of candidate genes for the missing steps of a pathway, notably for enzyme coded by large multigene family. Albeit a consequent bench work is still required to validate candidate gene function, such type of approaches has permitted huge advances in deciphering plant specialized metabolisms as recently illustrated for MIAs (Dugé de Bernonville, Foureau, et al., 2015). Furthermore, the numerical complexity of pathways is frequently enriched by their complex organization in planta at both the cellular and subcellular levels. For instance, in *C. roseus*, no less than four distinct leaf cell types and more than five subcellular compartments host the almost 30–40 enzymatic steps required to synthesize MIAs (Courdavault et al., 2014). While the cellular organization of a pathway can be easily skirted for metabolic engineering approaches by coexpressing all genes in a single heterologous organism, much attention should be paid to subcellular organizations since the inherent metabolite translocations between organelles could impact metabolic fluxes and hampered the production of the desired compounds. This implies a rigorous characterization of the biosynthetic pathway subcellular architecture to foreshadow and remedy future potential bottlenecks in metabolite synthesis postgene transfers. In addition, several biases of localizations could arise when expressing plant proteins in yeast or bacteria necessitating the validation of each protein localization in the recipient organism.

With the intent to streamline/rationalize the production of valuable compounds through synthetic biology approaches, we describe hereafter a guideline of the main technical procedures that should be used to generate some prerequisites to the transfer of a biosynthetic pathway in heterologous organisms (Fig. 1). By using, as a framework, examples from the Madagascar periwinkle, we successively present (1) the basics of computational

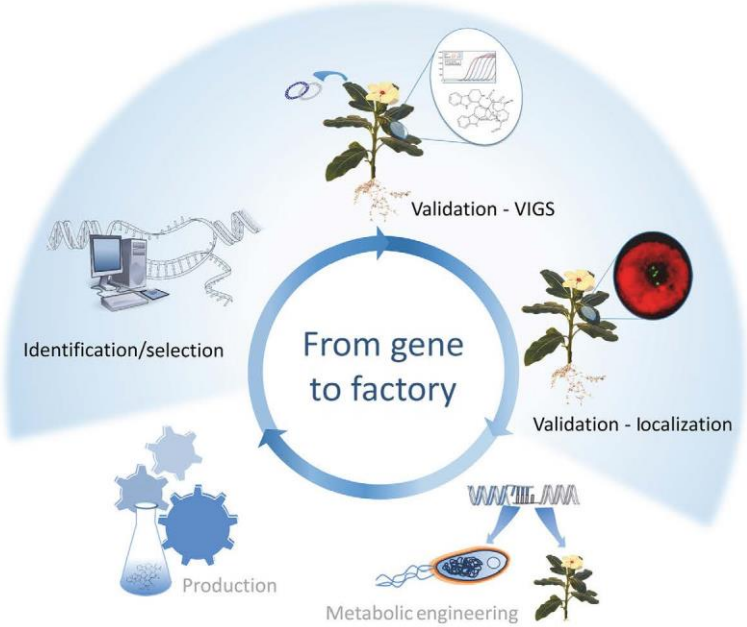


Fig. 1 Prequels to synthetic biology: from candidate gene identification and validation to their characterization and transfer into heterologous organisms.

applications to valorize transcriptomic data and to identify candidate genes, (2) a procedure of virus-induced gene silencing (VIGS) mediated by direct DNA transfer to get around resistance to *Agrobacterium*-delivery of viral vectors, (3) a procedure of protein subcellular localization study via particle bombardment and fluorescent protein (FP) imaging in planta, and (4) the validation of protein localization in heterologous organisms such as yeast.

2. IDENTIFICATION OF CANDIDATE GENES THROUGH TRANSCRIPTOMIC DATA MINING AND ANALYSIS

Identification of all the genes of a specific biosynthetic pathway is the first prerequisite to its transfer in heterologous organisms. However, this characterization remains partial for many valuable compounds and implies prediction and validation of candidate genes. Basically, it is admitted that genes of a similar biosynthetic pathway are potentially subjected to coregulation, at least in distinct subparts of this pathway. Such coregulation can thus lead to the coexpression of all these genes in a wide set of conditions, which can be uncovered through analysis of transcriptomic data.

By consequence, the comparison of expression profiles between previously identified genes of the investigated pathway, used as baits, and the transcriptome allows the selection of a restricted subset of candidate genes displaying the highest level of expression profile similarity. The efficiency of this type of analysis strongly depends on the availability of a large set of experimental conditions, which display various and specific levels of biosynthetic pathway gene expression. Furthermore, other/previous RNA-seq runs can be included in these analyses via straightforward procedures of data reuse. The following sections will describe the basics of transcriptomic data processing from assembly and transcript abundance to gene expression correlation analysis and hierarchical clustering by using examples based on a consensus transcriptome built for *C. roseus* with previous RNA-seq data obtained in distinct plant organs and experimental conditions (Dugé de Bernonville, Clastre, et al., 2015).

2.1 Transcriptome Assembly, Annotation, and Transcript Abundance Estimation

2.1.1 Transcriptome Assembly

Short-read sequencing of transcriptomes generates raw FASTq files containing a sequence for each spot on the lane together with base quality indices. Careful examination of quality and trimming of raw reads constitutes essential prerequisites before assembling them into larger contigs. Several assembly methods and protocols allowing de novo assembly of transcripts from short read paired-end sequencing have already been reported (Góngora-Castillo et al., 2012; Haas et al., 2013). De novo assemblers are designed to resolve de Bruijn graphs built with raw reads splitted into shorter words of length *k*, named *k*-mers (Martin & Wang, 2011). Strategies to prepare a transcriptome include single assembly from combined reads, combination of assemblies prepared with different *k*-mer length using a same set of reads (Velvet/Oases, Schulz, Zerbino, Vingron, & Birney, 2012), or combination of single assemblies obtained from individual samples. In the last two strategies, redundancy may be reduced by using assemblers such as TGICL (Pertea et al., 2003) or CD-HIT-EST (Huang, Niu, Gao, Fu, & Li, 2010). Assessment of the quality of transcript reconstruction has to be performed to ensure that parameters used in the assembly process are appropriate. This can be checked by (i) analyzing transcript lengths, (ii) annotating transcripts, and (iii) comparing transcripts to known EST or full-length mRNA belonging to the same studied species or closely related ones.

Genome-guided assembly is also possible, albeit plant genomic sequences are still scarce. In addition, de novo assembly may be preferred in the case of low quality or incomplete genomic sequences and to ease identification of alternatively spliced transcripts. Genome-guided assembly starts by mapping reads to the genomic sequences. De Bruijn graphs are then applied to resolve transcript structures. Several programs are designed for that purpose, eg, Cufflinks/TopHat (Trapnell et al., 2012), Trinity (Haas et al., 2013), and StringTie (Pertea et al., 2015).

2.1.2 Transcriptome Annotation

Annotation of the transcriptome assembly is a fundamental step which should (i) improve transcriptome quality (see above) and (ii) help for post-assembly analyses. The most common procedures include the research of homologies against a database by BLAST (either blastx on mRNA sequence or blastp on predicted peptide sequences) and detection of functional domains with HMMER (hmmerscan). The free software Trinotate (https://trinotate.github.io/) offers an easy interface to incorporate different annotation layers in a transcriptome. We also recommend the use of perl script HpcRunningGrid collection written by B. Haas to parallelize BLAST and HMMERscan and significantly improve annotation speed (http://hpcgridrunner.github.io/). The following protocol describes an annotation process performed with Trinotate.

1. Download Uniprot database for local homology search and PfamA: follow instructions found on the Trinotate webpage in section "2.Sequence Databases Required"

2. Prepare a script for transcriptome annotation: transdecoder, blastp, blastx and hmmerscan

Don't forget to configure the hpc_running_grid file to speed up analysis.

```
#!/bin/bash

#SBATCH J transdec1
#SBATCH o transdec1.%j
#SBATCH e transdec1.%j
#SBATCH n 1
#SBATCH t 24:00:00
#SBATCH p defq
```

```
module purge
module load shared
module load gcc/4.9.0
module load slurm/14.03.0

export PATH=$BLAST+DIR/bin:$PATH

#predict ORFs
$HOME/TransDecoder_2.0/TransDecoder.LongOrfs -t CDF97.fa

#run blastp
mkdir /scratch/blastp_CDF97
cp CDF97.fa.transdecoder_dir/longest_orfs.pep /scratch/blastp_CDF97
cp uniprot* /scratch/blastp_CDF97
cd /scratch/blastp_CDF97

#parallelize search (configure SLURM.test.conf accordingly)
perl $HOME/HpcGridRunner-1.0.0/BioIfx/hpc_FASTA_GridRunner.pl
 --cmd_template "blastp -query __QUERY_FILE__ -db uniprot_sprot.
trinotate_v2.0.pep -max_target_seqs 1 -outfmt 6 -evalue 1e-5
 -num_threads 2" --query_fasta longest_orfs.pep -G $HOME/SLURM.test.
conf -N 1000 -O CDF97_blastp

#group results and move them to $HOME directory
find CDF97_blastp/ -name "*.fa.OUT" -exec cat {} \; > CDF97.blastp.out
mv CDF97.blastp.out $HOME
cd $HOME
rm -fr /scratch/blastp_rau_CDF97_v2

#run blastx
mkdir /scratch/blastx_CDF97
cp CDF97.fa /scratch/blastx_CDF97
cp uniprot* /scratch/blastx_CDF97
cd /scratch/blastx_CDF97

#parallelize search (configure SLURM.test.conf accordingly)
perl $HOME/HpcGridRunner-1.0.0/BioIfx/hpc_FASTA_GridRunner.pl
 --cmd_template "blastx -query __QUERY_FILE__ -db uniprot_sprot.
trinotate_v2.0.pep -max_target_seqs 1 -outfmt 6 -evalue 1e-20
 -num_threads 2" --query_fasta longest_orfs.pep -G $HOME/SLURM.test.
conf -N 1000 -O CDF97_blastx
```

```
#group results and move them to $HOME directory
find CDF97_blastx/ -name "*.fa.OUT" -exec cat {} \; > CDF97.blastx.out
mv CDF97.blastx.out $HOME
cd $HOME
rm -fr /scratch/blastx_rau_CDF97_v2

#run hmmerscan
mkdir /scratch/CDF97
cp $HOME/CDF97.fa.transdecoder_dir/longest_orfs.pep /scratch/hmmer_CDF97
cp Pfam-A.* /scratch/hmmer_CDF97
cd /scratch/hmmer_CDF97

perl $HOME/HpcGridRunner-1.0.0/BioIfx/hpc_FASTA_GridRunner.pl
--cmd_template "hmmscan --cpu 2 --domtblout __QUERY_FILE__.domtblout
Pfam-A.hmm __QUERY_FILE__" --query_fasta longest_orfs.pep -G $HOME/
SLURM.test.conf -N 1000 -O hmmer_CDF97

find hmmer_CDF97/ -name "*.fa.domtblout" -exec cat {} \; > CDF97.
hmmer.out

mv CDF97.hmmer.out $HOME
cd $HOME
rm -fr /scratch/hmmer_CDF97

#end

3. run trinotate and output to xls file: initialize MySQL Database and
load annotation results in it by following Trinotate basics instructions.
```

2.1.3 Transcript Abundance Estimation

The measurement of gene expression evaluated by transcript abundance estimation is essential for coexpression analysis in order to identify new genes associated within a same biosynthetic pathway. In addition, measuring transcript abundance is also an important complementary step of the assembly process. Indeed, setting appropriate abundance thresholds may significantly improve transcriptome quality by removing nonexpressed chimeric transcripts. Several algorithms are available to estimate transcript abundance. Cufflinks (Trapnell et al., 2010), eXpress (Roberts, Feng, & Pachter, 2013), and RSEM (Li & Dewey, 2011) are among the most commonly used.

The starting point of abundance estimation is the mapping of raw reads to the reference assembly by Burrows-Wheeler schemes like Bowtie (Langmead, Trapnell, Pop, & Salzberg, 2009), Bowtie2 (Langmead & Salzberg, 2012), or BWA algorithms (Li & Durbin, 2009). Expression levels are next estimated after determining the most probable contig origin in case of multiple read mapping, in particular by applying expectation-minimization algorithms.

In the next sections, we will describe how to use RSEM to estimate transcript abundance in a transcriptome (RSEM v1.2.15 procedure for CDF97 with paired-end reads according to Dugé de Bernonville, Clastre, et al., 2015). Readers are invited to refer to original articles for more explanations about the different algorithms presented here.

The following procedures are based on a transcriptome resulting from a clustering with CD-HIT-EST of single Trinity assemblies obtained for each RNA-seq run available on NCBI Sequence Read Archive (SRA) for *C. roseus*. To take into account polymorphisms linked to the diversity of samples, CD-HIT-EST representative cluster sequences were used as "gene" sequences and the remaining contigs in the clusters as "transcripts" corresponding to these genes. This allowed attributing correct expression levels to "genes" given the more exact mapping of reads on each "transcript." In the following example, publicly available data were reused. SRA SRR accessions were downloaded by ftp (access with wget, for example) and the resulting .sra files convert to fastq with the SRA Toolkit (./fastq-dump -I –split-files SRRxxxxx.sra).

TIP: Transcriptome assembly and read assignment to transcripts intensively consume computational resources that cannot be efficiently managed on a single computer. The optimal situation is to use a computing grid with several nodes containing several CPUs with memory. For the following parts, computations were performed on the CCSC Computing Grid of Orléans, France. Nodes are composed of 20 CPUs (Intel Xeon processors) with 64 Go of RAM. Jobs are scheduled with SLURM under Scientific Linux 6. In addition, computations were performed on a high-speed reading partition named "scratch" when possible.

2.1.3.1 Prepare Reference for Abundance Estimation on CDF97

The input file shall be a transcriptome containing all transcript (not only gene) sequences. For the CDF97 example, we used a multifasta file (CDF97_allcontigs.fasta) containing all contig sequences (not only representative sequences).

```
#!/bin/bash
### This is a SLURM submission file.
```

```
#SBATCH J prep_ref
#SBATCH o out_prep_ref.%j
#SBATCH e err_prep_ref.%j
#SBATCH  ntasks=1
#SBATCH  cpus per task=5
#SBATCH  tasks per node=1
#SBATCH t 96:00:00
#SBATCH p kernel3

export PATH=$BOWTIE_DIR/bowtie2 2.2.4:$PATH
$RSEM_DIR/rsem prepare reference     bowtie2     transcript to
gene map CDF97_reference_map CDF97_allcontigs.fasta CDF97_allcontigs

#copy reference to the «scratch» partition
mkdir /scratch/ref/
cp CDF97_allcontigs* /scratch/ref/

#end
```

In this script, the "--transcript-to-gene-map" corresponds to a two-column text file. The second column contains each «transcript» represented once, and the first one includes «genes» for which are found the corresponding «transcript». In the case of CDF97 example, «genes» correspond to CD-HIT-EST representative sequences and «transcripts» to other sequences found in sequence clusters. For other transcriptomes, current assemblers such as Trinity provide both genes and transcripts, as well as the corresponding map.

2.1.3.2 Align Paired-End Reads to Reference Transcriptome with bowtie2

This step is the most resource-consuming process; the processing time depends especially on the number of sequences in the transcriptome. In the following step, we used an array-like job to submit one alignment (align all reads for a given sample to the reference transcriptome) per node, using 20 CPUs on each. The following script contains instructions to copy fastq files to the scratch partition, run RSEM, and retrieve .isoforms.results and .genes.results files.

First, prepare a folder containing all fastq files (SRRxxxxx_1.fastq and SRRxxxxxx_2.fastq) and a "pe_sample" file containing accession names (SRRxxxxx). In the present example, we had 12 paired-end runs and each will be treated in a separate job.

```
#!/bin/bash

### This is a SLURM submission file.
```

```
#SBATCH J b2_alignreads
#SBATCH o out_alignreads.%A_%a
#SBATCH e err_alignreads.%A_%a
#SBATCH  ntasks=1
#SBATCH  cpus per task=20
#SBATCH  tasks per node=1
#SBATCH  array=1 12
#SBATCH t 96:00:00
#SBATCH p kernel3

export PATH=$BOWTIE_DIR/bowtie2 2.2.4:$PATH

#load required modules
module purge
module load shared
module load gcc/4.9.0
module load slurm/14.03.0

echo Start time: $(date)

number=$SLURM_ARRAY_TASK_ID
paramfile=pe_samples
in=`sed  n ${number}p $paramfile | awk '{print $1}'`

if [  e /scratch/align/CDF97_allcontigs_$in ]; then
  rm  fr /scratch/align/CDF97_allcontigs_$in
fi

mkdir /scratch/align/CDF97_allcontigs_$in
cp /path/to/fastq_files/${in}_*.fastq /scratch/tduge/CDF97_allcontigs_
$in/
cd /scratch/tduge/CDF97_allcontigs_$in

$RSEM_DIR/rsem calculate expression    bowtie2   p  20    bowtie
chunkmbs 1024   paired end ${in}_1.fastq ${in}_2.fastq /scratch/ref/
CDF97_allcontigs $in

cd $HOME/CDF97_allcontigs/
cp /scratch/align/CDF97_allcontigs_$in/*.results .
rm  fr /scratch/align/CDF97_allcontigs_$in

echo End time: $(date)
```

Submit job to your computing grid. Modify if required to suit scheduler (LSF, SGE, etc.) specificities. Each array job will produce two files named

".isoforms.results" and ".genes.results" according to the gene map that has been provided to build the reference.

TIP: To analyze single-end reads, just modify the rsem-calculate-expression parameters to –single-end and indicate the corresponding single fastq file. Using a similar array task, provide a "se_sample file" containing single-end accession.

2.1.3.3 Combine RSEM Results Files to Generate Raw Count or FPKM Matrix

```
#Rscript
    library(parallel)

    table.list< list.files(pattern=".isoforms.results")

    #get raw counts
    #read tables and store count columns
    listing.tpm< mclapply(table.list, function(x){
         tmp< read.table(x, header=T)
         tmp[,6]}, mc.cores=20)

    #temporary table
    big.table< do.call("cbind", listing.tpm)
    rm(listing.tpm)

    #attribute sample names
    tmp< read.table(table.list[1], header=T)
    colnames(big.table)< gsub(".isoforms.results", "", table.list)
    rownames(big.table)< as.vector(tmp[,1])

    #write to external file
    write.table(big.table, "fpkm_table")

    #get fpkm
    listing.fpkm< mclapply(table.list, function(x){
         tmp< read.table(x, header=T)
         tmp[,7]}, mc.cores=20)
    big.table< do.call("cbind", listing.fpkm)
    rm(listing.fpkm)
    tmp< read.table(table.list[1], header=T)
    colnames(big.table)< gsub(".isoforms.results", "", table.list)
    rownames(big.table)< as.vector(tmp[,1])
    write.table(big.table, "fpkm_table")
```

TIP: Comparison of qPCR data with in silico digital expression can be useful to assess both quality of transcript reconstruction and abundance estimation.

2.2 Transcriptome Postassembly Analysis

Once assembly and abundance estimation procedures completed, the dataset can be used to identify candidate genes associated with a specific biosynthetic pathway. To this aim, the postassembly analysis may be directed to cluster genes with or without a priori. It is often appropriated to combine these strategies to identify candidate genes. The dataset is presented as a matrix where rows are genes and columns are samples. In the following section, we will describe how to apply such procedures with R and specific bioconductor packages.

```
#read the data
tpm.table< read.table("tpm_table", header=T, row.names=1)
fpkm.table< read.table("fpkm_table", header=T, row.names=1)

#load annotations
annot< read.delim("annotation/Trinotate_output.xls", header=T)
annot[,1]< gsub("|", "_", as.vector(annot[,1]), fixed=T)
annot.ok< cbind(as.vector(annot[,1]) ,do.call("c", lapply(strsplit
(as.vector(annot[,3]), ";", fixed=T), function(x)x[1])), as.vector
(annot[,10]))

transcrit.1.orf< annot[which(duplicated(annot[,2])==FALSE),]
rownames(transcrit.1.orf )< transcrit.1.orf[,1]
transcrit.1.orf< transcrit.1.orf[which(rownames(transcrit.1.
orf ) %in% rownames(fpkm.table)),]

#prepare objects to be used in GO term enrichment analysis
#the main objective is to get all represented GO terms in the
transcriptome
#and their effective (number of transcripts with a given annotation)
GO.tot< lapply(1:nrow(transcrit.1.orf ),
function(x)unlist(strsplit(as.vector(transcrit.1.orf[x,14]), "`")))
names(GO.tot)< rownames(transcrit.1.orf )
all.GO< unique(unlist(GO.tot))

#GO annotations in Trinotate xls output are combined and separated by a
"^" character
#we used this to separate individual terms
```

```
all.GO.2< do.call("rbind", strsplit(all.GO, "^", fixed=T))
rownames(all.GO.2)< all.GO.2[,1]
all.GO.2[,3]< do.call("rbind", lapply(strsplit(all.GO.2[,3], " "),
function(x)ifelse(length(x)==1, x[1],
ifelse(length(x)==2, paste(x[1], x[2]),
ifelse(length(x)==3, paste(x[1], x[2], x[3]),
ifelse(length(x)>3, paste(x[1], x[2], x[3], x[4])))))))

all.GO.factor< factor(unlist(GO.tot, use.names=FALSE))
length.GO.terms< summary(all.GO.factor, maxsum=length(all.GO))
#we can remove GO terms with very few genes in order to have a more
general annotation
length.GO.terms< length.GO.terms[which(length.GO.terms>4)]
```

2.2.1 Differential Expression

Basically, identifying differentially expressed genes in RNA-seq datasets is similar to the strategy applied for microarray analysis. The objective is to perform a statistical analysis of the transcript mean or median expression to identify those which are significantly more or less expressed in comparison to the remaining transcripts in an experimental condition. Such analysis can be conducted using common R packages including edgeR (Robinson, McCarthy, & Smyth, 2010) and DESeq (Anders & Huber, 2010), available at Bioconductor (http://bioconductor.org/). The resulting *P*-values obtained from linear models are then adjusted by False Discovery Rate (FDR) or family-wise error rate (Bonferroni correction) and allow comparing gene expression. The sets of differentially expressed genes may next be used in gene set enrichment analysis to have better insights in the biological processes found in a given sample, provided that sufficient annotation information has been attributed to each transcript. Although the statistical analysis may be conducted by comparing single samples, it is more appropriated to have two or more biological replicates to strengthen the different tests.

Here is an example using edgeR (freely available at Bioconductor; http://bioconductor.org/packages/release/bioc/html/edgeR.html) to compare expression levels obtained in two conditions.

```
#Rscript
#edgeR
#load packages
library(edgeR)
```

```
rnaseqMatrix <  fpkm.table[rowSums(fpkm.table)>=2,]
#declare the number of biological repeats.
#although not recommended, this works for only one replicate per
condition
#be sure that expression matrix columns are ordered accordingly
Conditions <  factor(c(rep("Condition1", 1), rep("Condition2", 1)))

exp_study <  DGEList(counts=rnaseqMatrix, group=conditions)
exp_study <  calcNormFactors(exp_study)
et <  exactTest(exp_study, dispersion=0.1)
tTags <  topTags(et,n=NULL)

#get genes significantly (FDR corrected p value<0.05) up regulated
in condition 1
DE.geneList< rownames(tTags$table)[which(tTags$table[,4]<0.05  &
tTags$table[,1]<0)]
```

For automatic multiple pairwise comparisons, we recommend to use a Trinity wrapper script named run_DE_analysis.pl. as follow:

```
#calculate enrichment of GO terms from differentially expressed gene
lists
GO.tot.DEG< GO.tot[DE.geneList]
GO.factor.DEG<< factor(unlist(GO.tot.DEG, use.names=FALSE))
length.GO.a.terms< summary(GO.factor.DEG, maxsum=length(all.GO))
table.count< matrix(nrow=nlevels(GO.factor.DEG), ncol=2)
for (i in 1:length(length.GO.a.terms)){
  i.tmp< names(length.GO.a.terms[i])
  table.count[i, 1]< length.GO.terms[i.tmp]
  table.count[i, 2]< length.GO.a.terms[i.tmp]
}
rownames(table.count)< names(length.GO.a.terms)
res.pval< sapply(1:nrow(table.count), function(y){
  q=table.count[y,2]
  m=table.count[y,1]
  n=nrow(fpkm.table) table.count[y,1]
  k=length(DE.geneList)
  phyper(q, m, n, k, lower.tail=F)})
res.pval< p.adjust(res.pval, method="BH")
GO.table< cbind(table.count, "Pval"=res.pval)
```

TIP: this procedure may be used for any gene list stored in a vector object (such as those that may be obtained in the following procedures).

2.2.2 Correlation Analysis

Identification of candidate genes from a biosynthetic pathway can be carried out by correlation analyses. This procedure aims to identify a set of genes with the highest level of coexpression with a list of previously characterized genes of the corresponding pathway, used as baits. Correlation may be calculated from linear models or following the Pearson statistic (Pearson correlation coefficient, PCC). Calculating correlations for a transcript with the entire transcriptome may be computationally intensive. In biosynthetic or signaling pathways, it may be useful to (i) calculate PCC of each candidate gene with all other genes and (ii) determine intersections between lists of coexpressed genes to identify missing elements. In such a case, the lists of coexpressed genes are determined by setting an appropriate PCC threshold.

```
#Rscript
#collect candidate names stored in one given file, one name per row
candidates.list< scan("genes_of_interest", what="character")
correlation.list< lapply(candidates.list,
  function(x){sapply(1:nrow(fpkm.table),
    function(y)cor(as.numeric(fpkm.table[x,]), as.numeric(fpkm.
table[y,]))})
###use mclapply from 'parallel' package when several CPUs are
available
library('parallel')
cpu.numbers< detectCores()
correlation.list< mclapply(candidates.list,
  function(x){sapply(1:nrow(fpkm.table),
    function(y)cor(as.numeric(fpkm.table[x,]), as.numeric(fpkm.
table[y,]))}, mc.cores=cpu.numbers)

#get co expressed gene list for r>0.8
best.correlated.list< lapply(correlation.list,
function(x)names(which(x>0.8)))
names(best.correlated.list)< candidates.list

#determine and visualize intersections
#for 2 intersections
library(gplots)
intersection.res< venn(list("geneA"=best.correlated.list
[["geneA"]],
            "geneB"=best.correlated.list[["geneB"]]))
```

```
   #for 3 intersections
   intersection.res< venn(list("geneA"=best.correlated.list
[["geneA"]],
            "geneB"=best.correlated.list[["geneB"]],
            "geneC"=best.correlated.list[["geneC"]]))

   plot(intersection.res)
   #get intersection information
   attr(intersection.res, "intersections")

   #retrieve annotation for a given intersection
   tmp.list< attr(intersection.res, "intersections")[[1]]
   annot.tmp< sapply(tmp.list, function(x)which(annot.ok[,1] %in% x)
[1])
   annot.tmp.2< annot.ok[annot.tmp,]
   #export as a csv file, readable in any Microsoft Office Excel or
LibreOffice Calc
   write.csv(annot.tmp.2, "intersect_1.csv")
```

2.2.3 Clustering Procedures

Although correlation analyses (PCC calculations for example) constitute an efficient approach to identify candidate genes, application of gene clustering procedures renders gene discrimination even more stringent. A basic guideline of these strategies is described later.

2.2.3.1 Partitioning

The partitioning methods aim at creating groups of genes with the lowest variance within each group. The reference method is the *k*-means clustering. With a given *k* value which indicate the final number of groups, *k*-means algorithm tries to associate genes within each group such as the sum of their square value is minimized. The choice of *k* is intricate but can be guided by graphical inspection. For example, one may plot *k* values (eg, 1–100) against within group sum of squares for those different *k* values. The R function *k*-means as well as other functions of Cluster package can be used ("fanny" function for example).

```
   #Rscript
   #test within sum of squares for different k values
   withinss.values< sapply(1:50, function(x)kmeans(fpkm.mat, x, nstart
=25, iter.max=25)$tot.withinss)
```

```
#plot total within sum of squares for each k value
plot(1:50, withinss.values)
#this plot gives a visual inspection of appropriate k values, which
correspond to the lowest values minimizing the total within group sum of
squares.

#partition dataset according to the optimal k value (may also be
determined using silhouette plots)
kmeans.res< kmeans(fpkm.mat, k, nstart=25, iter.max=100)

#be careful, cluster research is a randomized process, so clusters will
have different names #but not composition if the command is re run. One
might set a set.seed() value for #reproducibility purposes.

library(ggplot2)
library(reshape2)

#plot cluster expression profiles : use cluster centers in kmeans.res
object
#first reformat table for ggplot plotting
melted.kmeans.res< melt(cbind.data.frame("Cluster"=paste
("cluster", 1:k, kmeans.res$size, sep="_"), kmeans.res$centers))

ggplot(melted.kmeans.res, aes(variable, value))+geom_point()
+geom_line(aes(group=1))+facet_wrap(~Cluster,scales="free_y",
ncol=2)+theme(axis.text.x=element_text(angle=330, hjust=0))

#retrieve annotation for a given cluster
tmp.list< names(which(kmeans.res$cluster==clusternumber))
annot.tmp< sapply(tmp.list, function(x)which(annot.ok[,1] %in% x)
[1])
annot.tmp.2< annot.ok[annot.tmp,]
#export as a csv file, readable in any Microsoft Office Excel or
LibreOffice Calc
write.csv(annot.tmp.2, "annotation_genes_clusternumber.csv")
```

The MB-RNA-seq cluster package (Si, Liu, Li, & Brutnell, 2014) was specifically designed for RNA-seq data. It adapts a model-based (non binomial or negative Poisson) distribution to an initial *k*-means partitioning.

2.2.3.2 Hierarchical Clustering

Given a dissimilarity matrix (computed by euclidean distances, for example), a hierarchical clustering procedure acts iteratively to cluster similar

individuals (genes) by joining most similar individuals. New dissimilarity measures (Ward, UPGMA, Lance-William, etc.) are calculated at each iteration between the new formed cluster and the remaining genes. In the final tree, genes with similar expression patterns are grouped together.

```
#Rscript
#calculate dissimilarity matrix
d.mat< dist(fpkm.table, method="euclidean")
#cluster genes
hclust.dmat< hclust(d.mat, method="ward.D2")
#plot tree
plot(hclust.dmat)
#create cluster by cutting tree
#first, observe how the tree is cut for different thresholds; try
different k values
rect.hclust(hclust.dmat, k=5)
#get cluster composition and size
cluster.hclust.dmat< cutree(hclust.dmat, k=5)
cluster.size< sapply(levels(as.factor(cluster.hclust.dmat)),
function(x)length(which(cluster.hclust.dmat==x)))
names(cluster.size)< paste("Cluster", levels(as.factor(cluster.
hclust.dmat)), sep="")

#plot cluster expression profiles
mean.clust< lapply(1:nlevels(as.factor(cluster.hclust.dmat)),
function(x){
 cbind.data.frame("Mean"=apply(fpkm.table[which(cluster.hclust.
dmat==x),], 2, mean),"Sample"=colnames(fpkm.table), "Cluster"=rep
(paste(x, cluster.size[x], sep="_"), ncol(fpkm.table)))})
mean.clust.table< do.call("rbind", mean.clust)

p< ggplot(mean.clust.table, aes(x=Sample, y=Mean))
p+geom_point()+geom_line(aes(group=Cluster))+facet_wrap
(~Cluster, ncol=4)

#retrieve annotation for a given cluster
tmp.list< names(which(cluster.hclust.dmat==clusternumber))
annot.tmp< sapply(tmp.list,function(x)which(annot.ok[,1]%in%x)[1])
annot.tmp.2< annot.ok[annot.tmp,]
#export as a csv file, readable in any Microsoft Office Excel or
LibreOffice Calc
write.csv(annot.tmp.2, "annotation_genes_clusternumber.csv")
```

In addition, agglomerative clustering may also be tested for investigation purposes with the "agnes" function from the Cluster package. Many dissimilarity measures are available with parameters (arguments to the dissimilarity methods) that may be fine-tuned to improve clustering.

2.2.3.3 HOPACH

This function associates an initial *k*-means based partitioning and a final hierarchical clustering procedure to order similar genes within a sum cluster (van der Laan & Pollard, 2003). Clusters are used to construct tree branches. The final order of genes is used to group genes displaying very correlated expression levels.

```
#Rscript
library(hopach)
gene.dist< distancematrix(fpkm.mat,"cosangle")
gene.hobj< hopach(fpkm.mat.sorted,dmat=gene.dist)
rm(gene.dist)akeoutput(fpkm.mat.sorted, gene.hobj, bootobj=NULL,
file="HOPACH.out",
        gene.names=rownames(fpkm.mat.sorted))

res.hopach< read.table("HOPACH.out", header=T)
rownames(res.hopach)< as.vector(res.hopach[,2])
#in this object, we have to look at the order of genes: this corresponds
to the order in the final tree, and genes with close expression patterns are
found near each other.
position.of.interest< res.hopach["GeneOfInterest",1]
#this returns the position of GeneOfInterest; the further step is to
analyse functions of #genes located in the neighborhood of this position
(eg, +/  100)
names(res.hopach[(position.of.interest 100): (position.of.interest
+100),1])
```

3. VALIDATION OF CANDIDATE GENE FUNCTION BY BIOLISTIC-MEDIATED VIGS

Whatever the efficiency of the procedures of candidate gene identification, functional validation of each candidate gene is required before considering their transfer into heterologous organisms to reconstitute the biosynthetic pathway of a desired valuable compound. While direct functional approaches including protein expression and biochemical assays are still frequently undertaken, rapid and potent screening of multiple candidates

performed through transient gene invalidations mediated by VIGS are increasingly popular. Based on reverse genetic principles, this approach aims at transiently silencing a specific gene in planta, by using the RNA degradation system that plants deploy to respond to viral infections, and at studying its consequence on multiple biological processes such as the biosynthesis of a specialized metabolites. Over the last 15 years, plenty protocols of VIGS have been described for several plants. However, with only few exceptions, all these protocols rely on the inoculation of the viral genome through *Agrobacterium tumefaciens*-mediated transformations. However, this biological delivery strategy is subject to host specificity restriction as well as the induction of plant defense responses, which are commonplace for medicinal plants. For instance, in *C. roseus*, such type of reactions has precluded, for long, the development of an efficient procedure of plant agrotransformation. To date, three VIGS protocols have been described for *C. roseus*, which are all based on the inoculation of tobacco rattle virus (TRV) vectors using *Agrobacterium* by mechanical inoculation through piercing or pinching the stem below the meristem or by seedling infiltration (De Luca, Salim, Levac, Atsumi, & Yu, 2012; Liscombe & O'Connor, 2011; Sung, Lin, & Chen, 2014). Recently, we described a distinct delivery method relying on the transfer of the TRV vector (pTRV1 and pTRV2) by a biolistic-mediated transformation of *C. roseus* plantlets (Carqueijeiro et al., 2015). By eliminating *Agrobacterium* as shuffling vector avoiding thus host specificity problems, this strategy potentially constitutes a transferable tool for other plant specifies recalcitrant to *Agrobacterium*.

3.1 Plant Material and Growth Condition Pretransformation

Seeds of *C. roseus* (Little Bright Eye or Apricot sunstorm) were germinated and cultivated at 23°C using loam as substrate, in a green house, under a 16-h light/8-h dark cycle, with white fluorescent light (maximum intensity of 70 μmol m^{-2} s^{-1}). At the cotyledon stage, plants were individually potted and grown until the first leaf pair reached full development and the second pair just emerged (Fig. 2A).

TIP: To allow an accurate analysis of the silencing results, we recommend preparing around 10 control plants (transformed with empty vectors), 10 plants per silenced candidate gene, and 10 plants transformed with constructs generating easily identifiable phenotype modifications. Monitoring the kinetic of phenotype appearance allows determining the optimal period for sample harvesting. Silencing of genes encoding phytoene desaturase

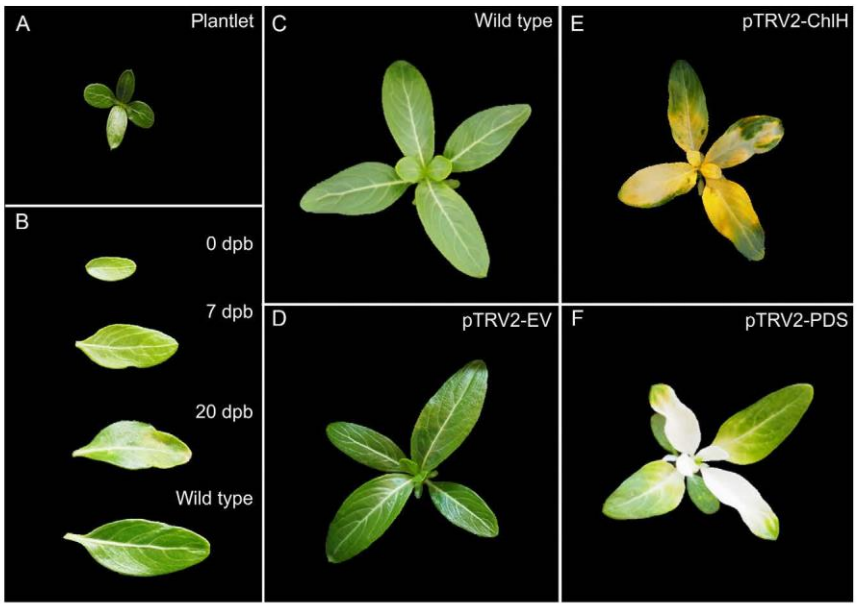


Fig. 2 Virus-induced gene silencing in *Catharanthus roseus* by biolistic transformation (VIGS). 2 weeks old *C. roseus* plantlets presenting one pair of fully expanded leaves were used to perform the particle bombardment. (A) Time table representing the development of the leaves following transformation of VIGS vectors (B) prebombardment (0 dpb), 7 (7 dpb), and 20 (20 dpb) days after bombardment, as compared with a 20 days old nontransformed leaf (wild type). (C–F) Phenotypic aspect of *C. roseus* plants portraying different conditions including wild-type plants (C), plants transformed with empty vector pTRV2-EV (D), protoporphyrin IX magnesium chelatase pTRV2-ChlH depicting the characteristic yellow pigmentation (E), or pTRV2-PDS exhibiting the bleaching of the leaves (F), at 30 dpb.

(PDS; De Luca et al., 2012) or protoporphyrin IX magnesium chelatase (ChlH; Liscombe & O'Connor, 2011) usually leads to useful results.

3.2 Silencing Constructs for VIGS

pTRV vectors (pTRV1 and pTRV2-MCS) expressing the two components of the TRV genome were obtained from the Arabidopsis Biological Resource Center (ABRC) and were used to propagate the virus within plantlets. Fragment of 200–400 bp of the target genes is usually cloned into appropriated restriction sites of the pTRV2 multiple cloning site using classical endonuclease-based DNA manipulation. Supercoiled plasmids used for particle bombardment were isolated from *E. coli* cultures using Nucleospin Plasmid kit (Macherey-Nagel) following manufacturer's instructions.

TIP: Before each VIGS experiment, we advise to confirm plasmid integrity by basic analytical electrophoresis gel. Avoid proceed in case of plasmid degradation.

3.3 Biolistic-Mediated Transformation of *C. roseus*

3.3.1 Particle Preparation

1. Weigh 30 mg of 1 μm gold particles (Bio-Rad) in a glass tube and dry heat at 180°C for 8 h.
2. Wash the gold particles, 5 min with 1 mL of fresh 70% ethanol using vortex and sonication bath. Transfer to 2 mL sterile microcentrifuge tube.
3. Centrifuge the gold particles at 16,000 × *g* for 5 s.
4. Remove the supernatants; wash the pellets three times with 1 mL of sterilized milliQ water.
5. Centrifuge the gold particles at 16,000 × *g* for 5 s and resuspend the gold particles in 500 μL of 50% glycerol (p/v) sterile.

3.3.2 Coating of Plasmids onto Particles

1. For 10 bombardments, 10 μg of each plasmid are coated and precipitated onto 6.25 mg of gold particles. Plasmid solutions usually displayed a 1 μg/μL final concentration.
2. Mix 50 μL of 0,1 *M* spermidine with 100 μL of glycerol stocked gold particles in a sterilized tube and homogenize simultaneously with vortex and short pulses of ultrasounds with sonication bath.
3. Add the appropriated volume of purified DNA plasmid (not exceeding 10 μL for each plasmid) and mix. Maximize the coating efficiency by allowing the binding for 3 min and by vortexing each minute.
4. While homogenizing the solution with vortex, add 60 μL of 2.5 *M* $CaCl_2$ and mix for additional 15 min with a vortex at constant speed (around 1000 rpm).
5. Spin the tubes for 2 s at 16,000 × *g* to pellet gold particles coated with plasmids and remove the supernatant.
6. Wash plasmid coated gold particle pellets successively with 500 μL of 70% ethanol and 500 μL of 100% ethanol without resuspending gold particles.
7. Remove all supernatant and resuspend particles in 100 μL of 100% ethanol.
8. Spread 10 μL of coated gold particles onto each macrocarrier (Bio-Rad) to allow drying before transformation.

3.3.3 Particle Bombardment Procedure

Transformations of *C. roseus* plantlets are performed with the Bio-Rad PDS1000/He delivery system according to manufacturer's recommendations, using 1100 psi rupture disks (Bio-rad) under a vacuum pressure of 28 in. of Hg, at a stopping-screen-to-target distance of 9 cm with a 1-cm distance-of-flight of the macrocarriers. A single potted plant is placed in the biolistic device and a unique bombardment is achieved.

3.4 Posttransformation Treatments and Analysis

Following bombardment, plants are replaced in greenhouse and are cultivated under similar conditions until appearance of the phenotype of the ChlH- or PDS-silenced plants used as positive controls of gene silencing. Using these transformation conditions, up to 90–100% of the bombarded periwinkle plants display gene silencing. While limited variations arise, leaf photobleaching or yellowing typically begins around 7–10 days after bombardment while fully bleached neo-formed leaves can be retrieved 21–25 days posttransformation (Fig. 2). Leaves of silenced candidate genes can be thus harvested in the same time laps for further analysis including evaluation of gene silencing by quantitative PCR as well as measurement of alkaloid contents by HPLC analysis (Besseau et al., 2013) in order to identify candidate gene function.

4. STUDYING THE SUBCELLULAR LOCALIZATION OF BIOSYNTHETIC PATHWAY ENZYMES IN PLANT AND YEAST CELLS TO ALLEVIATE BOTTLENECKS IN BIOENGINEERING APPROACHES

Most of the biosynthetic pathways of specialized metabolites, particularly in plants, exhibit a complex intracellular compartmentalization relying on the targeting of their enzymes to distinct organelles. An overview of the level of complexity that this type of organization can reach has been recently depicted in *C. roseus* (Courdavault et al., 2014). In this plant, the distribution of the alkaloid biosynthetic pathways in numerous subcellular compartments involves, as a corollary, manifold transmembrane transports of metabolic intermediates that could impact biosynthetic fluxes. While plants can deploy these exchanges to fine-tune the regulation of metabolites biosynthesis, a negative impact can be engendered in metabolic engineering applications based on the reconstitution of biosynthetic pathways in heterologous organisms. Such reconstitution and more generally, the understanding of the

general plant physiology thus require a complete knowledge of pathway organization. In the following sections, we describe a procedure of protein subcellular localization study based on the creation of fusions with FPs and on their expression in planta through transient biolistic-mediated transformation. This procedure allows the quick and standardized obtainment of robust results of localization, which can be useful when a biosynthetic pathway is composed of numerous enzymes. Given the differences between some of the plant and yeast targeting sequences, we also present a similar strategy allowing the validation of protein subcellular localizations in yeast that is required to avoid enzyme mislocalization and the inherent disruption of metabolite biosynthesis following pathway reconstitution. The detailed protocols are illustrated with localization of enzymes of the MIA biosynthetic pathway of *C. roseus*, including strictosidine synthase (STR) and strictosidine β-D-glucosidase (SGD), two enzymes acting consecutively in the pathway in distinct subcellular compartments (Guirimand et al., 2010).

4.1 Protein Subcellular Localization in *C. roseus* Cells

While protein subcellular localization can be analyzed by biochemical approaches involving assay for enzyme activity/protein immunodetection following cell fractionation or observed directly with electron microscopy of immunogold-labeled sections, expression of fusions with FPs is the most rapid and popular approach. Following stable and/or transient expression of these proteins in plant cells, protein targeting can be determined by visualization of the subcellular fluorescent profiles using epifluorescence or confocal microscopy and simultaneously compared with the fluorescent profiles of markers for each subcellular compartment.

4.1.1 Constructs Expressing Fusions with Fluorescent Proteins in Plant Cells

A myriad of color variants of FPs is now available. The choice of the FPs used to generate fusions can be guided by the intrinsic properties of each variant (pK_a, brightness, etc.) but also by the microscopy equipment utilized for fluorescence visualization. Since the capacities of epifluorescence microscopes (eFM) to discriminate the different FPs are usually lower than those of confocal microscopes, we recommend combining FPs with no overlapping excitation and emission spectra when eFM are used. In such a case, combination of yellow FP (YFP) with cyan FP (CFP) as well as green FP (GFP) with red FP (RFP) or mcherrry provides the best results. We have developed a set of plasmids with each variant color, based on the pSCA-YFP scaffold

(Guirimand et al., 2009), allowing overexpression of fusion proteins in plant cell under the dependence of the constitutive CaMV 35S promoter. These plasmids, available upon request, display multiple cloning sites at both the 5′ and 3′ ends of the FP coding sequence, with sites of compatible restriction enzymes, enabling cloning of the coding sequence of the studied enzymes at each end of the FP with the same cDNA.

One of the main pitfalls in localization studies with fusion proteins is the masking of the targeting sequence in the studied protein by the FP. For instance, if a protein possesses a plastid transit peptide at its N-terminal end, fusion with the C-terminal end of the FP (yielding FP-protein orientation) have to be avoided. To prevent artifactual localizations, a careful analysis of the putative targeting/anchoring sequences of each protein should be performed to select the most suitable orientation.

1. Predictions of the putative targeting sequence are routinely carried out with PSORT, TargetP, Predotar, MitoProt, PredPlant PTS1, NLS mapper, TMHMM algorithms, for instance. *When applied to STR and SGD, this analysis led to the identification of a putative vacuolar/secretory signal peptide at the N-terminal end of STR (1-MANFSESKSM-MAVFFMFFLLLLSSSSSSSSSPIL-35) and of a putative bipartite nuclear localization sequence (NLS) located at the C-terminal end of SGD (537-KKRFREEDKL-VELVKKQKY-555).*
2. Amplify the coding sequence of the studied proteins with high fidelity DNA polymerases and introduce appropriated restriction sites at both ends of the cDNA to allow cloning in the pSCA-vectors. *In our example, STR coding sequence was amplified with primers STR-for (5′-CTGAGA ACTAGTATGGCAAACTTTTCTGAATCTAAA-3′) and STR-rev (5′-CTGAGAACTAGTGCTAGAAACATAAGAATTTCCCTT-3′) that introduce the SpeI restriction site at both extremities to allow cloning into either the SpeI or NheI sites of the pSCA-CFP vector in order to express STR-CFP and CFP-STR fusions, respectively. Similarly, SGD was amplified with primers SGD-for (5′-CTGAGATCTAGAATGGGATCTAAAG ATGATCAGTCC-3′) and SGD-rev (5′-CTGAGATCTAGATTAGT ATTTTTGCTTCTTGACTAACTCAACT-3′) introducing a XbaI site (compatible with SpeI and NheI) to express the SGD-YFP and YFP-SGD fusion proteins.*
3. Following cloning into the pSCA-YFP vectors, extract and sequence the recombinant plasmids to ensure that constructs are exempt from mutations. For plant cell transformations, concentrate plasmids to final 1 $\mu g\ \mu L^{-1}$ concentration before transformation. *Note that supercoiled*

plasmids freshly extracted from E. coli cultures (using Nucleospin Plasmid kit for example) usually provide the best expression levels.

4.1.2 Cell Culture and Plating

Expression of the fusion proteins can be achieved in *C. roseus* cells and we preconize the use of the *C. roseus* C20 strain cell that is suitable for subcellular localization of proteins from the periwinkle or from other plant species.

1. *C. roseus* C20 cell suspensions are cultivated in the dark at 24°C under continuous shaking (100 rpm) for 7 days as previously described (Mérillon, Doireau, Guillot, Chénieux, & Rideau, 1986).
2. At the third day of culture, pour 4 mL of the homogenized cell suspension onto a circular piece of filter paper (45 mm diameter—Fisherbrand A70.70000) in a filtration funnel and apply weak air suction.
3. Transfer plated cells onto solid Gamborg B5 medium (8 g L^{-1} agar) (Gamborg, Miller, & Ojima, 1968) supplemented with 10 μ*M* naphthalene acetic acid (NAA) in a 45-mm Petri dish and cultivate at 24°C for 48 h in the dark before transformation.

4.1.3 Transient Cell Transformation by Biolistic

Studies of protein subcellular localizations can be carried out following stable and/or transient transformations of plant cells. While stable transformation allows performing long-term studies on selected and propagated transformed cells, this approach is more time consuming in particular because the analysis of several transformed cell lines is required to validate localization results. By contrast, transient transformations rapidly generate localization results (within 1–2 days) and allow the observation of thousands of independent transformation events in a single cell plate. As a consequence, transient transformations are suitable to characterize the localization of enzymes from biosynthetic pathways of specialized metabolites. For such transformations, numerous protocols using protoplasts have been described and notably in *C. roseus* (Duarte, Memelink, & Sottomayor, 2010). However, to avoid artifacts of localization caused by the cell stress induced by cell wall removal, we preconize the use of particle bombardments that results in the entry of a single small particle for most of the transformed cells. This constitutes a weakly traumatic situation adapted for localization studies. The protocol described hereafter has been developed according to Guirimand et al. (2009).

1. For each bombardment, 400 ng (or up to 1 μg) of purified plasmid is coated and precipitated onto 500 μg of gold particles prepared as

described in Section 3.3.1. When plasmid cotransformations are performed (with plasmids encoding subcellular markers for instance), prepare a stoichiometric mix of each plasmid. A useful set of plasmids encoding markers of each plat cell compartments has been described in Nelson, Cai, and Nebenführ (2007).

2. Mix 5 µL of 0.1 *M* spermidine with 100 µL of glycerol stocked gold particles in a sterilized tube and homogenize simultaneously with vortex and short pulses of ultrasounds (15 s; 40 W) with sonication bath.
3. Add the appropriate volume of purified DNA plasmid (usually not exceeding 3 µL for each plasmid) and mix.
4. While homogenizing the solution with vortex, add 5 µL of 2.5 *M* $CaCl_2$ and mix for additional 15 min.
5. Spin the tubes for 2 s at 16,000 × *g* to pellet gold particles coated with plasmids and remove the supernatant.
6. Wash plasmid coated gold particles pellets successively with 150 µL of 70% ethanol and 150 µL of 100% ethanol without resuspending gold particles.
7. Remove all supernatant and resuspend particles in 8 µL of 100% ethanol.
8. Spread the 8 µL of coated gold particles on each macrocarrier (Bio-Rad) and allow drying before transformation.
9. Transformations are performed with the Bio-Rad PDS100/He delivery system according to Section 3.3.3, using 1100 psi rupture disks, under a vacuum pressure of 28 in. of Hg, at a stopping-screen-to-target distance of 6 cm with a 1-cm distance-of-flight of the macrocarrier.
10. A single transformation per Petri dish of plated cells is performed and cells are cultivated for 16 h in the dark at 24°C before observation.

4.1.4 Fluorescent Protein Imaging and Epifluorescence Microscopy

Fluorescence profiles of the fusion proteins can be usually observed from 16 to 72 h postbombardment by harvesting transformed cells from the Petri dish and mounting them between slide and cover. Evolution of protein localizations has to be checked during all this period since the kinetic of targeting to each subcellular compartment greatly differs. When a fusion protein is coexpressed with a subcellular compartment fluorescent marker (a protein known to be targeted to a specific subcellular compartment and fused to a different FP) or with a second protein fused to a distinct FP, superimposition of the two distinct fluorescent signals has to be evaluated in order to definitely establish the localization of the studied protein. In

the example described later, image captures of *C. roseus* transiently transformed cells expressing FP-fused proteins are performed with an Olympus BX51 eFM equipped with the Olympus DP71 digital camera with CellD imaging software (Soft Imaging System, Olympus). The YFP and CFP fluorescence signals emitted from fusion proteins are visualized using a YFP filter set (Chroma#31040, 500–520 nm excitation filter, 540–580 nm emission filter) and a Cyan GFP filter set (Chroma#31044v2, 426–446 excitation filter, 460–500 nm band pass emission filter), respectively. YFP and CFP fluorescence are successively acquired and merged with the CellD imaging software while the morphology of transformed cells is observed with differential interference contrast (DIC). Fig. 3 illustrates the sequential image capture process carried out for *C. roseus* cells cotransformed with plasmids expressing the YFP-SGD (Fig. 3A) and STR-CFP (Fig. 3B) fusion proteins,

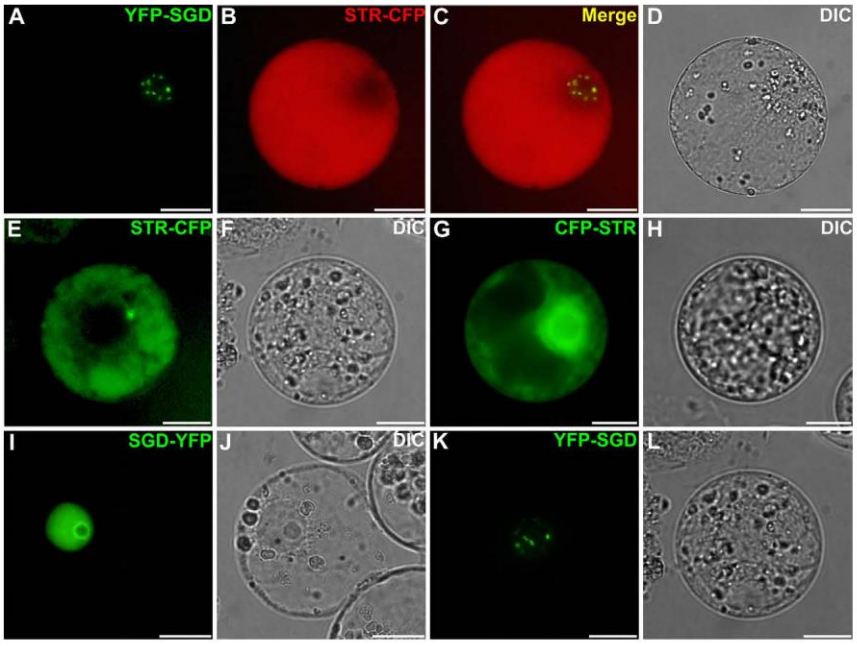


Fig. 3 Subcellular localization of STR and SGD expressed as FP fusions in plant cells. *C. roseus* cells were transiently cotransformed (A–D) or transformed (E–L) with constructs expressing YFP-SGD (A; K), STR-CFP (B; E), CFP-STR (G) of SGD-YFP (I). For cotransformation, superimposition of the two fluorescence signals appears on the merged image (C). Cell morphology (D; F; H; J; L) was observed with differential interference contrast (DIC). Bars = 10 μm.

their numerical superimposition (Fig. 3C), and the observation of cell morphology by DIC (Fig. 3D).

Depending on the cell strain used for transformation, cell aggregation can occur after platting on solid medium rendering more difficult mounting between slide and cover. In such a case, protoplasts of transformed cells can be prepared before observation as described later. Such treatments performed posttransformation and during a short period are not likely to induce modifications of localization.

1. Harvest around 250 mg of transformed cells with an inoculation loop in a tube containing 800 µL of MM Buffer (MES 20 m*M*, Mannitol 0,4 *M*).
2. Allow to decant for 5 min at room temperature, remove supernatant and resuspend cell in 800 µL of MM Buffer. Repeat this washing step twice.
3. Resuspend cell pellet in 1 mL of digestion MM Buffer (MM Buffer containing cellulase R-10 2%, Macerozyme r-10 0.3%, pectolyase 0.2%).
4. Transfer in a Petri dish (diameter 45 mm) and incubate for 2 h in the dark with slow agitation to generate protoplasts.
5. Harvest solution in a 1.7-mL Eppendorf tube, allow decanting for 5 min and remove supernatant.
6. Wash carefully protoplasts as described in step 2 and gently resuspend in a final volume of 400 µL of MM Buffer before mounting.

4.1.5 The Importance of Being Correctly Fused

As mentioned earlier, masking of targeting signals by improper orientation of fusion with FP is a common pitfall observed during subcellular localization studies and may result in artifactual protein targeting. As an illustration of this difficulty, Fig. 3 shows the analysis of the localization of STR that displays a predicted targeting sequence at its N-terminal end. In *C. roseus* transiently transformed cells, the STR-CFP fusion protein is efficiently targeted to the vacuole (Fig. 3E–F) while expression of the CFP-STR fusion protein results in protein mislocalization in the cytosol (Fig. 3G–H) probably caused by the masking of the targeting sequence that renders it non accessible and/or nonfunctional. Furthermore, the effects of an incorrect orientation of fusion with FP can be more tenuous and thus less easily identifiable. For SGD that bears a bipartite NLS at the C-terminal end, fusion with the FP does not inactivate this sequence that is still able to direct the SGD-YFP fusion in the nucleus as a soluble protein (Fig. 3I–J). However, such orientation of fusion annihilates the propensity of SGD to self-interact and to form high molecular weight complexes (appearing as small dots in the nucleus), which can be only observed when the C-terminal end of SGD

is free such as in the YFP-SGD fusion protein (Fig. 3K–L). More complex situations can be also encountered when the studied protein possesses targeting sequences at both extremities as recently depicted for the *C. roseus* isopentenyl diphosphate isomerase (IDI). This protein is characterized by the presence of a N-terminal dual plastid/mitochondria targeting peptide and by a C-terminal type 1 peroxisome targeting sequence (PTS1) requiring the expression of an YFP internal fusion protein to observe the triple localization to plastid, mitochondria, and peroxisome (Guirimand, Guihur, et al., 2012; Guirimand, Simkin, et al., 2012).

4.2 Protein Subcellular Localization in Yeast Cells

While the study of the subcellular localization in the plant cells of multiple enzymes of a biosynthetic pathway of interest allows identifying potential bottlenecks associated to transmembrane exchanges of intermediates, a subsequent confirmation of the correct protein localization can be required following the transfer of the whole biosynthetic pathway in heterologous organisms. Indeed, specific plant-targeting sequences (eg, plastid targeting peptide) are inoperative in yeasts or bacteria and common localization sequences can also be subjected to misinterpretations leading to undesired protein targeting. Such heterologous mistargeting has been depicted for the expression of STR in yeast that undergone a massive secretion in the medium instead of an expected vacuolar localization, due to the promiscuity of vacuolar/secretion targeting sequences (Geerlings et al., 2001). As a consequence, we recommend validating protein localization and, if needed, to replace inefficient plant localization peptides by sequences adapted to the host organisms. Our protocol of protein subcellular localization in yeast is described later.

4.2.1 Constructs Expressing Fusions with Fluorescent Proteins in Yeast Cells

Numerous plasmids dedicated to the expression of FP fusion proteins in yeast are now available. Most of these plasmids harbor codon optimized FP coding sequences under the control of strong constitutive promoters such as ACT1 or TEF1. To avoid potential problems caused by the continuous expression of FP fusions and a possible mislocalization caused by a massive protein overexpression, plasmids bearing inducible promoters should preferentially be used. This type of plasmids can be easily generated by using skeleton of commercial plasmids such as those of the pESC series (Agilent

Technologies) possessing the pGAL1/pGAL10 inducible promoters but are also available upon request.

1. Amplify yeast codon optimized coding sequences of FP with high fidelity DNA polymerases. For instance, yeYFP and yeCFP coding sequences can be amplified with a similar primer couple composed of FLUSC1 (5′-CTGAGGTCTAGA*AGATCT*ACTAGTATGTCTAA AGGTGAAGAATTAT-3′ introducing *Xba*I, *Bgl*II, and *Spe*I restriction sites) and FLUSC2 (5′-CTGAGAGGATCCTTACCTAGGTTTGTA CAATTCATCCATACCA-3′ introducing *BamH*I and *Avr*II restriction sites).
2. Digest PCR products by *Xba*I and *BamH*I and clone into pESC-LEU and/or pESC-TRP linearized by *Spe*I and *Bgl*II that generate compatible extremities with *Xba*I and *BamH*I, respectively. *In the following example, we cloned yeYFP and yeCFP coding sequences into pESC-LEU and pESC-TRP to generate pESC-LEU-YFP and pESC-TRP-CFP, respectively.*
3. Extract plasmids and sequence. The *Spe*I and *Bgl*II restriction sites initially present in the plasmid sequence have been disrupted through their annealing with *Xba*I and *BamH*I compatible extremities. The coding sequences of the studied proteins can now be cloned into the *Bgl*II or *Spe*I restriction sites (introduced by PCR) to generate a protein fused to the N-terminus of YFP/CFP or into *Avr*II to express a protein fused to the C-terminal end of YFP/CFP. *These restriction sites have been selected for compatibility with cloning strategy used to study the protein localization in plant cells using the pSCA-YFP vectors. The coding sequences of STR and SGD (Section 4.1.1) have thus been cloned in pESC-LEU-YFP and pESC-TRP-CFP, accordingly.*

4.2.2 Preparation of Yeast Competent Cells

The following protocol of yeast competent cell preparation is restricted to transformation by electroporation as described in the subsequent section.

1. Streak a *S. cerevisiae* reference strain (WT303 for instance, auxotroph to leucine and tryptophan) on solid YPD medium (10 g L^{-1} yeast extract, 20 g L^{-1} peptone, 20 g L^{-1} dextrose, and 15 g L^{-1} agar) and grow at 30°C for 48 h.
2. Pick a single colony, inoculate 10 mL of liquid YPD medium and grow at 30°C overnight.
3. Inoculate 50 mL of liquid YPD medium with 500 µL of the preculture and grow around 4 h at 30°C until reaching the end of the exponential growth (absorbance at 600 nm of 0.6–1.5).

4. Chill the yeast culture 15 min on ice before transferring into sterile 50 mL Falcon tubes. Centrifuge at 3000 × *g* for 10 min at 4°C to pellet yeast.
5. Remove the supernatant and resuspend yeast pellet in 40 mL of ice-cold DTT-supplemented lithium acetate solution (lithium acetate 100 m*M*, DTT 10 m*M*). Incubate at 28°C for 1 h.
6. Centrifuge at 3000 × *g* for 10 min at 4°C in sterile centrifuge tubes, discard the supernatant, and wash the pellet twice with 20 mL sorbitol 1 *M*.
7. Resuspend yeast pellet in 5 mL of 1 *M* sorbitol and distribute in pre-chilled eppendorf tubes before flash freezing in liquid nitrogen. Store at −80°C.

4.2.3 Protocol of Yeast Cell Transformation

Several protocols of yeast transformation have been described either by heat shock treatment or by electroporation. We recommend using this last one due to a higher efficiency of transformation.

1. Mix 0.5–2 μg of DNA Purified Plasmid (pESC-LEU-YFP and/or pESC-LEU-TRP) with 200 μL of yeast competent cells in a sterile Eppendorf tube.
2. Incubate 10 min on ice and transfer into a 0.2-cm gap width electroporation cuvette.
3. Perform electroporation by applying a 5-ms electric pulse of 1.5 kV (Bio-Rad MicroPulser Electroporator—program Sc2).
4. Plate transformed cells on CSM-LEU (YNB with ammonium 6.7 g L^{-1}, dextrose 20 g L^{-1}, DOB-LEU 500 mg L^{-1}, agar 20 g L^{-1}), CSM-TRP (YNB with ammonium 6.7 g L^{-1}, dextrose 20 g L^{-1}, DOB-TRP 500 mg L^{-1}, agar 20 g L^{-1}), or CSM-LEU-TRP (YNB with ammonium 6.7 g L^{-1}, dextrose 20 g L^{-1}, DOB-LEU-TRP 500 mg L^{-1}, agar 20 g L^{-1}) selective solid medium, depending on the combination of selection marker used in the cloning vectors and incubate at 30°C for 3–5 days.
5. Streak independently transformed yeast colonies onto CSM-LEU, CSM-TRP, or CSM-LEU-TRP containing 2% galactose to induce protein expression and grow two additional days at 30°C.
6. Resuspend transformed yeast independently in 100 μL of water and analyze fluorescence using eFM in the conditions described in Section 4.1.4.

4.2.4 Correct and Incorrect Plant Protein Targeting in Yeast

Since STR and SGD display a complete sequestration at the subcellular level, being targeted to vacuole and nucleus, respectively, in plant cells, we illustrated the validation of the localization of both enzymes in yeast though their expression as fusions with FP (Section 4.2.1). As compared to nonfused FP exhibiting a classical nucleocytosolic localization (Fig. 4A–B), STR-CFP was efficiently targeted to the vacuole in our experimental conditions, with a negligible secretion in the medium (Fig. 4C–D). Such localization is thus consistent with that observed in plant cells. By contrast, SGD localization in yeast produced a more complex situation. For the SGD-YFP fusion, a unique and unexpected targeting to the vacuole was observed (Fig. 4E–F) as revealed with cotransformation with STR-CFP (Fig. 5A–D). Furthermore, we observed that YFP-SGD fusions were targeted, in similar proportions, to the nucleus (Figs. 4G–H and 5E–H) or to the vacuole (Figs. 4I–J and 5I–L) and less frequently to both compartments (Figs. 4K–L and 5M–P). In this case, targeting of STR and SGD to the

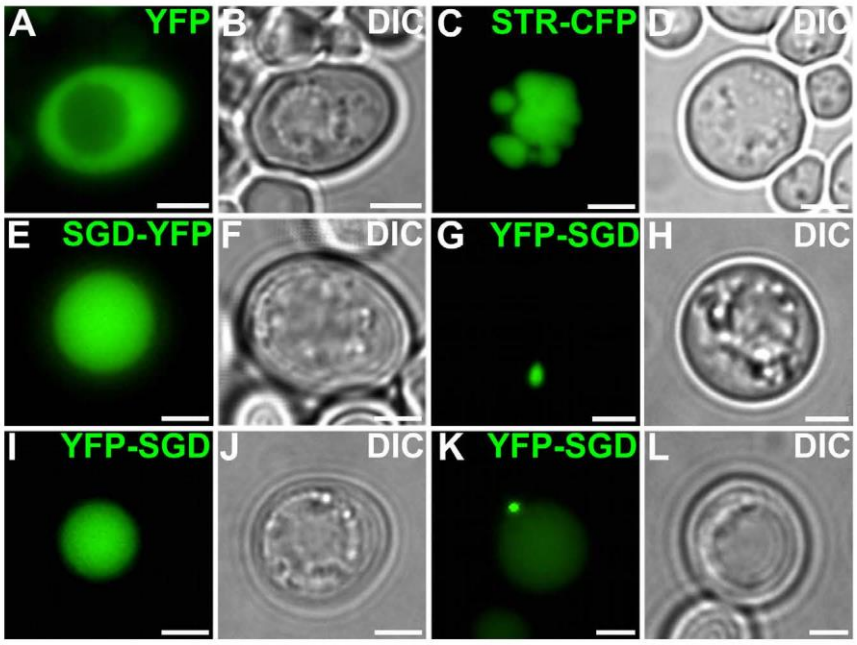


Fig. 4 Subcellular localization of STR and SGD expressed as FP fusions in yeast cells. Yeast cells were transformed with constructs expressing unfused YFP (A), STR-CFP (C), SGD-YFP (E), or YFP-SGD (G; I; K). Cell morphology (B; D; F; H; J; L) was observed with differential interference contrast (DIC). Bars = 2 μm.

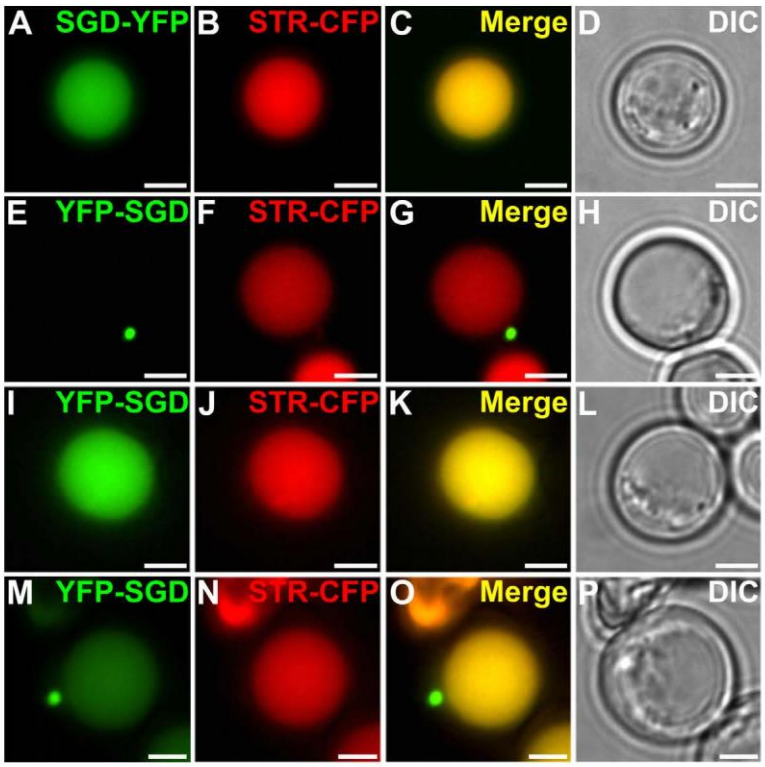


Fig. 5 Colocalization of STR and SGD expressed as FP fusions in yeast cells. Yeast cells were co transformed with constructs expressing SGD-YFP and STR-CFP (A–D) or YFP-SGD and STR-CFP (E–P). Colocalization of the two fluorescence signals appears on the merged images (C; G; K; O). Cell morphology (D; H; L; P) was observed with differential interference contrast (DIC). Bars = 2 μm.

vacuole can result in potential undesirable effects in yeast since it can lead to a massive deglucosylation of stricosidine that might be toxic as a result of protein reticulation (Guirimand et al., 2010). Albeit these mislocalizations potentially result from the fusion with FPs that could alter the functionality of the targeting sequences, it also highlights the differences of protein behavior in plant and yeast cells and the importance to validate protein. It should also be taken into consideration when small tags are added to proteins to monitor their expression in heterologous organisms.

5. CONCLUDING REMARKS

While synthetic biology now technically offers the possibility of transferring and controlling the whole biosynthetic pathway of a valuable

compound in heterologous organisms, identification of all the enzymes of the pathway and of its subcellular organization still constitutes essential prerequisites to the achievement of bioengineered productions. With the recent availability of massive transcriptomics and genomics data concerning organisms producing metabolites of interest, notably for plants, and with the development of efficient tools for candidate gene prediction, validation, and characterization, such complete and intensive deciphering of pathways are more than ever right at our fingertips. As such, the protocols described earlier are part and parcel of the technical arsenal that can be deployed to attain pathway characterization. Albeit being initially developed for the Madagascar periwinkle, most of them are applicable to other plant species. We expect for instance that direct-transformation of vectors for VIGS will expand the availability of this powerful functional approach outside the host range of Agrobacterium-delivered VIGS. By showing unforeseen localization obtained with enzymes of the MIA pathway, we also highlighted the importance to take account of the subcellular localization in metabolic engineering pathways, in particular following transfer to microbial systems which may not properly process plant-targeting signals. Thereby, the study of protein subcellular localization still constitutes a milestone in the early steps of synthetic biology approaches dedicated to metabolic engineering.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the Région Centre (France, ABISAL grant, Doctoral Fellowship to F.L. and Post-Doctoral Fellowship to I.C.). E.F. was financed by a fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche (France). We also thank M.A. Marquet, E. Danos, and E. Marais for maintenance of cell cultures. We also acknowledge the Cascimodot Fédération (CCSC, Orléans) for access to the Région Centre computing grid.

REFERENCES

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*, R106.

Besseau, S., Kellner, F., Lanoue, A., Thamm, A. M. K., Salim, V., Schneider, B., et al. (2013). A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in *Catharanthus roseus*. *Plant Physiology, 163*, 1 12.

Brown, S., Clastre, M., Courdavault, V., & O'Connor, S. E. (2015). De novo production of the plant-derived alkaloid strictosidine in yeast. *Proceedings of the National Academy of Sciences of the United States of America, 112*, 3205 3210.

Carqueijeiro, I., Masini, E., Foureau, E., Sepúlveda, L. J., Marais, E., Lanoue, A., et al. (2015). Virus-induced gene silencing in *Catharanthus roseus* by biolistic inoculation of tobacco rattle virus vectors. *Plant Biology, 17*, 1242 1246.

Cassels, K. B., Asencio, M., Conget, P., Speisky, H., Videla, A. L., & Lissi, A. E. (1995). Structure-antioxidative activity relationships in benzylisoquinoline alkaloids. *Pharmacology Research, 31*, 103 107.

Chang, M. C., Eachus, R. A., Trieu, W., Ro, D. K., & Keasling, J. D. (2007). Engineering *Escherichia coli* for production of functionalized terpenoids using plant P450s. *Nature Chemical Biology, 3*, 274 277.

Courdavault, V., Papon, N., Clastre, M., Giglioli-Guivarc'h, N., St-Pierre, B., & Burlat, V. (2014). A look inside an alkaloid multisite plant: The Catharanthus logistics. *Current Opinion in Plant Biology, 19*, 43 50.

De Luca, V., Salim, V., Levac, D., Atsumi, S. M., & Yu, F. (2012). Discovery and functional analysis of monoterpenoid indole alkaloid pathways in plants. In D. A. Hopwood (Ed.), *Methods in enzymology: Vol. 515. Natural product biosynthesis by microorganisms and plants, part A* (pp. 207 229). Waltham: Academic Press—Elsevier.

Drewes, S. E., George, J., & Khan, F. (2003). Recent findings on natural products with erectile-dysfunction activity. *Phytochemistry, 62*, 1019 1025.

Duarte, P., Memelink, J., & Sottomayor, M. (2010). Fusion with fluorescent proteins for subcellular localization of enzymes involved in plant alkaloid biosynthesis. In E. Fett-Neto & A. Germano (Eds.), *Methods in molecular biology: Vol. 643. Plant secondary metabolism engineering* (pp. 275 290). New York: Humana Press.

Dugé de Bernonville, T., Clastre, M., Besseau, S., Oudin, A., Burlat, V., Glévarec, G., et al. (2015). Phytochemical genomics of the Madagascar periwinkle: Unravelling the last twists of the alkaloid engine. *Phytochemistry, 113*, 9 23.

Dugé de Bernonville, T., Foureau, E., Parage, C., Lanoue, A., Clastre, M., Londono, M. A., et al. (2015). Characterization of a second secologanin synthase isoform producing both secologanin and secoxyloganin allows enhanced de novo assembly of a Catharanthus roseus transcriptome. *BMC Genomics, 16*, 619.

Gamborg, O. L., Miller, R. A., & Ojima, K. (1968). Nutrient requirements of suspension cultures of soybean root cells. *Experimental Cell Research, 50*, 151 158.

Geerlings, A., Redondo, F., Contin, A., Memelink, J., van Der Heijden, R., & Verpoorte, R. (2001). Biotransformation of tryptamine and secologanin into plant terpenoid indole alkaloids by transgenic yeast. *Applied Microbiology and Biotechnology, 56*, 420 424.

Góngora-Castillo, E., Fedewa, G., Yeo, Y., Chappell, J., DellaPenna, D., & Buell, C. R. (2012). Genomic approaches for interrogating the biochemistry of medicinal plant species. *Methods in Enzymology, 517*, 139 159.

Guirimand, G., Burlat, V., Oudin, A., Lanoue, A., St-Pierre, B., & Courdavault, V. (2009). Optimization of the transient transformation of Catharanthus roseus cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. *Plant Cell Reports, 28*, 1215 1234.

Guirimand, G., Courdavault, V., Lanoue, A., Mahroug, S., Guihur, A., Blanc, N., et al. (2010). Strictosidine activation in Apocynaceae: Towards a "nuclear time bomb"? *BMC Plant Biology, 10*, 182.

Guirimand, G., Guihur, A., Phillips, M. A., Oudin, A., Glévarec, G., Melin, C., et al. (2012). A single gene encodes isopentenyl diphosphate isomerase isoforms targeted to plastids, mitochondria and peroxisomes in *Catharanthus roseus*. *Plant Molecular Biology, 79*, 443 459.

Guirimand, G., Simkin, A. J., Papon, N., Besseau, S., Burlat, V., St-Pierre, B., et al. (2012). Cycloheximide as a tool to investigate protein import in peroxisomes: A case study of the subcellular localization of isoprenoid biosynthetic enzymes. *Journal of Plant Physiology, 169*, 825 829.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols, 8*, 1494 1512.

Hanson, J. R. (2003). Natural products: Secondary metabolites. In E. W. Abel (Ed.), *Tutorial chemistry texts: Vol. 17* (pp. 3 18). London: The Royal Society of Chemistry.

Hawkins, K. M., & Smolke, C. D. (2008). Production of benzylisoquinoline alkaloids in Saccharomyces cerevisiae. *Nature Chemical Biology*, *4*, 564 573.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, *26*, 680 682.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357 359.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*, R25.

Lee, M. (2011). The history of Ephedra (ma-huang). *Journal of the Royal College of Physicians of Edinburgh*, *41*, 78 84.

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323.

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

Li, L. P., Liu, W., Liu, H., Zhu, F., Zhang, D. Z., Shen, H., et al. (2015). Synergistic antifungal activity of berberine derivative B-7b and fluconazole. *PLoS One*, *10*, e0126393.

Liscombe, D. K., & O'Connor, S. E. (2011). A virus-induced gene silencing approach to understanding alkaloid metabolism in Catharanthus roseus. *Phytochemistry*, *72*, 1969 1977.

Mano, M. (2006). Vinorelbine in the management of breast cancer: New perspectives, revived role in the era of targeted therapy. *Cancer Treatment Reviews*, *32*, 106 118.

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*, 671 682.

Mérillon, J. M., Doireau, P., Guillot, A., Chénieux, J. C., & Rideau, M. (1986). Indole alkaloid accumulation and tryptophan decarboxylase activity in *Catharanthus roseus* cells cultured in three different media. *Plant Cell Reports*, *5*, 23 26.

Miller, M. L., & Ojima, I. (2001). Chemistry and chemical biology of taxane anticancer agents. *Chemical Records*, *1*, 195 211.

Nelson, B. K., Cai, X., & Nebenführ, A. (2007). A multicolored set of in vivo organelle markers for co-localization studies in Arabidopsis and other plants. *The Plant Journal*, *51*, 1126 1136.

Newman, D. J., & Cragg, G. M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of Natural Products*, *75*, 311 335.

Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., et al. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, *496*, 528 532.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., et al. (2003). TIGR Gene Indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics*, *19*, 651 652.

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

Pollier, J., Moses, T., & Goossens, A. (2011). Combinatorial biosynthesis in plants: A (p)review on its potential and future exploitation. *Natural Product Reports*, *28*, 1897 1916.

Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., et al. (2006). The path forward for biofuels and biomaterials. *Science*, *311*, 484 489.

Ro, D. K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, *440*, 940 943.

Roberts, A., Feng, H., & Pachter, L. (2013). Fragment assignment in the cloud with eXpress-D. *BMC Bioinformatics*, *14*, 358.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139 140.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*, 1086 1092.

Si, Y., Liu, P., Li, P., & Brutnell, T. P. (2014). Model-based clustering for RNA-seq data. *Bioinformatics*, *30*, 197 205.

Stermitz, F. R., Lorenz, P., Tawara, J. N., Zenewicz, L. A., & Lewis, K. (2000). Synergy in a medicinal plant: Antimicrobial action of berberine potentiated by 5′-methoxyhydnocarpin, a multidrug pump inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, *15*, 1433 1437.

Sung, Y. C., Lin, C. P., & Chen, J. C. (2014). Optimization of virus-induced gene silencing in *Catharanthus roseus*. *Plant Pathology*, *63*, 1159 1167.

Tarselli, M. A., Raehal, K. M., Brasher, A. K., Groer, C. E., Cameron, M. D., Bohn, L. M., et al. (2011). Synthesis of conolidine, a potent non-opioid analgesic for tonic and persistent pain. *Nature Chemistry*, *3*, 449 453.

Thomas, C. J., Rahier, N. J., & Hecht, S. M. (2004). Camptothecin: Current perspectives. *Bioorganic & Medicinal Chemistry*, *12*, 1585 1604.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7, 562 578.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*, 511 515.

van der Laan, M. J., & Pollard, K. S. (2003). Hybrid clustering of gene expression data with visualization and the bootstrap. Journal of Statistical Planning and Inference, 117, 275 303.

Wink, M. (1999). Plant secondary metabolism: Biochemistry, function, and biotechnology. In M. Wink (Ed.), *Biochemistry of plant secondary metabolism* (pp. 1 16). Shefield: Shefield Academic Press.

Zhao, J., & Dixon, R. A. (2009). MATE transporters facilitate vacuolar uptake of epicatechin 3′-O-glucoside for proanthocyanidin biosynthesis in Medicago truncatula and Arabidopsis. *Plant Cell Reports*, *21*, 2323 2340.

Partie 2: Les cytochromes P450 et leurs cytochromes P450 réductases

Article 2: Characterization of a second secologanin synthase isoform producing both secologanin and secoxyloganin allows enhanced de novo assembly of a Catharanthus roseus transcriptome

Article 3: Class II Cytochrome P450 reductase governs alkaloid biosynthesis in Madagascar periwinkle

2.1 Généralités sur les cytochromes P450

Les cytochromes P450 (dénommées CYP ou P450), forment une large superfamille d'hémoprotéines ubiquitaires que l'on retrouve chez la majorité des êtres vivants (procaryotes et eucaryotes). Leur nombre varie selon les espèces. Elles sont minoritaires chez *Saccharomyces cerevisiae*, qui ne possède que 3 gènes codant des P450 (Nelson et *al*., 2004b). On dénombre 55 gènes et 25 pseudogènes chez l'Homme (Werck-Reichhart et Feyereisen, 2000) mais plus de 244 gènes chez *Arabidopsis thaliana (*Bak et *al*., 2011). Cette spectaculaire diversification des P450 chez les plantes, peut s'expliquer par la nécessité de produire de nouvelles molécules leur permettant de s'adapter au milieu environnant. Les P450 sont très souvent recrutées pour la biosynthèse des métabolites secondaires des plantes et sont considérées comme l'une des familles clés de la genèse de la diversité de ces métabolites (Mizutani et Sato, 2011 ; Schuler, 2011 ; Renault et *al*., 2014). Ils sont notamment impliqués dans la biosynthèse de phytohormones, de molécules de défenses (phytoalexines), dans une variété de métabolites secondaires de la famille des terpènes, des alcaloïdes, des phénylpropanoïdes, et peuvent participer à la détoxification d'agents chimiques exogènes comme les herbicides (Werck-Reichhart et *al*., 2000; Nelson, 2011 ; Renault et *al*., 2014).

La réaction la plus couramment catalysée par les P450 est la réaction de mono-oxygénation avec l'insertion d'un atome d'oxygène en position aliphatique d'un substrat organique, tandis que l'autre atome d'oxygène est réduit pour donner une molécule d'eau. Dans la plupart des cas l'activation et le clivage du dioxygène O_2, nécessite un apport d'électrons provenant du NAD(P)H (NADPH ou NADH) transférés aux P450 par des protéines partenaires rédox parmi lesquelles figure les cytochromes P450 réductases (Jensen et Moller, 2010 ; Renault et *al*., 2014) dont il est question dans le chapitre suivant. Les réactions catalysées par les P450 sont très diverses. Si elles sont généralement associées à des

réactions d'hydroxylation (Werck-Reichhart et Feyereisen, 2000), elles peuvent cependant aussi catalyser d'autres réactions, comme des réactions d'époxidation, de déalkylation, de déhydratation et même de clivage de liaison carbone-carbone comme dans le cas de la SLS dont il est question ci-dessous au paragraphe 2 (Werck-Reichhart et Feyereisen, 2000 ; Schuler et Werck-Reichhart, 2003; Ortiz de Montellano, 2005).

Ces hémoprotéines sont constituées d'une protoporphyrine IX avec un noyau fer (hème) et d'une apoprotéine d'un poids moléculaire compris entre 45 et 60 kDa. Cette protoporphyrine est liée de façon « non covalente » à l'apoprotéine par l'intermédiaire d'une cystéine. Le noyau fer est lié aux quatres atomes d'azotes de la protoporphyrine et une cinquième liaison s'effectue avec le groupement thiolate (SH) de la cystéine de l'apoprotéine (figure 18).

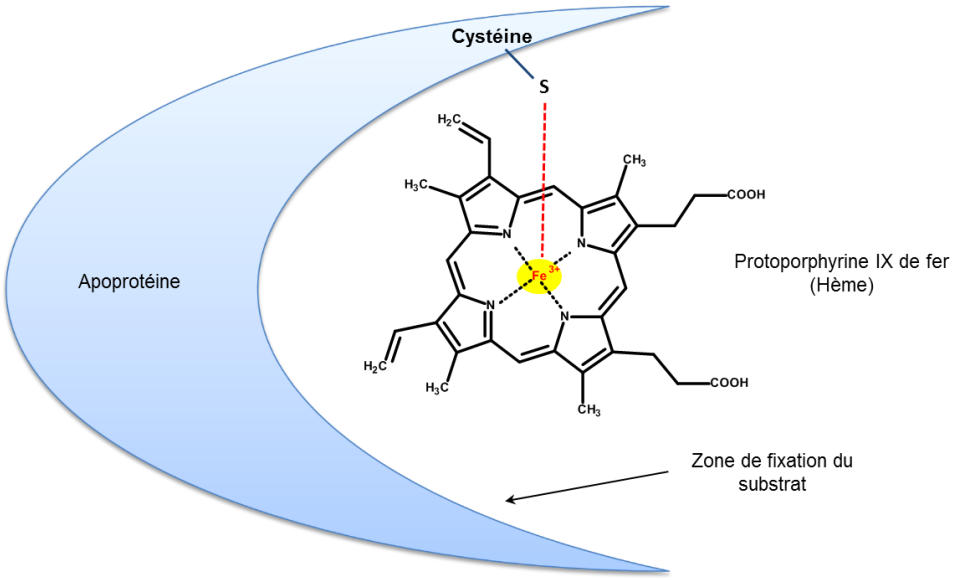


Figure 18 : Représentation schématique d'un cytochrome P450.

Contrairement aux P450 cytosoliques des procaryotes, la grande majorité des cytochromes P450 eucaryotes possèdent dans leur région N-terminal, une hélice transmembranaire leur permettant un ancrage sur la face cytosolique du RE (réticulum endoplasmique). Peu de données existent concernant la répartition des P450 au sein du RE chez les plantes. Chez l'homme 12 à 15% du RE est composé de cytochromes P450 pour un peu moins d'1% du poids total d'un hépatocyte (Ruckpaul et Rein, 1984). Par ailleurs, des études stoechiométriques (Finch et Stier, 1991 ; Schwarz, 1991) ont montrées qu'au sein de la

membrane du RE, les P450 se regrouperaient sous la forme d'unités hexamériques (ou octamériques) autour d'une réductase (un partenaire fournisseur en électrons).

Les mécanismes enzymatiques des cytochromes P450 sont encore mal caractérisés. Cependant les structures conservées de leur site actif permettent de décrire un mécanisme réactionnel commun en une succession de 8 étapes enzymatiques (Werck-Reichhart et Feyereisen, 2000) (figure 19).

Dans un premier temps, la fixation d'un substrat au niveau du site catalytique du cytochrome P450 entraine le départ d'une molécule d'H_2O fixée à l'atome de fer de l'hème. Dans l'étape suivante, le transfert d'un électron provenant du NADPH vers le noyau fer du P450 s'effectuant grâce à une NADPH cytochrome P450 réductase (CPR) entraine la réduction du noyau fer ferrique Fe(III) de l'hème en fer ferreux Fe(II). Ce dernier peut alors réagir avec l'oxygène moléculaire en fixant une molécule de dioxygène (O_2) et établir un complexe intermédiaire Fe(III)-O_2. Le transfert d'un second électron, par la CPR vers ce complexe vient réduire l'oxygène moléculaire et l'active. Une double protonation suivie du départ d'une molécule d'eau après clivage du dioxygène aboutit à la formation d'un ion oxoferryle. Le P450 va ensuite réaliser l'oxydation du produit en insérant un atome d'oxygène sur celui-ci (sur un groupement carboné). Le produit est ensuite libéré de la poche du site actif du P450 et le fer retrouve son état initial avec la fixation d'une molécule d'eau (Werck-Reichhart et Feyereisen, 2000 ; Munro et *al*., 2013) (figure 19).

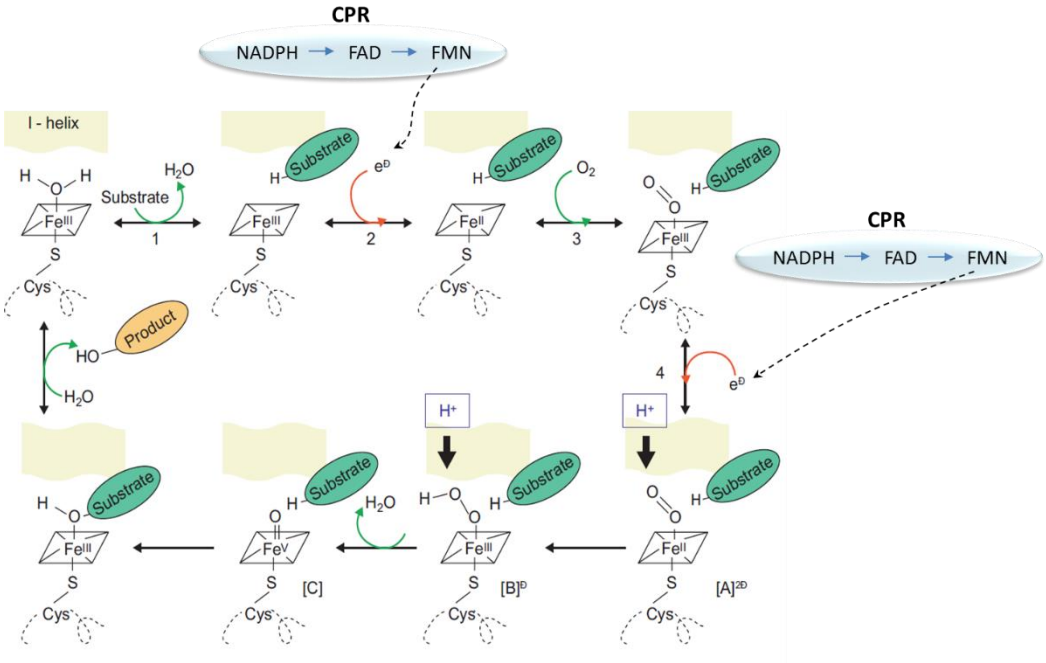


Figure 19 : Mécanisme enzymatique des cytochromes P450 chez les plantes (d'après Werck-Reichhart et Feyereisen, 2000)**.**

2.2 Caractérisation d'une nouvelle isoforme de cytochrome P450 chez *C. roseus*

La voie de biosynthèse des AIM de *C. roseus*, se compose de plus d'une trentaine d'étapes enzymatiques réparties au sein de 6 compartiments subcellulaires (Courdavault et *al*., 2014). La complexité de cette voie de biosynthèse repose sur son architecture tissulaire et subcellulaire impliquant un grand nombre, ainsi qu'une grande diversité des enzymes intervenant dans des réactions biochimiques clés de la synthèse des AIM (Courdavault et *al*., 2014). Parmi celles-ci, la famille des cytochromes P450 est bien représentée. Ainsi on dénombre quatre cytochromes P450 dans la voie menant du GPP à la sécologanine (G10H, IO, 7DLH, SLS) (figure 11, figure 14) (Collu et *al*., 2001 ; Irmler et *al*., 2000 ; Geu-Flores et *al*., 2012 ; Miettinen et *al*., 2014).

Un autre aspect ayant trait à la complexité du métabolisme des AIM tient au fait que l'on a découvert récemment qu'il peut exister des isoformes enzymatiques qui présentent des profils d'expression différents. C'est le cas des cytochromes P450 T16H1 et T16H2 catalysant la réaction d'hydroxylation de la tabersonine en position 16 (Besseau et *al*., 2013). Ces recherches ont montré que dans les feuilles, c'est le gène T16H2 qui possède un profil d'expression similaire aux autres gènes connus codant les enzymes de la voie de la vindoline (figure 12).

Dans l'article qui suit, nous rapportons la découverte d'une seconde isoforme enzymatique impliquée dans la formation de la sécologanine à partir de la loganine. Dénommée SLS2, par rapport à l'isoforme SLS1 identifiée antérieurement (Irmler et *al*., 2000), elle appartient également à la grande famille des P450. Nous avons montré que ces deux isoformes sont capables de produire de la sécologanine mais également de la sécoxyloganine lorsqu'elles sont exprimées dans la levure. L'isoforme SLS1 est majoritairement exprimée dans les racines tandis que SLS2 est exprimée dans les parties aériennes de la plante suggérant que ces deux isoformes, de façon complémentaire, permettent la production de sécologanine dans les organes différents de la plante.

La découverte des isoformes T16H et SLS nous a conduit à réactualiser les données transcriptomiques qui étaient disponibles dans les bases de données en vue de prédire s'il pouvait exister d'autres isoformes enzymatiques impliquées dans les étapes de la voie de biosynthèse des AIM. Ainsi, la seconde partie de l'article décrit la reconstruction d'un transcriptome qui se veut plus abouti que les précédents transcriptomes de *C. roseus* qui avaient été élaborés par trois consortiums (Medicinal Plant Genomic Resources, http://medicinalplantgenomics. msu.edu; PhytoMetaSyn, http://www.phytometasyn.ca; Cathacyc http://www.cathacyc.org) (Gongora-Castillo et *al*., 2012 ; Xiao et *al*., 2013 ; Van Moerkercke et *al*., 2013). Ce transcriptome optimisé est également la base sur laquelle s'appuient les analyses de corrélations d'expression de gènes en vue d'identifier de nouveaux gènes codant les enzymes non encore élucidées du métabolisme alcaloïdique de *C. roseus*. Il a, entre autre, servi à caractériser la tabersonine 3-réductase (T3R) dont il est question dans la partie 4 des résultats.

2.3 Généralités sur les cytochromes P450 réductases

Les P450 ne sont pas des enzymes autonomes et leur activité catalytique dépend strictement de partenaires rédox donneurs d'électrons, et principalement des NADPH cytochromes P450 réductases (CPR), auxquelles sont quelquefois associées des cytochromes b5 réductases (Renaul et *al*., 2014). La majorité des P450 chez les eucaryotes et notamment les plantes, sont des P450 de classe II associés à des CPR dépendantes du NADPH ancrées à la membrane du RE. La formation des complexes protéiques cytochrome P450-CPR s'effectue par des interactions ioniques entre résidus d'acides aminés (Hasemann et *al*., 1995) (figure 20 A).

Les CPR appartiennent à la famille des diflavines réductases (Jensen et Møller, 2010). Ils possèdent deux domaines renfermant chacun un groupement prosthétique séparés par un domaine « linker » (Wang et *al*., 1997 ; Jensen et Møller, 2010) assurant une certaine flexibilité à l'ensemble. Le premier domaine prosthétique renferme une flavine adénine dinucléotide (FAD) et possède une activité oxidoréductase. Le second domaine prosthétique est un domaine transporteur renfermant une flavine mononucléotide (FMN). Le flux d'électrons depuis le NADPH vers le dioxygène en transitant par l'hème du P450 s'effectue dans l'ordre suivant : NADPH → FAD → FMN → P450 → O2 (Jensen et Møller, 2010) (figure 20 A).

Ces complexes cytochrome P450-CPR peuvent parfois s'associer avec des cytochromes b5 favorisant ainsi la réduction des cytochromes P450 et leur stabilité vis-à-vis de leurs substrats (Paddon et *al*., 2013). Dans ce cas, une NADPH cytochrome b5 réductase assure le transfert du deuxième électron, le premier étant transféré via la CPR. (Schenkman et Jansson, 2003) (figure 20B).

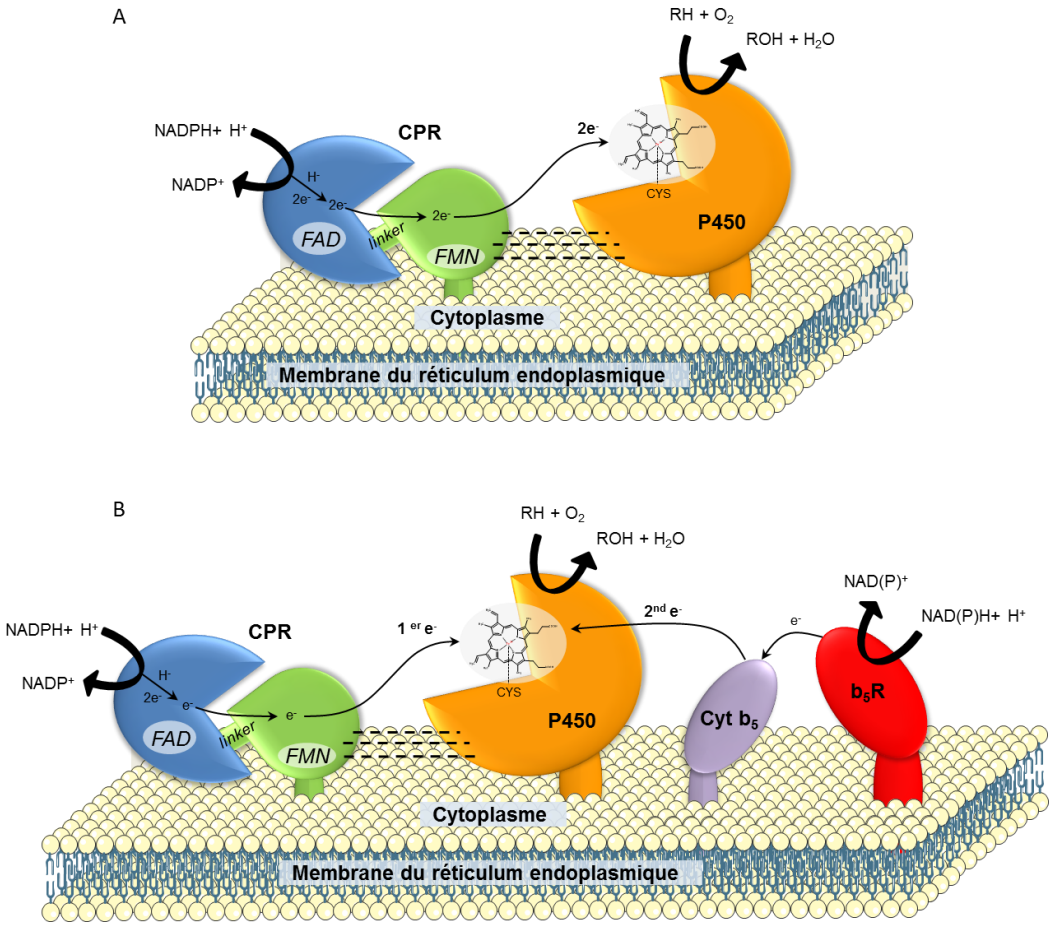


Figure 20 : Représentation schématique des complexes CPR-P450, CPR-P450-Cyt b5-b5R chez les plantes. (A) Mécanisme réactionnel d'un complexe CPR-P450 de classe II chez les plantes. Les deux électrons requis pour la réaction effectuée par le P450 lui sont transférés le plus souvent par une CPR. (B) Mécanisme réactionnel d'un complexe CPR-P450-Cyt b5-b5R. Dans quelques cas, le transfert des deux électrons proviennent à la fois d'une CPR et d'un complexe formé par un cytochrome b5 et d'une réductase b5.

2.4 Caractérisation des CPR de *C. roseus*

A l'origine, un ADNc codant une CPR avait été identifiée chez *C. roseus*. Les auteurs ont montré que son activité était liée à celle des activités G10H et C4H (cinnamate 4-hydroxylase) et que par conséquent elle devait être impliquée, dans le métabolisme des AIM et celui des phénylpropanoïdes. (Meijer et *al*., 1993).

Basée sur une analyse des ressources transcriptomiques, l'article qui suit présente l'identification d'une seconde CPR nommée CPR1, la CPR originale étant renommée CPR2 ainsi qu'une CPR-like qui est en fait une diflavine réductase (DFR). La caractérisation de ces 2 CPR a été réalisée de manière approfondie. Elle repose sur des études :

- de corrélation d'expression de gènes
- de localisation tissulaire et subcellulaire
- d'activités enzymatiques (couplées aux activités de cytochromes P450)
- d'interactions avec les cytochromes P450 (techniques BiFC)
- d'extinction de gènes (technique VIGS) corrélé au taux des AIM

Il en ressort que *C. roseus* possède 2 CPR vraies et que CPR2 est la réductase préférentiellement associée au métabolisme spécialisé dont celui des AIM alors que CPR1 est la réductase plutôt associée à des fonctions cellulaires de base.

Dugé de Bernonville *et al. BMC Genomics* (2015) 16:619
DOI 10.1186/s12864-015-1678-y

BMC Genomics

RESEARCH ARTICLE Open Access

Characterization of a second secologanin synthase isoform producing both secologanin and secoxyloganin allows enhanced *de novo* assembly of a *Catharanthus roseus* transcriptome

Thomas Dugé de Bernonville[1†], Emilien Foureau[1†], Claire Parage[1†], Arnaud Lanoue[1], Marc Clastre[1], Monica Arias Londono[1,2], Audrey Oudin[1], Benjamin Houillé[1], Nicolas Papon[1], Sébastien Besseau[1], Gaëlle Glévarec[1], Lucia Atehortùa[2], Nathalie Giglioli-Guivarc'h[1], Benoit St-Pierre[1], Vincenzo De Luca[3], Sarah E. O'Connor[4] and Vincent Courdavault[1*]

Abstract

Background: Transcriptome sequencing offers a great resource for the study of non-model plants such as *Catharanthus roseus*, which produces valuable monoterpenoid indole alkaloids (MIAs) via a complex biosynthetic pathway whose characterization is still undergoing. Transcriptome databases dedicated to this plant were recently developed by several consortia to uncover new biosynthetic genes. However, the identification of missing steps in MIA biosynthesis based on these large datasets may be limited by the erroneous assembly of close transcripts and isoforms, even with the multiple available transcriptomes.

Results: Secologanin synthases (SLS) are P450 enzymes that catalyze an unusual ring-opening reaction of loganin in the biosynthesis of the MIA precursor secologanin. We report here the identification and characterization in *C. roseus* of a new isoform of SLS, SLS2, sharing 97 % nucleotide sequence identity with the previously characterized SLS1. We also discovered that both isoforms further oxidize secologanin into secoxyloganin. SLS2 had however a different expression profile, being the major isoform in aerial organs that constitute the main site of MIA accumulation. Unfortunately, we were unable to find a current *C. roseus* transcriptome database containing simultaneously well reconstructed sequences of SLS isoforms and accurate expression levels. After a pair of close mRNA encoding tabersonine 16-hydroxylase (T16H1 and T16H2), this is the second example of improperly assembled transcripts from the MIA pathway in the public transcriptome databases. To construct a more complete transcriptome resource for *C. roseus*, we re-processed previously published transcriptome data by combining new single assemblies. Care was particularly taken during clustering and filtering steps to remove redundant contigs but not transcripts encoding potential isoforms by monitoring quality reconstruction of MIA genes and specific SLS and T16H isoforms. The new consensus transcriptome allowed a precise estimation of abundance of SLS and T16H isoforms, similar to qPCR measurements.

Conclusions: The *C. roseus* consensus transcriptome can now be used for characterization of new genes of the MIA pathway. Furthermore, additional isoforms of genes encoding distinct MIA biosynthetic enzymes isoforms could be predicted suggesting the existence of a higher level of complexity in the synthesis of MIA, raising the question of the evolutionary events behind what seems like redundancy.

Keywords: *Catharanthus roseus*, Transcriptome assembly, Isoform, Secologanin synthase, Secoxyloganin

* Correspondence: vincent.courdavault@univ-tours.fr
†Equal contributors
[1]Université François-Rabelais de Tours, EA2106 "Biomolécules et Biotechnologies Végétales", UFR Sciences et Techniques, 37200 Tours, France
Full list of author information is available at the end of the article

Background

Monoterpenoid Indole Alkaloids (MIAs) constitute a remarkable class of specialized metabolites, with a huge chemical diversity, and a source of several active compounds, including important pharmacophores. Some of the most active anticancer drugs are based on this type of skeleton including camptothecans and Vinca alkaloids. The later compounds are present in minute amounts in the leaves of the Madagascar periwinkle, *Catharanthus roseus*, and result from a complex metabolic pathway, which is the target of expanding research efforts in the phytochemical genomic era [1, 2].

MIAs stem from a unique polyvalent skeleton named strictosidine. This central precursor is the condensation product of a tryptophan-derived amine coupled to an extensively modified monoterpenoid moiety (Fig. 1). While tryptamine is derived from tryptophan by a single reaction catalyzed by tryptophan decarboxylase (TDC) [3], the assembly of the monoterpene secoiridoid moiety, requires several reactions to convert the methyl-erythritol phosphate (MEP) pathway-derived monoterpenoid skeleton into secologanin (Fig. 1) [4].

Recently, the elusive reaction scheme of secologanin biosynthesis has been elucidated in *C. roseus*. In the plastid-localized MEP pathway, glyceraldehyde 3-phosphate (GAP) and pyruvate are converted into the universal isoprenoid precursors, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), through seven enzymatic reactions. The subsequent conversion of these primary metabolites into secologanin, by the monoterpene secoiridoid pathway, requires ten more enzymes.

First, the prenyl-transfer of IPP on DMAPP by geranyl diphosphate synthase (GPPS) [5], is followed by formation of the monoterpene geraniol by geraniol synthase (GES) [6] (Fig. 1). Subsequently, geraniol is hydroxylated into the diol 10-hydroxygeraniol (alternative nomenclature:

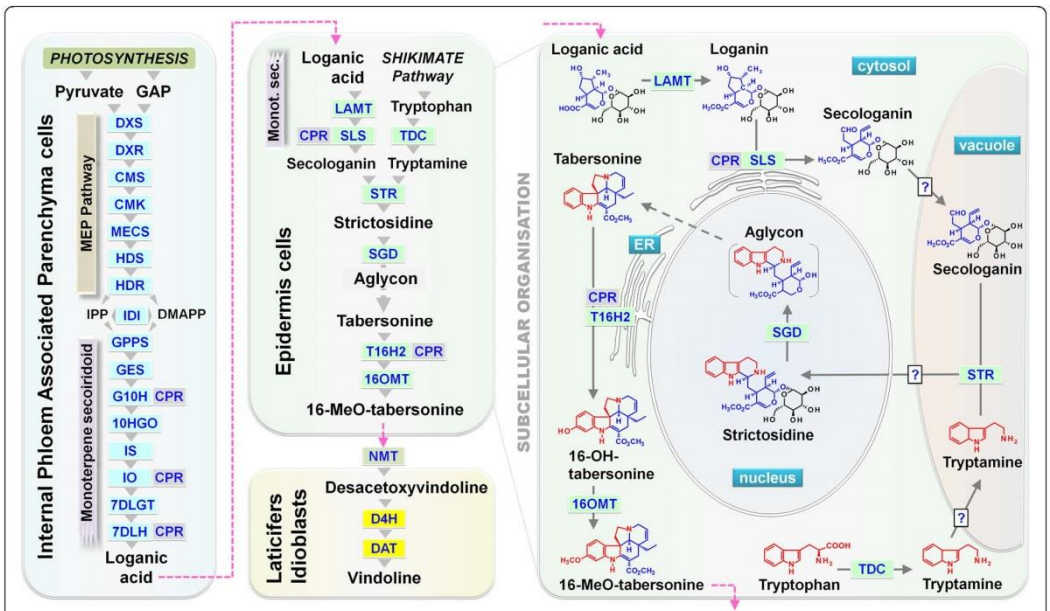


Fig. 1 The biosynthetic pathway of MIA in *C. roseus* leaves. Simplified representation of the MIA biosynthesis in *C. roseus* highlighting the subcellular organization of the central steps of the pathway. Known single enzymatic steps in each cell type are indicated by grey arrows and abbreviation of enzyme names. Broken grey arrows and broken pink arrows indicate unknown enzymatic steps and metabolite translocation, respectively. DXS, 1-deoxy-D-xylulose-5-phosphate (DXP) synthase; DXR, DXP reductoisomerase; CMS, 4-(cytidine 5′-diphospho)-2*C*-methyl-D-erythritol (CM) synthase; CMK, CM kinase; MECS, 2*C*-methyl-D-erythritol-2,4-cyclodiphosphate (MEC) synthase; HDS, hydroxymethylbutenyl 4-diphosphate (HD) synthase; HDR, HD reductase; IDI, isopentenyl diphosphate isomerase; GPPS, geranyl diphosphate synthase; GES, geraniol synthase; G10H (CYP7686), geraniol 10-hydroxylase; CPR, cytochrome P450-reductase; 10HGO, 10-hydroxygeraniol oxidoreductase; IO, iridoid oxidase; IS, iridoid synthase; 7DLGT, 7-deoxyloganetic acid glucosyltransferase; 7DLH, 7-deoxyloganic acid 7-hydroxylase; LAMT, loganic acid *O*-methyltransferase; SLS (CYP72A1), secologanin synthase; TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGD, strictosidine β-glucosidase; T16H2 (CYP71D351), tabersonine 16-hydroxylase 2; 16OMT, 16-hydroxytabersonine *O*-methyltransferase; NMT, 16-methoxy-2,3-dihydrotabersonine N-methyltransferase; D4H, desacetoxyvindoline 4-hydroxylase; DAT, deacetylvindoline 4-*O*-acetyltransferase. DMAPP, dimethylallyl diphosphate; GAP, glyceraldehyde 3-phosphate; IPP, isopentenyl diphosphate

8-hydroxygeraniol) and further oxidized into 10-oxogeraniol by the bifunctional P450 CYP76B6 named geraniol 10-hydroxylase (G10H [7]; also renamed G8O; [8]). The third oxidation into the dialdehyde 10-oxogeranial requires a specific alcohol dehydrogenase, 10-hydroxygeraniol oxidoreductase (10HGO, also named 8HGO [9]). Thereafter, iridoid synthase (IS) performs the key step for the assembly of the iridoid heterocyclic ring structure. IS uses 10-oxogeranial and probably couples an initial NAD(P)H-dependent reduction with a subsequent cyclization step to form the ring structure of *cis-trans*-nepetelactol [10]. A second P450 enzyme CYP76A26, 7-deoxyloganetic acid synthase, also named iridoid oxidase (IO), catalyzes a key 3-step oxidation of *cis-trans*-nepetelactol to form 7-deoxyloganetic acid [9, 11]. The latter compound is linked to a glucosyl residue by a substrate specific UDP-glucose glycosyltransferase (UGT), 7-deoxyloganetic acid glucosyltransferase (7DLGT) [9, 12]. The resulting product 7-deoxyloganic acid is hydroxylated at the C-7 position by 7-deoxyloganic acid 7-hydroxylase (7DLH, CYP72A224) [9, 13] to yield loganic acid, which is methylated into loganin by a S-adenosyl-L-methionine: loganic acid methyl transferase (LAMT) [14]. Finally, the ring-opening reaction of loganin in the biosynthesis of secologanin is catalyzed by the fourth P450 of this pathway, secologanin synthase (SLS, CYP72A1) [15]).

Following assembly of the monoterpenoid and indole precursors, formation of the MIA basic skeleton is initiated by strictosidine synthase (STR), which catalyzes the stereospecific condensation of tryptamine with secologanin to form 3α (*S*)-strictosidine [16, 17]. Strictosidine β-D-glucosidase (SGD), catalyzing deglucosylation of strictosidine, produces the last common intermediate in the biosynthesis of the thousand existing MIAs, since the resulting aglycone is the starting point for many different skeletons [18, 19]. The later conversion of the strictosidine aglycone into tabersonine has not been elucidated, but most steps in the final conversion of tabersonine into vindoline have been described. Following 16-methoxylation of tabersonine, performed by the sequential action of tabersonine 16-hydroxylase (T16H) [20-22] and 16-hydroxytabersonine O-methyltransferase (16OMT) [23, 24], 16-methoxytabersonine undergoes an uncharacterized hydration reaction followed by N-methylation, hydroxylation and acetylation carried out by 16-methoxy-2,3-dihydrotabersonine N-methyltransferase (NMT) [25-27], desacetoxyvindoline-4-hydroxylase (D4H) [28, 29] and deacetylvindoline-4-*O*-acetyltransferase (DAT) [30], respectively.

The MIA biosynthetic pathway displays one of the most complex and elaborated forms of compartmentalization described to date (Fig. 1). It was shown to require the coordinated implication of at least four different cell types, implying specific intercellular translocations of metabolite whose identifications are underway. The biosynthesis of MIAs is initiated within internal phloem associated parenchyma (IPAP) cells which host the initial steps leading to secologanin, i.e. the whole MEP pathway together with the eight first reactions of monoterpene-secoiridoid pathway [1, 2, 6, 9-12, 31-33]. The central steps occur in leaf epidermis with conversion of loganic acid into secologanin, after its translocation from IPAP cells. This latter is next conjugated to tryptamine to yield strictosidine, whose corresponding aglycone serves as the primary precursor for complex alkaloids [14, 15, 22, 24, 34-36]. Following translocation of 16-methoxytabersonine or a downstream intermediate, vindoline biosynthesis is completed in laticifers and idioblast cells hosting D4H and DAT activities [34, 37]. In addition, all these biosynthetic steps are marked by a complex subcellular distribution pattern: soluble cytosolic enzymes (TDC, IS, 7DLGT, LAMT, D4H and DAT), endoplasmic reticulum anchored enzymes (G10H, IO, 7DLH, SLS, T16H1 and T16H2), plastidial enzymes (MEP pathway enzymes, GPPS, GES), vacuolar enzymes (STR and PEX1) and nuclear SGD [5, 6, 9, 10, 22, 35-39]. However, despite the existence of multiple intra- and intercellular transports, only one transporter of the MIA pathway has been characterized to date, TPT2, that mediates the specific excretion of catharanthine at the leaf epidermis [40].

Recently, high throughput sequencing approaches (RNA-seq) have been used to provide an access to full *C. roseus* transcriptomes and to help in identifying new genes of the MIA biosynthetic pathway. Such transcriptomes were released by three main initiatives, the Medicinal Plant Genomics Resource (MPGR) [41], Cathacyc and ORCAE [42] (ccOrcae) and Phytometasyn (PMS) [43], as well as other independent studies [44, 45]. These data were generated from the sequencing of libraries prepared from whole-organs and specific experimental conditions. The resulting sequences have been used in orthology and gene clustering allowing the identification of new genes, such as 7DLH and 7DLGT (reviewed in [2]). However, new results have pinpointed the involvement of multiple enzyme isoforms in this highly compartmentalized pathway of MIA biosynthesis, adding thus an additional layer of complexity. Indeed, we have recently described two isoforms of T16H (T16H1 and T16H2), encoded by two distinct genes displaying different tissue-specific expression patterns [22]. However, it should be noted that the currently available *C. roseus* transcriptome resources failed to correctly integrate these isoforms, which could result from improper *de novo* assembly or insufficient sequencing depth of samples. Hence, browsing the current *C. roseus* transcriptome resources might miss important information, highlighting the need for a more exhaustive transcriptome.

Based on this ascertainment, the objective of the present study was to generate a consensus transcriptome

containing an exhaustive library of *C. roseus* transcripts with expression level information. Different strategies have been previously employed to generate transcriptome assemblies for non-model animal and plant species. Most of them rely on the combination of assemblies resulting from different assemblers such as Trinity [46], Oases [47], TransAbyss [48] and SOAPdenovo-Trans [49] with eventually different *k*-mer lengths since this criteria is expected to bypass the uneven distribution of transcript abundance [50]. Such strategy was successfully conducted for *Anas platyrhynchos domestica* [51] and *Nicotiana benthamiana* [52], for which assemblies were performed on a unique library, but also for wheat, with assemblies performed on a mix of 4 libraries [53]. In each case, redundancy caused by the merging of different assemblies was decreased by using clustering tools such as CD-HIT-EST [54] or TGICL [55]. In *C. roseus*, a recent study compared assemblies generated by Abyss, Velvet and Oases running with different *k*-mer values on a mix of 3 libraries prepared from different organs and merged the best result with the previous assembly prepared by MPGR [44]. In such a case, mixing samples is expected to increase the possibility to find lowly expressed genes and isoforms, due to the diversity of reads sequenced from diverse tissues/experimental conditions, for instance. However, combining libraries prepared from different sources, such as plant cultivars, may also generate more potential isoforms due to genetic polymorphisms. In the present study, we built a new consensus transcriptome for *C. roseus* using already published data. We generated assemblies for every available sample to take advantage of the diversity of tissues/experimental conditions, combined them and tested different thresholds to cluster homologous contigs. Special attention was taken to reduce the redundancy without affecting transcript quality. Optimization of this consensus assembly was performed by monitoring reconstruction quality of all MIA biosynthetic genes, with a particular emphasis on the two previously described T16H isoforms [22] and on a newly identified SLS isoform whose functional validation is also depicted. The reconstruction of such a *C. roseus* consensus transcriptome is expected to facilitate the identification of the missing MIA biosynthetic enzymes by studying the clustering of gene expression for instance, but also the characterization of new isoforms whose existence could be predicted through this work.

Results and discussion

Identification and characterization of a second SLS isoform

While amplifying the coding sequence of SLS (CYP71A1, Genbank accession number L10081) [14], sequencing of the PCR products revealed the presence of a second putative isoform exhibiting 96 % identity with the original SLS isoform. Interrogation of the *C. roseus* transcriptomic databases (Medicinal Plant Genomics Resource, CathaCyc/Orcae and Phytometasyn) led to the identification of identical but partial sequences confirming thus the existence of this new SLS sequence that has been recently deposited to Genbank under accession number KF415117. The corresponding P450 also displayed a high level of identity (97 %) with the first SLS isoform (Additional file 1: Figure S1) suggesting that it could also catalyze the oxidative ring cleavage of loganin to produce secologanin. To test this hypothesis. The original and the new putative SLS isoforms were individually expressed in the *Saccharomyces* WAT11 strain that overexpresses the Arabidopsis NADPH P450 reductase [56]. Crude extracts of galactose-induced yeasts transformed with the pYeDP60 empty control vector or the pYeDP60 expressing each P450 were subsequently incubated with NADPH, H^+ and loganin, and analyzed by UPLC-MS (Fig. 2). While no modification of loganin occurred with the empty vector crude extract, a conversion of loganin into secologanin was observed with the crude extract of each enzyme. This established that the putative SLS isoform truly corresponds to a new SLS isoform, named SLS2 as reference to CYP71A1, renamed SLS1.

Interestingly, we noted that both SLS1 and SLS2 also convert loganin into a more polar compound identified as secoxyloganin according to UV and MS spectra and a comparison with a pure authentic standard (Fig. 2; Additional file 2: Figure S2). By contrast, this product was not produced using the empty vector crude extract suggesting that it results from a reaction catalyzed by SLS. For SLS1 and SLS2, time course reactions showed a decrease of the loganin content accompanied by the formation of both secologanin and secoxyloganin (Fig. 3). Since secoxyloganin corresponds to the acidic form of secologanin, it may result from the oxidation of the aldehyde function of secologanin. Therefore, we tested the capacity of both SLS1 and SLS2 to convert secologanin into secoxyloganin. While no formation of secoxyloganin was monitored by incubating secologanin with the empty vector crude extract, both SLS1 and SLS2 directly produce secoxyloganin from secologanin in a stoichiometric manner at least during the early times of the reaction (Fig. 2, Fig. 3). As a consequence, these results suggest that SLS1 and SLS2 not only catalyze the oxidative ring cleavage of loganin to produce secologanin but also perform the oxidation of secologanin into secoxyloganin. Besides G10H and IO, SLS1 and SLS2 constitute the third type of P450 from the seco-iridoid pathway performing more than one catalytic reaction [8, 11]. Interestingly, the additional reaction catalyzed by SLS1 and SLS2 is similar to the third oxidation performed by IO to generate 7-deoxyloganetic acid, suggesting that regiospecific multioxidation is rather common to P450s acting in secoiridoid biosynthesis. The occurrence of sequential oxidations has been reported for several P450s [57] but the dissociation of intermediates is still a question of debate since it ranges

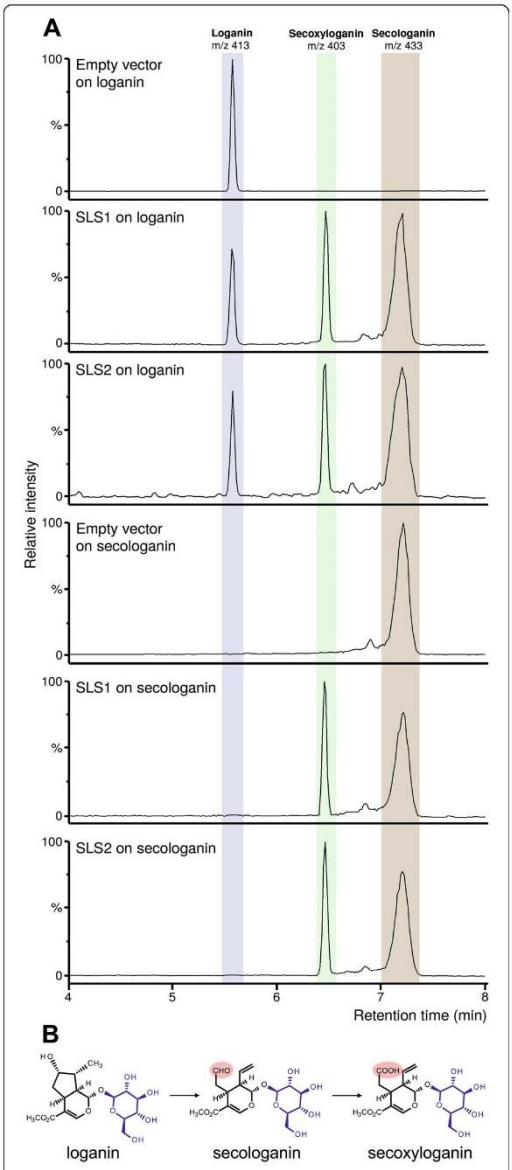


Fig. 2 SLS1 and SLS2 catalyze the conversion of loganin in secologanin and secoxyloganin. **a** LC-MS chromatograms using selected ion monitoring (loganin, mass-to-charge ratio 413; secologanin, mass-to-charge ratio 433; secoxyloganin, mass-to-charge ration 403) of the reaction products of yeast extracts from cell cultures expressing either SLS1 or SLS2 or containing the empty pYeDP60 vector, incubated with loganin or secologanin. **b** Schematic reaction catalyzed by SLS1 and SLS2 highlighting the aldehyde to acid conversion

from an absence of dissociation for P450 11B2 [58] to a dissociation of 85 % for P450 2C11 [59]. In the absence of pulse-chase experiments, we are not able to propose a reaction scheme for both SLS1 and SLS2 concerning secologanin release.

While important amounts of secologanin can be measured in the different organs of *C. roseus* [22] we never detected the presence of secoxyloganin (data not shown), raising the question of the physiological signification of this compound production but also suggesting that secologanin could be release from the SLS catalytic site during the sequential oxidation process. Secoxyloganin is an acidic compound derived from secologanin that is no longer able to be condensed with tryptamine by STR in the vacuole, due to the absence of the aldehyde function. If secoxyloganin formation occurs *in vivo*, the resulting depletion of the secologanin pool would be deleterious for the subsequent synthesis of MIAs. Although we cannot exclude that additional enzymes might convert secoxyloganin back to secologanin, the subcellular compartmentation of secologanin biosynthesis may also limit secoxyloganin formation *in planta*. Subcellular localization studies showed that SLS2 is located to the endoplasmic reticulum as previously observed for SLS1 [36] (Fig. 4). Since both SLS1 and SLS2 are anchored to this subcellular compartment and release their product in the cytosol, active transport of secologanin to the vacuolar compartment with high efficiency might allow its import before additional oxidation by SLS1 or SLS2. This would be a direct and interesting consequence of the complex subcellular organization of the MIA biosynthetic pathway regarding regulation of the metabolic flux.

Besides T16H and IS, SLS corresponds to the third type of enzymes from the MIA biosynthetic pathway displaying more than one isoform [22, 60]. While T16H1 and T16H2 have distinct organ specific-roles in MIA biosynthesis, IS4 and IS5 display somewhat redundant functions. To gain insight into the respective involvement of the two SLS isoforms, SLS1 and SLS2 gene expression was measured in the main *C. roseus* organs (Fig. 5). SLS2 expression was detected in all the tested organs and reached maxima in those directly associated with MIA biosynthesis including roots, flower buds and leaves. By contrast, SLS1 transcripts were barely detectable in all organs except in roots where expression was three-fold lower than SLS2. It is interesting to note that SLS1 was initially characterized from a cell suspension culture cDNA library [15]. Therefore, these results suggest that SLS1 and SLS2 can contribute concomitantly to secologanin biosynthesis in roots while SLS2 can be the prominent isoform of secologanin biosynthesis in the aerial parts of the plant.

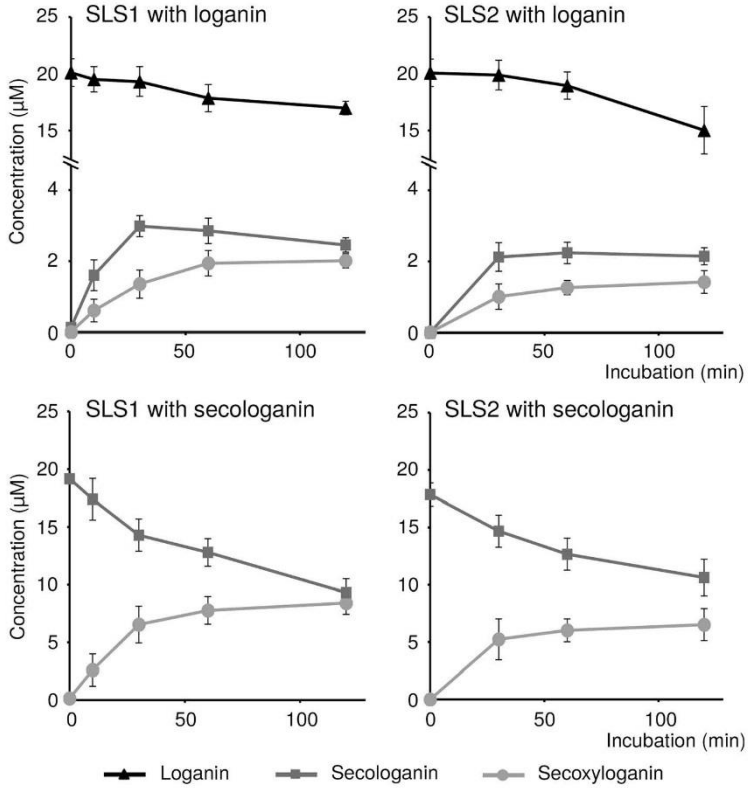


Fig. 3 Time course analyses of the reaction catalyzed by SLS1 and SLS2. Proteins extracts from yeast cells expressing either SLS1 or SLS2 were incubated at 30 °C with loganin or secologanin. Formation of the resulting products was monitored by LC-MS analysis. Black triangles, loganin; dark grey squares, secologanin; light circles, secoxyloganin

Reconstruction of individual MIA genes in current assemblies and new single assemblies generated with Trinity

Our functional approach clearly demonstrates the existence of MIA enzyme isoforms potentially displaying specific catalytic parameters and/or expression patterns. The identification of new enzyme isoforms is of importance and may notably help in isolating the most efficient enzymes that could be used in synthetic biology approaches as recently highlighted for strictosidine production [61]. RNA-seq data for *C. roseus* provide an important opportunity to retrieve such isoforms by analyzing sequences sharing high identities. In addition, combining gene expression levels from different experimental conditions will improve the quality

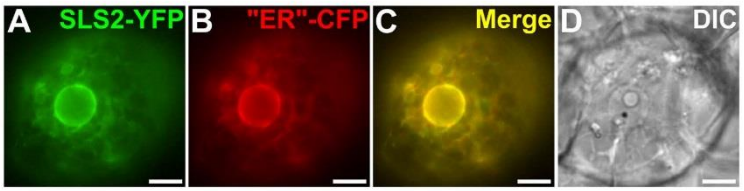


Fig. 4 SLS2 is located to the endoplasmic reticulum. *C. roseus* cells were transiently transformed with the SLS2-YFP-expressing vector (SLS2-YFP; (**a**) in combination with the plasmid expressing an ER-CFP marker ("ER"-CFP; (**b**). Colocalization of the two fluorescence signals appears on the merged images (**c**). Cell morphology (**d**) was observed with differential interference contrast (DIC). Bars = 10 µm

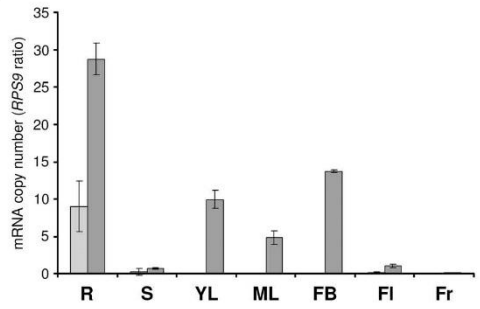


Fig. 5 Expression of SLS1 and SLS2 in *C. roseus* organs. SLS1 (light gray bars) and SLS2 (dark gray bars) transcript levels were determined by real-time RT-PCR analyses performed on total RNA extracted from *C. roseus* organs and transcript copy numbers were normalized using CrRPS9. R, Roots; S, stems; YL, young leaves; ML, mature leaves; FB, flower buds; Fl, flowers; Fr, fruits

of transcript expression patterns which allow conducting gene clustering analyses. However, this requires a correct reconstruction of each isoform.

We performed a detailed inspection of the current transcriptomic resources, available from Medicinal Plant Genomic Resources [41], PhytoMetaSyn [43] (with Illumina reads or with 454 reads), Cathacyc/Orcae [42] and a newly prepared transcriptome by a NIPGR research team [44]. The corresponding datasets will be thereafter named mpgrCra, PMSIllu (Illumina reads), PMS454 (454 reads), ccOrcae and NIPGR, respectively (Additional file 3: Table S1). Our analysis revealed that correct reconstruction of MIA genes was not systematic. Reference sequences of MIA genes available on NCBI were blasted against those assemblies and the bitscore of best hit was compared to that of an ideal reconstruction, i.e. the bitscore of the reference sequence against itself. On the whole, quality of reconstruction was quite unequal between assemblies (Fig. 6). PMS454 and ccOrcae assemblies displayed the best sequences while PMSIllu was of weaker quality (see for example 10HGO, 16OMT, CMK, HDS, IDI1 and IO). NIPGR and mpgrCra assemblies were quite similar in content, probably due to the construction design of the NIPGR assembly (independent libraries assembled and mixed with mpgrCra before filtering). Classically, discrepancies between assemblies might be due to natural polymorphisms, sequencing and/or reconstruction errors. When looking at very well reconstructed genes such as 7DLH and LAMT, it appeared that small differences are related to single-base variations. For 7DLH, such a variation was

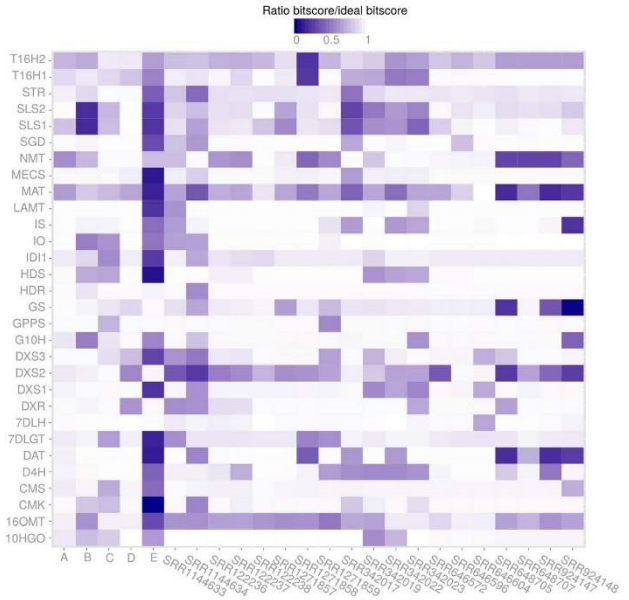


Fig. 6 Reconstruction quality of MIA genes in different assemblies. Current resources (A = ccOrcae (Smartcell), B = mpgrCra (Medicinal Plant Genomics Resource), C = NIPGR, D = PMS454 (PhytoMetaSyn), E = PMSIllu (PhytoMetaSyn) and new assemblies (19 SRR with PE sequencing design) were used as databases to identify homologs of MIA genes, and the resulting bitscore obtained by BLAST was compared to that of an ideal sequence (bitscore of the reference sequence against itself, i.e. bitscore ratio = 1)

observed at the position 564 of the reference sequence (KF415115) in the two assemblies mprgCra and NIPGR where a C was changed to A. This variation could be a true SNP (Single Nucleotide Polymorphism) as the reference sequence was obtained with another cultivar (Little Delicata). Concerning isoforms of T16H (T16H1 and T16H2) and SLS (SLS1 and SLS2), it appeared that current assemblies failed to present high quality sequences of the 4 transcripts (T16H1, T16H2, SLS1 and SLS2) simultaneously. For example, while both SLS isoforms were well reconstructed (bitscore/ideal bitscore >0.99) in PMS454, it was not the case for T16H1 (0.78). The best reconstructions of T16H1 and T16H2 were found in mpgrCra (0.92 for T16H1) and NIPGR (0.92 for T16H2), respectively. This result prompted us to try new assembly strategies in order to produce a more complete transcriptome.

To this aim, we next compared new single assemblies of each sample prepared with the Trinity pipeline [46, 62], since the resulting diversity is expected to reveal more MIA biosynthetic genes and isoforms. For this approach, a total of 19 samples were retained because of their paired-end sequencing design as it is expected to improve the quality of reconstruction (Additional file 3: Table S1). Before running Trinity, read content in each sample was normalized by analyzing *k*-mer content (*k*-mer size = 25, maximum coverage = 30) to remove reads being overrepresented or displaying abnormal distribution and their quality was assessed with FastQC (Additional file 4: Table S2). Again, a wide range of reconstruction quality was observed. All MIA biosynthetic genes had a high reconstruction quality (>0.85, minimal highest quality observed T16H2 in SRR1271857 with 0.88) in at least one single assembly. However, as observed for current resources, we did not observe the simultaneous presence of SLS and T16H isoforms, despite the use of mixed libraries (SRR122236, SRR122237 and SRR122238). Interestingly, the base variation within 7DLH described above was no more observable in the alignment of best hits (data not shown) in each single assembly with the 7DLH reference sequence. Hence the variations observed in NIPGR and mpgrCra assemblies probably result from reconstruction errors.

Preparation of a consensus transcriptome for *C. roseus*

The quality of reconstruction of MIA biosynthetic genes in single assemblies suggests that raw resources might contain enough information to construct a consensus transcriptome since most of genes displayed a good reconstruction (>0.8) in at least one sample (Fig. 6). Because samples were sequenced at different depth, it may be possible that partial transcripts were also reconstructed in single assemblies. Therefore, two strategies based on the combination of samples were then tested to correctly assemble isoforms and MIA biosynthetic genes (Fig. 7): we first tried to combine all reads and generate a new assembly, while, in the second approach, the individual transcriptomes were combined and the resulting dataset clustered (using CD-HIT-EST) and subsequently filtered (to ensure the removal of clusters with weak representation by reads and in single assemblies).

In the first approach, reads from each paired-end samples were normalized (k-mer size = 25, maximum coverage = 30; Additional file 4: Table S2), combined and the resulting read set was normalized again. The resulting transcriptome contained 353,245 transcripts with a N50 value of 2,036 nt (median 564 nt). This important number of transcripts suggested the presence of fragmented transcripts and CD-HIT-EST was employed with increasing thresholds of sequence identity (90 to 100 %). However, neither this transcriptome nor its corresponding clustered subsets contained high quality sequences (Additional file 5: Figure S3).

In the second approach, we merged all single Trinity assemblies (see above) and ran different filtering procedures in order to decrease the resulting redundancy without altering transcript quality. A total of 3,145,245 contigs from single assemblies were then combined. This allowed combining very high quality transcripts within one new assembly which however, contained an evident redundancy due to the merging procedure. Indeed, the resulting large dataset is expected to cover a large number of isoforms. These isoforms may be real transcripts such as isoforms of SLS and T16H that have to be differentiated, or alleles of different cultivars, which should be integrated into a reference sequence. Running CD-HIT-EST with different sequence identity thresholds succeeded in combining contigs into clusters (Fig. 8a). This algorithm clusters similar sequences and uses one of them as a representative one. A weak decrease in sequence quality was observed with lower identity thresholds for 16OMT (bitscore/ideal bitscore in non-clustered dataset, 0.94; at clustering threshold 98 %, 0.92), IS and SLS2 for clustering thresholds lower than 0.94 (Fig. 8b). The transcript with lowest quality was T16H2 (0.88 for clustering threshold above 0.96). However, its quality was quite similar with that of the best reconstruction in current resources (0.92 in NIPGR). Two other genes, IDI1 and STR did not display ideal reconstruction, according to the reference sequence. IDI1 was slightly better reconstructed in PMS454 and STR was better in NIPGR and PMS454. The origin of those discrepancies are unclear but might have been caused by a higher polymorphism, leading to a different reference sequence in comparison to the representative clusters obtained here. According to the quality of MIA biosynthetic gene reconstruction, we further retained the clustered dataset obtained with a sequence identity threshold of 97 %. This threshold should be permissive enough to combine alleles differing by only few SNPs.

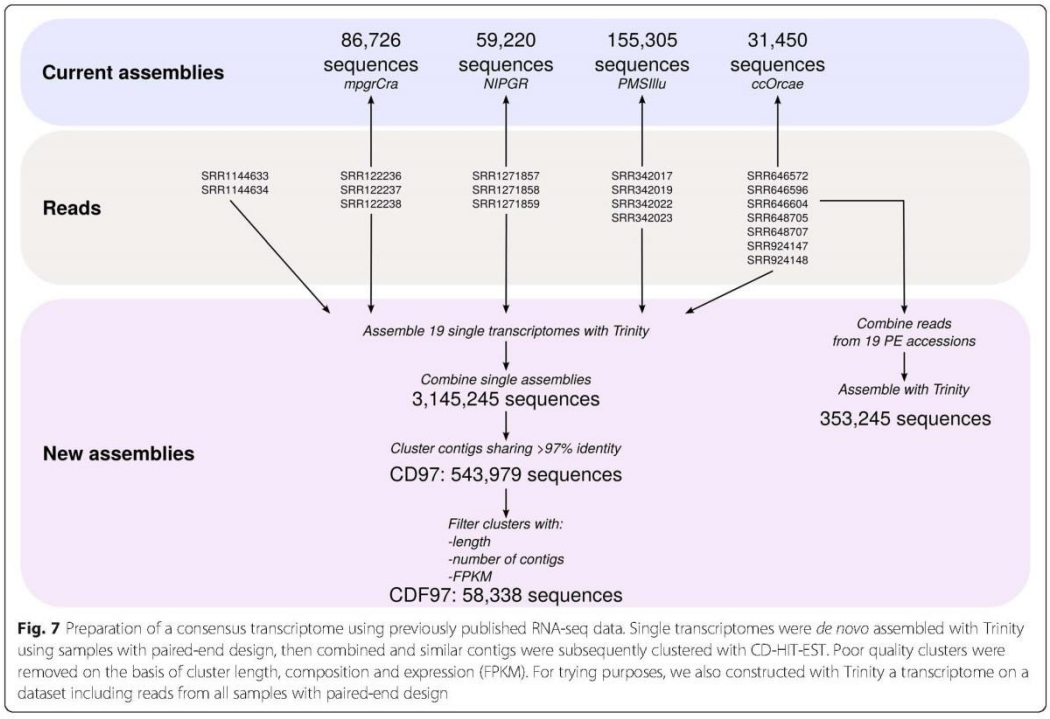


Fig. 7 Preparation of a consensus transcriptome using previously published RNA-seq data. Single transcriptomes were *de novo* assembled with Trinity using samples with paired-end design, then combined and similar contigs were subsequently clustered with CD-HIT-EST. Poor quality clusters were removed on the basis of cluster length, composition and expression (FPKM). For trying purposes, we also constructed with Trinity a transcriptome on a dataset including reads from all samples with paired-end design

The resulting clustered dataset, thereafter renamed CD97, was composed of a total of 534,979 clusters, 357,652 being singletons (a contig displaying no sufficient identity with other contigs) and 177,327 being real clusters, containing more than two contigs (which may originate from the same single assembly or from different single assemblies). A total of 249,423 sequences had identities (e-value < 1e-20) with sequences of the Uniprot database (Blastx), and 9,283 proteins found in this database were represented at 90 % of their length by at least one cluster in CD97.

Participation of initial single assemblies in CD97 clusters was homogeneous, except for SRR122238 for which only 10 % of contigs (4.3 % in clusters, 5.7 % in singletons; Additional file 4: Table S2) were used by CD-HIT-EST. Concerning SRR122238 single assembly, the low proportion of reads used in CD97 was probably due to its very high number of contigs (1,666,984) in comparison with the other assemblies. For other single assemblies, more than 90 % of contigs were used by CD-HIT-EST, with at least 50 % in true clusters (Fig. 9a; Additional file 4: Table S2). Composition of true clusters revealed a somewhat preferential association of contigs from single assemblies obtained in a same study (Fig. 9b). Correlation coefficients calculated on the pattern of participation of each single assembly in true clusters were higher for 4 groups of samples: (i) SRR1144633 and SRR1144634 (SRP035766, leafy flower transition study), (ii) SRR646596, SRR646604 and SRR646572 (SRP017832, MeJA treatments on shoots), (iii) SRR122237 and SRR122236 (SRP005953, mixed libraries from different organs) and (iv) SRR924147, SRR924148, SRR648707 and SRR648705 (SRP026417 and SRP017947, cell suspension MeJA and ORCA overexpression). This preferential association is more likely to be due to the inherent genetic diversity between *C. roseus* cultivars than experimental conditions. However, high coefficient correlations (>0.6) were also observed for independent studies, as exemplified between SRR122236 and SRR1144634. The strongest differences were observed for samples of the NIPGR study (SRR1271857, SRR1271858 and SRR1271859) and for SRR122238. For the latter, this might be linked to its higher participation in singletons than in true clusters (Fig. 9a). In CD97, 105,730 clusters contained contigs from 2 to 5 different single assemblies, 31,055 clusters contained contigs from more than 10 single assemblies and 3,506 clusters were composed of contigs from the 19 single assemblies (Fig. 9c). These 31,055 clusters might represent the core transcriptome of *C. roseus*. Indeed, 25,692 had significant (e-value < 1e-20) identities with proteins of the UniprotKB database (Blastx) (Table 1).

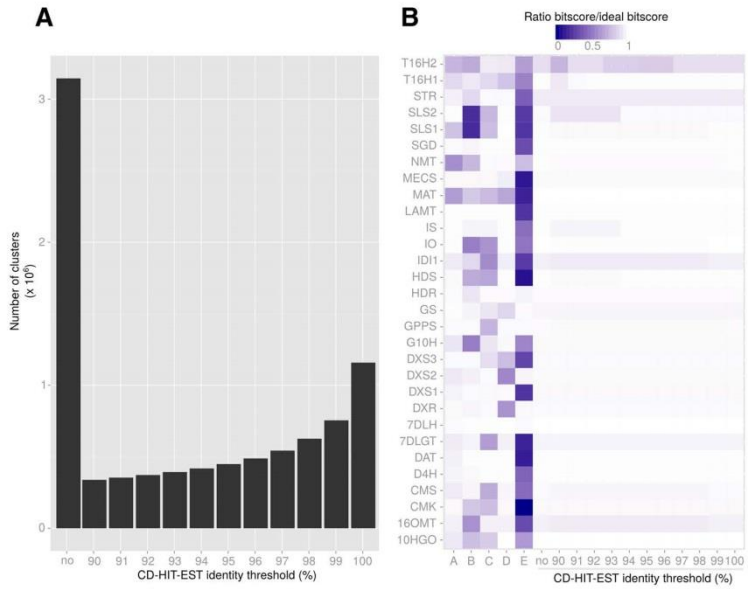


Fig. 8 Clustering of redundant contigs in the dataset resulting from the combination of all single assemblies. Contigs sharing a given % of identity were clustered with CD-HIT-EST. **a** Number of clusters after CD-HIT-EST at % identity thresholds fixed from 90 to 100 %. **b** Reconstruction quality of MIA genes in the current resources (A = ccOrcae, B = mpgrCra, C = NIPGR, D = PMS454, E = PMSIllu) and the datasets resulting from the clustering by CD-HIT-EST at % identity thresholds. Reference MIA gene sequences were BLASTed against each assembly and the resulting bitscore was compared to that of an ideal sequence (bitscore of the reference sequence against itself, i.e. bitscore ratio = 1)

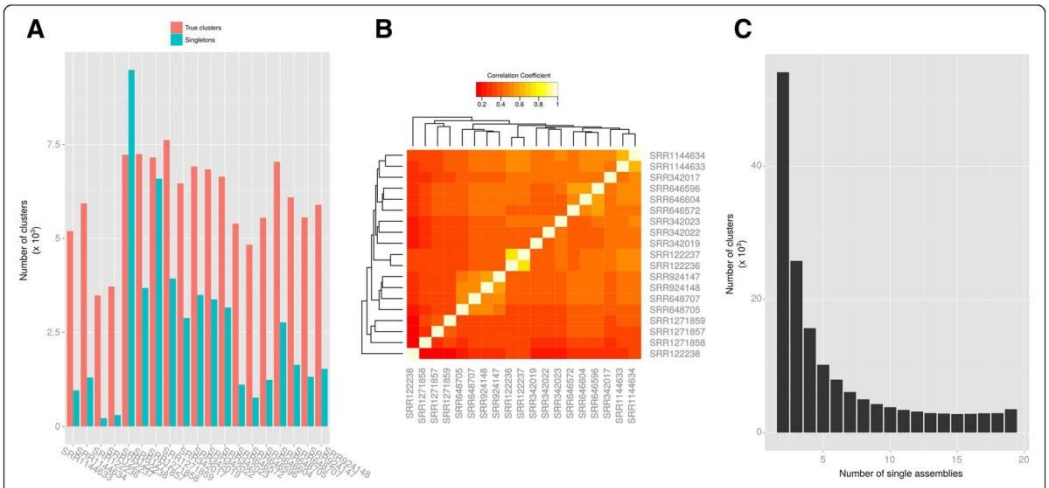


Fig. 9 Composition of the clustered dataset (CD97) resulting from the processing of the combination of all single assemblies with CD-HIT-EST at 97 % identity. **a** Integration of contigs from single assemblies into clusters. Contigs which cannot be grouped with others are called singletons. True clusters, i.e. containing at least two different contigs, may have been formed by the combination of contigs from one or more initial single assemblies. **b** Correlation plot of single assemblies. Contigs found in each cluster (singletons and true clusters) were identified and counted per initial assembly. Two initial assemblies are therefore strongly correlated (Pearson Correlation Coefficient) if their contigs are found in the same clusters. **c** Composition of true clusters. This graph shows how many single assemblies are represented within clusters

Table 1 Transcript annotation and analysis of full-length transcript reconstruction

Assembly	ccOrcae	mpgrCra	NIPGR	PMS454	PMSIllu	CD97	CD97 best clusters[c]	CDF97
Number of transcripts	31,450	86,726	59,220	26,804	155,305	543,979	31,055	58,338
Total annotated transcripts[a]	16,727	55,073	24,142	15,868	22,716	249,413	25,692	49,128
% Annotated transcripts[a]	53.19	63.50	40.77	59.20	14.63	45.85	82.73	84.21
	Cumulative number of proteins[b]							
% Length coverage								
100	4,539	4,115	3,540	1,546	3,168	7,155	5,006	5,734
90	5,686	5,547	5,124	2,127	4,125	9,283	6,198	7,110
80	6,370	6,626	6,377	2,624	4,783	10,903	6,913	7,926
70	6,855	7,641	7,469	3,193	5,381	12,410	7,466	8,572
60	7,298	8,703	8,372	3,827	5,940	14,048	7,958	9,135
50	7,724	9,670	9,119	4,575	6,574	16,085	8,350	9,640
40	8,102	10,528	9,686	5,457	7,177	18,610	8,671	10,033
30	8,479	11,282	10,141	6,497	7,793	21,831	8,908	10,489
20	8,767	11,744	10,393	7,367	8,303	24,619	9,021	10,520
10	8,816	11,830	10,438	7,588	8,430	25,229	9,044	11,137

[a],Blastx analysis vs UniprotKB/Swiss-Prot. A transcript was considered to be annotated if it matches a protein in the database at a e-value threshold of 1e-20
[b],Reports the cumulative number of proteins in UniprotKB/Swiss-Prot matched by at least one transcript in the corresponding assembly at a given % coverage
[c],Clusters in CD97 having contigs from at least ten different single assemblies

To further clean CD97, all putative clusters were tested for 3 criteria: (i) length, (ii) number of contigs and (iii) expression level (sum of Fragment per Kilobase per Million of reads (FPKM) calculated on the 42 samples (19 paired-end and 23 single-end, see Additional file 3: Table S1). Visual inspection of the number of clusters potentially removed by each filter (Fig. 10) was used to choose appropriate values. The objective was to eliminate clusters with poor representation which could be reconstruction artefacts. Choosing low thresholds of number of contigs and FPKM quickly removed a high number of clusters (427,494 with less than 3 contigs and 315,357 with sum FPKM < 5; Fig. 10a and b). For these two filters, we choose to retain values at which changes in the number of removed clusters displayed lower variation:10 contigs per cluster and sum of FPKM >50. Concerning cluster length, the distribution was more graduated (Fig. 10c). In order to avoid removing weakly expressed or small genes, we chose to discard clusters that do not meet at least two of the three filters (Fig. 10d). We expected that this procedure could reduce the loss of weakly expressed genes or rare isoforms. By fixing a minimal length of 500 bp, a number of contigs > 10 and a sum of FPKM > 50, we found that a large number of sequences (245,395) did not pass the three filters. This indicated that many clusters which size was < 500 bp have both poor representation and weak expression levels. We also found 233,752 clusters which had a sum of FPKM < 50 and contained less than 10 contigs. All sequences having a null sum of FPKM fell in this class. Out of the 543,979 clusters of CD97, a total of 485,641 sequences did not pass the filters. The resulting dataset, which contained 58,338 clusters, was retained and called CDF97.

Out of these 58,338 clusters in CDF97, 49,128 sequences had a significant (e-value < 1e-20) identity within the UniprotKB database (Blastx) (Table 1). Our filtering process obviously decreased the number of potentially annotated proteins (249,423 sequences in CD97), and to a lower extent the number of full-length proteins in Uniprot represented by more 90 % of their length (9,283 in CD97, 7,110 in CDF97). However, given the large number of clusters that were removed, we considered that this loss was limited. Very rare isoforms may have been discarded in CDF97 by the filtering procedure but only if they had weak single assembly representation and low expression level. In addition, the number of full-length proteins in CDF97 was still higher than in the previously published datasets (Table 1). Looking for such isoforms will therefore require a more detailed examination of CD97 (non-filtered dataset). The total number of clusters in CDF97 was higher than other assemblies (31,450 in ccOrcae and 26,804 in PMS454) but similar to that of 59,220 in NIPGR and lower than that of mpgrCra (86,726). Therefore, it is likely that all redundancy has not been removed in CDF97. As a consequence, detailed inspection of expression levels together with functional studies will be required to further clean CDF97.

Validation of the consensus FPKM-filtered-CDHIT-94 transcriptome through transcript abundance estimation

The gene expression levels of SLS and T16H isoforms were determined by qPCR and compared to FPKM values

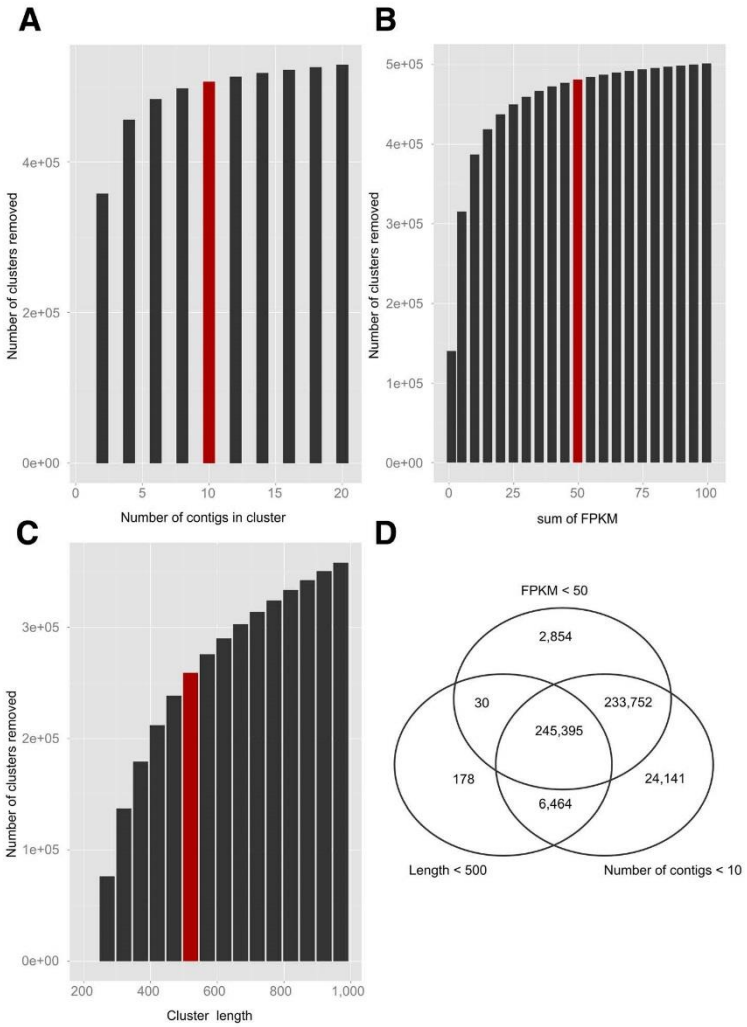


Fig. 10 Characterization of clusters in the clustered dataset (CD97). These graphs represent the number of clusters falling below different threshold values for the number of contigs (**a**), the summed FPKM (on the 42 available samples) (**b**) and cluster length (**c**). Red bars show the thresholds that were retained to filter poorly supported sequences from CD97. Sequences which met at least two of those criteria were discarded (**d**) (total of 485,641)

calculated according to the RSEM procedure for each transcript within each assembly. As each isoform has apparent specific expression patterns (Fig. 11b; [22]), the correct alignment of reads to high quality sequences should be able to yield similar results.

More than 1.1 billion reads from the 42 *C. roseus* samples were mapped back with Bowtie2 [63] to current assemblies, as well as our datasets CD97 and CDF97. Substantial differences were observed in the total number of correctly aligned reads between assemblies: 81.69 % on ccOrcae, 61.75 % on mpgrCra, 73.54 % on NIPGR, 61.75 % on PMS454, 53.54 % on PMSIllu, 90.32 % on CD97 and 89.01 % on CDF97. To estimate abundance of clusters in CD97 and CDF97, a contig to cluster map was prepared, similarly to the procedure used in RSEM relying on a transcript to gene map. This was expected to ensure a correct estimation of expression, in particular for rare isoforms (e.g. found in only one sample) displaying slight sequence difference from the cluster representative sequence. Best hits for SLS and T16H isoforms were selected within each

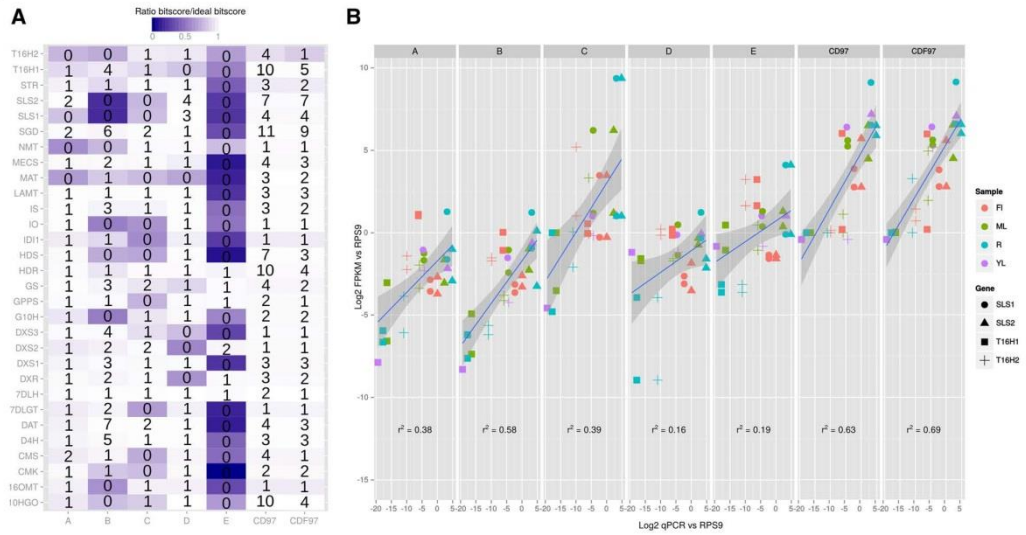


Fig. 11 CDF97 performance in quality reconstruction and transcript expression level. **a** Current resources (A = ccOrcae, B = mpgrCra, C = NIPGR, D = PMS454, E = PMSIllu), CD97 and CDF97 were used as databases to identify homologs of MIA genes, and the resulting bitscore obtained by BLAST was compared to that of an ideal sequence (bitscore of the reference sequence against itself, i.e. bitscore ratio = 1). Values inside each cell indicate the number of hits having a ratio >0.8. **b** Comparison of log2 expression levels measured by qPCR and RNA-seq for SLS and T16H isoforms in current assemblies, CD97 and CDF97. RNA-seq samples used for this comparison were: flowers (FI, SRR122239 and SRR1271859), mature leaves (ML, SRR122251 and SRR1271857), roots (R, SRR122254 and SRR1271858) and young leaves (YL, SRR122252). Expression levels were normalized by RPS9, in both qPCR and RNA-seq measurements. The linear regression line is shown in blue with shading corresponding to 95 % confidence intervals. The r^2 values indicate correlation coefficients of the linear models

assembly. SLS1, SLS2, T16H1 and T16H2 sequences were respectively: Caros007144, Caros020659, Caros001600 and Caros025399 in ccOrcae; cra_locus_10318_iso_2_len_357_-ver_3, cra_locus_1389_iso_1_len_312_ver_3, cra_locus_6184_iso_8_len_1687_ver_3 and cra_locus_6184_iso_10_len_1687_ver_3 in mpgrCra; Cr_TC01142 (SLS1 and SLS2), Cr_TC26727 and Cr_TC35206 in NIPGR; CROWL1VD_rep_c387, CROWL1VD_rep_c782, CROWL1VD_rep_c1347 (T16H1 and T16H2) in PMS454; cro. CRO1L1VD_velvet–Contig19748 (SLS1 and SLS2) and cro. CRO1L1VD_velvet–Contig11543 (T16H1and T16H2) in PMSIllu; SRR648707|TR19558|c0_g2_i4, SRR646572|TR30446|c0_g1_i1, SRR648707|TR1325|c0_g1_i2 and SRR1271857|TR29335|c0_g2_i12 in CD97; SRR648707|TR19558|c0_g2_i4, SRR646572|TR30446|c0_g1_i1, SRR648707|TR1325|c0_g1_i2 and SRR342019|TR37243|c0_g2_i1 in CDF97. Interestingly, the best hit for T16H2 differed between CD97 and CDF97. This indicated that the best form in CD97 was either not sufficiently expressed or not represented in the single assemblies to be conserved in CDF97. Hence in CDF97, a slightly less similar sequence for T16H2 was retained albeit being better represented in reads and assemblies (Fig. 11a). Expression levels were calculated as FPKM and normalized to the best hit of *C. roseus* RPS9 gene sequence (Caros004092, for ccOrcae; cra_locus_1407_iso_8_len_924_ver_3 for mpgrCra; Cr_TC 15537 for NIPGR; CROWL1VD_rep_c282 for PMS454; cro. CRO1L1VD_velvet–Contig5697 for PMSIllu; SRR646572|TR4777|c0_g1_i1 for CD97 and CDF97).

The expression of each isoform was monitored in immature leaves, mature leaves, flowers and roots. Despites the use of different matrices for FPKM and qPCR analyses, interesting results were observed by comparing both types of measurement. Using current assemblies, we found that expression levels measured on mpgrCra displayed the best correlation with qPCR measurements (linear regression, $r^2 = 0.58$; Fig. 11b). For this assembly, T16H2 expression was mixed with that of T16H1 since FPKM indicated similar expressions in flower and immature leaves, while we previously showed by qPCR that T16H1 accumulated in flowers but not T16H2 [22]. The highest correlation was observed for CDF97 ($r^2 = 0.69$; Fig. 11b). The contig-to-cluster map used for CD97 and CDF97 (similar to the transcript to gene procedure in RSEM) apparently allowed a more precise calculation of cluster expression values. The higher correlation coefficient obtained with CD97 and CDF97 were probably due to this procedure which was expected to encompass slight

polymorphisms that would have impeded read alignment on representative sequences of clusters. Because both the contig and representative sequence (which came from a single assembly) belong to the same cluster, expression levels are calculated for each contig and subsequently attributed to the representative sequence. Contigs specifically expressed in one given sample, having punctuate sequence variation (e.g., due to genetic diversity, but corresponding to the same entity) are thereby used to estimate the expression level of the representative sequence in this sample. Taken together, these results are good indicators of the validity of our CDF97dataset, in both sequence reconstruction and expression levels.

Exploitation of CDF97 transcriptome for prediction of other MIA biosynthetic gene isoforms

Our finding that gene isoforms may encode similar enzymes (SLS, this study; T16H [22]) potentially add another layer of complexity to the MIA biosynthetic pathway. To predict putative new isoforms, we therefore looked at the number of hits having a score ratio > 0.8 within each assembly. This threshold was empirically determined as lower enough to reveal potential isoforms. According to Fig. 11a, SLS isoforms are identified in PMS454, CD97 and CDF97, as indicated by the presence of more than 1 hit for the corresponding gene. Similarly, isoforms of T16H are observable in NIPGR, CD97 and CDF97. It also appeared that our dataset CDF97 still displayed redundancy among sequences as the number of hits for some genes was still important (10HGO, HDR, HDS, SGD and SLS2). Rather than being true isoforms, it is probable that our clustering and filtering procedures were not stringent enough to remove all redundant sequences. However, in our approach, a special attention was given to avoid discarding rare isoforms. Therefore, genes having 2 or 3 hits with a score ratio > 0.8 (e.g., LAMT or G10H) would merit further studies to determine whether they correspond to true isoforms (separate genes or alternative transcripts) or only simple alleles of different cultivars. For instance, two similar sequences were predicted for IS in CDF97, which correspond to the recently reported IS homologs in *C. roseus* [60].

Conclusions

Besides T16H and IS, the identification of a second SLS isoform shade light anew on the existence of multiple isoforms of MIA biosynthetic enzymes in *C. roseus*. Apart from the complex cellular and subcellular organisation of the MIA pathway, such a potential enzyme multiplicity constitutes another element implemented along evolution to ensure an efficient and modular production of MIA. It also raises interesting questions regarding the regulation of the MIA metabolic fluxes as suggested by the capacity of both SLS1 and SLS2 to produce secologanin and secoxyloganin, but also regarding evolution of P450s from the seco-iridoid pathway that catalyze more than one reaction.

All these questions strengthened the necessity to develop new tools facilitating the identification of MIA biosynthetic enzymes as well as their potential isoforms. Our reconstructed assembly constitutes thus one of the most optimized transcriptomic resources for *C. roseus* that will facilitate future identification of homologs in the MIA biosynthetic pathway enzymes as well the discovery of uncharacterized enzymes through analyses of gene expression correlation as recently described [1, 2] and demonstrated [64, 65]. This resource opens new perspectives toward the understanding of the whole MIA biosynthetic pathway and remains complementary to genomic sequence analysis.

Methods

Heterologous expression of SLS1 and SLS2 in yeast

Full length SLS1 and SLS2 cDNA were amplified using the pair of primers SLS1for (CTGAGAAGATCTATGGAGATGGATATGGATACCATTAG)/SLS1rev (CTGAGAAGATCTCTAGCTCTCAAGCTTCTTGTAGATG) and SLS2-pYEfor (CTGAGAAGATCTATGGAGATGGATATGGATATCATTAGAAAG)/SLS2-pYErev (CTGAGAAGATCTTTAAAAATTCTGTCTCTCAAGCTTCTTGTAGATA), respectively. Both primer couples include *Bgl*II restriction sites at both extremities to allow cloning of the resulting PCR product in the *Bam*HI site of pYeDP60. Both recombinant plasmids and the empty plasmid were independently used to transform the *S. cerevisiae* strain WAT11 expressing the *A. thaliana* NADPH P450 reductase 1 [56]. Yeasts were grown in 10 ml of CSM medium (Yeast Nitrogen Base 0.67 %, dextrose 2 %, drop-out mix without adenine and uracil 0.05 %) until reaching the stationary phase of culture and prior being harvested by centrifugation. Protein expression was induced by cultivating the harvested yeast in 50 ml of YPGal medium (1 % bacto peptone, 1 % yeast extract, and 2 % Galactose) for 6 h as described in [22].

Enzyme assays

Following induction of protein expression, 50 mL of yeast culture were harvested by centrifugation and resuspended in 2 mL of buffer R (Tris–HCl pH7.5, 50 mM; EDTA 1 mM) in a 50 ml centrifugation tube. An equal volume of glass beads were added (425–600 μm, Sigma) and cells were broken by vigorous shaking. Briefly, tubes were shaked by hand during 30 s in a cold room (4 °C) before being put on ice for 30 additional seconds. This operation was repeated ten times before the addition of two volumes of buffer R allowing the recovering of the yeast crude extracts prior to protein quantification using the Bio-Rad protein microassay. SLS1 and SLS2 activities were analyzed in a final volume of 100 μl containing

600 μg of proteins, 200 μM of NADPH,H^+ and either 20 μM of loganin or secologanin. Reactions were initiated by addition of NADPH,H^+, incubated at 30 °C during 10, 30, 60 or 120 min and quenched by addition of 100 μl of methanol prior to ultra-performance liquid chromatography-mass spectrometry analysis (UPLC-MS).

UPLC-MS analyses

All samples were centrifuged and the supernatants were stored at 4 °C prior to injection. UPLC chromatography system consisted in an ACQUITY UPLC (Waters, Milford, MA, USA). Separation was performed using a Waters Acquity HSS T3 C18 column (150 mm × 2.1 mm, 1.8 μm) with a flow rate of 0.4 mL/min at 55 °C. The injection volume was 5 μL. The mobile phase consisted of solvent A (0.1 % formic acid in water) and solvent B (0.1 % formic acid in acetonitrile). Chromatographic separation was achieved using an 8-min linear gradient from 10 to 24 % solvent B. MS detection was performed by using a SQD mass spectrometer equipped with an electrospray ionization (ESI) source controlled by Masslynx 4.1 software (Waters, Milford, MA). The capillary and sample cone voltages were 3,000 V and 30 V, respectively. The cone and desolvation gas flow rates were 60 and 800 Lh^{-1}. Data collection was carried out in negative mode for secoxyloganin ($[M\text{-}H]^-$ = 403, RT = 6.42 min) and secologanin ($[M + HCOOH\text{-}H]^-$ = 433, RT = 7.12 min) and in positive mode for loganin ($[M + Na]^+$ = 413, RT = 5.61 min). Standard calibration curves for secoxyloganin, secologanin and loganin were prepared (1–25 μM) with known pure standards from Chemtek (Worcester, MA, USA), Phytoconsult (Leiden, The Netherlands) and Extrasynthese (Genay, France), respectively.

Subcellular localization of SLS2

The subcellular localization of SLS2 was determined according to the procedures described in [39]. Briefly, the SLS2 coding sequence was amplified by PCR using primers SLS2-YFP-for (CTGAGAACTAGTATGGAGATGGATATGGATATCATTAGAAAG) and SLS2-YFP-rev (CTGAGAACTAGTAAAATTCTGTCTCTCAAGCTTCTTGTAGATA) and cloned into the *Spe*I restriction sites pSC-A cassette YFPi plasmid in frame with the 5′ extremity of the YFP coding sequence. The resulting plasmid was used for transient transformation of *C. roseus* cells by particle bombardment in combination with a plasmid expressing the ER-CFP marker [66].

Gene expression analysis

SLS1 (L10081) and SLS2 (KF415117) expression was measured by real-time RT-PCR using primers SLS1_QPCR1-for (TAAACCTGAGTTTGAACGCTTAAATCAC)/SLS1-QPCR1-rev (GACAATCTTTGTTAGATCAATCACTGGT) and SLS2_QPCR1-for (CAAGCCTGAATTTGAACGCTTGAATCAT) and SLS2_QPCR1-rev (AATAATCTTGGTCAGATCAATAACTGGC). PCR products were cloned in pGEM-Teasy according to the manufacturer protocol and Sanger sequenced to ensure the specificity of amplification. Primer efficacity and cross-amplification was tested on plasmids containing either SLS1 or SLS2 coding sequence. Different *C. roseus* organs (such as roots, stems, young and mature leaves, flower buds, flowers, and fruits – Apricot sunstorm cultivar) were immediately frozen in liquid nitrogen after sampling. Samples (50 mg) were ground with a mortar and a pestle in liquid nitrogen and total RNA were extracted with the RNeasy Plant mini kit (Qiagen), controlled with a Nanodrop spectrophotometer (ThermoFisher) and treated (1.5 μg) with RQ1 RNase-free DNase (Promega) before being used for first-strand cDNA synthesis by priming with oligo (dT) 18 (0.5 μM). Retro-transcription (RT) of 1.5 μg of total RNA was carried out using the SuperScript III reverse transcriptase kit (Invitrogen) at 50 °C during 1 h according to manufacturer's instructions. Real-time PCR was run on a CFX96 Touch Real- Time PCR System (Bio-Rad) using the SYBR Green I technology. Each reaction was performed in a total reaction volume of 25 μL containing an equal amount of cDNAs (1/3 dilution), 0.05 μM forward and reverse primers, and 1 × DyNAmo™ ColorFlash Probe qPCR Kit (Termo Fisher Scientific). The amplication program was 95 °C for 7 min (polymerase heat activation), followed by 40 cycles containing 2 steps, 95 °C for 10 sec and 60 °C for 40 sec. At the end of the amplification, a melt curve was performed to check amplification specificity. Absolute quantification of transcript copy number was performed with calibration curves and normalization with the *C. roseus* 40S Ribosomal protein S9 (RPS9 – primers qRPS9for TTACAAGTCCCTTCGGTGGT and qRPS9rev TGCTTATTCTTCATCCTCTTCATC) reference gene (Genbank accession AJ749993.1). All amplifications were performed in triplicate and repeated at least on two independent biological repeats.

Publicly available datasets and *de novo* transcriptome assembly

Sequencing files of *C. roseus* samples (project accessions: SRP035766, SRP005953, SRP017832, SRP041695 and SRP008096) were downloaded from the ftp server of the NCBI SRA database. Exhaustive description of all files is provided in Additional file 3: Table S1. Files were converted to fastq files with the NCBI SRA toolkit (v2.3.4-2) and checked with FastQC (v0.11.2). Overall quality was good (Additional file 4: Table S2) but reads from left and right sequencing of paired-end samples as well as single end samples were treated to remove aberrant fragments and adaptors with Trimmomatic v.0.32 [67]. Parameters were the following: Illuminaclip = 2:30:10, Leading = 3, Trailing = 3, Sliding Window =

4:15 and Minimum Length = 36. Adapter sequences were trimmed according to the library design used (GIIx or HiSeq2000). Correct reads were subsequently subjected to *in silico* normalization after being converted into the fasta format with Fastools. Trinity's *in silico* normalization (v2.0.4) relies on the processing of a *k*-mer library (with Jellyfish v2.1.4, $k = 25$) obtained from reads and was used to discard reads having aberrant k-mer abundance and those which coverage (abundance in a given transcript) exceeded 50 (max_cov = 50). This step aims at reducing overrepresented reads that may impede the reconstruction process. Detailed description of the number of processed reads (trimmed and normalized) is available in Additional file 4: Table S2. Paired-end normalized samples were then assembled with Trinity (v2.0.4) [46, 62]. For testing purposes, Trinity was also performed on all the reads combined from ever paired-end normalized sample, with respecting read orientation. Before processing this large number of reads, a second normalization step was performed with the same parameters as described above. Parameters for Inchworm, Chrysalis and Butterfly were defaults. All the steps were conducted on the CCSC computer grid facility (Orléans, France) using the SLURM scheduler running under on a Linux ×86_64 architecture.

Clustering of similar sequences

Different homology thresholds (–c parameter, word length–n 9) were tested for CD-HIT-EST [54] (multi-threaded revised version 784a6f1b5e11 which supports longer fragments) to evaluate the ability to combine similar sequences while preserving isoforms from being assembled in a same contig. This program returns a '.clstr' files containing the contig composition of each cluster and a multi-fasta file containing the representative sequences of each cluster. A representative sequence is the contig which matched best the other contigs found in the same cluster.

Sequence alignment and annotation

BLAST analyses against given databases were performed with the stand alone application v2.29 [68]. Annotation of transcripts was performed with Blastx against the UniprotKB/Swiss-prot database and analyze of hit coverage was done with Trinity perl script analyze_blastPlus_topHit_coverage.pl. This analysis focuses on the number of proteins which are matched at least once by sequences in a given transcriptome [62]. Quick evaluation of transcriptome assemblies was done by setting a local BLAST server with SequenceServer (Pryiam et al., unpublished) and analyzing candidate sequences from the MIA pathway with Blastn.

Estimation of transcript abundance

Estimation of transcript abundance was performed with RSEM v1.2.15 after aligning reads to target transcriptome with Bowtie2 [63] (v2.2.5) with default parameters for both programs and the–no-polyA option. FPKM for the CD97 and CDF97 datasets were computed after preparing respective contig-to-cluster maps (using '.clstr' file generated with CD-HIT-EST) for the rsem-prepare-reference procedure in order to get expression values reflecting abundance in all samples. This is expected to allow alignment of cultivar-specific reads to its cognate contig, while using its expression level for the whole cluster. Expression tables for CD97 and CDF97 were prepared by merging '.genes.results' files. To compare expression levels between qPCR and FPKM in other assemblies, expression levels were re-calculated using the above described procedure but without contig-to-cluster or transcript-to-gene information.

Data processing

The R software (3.1.0, [69]) was used with the GUI interface RStudio v0.98.1091 or in the command line interface for high multicore parallelization. All operations outside dedicated programs were performed with R. The Bioconductor package 'SRA.db' [70] (v1.22.0) was used to retrieve sample information from the SRA. We used the 'seqinr' [71] package (v3.1-3) to remove poorly represented clusters from CD97 after identification of appropriate thresholds for cluster length, abundance and contig number. Graphs for the transcriptomic analysis were generated with the 'ggplot2' package [72] (v1.0.1). Correlations between qPCR and FPKM data were calculated as the adjusted r^2 of a linear model built with the "lm" function. The correlation was established for each dataset using their own expression data (see above).

Availability of supporting data

The dataset (CDF97) supporting the results of this article is available at the LabArchives, LLC, repository, with DOI number 10.6070/H4DR2SG9 and open access at http://dx.doi.org/10.6070/H4DR2SG9. It is also freely available on our website http://bbv-ea2106.sciences.univ-tours.fr/. The non-filtered dataset (CD97) as well as all new assemblies described in this study are available upon request.

Additional files

Additional file 1: Figure S1. Alignment of the amino-acid sequence of SLS1 and SLS2. Sequence identity and similarity are highlighted by black and grey shading, respectively. Red bars denote the position of a predicted transmembrane helix as described in [36].

Additional file 2: Figure S2. Identification of secoxyloganin in enzymatic assays (left) by comparison with a pure authentic standard using UV spectrum (A) and MS spectra in negative (B) and positive (C) modes.

Additional file 3: Table S1. Details for each RNA-seq sample used in this study.

Additional file 4: Table S2. Quality control (FastQC) of paired-end samples, number of reads before and after trimming/normalization procedures and characteristics of individual assemblies.

Additional file 5: Figure S3. Quality of reconstruction of MIA genes in the assembly constructed with reads from all 19 paired-end samples. As a large number of sequences were obtained for this assembly (353,245), redundant sequences were clustered with CD-HIT-EST using different % identity thresholds; A = ccOrcae, B = mpgrCra, C = NIPGR, D = PMS454, E = PMSIlu, no = no clustering.

Abbreviations

MIA: Monoterpenoid indole alkaloids; TDC: Tryptophan decarboxylase; MEP: Methyl-erythritol phosphate; GAP: Glyceraldehyde 3-phosphate; IPP: Isopentenyl diphosphate; DMAPP: Dimethylallyl diphosphate; GPP: Geranyl diphosphate; GES: Geraniol synthase; G10H: Geraniol 10-hydroxylase; 10HGO: 10-Hydroxygeraniol oxidoreductase; IS: Iridoid synthase; IO: Iridoid oxidase; UGT: UDP-glucose glycosyltransferase; 7DLGT: 7-Deoxyloganetic acid glucosyltransferase; 7DLH: 7-Deoxyloganic acid 7-hydroxylase; LAMT: S-adenosyl-L-methionine: loganic acid methyl transferase; SLS: Secologanin synthase; STR: Strictosidine synthase; SGD: Strictosidine β-D-glucosidase; T16H: Tabersonine 16-hydroxylase; 16OMT: 16-Hydroxytabersonine O-methyltransferase; NMT: 16-Methoxy-2, 3-dihydrotabersonine N-methyltransferase; D4H: Desacetoxyvindoline-4-hydroxylase; DAT: Deacetylvindoline-4-*O*-acetyltransferase; IPAP: Internal phloem associated parenchyma; MPGR: Medicinal plant genomics resource; CC: Cathacyc; PMS: Phytometasyn; UPLC-MS: Ultra-performance liquid chromatography-mass spectrometry; YFP: Yellow fluorescent protein; FPKM: Fragments per kilobase of exon per million fragments mapped; UTR: Untranslated region.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TDDB performed the reconstruction of the *C. roseus* transcriptome; CP and VDL made the initial discovery of SLS2; EF, CP, MAL, MC, BH, ML conducted enzyme assays; EF, NP, SB performed subcellular localization; AO, GG analyzed SLS1 and SLS2 gene expression; NGG, BSt-P, LA assisted in the supervision of this work; SEO supervised analysis of the transcriptomic data; VC conceived and coordinated this study. TDDB, BSt-P, MC, SEO, VDL, VC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We gratefully acknowledge support from the "Région Centre" (France, ABISAL grant and Post-Doctoral Fellow attributed to C. P.). We would like also to acknowledge the Fédération CaSciModOT (CCSC, Orléans, France) and Laurent Catherine for access and help to the Région Centre computing grid.

Author details

[1]Université François-Rabelais de Tours, EA2106 "Biomolécules et Biotechnologies Végétales", UFR Sciences et Techniques, 37200 Tours, France. [2]Universidad de Antioquia, Laboratorio de Biotecnología, Sede de Investigación Universitaria, Medellín, Colombia. [3]Department of Biological Sciences, Brock University, 500 Glenridge Avenue, St Catharines, Ontario L2S 3A1, Canada. [4]Department of Biological Chemistry, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK.

Received: 24 October 2014 Accepted: 1 June 2015
Published online: 19 August 2015

References

1. Courdavault V, Papon N, Clastre M, Giglioli-Guivarc'h N, St-Pierre B, Burlat V. A look inside an alkaloid multisite plant: the *Catharanthus* logistics. Curr Opin Plant Biol. 2014;19:43–50.
2. Dugé de Bernonville T, Clastre M, Besseau S, Oudin A, Burlat V, Glévarec G, et al. Phytochemical genomics of the Madagascar periwinkle: Unravelling the last twists of the alkaloid engine. Phytochemistry. 2015;113:9–23.
3. De Luca V, Marineau C, Brisson N. Molecular cloning and analysis of cDNA encoding a plant tryptophan decarboxylase: comparison with animal dopa decarboxylases. Proc Natl Acad Sci. 1989;86:2582–6.
4. Oudin A, Mahroug S, Courdavault V, Hervouet N, Zelwer C, Rodríguez-Concepción M, et al. Spatial distribution and hormonal regulation of gene products from methyl erythritol phosphate and monoterpene-secoiridoid pathways in *Catharanthus roseus*. Plant Mol Biol. 2007;65:13–30.
5. Rai A, Smita SS, Singh AK, Shanker K, Nagegowda DA. Heteromeric and homomeric geranyl diphosphate synthases from *Catharanthus roseus* and their role in monoterpene indole alkaloid biosynthesis. Mol Plant. 2013;6:1531–49.
6. Simkin AJ, Miettinen K, Claudel P, Burlat V, Guirimand G, Courdavault V, et al. Characterization of the plastidial geraniol synthase from Madagascar periwinkle which initiates the monoterpenoid branch of the alkaloid pathway in internal phloem associated parenchyma. Phytochemistry. 2013;85:36–43.
7. Collu G, Unver N, Peltenburg-Loornan AMG, van der Heijden R, Verpoorte R, Memelink J. Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. FEBS Lett. 2001;508:215–20.
8. Höfer R, Dong L, André F, Ginglinger J-F, Lugan R, Gavira C, et al. Geraniol hydroxylase and hydroxygeraniol oxidase activities of the CYP76 family of cytochrome P450 enzymes and potential for engineering the early steps of the (seco)iridoid pathway. Metab Eng. 2013;20:221–32.
9. Miettinen K, Dong L, Navrot N, Schneider T, Burlat V, Pollier J, et al. The seco-iridoid pathway from *Catharanthus roseus*. Nat Commun. 2014;5:3606.
10. Geu-Flores F, Sherden NH, Courdavault V, Burlat V, Glenn WS, Wu C, et al. An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. Nature. 2012;492:138–42.
11. Salim V, Wiens B, Masada-Atsumi S, Yu F, De Luca V. 7-deoxyloganetic acid synthase catalyzes a key 3 step oxidation to form 7-deoxyloganetic acid in *Catharanthus roseus* iridoid biosynthesis. Phytochemistry. 2014;101:23–31.
12. Asada K, Salim V, Masada-Atsumi S, Edmunds E, Nagatoshi M, Terasaka K, et al. A 7-deoxyloganetic acid glucosyltransferase contributes a key step in secologanin biosynthesis in Madagascar periwinkle. Plant Cell. 2013;25:4123–34.
13. Salim V, Yu F, Altarejos J, De Luca V. Virus-induced gene silencing identifies *Catharanthus roseus* 7-deoxyloganic acid-7-hydroxylase, a step in iridoid and monoterpene indole alkaloid biosynthesis. Plant J. 2013;76:754–65.
14. Murata J, Roepke J, Gordon H, De Luca V. The leaf epidermome of *Catharanthus roseus* reveals its biochemical specialization. Plant Cell. 2008;20:524–42.
15. Irmler S, Schröder G, St-Pierre B, Crouch NP, Hotze M, Schmidt J, et al. Indole alkaloid biosynthesis in *Catharanthus roseus*: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. Plant J. 2008;24:797–804.
16. Kutchan TM, Hampp N, Lottspeich F, Beyreuther K, Zenk MH. The cDNA clone for strictosidine synthase from *Rauvolfia serpentina* DNA sequence determination and expression in *Escherichia coli*. FEBS Lett. 1988;237:40–4.
17. McKnight TD, Roessner CA, Devagupta R, Scott AI, Nessler CL. Nucleotide sequence of a cDNA encoding the vacuolar protein strictosidine synthase from *Catharanthus roseus*. Nucleic Acids Res. 1990;18:4939.
18. Geerlings A, Ibañez MM, Memelink J, van Der Heijden R, Verpoorte R. Molecular cloning and analysis of strictosidine beta-D-glucosidase, an enzyme in terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. J Biol Chem. 2000;275:3051–6.
19. St-Pierre B, Besseau S, Clastre M, Courdavault V, Courtois M, Crèche J, et al. Deciphering the evolution, cell biology and regulation of monoterpene indole alkaloids. Adv Bot Res. 2013;68:73–109.
20. St-Pierre B, De Luca V. A cytochrome P-450 monooxygenase catalyzes the first step in the conversion of tabersonine to vindoline in *Catharanthus roseus*. Plant Physiol. 1995;109:131–9.
21. Schröder G, Unterbusch E, Kaltenbach M, Schmidt J, Strack D, De Luca V, et al. Light-induced cytochrome P450-dependent enzyme in indole alkaloid biosynthesis: tabersonine 16-hydroxylase. FEBS Lett. 1999;458:97–102.
22. Besseau S, Kellner F, Lanoue A, Thamm AMK, Salim V, Schneider B, et al. A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in *Catharanthus roseus*. Plant Physiol. 2013;163:1792–803.
23. De Luca V, Fernandez JA, Campbell D, Kurz WGW. Developmental regulation of enzymes of indole alkaloid biosynthesis in *Catharanthus roseus*. Plant Physiol. 1988;86:447–50.
24. Levac D, Murata J, Kim WS, De Luca V. Application of carborundum abrasion for investigating the leaf epidermis: molecular cloning of *Catharanthus roseus* 16-hydroxytabersonine-16-O–methyltransferase. Plant J. 2008;53:225–36.

25. De Luca V, Cutler AJ. Subcellular localization of enzymes involved in indole alkaloid biosynthesis in *Catharanthus roseus*. Plant Physiol. 1987;85:1099–102.
26. Dethier M, De Luca V. Partial purification of an N-methyltransferase involved in vindoline biosynthesis in *Catharanthus roseus*. Phytochemistry. 1993;32:673–8.
27. Liscombe DK, Usera AR, O'Connor SE. Homolog of tocopherol C methyltransferases catalyzes N methylation in anticancer alkaloid biosynthesis. Proc Natl Acad Sci U S A. 2010;107:18793–8.
28. De Carolis E, Chan F, Balsevich J, De Luca V. Isolation and characterization of a 2-oxoglutarate dependent dioxygenase involved in the second-to-last step in vindoline biosynthesis. Plant Physiol. 1990;94:1323–9.
29. Vazquez-Flota F, De Carolis E, Alarco AM, De Luca V. Molecular cloning and characterization of desacetoxyvindoline-4-hydroxylase, a 2-oxoglutarate dependent-dioxygenase involved in the biosynthesis of vindoline in *Catharanthus roseus* (L.) G. Don. Plant Mol Biol. 1997;34:935–48.
30. St-Pierre B, Laflamme P, Alarco AM, De Luca V. The terminal O-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer. Plant J. 1998;14:703–13.
31. Burlat V, Oudin A, Courtois M, Rideau M, St-Pierre B. Co-expression of three MEP pathway genes and geraniol 10-hydroxylase in internal phloem parenchyma of *Catharanthus roseus* implicates multicellular translocation of intermediates during the biosynthesis of monoterpene indole alkaloids and isoprenoid-derive. Plant J. 2004;38:131–41.
32. Oudin A, Courtois M, Rideau M, Clastre M. The iridoid pathway in *Catharanthus roseus* alkaloid biosynthesis. Phytochem Rev. 2007;6:259–76.
33. Guirimand G, Guihur A, Phillips MA, Oudin A, Glévarec G, Melin C, et al. A single gene encodes isopentenyl diphosphate isomerase isoforms targeted to plastids, mitochondria and peroxisomes in *Catharanthus roseus*. Plant Mol Biol. 2012;79:443–59.
34. St-Pierre B, Vazquez-Flota F, De Luca V. Multicellular compartmentation of *Catharanthus roseus* alkaloid biosynthesis predicts intercellular translocation of a pathway intermediate. Plant Cell. 1999;11:887–900.
35. Guirimand G, Courdavault V, Lanoue A, Mahroug S, Guihur A, Blanc N, et al. Strictosidine activation in Apocynaceae: towards a "nuclear time bomb"? BMC Plant Biol. 2010;10:182.
36. Guirimand G, Guihur A, Ginis O, Poutrain P, Héricourt F, Oudin A, et al. The subcellular organization of strictosidine biosynthesis in *Catharanthus roseus* epidermis highlights several trans-tonoplast translocations of intermediate metabolites. FEBS J. 2011;278:749–63.
37. Guirimand G, Guihur A, Poutrain P, Héricourt F, Mahroug S, St-Pierre B, et al. Spatial organization of the vindoline biosynthetic pathway in *Catharanthus roseus*. J Plant Physiol. 2011;168:549–57.
38. Costa MMR, Hilliou F, Duarte P, Pereira LG, Almeida I, Leech M, et al. Molecular cloning and characterization of a vacuolar class III peroxidase involved in the metabolism of anticancer alkaloids in *Catharanthus roseus*. Plant Physiol. 2008;146:403–17.
39. Guirimand G, Burlat V, Oudin A, Lanoue A, St-Pierre B, Courdavault V. Optimization of the transient transformation of *Catharanthus roseus* cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. Plant Cell Rep. 2009;28:1215–34.
40. Yu F, De Luca V. ATP-binding cassette transporter controls leaf surface secretion of anticancer drug components in *Catharanthus roseus*. Proc Natl Acad Sci USA. 2013;110:15830–5.
41. Góngora-Castillo E, Childs KL, Fedewa G, Hamilton JP, Liscombe DK, Magallanes-Lundback M, et al. Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. PLoS One. 2012;7, e52506.
42. Van Moerkercke A, Fabris M, Pollier J, Baart GJE, Rombauts S, Hasnain G, et al. CathaCyc, a metabolic pathway database built from *Catharanthus roseus* RNA-Seq data. Plant Cell Physiol. 2013;54:673–85.
43. Xiao M, Zhang Y, Chen X, Lee E-J, Barber CJS, Chakrabarty R, et al. Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. J Biotechnol. 2013;166:122–34.
44. Verma M, Ghangal R, Sharma R, Sinha AK, Jain M. Transcriptome Analysis of *Catharanthus roseus* for Gene Discovery and Expression Profiling. PLoS One. 2014;9, e103583.
45. Liu L-YD, Tseng H-I, Lin C-P, Lin Y-Y, Huang Y-H, Huang C-K, et al. High-throughput transcriptome analysis of the leafy flower transition of *Catharanthus roseus* induced by peanut witches'-broom phytoplasma infection. Plant Cell Physiol. 2014;55:942–57.
46. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
47. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92.
48. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7:909–12.
49. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30:1660–6.
50. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14:157–67.
51. Moreton J, Dunham SP, Emes RD. A consensus approach to vertebrate de novo transcriptome assembly from RNA-seq data: assembly of the duck (*Anas platyrhynchos*) transcriptome. Front Genet. 2014;5:190.
52. Nakasugi K, Crowhurst R, Bally J, Waterhouse P. Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. PLoS One. 2014;9, e91776.
53. Duan J, Xia C, Zhao G, Jia J, Kong X. Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics. 2012;13:392.
54. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.
55. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003;19:651–2.
56. Pompon D, Louerat B, Bronine A, Urban P. Yeast expression of animal and plant P450s in optimized redox environments. Methods Enzymol. 1996;272:51–64.
57. Guengerich FP, Sohl CD, Chowdhury G. Multi-step oxidations catalyzed by cytochrome P450 enzymes: Processive vs. distributive kinetics and the issue of carbonyl oxidation in chemical mechanisms. Arch Biochem Biophys. 2011;507:126–34.
58. Imai T, Yamazaki T, Kominami S. Kinetic studies on bovine cytochrome p45011 beta catalyzing successive reactions from deoxycorticosterone to aldosterone. Biochemistry. 1998;37:8097–104.
59. Sugiyama K, Nagata K, Gillette J, Darbyshire J. Theoretical kinetics of sequential metabolism in vitro. Study of the formation of 16 alpha-hydroxyandrostenedione from testosterone by purified rat P450 2C11. Drug Metab Dispos. 1994;22:584–91.
60. Munkert J, Pollier J, Miettinen K, Van Moerkercke A, Payne R, Müller-Uri F, et al. Iridoid Synthase Activity Is Common among the Plant Progesterone 5β-Reductase Family. *Mol Plant* 2015;8:136–52.
61. Brown S, Clastre M, Courdavault V, O'Connor SE. De novo production of the plant-derived alkaloid strictosidine in yeast. Proc Natl Acad Sci U S A. 2015;112:3205–10.
62. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
63. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
64. Kellner F, Geu-Flores F, Sherden NH, Brown S, Foureau E, Courdavault V, et al. Discovery of a P450-catalyzed step in vindoline biosynthesis: a link between the aspidosperma and ebumamine alkaloids. Chem Commun. 2015;51:7626–8.
65. Stavrinides A, Tatsis EC, Foureau E, Caputi L, Kellner F, Courdavault V, et al. Unlocking the Diversity of Alkaloids in *Catharanthus roseus*: Nuclear Localization Suggests Metabolic Channeling in Secondary Metabolism. Chem Biol. 2015;22:336–41.
66. Nelson BK, Cai X, Nebenführ A. A multicolored set of in vivo organelle markers for co-localization studies in Arabidopsis and other plants. Plant J. 2007;51:1126–36.
67. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
68. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

69. Team R: R Development Core Team. *R A Lang Environ Stat Comput* 2013. Available: http://www.r-project.org/.
70. Zhu Y, Stephens RM, Meltzer PS, Davis SR. SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics. 2013;14:19.
71. Charif D, Lobry JR. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Struct approaches to Seq Evol. Springer. 2007. p. 207–32.
72. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer Science & Business Media; 2009.

Class II Cytochrome P450 reductase governs the biosynthesis of alkaloids in Madagascar periwinkle

Claire Parage[1], Emilien Foureau[1*], Franziska Kellner[2*], Vincent Burlat[3], Samira Mahroug[1], Arnaud Lanoue[1], Thomas Dugé de Bernonville[1], Monica Arias Londono[1,4], Inês Carqueijeiro[1], Audrey Oudin[1], Sébastien Besseau[1], Nicolas Papon[5], Gaëlle Glévarec[1], Lucia Atehortùa[4], Nathalie Giglioli-Guivarc'h[1], Benoit St-Pierre[1], Marc Clastre[1],Sarah E. O'Connor[2#], Vincent Courdavault[1#]*

[1]Université François-Rabelais de Tours, EA2106 "Biomolécules et Biotechnologies Végétales", Tours, France.

[2]Department of Biological Chemistry, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, United Kingdom.

[3]Université de Toulouse, UPS, UMR 5546, Laboratoire de Recherche en Sciences Végétales, BP 42617 Auzeville, F-31326 Castanet-Tolosan, France.

[4]Universidad de Antioquia, Laboratorio de Biotecnología, Sede de Investigación Universitaria, Medellin, Colombia.

[5]Université d'Angers, EA3142 "Groupe d'Etude des Interactions Hôte-Pathogène", Angers, France.

***These authors have contributed equally to this work**

[#]Corresponding authors

Vincent Courdavault

Université François-Rabelais de Tours - EA2106 "Biomolécules et Biotechnologies Végétales", UFR Sciences et Techniques, 37200, Tours, France

e-mail: vincent.courdavault@univ-tours.fr

Sarah E. O'Connor

Department of Biological Chemistry, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, United Kingdom

e-mail: Sarah.O'Connor@jic.ac.uk

Running title: Class II CPR governs alkaloid biosynthesis

Key-words : *Catharanthus*, cytochrome P450 reductase, specialized metabolism, alkaloids

Abbreviations: BiFC, bimolecular fluorescence complementation; C4H, cinnamate 4-hydroxylase; *C. roseus*, *Catharanthus roseus*; CPR, NADPH-cytochrome P450 reductase; CFP, cyan fluorescent protein; DFR, diflavin reductase; FMN, flavin mononucleotide; G8H,

geraniol 8-hydroxylase; IPAP, internal phloem associated parenchyma; IO, iridoid oxidase; PCC, Pearson correlation coefficient; SLS, secologanin synthase; T3O, 16-methoxytabersonine 3-oxygenase; T16H, tabersonine 16-hydroxylase; T19H, tabersonine 19-hydroxylase; YFP, yellow fluorescent protein; VIGS, virus-induced gene silencing; 7DLH, 7-deoxyloganic acid 7-hydroxylase.

INTRODUCTION

With more than 200,000 distinct molecules, the specialized metabolism of plants constitutes one of the main sources of natural compounds, also reflecting the wide capacities of environmental adaptation of these sessile organisms. This singular complexity of compounds is an evolutionary process that results from a dramatic diversification of plant metabolic pathways and of the genes encoding the associated metabolic enzymes. From this point of view, cytochromes P450 (P450s) are a prototypical example of ubiquitous enzymes encoded by a gene superfamily, carrying out multiple types of reactions ranging from hydroxylation, epoxidation, oxygenation, dealkylation, decarboxylation, C-C cleavage or ring opening (Bak et al., 2011; Guengerich and Munro, 2011). P450s catalyze a considerable array of challenging reactions in the biosynthesis of plant specialized metabolites including phenylpropanoids, terpenoids, cyanogenic glycosides or alkaloids (Mizutani and Ohta, 2010; Mizutani and Sato, 2011), in addition to their role in primary metabolism such as hormone biosynthesis (Bak et al., 2011; Takei et al., 2004). Most of characterized P450s perform single oxidations but growing evidence now point out the existence of P450-mediated multi-step oxidations (Guengerich et al., 2011). However, their catalytic cycle always requires two one-electron transfer steps from NADPH into the prosthetic heme of P450s. This transfer occurs through the flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN) domains of NADPH-cytochrome P450 reductases (CPR), albeit in some cases, the second electron transfer can also arise from cytochrome b5 (Munro et al., 2013). The absolute requirement of CPR thus makes this flavoprotein a cornerstone in P450 activities and, consequently, is an absolute requirement in specialized (and primary) metabolisms. As such, the physical interactions between P450s and CPR that guide electron shuffling directly influence P450 activities and are facilitated by membrane anchoring of both types of proteins mainly localized to endoplasmic reticulum (Hasemann et al., 1995; Ro et al., 2002; Denisov et al., 2007). The low CPR/P450 ratio, around 1:15, also implies a potential competition between P450s and as a consequence, some kind of logistic control must be employed to ensure the

coordinated operation of all P450s belonging to similar biosynthetic pathways (Shephard et al., 1983).

In contrast to yeasts and mammals that harbor only a single CPR, vascular plants evolved 2 or 3 CPR isoforms, as reported for instance in Jerusalem artichoke (Benveniste et al., 1991), poplar (Ro et al., 2002), parsley (Koopmann and Hahlbrock, 1997), cotton (Yang et al., 2010), winter cherry (Rana et al., 2013). The Arabidopsis genome contains two genes encoding functionally active CPR genes (Arabidopsis Thaliana P450 Reductase), named ATR1 and ATR2, and a third more distant gene (*ATR3*) whose expression product has been reported to reduce P450s (Urban et al., 1997; Mizutani and Ohta, 1998; Varadarajan et al., 2010). By contrast, poplar and *Nothapodytes foetida* possess three genes encoding genuine CPRs (Ro et al., 2002; Huang et al., 2012). Except for CPR-like (ATR3 type), flowering plant CPRs are highly conserved (65%-80%) and clustered into two distinct phylogenic groups on the basis of N-terminal hydrophobic sequences. The first cluster (class I) contains sequences from eudicotyledons while the second (class II) encompass monocotyledon and eudicotyledon (Ro et al., 2002). While CPRs from both clusters can reduce P450 with an apparent similar efficiency, strong differences characterize their expression profiles (Urban et al., 1997; Mizutani and Ohta, 1998; Ro et al., 2002). CPR1s belonging to class I, such as ATR1, are constitutively expressed, while CPR2s from class II (ATR2 for example) are inducible by environmental stimuli such as wounding, pathogen infection or light exposure (Koopmann and Hahlbrock, 1997; Mizutani and Ohta, 1998; Ro et al., 2002; Schwarz et al., 2009; Yang et al., 2010, Rana et al., 2013). These observations support a specification of the physiological roles of each CPR isoform that plants may deploy to meet the reductive demand of P450-mediated reactions. CPR-like (ATR3) are poorly characterized but appears to be essential for embryo development (Varadarajan et al., 2010). By contrast, it is now believed that constitutively expressed CPRs (CPR1s- class I) ensure basal P450s activities in primary metabolism or in constitutive synthesis of specialized metabolisms while inducible CPRs (CPR2s- class II) serve in adaptation mechanisms or in defense reactions including the elicited biosynthesis of specialized metabolites. This hypothesis has been partially confirmed by establishing a correlation between ATR2 activity and lignin biosynthesis in Arabidopsis as well as between *CPR1* expression and basal pungent alkaloid synthesis in *Capsicum* spp. (Mazourek et al., 2009; Sundin et al., 2014). However, more investigations are still required to firmly establish a functional specificity of CPRs in plants.

For more than forty years, specialized metabolisms have been illustrated through studies on the Madagascar periwinkle, *Catharanthus roseus* (*C. roseus*). This plant notably synthesizes alkaloids from the monoterpene indole alkaloid (MIA) family that encompass valuable compounds such as the antineoplastic vinblastine and vincristine. These MIAs result from a long and complex biosynthetic pathway whose characterization has made great progress over the last five years due to the development of large sets of transcriptomic data and a draft genome sequence (Góngora-Castillo et al., 2012; Van Moerkercke et al., 2013 ; Xiao et al., 2013; Dugé de Bernonville et al., 2015a; Kellner et al., 2015a). As a case study, within the 30-50 enzymatic steps from the MIA biosynthetic pathway, no less than eleven P450s have been successively identified to date including, ranked as per their order in MIA biosynthetic steps (**Figure 1**): geraniol 8-hydroxylase (G8H, CYP76B6; Collu et al., 2001), iridoid oxidase (IO, CYP76A26; Salim et al., 2014: Miettinen et al., 2014), 7-deoxyloganic acid 7-hydroxylase (7DLH, CYP72A224; Salim et al., 2013; Miettinen et al., 2014), four isoforms of secologanin synthase (SLS1-4, CYP72A1; Irmler et al., 2000; Dugé de Bernonville et al., 2015b; Brown et al., 2015), two isoforms of tabersonine 16-hydroxylase (T16H1-2, CYP71D12-CYP71D351; Schröder et al., 1999; Guirimand et al., 2011; Besseau et al., 2013), tabersonine 19-hydroxylase (T19H, CYP71BJ1; Giddings et al., 2011) and 16-methoxytabersonine 3-oxygenase (T3O, CYP71D1; Kellner et al., 2015b; Qu et al., 2015a). While some of these enzymes catalyze a single oxygenation reaction including hydroxylation (7DLH, T16H1, T16H2, T19H) or epoxidation (T3O), unusual reactions have been also reported such as the ring opening of loganin yielding secologanin (SLS) or the three step oxidation of nepetalactol to form 7-deoxyloganetic acid (IO). Besides this diversity of reactions, a complex spatiotemporal organization of these P450-mediated enzymatic steps has been also reported in periwinkle leaves with the first MIA biosynthetic conversions (G8H to 7DLH) occurring in internal phloem associated parenchyma (IPAP) and the remaining reactions until T3O and the next two enzymatic steps in epidermis (**Figure 1**; Courdavault et al., 2014; Salim et al., 2014; Miettinen et al., 2014; Qu et al., 2015a). As a matter of fact, this multicellular organization of the MIA biosynthetic pathway constitutes the first layer of the physiological processes regulating MIA formation.

Interestingly, the periwinkle CPR was one of the first plant CPRs to be purified and cloned (Madyastha and Coscia, 1979; Meijer et al., 1993). Due to the presence of a hydrophobic residue stretch at its N-terminal end, this protein groups with class II CPRs in agreement with its transcriptional regulation in response to fungal elicitor preparation and to

jasmonate (Meijer et al., 1993; Collu et al., 2001; Ro et al., 2002). Although only one CPR has been cloned in *C. roseus*, occurrence of multiple isoforms is expected and is now supported by transcriptomic and genomic data (Canto-Canché and Loyolas-Vargas, 2001). However, no formal relationship has been established between the periwinkle CPRs and the biosynthesis of MIA. This prompted us to take advantage of the richness of the *C. roseus* model for MIA biosynthesis to accurately explore the role of each class of CPRs in specialized metabolism. By combining biochemical characterization, protein interaction analyses, mapping of cellular gene expression profiles and gene silencing approaches, our data provide new evidences for the predominant role of class II CPR in MIA/specialized metabolism.

RESULTS

Identification of C. roseus CPR homologs

In addition to a contig corresponding to the CPR cDNA (X69791) cloned by Meijer et al. (1993), interrogation of *C. roseus* transcriptomic resources led to the identification of two additional contigs homologous to CPR (**Supplemental Table 1**), with proteins deduced from "contig CPR candidate 1" and "contig CPR candidate 2" displaying 67% and 25% of identity with CPR, respectively (**Supplemental Figure 1**). According to previously published classifications of CPRs (Ro et al., 2002; Jensen and Moller, 2010; Varadarajan et al., 2010), phylogenetic analyses demonstrated that each deduced protein clusters in distinct phylogenetic subgroups (**Supplemental Figure 2**). While the original CPR falls into class II CPR, encompassing CPRs potentially associated to specialized metabolism, "contig CPR candidate 1" falls into class I CPR, mostly dedicated to basal metabolism and "contig CPR candidate 2" clusters in CPR class III whose prominent member in Arabidopsis corresponds to ATR3. Based on this result and on the Arabidopsis nomenclature, "contig CPR candidate 1", the original CPR and "contig CPR candidate 2" were renamed CPR1, CPR2 and CPR3/diflavin reductase (DFR), respectively. *C. roseus* genome analysis also revealed that the three corresponding genes were present at one copy per haploid genome and spanned over 18 or 12 exons for CPR1 (CRO_T001672)/CPR2 (CRO_T031702) and CRP3 (CRO_T033752), respectively (**Supplemental Figure 3**). Genomic organization of CPR1 and CPR2 is similar regarding intron positions and intron/exon sizes except for the first intron of CPR1 that is roughly 7-fold longer than the first intron of CPR2. Such similarity may reflect the gene duplication event leading to both CPR appearances. Therefore, these results suggest

that *C. roseus* contains two CPRs, CPR1 (KJ701028) and CPR2 (X69791) potentially associated to basal and inducible/specialized metabolisms, respectively, and one more distant copy CPR3/DFR (KM111538), a likely ortholog of ATR3.

Sequence analysis and subcellular localization of C. roseus CPRs

Analysis of the deduced protein sequences of *C. roseus* CPRs revealed that both CPR1 and CPR2 are characterized by (i) the presence of conserved FMN-, FAD- and NADPH binding domains that have been implicated in electron transfer, and (ii) identical residues predicted to be involved in interactions with P450s (Jensen and Moller, 2010) (**Supplemental Figure 4, Supplemental Figure 5**). Both proteins also bear a predicted membrane spanning domain at their N-terminal end that is predicted to adopt a α-helical conformation (**Supplemental Figure 6**), and that has been shown to mediate endoplasmic reticulum (ER) anchoring in poplar (Ro et al., 2002). To test the functionality of this type of sequence and to investigate the subcellular localization of *C. roseus* CPRs, CPR1 and CPR2 were expressed as a C-terminal yellow fluorescent protein (YFP) fusion proteins (CPR1-YFP, CPR2-YFP) to avoid masking of the membrane spanning domain. In *C. roseus* transiently transformed cells, the fusion proteins exhibited a network-shaped fluorescent signal that perfectly merged with the signal of the "ER"-cyan fluorescent protein (CFP) marker (**Figure 2A-H**), suggesting that both CPR1 and CPR2 are anchored to ER in agreement with the classical localization pattern of P450s. When mutants of CPR1 and CPR2 lacking the predicted transmembrane domain were expressed as YFP labeled fusions, the previously observed localization pattern was disrupted, demonstrating the functionality of the predicted spanning membrane domain that was necessary and sufficient to ensure ER localization/anchoring (**Supplemental Figure 7**). In addition, in agreement with the assigned CPR classification, we noted that CPR1 and CPR2 differ in the length of the protein sequence preceding the membrane spanning domain; CPR1 only has a short stretch of residues while CPR2 exhibits an extended amino acid sequence, enriched in serine residues, which was initially, but wrongly, considered as a plastid targeting sequence (**Supplemental Figure 1**; Ro et al., 2002). CPR3/DFR also displays conserved FMN-, FAD- and NADPH binding domains but shows substantial differences in the P450 interacting region compared to CPR1 and CPR2 (**Supplemental Figure 5**). Moreover, CPR3/DFR lacks the N-terminal membrane spanning domain, explaining the nucleocytosolic localization observed in *C. roseus* cells transformed with CPR3-YFP fusion protein (**Figure 2I-L**). This nucleocytosolic localization has been previously observed with ATR3 (Varadarajan et al., 2010). Therefore, the distinct subcellular localization patterns, along with

the sequence differences in the P450 interacting domain in CPR1 and CPR2 compared to CPR3/DFR suggest the existence of distinct sets of interacting partners for these proteins.

CPR1 and CPR2 but not CPR3/DFR interact and reduce C. roseus P450s associated to specialized metabolism

To gain insight into some possible selective interactions among the P450s and CPRs of *C. roseus*, BiFC analyses were first conducted by co-expressing each of the three CPRs with several P450s from the MIA biosynthetic pathway, including G8H, SLS1, T16H1 and T16H2 (**Figure 1**; **Figure 3**). An additional test was also conducted with cinnamate 4-hydroxylase (C4H; CYP73A5) since this P450 corresponds to the first P450 of the phenylpropanoid biosynthetic pathway, acting thus in a distinct specialized metabolism. Moreover, the Arabidopsis C4H ortholog has been shown to be reduced with a similar efficiency by both ATR1 and ATR2, and thereby provides a reference (Hotze et al., 1995; Mizutani and Ohta, 1998). As described above, CPRs and P450s were fused to the N-terminal end of the split YFP fragments (N-terminal split YFP, YFP^N, for CPRs and C-terminal split YFP, YFP^C, for P450s) to preserve ER anchoring capacities following transient expression in *C. roseus* cells. Interestingly, while no self-interactions were observed for either CPR1 or CPR2 (**Figure 3A-D**), the apparition of a strong fluorescent signal in all the other conditions suggested that CPR1 and CPR2 were capable of interaction with each of the five tested P450s (**Figure 3E-X**). By contrast, no interactions of CPR3/DFR neither with itself nor with P450s were observed, as exemplified with T16H2 (**Figure 3Y-A2**), which may be link to the neofunctionalization of class III CPR and/or to the lack of the N-terminal membrane spanning domain in CPR3/DFR.

In addition, the capacity of the periwinkle CPR1 and CPR2 to reduce G8H, SLS1, T16H1, T16H2 and C4H was analyzed by conducting functional assays via protein expression in yeast. The resulting activities were compared to those measured in the WAT11 strain expressing an Arabidopsis CPR and to the activity engendered by the endogenous yeast CPR (**Table 1**; **Supplemental Table 2**). Since partial losses of activity were observed when periwinkle CPRs were expressed with tags as reported for CPR from *Camptotheca acuminata* (Qu et al., 2015b), assays were thus performed with untagged proteins. Interestingly, in agreement with BiFC results, we noted that both CPR1 and CPR2 were able to reduce the five tested P450s as revealed by specific substrate conversions. Moreover, although quantification of CPR/P450 activity, which requires purified protein, was not achievable, each of the P450 activities was among the same order of magnitude, suggesting that both CPR1 and CPR2 may

reduce P450s *in vitro* with a similar efficiency as previously reported for C4H in Arabidopsis (Mizutani and Ohta, 1998). Furthermore, we also noticed that CPR3/DFR was unable to reduce T16H, which is consistent with the lack of interaction observed during BiFC assays. To determine whether the inactivity resulted from a lack of ER membrane anchoring, the 74-first residues of CPR2 encompassing the membrane spanning domain was fused to the N-terminus of CPR3/DFR, allowing ER localization as revealed by YFP imaging in *C. roseus* cells (**Supplemental Figure 8**). However, this modification was unable to restore P450 activities (at least SLS1, T16H1 and T16H2) in yeast, strongly suggesting that CPR3/DFR reduces different enzymes and acts in different metabolic pathways. As a consequence, CPR3/DFR was renamed DFRs in agreement with the presence of two flavin reductase domains and the apparent lack of P450 reduction capacity.

CPR1 and CPR2 display ubiquitous expressions associated with specific sets of genes

The proposed/putative specialization of CPRs with MIA biosynthesis was then subsequently investigated at the whole plant level through gene expression analyses. By studying transcript abundance in *C. roseus* transcriptomic datasets (Góngora-Castillo et al., 2012; Dugé de Bernonville et al., 2015a), we first observed that *CPR1*, *CPR2* and *DFR* genes were expressed in all organs associated with MIA biosynthesis, including roots, stems, young/mature leaves and flowers although *DFR* expression was very low as previously reported for *ATR3* (**Figure 4**; Varadarajan et al., 2010). *CPR1* and *CPR2* exhibited a similar pattern of expression in organs but *CPR2* expression was always almost two-fold higher than *CPR1* expression. Furthermore, while *CPR1* was not regulated by methyl jasmonate (MeJA) treatment, *CPR2* was induced by this hormone, which is consistent with class II *CPR* responsiveness. This ubiquitous distribution of *CPR1* and *CPR2* transcripts in periwinkle organs was confirmed by qPCR and was also compared to the expression of *G10H*, *SLS1*, *T16H1*, *T16H2* and *C4H* (**Supplemental Figure 9**). Although we noted that CPR and P450 encoding genes were co-expressed in the different MIA producing organs with some profile specificities inherent to each P450, this co-occurrence of gene expression can only led us to speculate that *CPR1* as well as *CPR2* can both potentially reduce P450s associated to MIA biosynthesis.

As a consequence, by taking advantage of the optimized transcriptomic *C. roseus* dataset (Dugé de Bernonville et al., 2015a), a global analysis of the periwinkle genes co-expressed with each CPRs had been also undertaken *via* calculating the Pearson correlation coefficient (PCC) of either *CPR1*, *CPR2* or *DFR* with each of the 58,338 other transcripts

found in this assembly. This dataset was previously obtained by combining individual transcriptomes assembled from sequencing runs available on the SRA. It has the advantage to integrate expression measurements over a wider range of samples than previously reported by taking into account sequence polymorphisms. Different lists for each *CPR/DFR* were then established at different PCC values and we examined similarities between three selected *CPR*-specific lists (one per *CPR*) obtained at selected PCC cut-off values (**Figure 5A**). Interestingly, intersection sizes indicated that these co-expressed gene lists did not overlap unless considering genes with lower PCC values (< 0.2 for instance), arguing thus for a specialization of *CPR* expression. In addition, the number of correlated genes strongly differed between the two *CPRs* and *DFR*. For instance, at a PCC value > 0.4, we counted 12,506 transcripts for *CPR1*, 2,202 for *CPR2* and 236 for *DFR* (**Figure 5B**). In fact, *CPR1* had the highest number of strongly co-expressed genes (for PCC > 0.8, 1,383 transcripts for *CPR1*, 5 for *CPR2* and 0 for *CPR3*). This revealed once again that periwinkle *CPRs* and *DFR* are transcriptionally unrelated. The analysis of overexpressed Gene Ontology (GO) terms represented in co-expressed genes with a chosen low PCC value (> 0.4) and of keywords from UniProt also reinforces these differences (**Supplemental Figure 10A and B**).

While more genes co-expressed with *CPR1*, we noted that more P450s were associated to *CPR2* (79 out of 2,202 co-expressed genes) than to *CPR1* (48 out of 12,506) or to *DFR* (2 out of 236; Supplemental Table 3A; Supplemental Table 3B) based on a total of around 400 predicted P450 coding sequences (that may exhibit redundancy). A basic analysis of the predicted P450 functions revealed that more than 25 % of the *CPR1*-associated P450s (13/48 but 13/38 P450s displaying predicted functions) acted in primary metabolism such as hormone synthesis, whereas the remaining P450s were potentially involved in a specialized metabolism that appears to correspond to phenylpropanoid biosynthesis. By contrast, only one occurrence potentially associated to primary metabolism was found in the *CPR2* list whereas all the others (65/79 but 65/66 with predicted functions) were associated to specialized metabolism, including MIA biosynthesis. Furthermore, the fact that only two P450s were associated to *DFR* confirms that this protein is probably involved in the reduction of proteins distinct from P450s, in agreement with biochemical activity assays. Such discrepancies were also confirmed by the analysis of GO terms of the whole gene list of each *CPR/DFR*. Indeed, GO terms found in the co-expressed gene list of CPR1 correspondto large cellular functions (nucleotide binding, protein binding) supporting its involvement in a basal/primary metabolism (**Supplemental Figure 10**; **Supplemental Table 3A**; **Supplemental Table 3B**).

By contrast, many genes co-expressed with CPR2 were associated to oxido-reduction and metabolic processes, reflecting its transcriptional association with metabolic processes requiring many P450s such as specialized metabolisms. In fact, more than 20 genes from the MIA biosynthetic pathway (including P450s) were identified in the co-expressed gene list of *CPR2*, whereas a unique MIA gene was found in *CPR1* and *DFR* lists (**Figure 5C**; **Supplemental Table 3A**; **Supplemental Table 3B**). Altogether, these results emphasize the occurrence of specific metabolic associations for each CPR, *e.g. CPR1* with primary and basal specialized metabolisms; and *CPR2* with specialized metabolisms and notably the biosynthesis of MIA. Due to the absence of P450 gene expression association and reduction capacity, *DFR* was not investigated further in this study.

CPR2 is expressed in leaf and cotyledon tissues hosting MIA biosynthetic steps catalyzed by P450

Although our global analysis of gene co-expression provides valuable information regarding CPRs/metabolic pathway associations, these analyses cannot actually reflect the intricate gene co-expression networks occurring at the tissue and/or cellular levels and are required to ensure efficient P450 reduction. Therefore, since the MIA and phenypropanoid biosynthetic pathways display complex spatial organizations in *C. roseus* (e.g. St-Pierre et al., 1999; Burlat et al., 2004; Mahroug et al., 2006; reviewed in Courdavault et al., 2014), we next studied the cellular distribution of *CPR1* and CPR2 transcripts on cotyledons of germinating seedlings. Given the high expressions of *CPR1*/*CPR2* in this tissue, RNA *in situ* hybridization has been used to monitor cellular distribution. No specific zones of high accumulation of *CPR1* mRNAs were detected but a diffuse barely detectable signal was detected in the whole cotyledon tissues (mostly visible in the spongy parenchyma) with the antisense probe (**Figure 6A**), while no apparent signal was revealed with the sense probe (**Figure 6B**). By contrast, intense signals were observed with the *CPR2* antisense probe in distinct cotyledon cell types including IPAP, xylem and epidermis (**Figure 6C**), which were not revealed with the sense probe (**Figure 6D**). This complex organization of *CPR2* transcript distribution was further compared to the expression of representative *P450s* from both MIA and phenylpropanoid pathways that also exhibited compartmented expression such as *C4H*, *G8H* and *SLS1*. In young leaves, *CPR*2 mRNAs were detected in IPAP, epidermis and xylem once again (**Figure 7A**). Remarkably, this profile corresponds to a superimposition of the expression patterns of each P450 tested including *C4H* whose transcripts were detected in epidermis and xylem (**Figure 7B**), *G8H* expressed in IPAP and *SLS1* displaying mRNA accumulation in epidermis

(**Figure 7D**). Such observation was also confirmed in cotyledons of germinating seedlings (**Figure 7E-H**) and supports a preferential involvement of *CPR2* in the reduction of P450s expressed in tissues hosting high levels of MIA and phenylpropanoid biosynthesis. In contrast, the low ubiquitous expression of *CPR1* argues again for a preferential involvement in basal specialized and primary metabolism. These results also allow proposing that the 79 predicted P450 co-expressed with *CPR2* may be involved in metabolisms located to one of these three tissues (i.e. IPAP, xylem and/or epidermis), whereas for the P450s co-expressed with *CPR1*, no such conclusion may be drawn.

Silencing of CPR1 does not impact the biosynthesis of MIA

To determine the role of CPR1 and CPR2 in the reduction of P450s associated to specialized metabolism *in planta*, we used virus-induced gene silencing (VIGS) to target each protein. Given the relatively high sequence similarity between CPR1 and CPR2 and the presence of highly conserved domains, the identification of silencing sequences that did not result in cross-silencing remained tricky. For instance, these silencing sequences were first designed in the 3'untranslated region of each CPR but did not lead to substantial transcript down-regulation (data not shown). Therefore, we subsequently used partial coding sequences of CPR1 or CPR2 but it was difficult to avoid some degree of cross silencing, though the most divergent sequence regions were selected. qPCR of mRNA isolated from silenced plants indicated that *CPR1* could be silenced to approximately 60% of the empty vector control levels with minimal cross silencing of *CPR2* (**Figure 8A**). However, when *CPR2* was silenced (70% of empty vector control levels), we also observed a decrease in the levels of *CPR1* transcripts (40% of empty vector controls) (**Figure 8B**). Nevertheless, we used these plants to monitor the levels of four monoterpene indole alkaloids that are abundant in leaf, namely catharanthine, vindoline, vindorosine and serpentine (Figure 8B). First, we noted that CPR1 silencing did not result in any quantitative modifications of the MIA content of the silenced plants, strongly suggesting that CPR1 is not required to MIA biosynthesis and/or CPR2 can fully compensate the lack of CPR1 to reduce P450s associated to the production of MIA. By contrast, in CPR2 silenced plants, a 45% decrease of the total analyzed MIA content was observed, which mainly resulted from catharanthine and vindorosine amount reduction. Albeit we cannot exclude that this MIA decrease is a consequence of both *CPR1* and *CPR2* transcript down-regulation, the more pronounced CPR2 silencing suggests that CPR2 impacts MIA biosynthesis to a greater extent than CPR1.

DISCUSSION

While the specific functions of class I and class II CPR in plants have not been clearly deciphered to date, the Madagascar periwinkle and its complex MIA biosynthesis constitute an attractive model to study relationships between CPRs and specialized metabolism. In contrast to a single gene in fungi and animals, most vascular plant possesses two functional CPRs originating from an ancestral gene duplication, and a more distinct protein without transmembrane domain that does not reduce the P450 tested in this work. Given their distinct N-terminal sequences and jasmonate responsiveness, the two functional periwinkle CPRs cluster in class I and class II CPR and are assumed to play active role in basal primary metabolism/constitutive specialized metabolism and inducible specialized metabolism, respectively (Jensen and Moller, 2010). In agreement with this statement, functional assays and tests of interaction between CPR1 or CPR2 and G8H, SLS1, T16H1, T16H2 and C4H clearly demonstrated that both CPRs are able to interact and to reduce each P450 with an apparent similar efficiency suggesting that both types of CPR did not display any structural specificity towards P450s (**Figure 3**; **Table 1**). However, compilation of genes co-expressed with CPR1 or CPR2 shaded light on a more pronounced specialization of CPR2 expression towards specialized metabolisms and notably MIA biosynthesis. First, CPR1 expression is correlated to a higher number of genes than CPR2, eg 12,506 versus 2,202 transcripts with a few overlapping even at a low PCC value > 0.4, suggesting a broader implication of CPR1 in the general plant physiology (**Figure 5B**). Moreover, the gene list that co-expressed with each CPR do not overlap unless considering low PCC value arguing for a distinct specialization of CPR1 and CPR2 (**Figure 5A**) Such hypothesis is reinforced by the GO terms attributed to genes associated with CPR1 expression that mainly include protein, zinc ion, or nucleic acid binding, and strongly differ from those linked to CPR2 essentially relying on oxido-reduction processes. Furthermore, the uniprot keywords associated to genes correlated to CPR1 mainly include basal cellular functions (**Supplemental Figure 10**). In addition, among the P450s with expression profiles related to CPR1, almost 25% act in primary metabolism such as hormone biosynthesis while the remaining P450s are mostly associated to phenylpropanoid metabolism (**Supplemental Table 3A**; **Supplemental Table 3B**). This is in partial agreement with previous transcriptomic data analyses in Arabidopsis showing that both ATR1 and ATR2 were co-expressed with lignin biosynthetic genes during inflorescence stem development (Sundin et al., 2014). By contrast, almost all the P450s whose expression correlates with

CPR2 are associated with specialized metabolisms and especially with MIA metabolism (**Supplemental Table 3A**; **Supplemental Table 3B**). This is also in agreement with GO terms and unitprot keywords of genes associated to CPR2 that include oxidation-reduction processes and alkaloid metabolism (**Supplemental Figure 10**, **Supplemental Table 3A**; **Supplemental Table 3B**). The strong correlation between CPR2 expression profile and MIA biosynthesis is reinforced by the identification of nearly all the previously identified MIA biosynthetic genes in the gene list associated to CPR2 (**Figure 5C**; **Supplemental Table 3B**). Interestingly, although numerous P450s associated to other specialized metabolism were correlated to CPR2 such as sesquiterpene metabolism, only a few involved in phenylpropanoid metabolism were identified. This is slightly different from the situation described for ATR2 but in such case, the dataset were generated from Arabidopsis leaves subjected to cold treatment under high-light conditions or in lignin biosynthetic gene mutants (Soitamo et al., 2008; Sundin et al., 2014). Taken altogether, specific gene co-expression with each class of CPR adds another layer of complexity to CPR specialization towards specialized metabolisms: the proposed partition of class I and class II CPRs between basal and inducible specialized metabolism cannot be ruled out for phenylpropanoid biosynthesis but exclusive associations of class II CPRs to subclasses of specialized metabolisms and in particular MIA biosynthesis seem to have been acquired.

Our analysis of CPR genes expression in *C. roseus* also supports the specialization of CPR2 expression towards MIA biosynthesis. Although qPCR demonstrated that CPR1 and CPR2 were both expressed in the same organs, as well as G8H, SLS1 and both T16H, transcript level analysis from whole organ RNA preparation cannot discriminate cell-specific correlation. In fact, our RNA in situ hybridization established cell-specific co-localization of CPR2 transcripts, but not CPR1's, with the cumulative mRNA distribution of the tested P450s involved in specialized metabolism (**Figure 4**; **Figure 6; Figure 7**). The biosynthesis of MIA in *C. roseus* leaves is a highly compartmentalized process involving at least four distinct cell types, which include IPAP, hosting all the early steps of the monoterpene precursor synthesis up to loganic acid, epidermis that carries out the incorporation of loganic acid into MIA including desacetoxyvindoline and the specialized cells laticifers and idioblasts where take place the last two steps of vindoline formation. On the basis of this architecture, the expression of multiple P450s associated to MIA biosynthesis has been shown to be restricted to IPAP (G8H, IO, 7DLH) or to epidermis (SLS1-4, T16H1, T16H2, 16T3O), that constitute two leaf tissues of high P450 activity. In our localization studies, while the CPR1 transcripts

were barely detectable in cotyledons and leaves with an apparent general distribution, we showed that CPR2 mRNAs were highly accumulated, together with G8H, in IPAP; in epidermis, like SLS1 and T16H2 messengers; and to a lower extend in xylem, where C4H transcripts were also detected. The CPR2 expression profile appears thus highly compartmentalized, being superimposed to that of MIA and phenylpropanoid biosynthetic genes. This high CPR2 expression in IPAP suggests thus that type II CPR is required to G8H activity in this tissue, and by extension to IO and 7DLH, while type I CPR seems to play a house keeping role given its uniform distribution and low transcript level. The same rationale could be applied to leaf epidermis in which CPR2, SLS1, T16H2 are highly transcribed, reinforcing the specialization of CPR2 in MIA biosynthesis. Data concerning tissue specific expression of CPR are scarce. Therefore, demonstration of the specific co-occurrence of CPR and P450 transcripts in similar restricted tissues provides a plausible explanation of how CPR specialization is established *in planta*, based on similar transcriptional regulatory processes. This may also shed light on a predictive tissue-specific expression for the list of P450s coexpressed with CPR2, and to the foreseen gathering of various specialized metabolisms in given specialized cell types as exemplified for epidermis (Mahroug et al., 2006).

CPR specialization was also supported by the VIGS assays conduced in *C. roseus*. While characterization of CPR mutant has already been reported in Arabidopsis (Sundin et al., 2015), gene silencing approaches have not been applied to date to CPRs except in animal system (Mackenzie et al., 2010; Tang et al., 2012; Liu et al., 2015). Therefore, the silencing of CPR carried out in this work demonstrates its feasibility in plants. More importantly, it established the lack of effect of CPR1 silencing on MIA biosynthesis (**Figure 8**). Albeit we cannot exclude a possible compensation by CPR2 in CPR1 silenced plants, this result tempts to weaken once again the role of CPR1 in the reduction of P450s associated to MIA production. Compensation between the two classes of CPRs seems unlikely given their highly specific gene expression profiles (**Figure 6**; **Figure 7**) but also because it had not been reported in ATR2 mutants which presented altered lignin composition (Sundin et al., 2015). However, given sequence identity, we were not able to specifically down-regulate CPR2, resulting in a huge decrease of CPR2 transcripts accompanied by a less marked but real decrease (40%) of CPR1 messengers in CPR2 silenced plants (**Figure 8A**). Such transcript down-regulations engendered a significant decrease of the total leaf MIA content (around 45%) in which the main leaf MIAs were mostly affected (**Figure 8B**). Given its specific tissue expression profiles, the more pronounced transcript down-regulation and the absence of such

a decrease in CPR1 VIGS lines, we anticipate that this MIA decrease mainly resulted from CPR2 downregulation. Since class I CPRs have been suggested to play a role in basal specialized metabolism, one could speculate that part of the MIA decrease could also be a consequence of CPR1 downregulation by CPR2 silencing. However, since no effect of CPR1 silencing on MIA accumulation has been observed in our VIGS assays, this seems unlikely. As a consequence, it raises the question of the frontiers of the association of class I and class II CPRs with basal and/or inducible specialized metabolisms. In *C. roseus*, the biosynthesis of MIA appears as constitutive process that can also be stimulated by external stimuli or defense-stimulated phytohormones (El-Sayed and Verpoorte, 2007). Overall, our results tend to indicate that CPR2 has a distinct function in the specialized metabolism both at the basal and developmentaly-regulated level as well as in the response to environmental signals. By contrast, gene expression correlation reinforces the potential involvement of class I CPR with a different set of CYPs involved in other specialized metabolisms.

In summary, our study gives compelling evidence that acquisition of a second isoform of CPR (class II CPR) in the plant lineage provides an additional level of control of the redox power homeostasis. With respect to the hazard of CPR overexpression (ROS production, Bassard et al., 2012; Paddon et al, 2013), strict adjustment of cytochrome reducing power to the metabolic demands may take advantage of the transcriptional regulation of two independent genes, with their specific regulatory control. Class I CPR transcriptional control is compatible with reducing power required for low level, basal, primary and secondary metabolism, while class II CPR fulfills electrons for highly expressed CYPs involved in tissue-specific and intensive specialized metabolism. A such and with the indubitable requirement of class II CPR in the biosynthesis of MIA, our work provides pioneering hypotheses towards the so-far poorly characterized specialization of the different CPR classes. Albeit the *C. roseus* model has been particularly appropriated to conduct this study given the complexity and unicity of MIA biosynthesis, it only reflects one possible strategy of evolution arising in one clade and could not afford an exhaustive view of CPR specialization strategies. In this way, it has been very recently demonstrated in Apiales lineage that class I CPR genes have been lost and compensated by the acquisition of additional isoforms of class II CPR (Andersen et al, 2016).

MATERIALS AND METHODS

Chemicals

Secologanin and tabersonine were purchased from Phytoconsult (The Netherlands) and ChromaDex (USA). Cinnamic acid, geraniol and 10-hydroxygeraniol were purchase from Sigma-Aldrich.

Plant and cell culture growth

Fully expended *C. roseus* plants, cultivars Pacifica Pink and Sunstorm Apricot, were used for microscopy fixation (RNA *in situ* hybridization experiments) and RNA extraction (cloning experiments), respectively. Virus induced gene silencing assays were performed on the *C. roseus* Sunstorm Apricot cultivar. The *C. roseus* C20A cell suspension culture used for subcellular localization studies was propagated in Gamborg B5 medium (Duchefa, NL) at 24°C under continuous shaking (100 rpm) for 7 days as previously described (Guirimand et al., 2009).

Transcriptomic resources

Identification of CPR homologous sequences was achieved by interrogating *C. roseus* transcriptomic databases including the consensus *C. roseus* transcriptome (Dugé de Bernonville et al., 2015; http://bbv-ea2106.sciences.univ-tours.fr/images/files_to_download/BAFC94.fasta), the Medicinal Plant Genomic Resource database (Góngora-Castillo et al., 2012; http://medicinalplantgenomics.msu.edu) and the Phytometasyn database (Xia et al., 2013; http://www.phytometasyn.ca).

RNA extraction and reverse transcription

Total RNA was extracted from young leaves using the NucleoSpin RNA Plant kit (Macherey-Nagel, France). First-strand cDNA was synthesized from 0.5 µg of total RNA using oligo(dT)18 primers (0.5 µM) and 15 units of Thermoscript reverse transcriptase (Invitrogen). Following reverse transcription, complementary RNA was removed by treatment with *E. coli* RNase H (Invitrogen) for 20 min at 37°C.

Subcellular localization, fusion/deletion experiments and protein interaction studies

The subcellular localizations of CPR1, CPR2 and CPR3 were determined by expressing YFP-fusion proteins. The full length ORF of each CPR was amplified using specific couples of primers (**Supplemental Table 4**) and cloned into the *Spe*I restriction site

of the pSCA-cassette YFPi plasmid in frame with the 5' extremity of the YFP coding sequence, to generate the CPR1-YFP, CPR2-YFP and CPR3-YFP fusion proteins (Guirimand et al., 2009). Functionalities of the membrane anchoring domains of CPR1 and CPR2 were studied by fusion/deletion experiments as follows. The coding sequences of the 53- and 74-first residues of CPR1 and CPR2 were amplified using primers helixCPR1-for/helixCPR1-rev and helixCPR2-for/helixCPR2-rev (Supplemental Table X) and cloned into the *Spe*I restriction site of the pSCA-cassette YFPi plasmid to express the hxCPR1-YFP and hxCPR2-YFP fusion proteins. To express the CPR1 and CPR2 proteins deprived of their membrane anchoring domain, the coding sequence of the remaining parts of each protein (54-691 for CPR1 and 75-714 for CPR2) were amplified using primers delCPR1/CPRnewrev and delCPR2/CPRoldrev (**Supplemental Table 4**) and cloned into the *Spe*I restriction site of pSCA cassette YFPi. The addition of the CPR2 membrane anchoring domain to CPR3 was achieved through amplification of the coding sequence of the 74 first residues of CPR2 using primers helixCPR-for and helixCPR-rev, harboring a *Xba*I and *Spe*I restriction site at their extremity, respectively. The resulting PCR product was cloned into the *Spe*I site of pSCA-YFP to generate pSCA-helixYFP. This plasmid was subsequently linearized by *Spe*I to allow the introduction of the CPR3 coding sequence yielding the pSCA-hxCPR3-YFP plasmid. For protein interaction analyses, the same CPR amplification products were cloned into the *Spe*I restriction site of the pSCA-SPYNE 173 plasmid in frame with the 5' end of the sequence encoding the N-terminal YFP split fragment (YFP^N) to express the CPR1-YFP^N, CPR2-YFP^N, CPR3-YFP^N proteins (Guirimand et al., 2010). The coding sequences of the five tested P450 (G10H, SLS1, T16H1, T16H2 and C4H) were amplified using appropriated primers (Supplemental Table X) and cloned into the *Spe*I and/or *Bgl*II restriction sites of the pSCA-SPYCE(M) plasmid in frame with the 5'end of the sequence encoding the C-terminal YFP split fragment (YFP^C) to express G10H-YFP^C, SLS1-YFP^C, T16H1-YFP^C, T16H2-YFP^C and C4H-YFP^C. These recombinant plasmids were used for transient transformation of *C. roseus* cells by particle bombardment and YFP imaging according to Guirimand et al. (2009, 2010) and Foureau et al. (2016). Briefly, *C. roseus* plated cells were bombarded with DNA-coated gold particles (1 µm) and 1,100 psi rupture disc at a stopping-screen-to-target distance of 6 cm, using the Bio-Rad PDS1000/He system. Cells were cultivated for 16 h to 38 h before being harvested and observed. The subcellular localization was determined using an Olympus BX-51 epifluorescence microscope equipped with an Olympus DP-71 digital camera and a combination of YFP and CFP filters. The pattern of localization presented in this work is representative of circa 50 observed cells. The ER or nucleocytosolic localizations of the

different fusion proteins were confirmed by co-transformation experiments using the ER-CFP marker (CD3-954; Nelson et al., 2007) and the nucleocytosolic CFP marker (Guirimand et al., 2011). Such plasmid co-transformations were performed using 400 ng of each plasmid or 100 ng for BiFC assays.

Tissue fixation, embedding in paraffin and sectioning

RNase-free conditions were strictly observed for all steps. All glassware was baked for 8 h at 180°C and non-disposable plastic wares were incubated for 10 min in an aqueous 3% H_2O_2 solution before washing in diethylpyrocarbonate (DEPC) treated water. Leaves from mature *C. roseus* plants grown in green house and young germinating seedlings were rapidly fixed in formalin / acetic acid / alcohol and embedded in Paraplast (Dominique Dutscher, Brumath, France) as described previously (Mahroug et al., 2006; Guirimand et al., 2011b). Serial sections (10 µm) were spread on aminopropyltriethoxysilane -coated slides overnight at 40 °C, and paraffin was removed using xylene (twice for 15 min) before rehydration in an ethanol gradient series up to DEPC treated water.

In situ RNA hybridization of C. roseus leaves and cotyledons

Full-length CPR1 and CPR2 cDNAs cloned into pGEMT-easy vector (Promega) were used for the synthesis of sense and anti-sense RNA probes as previously described (Mahroug et al.,2006). For G8H, SLS1, C4H, previously described plasmids were used for the riboprobe *in vitro* transcription (Burlat et al., 2004; Irmler et al., 2000; Mahroug et al., 2006). Paraffin-embedded serial longitudinal sections of young leaves and cross sections of emerging cotyledons were hybridized with digoxigenin-labeled transcripts and localized with anti-digoxigenin–alkaline phosphatase conjugates according to Mahroug et al. (2006).

Heterologous expression of C. roseus CPRs and P450s in yeast

Full length CPR1, CPR2 and CPR3 cDNAs were amplified using the specific yeast expression primers described in **Supplemental Table 4**. Each couple of primers includes appropriated restriction sites at their extremities to allow cloning of the resulting PCR products into the *Spe*I site of pESC-Leu yeast expression plasmid harboring the LEU2 nutritional marker (Agilent Technologies). Addition of the CPR2 membrane anchoring domain to CPR3 was achieved as described in the previous section except that cloning were performed into the pESC-Leu plasmid. G8H, SLS1, T16H1, T16H2 and C4H coding sequences were amplified using primers including *Bgl*II restriction at their extremities to

allow cloning into the *BamHI* site of the pYeDP60 plasmid harboring URA3 and ADE2 nutritional markers (Pompon et al., 1996). pESC-Leu CPRs, pYeDP60 P450s recombinant vectors and/or empty plasmids were associated by pair and used to transform either the *Saccharomyces cerevisiae* strain WT303 (containing the endogenous yeast CPR) or the WAT11 expressing ATR1 (Pompon et al., 1996). Leu+, Ura+/Ade+ or Ura+/Ade+/Leu+ transformants were selected onto solid complete synthetic medium (CSM) plates containing 0.67% Yeast Nitrogen Base (YNB), 2% agar, 2% dextrose and 0.05% DOB-Leu, DOB-Ura-Ade or DOB-Ura-Ade-Leu as required. Yeast transformants were grown in 10 ml of appropriate CSM liquid medium until reaching the stationary phase of culture and prior being harvested by centrifugation. Protein expression was induced by cultivating the harvested yeast in 50 ml of liquid YPGal medium (1% bactopeptone, 1% yeast extract, and 2% galactose) for 6 h as described in Besseau et al. (2013).

Enzyme assays

Following induction of protein expression, 50 mL of yeast culture were harvested by centrifugation and resuspended in 2 mL of buffer R (Tris–HCl pH7.5, 50 mM; EDTA 1 mM) in a 50 ml centrifugation tube. An equal volume of glass beads were added (425–600 µm, Sigma) and cells were broken by vigorous shaking. For this purpose, tubes were shaken by hand during 30 s in a cold room (4 °C) before being put on ice for 30 additional seconds. This operation was repeated ten times before the addition of two volumes of buffer R allowing the recovering of the yeast crude extracts prior to protein quantification using the Bio-Rad protein microassay. P450s activities were analyzed in a final volume of 100 µl containing 300 µg of proteins, 100 µM of NADPH,H^+ and 20 µM of either loganin for SLS1, tabersonine for T16H1 and T16H2 or cinnamate for C4H. Reactions were initiated by addition of NADPH,H+, incubated at 30 °C during 0, 5, 15, 30 min (T16H1, T16H2 and C4H) or during 10, 30, 60 or 120 min (SLS1) and quenched by addition of 100 µl of methanol prior to ultra-performance liquid chromatography-mass spectrometry analysis (UPLC-MS). G10H activity was tested using microsomal membranes purified according to Heitz et al. (2012). Enzymatic assays were conducted according to Höfer et al. (2013) in a final volume of 100 µl of citrate-phosphate buffer pH 7.4, 20 mM, containing 400 µM of NADPH, 200 µM of geraniol and normalized amounts of microsomes. After 60 min of reaction, products were extracted with 500 µl of ethyl acetate. Ethyl acetate phase was collected with a glass pipette, dried under a nitrogen gas flow and products were analyzed on GC-FID. For all P450 tests performed in the WT303 yeast strain including CPR1, CPR2 or CPR3, activities resulting from the yeast

endogenous CPR was subtracted from total activity to estimate the activity resulting from each periwinkle CPR. All results were expressed as substrate conversion rate.

UPLC-MS analyses

All samples were centrifuged and the supernatants were stored at 4 °C prior to injection. UPLC chromatography system consisted in an ACQUITY UPLC (Waters, Milford, MA, USA). Separation was performed using a Waters Acquity HSS T3 C18 column (150 mm × 2.1 mm, 1.8 μm) with a flow rate of 0.4 mL/min at 55 °C. The injection volume was 5 μL. The mobile phase consisted of solvent A (0.1 % formic acid in water) and solvent B (0.1 % formic acid in acetonitrile). Chromatographic separation was achieved using an 8-min linear gradient from 10 to 24 % solvent B. MS detection was performed by using a SQD mass spectrometer equipped with an electrospray ionization (ESI) source controlled by Masslynx 4.1 software (Waters, Milford, MA). The capillary and sample cone voltages were 3,000 V and 30 V, respectively. The cone and desolvation gas flow rates were 60 and 800 Lh−1. Data collection was carried out in negative mode for secologanin ([M + HCOOH-H]− = 433, RT = 7.12 min), cinnamate ([M - H]-=147), 4-hydroxycinnamate ([M - H]-=163) and in positive mode for loganin ([M + Na]+ = 413, RT= 5.61 min), tabersonine ([M + H]+ = 337, RT = 4.72 min), and 16-hydroxytabersonine ([M + H]+ = 353, RT = 3.87 min).

GC analyses

Monoterpenes were analyzed by a GC-FID apparatus (Alpha-MOS). Samples were injected in the split mode (50 mL/min) and compounds were separated on the BPX5 capillary column. The injector was heated at 250°C and the oven was set at 45°C for 60s. The temperature was next increased up to 250°C at a rate of 8°C/min, followed by an increase to 320°C at a rate of 30°C/min. The oven was maintained at 320°C for 1 min before the end of analysis. The Flame Ionisation Detector was set at 280°C. Peaks were identified by comparing the retention times of authentic standards.

Gene expression correlation analyses

Pearson correlation coefficients were calculated with R (R Development Core Team ,2015) for each CPR with each transcript predicted in the CDF97 consensus assembly and over a total of 16 PE-samples (SRR1144633, SRR1144634, SRR1271857, SRR1271858, SRR1271859, SRR342017, SRR342019, SRR342022, SRR342023, SRR646572, SRR646596, SRR646604, SRR648705, SRR648707, SRR924147, SRR924148) and 23 SE-

samples (SRR122239, SRR122240, SRR122241, SRR122242, SRR122243, SRR122244, SRR122245, SRR122246, SRR122247, SRR122248, SRR122249, SRR122250, SRR122251, SRR122252, SRR122253, SRR122254, SRR122255, SRR122256, SRR122257, SRR122258, SRR122259, SRR122260, SRR122261) as described in Dugé de Bernonville et al 2015a. Annotation of transcripts was performed by following the Trinotate pipeline. Intersections between lists of co-expressed gene lists were visualized with the 'venn' function of the 'gplots' R package. GO terms were associated according to the Uniprot database. Gene set enrichment analyses were made by comparing the observed representation of GO terms with a theorical hypergeometric distribution ('phyper' function in R).

Gene expression analyses

CPR1, CPR2 and CPR3 expression was first determined by comparing transcript abundance in *C. roseus* transcriptomic datasets (Góngora-Castillo et al., 2012; Dugé de Bernonville et al., 2015) and subsequently measured by real-time RT-qPCR using primers listed in **Supplemental Table 5**. The corresponding PCR products were cloned in pGEM-Teasy (Promega) and sequenced to ensure the specificity of amplification. Primer efficacy was evaluated on plasmids containing the appropriated cDNA. Distinct *C. roseus* organs (such as roots, stems, young and mature leaves, petals – Apricot sunstorm cultivar) were immediately frozen in liquid nitrogen after sampling. Samples (50 mg) were ground with a mortar and a pestle in liquid nitrogen and total RNA were extracted with the Nucleospin RNA plant (Macherey-Nagel), quantified with a Nanodrop spectrophotometer (ThermoFisher) and treated (1.5 µg) with RQ1 RNase-free DNase (Promega) before being used for first strand cDNA synthesis by priming with oligo (dT) 18 (0.5 µM). Retro-transcription (RT) of 1.5 µg of total RNA was carried out using the SuperScript III reverse transcriptase kit (Invitrogen) at 50 °C during 1 h according to manufacturer's instructions. Real-time PCR was run on a CFX96 Touch Real- Time PCR System (Bio-Rad) using the SYBR Green I technology. Each reaction was performed in a total reaction volume of 25 µL containing an equal amount of cDNAs (1/3 dilution), 0.05 µM forward and reverse primers, and 1 × DyNAmo™ ColorFlash Probe qPCR Kit (Termo Fisher Scientific). The amplication program was 95 °C for 7 min (polymerase heat activation), followed by 40 cycles containing 2 steps, 95 °C for 10 sec and 60 °C for 40 sec. At the end of the amplification, a melt curve was performed to check amplification specificity. Relative quantification of transcripts was performed with calibration curves and normalization with the *C. roseus* 40S Ribosomal protein S9 (RPS9, AJ749993.1)

reference gene. All amplifications were performed in triplicate and repeated at least on two independent biological repeats.

Virus Induced Gene Silencing

CPR1 and CPR2-silencing fragments were amplified using the primers described in Supplemental Table I and cloned into the pTRV2u vector described by Geu-Flores et al. (2012). The resulting plasmids and the empty vector were used to perform the VIGS assays on *C. roseus* seedlings as described by Liscombe and O'Connor (2011) or Carqueijeiro et al. (2015). Leaves from the first two leaf pairs to emerge following inoculation were harvested from eight plants transformed with each construct and subjected to gene expression analysis by real-time RT-PCR (primers are given in **Supplemental Table 5**). The alkaloid content of silenced leaves was determined by LC-MS as described previously (Liscombe and O'Connor, 2011; Geu-Flores et al., 2012).

AUTHOR CONTRIBUTIONS

C.P., E.F., A.L., M.A.L, I.C., N.P., M.C. performed the biochemical characterization of CPRs; E.F. studied protein subcellular localization and interaction; F.K. carried out silencing experiments; V.B., S.M., B.S. analyzed transcript distribution by *in situ* hybridization; TDDB, S.B. achieved bioinformatics analyses; A.O., G.G. conducted analysis of gene expression; N.G.G., B.St-P., L.A., M.C. assisted in the supervision of this work; S.E.O., V.C. conceived, supervised and coordinated this study and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Marie-Antoinette Marquet, Marie-Françoise Aury, Evelyne Danos, Cédric Labarre (EA2106 Biomolécules et Biotechnologies Végétales) for help in maintaining cell cultures and plants. We also thank Emelyne Marais and Céline Melin for their precious technical assistance. We gratefully acknowledge support from the "Région Centre" (France, ABISAL grant and Post-Doctoral Fellow attributed to C. P. and I. C.) and from the University of Tours. We would like also to acknowledge the Fédération CaSciModOT (CCSC, Orléans, France) for access the Région Centre computing grid.

FUNDINGS

This research has received funding from "Région Centre" of France (ABISAL Project). C.P. and I.C. were supported by a postdoctoral fellowship from "Région Centre". E.

F. was financed by a fellowship from the Ministère de l'Education Nationale, de l'Enseignement Supérieu r et de la Recherche (France).

FIGURE LEGENDS

Figure 1. Main characterized steps of MIA biosynthesis in *C. roseus* aerial organs highlighting several steps involving P450s. Color code indicates the cellular organization of the pathway as follows: blue, pink and yellow rectangles for compartmentation in IPAP, epidermis and laticifers/idioblasts, respectively. Noteworthy metabolites are depicted such as geranyl-PP that constitutes the entry point of MIA biosynthesis, loganic acid and desacetoxyvindoline transported from IPAP to epidermis and from epidermis to laticifers/idioblasts respectively, strictosidine the first MIA and vindoline which is the main MIA accumulated in leaves. P450s are highlighted by red rectangles and reducing power by CPR indicated by red arrow lines. White arrowhead indicates multiple not yet discovered enzymes. GPPS, geranyl diphosphate synthase; GES, geraniol synthase; G10H (CYP76B6), geraniol 10- hydroxylase; CPR, cytochrome P450-reductase; 10HGO, 10-hydroxygeraniol oxidoreductase; IS, iridoid synthase; 7DLGT, 7-deoxyloganetic acid glucosyltransferase; 7DLH (CYP72A224), 7-deoxyloganic acid hydroxylase; LAMT, loganic acid O-methyltransferase; SLS1 to 4 (CYP72A1), secologanin synthase isoform 1 to 4; STR, strictosidine synthase; SGD, strictosidine β-glucosidase; T16H1-2 (CYP71D351), tabersonine 16-hydroxylase isoform 1 and 2; 16OMT, 16-hydroxytabersonine O-methyltransferase; NMT, 3-hydroxy-16-methoxy-2,3-dihydrotabersonine N-methyltransferase; T3O (CYP71D1) , tabersonine 3-oxygenase; T3R, tabersonine 3-reductase; D4H, desacetoxyvindoline 4-hydroxylase; DAT, deacetylvindoline 4- acetyltransferase.

Figure 2. CPR1 and CPR2 are located to the endoplasmic reticulum while CPR3 is a nucleocytosolic protein. *C. roseus* cells were transiently transformed with the CPR1-YFP (A), CPR2-YFP (E) or CPR3-YFP (I) expressing vectors in combination with plasmids expressing either an ER-CFP marker ("ER"-CFP; B, F) or a nucleocytosolic marker (CFP; J). Co-localization of the fluorescent signals appears on the merged images (C, G, K). Cell morphologies (D, H, L) were observed with differential interference contrast (DIC). Bars, 10 μm.

Figure 3. CPR1, CPR2 but not CPR3 interact with G10H, SLS1, T16H1, T16H2 and C4H. CPRs and P450s interactions were analyzed by BiFC in *C. roseus* cells transiently transformed by plasmids encoding fusions indicated on the top (CPR1 or CPR2 fused to the

split YFP^N fragment) and on the left (CPR1, CPR2, G10H SLS1, T16H1, T16H2 and C4H fused to the split YFP^C fragment). The YFP signal emanating from BiFC complex reformation is show in green false color (A, C, E, G, I, K, M, O, Q, S, U, W) and cell morphology is observed with differential interference contrast (DIC; D, D, F, H, J, L, N, P, R, T, V, X). Bars, 10 µm.

Figure 4. CPR1, CPR2 and DFR transcript abundance in *C. roseus* organs and response to methyljasmonate treatments. The abundance of transcripts was determined by comparing fpkm values of CPR1, CPR2 and CPR3 locus from the main *C. roseus* transcriptomic dataset. MeJA, methyljasmonate. CPR1: open bars, CPR2: grey bars and CPR3: black bars

Figure 5. Analysis of CPR1, CPR2 or DFR gene co-expression correlation. Sizes of intersections between co-expressed gene lists (A). Lists of co-expressed genes were obtained after calculating PCC of *C. roseus* CPR expression levels with each other transcript found in CDF97v2 assembly and setting different PCC threshold values. Sizes of co-expressed gene lists for each CPR at different PCC thresholds (B). Functional composition of co-expressed genes with PCC > 0.4 with each CPR (C). Pearson correlation coefficients of transcripts related to alkaloid metabolism (according to Uniprot annotation) with CPR1 and CPR2. When annotation did not correspond to a deposited protein from Catharanthus, the corresponding protein name is followed by the name of initial plant (RAUSE, *Rauwolfia serpentina*; PAPSO, *Papaver somniferum*; ESCCA, *Eschscholtzia californica*; TAXCU, *Taxus cuspidata*). See **Supplemental Table** 3B for full gene information.

Figure 6. Localization of *CPR1* and *CPR2* transcripts in cotyledons of *C. roseus* germinating seedlings. The analysis of *CPR1* and *CPR2* transcript distribution was performed by *in situ* RNA hybridization. Serial sections of germinating seedlings were hybridized either with *CPR1* antisense (AS) probes (A), *CPR2* antisense probes (C), *CPR1* sense (S) probes (B) or *CPR2* sense probes (D) used as negative controls. Ep, epidermis; IPAP, Internal Phloem Associated Parenchyma; X, Xylem. Bars = 100 mm.

Figure 7. *CPR2* is expressed in leaf and cotyledon tissues hosting transcripts of P450s involved in MIA and phenylpropanoid metabolisms. *CPR2*, *C4H*, *G8H* and *SLS1* transcript localizations were carried out by RNA *in situ* hybridization performed on young leaves (A-D)

and cotyledons (E-H). Serial sections were hybridized either with *CPR2* antisense probes (A, E), C4H antisense probes (B, F), *G8H* antisense probes (C, G) or *SL S1* antisense probes (D, H). Ep, epidermis; IPAP, Internal Phloem Associated Parenchyma; X, Xylem. Bars = 100 mm.

Figure 8. Silencing of CPR1 does not impact MIA biosynthesis. A, Down-regulation of CPR1 and CPR2 transcript by VIGS. The relative expression of each gene was determined by real-time RT-PCR analyses performed on total RNA extracted from *C. roseus* leaves of CPR1 or CPR2-silenced plants (dark gray and light gray bars, respectively) or plants transformed with an empty vector control (EV; dark bars). CrRPS9 was used as a reference gene. Data were normalized to CrRPS9 expression and correspond to average values (n = 8) ± SD of independent transformed plants. Letters indicate statistical classes (Wilcoxon rank sum test, FDR-adjusted *p*-value<0.05). B, Relative MIA content of CPR1- and CPR2-VIGS plants expressed relatively to that of EV plants. The relative MIA content was determined by quantification of catharanthine (white), vindorosine (drak grey), vindoline (light grey) and serpentine (dark)performed by LC-MS. The amount of each MIA in silenced plants (8 plants per gene) was expressed relatively to that measured in EV plants (8 plants; normalized to 100%). Asterisks denote statistical significance (*P,0.005, **P,0.0005, ***P,0.00005).

Supplemental Figure 1. Alignment of periwinkle CPR1, CPR2 and CPR3 deduced amino-acid sequences. Identity and similarity are show with black and grey highlighting respectively.

Supplemental Figure 2. Tree

Supplemental Figure 3. Gene organizations of CPR1, CPR2 and CPR3/DFR. Exons and introns are indicated by boxes and solid lines, respectively.

Supplemental Figure 4. Alignment of CPR1 and CPR2 amino-acid sequences highlighting characteristic membrane anchor, FMN, FAD, NADPH and P450-binding domains.

Supplemental Figure 5. Alignment of periwinkle CPR1, CPR2 and CPR3 deduced amino-acid sequences and determination of the putative functional domains of CPR3 according to Varadarajan et al. (2010).

Supplemental Figure 6. Detection of a putative transmembrane helix at the N-terminal end of CPR1 (A), CPR2 (B) and CPR3/DFR (C). Probability of a residue to belong to a transmembrane helix as calculated for the 100-first amino acids of each CPR with a Markov model by the TMHMM server. Projection of the helical wheel has been done using http://www-nmr.cabm.rutgers.edu/bioinformatics/Proteomic_tools/Helical_wheel/.

Supplemental Figure 7. Characterization of the membrane anchoring domain of CPR1 and CPR2. *C. roseus* cells were transiently transformed with plasmids expressing hxCPR1-YFP (A) corresponding to the first 53 residues of CPR1 fused to YFP, delCPR1-YFP corresponding to the remaining part of CPR1 (AA 54-691) fused to YFP (E), hxCPR2-YFP (I) corresponding to the first 74 residues of CPR2 fused to YFP or delCPR2-YFP corresponding to the remaining part of CPR2 (AA75-715) fused to YFP (M) in combination with plasmids expressing either an ER-CFP marker ("ER"-CFP; B, I) or a nucleocytosolic marker (CFP; F, N). Colocalization of the fluorescent signals appears on the merged images (C, G, K, O). Cell morphologies (D, H, L, P) were observed with differential interference contrast (DIC). Bars, 10 μm.

Supplemental Figure 8. Addition of the CPR2 membrane anchoring domain to CPR3 enables ER anchoring. *C. roseus* cells were transiently transformed with the plasmid pSCA-hxCPR3-YFP hxCPR1-YFP (A) in combination with the plasmid expressing the ER-CFP marker ("ER"-CFP; B). Colocalization of the fluorescent signals appears on the merged images (C). Cell morphology (D) was observed with differential interference contrast (DIC). Bars, 10 μm.

Supplemental Figure 9. Transcript distribution of CPR1, CPR2, C4H, G8H, SLS1, T16H1 and T16H2 in various *C. roseus* organs. Relative expression of each gene was determined by real-time RTPCR analyses performed on total RNA extracted from various *C. roseus* organs including roots (R), stems (S), young leaves (YL), mature leaves (ML) and petals (P). RPS9 was used as a reference gene.

Supplemental Figure 10. Functional classification of nearest co-expressed genes with *C. roseus* CPR. (A). GO Molecular functions. GO terms were obtained after Pfam domain attribution with hmmerscan and with GO annotation in Uniprot. This graph presents for each CPR the 10 GO terms with more than 3 genes that were significantly enriched

(hypergeometric distribution, FDR-adjusted p-value). (B). Keywords from Uniprot. Homologies to Uniprot entries were obtained by Blastx with the Uniprot database. Keywords from the resulting sequences were retrieved by mapping their names to the database (http://www.uniprot.org/uploadlists/).

Table 1. CPR1 and CPR2 reduce C4H, G8H, SLS1, T16H1 and T16H2 with an apparent similar efficiency. Substrate conversion rates (%) were determined using crude protein extract of WT303 yeast strain expressing each CPR/P450 pairs (C4H, SLS1, T16H1 and T16H2 assays) or microsomes (G8H assays) by addition of NADPH,H+ as electron donor. Yeast endogenous CPR activity was estimated by measuring P450 activity in similar conditions without expression of the periwinkle CPRs, and was subtracted from activities measured with CPR1 and CPR2. All assays were conducted independently 3 times with at least three technical replicates.

Supplemental Table 1. Identification of contigs potentially encoding CPR candidates in the MPGR database (http://medicinalplantgenomics.msu.edu/index.shtml), phytometasyn database (http://www.phytometasyn.ca/) and in the *C. roseus* consensus transcriptome (Dugé de Bernonville et al., 2015a).

Supplemental Table 2. Evaluation of the efficiency of P450 reduction by CPR1, CPR2, CPR3/DFR, hxCPR3/DFR, the yeast endogenous CPR (WT303 yeast strain) and the codon-optimized ATR1 of the WAT11 yeast strain. Substrate conversion rates (%) were determined using crude protein extract of WT303 yeast strain expressing each CPR/P450 pairs (C4H, SLS1, T16H1 and T16H2 assays) or microsomes (G8H assays) by addition of NADPH,H+ as electron donor. The hxCPR3 was created by fusion of the 74 first residues of CPR2 including the membrane spanning domain. Similar reactions were performed in the WAT11 yeast strain expressing the codon optimized ATR1 from Arabidopsis. Yeast endogenous CPR activity was also estimated by measuring P450 activity in similar conditions without expression of the periwinkle CPRs, and was subtracted from activities measured with CPR1 and CPR2. Control reactions aiming at evaluating the potential consumption of substrates by yeast endogenous enzymes were carried out using yeast strains transformed with the empty pYeDP60 vector. All assays were conducted independently 3 times with at least three technical replicates. nt, not tested.

Supplemental Table 3. Lists of P450 genes co-expressed with CPR1, CPR2 or DFR. Pink and green highlighting indicates potential association to specialized and primary metabolisms, repectively.

Supplemental Table 3B. Whole list of genes co-expressed with each CPR.

Supplemental Table 4. Primers used for cDNA cloning.

Supplemental Table 5. Primers used for qPCR analyses.

REFERENCES

Andersen, T.B., Hansen, N.B., Laursen, T., Weitzel, C., Simonsen, H.T. (2016) Evolution of NADPH-cytochrome P450 oxidoreductases (POR) in Apiales – POR 1 is missing. Mol. Phylogenet. Evol. **98**:21-28.

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., Werck-Reichhart, D. (2011) Cytochromes p450. Arabidopsis Book. **9**:e0144.

Bassard, J.E., Mutterer, J., Duval, F., Werck-Reichhart, D. (2012) A novel method for monitoring the localization of cytochromes P450 and other endoplasmic reticulum membrane associated proteins: a tool for investigating the formation of metabolons. FEBS J. **279**:1576-1583.

Benveniste, I., Lesot, A., Hasenfratz, M.P., Kochs, G., Durst, F. (1991) Multiple forms of NADPH-cytochrome P450 reductase in higher plants. Biochem. Biophys. Res. Commun. **177**:105-112.

Besseau, S., Kellner, F., Lanoue, A., Thamm, A.M., Salim, V., Schneider, B., Geu-Flores. F., Höfer. R., Guirimand, G., Guihur, A., Oudin, A., Glevarec, G., Foureau, E., Papon, N., Clastre, M., Giglioli-Guivarc'h N., St-Pierre, B., Werck-Reichhart, D., Burlat, V., De Luca, V., O'Connor, S.E., Courdavault, V. (2013) A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in Catharanthus roseus. Plant Physiol. **163**:1792-1803.

Brown, S., Clastre, M., Courdavault, V., O'Connor, S.E. (2015) De novo production of the plant-derived alkaloid strictosidine in yeast. Proc. Natl. Acad. Sci. U S A. **112**:3205-3210.

Burlat, V., Oudin, A., Courtois, M., Rideau, M., St- Pierre, B. (2004) Co- expression of three MEP pathway genes and geraniol 10- hydroxylase in internal phloem parenchyma of Catharanthus roseus implicates multicellular translocation of intermediates during the biosynthesis of monoterpene indole alkaloids and isoprenoid- derived primary metabolites. Plant J. **38**:131-141.

Canto-Canché, B.B., Loyola-Vargas, V.M. (2001) Multiple forms of NADPH-cytochrome P450 oxidoreductases in the madagascar periwinkle Catharanthus roseus. In Vitro Cellular & Developmental Biology-Plant. **37**:622-628.

Carqueijeiro, I., Masini, E., Foureau, E., Sepúlveda, L. J., Marais, E., Lanoue, A., Besseau, S., Papon, N., Clastre, M., Dugé de Bernonville, T., Glévarec, G., Atehortùa, L., Oudin, A., Courdavault, V. (2015) Virus- induced gene silencing in Catharanthus roseus by biolistic inoculation of tobacco rattle virus vectors. Plant Biol. **17**:1242-1246.

Collu, G., Unver, N., Peltenburg-Looman, A.M., van der Heijden, R., Verpoorte, R., Memelink, J. (2001) Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. FEBS Lett. **508**:215-220.

Courdavault, V., Papon, N., Clastre, M., Giglioli-Guivarc'h, N., St-Pierre, B., Burlat, V. (2014) A look inside an alkaloid multisite plant: the Catharanthus logistics. Curr. Opin. Plant Bio. **19**:43-50.

Dugé de Bernonville, T. D., Clastre, M., Besseau, S., Oudin, A., Burlat, V., Glévarec, G., Lanoue, A., Papon, N., Giglioli-Guivarc'h, N., St-Pierre, B., Courdavault, V. (2015a) Phytochemical genomics of the Madagascar periwinkle: Unravelling the last twists of the alkaloid engine. Phytochemistry. **113**: 9-23.

Dugé de Bernonville, T. D., Foureau, E., Parage, C., Lanoue, A., Clastre, M., Londono, M. A., Oudin, A., Houillé, B., Papon, N., Besseau, S., Glévarec, G., Atehortùa, L., Giglioli-Guivarc'h, N., St-Pierre, B., De Luca, V., O'Connor, S.E., Courdavault, V. (2015b). Characterization of a second secologanin synthase isoform producing both secologanin and secoxyloganin allows enhanced de novo assembly of a Catharanthus roseus transcriptome. BMC Genomics. **16**:619.

El-Sayed, M., Verpoorte, R. (2007) Catharanthus terpenoid indole alkaloids: biosynthesis and regulation. Phytochem. Rev. **6**:277-305.

Foureau, E., Carqueijeiro, I., de Bernonville, T. D., Melin, C., Lafontaine, F., Besseau, S., Lanoue, A., Papon, N., Oudin, A., Glévarec, G., Clastre, M., St-Pierre, B., Giglioli-Guivarc'h, N., Courdavault, V. (2016) Prequels to Synthetic Biology: From Candidate Gene Identification and Validation to Enzyme Subcellular Localization in Plant and Yeast Cells. Method. Enzymol. In press- doi:10.1016/bs.mie.2016.02.013.

Geu-Flores, F., Sherden, N.H., Courdavault, V., Burlat, V., Glenn, W.S., Wu, C., Nims, E., Cui, Y., O'Connor, S.E. (2012) An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. Nature. **492**:138-142.

Giddings, L.A., Liscombe, D.K., Hamilton, J.P., Childs, K. L., DellaPenna, D., Buell, C.R., O'Connor, S.E. (2011) A stereoselective hydroxylation step of alkaloid biosynthesis by a unique cytochrome P450 in *Catharanthus roseus*. J. Biol. Chem. **286**:16751-16757.

Góngora-Castillo, E., Childs, K.L., Fedewa, G., Hamilton, J.P., Liscombe, D.K., Magallanes-Lundback, M., Mandadi, K.K., Nims, E., Runguphan, W., Vaillancourt, B., Varbanova-Herde, M., Dellapenna, D., McKnight, T.D., O'Connor, S.E., Buell, C.R. (2012) Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. PloS one. **7**:e52506.

Guengerich, F.P., Munro, A.W. (2013) Unusual cytochrome P450 enzymes and reactions. J. Biol. Chem. **288**:17065-17073.

Guengerich, F.P., Sohl, C.D., Chowdhury, G. (2011) Multi-step oxidations catalyzed by cytochrome P450 enzymes: Processive vs. distributive kinetics and the issue of carbonyl oxidation in chemical mechanisms. Arch. Biochem. Biophys. **507**:126-134.

Guirimand, G., Burlat, V., Oudin, A., Lanoue, A., St-Pierre, B., Courdavault, V. (2009) Optimization of the transient transformation of *Catharanthus roseus* cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. Plant Cell. Reports. **28**:1215-1234.

Guirimand, G., Courdavault, V., Lanoue, A., Mahroug, S., Guihur, A., Blanc, N., Giglioli-Guivarc'h, N., St-Pierre, B., Burlat, V. (2010) Strictosidine activation in *Apocynaceae*: towards a" nuclear time bomb"?. BMC Plant Biol. **10**:182.

Guirimand, G., Guihur, A., Ginis, O., Poutrain, P., Hericourt, F., Oudin, A., Lanoue, A., St-Pierre, B., Burlat, V., Courdavault, V. (2011) The subcellular organization of strictosidine biosynthesis in *Catharanthus roseus* epidermis highlights several trans- tonoplast translocations of intermediate metabolites. FEBS J. **278**:749-763.

Guirimand, G., Guihur, A., Poutrain, P., Héricourt, F., Mahroug, S., St-Pierre, B., Burlat, V., Courdavault, V. (2011) Spatial organization of the vindoline biosynthetic pathway in *Catharanthus roseus*. J. Plant Physiol. **168**:549-557.

Höfer, R., Dong, L., André, F., Ginglinger, J. F., Lugan, R., Gavira, C., Grec, S., Lang, G., Memelink, J., Van der Krol. S., Bouwmeester, H., Werck-Reichhart, D. (2013) Geraniol hydroxylase and hydroxygeraniol oxidase activities of the CYP76 family of cytochrome P450 enzymes and potential for engineering the early steps of the (seco) iridoid pathway. Metab. Eng. **20**:221-232.

Heitz, T., Widemann, E., Lugan, R., Miesch, L., Ullmann, P., Désaubry, L., Holder, E., Grausem, B., Kandel, S., Miesch, M., Werck-Reichhart, D., Pinot, F. (2012) Cytochromes

P450 CYP94C1 and CYP94B3 catalyze two successive oxidation steps of plant hormone jasmonoyl-isoleucine for catabolic turnover. J. Biol. Chem. **287**:6296-6306.

Hotze, M., Schröder, G., Schröder, J. (1995) Cinnamate 4-hydroxylase from *Catharanthus roseus* and a strategy for the functional expression of plant cytochrome P450 proteins as translational fusions with P450 reductase in *Escherichia coli*. FEBS Lett. **374**:345-350.

Irmler, S., Schröder, G., St- Pierre, B., Crouch, N.P., Hotze, M., Schmidt, J., Strack, D., Matern, U., Schröder, J. (2000) Indole alkaloid biosynthesis in *Catharanthus roseus*: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. Plant J. **24**:797-804.

Kellner, F., Geu-Flores, F., Sherden, N.H., Brown, S., Foureau, E., Courdavault, V., O'Connor, S.E. (2015) Discovery of a P450-catalyzed step in vindoline biosynthesis: a link between the aspidosperma and eburnamine alkaloids. Chem. Commun. (Camb). **51**:7626-7628.

Kellner, F., Kim, J., Clavijo, B.J., Hamilton, J.P., Childs, K.L., Vaillancourt, B., Cepela, J., Habermann, M., Steuernagel, B., Clissold, L., McLay, K., Buell, C.R., O'Connor, S.E. (2015) Genome- guided investigation of plant natural product biosynthesis. Plant J. **82**:680-692.

Liscombe, D.K., O'Connor, S.E. (2011) A virus-induced gene silencing approach to understanding alkaloid metabolism in Catharanthus roseus. Phytochemistry. **72**:1969-1977.

Liu, S., Liang, Q.M., Zhou, W.W., Jiang, Y.D., Zhu, Q.Z., Yu, H., Zhang, C.X., Gurr, G.M., Zhu, Z.R. (2015) RNA interference of NADPH–cytochrome P450 reductase of the rice brown planthopper, Nilaparvata lugens, increases susceptibility to insecticides. Pest Manag. Sci. **71**:32-39.

Mackenzie, N.C., Lillico, S.G., Brown, K., Wolf, C.R., Whitelaw, C.B. (2010) Evaluation of RNA-knockdown strategies for modulation of cytochrome P450 reductase activity in mouse hepatocytes. J. RNAi Gene Silencing. **6**:416-421.

Madyastha, K.M., Coscia, C.J. (1979) Detergent-solubilized NADPH-cytochrome c (P-450) reductase from the higher plant, Catharanthus roseus. Purification and characterization. J. Biol. Chem. **254**: 2419-2427.

Mahroug, S., Courdavault, V., Thiersault, M., St-Pierre, B., Burlat, V. (2006) Epidermis is a pivotal site of at least four secondary metabolic pathways in Catharanthus roseus aerial organs. Planta. **223**: 1191-1200.

Mazourek, M., Pujar, A., Borovsky, Y., Paran, I., Mueller, L., Jahn, M.M. (2009) A dynamic interface for capsaicinoid systems biology. Plant Physiol. **150**:1806-1821.

Meijer, A.H., Cardoso, M., Voskuilen, J.T., Waal, A., Verpoorte, R., Hoge, J.H.C. (1993) Isolation and characterization of a cDNA clone from Catharanthus roseus encoding NADPH: cytochrome P- 450 reductase, an enzyme essential for reactions catalysed by cytochrome P-450 mono- oxygenases in plants. Plant J. **4**:47-60.

Miettinen, K., Dong, L., Navrot, N., Schneider, T., Burlat, V., Pollier, J., Woittiez, L., van der Krol, S., Lugan, R., Ilc, T., Verpoorte, R., Oksman-Caldentey, K. M., Martinoia, E., Bouwmeester, H., Goossens, A., Memelink, J., Werck-Reichhart, D. (2014). The seco-iridoid pathway from Catharanthus roseus. Nat.Commun. **5**:3606.

Mizutani, M., Ohta, D. (2010) Diversification of P450 genes during land plant evolution. Annu. Rev. Plant Biol. **61**:291-315.

Mizutani, M., Sato, F. (2011) Unusual P450 reactions in plant secondary metabolism. Arch. Biochem. Biophys. **507**:194-203.

Munro, A. W., Girvan, H.M., Mason, A.E., Dunford A. J., McLean, K.J. (2013) What makes a P450 tick? Trends Biochem. Sci. **38**:140-150.

Paddon, C.J, Westfall, P.J, Pitera, D.J, Benjamin, K., Fisher, K., McPhee, D., Leavell, M.D., Tai, A., Main, A., Eng, D., Polichuk, D.R., Teoh, K.H., Reed, D.W., Treynor, T., Lenihan, J., Jiang, H., Fleck, M., Bajad, S., Dang, G., Dengrove, D., Diola, D., Dorin, G., Ellens, K.W., Fickes, S., Galazzo, J., Gaucher, S.P., Geistlinger, T., Henry, R., Hepp, M., Horning, T., Iqbal, T., Kizer, L., Lieu B., Melis, D., Moss, N., Regentin, R., Secrest, S., Tsuruta, H., Vazquez, R., Westblade, L.F., Xu, L., Yu, M., Zhang, Y., Zhao L., Lievense, J., Covello, P.S., Keasling, J.D., Reiling, K.K., Renninger N.S., Newman, J.D. High-level semi-synthetic production of the potent antimalarial artemisinin. (2013) Nature. **496**:528-532.

Qu, Y., Easson, M.L., Froese, J., Simionescu, R., Hudlicky, T., De Luca, V. (2015a) Completion of the seven-step pathway from tabersonine to the anticancer drug precursor vindoline and its assembly in yeast. Proc. Natl. Acad. Sci. U S A. **112**:6224-6229.

Qu, X., Pu, X., Chen, F., Yang, Y., Yang, L., Zhang, G., Luo, Y. (2015b) Molecular Cloning, Heterologous Expression, and Functional Characterization of an NADPH-Cytochrome P450 Reductase Gene from Camptotheca acuminata, a Camptothecin-Producing Plant. PLoS One. **10**:e0135397.

Rana, S., Lattoo, S.K., Dhar, N., Razdan, S., Bhat, W.W., Dhar, R.S., Vishwakarma, R. (2013) NADPH-cytochrome P450 reductase: molecular cloning and functional characterization of two paralogs from Withania somnifera (L.) dunal. PLoS One. **8**:e57068.

Ro, D.K., Ehlting, J., Douglas, C.J. (2002) Cloning, functional expression, and subcellular localization of multiple NADPH-cytochrome P450 reductases from hybrid poplar. Plant Physiol. **130**:1837-1851.

Schröder, G., Unterbusch, E., Kaltenbach, M., Schmidt, J., Strack, D., De Luca, V., Schröder, J. (1999) Light-induced cytochrome P450-dependent enzyme in indole alkaloid biosynthesis: tabersonine 16-hydroxylase. FEBS Lett. **458**:97-102.

Shephard, E.A., Phillips, I.R., Bayney, R.M., Pike, S.F., Rabin, B.R. (1983) Qantification of NADPH: cytochrome P-450 reductase in liver microsomes by a specific radioimmunoassay technique. Biochem. J. **211**:333-340.

Salim, V., Wiens, B., Masada-Atsumi, S., Yu, F., De Luca, V. (2014) 7-deoxyloganetic acid synthase catalyzes a key 3 step oxidation to form 7-deoxyloganetic acid in Catharanthus roseus iridoid biosynthesis. Phytochemistry. **101**:23-31.

Salim, V., Yu, F., Altarejos, J., De Luca, V. (2013) Virus-induced gene silencing identifies Catharanthus roseus 7-deoxyloganic acid-7-hydroxylase, a step in iridoid and monoterpene indole alkaloid biosynthesis. Plant J. **76**:754-765.

Soitamo, A.J., Piippo, M., Allahverdiyeva, Y., Battchikova, N., Aro, E.M. (2008) Light has a specific role in modulating Arabidopsis gene expression at low temperature. BMC Plant Biol. **8**:13.

St-Pierre, B., Vazquez-Flota, F.A., De Luca, V. (1999) Multicellular compartmentation of *Catharanthus roseus* alkaloid biosynthesis predicts intercellular translocation of a pathway intermediate. Plant Cell. **11**:887-900.

Sundin, L., Vanholme, R., Geerinck, J., Goeminne, G., Höfer, R., Kim, H., Ralph, J., Boerjan, W. (2014) Mutation of the inducible ARABIDOPSIS THALIANA CYTOCHROME P450 REDUCTASE2 alters lignin composition and improves saccharification. Plant Physiol. **166**:1956-1971.

Takei, K., Yamaya, T., Sakakibara, H. (2004) Arabidopsis CYP735A1 and CYP735A2 encode cytokinin hydroxylases that catalyze the biosynthesis of trans-Zeatin. J. Biol. Chem. **279**:41866-41872.

Tang, T., Zhao, C., Feng, X., Liu, X., Qiu, L. (2012) Knockdown of several components of cytochrome P450 enzyme systems by RNA interference enhances the susceptibility of Helicoverpa armigera to fenvalerate. Pest Manag. Sci. **68**:1501-1511.

Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G.J., Rombauts, S., Hasnain, G., Rischer, H., Memelink, J., Oksman-Caldentey, K.M., Goossens, A. (2013) CathaCyc, a metabolic

pathway database built from Catharanthus roseus RNA-Seq data. Plant Cell. Physiol. **54**:673-685.

Varadarajan, J., Guilleminot, J., Saint-Jore-Dupas, C., Piégu, B., Chabouté, M.E., Gomord, V., Coolbaugh, R.C., Devic, M., Delorme, V. (2010) ATR3 encodes a diflavin reductase essential for Arabidopsis embryo development. New Phytol. **187**:67-82.

Xiao, M., Zhang, Y., Chen, X., Lee, E.J., Barber, C.J., Chakrabarty, R., Desgagné-Penix, I., Haslam, T.M., Kim, Y.B., Liu, E., MacNevin, G., Masada-Atsumi, S., Reed, D.W., Stout, J.M., Zerbe, P., Zhang, Y., Bohlmann, J., Covello, P.S., De Luca, V., Page, J.E., Ro, D.K., Martin, V.J., Facchini, P.J., Sensen, C.W. (2013) Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. J. Biotechnol. **166**:122-134.

Yang, C.Q., Lu, S., Mao, Y.B., Wang, L.J., Chen, X.Y. (2009) Characterization of two NADPH: cytochrome P450 reductases from cotton (Gossypium hirsutum). Phytochemistry. **71**:27-35.

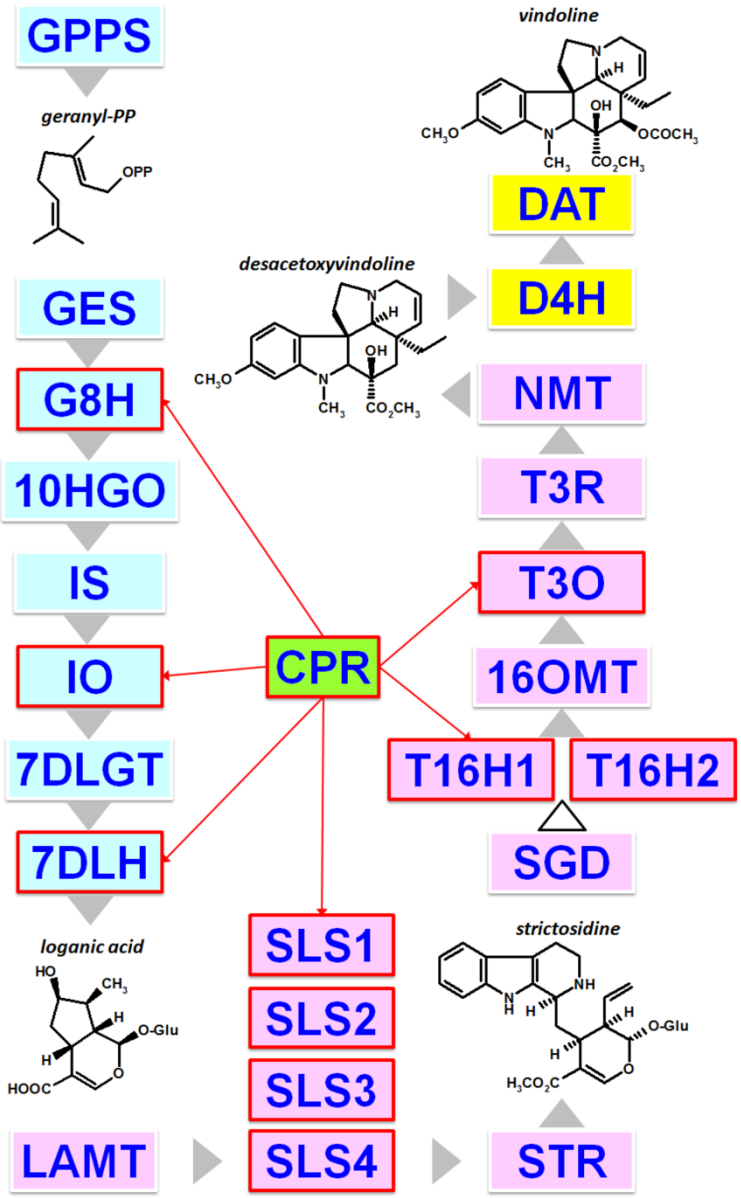


Figure 1. Main characterized steps of MIA biosynthesis in *C. roseus* aerial organs highlighting several steps involving P450s. Color code indicates the cellular organization of the pathway as follows: blue, pink and yellow rectangles for compartmentation in IPAP, epidermis and laticifers/idioblasts, respectively. Noteworthy metabolites are depicted such as geranyl-PP that constitutes the entry point of MIA biosynthesis, loganic acid and desacetoxyvindoline transported from IPAP to epidermis and from epidermis to laticifers/idioblasts respectively, strictosidine the first MIA and vindoline which is the main MIA accumulated in leaves. P450s are highlighted by red rectangles and reducing power by CPR indicated by red arrow lines. White arrowhead indicates multiple not yet discovered enzymes. GPPS, geranyl diphosphate synthase; GES, geraniol synthase; G10H (CYP76B6), geraniol 10- hydroxylase; CPR, cytochrome P450-reductase; 10HGO, 10-hydroxygeraniol

oxidoreductase; IS, iridoid synthase; 7DLGT, 7-deoxyloganetic acid glucosyltransferase; 7DLH (CYP72A224), 7-deoxyloganic acid hydroxylase; LAMT, loganic acid O-methyltransferase; SLS1 to 4 (CYP72A1), secologanin synthase isoform 1 to 4; STR, strictosidine synthase; SGD, strictosidine β-glucosidase; T16H1-2 (CYP71D351), tabersonine 16-hydroxylase isoform 1 and 2; 16OMT, 16-hydroxytabersonine O-methyltransferase; NMT, 3-hydroxy-16-methoxy-2,3-dihydrotabersonine N-methyltransferase; T3O (CYP71D1) , tabersonine 3-oxygenase; T3R, tabersonine 3-reductase; D4H, desacetoxyvindoline 4-hydroxylase; DAT, deacetylvindoline 4- acetyltransferase.

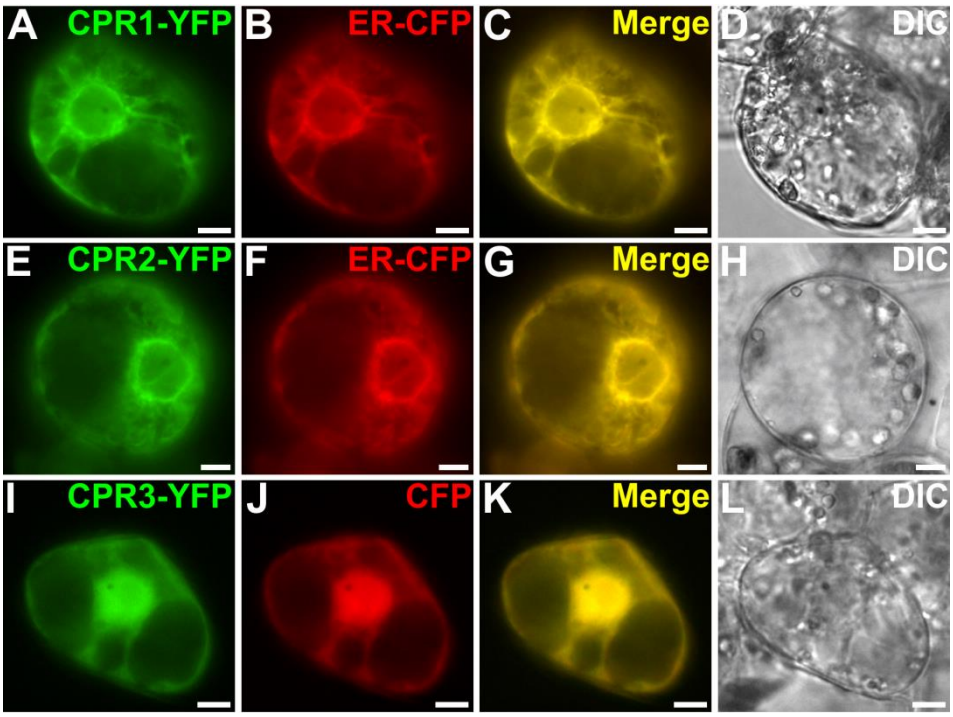


Figure 2. CPR1 and CPR2 are located to the endoplasmic reticulum while CPR3 is a nucleocytosolic protein. *C. roseus* cells were transiently transformed with the CPR1-YFP (A), CPR2-YFP (E) or CPR3-YFP (I) expressing vectors in combination with plasmids expressing either an ER-CFP marker ("ER"-CFP; B, F) or a nucleocytosolic marker (CFP; J). Co-localization of the fluorescent signals appears on the merged images (C, G, K). Cell morphologies (D, H, L) were observed with differential interference contrast (DIC). Bars, 10 μm.

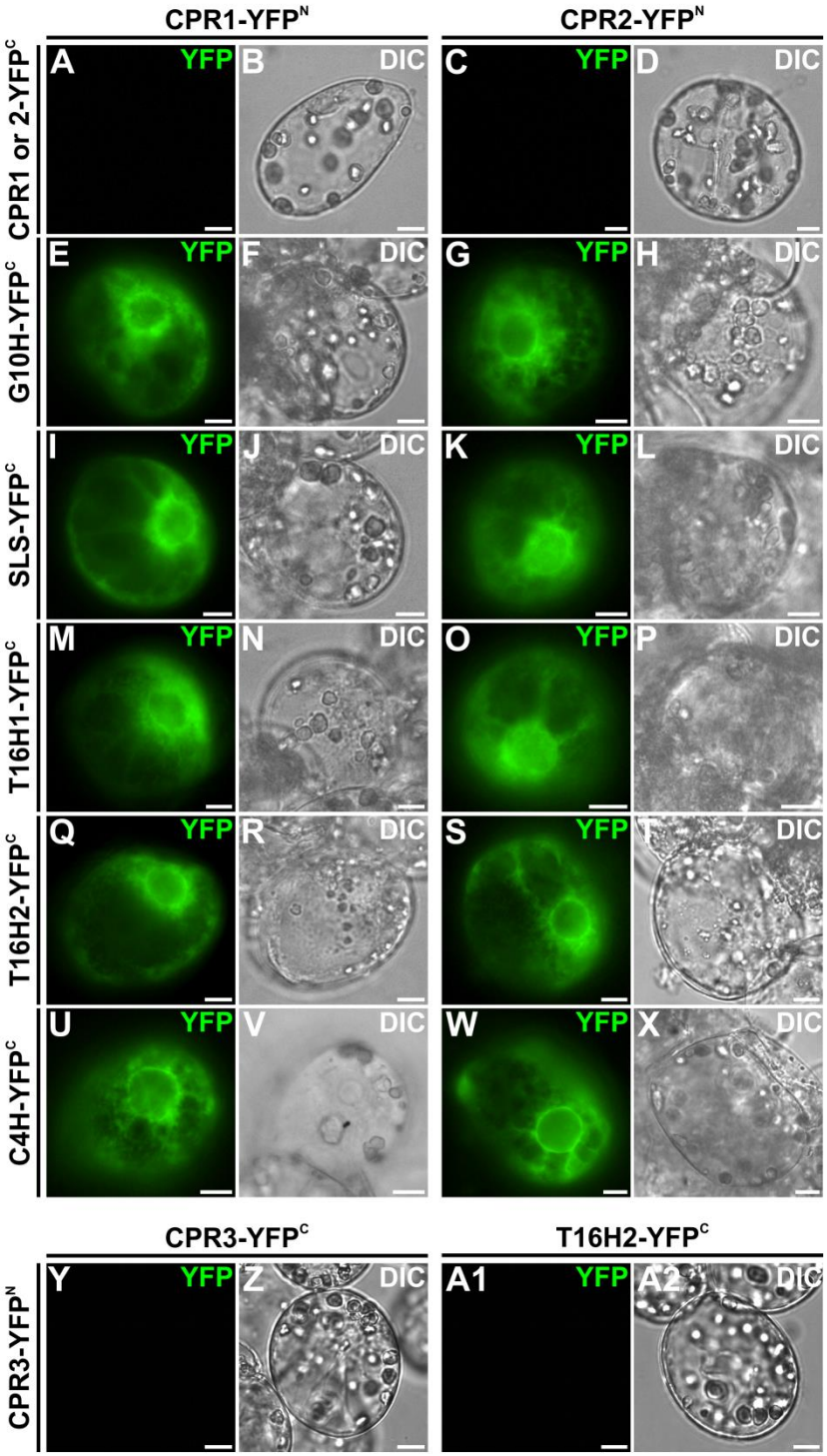


Figure 3. CPR1, CPR2 but not CPR3 interact with G10H, SLS1, T16H1, T16H2 and C4H. CPRs and P450s interactions were analyzed by BiFC in *C. roseus* cells transiently transformed by plasmids encoding fusions indicated on the top (CPR1 or CPR2 fused to the split YFP^N fragment) and on the left (CPR1, CPR2, G10H SLS1, T16H1, T16H2 and C4H fused to the split YFP^C fragment). The YFP signal emanating from BiFC complex reformation is show in green false color (A, C, E, G, I, K, M, O, Q, S, U, W) and cell morphology is observed with differential interference contrast (DIC; D, D, F, H, J, L, N, P, R, T, V, X). Bars, 10 µm.

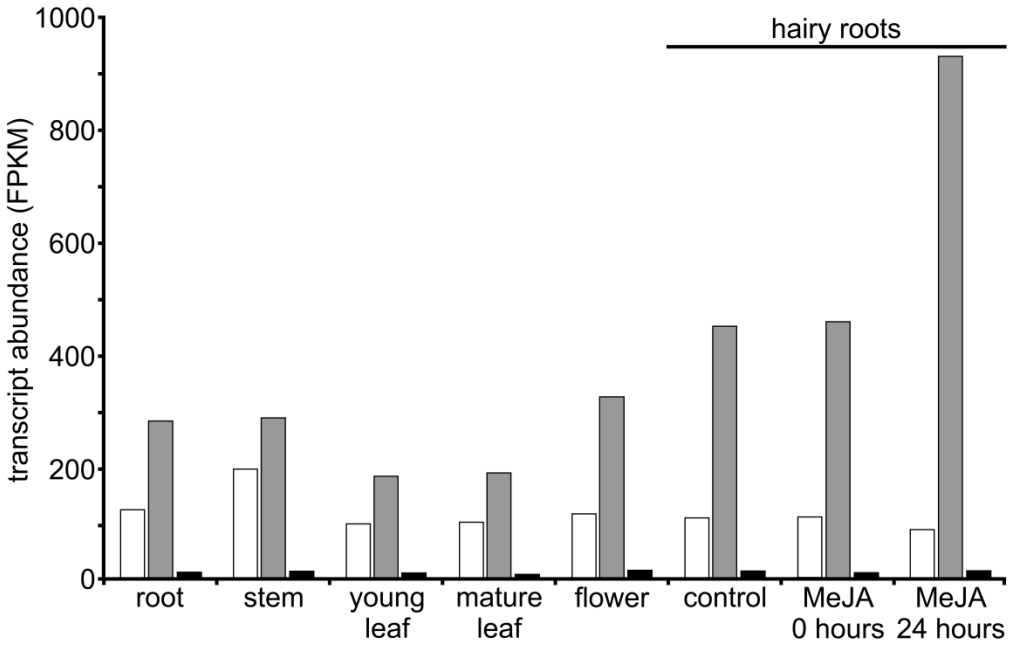


Figure 4. CPR1, CPR2 and DFR transcript abundance in *C. roseus* organs and response to methyljasmonate treatments. The abundance of transcripts was determined by comparing fpkm values of CPR1, CPR2 and CPR3 locus from the main *C. roseus* transcriptomic dataset. MeJA, methyljasmonate. CPR1: open bars, CPR2: grey bars and CPR3: black bars

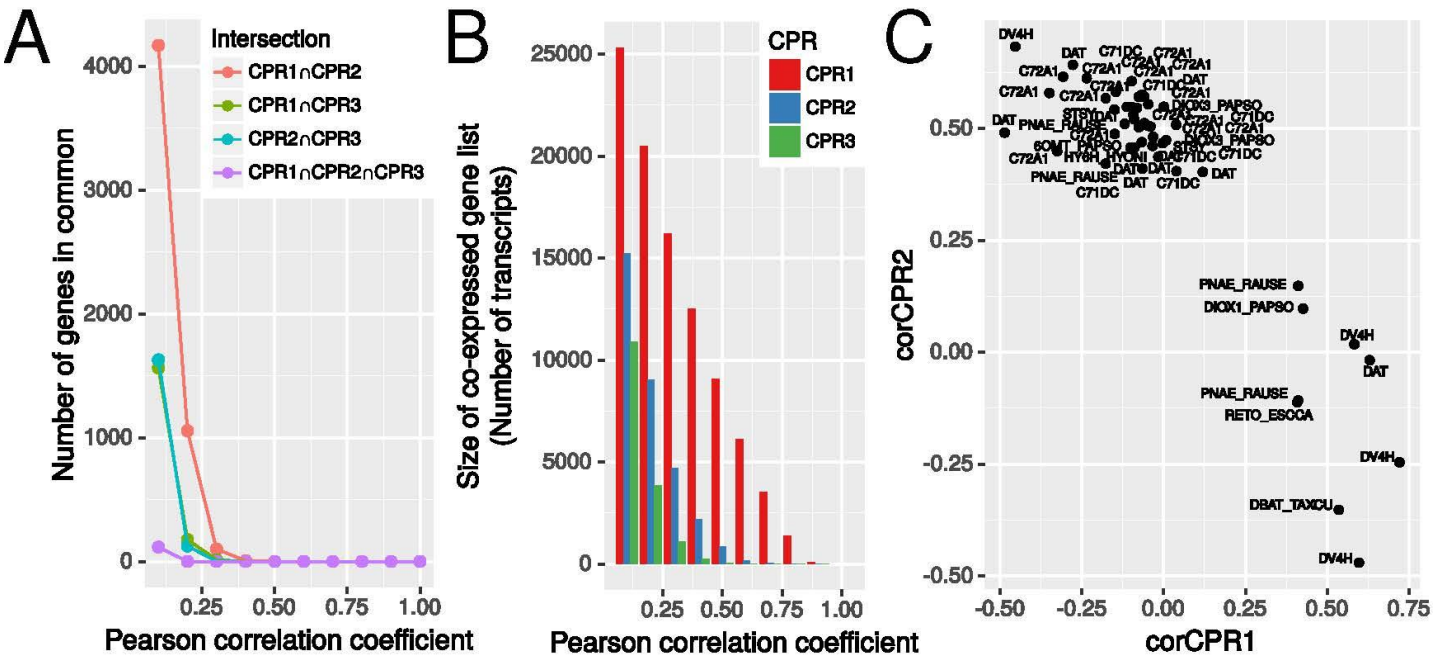


Figure 5. Analysis of CPR1, CPR2 or DFR gene co-expression correlation. Sizes of intersections between co-expressed gene lists (A). Lists of co-expressed genes were obtained after calculating PCC of *C. roseus* CPR expression levels with each other transcript found in

CDF97v2 assembly and setting different PCC threshold values. Sizes of co-expressed gene lists for each CPR at different PCC thresholds (B). Functional composition of co-expressed genes with PCC > 0.4 with each CPR (C). Pearson correlation coefficients of transcripts related to alkaloid metabolism (according to Uniprot annotation) with CPR1 and CPR2. When annotation did not correspond to a deposited protein from Catharanthus, the corresponding protein name is followed by the name of initial plant (RAUSE, *Rauwolfia serpentina*; PAPSO, *Papaver somniferum*; ESCCA, *Eschscholtzia californica*; TAXCU, *Taxus cuspidata*). See **Supplemental Table** 3B for full gene information.

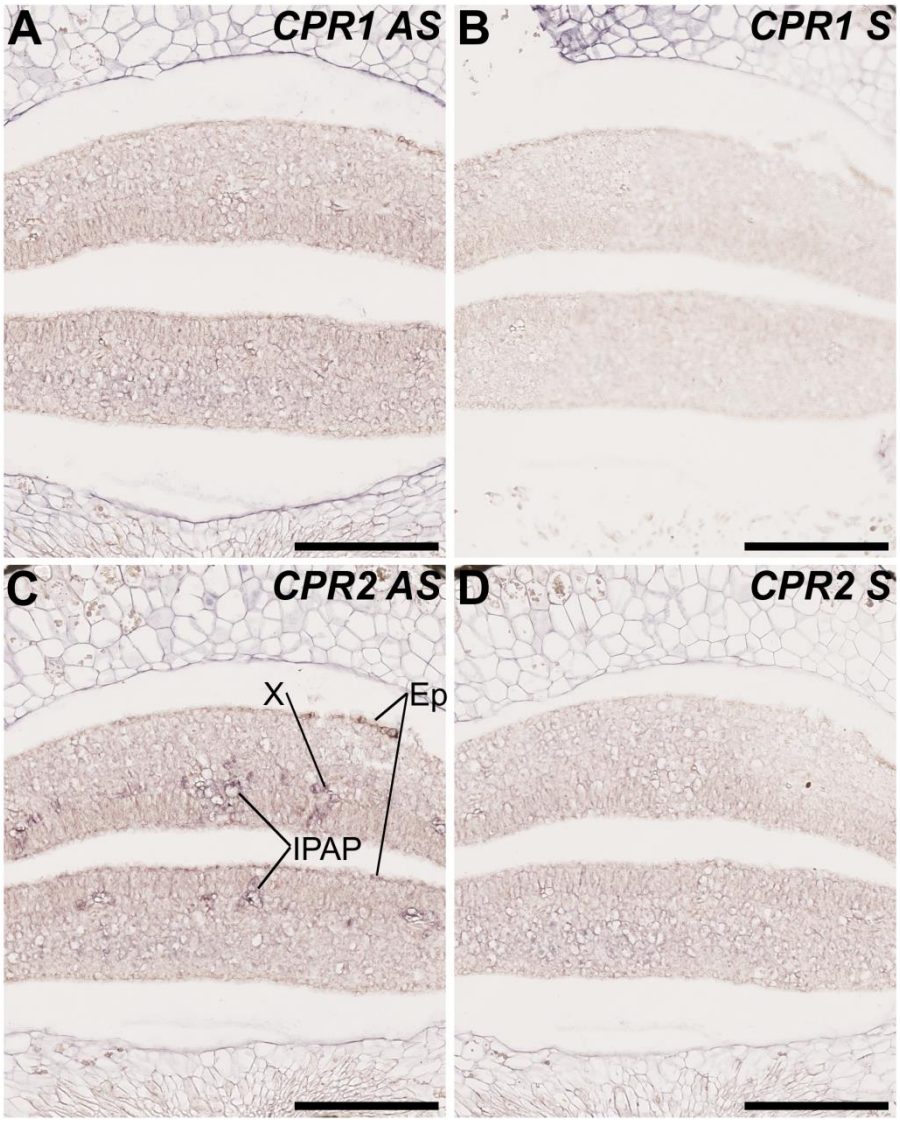


Figure 6. Localization of *CPR1* and *CPR2* transcripts in cotyledons of *C. roseus* germinating seedlings. The analysis of *CPR1* and *CPR2* transcript distribution was

performed by *in situ* RNA hybridization. Serial sections of germinating seedlings were hybridized either with *CPR1* antisense (AS) probes (A), *CPR2* antisense probes (C), *CPR1* sense (S) probes (B) or *CPR2* sense probes (D) used as negative controls. Ep, epidermis; IPAP, Internal Phloem Associated Parenchyma; X, Xylem. Bars = 100 mm.

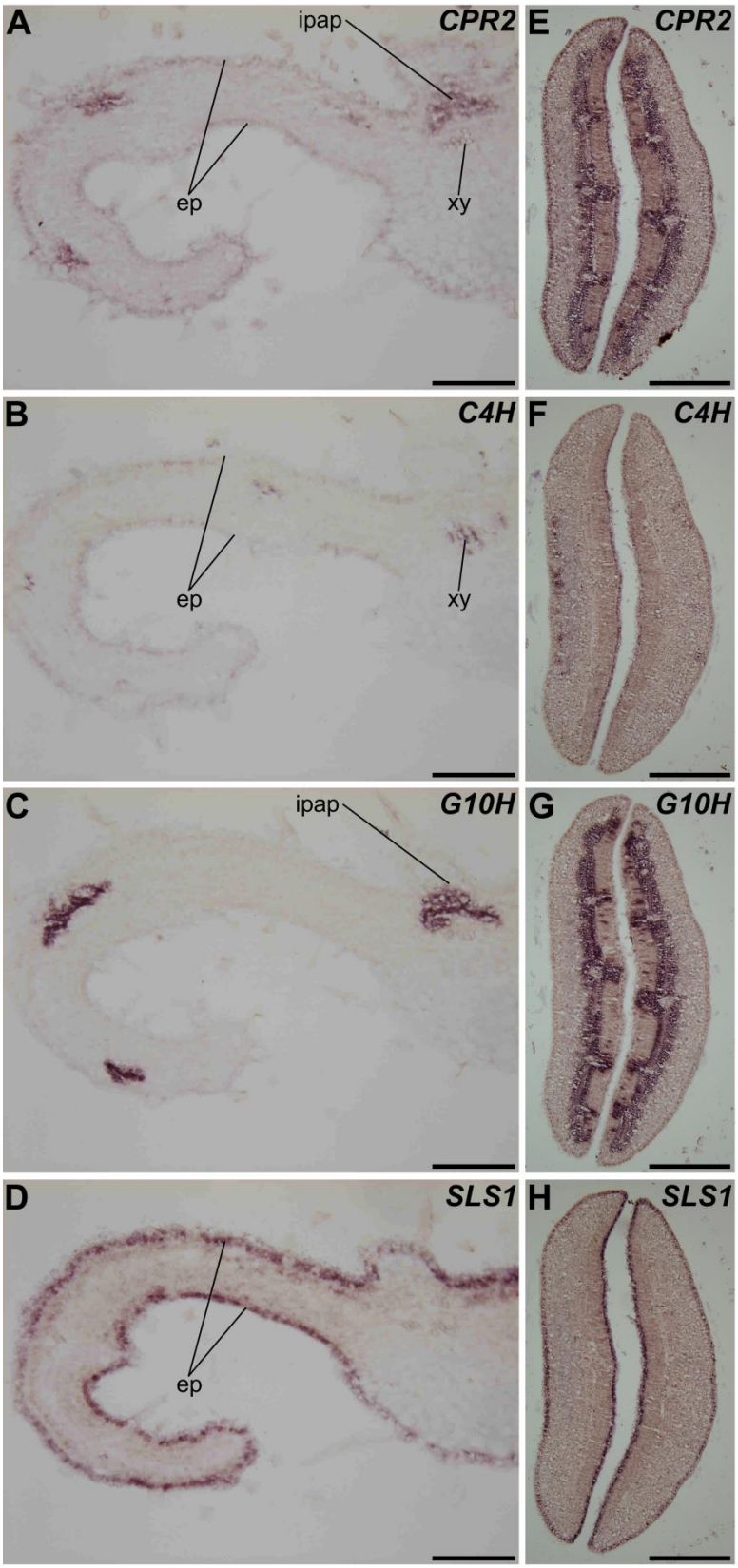


Figure 7. *CPR2* is expressed in leaf and cotyledon tissues hosting transcripts of P450s involved in MIA and phenylpropanoid metabolisms. *CPR2*, *C4H*, *G8H* and *SLS1* transcript localizations were carried out by RNA *in situ* hybridization performed on young leaves (A-D) and cotyledons (E-H). Serial sections were hybridized either with *CPR2* antisense probes (A, E), C4H antisense probes (B, F), *G8H* antisense probes (C, G) or *SL S1* antisense probes (D, H). Ep, epidermis; IPAP, Internal Phloem Associated Parenchyma; X, Xylem. Bars = 100 mm

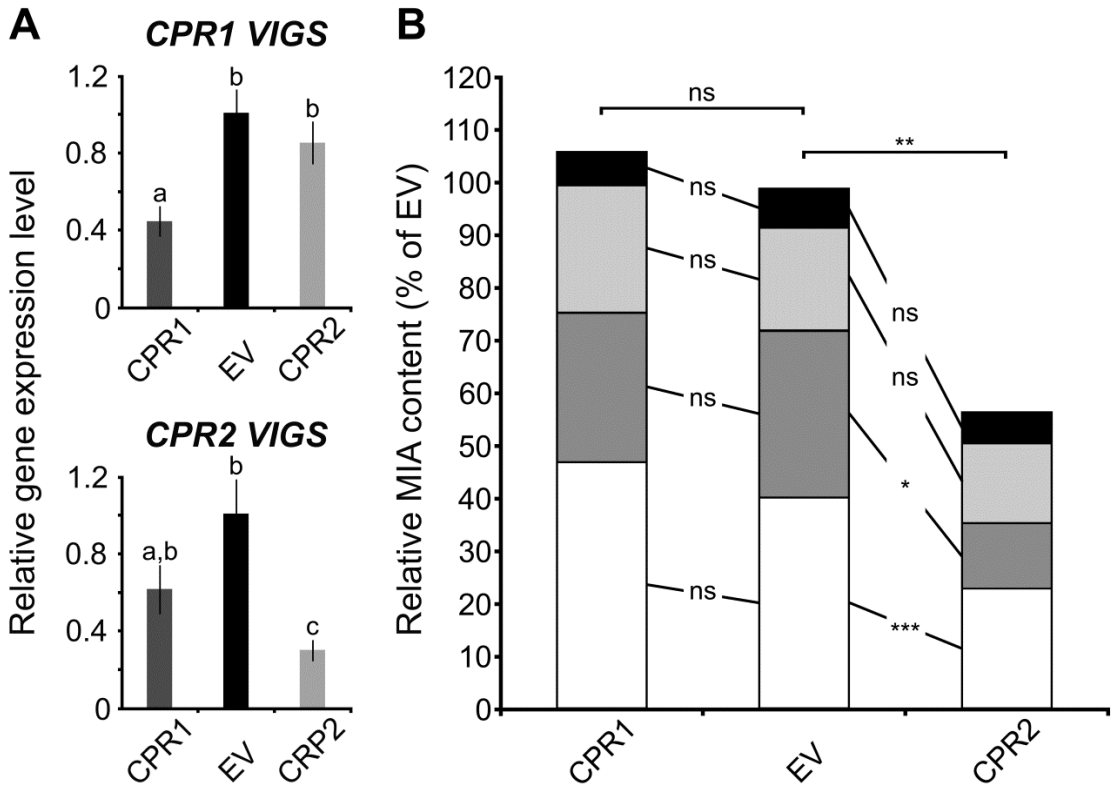


Figure 8. Silencing of CPR1 does not impact MIA biosynthesis. A, Down-regulation of CPR1 and CPR2 transcript by VIGS. The relative expression of each gene was determined by real-time RT-PCR analyses performed on total RNA extracted from *C. roseus* leaves of CPR1 or CPR2-silenced plants (dark gray and light gray bars, respectively) or plants transformed with an empty vector control (EV; dark bars). CrRPS9 was used as a reference gene. Data were normalized to CrRPS9 expression and correspond to average values (n = 8) ± SD of independent transformed plants. Letters indicate statistical classes (Wilcoxon rank sum test, FDR-adjusted *p*-value<0.05). B, Relative MIA content of CPR1- and CPR2-VIGS plants expressed relatively to that of EV plants. The relative MIA content was determined by quantification of catharanthine (white), vindorosine (drak grey), vindoline (light grey) and serpentine (dark)performed by LC-MS. The amount of each MIA in silenced plants (8 plants

per gene) was expressed relatively to that measured in EV plants (8 plants; normalized to 100%). Asterisks denote statistical significance (*P,0.005, **P,0.0005, ***P,0.00005).

Partie 3: Les déshydrogénases/réductases

Article 4: Unlocking the Diversity of Alkaloids in *Catharanthus roseus*: Nuclear Localization Suggests Metabolic Channeling in Secondary Metabolism

Article 5: The structural basis of ajmalicine biosynthesis: Active site elements to control stereoselectivity in Corynanthe alkaloid biosynthesis

Au cours de ces cinq dernières années, les étapes manquantes de la voie de biosynthèse des AIM conduisant du GPP à la strictosidine ont toutes été identifiées, les unes après les autres (l'ensemble de ces travaux est relaté dans les articles de synthèse: Dugé de Bernonville et *al.*, 2015, Courdavault et *al.*, 2014). L'enzyme suivante, la SGD conduit à l'aglycone de strictosidine, composé instable subissant des réarrangements spontanés. Ce composé est décrit comme étant la plaque tournante pour la formation des divers types d'alcaloïdes: *corynanthe/hétéroyohimbine*, *iboga* (catharanthine) et *aspidosperma* (vindoline).

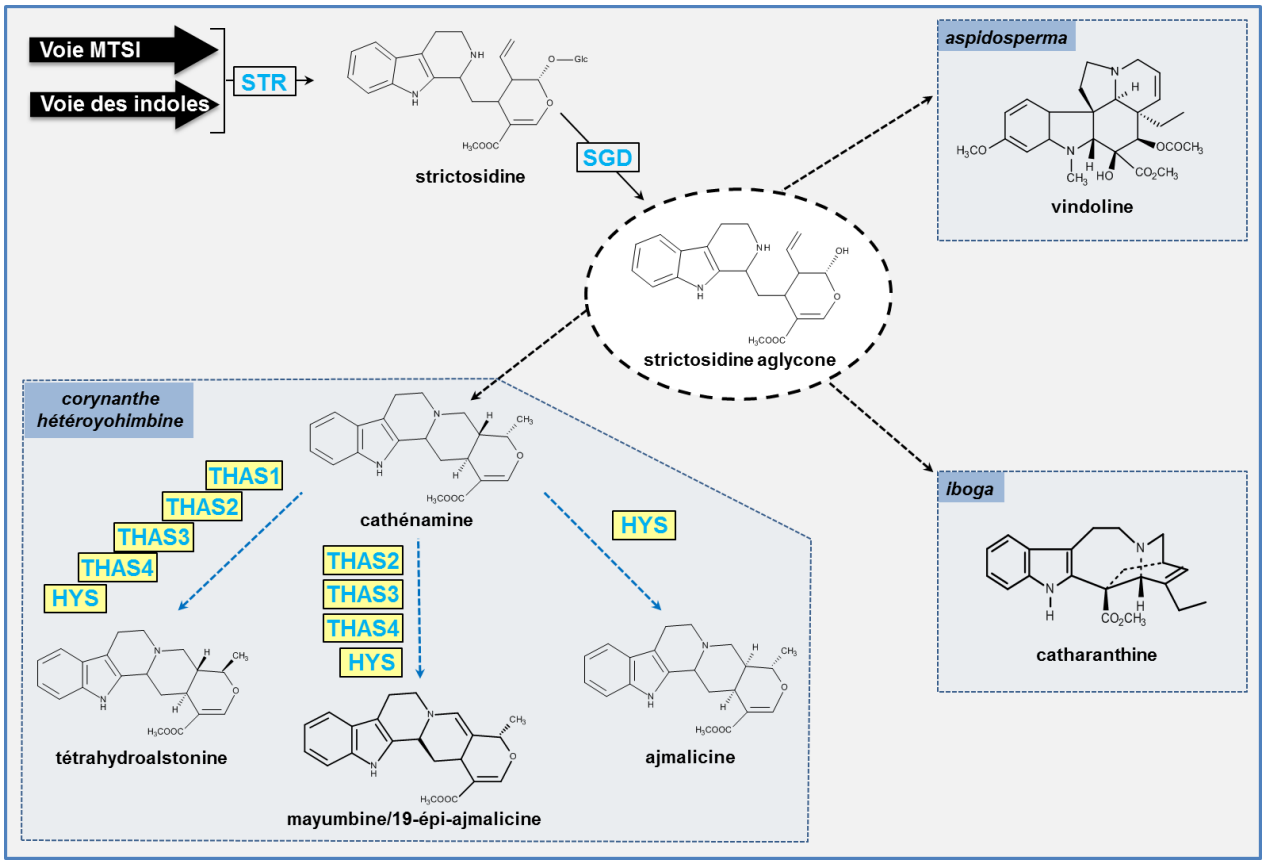


Figure 21 : Schéma simplifié de la voie de biosynthèse des AIM à partir de la strictosidine aglycone chez *C. roseus*. Les enzymes THAS1 à THAS4 et HYS sont celles nouvellement identifiées dans nos travaux. MTSI : voie des monoterpènes sécoiridoïdes.

Les alcaloïdes de type hétéroyohimbine dérivent de la cathénamine et de l'épi-cathénamine, tous deux issus de réarrangements de l'aglycone de strictosidine. Les deux articles suivants portent sur l'identification de nouvelles enzymes impliquées dans les alcaloïdes de type hétéroyohimbine. Il s'agit de réductases dépendantes du NADPH et annotées, à l'origine, dans les bases de données, comme alcool déshydrogénases (ADH) du fait de leur similarité de séquences avec ces familles d'enzymes.

Classiquement, les alcools déshydrogénases (ADH), catalysent des réactions de conversion d'alcools en aldéhydes, en utilisant les NAD(P)H comme cofacteurs réactionnels (Jörnvall et *al*., 2010 ; Strommer, 2011). Ubiquitaires, ces enzymes participent à la conversion des alcools toxiques en composés moins toxiques (aldéhydes, cétones) chez l'homme, tandis que chez les plantes, les levures, et de nombreuses bactéries, les ADH fonctionnent de façon réversible en produisant de l'alcool à partir d'aldéhydes en générant du NAD^+ ou *vice versa*. Sur la base de leur structure protéique, on distingue plusieurs familles d'ADH qui appartiennent à des superfamilles différentes. Ainsi, certaines familles d'ADH font partie de la superfamille des déshydrogénases/réductases à chaine courte (SDR : short-chain deshydrogenase/reductase), d'autres de la superfamille des déshydrogénases/réductases à chaine moyenne (MDR : medium-chain deshydrogenase/reductase), renfermant ou non un atome de Zinc (Jörnvall et al., 2015).

Parmi les superfamilles des SDR et MDR, majoritairement constituées de familles d'ADH, on trouve également des enzymes catalysant d'autres types de réaction, y compris des réductions de doubles liaisons C=C et C=N (Kavanagh et *al*., 2008 ; Kurata et *al*., 2005).

A la suite d'un travail collaboratif, les nouvelles enzymes caractérisées dans les articles suivant appartiennent à la superfamille des MDR et réduisent les formes iminium de la cathénamine et de l'épi-cathénamine.

Le premier article porte sur la caractérisation, pour la première fois, d'une enzyme opérant en aval de l'étape catalysée par la SGD. Il s'agit de la tétrahydroastonine synthase (THAS) générant la tétrahydroalstonine et nommée THAS1. Le second article s'inscrit dans la continuité du premier et décrit la caractérisation de quatre nouvelles enzymes réduisant l'aglycone de strictosidine et formant des alcaloïdes de type hétéroyohimbine. Parmi ces enzymes, trois sont de nouvelles THAS (annotées THAS2, THAS3 et THAS4) produisant

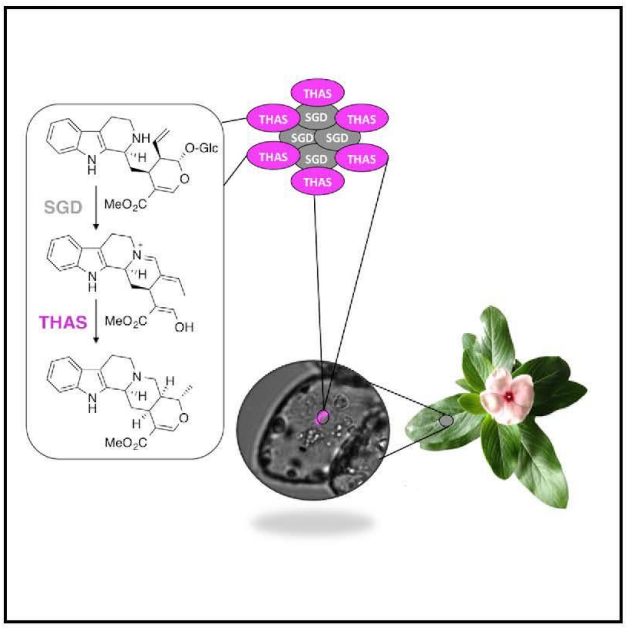
majoritairement de la tétrahydroalstonine. La quatrième enzyme se comporte différemment. Elle a été nommée hétéroyohimbine synthase (HYS) car elle produit un mélange d'alcaloïdes de type hétéroyohimbine : ajmalicine, tétrahydroalstonine et mayumbine. Les enzymes THAS1, THAS2 et HYS ont été cristallisées et l'étude de leur structure tridimentionnelle a permis d'élucider leur mécanisme réactionnel et leur stéréosélectivité. Par ailleurs, les études utilisant la technique BiFC ont montré que les trois protéines pouvaient interagir avec SGD et former des aggrégats dans le noyau, soulevant la question du recrutement par SGD de diverses enzymes opérant après la formation de l'aglycone de strictosidine.

Brief Communication

Chemistry & Biology

Unlocking the Diversity of Alkaloids in *Catharanthus roseus*: Nuclear Localization Suggests Metabolic Channeling in Secondary Metabolism

Graphical Abstract



Authors

Anna Stavrinides,
Evangelos C. Tatsis, ...,
Vincent Courdavault,
Sarah E. O'Connor

Correspondence

sarah.oconnor@jic.ac.uk (S.E.O.),
vincent.courdavault@univ-tours.fr (V.C.)

In Brief

How plants transform the central biosynthetic intermediate strictosidine into thousands of divergent alkaloids has remained unresolved. Stavrinides et al. discover a nuclear-localized alcohol dehydrogenase homolog responsible for conversion of strictosidine aglycone to tetrahydroalstonine that appears to interact with an upstream pathway enzyme.

Highlights

- Tetrahydroalstonine synthase catalyzes the formation of a plant-derived alkaloid
- Tetrahydroalstonine synthase is localized to the nucleus
- Tetrahydroalstonine synthase and the preceding pathway enzyme interact
- Discovery of a gene controlling structural diversity of monoterpene indole alkaloids

Stavrinides et al., 2015, Chemistry & Biology *22*, 1–6
March 19, 2015 ©2015 The Authors
http://dx.doi.org/10.1016/j.chembiol.2015.02.006

Please cite this article in press as: Stavrinides et al., Unlocking the Diversity of Alkaloids in *Catharanthus roseus*: Nuclear Localization Suggests Metabolic Channeling in Secondary Metabolism, Chemistry & Biology (2015), http://dx.doi.org/10.1016/j.chembiol.2015.02.006

Chemistry & Biology

Brief Communication

Unlocking the Diversity of Alkaloids in *Catharanthus roseus*: Nuclear Localization Suggests Metabolic Channeling in Secondary Metabolism

Anna Stavrinides,[1] Evangelos C. Tatsis,[1] Emilien Foureau,[2] Lorenzo Caputi,[1] Franziska Kellner,[1] Vincent Courdavault,[2,*] and Sarah E. O'Connor[1,*]

[1]Department of Biological Chemistry, The John Innes Centre, Colney, Norwich NR4 7UH, UK

[2]Université François Rabelais de Tours, EA2106 "Biomolécules et Biotechnologies Végétales", 37200 Tours, France

*Correspondence: sarah.oconnor@jic.ac.uk (S.E.O.), vincent.courdavault@univ-tours.fr (V.C.)

http://dx.doi.org/10.1016/j.chembiol.2015.02.006

SUMMARY

The extraordinary chemical diversity of the plant-derived monoterpene indole alkaloids, which include vinblastine, quinine, and strychnine, originates from a single biosynthetic intermediate, strictosidine aglycone. Here we report for the first time the cloning of a biosynthetic gene and characterization of the corresponding enzyme that acts at this crucial branchpoint. This enzyme, an alcohol dehydrogenase homolog, converts strictosidine aglycone to the heteroyohimbine-type alkaloid tetrahydroalstonine. We also demonstrate how this enzyme, which uses a highly reactive substrate, may interact with the upstream enzyme of the pathway.

INTRODUCTION

The monoterpene indole alkaloids (MIAs) are a highly diverse family of natural products that are produced in a wide variety of medicinal plants. Over 3000 members of this natural product class, which includes compounds such as quinine, vinblastine, reserpine, and yohimbine, are derived from a common biosynthetic intermediate, strictosidine aglycone (O'Connor and Maresh, 2006). How plants transform strictosidine aglycone into divergent structural classes has remained unresolved.

The recent availability of transcriptome and genome data has dramatically accelerated the rate at which new plant biosynthetic genes are discovered. All genes that lead to strictosidine aglycone have been recently cloned from the well-characterized medicinal plant *Catharanthus roseus*, which produces over 100 MIAs (De Luca et al., 2014). However, gene products that act on strictosidine aglycone have not been identified in any plant, despite decades of effort. Attempts have been hampered in part by the reactivity and instability of strictosidine aglycone. In *C. roseus*, there are at least two major pathway branches derived from strictosidine aglycone (O'Connor and Maresh, 2006). One pathway is hypothesized to lead to the aspidosperma and the iboga classes to yield the precursors of vinblastine, while the other is expected to lead to alkaloids of the heteroyohimbine type (Figure 1A). These alkaloids have diverse biological activities: vinblastine is used as an anticancer agent (Kaur et al., 2014) and the heteroyohimbines have a range of pharmacological uses (Costa-Campos et al., 1998; Elisabetsky and Costa-Campos, 2006). While it is unknown how many *C. roseus* enzymes use strictosidine aglycone as a substrate, there is clearly more than one enzyme that acts at this crucial branchpoint.

The biochemical pathway leading from strictosidine aglycone to the heteroyohimbine alkaloids has been previously investigated using both crude plant extracts and biomimetic chemistry. Reduction of strictosidine aglycone with $NaBH_4$ or $NaCNBH_3$ yielded the heteroyohimbines ajmalicine (raubasine), tetrahydroalstonine, and 19-epi-ajmalicine, which differ only in the stereochemical configuration at carbons 15, 19, and 20, in various ratios (Figure 1B) (Brown et al., 1977; Kan-Fan and Husson, 1978, 1979, 1980). These three diastereomers were again observed, also in varying relative amounts, when crude *C. roseus* protein extracts were incubated with strictosidine aglycone and NADPH, but not in the absence of NADPH (Rueffer et al., 1979; Stoeckigt et al., 1976, 1977, 1983; Zenk, 1980). Collectively, these observations indicate that the heteroyohimbines result directly from the reduction of strictosidine aglycone and that an NADPH-dependent enzyme is implicated in this process. However, no gene encoding such an enzyme has been identified. Here we report the discovery of a reductase that converts strictosidine aglycone to the heteroyohimbine alkaloid tetrahydroalstonine.

RESULTS AND DISCUSSION

Given that heteroyohimbine biosynthesis likely requires reduction of an iminium present in strictosidine aglycone (Figure 1B), we used a publically available RNA-seq database that we recently generated (Gongora-Castillo et al., 2012) to search for *C. roseus* candidates displaying homology to enzyme classes known to reduce the carbonyl functional group. The alcohol dehydrogenases (ADHs), enzymes that reduce aldehydes and ketones to alcohols, were chosen as the initial focus. As part of a screen of ADHs that are upregulated in response to methyl jasmonate (Gongora-Castillo et al., 2012), a hormone known to upregulate alkaloid biosynthesis, we identified a candidate annotated as sinapyl alcohol dehydrogenase (Supplemental Information). When heterologously expressed and purified from *E. coli* (Figure S1), and assayed with strictosidine aglycone and NADPH, this candidate yielded a product with a mass consistent with a heteroyohimbine (*m/z* 353.1855), thereby implicating this

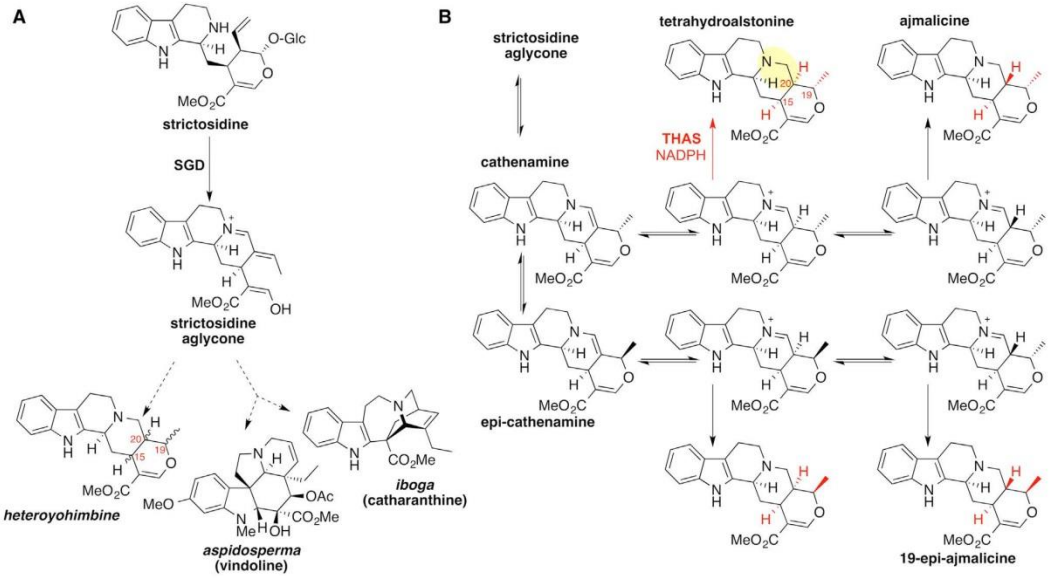


Figure 1. The Monoterpene Indole Alkaloids

(A) Representative monoterpene indole alkaloids derived from strictosidine and strictosidine aglycone found in *Catharanthus roseus*.

(B) Heteroyohimbine biosynthesis.

enzyme in the important structural branchpoint of the MIA biosynthetic pathway (Figure 2A).

To determine the identity of the alkaloid product, the enzyme was incubated with purified strictosidine (4.3 mg) in the presence of strictosidine glucosidase (SGD), which generated strictosidine aglycone in situ to best mimic physiologically relevant conditions. The major product (approximately 1 mg) was isolated by preparative thin-layer chromatography and exhibited an ^{1}H-NMR and ^{13}C-NMR spectrum matching an authentic standard of tetrahydroalstonine (Figure 2B; Figure S2). Hemscheidt and Zenk (1985) previously reported the isolation of an enzyme that produced tetrahydroalstonine, although this protein was purified only 35-fold from *C. roseus* cell cultures. Consistent with Hemscheidt and Zenk's (1985) nomenclature, we named this enzyme tetrahydroalstonine synthase (THAS). A minor enzymatic product was produced in yields too low for NMR characterization, but had a mass and R_f value consistent with ajmalicine, a stereoisomer of tetrahydroalstonine (Figure S2). When applied to normal phase liquid chromatography conditions, ajmalicine and tetrahydroalstonine could be resolved, indicating that the enzyme produces approximately 95% tetrahydroalstonine (Figure 3; Supplemental Information). We also silenced this gene in *C. roseus* seedlings using virus-induced gene silencing (VIGS) (Liscombe and O'Connor, 2011). LC-mass spectrometry (MS) analysis of the silenced leaf tissue showed a statistically significant decrease (approximately 50%) of a peak with a mass and retention time consistent with a heteroyohimbine, suggesting that this enzyme is involved in this biosynthetic pathway branch in vivo (Figure S2). A 50% reduction in product levels upon silencing has been observed for other physiologically relevant biosynthetic genes using the VIGS approach in both *C. roseus* (Asada et al., 2013; Geu-Flores et al., 2012) and another well-studied medicinal plant, opium poppy (Desgagne-Penix and Facchini, 2012; Chen and Facchini, 2014). Therefore, THAS is likely a major producer of tetrahydroalstonine in vivo, although additional, undiscovered *C. roseus* enzymes could also contribute to production of this compound. While we could not resolve tetrahydroalstonine and its stereoisomer ajmalicine in the silenced crude extracts, the levels of the ajmalicine-derived alkaloid serpentine remain the same, suggesting that silencing of THAS does not substantially affect ajmalicine levels and consequently that THAS does not play a major role in the biosynthesis of ajmalicine in planta.

Small-scale assays using LC-MS to monitor product formation indicated that NADPH was required for the reaction, although NADH could also be utilized (Figure S1). Efforts to accurately measure the steady state kinetic constants of this enzyme were complicated because strictosidine aglycone reacts with nucleophiles, opening the possibility that the substrate reacts with components in the reaction or the enzyme. This reactivity has already been associated with a plant defense mechanism involving strictosidine aglycone-mediated aggregation of proteins in *C. roseus* (Guirimand et al., 2010). Nevertheless, we obtained estimated K_m and k_{cat} values (Figure S1). To support these kinetic data, we also performed isothermal titration calorimetry (ITC) with THAS in the presence of NADPH and strictosidine aglycone. Titration of THAS with NADPH indicated that the co-substrate binds first with a K_d of 1.5 ± 0.1 μM (ΔH (cal/mol)

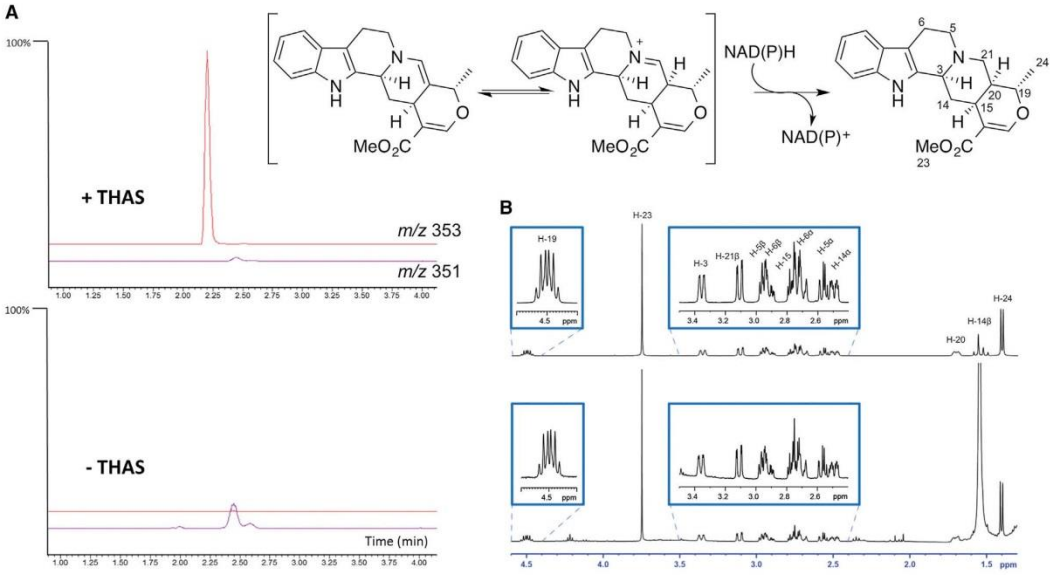


Figure 2. Activity Assays of THAS

Enzyme reactions were performed at 25°C for 30 min and assayed using a mass spectrometer in tandem with ultraperformance liquid chromatography.

(A) The total ion chromatogram for *m/z* 353 (red trace) and *m/z* 351 (purple trace) from 1 to 4 min is shown. Top trace: THAS (50 nM), SGD (6 nM), strictosidine (200 μM), NADPH (200 μM); bottom trace: same reaction in the absence of THAS. The *y* axis represents normalized ion abundance as a percentage relative to $1.00e^8$ detected by selected ion monitoring at *m/z* 353 and 351.

(B) Portion of the ^{1}H-NMR spectrum of the isolated enzymatic product compared with an authentic standard of tetrahydroalstonine.

2310 ± 123.2; Δ*S* (cal/mol/deg) 34.2 ± 0.3) (Figure S1). The aglycone substrate does not appear to bind in the absence of NADPH, suggesting that the enzyme utilizes an ordered binding mechanism in which NADPH binds first. However, titration of the THAS-NADPH complex with strictosidine aglycone led to formation of a precipitate when concentrations of strictosidine aglycone exceeded 60 μM, preventing calculation of an accurate K_d. Collectively, the ITC data for THAS are consistent with an ordered Bi-Bi mechanism, a kinetic mechanism that has been reported for similar ADHs such as cinnamyl alcohol dehydrogenase (Charlier and Plapp, 2000; Lee et al., 2013).

The amino acid sequence of THAS was subjected to a BLAST alignment against the *C. roseus* transcriptome (Gongora-Castillo et al., 2012), as well as the NCBI (Figure S3). The closest characterized homologs of THAS are sinapyl alcohol dehydrogenase (*Populus tremuloides*, 64% amino acid identity), cinnamyl alcohol dehydrogenase (*Populus tomentosa*, 64%) and 8-hydroxygeraniol dehydrogenase (*C. roseus*, 63%), which are zinc-containing medium chain ADHs (Bomati and Noel, 2005; Lee et al., 2013).

Strictosidine aglycone can rearrange into several isomers (Figure 1B), and while it has been reported that the dominant isomer is cathenamine (Gerasimenko et al., 2002; Stoeckigt et al., 1977), equilibration in solution with other isomers occurs (Brown and Leonard, 1979; Stoeckigt et al., 1983). Reduction of cathenamine or epi-cathenamine (Figure 1B) by a reductase would require reduction of the carbon-carbon double bond of an enamine; alternatively, Stoeckigt et al. (1983) and Zenk (1980) suggested that the iminium isomer is reduced (Figure 1B). THAS may catalyze the stereoselective formation of tetrahydroalstonine by selectively binding the correct isomer of the substrate for reduction, thereby relying on the inherent propensity for the enamine and imine to tautomerize under physiological conditions. Given that three diastereomers, ajmalicine, tetrahydroalstonine, and 19-epi-ajmalicine, can be obtained from chemical reduction of strictosidine aglycone, this is a chemically reasonable proposal. An alternative hypothesis is that THAS catalyzes enamine-imine tautomerization in addition to reduction. The difficulties associated with obtaining accurate kinetic data in this system, as well as the inherent reactivity of the strictosidine aglycone, make answering these questions using enzymology approaches challenging. However, identification and comparison with enzymes that generate other heteroyohimbine diastereomers will likely provide the basis for a more definitive mechanism of product specificity.

Recent research has highlighted that plant secondary metabolite biosynthetic pathways often are compartmentalized in different subcellular locations. While microscopy experiments have demonstrated that most of the early steps of monoterpene indole alkaloid biosynthesis in *C. roseus* take place in the cytosol (Courdavault et al., 2014), the enzyme that synthesizes strictosidine is located in the vacuole, and the enzyme SGD, which deglycosylates strictosidine, contains a nuclear localization signal and is in the nucleus, a highly unusual site for secondary

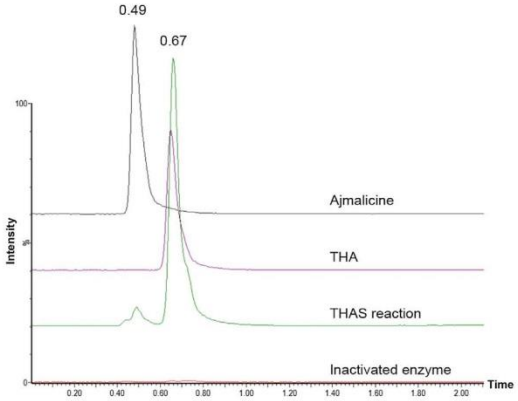


Figure 3. LC-MS Performed under Normal Phase Conditions (Hydrophilic Interaction Liquid Chromatography) Showing Separation of Ajamlicine (Retention Time of 0.49 min) and Tetrahydroalstonine (THA, Retention Time 0.67 min)

THAS produces approximately 95% of the tetrahydroalstonine (THA) diastereomer. The *y* axis represents normalized ion abundance as a percentage detected by selected ion monitoring at *m/z* 353.

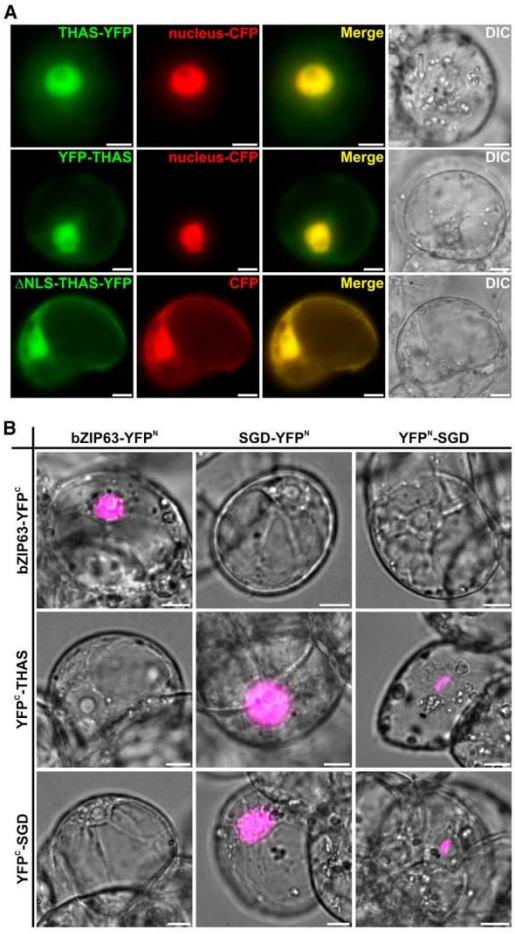


Figure 4. THAS Is Targeted to the Nucleus via a Monopartite Nuclear Localization Signal (NLS) and Interacts with SGD

(A) *C. roseus* cells were transiently cotransformed with plasmids expressing either THAS-YFP (upper row), YFP-THAS (middle row), or the NLS deleted version of THAS (lower row) and plasmids encoding the nuclear CFP marker or the nucleocytosolic CFP marker (second column). Colocalization of the fluorescence signals appears in yellow when merging the two individual (green/red) false color images (third column). Cell morphology is observed with differential interference contrast (DIC) (fourth column).

(B) THAS and SGD interactions were analyzed by BiFC in *C. roseus* cells transiently transformed by plasmids encoding fusions indicated on the top (fusion with the split YFP^N fragment) and on the left (fusion with split YFP^C fragment). bZIP63 was used as a positive BiFC control and to evaluate the specificity of THAS and SGD interactions. The images are merges of the YFP BiFC channel (magenta false color) with the DIC channel to show the nuclear localization of the interactions. Bars, 10 μm.

metabolite biosynthesis (Guirimand et al., 2010). Notably, a motif resembling a class V nuclear localization sequence (Kosugi et al., 2008) was observed in THAS ($K_{214}K_{215}K_{216}R_{217}$). Microscopy of *C. roseus* cells transformed with YFP-tagged THAS confirmed the nuclear location of this enzyme, while deletion of the KKKR sequence disrupted the localization (Figure 4A; Figure S4). This is one of the very few examples of secondary metabolism that is localized to the nucleus (Saslowsky et al., 2005).

Given the reactivity of strictosidine aglycone (Guirimand et al., 2010), metabolic channeling via a protein-protein interaction between SGD and the enzyme immediately downstream may be necessary to protect the substrate. Pull down experiments between SGD and THAS gave partially positive but inconclusive results (Figure S4). However, when we used bimolecular fluorescence complementation (BiFC) in *C. roseus* cells, we observed an interaction between SGD and THAS (Figure 4B). While this interaction generated a diffuse nuclear fluorescent signal when the C-terminal end of SGD was fused to the split-YFP fragment, a sickle-shaped signal was observed when both SGD and THAS were expressed with free C-terminal ends (YFP^N-SGD and YFP^C-THAS). Such a signal was also observed for SGD self-interactions (Guirimand et al., 2010) and likely results from the formation of SGD complexes over 1.5 MDa (Luijendijk et al., 1998). Similar experiments with SGD and an upstream MIA biosynthetic enzyme, loganic acid methyl transferase, failed to show an interaction, highlighting the specificity of this interaction (Figure S4). The fact that THAS interacts with SGD provides further support for the physiological relevance of THAS in planta. As strictosidine aglycone is reactive and most likely toxic in vivo, it has been proposed that this molecule is produced by the plant in response to attack (Guirimand et al., 2010). The nuclear localization of THAS might be an evolutionary mechanism designed to channel this molecule into a more stable product when no such defense is required. Identification of additional nuclear-localized biosynthetic enzymes in *C. roseus* and other heteroyohimbine

producing plants may provide more insight into the reasons for this unusual localization pattern.

SIGNIFICANCE

Many of the monoterpene indole alkaloid structural classes are generated at the SGD junction. Here we report the first identification of a biosynthetic gene that acts directly downstream of SGD. The enzyme, an ADH homolog, generates a heteroyohimbine alkaloid by reducing one of the isomers of strictosidine aglycone. Unusually, this enzyme is located in the nucleus and may interact with its upstream partner, SGD. The discovery of the THAS gene represents the completion of a major branch of monoterpene indole alkaloid biosynthesis, which will now allow reconstruction of heteroyohimbines and heteroyohimbine analogs in heterologous hosts. This discovery is a crucial first step in understanding how the structural diversity of MIAs is controlled.

EXPERIMENTAL PROCEDURES

The THAS gene (accession number KM524258) was cloned into pOPINF and expressed in Rosetta 2 pLysS *E. coli* cells (Novagen) with induction of expression with 0.1 mM isopropyl β-D-1-thiogalactopyranoside. Cultures were grown at 18°C for 16 hr, with shaking at 200 rpm. His-tagged THAS was purified using a HisTrap FF 5-ml column (GE Healthcare). SGD expression and purification was done as described for THAS using the expression system described previously by Yerkes et al. (2008). Purified THAS and SGD were used in all assays. Strictosidine was enzymatically synthesized from tryptamine and a crude methanol extract of snowberries (*Symphoricarpos albus*) enriched in secologanin prepared as previously described (Geerlings et al., 2001). Strictosidine aglycone was generated in situ prior to addition of THAS by incubation of strictosidine and SGD in the appropriate solution for 10 min, at which time strictosidine was completely converted to the aglycone.

Steady state kinetic analyses were performed with 50 nM THAS and 6 nM SGD, 50 mM phosphate buffer (pH 7.5), 200 μM NADPH, and an internal caffeine standard (50 μM). All LC-MS measurements were performed on AQUITY ultra-performance liquid chromatography with a Xevo TQ-S mass spectrometer.

For VIGS, a 330-bp fragment of THAS was cloned into the pTRV2u vector as described (Geu-Flores et al., 2012). The resulting pTRV2u-THAS construct was used to silence THAS in *C. roseus* seedlings essentially as described (Liscombe and O'Connor, 2011).

The subcellular localization of THAS was studied by creating fluorescent fusion proteins using the pSCA-cassette YFPi plasmid (Guirimand et al., 2009, 2010). The capacity of interaction of THAS and SGD was characterized by BiFC assays using THAS PCR product cloned via *Spe*I into the pSPYCE(MR) plasmid (Waadt et al., 2008), which allows expression of THAS fused to the carboxy-terminal extremity of the split YFP^C fragment (YFP^C-THAS). The pSCA-SPYNE173-SGD and pSPYNE(R)173-SGD plasmids (Guirimand et al., 2010) were used to express SGD fused to the amino-terminal or carboxy-terminal extremity of the split YFP^N fragment (SGD-YFP^N and YFP^N-SGD, respectively). THAS self-interactions were analyzed via additional cloning of the THAS PCR product into the pSCA-SPYNE173 and pSCA-SPYCE(M) plasmids (Guirimand et al., 2010) to express THAS-YFP^N and THAS-YFP^C, respectively. Transient transformation of *C. roseus* cells by particle bombardment and fluorescence imaging were performed following the procedures previously described (Guirimand et al., 2009, 2010).

Complete experimental details are included in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and four figures and can be found with this article online at http://dx.doi.org/10.1016/j.chembiol.2015.02.006.

AUTHOR CONTRIBUTIONS

A.S. made the initial discovery of THAS activity and conducted all enzyme assays, kinetics, pulldown, and ITC; E.C.T. performed VIGS and assisted in the structural characterization of the enzyme product; E.F. performed the microscopy experiments; L.C. assisted in the purification of THAS and pulldown; F.K. provided initial genomic data that assisted in identification of the THAS candidate; V.C. conceived, initiated, and supervised all localization and BiFC experiments; S.E.O. supervised all enzymology experiments; A.S., V.C., S.E.O. wrote the manuscript.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the ERC (311363) and from the Région Centre (France, ABISAL grant). A.S. is supported by a BBSRC DTP studentship.

Received: November 7, 2014
Revised: January 24, 2015
Accepted: February 17, 2015
Published: March 12, 2015

REFERENCES

Asada, K., Salim, V., Masada-Atsumi, S., Edmunds, E., Nagatoshi, M., Terasaka, K., Mizukami, H., and De Luca, V. (2013). A 7-deoxyloganetic acid glucosyltransferase contributes a key step in secologanin biosynthesis in Madagascar periwinkle. Plant Cell *25*, 4123–4134.

Bomati, E.K., and Noel, J.P. (2005). Structural and kinetic basis for substrate selectivity in *Populus tremuloides* sinapyl alcohol dehydrogenase. Plant Cell *17*, 1598–1611.

Brown, R.T., and Leonard, J. (1979). Biomimetic synthesis of cathenamine and 19-epicathenamine, key intermediates to heteroyohimbine alkaloids. J. Chem. Soc. Chem. Commun. 877–879.

Brown, R.T., Leonard, J., and Sleigh, S.K. (1977). 'One-pot' biomimetic synthesis of 19β-heteroyohimbine alkaloids. J. Chem. Soc. Chem. Commun. 636–638.

Charlier, H.A., and Plapp, B.V. (2000). Kinetic cooperativity of human liver alcohol dehydrogenase γ2. J. Biol. Chem. *275*, 11569–11575.

Chen, X., and Facchini, P.J. (2014). Short-chain dehydrogenase/reductase catalyzing the final step of noscapine biosynthesis is localized to laticifers in opium poppy. Plant J. *77*, 173–184.

Costa-Campos, L., Lara, D.R., Nunes, D.S., and Elisabetsky, E. (1998). Antipsychotic-like profile of alstonine. Pharmacol. Biochem. Behav. *60*, 133–141.

Courdavault, V., Papon, N., Clastre, M., Giglioli-Guivarc'h, N., St-Pierre, B., and Burlat, V. (2014). A look inside an alkaloid multisite plant: the *Catharanthus* logistics. Curr. Opin. Plant Biol. *19*, 43–50.

De Luca, V., Salim, V., Thamm, A., Masada, S.A., and Yu, F. (2014). Making iridoids/secoiridoids and monoterpenoid indole alkaloids: progress on pathway elucidation. Curr. Opin. Plant Biol. *19*, 35–42.

Desgagne-Penix, I., and Facchini, P.J. (2012). Systematic silencing of benzylisoquinoline alkaloid biosynthetic genes reveals the major route to papaverine in opium poppy. Plant J. *72*, 331–344.

Elisabetsky, E., and Costa-Campos, L. (2006). The alkaloid alstonine: a review of its pharmacological properties. Evidence-based complementary and alternative medicine. Evid. Based Complement. Alternat. Med. *3*, 39–48.

Geerlings, A., Redondo, F.J., Contin, A., Memelink, J., van der Heijden, R., and Verpoorte, R. (2001). Biotransformation of tryptamine and secologanin into plant terpenoid indole alkaloids by transgenic yeast. Appl. Microbiol. Biotechnol. *56*, 420–424.

Gerasimenko, I., Sheludko, Y., Ma, X., and Stockigt, J. (2002). Heterologous expression of a *Rauvolfia* cDNA encoding strictosidine glucosidase, a biosynthetic key to over 2000 monoterpenoid indole alkaloids. Eur. J. Biochem. *269*, 2204–2213.

Geu-Flores, F., Sherden, N.H., Courdavault, V., Burlat, V., Glenn, W.S., Wu, C., Nims, E., Cui, Y., and O'Connor, S.E. (2012). An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. Nature *492*, 138–142.

Gongora-Castillo, E., Childs, K.L., Fedewa, G., Hamilton, J.P., Liscombe, D.K., Magallanes-Lundback, M., Mandadi, K.K., Nims, E., Runguphan, W., Vaillancourt, B., et al. (2012). Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. PLoS One *7*, e52506.

Guirimand, G., Burlat, V., Oudin, A., Lanoue, A., St-Pierre, B., and Courdavault, V. (2009). Optimization of the transient transformation of *Catharanthus roseus* cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. Plant Cell Rep. *28*, 1215–1234.

Guirimand, G., Courdavault, V., Lanoue, A., Mahroug, S., Guihur, A., Blanc, N., Giglioli-Guivarc'h, N., St-Pierre, B., and Burlat, V. (2010). Strictosidine activation in Apocynaceae: towards a "nuclear time bomb"? BMC Plant Biol. *10*, 182.

Hemscheidt, T., and Zenk, M.H. (1985). Partial purification and characterization of a NADPH dependent tetrahydroalstonine synthase from *Catharanthus roseus* cell suspension cultures. Plant Cell Rep. *4*, 216–219.

Kan-Fan, C., and Husson, H.P. (1978). Stereochemical control in the biomimetic conversion of heteroyohimbine alkaloid precursors. Isolation of a novel key intermediate. J. Chem. Soc. Chem. Commun. 618–619.

Kan-Fan, C., and Husson, H.P. (1979). Isolation and biomimetic conversion of 4,21-dehydrogeissoschizine. J. Chem. Soc. Chem. Commun. 1015–1016.

Kan-Fan, C., and Husson, H.P. (1980). Biomimetic synthesis of yohimbine and heteroyohimbine alkaloids from 4,21-dehydrogeissoschizine. Tetrahedron Lett. *21*, 1463–1466.

Kaur, R., Kaur, G., Gill, R.K., Soni, R., and Bariwal, J. (2014). Recent developments in tubulin polymerization inhibitors: an overview. Eur. J. Med. Chem. *87C*, 89–124.

Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Miya, E., Tomita, M., and Yanagawa, H. (2008). Six classes of nuclear localization signals specific to different binding grooves of importin alpha. J. Biol. Chem. *284*, 478–485.

Lee, C., Bedgar, D.L., Davin, L.B., and Lewis, N.G. (2013). Assessment of a putative proton relay in *Arabidopsis* cinnamyl alcohol dehydrogenase catalysis. Org. Biomol. Chem. *11*, 1127–1134.

Liscombe, D.K., and O'Connor, S.E. (2011). A virus-induced gene silencing approach to understanding alkaloid metabolism in *Catharanthus roseus*. Phytochemistry *72*, 1969–1977.

Luijendijk, T.J.C., Stevens, L.H., and Verpoorte, R. (1998). Purification and characterisation of strictosidine β-D-glucosidase from *Catharanthus roseus* cell suspension cultures. Plant Physiol. Biochem. *36*, 419–425.

O'Connor, S.E., and Maresh, J.J. (2006). Chemistry and biology of monoterpene indole alkaloid biosynthesis. Nat. Prod. Rep. *23*, 532–547.

Rueffer, M., Kan-Fan, C., Husson, H.P., Stoeckigt, J., and Zenk, M.H. (1979). 4,21-Dehydrogeissoschizine, an intermediate in heteroyohimbine alkaloid biosynthesis. J. Chem. Soc. Chem. Commun. 1016–1018.

Saslowsky, D.E., Warek, U., and Winkel, B.S. (2005). Nuclear localization of flavonoid enzymes in *Arabidopsis*. J. Biol. Chem. *280*, 23235–23740.

Stoeckigt, J., Treimer, J., and Zenk, M.H. (1976). Synthesis of ajmalicine and related indole alkaloids by cell free extracts of *Catharanthus roseus* cell suspension cultures. FEBS Lett. *70*, 267–270.

Stoeckigt, J., Husson, H.P., Kan-Fan, C., and Zenk, M.H. (1977). Cathenamine, a central intermediate in the cell free biosynthesis of ajmalicine and related indole alkaloids. J. Chem. Soc. Chem. Commun. 164–166.

Stoeckigt, J., Hemscheidt, T., Hoefle, G., Heinstein, P., and Formacek, V. (1983). Steric course of hydrogen transfer during enzymatic formation of 3α-heteroyohimbine alkaloids. Biochemistry *22*, 3448–3452.

Waadt, R., Schmidt, L.K., Lohse, M., Hashimoto, K., Bock, R., and Kudla, J. (2008). Multicolor bimolecular fluorescence complementation reveals simultaneous formation of alternative CBL/CIPK complexes in planta. Plant J. *56*, 505–516.

Yerkes, N., Wu, J.X., McCoy, E., Galan, M.C., Chen, S., and O'Connor, S.E. (2008). Substrate specificity and diastereoselectivity of strictosidine glucosidase, a key enzyme in monoterpene indole alkaloid biosynthesis. Bioorg. Med. Chem. Lett. *18*, 3095–3098.

Zenk, M.H. (1980). Enzymic synthesis of ajmalicine and related indole alkaloids. J. Nat. Prod. *43*, 438–451.

The structural basis of ajmalicine biosynthesis: Active site elements that control stereoselectivity in alkaloids

Anna Stavrinides[1,2], Evangelos C. Tatsis[1,2], Lorenzo Caputi[2], Emilien Foureau[3], Clare E. M. Stevenson[2], David M. Lawson[2], Vincent Courdavault[3]*, Sarah E. O'Connor[2]*

[1] These authors contributed equally

[2] The John Innes Centre, Department of Biological Chemistry, Norwich NR4 7UH, UK

[3] Université François-Rabelais de Tours, EA2106 'Biomolécules et Biotechnologies Végétales', Tours, France

sarah.oconnor@jic.ac.uk

vincent.courdavault@univ-tours.fr

Keywords:

Monoterpene indole alkaloid, ajmalicine, tetrahydroalstonine, Corynanthe, medium chain dehydrogenase/reductase, *Catharanthus roseus*

Abstract

Plants produce an enormous array of biologically active metabolites, often with stereochemical variations on the same molecular scaffold. These changes in stereochemistry dramatically impact biological activity. Notably, the stereoisomers of the heteroyohimbine alkaloids show diverse pharmacological activities. We reported a medium chain dehydrogenase/reductase from *Catharanthus roseus* that catalyzes formation of a heteroyohimbine isomer. Here we report the discovery of additional heteroyohimbine synthases, one of which produces a mixture of diastereomers. The crystal structures for three heteroyohimbine synthases have been solved, providing insight into the mechanism of

reactivity and stereoselectivity, with mutation of one loop transforming product specificity. Localization and gene silencing experiments provide a basis for understanding the function of these enzymes in vivo. This work sets the stage to explore how medium chain dehydrogenase/reductases evolved to generate structural and biological diversity in specialized plant metabolism and opens the possibility for metabolic engineering of new compounds based on this scaffold.

Introduction

Heteroyohimbines are a prevalent subclass of the monoterpene indole alkaloids (Corynanthe type skeleton), having been isolated from many plant species, primarily from the Apocynaceae and Rubiaceae families.[1] These alkaloids exhibit numerous biological activities: ajmalicine is an α1-adrenergic receptor antagonist,[2-5] and mayumbine (19-epi-ajmalicine) is a ligand for the benzodiazepine receptor (**Figure 1**).[6] Oxidized beta-carboline heteroyohimbines also exhibit potent pharmacological activity: serpentine has shown topoisomerase inhibition activity[7] and alstonine has been shown to interact with 5-HT2A/C receptors and shows promise as an anti-psychotic agent.[8-13] Additionally, heteroyohimbines are biosynthetic precursors of many oxindole alkaloids, which also display a wide range of biological activities.[14] A total of 16 heteroyohimbine stereoisomers are possible, though only eight are reportedly found in nature (C3, C19, C20, **Fig. 1**).[14-20] How and why the stereoselectivity is controlled in the biosynthesis of these alkaloids remains unclear.

The medicinal plant *Catharanthus roseus* produces three of these isomers, ajmalicine (raubasine), tetrahydroalstonine and 19-epi-ajmalicine (mayumbine) (**Fig. 1**).[21] These heteroyohimbines, along with the majority of monoterpene indole alkaloids, are derived from deglycosylated strictosidine (strictosidine aglycone).[22] The removal of a glucose unit from strictosidine by strictosidine glucosidase (SGD) forms a reactive dialdehyde intermediate that can rearrange to form numerous isomers.[23] The stabilization of these isomers by enzyme-catalyzed reduction is hypothesized to be the stepping stone for the extensive chemical diversity observed in the monoterpene indole alkaloids (**Fig. 1**).[21,22] We recently

reported the first cloning of a biosynthetic gene encoding an enzyme that acts on strictosidine aglycone. This zinc-dependent medium chain dehydrogenase/reductase (MDR), named tetrahydroalstonine synthase (THAS), produces the heteroyohimbine tetrahydroalstonine (**Fig. 1**).[24]

In this study, we assayed 14 MDR homologues identified from the *C. roseus* transcriptome[25,26] that have homology to THAS (Cr_024553). This screen revealed three additional enzymes with THAS activity (Cr_010119, Cr_021691, Cr_032583a), and, importantly, an enzyme that produced a mixture of heteroyohimbine diastereomers (Cr_032583b). To understand how these enzymes synthesize the heteroyohimbine scaffold, we solved the crystal structure of THAS (here referred to as THAS1), a second representative THAS (Cr_021691, THAS2) and the structure of the promiscuous homologue (Cr_032583b, heteroyohimbine synthase, HYS). Based on the structures and sequences of these enzymes, we designed mutants that revealed key residues that control the stereochemistry of the product profiles. Studies with isotopically labelled substrates suggested the identity of reaction intermediates and the stereochemical course of reduction. Virus induced gene silencing was used to explore the physiological relevance of these genes. Notably, analysis of the sub-cellular localization of some of these heteroyohimbine synthases indicates an unusual nuclear localization pattern and an interaction with the previous enzyme, SGD. In summary, the discovery of a promiscuous heteroyohimbine synthase homologue, along with the delineation of the structure and localization of three members of this enzyme class, provides insight into the mechanism and evolution of a crucial branch point in a specialized metabolic pathway with pharmacological and evolutionary importance.

Results

Discovery of heteroyohimbine synthases

Guided by our initial discovery of THAS1[24] we identified candidates from the MDR protein family in the *C. roseus* transcriptome[25,26] based on amino acid similarity to this enzyme

(Supplementary Information, Supplementary Table 1). Each of these candidates was cloned from *C. roseus* cDNA and expressed in *E. coli*, with the exception of Cr 017994, which could not be expressed and was not considered further. The remaining candidates were assayed with the substrate strictosidine aglycone, and product formation was monitored by LC-MS. Of these, four (Cr_010119, Cr_021691, Cr_032583a, Cr_032583b) reduced strictosidine aglycone to a product corresponding to one of the heteroyohimbines (**Fig. 2**, **Supplementary Fig. 1**). The products of the enzymatic reactions were identified based on liquid chromatography mass spectrometry (LC-MS) data and comparison to authentic standards (**Supplementary Fig. 2**). Enzymes that failed to produce a heteroyohimbine product were not studied further (**Supplementary Fig. 1**). Three of the enzymes (Cr_021691, THAS2; Cr_010119, THAS3; Cr_032583a, THAS4) produced tetrahydroalstonine in approximately 85% yield, with small amounts of 19-epi-ajmalicine (mayumbine) (< 15%) also observed in these reactions, similar to the previously reported THAS1. Notably, one enzyme (Cr_032583b, heteroyohimbine synthase (HYS)) produced a dramatically different product profile consisting of a mixture of ajmalicine: tetrahydroalstonine: mayumbine (55: 27: 15, at pH 6) (**Fig. 2**). The discovery of this enzyme, HYS, now provides a molecular basis to understand the generation of stereochemical diversity in this alkaloid family.

Crystallography of three heteroyohimbine synthases

To understand the mechanism of stereochemical control at this crucial biosynthetic branch point, we crystallized three heteroyohimbine synthases. THAS1 and THAS2, which produce predominantly tetrahydroalstonine, were both crystallized, since their amino acid sequence identity is relatively low (55%) and the predicted active sites have numerous differences (**Fig. 3**). HYS, which has a distinctly different product profile, was also crystallized to explore the structural basis behind this distinct stereochemical outcome. Structures were obtained for THAS1 and THAS2 in both apo form (THAS1, 2.25 Å resolution; THAS2, 2.05 Å resolution) and with $NADP^+$ bound (THAS1, 1.05 Å resolution; THAS2, 2.10 Å resolution),

while HYS could only be crystallized in the apo form (2.25 Å resolution) (**Fig. 4**, **Supplementary Fig. 3-5**, **Supplementary Table 3**).

Structural features of heteroyohimbine synthase active sites

The five heteroyohimbine synthase structures described here are similar to sinapyl alcohol dehydrogenase (SAD; PDB accession codes 1YQX and 1YQD) and the SAD homolog cinnamyl alcohol dehydrogenase (CAD; PDB accession codes 2CF5 and 2CF6),[27-29] which reduce the aldehyde moiety of lignin precursors. Indeed, pairwise superpositions of subunits from these structures gave RMSD values of less than 2 Å (**Supplementary Table 4**). The biological unit is an elongated homodimer, with each subunit divided into a substrate and cofactor binding domain; the latter also being responsible for forming the dimer interface (**Supplementary Fig. 3**).

The active site cavities of the heteroyohimbine synthases are framed by helix α2, the catalytic zinc coordination sphere, and loops 1 and 2, with the NADP(H) co-substrate binding at the base of the active site (**Fig. 3**, **4**, **Supplementary Fig. 4**). Loop 2, which is positioned above the active site, is highly variable in length and sequence (**Fig. 3**, **4**). In both THAS1 and THAS2, a network of amino acids holds $NADP^+$ in place (**Fig. 4**). Most notably, Glu59 of THAS1 anchors NADP(H) through a bidentate interaction with both ribose hydroxyls, with His59 playing a comparable role in SAD, although here the interaction is with the 3' OH only. Glu59 is conserved in HYS, but an aspartate residue (THAS3 and THAS2) or a tyrosine residue (THAS4) serves this role in other homologues. MDRs usually contain two zinc ions,[30] a distal "structural" zinc ion, which in this case is coordinated by four cysteine residues, and a proximal "catalytic" zinc ion near the active site, which is coordinated by two cysteines, one histidine and one glutamate residue (**Fig. 3**, **4**).[27,31] The proximal zinc of THAS1 is approximately 2 Å further away from the cofactor relative to SAD and thus may play no direct role in catalysis (**Supplementary Fig. 4**). However, it may have a function in maintaining the tertiary structure since three of the liganding residues are in the substrate binding domain and the fourth is in the cofactor binding domain. SAD/CAD utilize an active site serine that

protonates the alkoxide that results from reduction of the aldehyde substrate;[27,29] this serine is replaced with a tyrosine residue in THAS1 (Tyr56) and HYS (Tyr53) (**Fig. 3**). In THAS2, this tyrosine on helix α2 is replaced with a tryptophan residue, but a tyrosine at position 120 points into the active site. Interestingly, a non-proline *cis*-peptide is present in the THAS1 $NADP^+$ and HYS apo structures (**Supplementary Fig. 5**). Closer inspection of this region in THAS1 shows that when this bond is in the *trans* conformation, the side chain of Asp340 is projected into the cofactor binding site such that it would prevent NADPH binding.

Strictosidine aglycone binding

Despite extensive efforts, both product and substrate failed to co-crystallize with any of the enzymes. Therefore, molecular docking was used to visualize the position of strictosidine aglycone in THAS1 (**Fig. 4**). To ensure that the correct substrate tautomer was used for docking, we identified the most predominant strictosidine aglycone isomer that forms in solution. Although product precipitation prevented monitoring the SGD reaction *in situ* under aqueous conditions (Methods), $^{1}H,^{15}N$-HMBC NMR showed that an enamine species was the most predominant product in aqueous methanolic solution (**Supplementary Fig. 6**). This is consistent with literature reports that cathenamine is the major product of SGD, and is the proposed precursor of ajmalicine and tetrahydroalstonine (**Fig. 1**).[21,23] *In silico* docking with THAS1 positions cathenamine between the nicotinamide of the $NADP^+$ and Tyr56, which is located on helix α2 (**Fig. 4**). THAS1 loop 1 contains Phe65 that also projects into the active site and may interact with the aromatic cathenamine substrate.

Mechanism of reduction and heteroyohimbine formation

In tetrahydroalstonine biosynthesis, we hypothesize that cathenamine tautomerizes to the iminium form by protonation at C20, followed by addition of the hydride at C21. Protonation at C20 must occur from the bottom face to yield the S stereochemistry observed at this position (**Fig. 5**). While there does not appear to be an appropriately positioned active site residue to perform this role, the crystal structures of these enzymes reveal the presence of

numerous water molecules in the active site that could potentially protonate this carbon (**Fig. 4A**). $^{1}H,^{15}N$-HMBC measurements of strictosidine aglycone at different pH values shows formation of the iminium species in solution when the pH was reduced to approximately 3.5, indicating that this tautomer can readily form in the presence of an acidic moiety (**Supplementary Fig. 6**).

To elucidate the stereo and regioselectivity of reduction by NADPH, we isolated tetrahydroalstonine from reactions using THAS1 and pro-*R* NADPD. Analysis by ^{1}H-NMR showed that tetrahydroalstonine is labeled with deuterium in the pro-*R* position at C21, consistent with previously reported experiments performed in crude cell extracts (**Fig. 6**).[32] It is possible that THAS1 could reduce the enamine directly, in which case hydride addition would occur at C21, followed by protonation at C20 by a water molecule as described above. The presence of mayumbine/19-epi-ajmalicine in some of the enzymatic reactions suggests that small amounts of cathenamine can open and form 19-epi-cathenamine, either in solution or in the active site.

In the case of HYS, which produces both ajmalicine (*R* C20) and tetrahydroalstonine (*S* C20), protonation must also occur from the opposite face to yield *R* stereochemistry at C20. Products of HYS generated with pro-*R* NADPD were also isolated and analyzed by ^{1}H-NMR, and as for tetrahydroalstonine from THAS1, in each case showed deuterium labelling in the pro-*R* position at C21 (**Fig. 6**). Therefore, the stereochemical course of hydride addition is not altered in HYS compared to THAS1.

The major difference between HYS and THAS1/THAS2 appears to be the extended loop over the HYS active site (D125-GHFGNN-F132 in HYS and D128-SN-Y131 in THAS1, loop 2 in **Fig. 3**). The histidine residue in HYS loop 2 (His127) appears to be positioned appropriately to provide an alternative proton source for the opposite ("top") face of the substrate, which could explain the appearance of ajmalicine in the product profile of HYS. Reactions with THAS1 and HYS performed at different pH conditions (5 to 8) revealed that while changes in pH did not substantially impact the product profile of THAS1, HYS

produced increased amounts of ajmalicine relative to tetrahydroalstonine at pH 6 compared to higher pH values (**Supplementary Fig. 7**). The increased level of ajmalicine in HYS at lower pH values is consistent with the pKa value of the histidine side chain and supports the role of this histidine in ajmalicine biosynthesis, though attributing pH dependence to specific residues must be approached with caution.[33]

Switching stereoselectivity of heteroyohimbine synthases

Since the major sequence and structural difference between HYS and THAS1 is the extended loop over the HYS active site, loop 2 (**Fig. 3**), we swapped these loop regions in THAS1 and HYS to determine whether the stereochemical product profile could be switched. Loop1, which is near the active site, was also swapped. While the THAS1 mutant containing the swaps displayed reduced activity rather than an altered product profile, the HYS mutant containing the shorter THAS1 loop2 resulted in a product profile similar to that of THAS1 (**Fig. 7**, **Supplementary Fig. 8**). Since His127 is the only ionizable residue in this loop, we hypothesized that this residue protonates C20, as discussed above. Therefore, we mutated this histidine to alanine or asparagine in HYS. Both of these mutants gave the same THAS-like profile, suggesting that His127 is required for producing the ajmalicine (*R* C20) stereochemistry (**Supplementary Fig. 9**). Mutation of other conserved ionizable residues in the THAS1 active site (Tyr56, Ser102 and Thr166) did not result in substantial changes in the distribution of products (**Supplementary Table 6**, **Supplementary Fig. 10**, **11**). Mutations to Glu59, which anchors the NADPH cofactor, resulted in a slight increase in product promiscuity (**Supplementary Fig. 11**), perhaps by causing a shift in the cofactor position. The reactivity of the substrate[34] makes obtaining accurate kinetic constants challenging,[24] so endpoint assays were used to measure activity levels of mutant enzymes (**Supplementary Table 6**)

In planta localization of heteroyohimbine synthases

Plants use spatial organization on the organ, tissue and intracellular levels to control product distribution. This additional layer of control is critical for establishing the product profiles in the whole organism. Notably, expression profile data reveal variations in the expression levels of heteroyohimbine synthase transcripts (**Supplementary Fig. 12**). At the subcellular level, we previously showed that THAS1 has an unusual nuclear localization pattern,[24] which is also where SGD, the enzyme that synthesizes strictosidine aglycone, is localized.[34] Physical interactions using Bimolecular Fluorescence Complementation (BiFC) were observed between these two enzymes.[24] THAS2 and HYS localization were similarly investigated by expressing yellow fluorescent protein (YFP) fusions in *C. roseus* cells. Microscopy of transiently transformed cells revealed that THAS2-YFP was located in both the cytosol and the nucleus while HYS-YFP, similar to THAS1, displayed a preferential nuclear localization (**Fig. 8** and **Supplementary Fig. 13**). As reported for THAS1, this localization relies on the presence of a class V nuclear localization sequence in HYS (215-KKKR-218) that is absent from THAS2 (**Fig. 3**). BiFC assays revealed that both THAS2 and HYS are capable of self-interactions (**Supplementary Fig. 14**).

BiFC assays were used to determine whether THAS2 and HYS also interact with SGD (**Fig. 9**). N-terminal split-YFP fragment fusions of both enzymes (THAS2-YFP^N and HYS-YFP^N) were co-transformed with SGD that was fused to a C-terminal split-YFP fragment (YFP^C-SGD). The formation of a nuclear BiFC complex suggests that both of these enzymes interact with SGD in the nucleus (**Fig. 9 A-D**). Interestingly, the emitted fluorescent signal exhibited a punctated, sickle shaped aspect as previously observed for the THAS1/SGD interaction (**Fig. 9 E-F**) and for SGD localization.[34] In contrast, no interactions were detected when the BiFC assay was conducted with a downstream enzyme from this biosynthetic pathway, 16-hydroxytabersonine 16-O-methyltransferase (16OMT), that is not expected to interact with SGD (**Fig. 9 G-H**).

Double BiFC assays were performed to combine the study of THAS2 and HYS interactions as well as their interactions with SGD. After transformation into plant cells (16 hours), we noted the formation of a dual fluorescent signal for THAS2, both in the cytosol and as

punctates in the nucleus that may correspond to the superposition of the signal observed for THAS2 self-interactions and THAS2-SGD interaction (**Fig. 9 I-J**) as confirmed by multicolor BiFC assays (**Supplemental Fig. 15**). Increased time of expression (36 hours) progressively resulted in the disappearance of the cytosolic signal, and it is intriguing to speculate that this implies a recruitment of THAS2 by SGD (**Fig. 9 Q-R**). A similar phenomenon was observed for HYS and THAS1 (**Fig. 9 K-N**; **S-V- Supplemental Fig. 15**). While self-interactions of 16OMT were detected, no nuclear signal was recovered, confirming the specificity of THAS1-, THAS2-, HYS-SGD interactions (**Fig. 9 O-P**). The interaction of THAS1, THAS2 and HYS with SGD reinforces the hypothesis of an evolutionary mechanism deployed by strictosidine accumulating plants to manage the reactivity of the strictosidine aglycone generated by SGD. It also raises the question of a possible competition between heteroyohimbine synthases for recruitment by SGD when distinct enzymes are co-expressed in the same tissue/cells.

In planta silencing of heteroyohimbine synthases

The genes encoding active heteroyohimbine synthases were silenced to establish whether any of them synthesize the expected metabolic product *in planta*. For many medicinal plants, including *C. roseus*, Virus Induced Gene Silencing (VIGS) is the only established method to silence genes in the whole plant. In *C. roseus,* the effect of VIGS is temporally and spatially limited to the first two leaves that emerge immediately after infection.[35] Each of the genes encoding a biochemically active enzyme (THAS1, THAS2, THAS3, THAS4 and HYS) was subjected to VIGS in *C. roseus* seedlings and the effect on alkaloid production was monitored by mass spectrometry. Since HYS and THAS4 were too similar to silence separately, one common gene fragment was used for silencing both genes simultaneously. Successful silencing of the genes was confirmed by qRT-PCR (**Supplementary Fig. 16**). Aside from a small degree of cross-silencing between THAS2 and THAS3 (12%), all of the target genes were silenced selectively, as measured by qRT-PCR (**Supplementary Fig. 16**).

Due to the inherent reactivity of the heteroyohimbine synthase substrate (strictosidine aglycone), changes in the level of this compound *in planta* are difficult to accurately detect. Instead, the effect of silencing must be established by quantitatively measuring decreases in heteroyohimbine levels. Previously for THAS1, we measured the combined peak for heteroyohimbines, since the diastereomers were difficult to resolve on a reverse phase LC column under the reported conditions.[24] However, after substantial optimization (see Methods), an LC-MS method was developed to separate ajmalicine and tetrahydroalstonine in crude leaf extracts (19-epi-ajamlicine/mayumbine is not observed in *C. roseus* leaves, **Supplementary Fig. 17**). Unfortunately, ajmalicine, and particularly tetrahydroalstonine, are observed in low levels even in empty vector control samples, and heteroyohimbine composition varied substantially among individual leaves. Therefore, accurate measurement of decreases in ajmalicine and tetrahydroalstonine levels is challenging. There was no evidence for a decrease of ajmalicine or tetrahydroalstonine when THAS2 and THAS3 were silenced. However, for HYS, there was a statistically significant decrease (t-test 0.0275) in ajmalicine, and no change in tetrahydroalstonine levels. Surprisingly, a statistically significant decrease in ajmalicine as well as tetrahydroalstonine (t-test 0.0277 and 0.0276 respectively) was also noted for THAS1 (**Supplementary Fig. 16**). While the results are statistically significant, the leaf-to-leaf variability, the low level of endogenous production, and the catalytic redundancy of these enzymes make it difficult to draw firm conclusions from these VIGS data. Additionally, regulatory factors that impact the ratio of ajmalicine and tetrahydroalstonine, such as transport mechanisms and/or further derivatization to other products, cannot be ruled out. Additional silencing systems, in different tissues, will be required to more firmly establish the physiological function of these enzymes. Nevertheless, we can state that silencing of HYS and THAS1 impacts alkaloid production in *C. roseus* leaves.

Discussion

Here we report several medium chain dehydrogenases/reductases that produce the heteroyohimbine stereoisomers ajmalicine and/or tetrahydroalstonine, thereby providing a framework to understand the enzymatic control over stereoselectivity in this metabolic pathway. It is notable that we have identified four enzymes that generate tetrahydroalstonine, yet ajmalicine is the more abundant isomer *in planta* (**Supplementary Fig. 17**). Expression profile data of the genes identified in this study suggest that HYS, which produces ajmalicine, is not expressed at higher levels than the other synthases (**Supplementary Fig. 12**). There may be additional ajmalicine synthases that are not related to the MDR superfamily homologues identified in this study. Alternatively, tetrahydroalstonine could be shuttled into another pathway or degraded, thereby resulting in the observed lower levels that accumulate *in planta*.

Importantly, the pharmacological activity of heteroyohimbines is impacted by the stereochemistry. Ajmalicine (raubasine) has recently been used in combination with almitrine in post-stroke treatments, though the side effects caused by almitrine resulted in widespread withdrawal of the drug in 2013.[4] While tetrahydroalstonine has no reported pharmacological function, its oxidized product, alstonine (**Fig. 1**), has recently been shown to act by a unique mechanism for modulating dopamine uptake and shows potential as an anti-psychotic drug.[13] The heteroyohimbines have excellent promise as a scaffold for pharmacological activity. The discovery of the heteroyohimbine synthases, along with recently developed heterologous production platforms for monoterpene indole alkaloids,[36] now allows the possibility of generating these alkaloids and unnatural derivatives through metabolic engineering/synthetic biology strategies.

While the *in planta* function of heteroyohimbines is unknown, degycosylated strictosidine is toxic and may act as a defense compound,[34] similar to the defense roles of the aglycones of the iridoids from which stictosidine is derived.[37,38] SGD is expressed in most tissues (**Supplementary Fig. 12**), suggesting that the plant must have evolved mechanisms to control the levels of the toxic strictosidine aglycone. In directed overflow metabolism, excess reactive intermediates are converted into non-reactive byproducts.[39] It is intriguing to

speculate that monoterpene indole alkaloid biosynthesis may have initially arisen as a mechanism for handling overflow of strictosidine aglycone. The heteroyohimbine synthases perform a single, chemically straightforward reduction reaction that immediately neutralizes the reactivity of strictosidine aglycone/cathenamine. The co-localization of SGD and at least some of the heteroyohimbine synthases supports this hypothesis. Whether the heteroyohimbines serve an active biological function in the plant, or whether they are simply the end product of directed overflow metabolism, or both, remains to be investigated. Regardless, it is clear that MDRs play an important role in the generation of a wide variety of chemical structures. Duplication of the evolutionary dehydrogenase ancestor may have given rise to multiple heteroyohimbine synthases, along with MDRs with other biosynthetic activities, such as tabersonine-3-reductase that is involved in the biosynthesis of the anti-cancer alkaloid vinblastine (**Supplementary Fig. 18**).[40]

Methods

Selection and Cloning of candidate MDRs. The nucleotide and the protein sequences of THAS1 were subjected to a BLAST search against the *Catharanthus roseus* Sunstorm Apricot V1.0 Transcript sequences (http://medicinalplantgenomics.msu.edu) and the MDR sequences with the highest identity to THAS1 at the active site and which showed non-negligible expression levels in young and mature leaves were selected as candidates for cloning and expression. The protein sequence of Sinapyl Alcohol Dehydrogenase (SAD) was blasted against the same database and MDRs were also selected based on their active site similarity to that of SAD. The genes coding the candidate MDRs were amplified from *C. roseus* leaf cDNA and cloned into the *E. coli* expression vector pOPINF using the In-Fusion cloning kit (Clontech Takara)[41] by using primers designed based on the transcript sequences (**Supplementary Table 1**).

Site directed mutagenesis of THAS1 and HYS. THAS1 mutants were generated by overlap extension PCR. Briefly, the codon to be mutated was selected and two primers, one reverse and one forward (**Supplementary Table 4**), were designed to overlap and introduce

the mutation. A first PCR was carried out using the reverse mutant primer and the 5' forward gene-specific primer (**Supplementary Table 1**), thus generating the 5' half of the gene carrying the mutation. In parallel, the 3' half of the mutated gene was generated by PCR using the forward mutant primer and the 3' reverse gene-specific primer (**Supplementary Table 4**). PCR products were gel purified and used for the second PCR overlap reaction for generation of the full-length mutated gene where the 5' and 3' halves of the mutated genes were mixed in a PCR reaction in equimolar amounts (approx. 100 ng per fragment) and 5 cycles of PCR were carried out without including primers. After the 5 overlap PCR cycles the forward and reverse gene-specific primers were added to the mix and a further 30 cycles were performed. Full-length PCR products were gel purified, ligated into pOPINF expression vector and transformed into competent *E. coli* Stellar strain cells (Clontech Takara). HYS point mutants were obtained as gene fragments (Integrated DNA Technologies, Belgium) with the H127 or F128 codons mutated (H127A CAT -> GCA; H127N CAT -> AAC; F128A TTT -> GCT; F128Y TTT -> TAC) and the pOPINF overhangs included at the 3' and 5' extremities. The THAS1 and HYS double loop mutants were generated by first making their loop1 mutant genes and then inserting the second loop2 swap following the same procedure described above. Mutant constructs were sequenced to verify the mutant gene sequence and correct insertion.

Enzyme activity assays. All candidate enzymes and mutants were expressed in SoluBL21 (DE3) *E. coli* cells (Genlantis) grown in 2xYT medium. Protein production was induced by addition of 0.2 mM IPTG and the cultures were shaken at 18°C for 16 h. Cells were collected by centrifugation, lysed by sonication in Buffer A (50 mM Tris-HCl pH 8, 50 mM glycine, 500 mM NaCl, 5% v/v glycerol, 20 mM imidazole) supplemented with EDTA-free protease inhibitor (Roche Diagnostics Ltd.) and 0.2 mg ml^{-1} lysozyme. Soluble proteins were purified on Ni-NTA agarose (Qiagen) and eluted with Buffer B (50 mM Tris-HCl pH 8, 50 mM glycine, 500 mM NaCl, 5% v/v glycerol, 500 mM imidazole). Eluates were analysed by SDS-PAGE to verify the purity and the molecular weight of the purified proteins. All proteins were dialysed in Buffer C (50 mM phosphate pH 7.6, 100 mM NaCl) and concentrated. Protein

concentration was measured with Bradford reagent (Sigma-Aldrich) according to the manufacturer's instructions. Purified proteins were divided in 20 µl aliquots, fast-frozen in liquid nitrogen and stored at -20°C.

Candidate MDR enzymes and the selected mutants were screened for activity against deglycosylated strictosidine. The substrate was generated by deglycosylating strictosidine (300 µM) by the addition of purified SGD in the presence of 50 mM phosphate buffer (pH 6.5) at room temperature for 10 minutes. The reactions were started by the addition of MDR enzyme (1 µM) and NADPH (5 mM). Caffeine (50 µM) was used as internal standard. All reactions were performed in triplicate. Aliquots of the reaction mixtures (10 µl) were sampled 1 minute and 30 min after addition of MDR enzyme. The reactions were stopped by the addition of 10 µl of 100% MeOH. Samples were diluted 1:5 in mobile phase (H_2O + 0.1% formic acid) and centrifuged for 10 minutes at 4000 *g* before UPLC-MS injection (1 µl). The activity of MDR enzymes and mutants was measured by UPLC-MS.

Protein crystallization. Proteins for crystallization were purified from 2 L cultures in 2xYT medium. Protein expression was induced by addition of IPTG and the cultures were grown for 16 h at 18°C. Cells were collected by centrifugation and lysed by sonication in 50 ml Buffer A supplemented with EDTA-free protease inhibitor and 10 mg of Lysozyme. Lysates were clarified by centrifugation at 17000 *g* for 20 min. 2D automated purification was performed on an ÄKTAxpress purifier (GE Healthcare). The IMAC step was performed on HisTrap HP 5 ml columns (GE Healthcare) equilibrated with Buffer A. Proteins were step-eluted with Buffer B and directly injected on a gel filtration column equilibrated with Buffer D (20mM HEPES, 150mM NaCl, pH 7.5). Fractions were collected and analysed by SDS-PAGE and those containing pure protein were pooled and concentrated in a 10 kDa membrane filter Millipore filter (Merck Millipore).

Purification of HYS required the addition of 1 mM DTT to all purification buffers and dialysis in Buffer D containing 0.5 mM *tris*(2-carboxyethyl)phosphine (TCEP) before crystallization and storage.

Crystallization screens were conducted by sitting-drop vapor diffusion in MRC2 96-well crystallization plates (Swissci) with a mixture of 0.3 µl well solution from the PEGs (Qiagen), PACT (Qiagen) and JCSG (Molecular Dimensions) suites and 0.3 µl protein solution. Protein concentrations were adjusted to 7-10 mg ml^{-1} whilst $NADP^+$ (Sigma Aldrich) was added to a final concentration of 1 mM for co-crystallization studies. Solutions were dispensed either by an OryxNano or an Oryx8 robot (Douglas Instruments).

THAS1 apo crystals were obtained from His_6-tag cleaved THAS1 (3C protease) in a solution containing 0.1 M MES, pH 6.5, 15% w/v PEG 2000. THAS1 $NADP^+$ crystals were obtained from a solution containing 0.2 M potassium/sodium tartrate with 20% w/v PEG 3350. THAS2 crystals (with and without $NADP^+$) were obtained from a condition containing 0.2 M lithium chloride and 20% w/v PEG 3350. HYS crystals were obtained after removal of the His_6-tag (using 3C protease) in 0.1 M MMT buffer, pH 5 and 15% w/v PEG 3350. All crystals were cryoprotected by soaking in crystallization solution containing 25% v/v ethylene glycol before flash-cooling in liquid nitrogen.

Data collection and structure determination. X-ray datasets were recorded on one of three beamlines at the Diamond Light Source (Oxfordshire, UK) (2FI3, I04; 2FI5, I03; 5H81 I04-1; 5H82, I04-1; 5H83, I04-1) at wavelengths of 0.9000-0.976 Å (2FI3, 0.900 Å; 2FI5, 0.976 Å; 5H81, 0.920 Å; 5H82, 0.920 Å; 5H83, 0.920 Å) using either a Pilatus 6M or 2M detector (Dectris) with the crystals maintained at 100 K by a Cryojet cryocooler (Oxford Instruments). Diffraction data were integrated using XDS[42] and scaled and merged using AIMLESS[43] via the XIA2 expert system;[44] data collection statistics are summarized in Supplementary Table 3. Initially the THAS1 $NADP^+$ dataset was automatically processed at the beamline by fast_dp[45] to 1.12 Å resolution and a structure solution was automatically obtained by single wavelength anomalous dispersion phasing using the SHELX suite[46] via the fast_ep pipeline (Winter, manuscript in preparation). Despite being collected at a wavelength somewhat remote from the zinc *K* X-ray absorption edge (theoretical wavelength 1.284 Å) the anomalous signal was sufficient for fast_ep to locate four zinc sites and calculate a very clear experimentally phased electron density map (**Fig. 4A**). This was

available to view at the beamline in the ISPyB database[47] via the SynchWeb interface[48] within a few minutes of completing the data collection. The map was of sufficient quality to enable 94% of the residues expected for a THAS1 homodimer to be automatically fitted using BUCCANEER.[49] The model was finalised by manual rebuilding in COOT[50] and restrained refinement using anisotropic thermal parameters in REFMAC5[51] against the same dataset reprocessed to a resolution of 1.05 Å as described above (Supplementary Table 3), and contained 97% of the expected residues, with one $NADP^+$ molecule and two zinc ions per subunit. All the remaining structures were solved by molecular replacement using PHASER.[52] In each case, the asymmetric unit corresponded to the biological dimer and the preliminary models were obtained by searching for two copies of a monomer template. For THAS1 apo, THAS2 $NADP^+$ and HYS apo, a THAS1 $NADP^+$ protein only monomer model was used as the basis for the template, although in the latter two cases a homology model of the target structure was generated from the THAS1 template using the Phyre2 server[53] (http://www.sbg.bio.ic.ac.uk/~phyre2) before running PHASER. For solving theTHAS2 apo structure, a THAS2 $NADP^+$ monomer was used as the template. In contrast to THAS1 $NADP^+$, these four structures were refined in REFMAC5 with isotropic thermal parameters and TLS group definitions obtained from the TLS-MD server.[54] Model geometries were validated with the MOLPROBITY[55] tool before submission to the PDB. The statistics of the final models are summarized in **Supplementary Table 3**. Additional statistics for R_{pim}: 5FI3, 0.020 (0.517); 5FI5, 0.038 (0.600); 5H81, 0.041 (0.349); 5H82, 0.033 (0.439); 5H83, 0.068 (0.664) and $CC_{½}$: 2FI3, 0.999 (0.510); 2FI5, 0.999 (0.523); 5H81, 0.998 (0.725); 5H82, 0.999 (0.639); 5H83, 0.996 (0.510) (where values in parentheses are for highest-resolution shell) were also noted. Ramachandran statistics (favored/allowed/outlier (%)) are 5FI3, 96.8/3.2/0.0; 5FI5, 96.0/4.0/0.0; 5H81, 96.2/3.8/0.0; 5H82, 96.1/3.9/0.0; 5H83, 96.6/3.1/0.3. All structural figures were prepared using CCP4mg.[56]

UPLC-MS and NMR analysis. UPLC-MS analysis was carried out on a UPLC (Waters) equipped with an Acquity BEH C18 1.7 µm 2.1 x 50 mm column connected to Xevo TQS (Waters). For fast dereplication of active enzymes and mutants, a linear gradient method

(Method 1) was used at a flow rate of 0.6 ml min^{-1} using a binary solvent system in which solvent A1 was 0.1% formic acid in water and solvent B1 was acetonitrile. The gradient profile was: 0 min, 5% B1; from 0 to 3.5 min, linear gradient to 35% B1; from 3.5 to 3.75 min, linear gradient to 100% B1; from 3.75 to 4 min, wash at 100% B1; back to the initial conditions of 5% B1 and equilibration for 1 min before the next injection. Column temperature was held at 30 °C. The injection volume for both the solutions of standard compounds and the samples was 1 µl. Samples were kept at 10 °C during the analysis.

For separation of the different heteroyohimbines, a different chromatographic method was applied that was adapted from the work of Sun J. *et al*.[57] In this method (Method 2) solvent A2 was 0.1% NH_4OH and solvent B2 was 0.1% NH_4OH in acetonitrile. A linear gradient from 0% to 65% B2 in 17.5 min was applied for separation of the compounds followed by an increase to 100% B2 at 18 min, a 2 min wash step and a re-equilibration at 0% B2 for 3 min before the next injection. The column was kept at 60 °C throughout the analysis and the flow rate was 0.6 ml min^{-1}.

MS detection was performed in positive ESI. Capillary voltage was 3.0 kV; the source was kept at 150 °C; desolvation temperature was 500 °C; cone gas flow, 50 L h^{-1} and desolvation gas flow, 800 L h^{-1}. Unit resolution was applied to each quadrupole.

Multiple Reactions Monitoring (MRM) signals were used for detection and quantification of caffeine (*m/z* 195 > 110, 138), heteroyohimbine alkaloids (353 > 117, 14).

NMR spectra were acquired using a Bruker Advance NMR instrument operating at 400 MHz for ^{1}H equipped with a BBFO plus 5 mm probe. The number of scans was depended on the concentration of the sample. $^{1}H,^{15}N$-HMBC experiment was acquired with a spectral width 6009 Hz in the F2 (^{1}H) dimension and 30410 in the F1 (^{15}N) and with an acquisition time 0.09 s and 360 scans per increment. The long range delay was optimized after a series of experiments with [4-^{15}N]-strictosidine using a range of different mixing times and finally was adjusted for a coupling of 5 Hz. The relaxation delay was 2.5 s, the data collection matrix was 1024×64, the t1 dimension was zero filled to 1k real data points and a π/2 square sine bell window was applied in both dimensions.

^{2}H labelling experiments. Deuterated Pro-R-NADPD was regenerated in solution by *Thermoanaerobacter brockii* alcohol dehydrogenase (50 units, Sigma) using 400 µM $NADP^+$ and 1% v/v [$^2H_6$]-isopropanol (CIL). The NADPD regeneration was monitored by UV spectroscopy at 340 nm. Strictosidine (19.9 mg) was incubated with 1.27 nM SGD in 94 ml of 50 mM phosphate buffer (pH 6.5). THAS1 enzyme was added to the reaction (final concentration of 1.65 µM) and the mixture was incubated at 35°C with shaking. The reaction was monitored for completeness by UPLC-MS and after 5 h no strictosidine or deglycosylated strictosidine was observed. The reaction was stopped by addition of 100 ml of methanol and reaction mixture was concentrated to dryness. The dried reaction mixture was resuspended in 15 ml H_2O and extracted with 3 x 15 ml of ethyl acetate and the EtOAc fraction was dried. [21α-$^2H_1$]-tetrahydroalstonine was isolated by preparative TLC separation on a nano-silica plate (Sigma-Aldrich), as previously described.[24] The band of [21α-$^2H_1$]-tetrahydroalstonine was excised from the plate, silica was crushed to powder and THA was extracted with EtOAc multiple times, (total volume 40 ml). The EtOAc fraction was filtered and dried using a high-vacuum pump overnight. The [21α-$^2H_1$]-tetrahydroalstonine was dissolved in 600 µl of $CDCl_3$ and ^{1}H NMR was measured.

Strictosidine (39.3 mg) was incubated with 1 nM of SGD and 500 µM $NADP^+$ with 50 units of *T. brockii* ADH and 1% v/v [$^2H_6$]-isopropanol in a total volume of 148 ml of 50 mM HEPES buffer (pH 7.5). HYS was added (final concentration 1.71 µM) and the reaction was incubated at 37°C with shaking and monitored for completeness by UPLC-MS. After 6 h the reaction was complete and was stopped by addition of 150 ml of methanol. The reaction mixture was concentrated to dryness and then was resuspended in 50 ml H_2O, basified with 2 ml triethylamine and extracted with 5 x 20 ml of ethyl acetate. [21α-$^2H_1$]-tetrahydroalstonine, [21α-$^2H_1$]-ajmalicine and [21α-$^2H_1$]-mayumbine were isolated by preparative TLC and ^{1}H NMR spectra measured as described above. ^{1}H NMR spectra of deuterated compounds were compared with those of corresponding standards.

^{15}N labelling experiments. *C. roseus* tryptophan decarboxylase (TDC) was cloned into pOPINF vector, expressed in *E. coli* and purified as described above for the MDRs. [alpha-^{15}N]-tryptophan (CIL, 50 mg) was incubated with 500 nM of TDC, 400 µM pyridoxal-5′-phosphate in 100 ml 50 mM phosphate buffer (pH 7.5) at 35°C. The reaction was monitored by mass spectrometry, continued through completion after 4 h and terminated by addition of 50 ml MeOH. [alpha-^{15}N]-tryptamine (34 mg) was isolated by preparative HPLC. The isolated [alpha-^{15}N]-tryptamine was incubated with 3 mM secologanin and 200 nM strictosidine synthase in 100 ml 50 mM phosphate buffer (pH 7.0) at 30°C overnight. The reaction was terminated by addition of 50 ml MeOH. [4-^{15}N]-strictosidine (62 mg) was isolated by preparative HPLC. [4-^{15}N]-strictosidine was then assayed with SGD and the product characterized by ^{1}H,^{15}N-HMBC as described above.

Subcellular localizations and analysis of protein-protein interactions by bimolecular fluorescence complementation (BiFC). Subcellular localization of THAS2 and HYS were studied by creating fluorescent fusion proteins using the pSCA-cassette YFPi plasmid.[58] The full-length open reading frame of THAS2 was amplified using the specific primers 5'-CTGAGAACTAGTATGTCTTCAAAATCAGCAAAACCAGTG-3' and 5'-CTGAGAACTAGTAGCAGATTTCAATGTGTTTTCTATGTCAAT-3', and HYS ORF with primers 5'-CTGAGAACTAGTATGGCTGCAAAGTCACCTGAAAATGTATAC-3' and 5'-CTGAGAACTAGTGAAAGATGGGGATTTGAGAGTGTTTCCTAC-3', which were designed to introduce the *Spe*I restriction site at both cDNA extremities. PCR products were sequenced and cloned at the 5'end of the yellow fluorescent protein (YFP) coding sequence to generate the THAS2-YFP, HYS-YFP fusion proteins or at the 3' end to express the YFP-THAS2 and YFP-HYS fusions.

The interaction of THAS2 and HYS with SGD were characterized by bimolecular fluorescence complementation (BiFC) assays using the previously amplified THAS2 and HYS PCR products cloned via *Spe*I into the pSPYCE (M) vector,[34] which allows expression of THAS2 and HYS fused to the amino-terminal extremity of the split-YFP^{C} fragment (THAS2-YFP^{C}, HYS-YFP^{C} respectively), and into the pSPYNE(R)173-SGD plasmid[34]

expressing SGD fused to the carboxy-terminal extremity of the split YFP^N fragment (YFP^N-SGD). Plasmids encoding THAS1-YFP^N, THAS1-YFP^C, YFP^C-THAS1 and plasmids expressing 16OMT-YFP^N and 16OMT-YFP^C were used as controls and were constructed previously.[24,59]

THAS2 and HYS self-interactions were analyzed via additional cloning of the THAS2 and HYS PCR products into the pSCA-SPYNE173, pSPYNE(R)173 and pSCA-SPYCE (MR) plasmids,[34,60] to express THAS2-YFP^N, HYS-YFP^C and YFP^C-THAS2, YFP^C-HYS, respectively.

The capacity of THAS2 and HYS to interact with SGD were also characterized by bimolecular fluorescence complementation (BiFC) and multicolor bimolecular fluorescence complementation (mBIFC). The previously amplified THAS2 and HYS PCR product was fused to the amino-terminal or carboxy-terminal of the split YFP fragments into the pSCA-SPYNE173, pSCA-SPYCE (M) and pSCA-SPYCE (MR) plasmids,[34,60] allowing expressing THAS2-YFP^N, YFP^C-THAS2 HYS-YFP^N, and HYS-YFP^C respectively. SGD was subsequently fused to the carboxy-terminal extremity of the split YFP^N fragment (YFP^N-SGD) and the CFP^N fragment (CFP^N-SGD).

Transient transformation of *C. roseus* cells by particle bombardment and fluorescence imaging were performed following the procedures previously described.[58] Briefly, *C. roseus* plated cells were bombarded with DNA-coated gold particles (1 µm) and 1,100 psi rupture disc at a stopping-screen-to-target distance of 6 cm, using the Bio-Rad PDS1000/He system. Cells were cultivated for 16 h to 38 h before being harvested and observed. The subcellular localization was determined using an Olympus BX-51 epifluorescence microscope equipped with an Olympus DP-71 digital camera and a combination of YFP and CFP filters. The pattern of localization presented in this work is representative of *circa* 50 observed cells. The nuclear localizations of the different fusion proteins were confirmed by co-transformation experiments using a nuclear-CFP marker.[34] Such plasmid transformations were performed using 400 ng of each plasmid or 100 ng for BiFC assays.

Agrobacterium-mediated Virus-Induced Gene Silencing and qPCR. The THAS1, THAS2, THAS3 and THAS4-HYS silencing fragments were amplified with primers (Supp. Table 6) and the resulting fragments were cloned into the pTRV2u vector as described.[61] Since THAS4 and HYS are ~ 91% identical it was not possible to design silencing fragments to avoid cross-silencing. Therefore, a common silencing fragment for both of the two genes was designed. The resulting pTRV2u constructs were used to silence the different tetrahydroalstonine synthases and heteroyohimbine synthase in *C. roseus* seedlings essentially as described before.[35] Leaves from the first two pairs to emerge following inoculation were harvested from eight plants transformed with the empty pTRV2u and pTRV2u carrying the silencing fragment. The collected leaves were frozen in liquid nitrogen, powdered using a pre-chilled mortar and pestle, and subjected to LC-MS and qRT-PCR analysis. The heteroyohimbine content of silenced leaves was determined by LC-MS. Leaves powder was weighed (10- 20 mg), extracted with methanol (2 ml) and vortexed for 1 min. After a 10-min centrifugation step at 17,000*g*, an aliquot of the supernatant (20 µl) was diluted to 200 µl with methanol, filtered through 0.2 µm PTFE filters and analysed on Waters Xevo TQ-MS. The chromatographic separation and MS measurements were carried out as described above (method 2).

Gene silencing was confirmed by qRT-PCR. qRT-PCR was also used to check the expression of the other heteroyohimbine synthase genes to ensure that no cross-silencing occurred. RNA extraction was performed using the RNeasy Plant Mini Kit (Qiagen). RNA (1 µg) was used to synthesize cDNA in 20-µl reactions using the iScript cDNA Synthesis Kit (Bio-Rad). The cDNA served as template for quantitative PCR performed using the CFX96 Real Time PCR Detection System (Bio-Rad) using the SSO Advanced SYBR Green Supermix (Bio-Rad). Each reaction was performed in a total reaction volume of 20 µl containing an equal amount of cDNA, 0.25 mM forward and reverse primers, and 1x SsoAdvanced SYBRGreen Supermix (Bio-Rad). The reaction was initiated by a denaturation step at 95°C for 10 min followed by 41 cycles at 95°C for 15 s and 60°C for 1 min. Melting

curves were used to determine the specificity of the amplifications. Relative quantification of gene expression was calculated according to the delta-delta cycle threshold method using the 40S ribosomal protein S9 (RPS9). All primer pairs (Supp. Table 7) efficiencies were between 98% and 108%, and the individual efficiency values were considered in the calculation of normalized relative expression, which was performed using the Gene Study feature of CFX Manager Software. All biological samples were measured in technical duplicates.

pH effect on product profile. Strictosidine was deglycosylated using purified SGD for 25 minutes at room temperature using assay conditions as described above. Strictosidine aglycone was then incubated at a final concentration of 300 µM at pH 5, 6, 7 and 8 in a buffer mix to avoid buffer ingredient effect on activity ((50mM Phosphate buffer, 50mM citric acid, 50mM HEPES). Caffeine (50 µM) was used as an internal standard.

At time zero the enzyme, either THAS1 or HYS (1 µM final concentration), premixed with NADPH (500 µM) was added to the substrate solution. In parallel, a chemical reducing agent, $NaBH_4$ (3 mM final concentration), was added to deglycosylated strictosidine as a control reaction. All reactions were carried out in triplicate. An end-point sample (10 µl) was taken for each assay and prepared for UPLC-MS by addition of 10 µl of 100% MeOH to stop the reaction, and then diluted 1 in 5 with H_2O, and centrifuged for 10 minutes at 4000 rpm. UPLC-MS and data collection were performed as described above for heteroyohimbine separation and quantification.

CD spectra and analysis. Far ultraviolet (UV) CD spectra of the wild-type enzymes THAS1 and HYS, as well as the loop mutants of THAS1 and HYS were recorded on a Chirascan Plus spectropolarimeter (Applied Photophysics) at 20°C in 10 mM potassium phosphate buffer pH 7.0. Samples were analysed from 180 nm to 260 nm using a 0.5 nm step at a speed of 1 s per step. Four replicate measurements were performed on each sample and baseline correction was applied to all data. Spectra are presented as the CD absorption coefficient calculated on a mean residue ellipticity (MRE) basis.

Melting curves of HYS and the HYS loop2 swap mutant were also acquired by CD. The samples were subjected to temperature ramping at the rate of 1°C min^{-1} from 20 °C to 90 °C. Data collection was done from 260 nm to 201 nm using a 1 nm step and 0.75 s time per point. Data were analysed using the Global 3 software. HYS melting point was measured as 61.0 ± 0.1 °C; enthalpy 351.5 ± 3.6 KJ/mol. HYS loop2 swap melting point was measured at 62.0 ± 0.1 °C; enthalpy 535.8 ± 4.5 KJ/mol.

Protein sequence alignments and phylogenetic tree. Protein sequence alignment was generated using ClustalW algorithm with Geneious v.8 (http://www.geneious.com).[62,63] The alignment was edited manually using Seaview V4[64] and secondary structure depiction was added using ESPript V3 (http://espript.ibcp.fr).[65] Phylogenetic analysis was performed using the Neighbor-Joining[66] algorithm and Bootstrap analysis with 1000 replicates.

Docking of cathenamine in THAS1 $NADP^+$ structure. Cathenamine was docked into the THAS1-$NADP^+$ crystal structure using Autodock 4.2.[67] The ligand (cathenamine) was prepared with 2 torsions at the C20, the rest of the molecule being rigid. The search space was defined by a 40x40x40 box, centred between the nicotinamide ring and the "catalytic" zinc in chain B. The genetic algorithm was used with the programme defaults, and the output was a Lamarckian Genetic Algorithm format. The lowest energy docking, as selected by the software, is used in the structures illustrated here.

Figure Legends

Figure 1. Heteroyohimbine biosynthesis. Heteroyohimbines with 3(*S*) stereochemistry derive from strictosidine aglycone. The three diastereomers found in *Catharanthus roseus*, are highlighted with red arrows. Alkaloids derived from heteroyohimbines are also shown.

Figure 2. LCMS analysis of product profiles of active MDR candidates against strictosidine aglycone. See **Fig. S1** for chromatograms of assays with inactive enzymes and negative controls.

Figure 3. Protein sequence alignment of *Catharanthus roseus* THAS1, THAS2, HYS and *Populus tremuloides* Sinapyl Alcohol Dehydrogenase (SAD). Numbering corresponds to HYS. Identical and similar amino acids are highlighted in red and yellow, respectively. Secondary structure elements of the HYS apo crystal structure are displayed. THAS1 and HYS active site amino acids (Y56/53 and E59/56) are indicated by blue dots, and THAS2 active site amino acids (Y120 and D49) are indicated by green dots. Ligands for catalytic and structural zinc ions are highlighted by black and gray dots, respectively. The nuclear localization signal of (THAS1 and HYS) and loops 1 and 2 are indicated in red. A non-proline *cis*-peptide bond that is observed in THAS1 holo, in one subunit of THAS1 apo (**Supplementary Fig. 5**), in HYS apo, and not at all in THAS2 is indicated with an orange dot. The substrate binding domain and the cofactor binding domain are indicated by blue and purple bars, respectively.

Figure 4. Crystal structures of heteroyohimbine synthases THAS1, THAS2 and HYS. A. Sample of automatically derived experimentally phased electron density from THAS1 (at 1.12 Å resolution) superimposed on the final model showing the active site region with the $NADP^+$ cofactor (green carbons) together with neighboring residues (magnolia carbons) and water molecules (small red spheres). B. THAS1 docked with cathenamine (pale blue carbons) with the protein shown in both cartoon (left) and space filling (right) modes. The $NADP^+$ cofactor is shown with green carbons; loop 1 is in orange and loop 2 is in cyan. Zinc ions are displayed as magenta spheres. The active site is largely contained within a single subunit (magnolia surface), although the mouth of the channel leading to the active site is partially bounded by the second subunit of the biological dimer (gray surface) C. Superposition of the apo structures of THAS1 (gold), THAS2 (pink) and HYS (blue). D.

Superposition of the holo ($NADP^+$ containing) structures of THAS1 (gold) and THAS2 (pink), with the cofactor of THAS1 shown as van der Waals spheres for emphasis. For C and D the structures were superposed onto the THAS1 structure, based on the upper subunit alone; only part of the lower subunit of the THAS1 structure is shown in gray for reference (see **Supplementary Fig. 3** for images of the full THAS1 dimer). The insets emphasise the differing lengths of loop 2 between the various structures; the central portion of loop 2 in apo THAS2 was disordered.

Figure 5. Mechanistic hypothesis for heteroyohimbine synthases. A. Proposed mechanism for formation of the tetrahydroalstonine (*S* C20) diastereomer. B. Proposed mechanism of formation of the ajmalicine (*R* C20) diastereomer that is observed in HYS, which contains a histidine residue near the active site.

Figure 6. Deuterium labeling of THAS1 and HYS products using pro-*R* NADPD. Comparison of selected regions of ^1H NMR spectra of labeled A. tetrahydroalstonine, B. ajmalicine, C. mayumbine. The spectra indicate that C21 is labelled with deuterium in the pro-*R* position.

Figure 7. Product profiles of THAS1 and HYS loop swap mutants. A. Shown is the apo THAS1 structure (magnolia) with loops 1 and 2 highlighted in orange and cyan, respectively. For clarity, only the corresponding loops of the HYS apo structure are shown in yellow after superposition. Similarly, only the cofactor from the superposed holo THAS1 structure is shown for reference (green carbons). The side chains of important residues are also shown. B. LC-MS chromatograms of assays with THAS1 mutants in which loop 1, loop 2 or both have been swapped with the corresponding sequences from HYS. C. LC-MS chromatograms of assays with HYS mutants in which loop 1, loop 2 or both have been swapped with the corresponding sequences from THAS1.

Figure 8. THAS2 displays nucleocytosolic localization while HYS is preferentially targeted to the nucleus. *C. roseus* cells were transiently cotransformed with plasmids expressing either THAS2-YFP (A), HYS-YFP (E) or YFP (I) and the plasmid encoding the nuclear CFP marker(B, F, J). Colocalization of the fluorescence signals appears in yellow when merging the two individual (green/red) false color images (C, F, K). Cell morphology is observed with differential interference contrast (DIC) (D, H, L). Bars, 10 µm.

Figure 9. THAS2 and HYS interact with SGD. THAS2/SGD (A, I, Q) and HYS/SGD (C, K, S) interactions were analyzed by BiFC in *C. roseus* cells transiently transformed by distinct combinations of plasmids encoding fusions with the two split YFP fragments, as indicated on each fluorescence picture. THAS1/SGD (E, M, U) and 16OMT/SGD (G, O, W) interactions were studied to evaluate the specificity of THAS2/SGD and HYS/ SGD interactions. Single BiFC assays showing interactions with SGD (upper row) and double BiFC assays highlighting both interactions with SGD and THAS2, HYS, THAS1, 16OMT self-interactions were conducted and observed 16h (middle row) and 36h (lower row) post-transformation. Cell morphology is observed with differential interference contrast (DIC) (B, D, F, H, J, L, N, P, R, T, V, X). Bars, 10 µm.

Acknowledgements

We gratefully acknowledge support from the ERC (311363) and a BBSRC Institute Strategic Programme grant (MET; BB/J004561/1) to S.E.O'C and from the Region Centre (France, ABISAL grant) to V. C. A.S. is supported by a BBSRC DTP studentship. The Diamond Light Source provided access to beamlines I03, I04 and I04-1 (proposal MX9475).

Author contributions

A.S. and E.T. and S.E.O'C. designed the project; A.S., E.T. and L.C. performed molecular cloning/enzyme assays; L.C., A.S., C.E.M.S. and D.M.L. assisted with crystallization, X-ray data acquisition and structure refinement; E.T. and L.C. performed VIGS; E.T. performed

NMR structural characterization; E.M. and V.C. performed all localization experiments; S.E.O'C. supervised the work; A.S. and E. T. and S.E.O'C. wrote the manuscript with input from all authors.

Accession codes

The atomic coordinates and structure factors of the five X-ray structures described in this manuscript have been deposited in the Protein Data Bank (http://www.pdb.org/), with accession codes 5FI3, 5FI5, 5H81, 5H82 and 5H83. GeneBank deposition numbers for all proteins assayed are listed in Supplementary Table 1.

References

1 Shamma, M. & Richey, J. M. The stereochemistry of the heteroyohimbine alkaloids. *J. Am. Chem. Soc.* **85**, 2507-2512 (1963).

2 Allain, H. & Bentue-Ferrer, D. Clinical efficacy of almitrine-raubasine. An overview. *Eur. Neurology* **39**, 39-44 (1998).

3 Benzi, G. Pharmacological features of an almitrine-raubasine combination. Activity at cerebral levels. *Eur. Neurology* **39**, 31-38 (1998).

4 Li, S. et al. Assessment of the therapeutic activity of a combination of almitrine and raubasine on functional rehabilitation following ischaemic stroke. *Curr. Med. Res. Opin.* **20**, 409-415 (2004).

5 Roquebert, J. & Demichel, P. Inhibition of the a1- and a2-adrenoceptor-mediated pressor response in pithed rats by raubasine, tetrahydroalstonine and akuammigine. *Eur. J. Pharmacology* **106**, 203-205 (1984).

6 Ai, J., Dekermendjian, K., Nielsen, M. & Witt, M. R. The heteroyohimbine mayumbine binds with high affinity to rat brain benzodiazepine receptors in vitro. *Nat. Prod. Lett.* **11**, 73-76 (1997).

7 Dassonneville, L. *et al.* Stimulation of topoisomerase II-mediated DNA cleavage by three DNA-intercalating plant alkaloids: Cryptolepine, Matadine, and Serpentine. *Biochemistry* **38**, 7719-7726 (1999).

8 Costa-Campos, L., Iwu, M. & Elisabetsky, E. Lack of pro-convulsant activity of the antipsychotic alkaloid alstonine. *J. Ethnopharmacology* **93**, 3017-3310 (2004).

9 Elisabetsky, E. & Costa-Campos, L. The alkaloid alstonine: A review of Its pharmacological properties. *eCAM* **3**, 39-48 (2006).

10 Herrmann, A. P. *et al.* Effects of the putative antipsychotic alstonine on glutamate uptake in acute hippocampal slices. *Neurochem. Int.* **61**, 1144-1150 (2012).

11 Linck, V. M. *et al.* Alstonine as an antipsychotic: Effects on brain amines and metabolic changes. *eCAM*, 418597 (2011).

12 Linck, V. M. *et al.* 5-HT2A/C receptors mediate the antipsychotic-like effects of alstonine. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **36**, 29-33 (2012).

13 Linck, V. M. *et al.* Original mechanisms of antipsychotic action by the indole alkaloid alstonine (Picralima nitida). *Phytomedicine* **22**, 52-55 (2015).

14 Saxton, J. E. *The chemistry of heterocyclic compounds, indoles: The monoterpenoid indole alkaloids,* (Wiley, 2009).

15 Amer, M. A. & Court, W. E. P. Alkaloids of Rauwolfia nitida root bark. *Phytochemistry* **20**, 2569-2573 (1981).

16 Hochstein, F. A. Alkaloids of Rauwolfia sellowii. *J. Am. Chem. Soc.* **77**, 5744-5745 (1955).

17 Melchio, J., Bouquet, A., Pais, M. & Goutarel, R. Alcaloides indoliques CVI (1) Identité de la mayumbine et de l'épi-19 ajmalicine. L'iso-3 rauniticine, un nouvel alcaloïde extrait du Corynanthe mayumbensis (R. Good) N. Hallé. *Tetrahedron Lett.* **18**, 315-316 (1977).

18 Phillipson, J. D. & Supavita, N. Alkaloids of uncaria elliptica. *Phytochemistry* **22**, 1809-1813 (1983).

19 Ponglux, D., Supavita, T., Verpoorte, R. & Phillipson, D. P. Alkaloids of Uncaria attenuata from Thailand. *Phytochemistry* **19**, 2013-2016 (1980).

20 Robinson, R. & Thomas, A. F. The alkaloids of picralima nitida, Stapf, Th. and H. Durand. Part I. The structure of alkuammigine. *J. Chem. Soc.*, 3479-3482 (1954).

21 Stoeckigt, J., Husson, H. P., Kan-Fan, C. & Zenk, M. H. Cathenamine, a central intermediate in the cell free biosynthesis of ajmalicine and related indole alkaloids. *J. Chem. Soc., Chem. Commun.*, 164-166 (1977).

22 O'Connor, S. E. & Maresh, J. J. Chemistry and biology of monoterpene indole alkaloid biosynthesis. *Nat. Prod. Rep.* **23**, 532-547 (2006).

23 Gerasimenko, I., Sheludko, Y., Ma, X. & Stöckigt, J. Heterologous expression of a Rauvolfia cDNA encoding strictosidine glucosidase, a biosynthetic key to over 2000 monoterpenoid indole alkaloids. *Eur. J. Biochem.* **269**, 2204-2213 (2002).

24 Stavrinides, A. *et al.* Unlocking the diversity of alkaloids in Catharanthus roseus: nuclear localization suggests metabolic channeling in secondary metabolism. *Chem. Biol.* **22**, 336-341 (2015).

25 Gongora-Castillo, E. *et al.* Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PloS One* **7**, e52506 (2012).

26 Kellner, F. *et al.* Genome-guided investigation of plant natural product biosynthesis. *Plant J.* **82**, 680-692 (2015).

27 Bomati, E. K. & Noel, J. P. Structural and kinetic basis for substrate selectivity in Populus tremuloides sinapyl alcohol dehydrogenase. *Plant Cell* **17**, 1698-1611 (2005).

28 Pan, H. *et al.* Structural studies of cinnamoyl-CoA reductase and cinnamyl-alcohol dehydrogenase, key enzymes of monolignol biosynthesis. *Plant Cell* **26**, 3709-3727 (2014).

29 Youn, B. *et al.* Crystal structures and catalytic mechanism of the Arabidopsis cinnamyl alcohol dehydrogenases AtCAD5 and AtCAD4. *Org. Biomol. Chem.* **4**, 687-1697 (2006).

30 Auld, D. S. & Bergman, T. Medium- and short-chain dehydrogenase/reductase gene and protein families: The role of zinc for alcohol dehydrogenase structure and function. *Cell. Mol. Life Sci.* **65**, 3961-3970 (2008).

31 Eklund, H. & Ramaswamy, S. Medium- and short-chain dehydrogenase/reductase gene and protein families: Three-dimensional structures of MDR alcohol dehydrogenases. *Cell. Mol. Life Sci.* **65**, 3907-3917 (2008).

32 Stoeckigt, J., Hemscheidt, T., Hoefle, G., Heinstein, P. & Formacek, V. Steric course of hydrogen transfer during enzymatic formation of 3α-heteroyohimbine alkaloids. *Biochemistry* **22**, 3448–3452 (1983).

33 Knowles, J. R. & Jencks, W. P. The intrinsic pka-values of functional groups in enzymes: Improper deductions from the ph-dependence of steady-state parameter. *Crit. Rev. Biochem.* **4**, 165-173 (1976).

34 Guirimand, G. *et al.* Strictosidine activation in Apocynaceae: towards a "nuclear time bomb"? *BMC Plant Biol.* **10**, 182 (2010).

35 Liscombe, D. K. & O'Connor, S. E. A virus-induced gene silencing approach to understanding alkaloid metabolism in Catharanthus roseus. *Phytochemistry* **72**, 1969-1977 (2011).

36 Brown, S., Clastre, M., Courdavault, V. & O'Connor, S. E. De novo production of the plant-derived alkaloid strictosidine in yeast. *Proc. Natl. Acad. Sci. USA* **112**, 3205-3210 (2015).

37 Konno, K., Hirayama, C., Yasui, H. & Nakamura, M. Enzymatic activation of oleuropein: A protein crosslinker used as a chemical defense in the privet tree. *Proc. Natl. Acad. Sci. USA* **96**, 9159-9164 (1999).

38 Pankoke, H., Buschmann, T. & Muller, C. Role of plant beta-glucosidases in the dual defense system of iridoid glycosides and their hydrolyzing enzymes in Plantago lanceolata and Plantago major. *Phytochemistry* **94**, 99-107 (2013).

39 Frelin, O. *et al.* A directed-overflow and damage-control N-glycosidase in riboflavin biosynthesis. *Biochem. J.* **466**, 137-145 (2015).

40 Qu, Y. *et al.* Completion of the seven-step pathway from tabersonine to the anticancer drug precursor vindoline and its assembly in yeast. *Proc. Natl. Acad. Sci. USA* **112**, 6224-6229 (2015).

41 Berrow, N. S. *et al.* A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res.* **35**, e45 (2007).

42 Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125-132 (2010).

43 Evans, P. R. & Murshudov, G. N. Scaling and assessment of data quality. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204-1214 (2013).

44 Winter, G. Xia2: an expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186-190 (2010).

45 Winter, G. & McAuley, K. E. Automated data collection for macromolecular crystallography. *Methods* **55**, 81-93 (2011).

46 Sheldrick, G. M. A short history of SHELX. *Acta Crystallogr. Sect. A* **64**, 112-122 (2008).

47 Delageniere, S. *et al.* ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics* **27**, 3186-3192 (2011).

48 Fisher, S. J., Levik, K. E., Williams, M. A., Ashton, A. W. & McAuley, K. E. SynchWeb: a modern interface for ISPyB. *J. Appl. Crystallogr.* **48**, 927-932 (2015).

49 Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. Sect. D* **62**, 1002-1011 (2006).

50 Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486-501 (2010).

51 Winn, M. D., Murshudov, G. N. & Papiz, M. Z. Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods Enzymol.* **374**, 300-321 (2003).

52 McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658-674 (2007).

53 Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845-858 (2015).

54 Painter, J. & Merritt, E. A. TLSMD web server for the generation of multi-group TLS models. *J. Appl. Crystallogr.* **39**, 109-111 (2006).

55 Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12-21 (2010).

56 McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallogr D Biol Crystallogr* **67**, 386-394, doi:10.1107/S0907444911007281 (2011).

57 Sun, J., Baker, A. & Chen, P. Profiling the indole alkaloids in yohimbe bark with ultra-performance liquid chromatography coupled with ion mobility quadrupole time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **25**, 2591-2602 (2011).

58 Guirimand, G. *et al.* Optimization of the transient transformation of Catharanthus roseus cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. *Plant Cell Rep.* **28**, 1215-1234 (2009).

59 Guirimand, G. *et al.* Spatial organization of the vindoline biosynthetic pathway in Catharanthus roseus. *J. Plant Physiology* **168**, 549-557 (2011).

60 Waadt, R. *et al.* Multicolor bimolecular fluorescence complementation reveals simultaneous formation of alternative CBL/CIPK complexes in planta. *Plant J.* **56**, 505-516 (2008).

61 Geu-Flores, F. *et al.* An alternative route to cyclic terpenes by reductive cyclization. *Nature* **492**, 138-142 (2012).

62 Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).

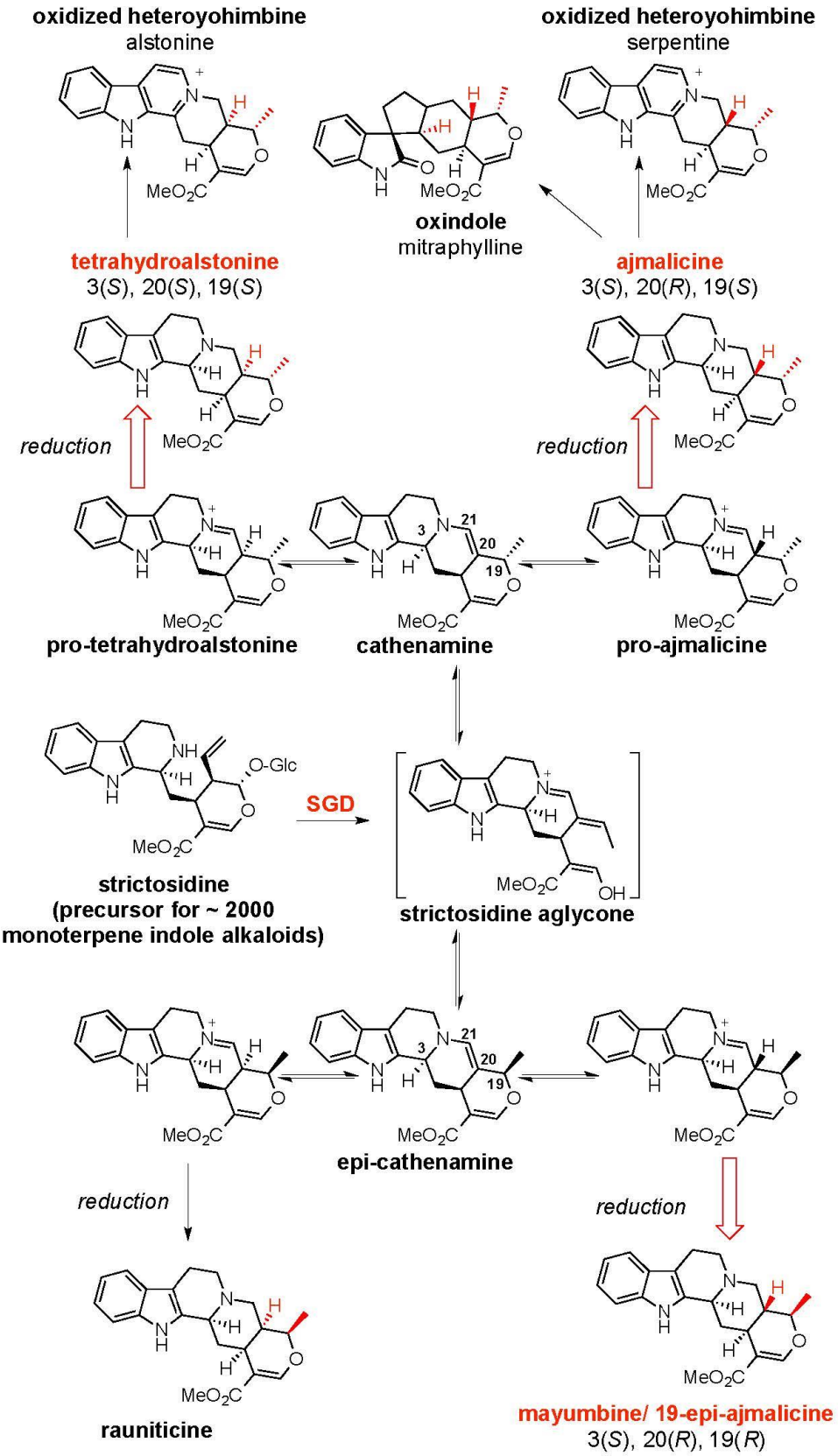
63 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

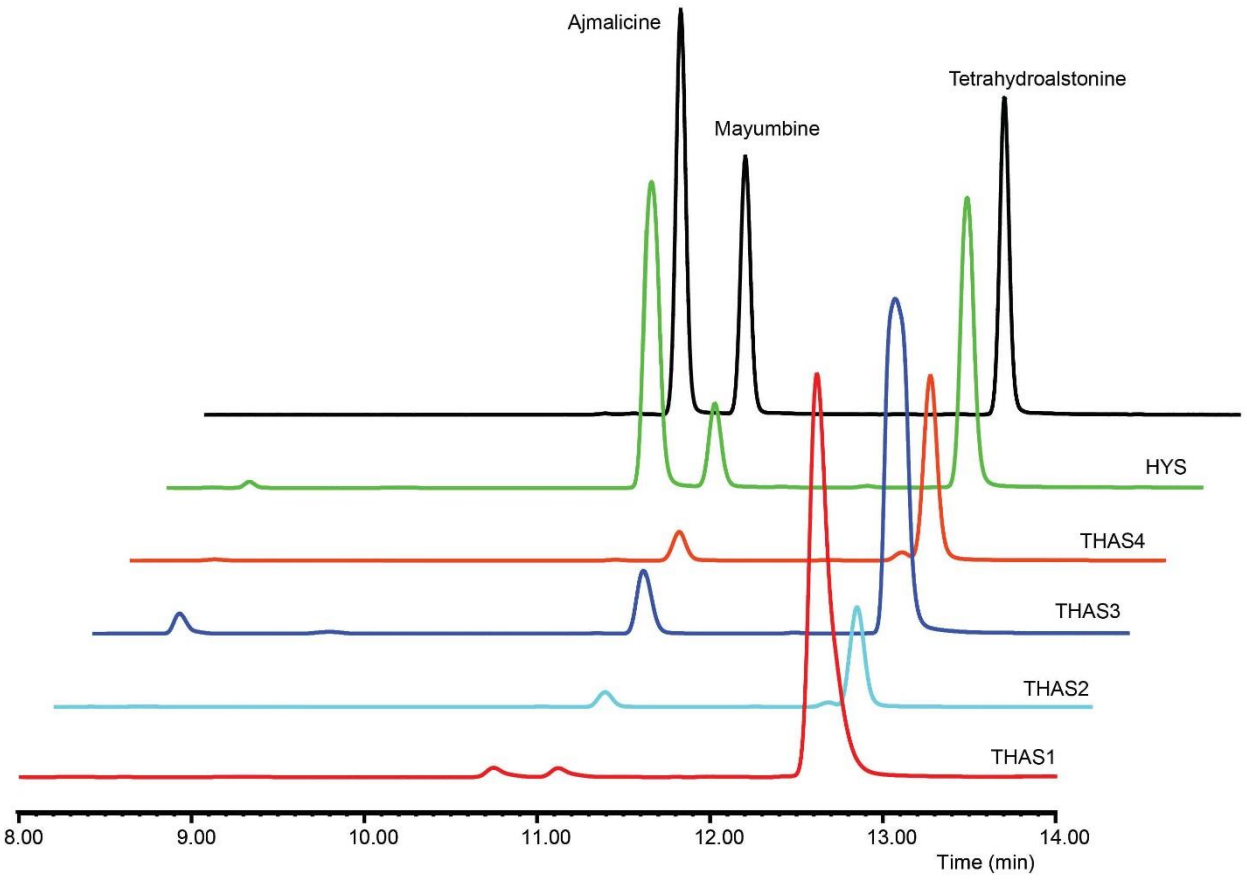
64 Gouy, M., Guindon, S. & Gascuel, O. Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221-224 (2010).

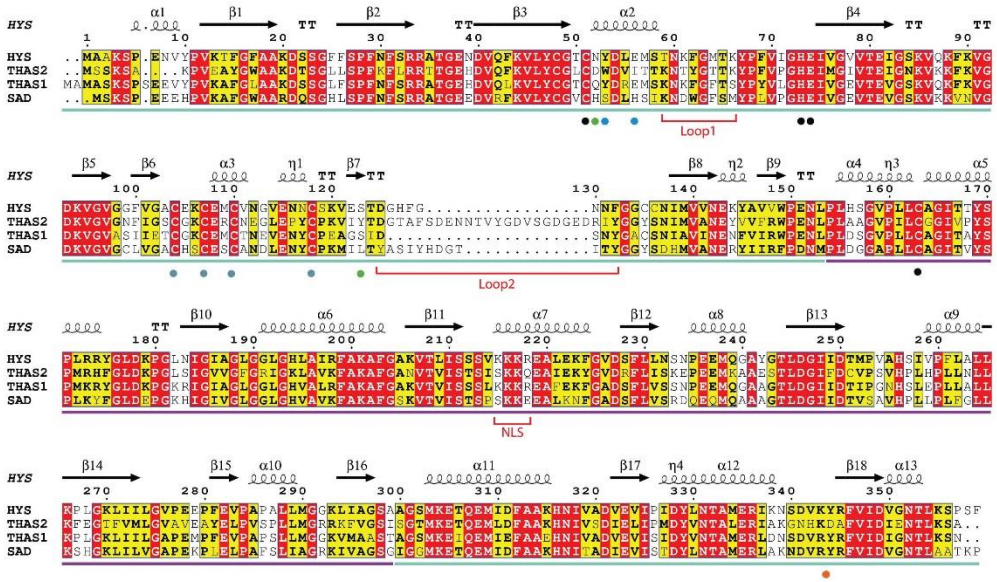
65 Gouet, P., Robert, X. & Courcelle, E. ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* **31**, 3320-3323 (2003).

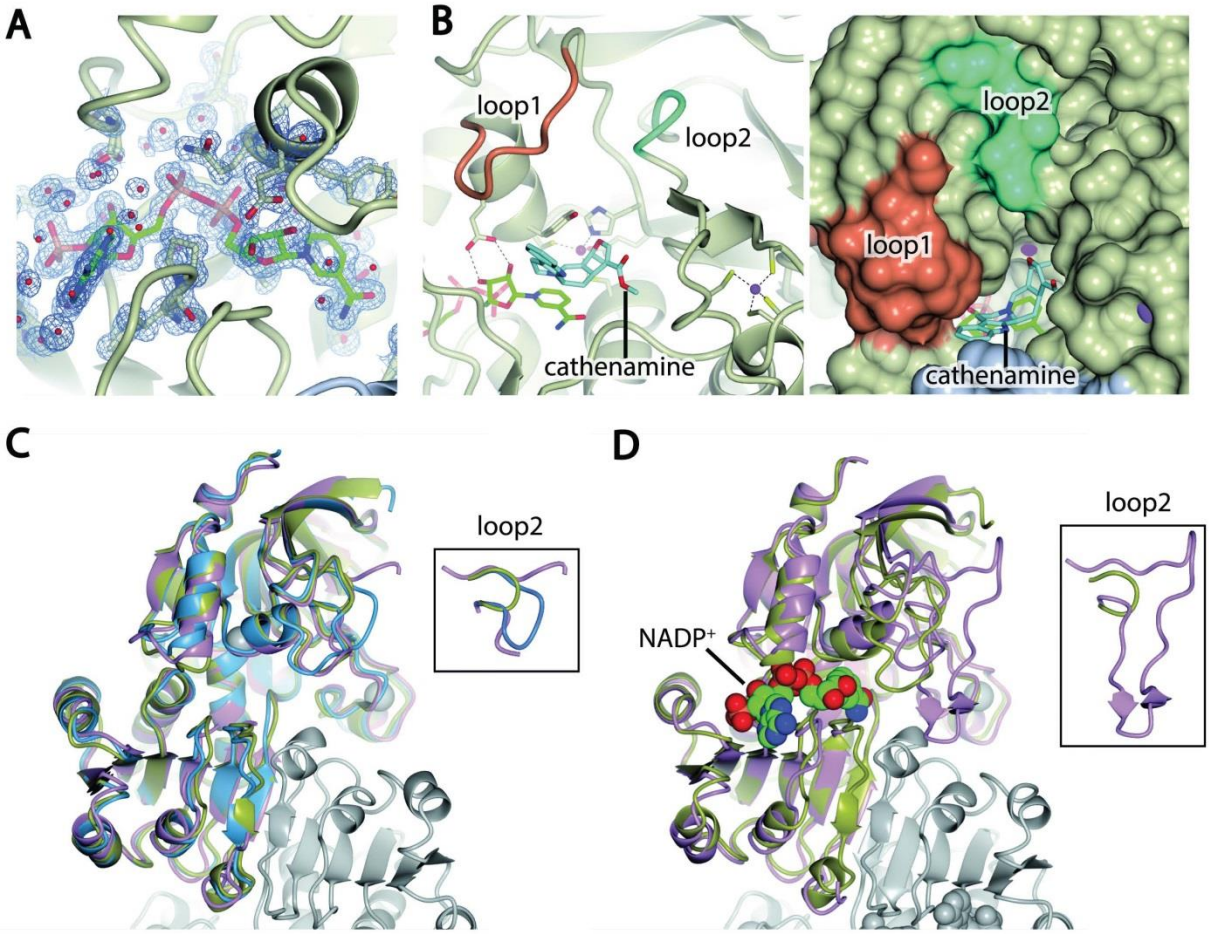
66 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).

67 Morris, G. M. *et al.* Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785-2791 (2009).

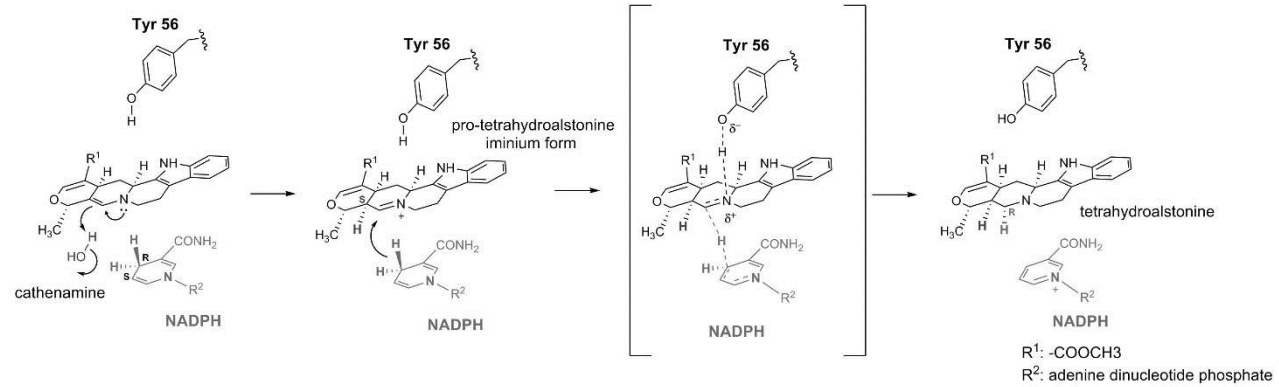

oxidized heteroyohimbine
alstonine
oxidized heteroyohimbine
serpentine
oxindole
mitraphylline
tetrahydroalstonine
3(S), 20(S), 19(S)
ajmalicine
3(S), 20(R), 19(S)
reduction
reduction
pro-tetrahydroalstonine
cathenamine
pro-ajmalicine
SGD
strictosidine
(precursor for ~ 2000
monoterpene indole alkaloids)
strictosidine aglycone
epi-cathenamine
reduction
reduction
rauniticine
mayumbine/ 19-epi-ajmalicine
3(S), 20(R), 19(R)


Ajmalicine
Mayumbine
Tetrahydroalstonine
HYS
THAS4
THAS3
THAS2
THAS1
8.00
9.00
10.00
11.00
12.00
13.00
14.00
Time (min)

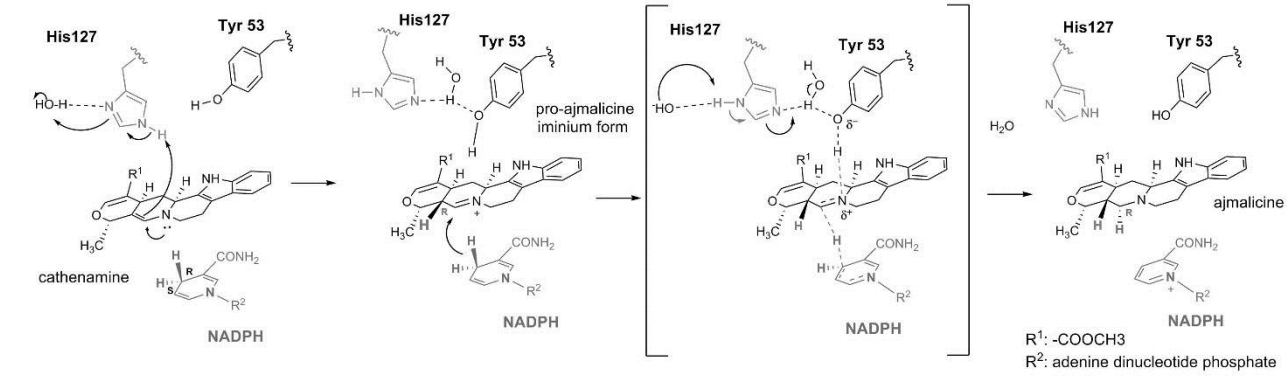

HYS
HYS
THAS2
THAS1
SAD
Loop1
Loop2
NLS

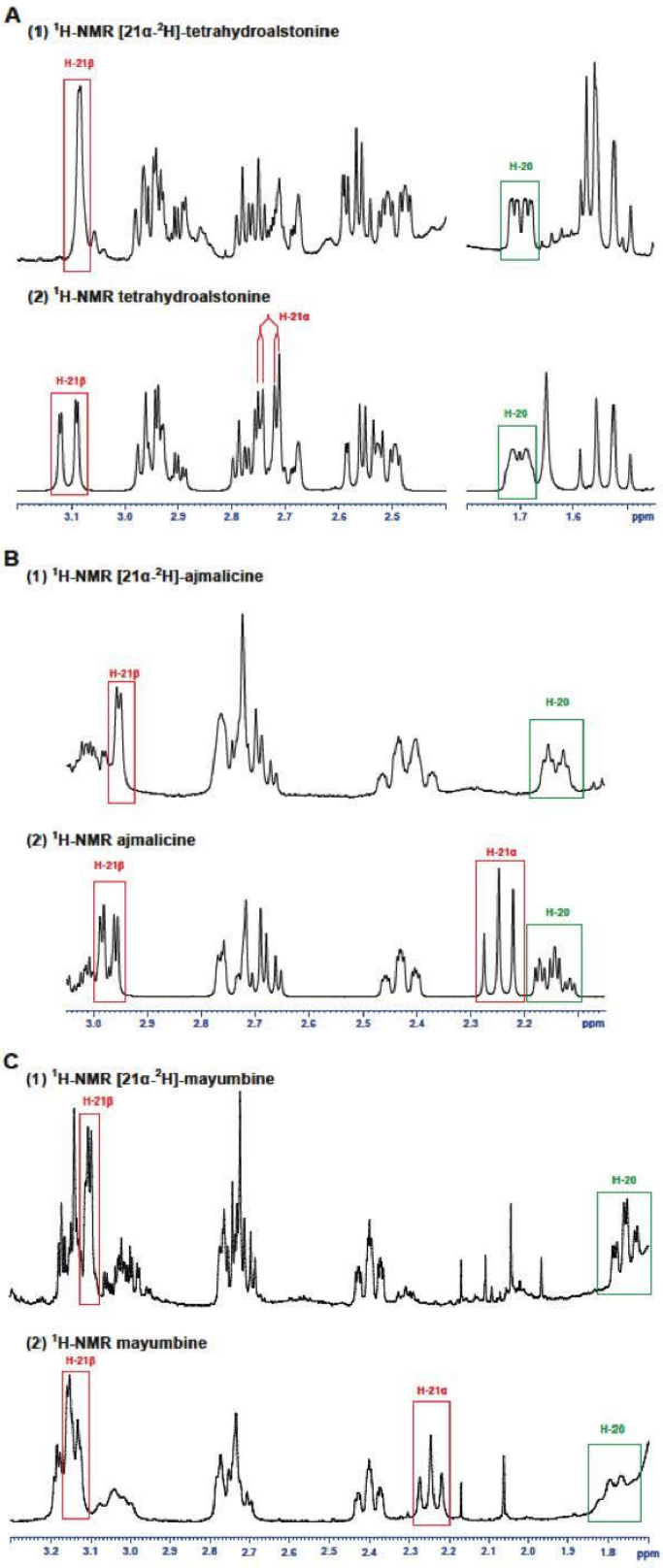

A
B
loop1
loop2
cathenamine
loop2
loop1
cathenamine
C
loop2
D
NADP+
loop2

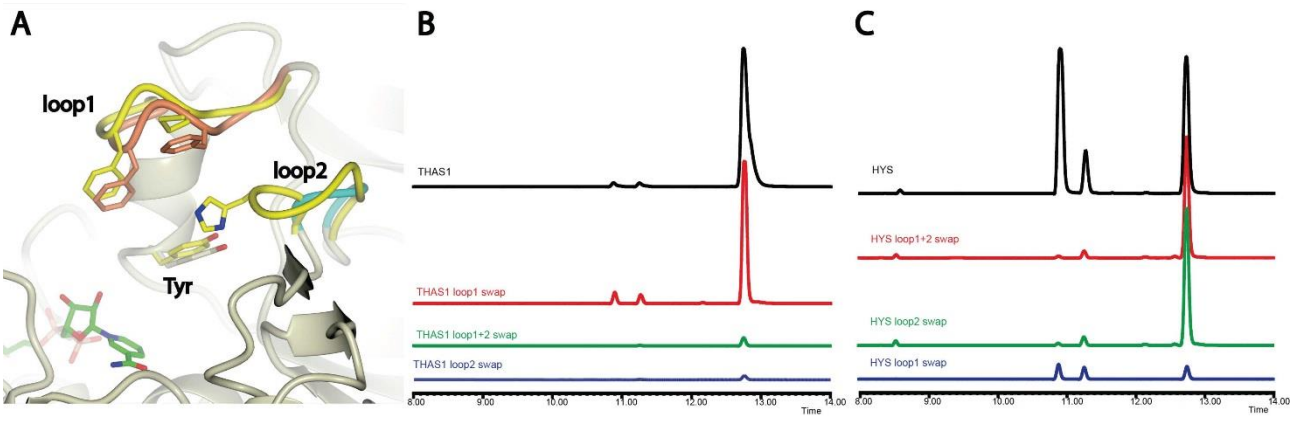
A. Tetrahydroalstonine biosynthesis mechanism in THAS and HYS

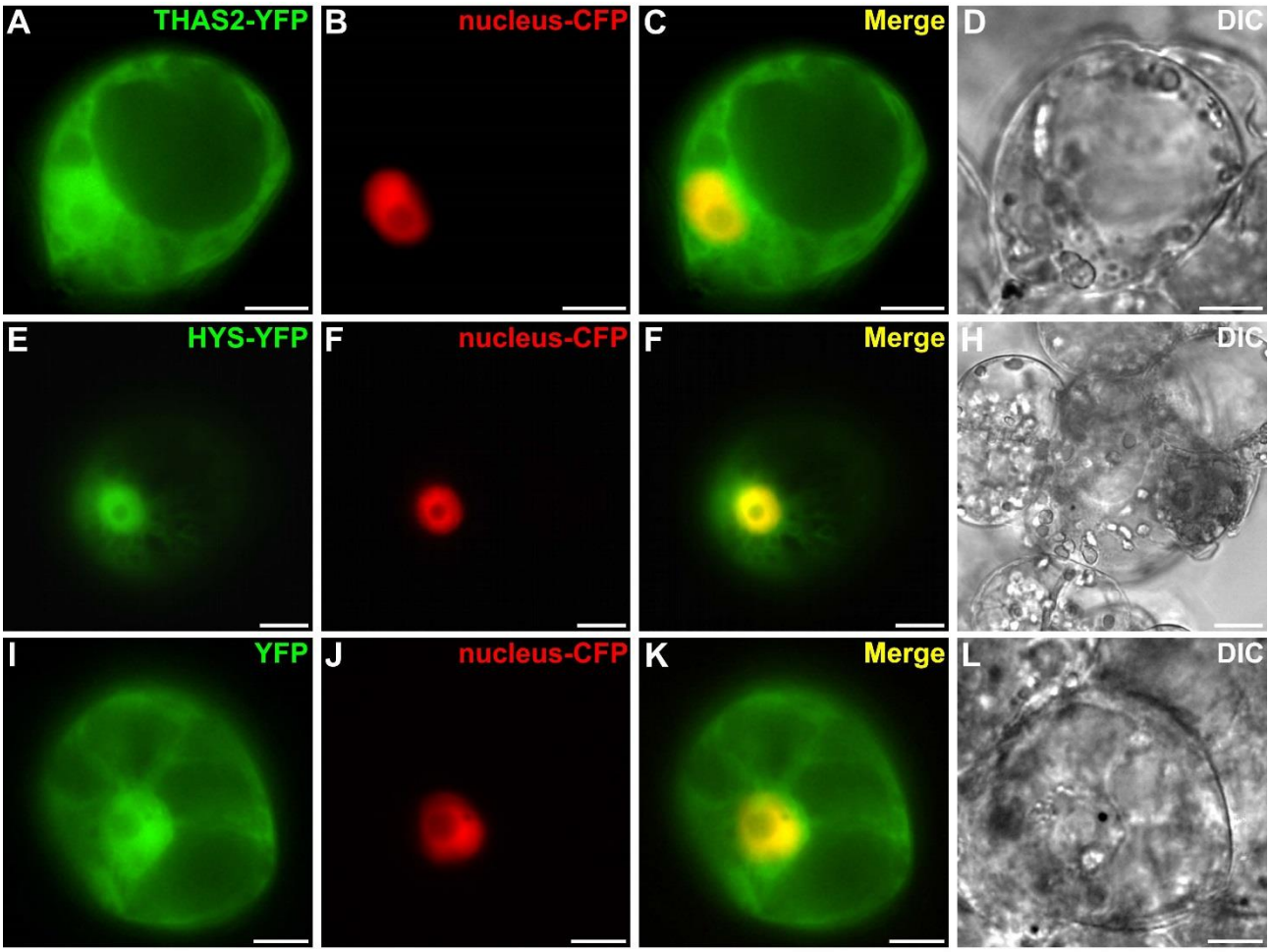


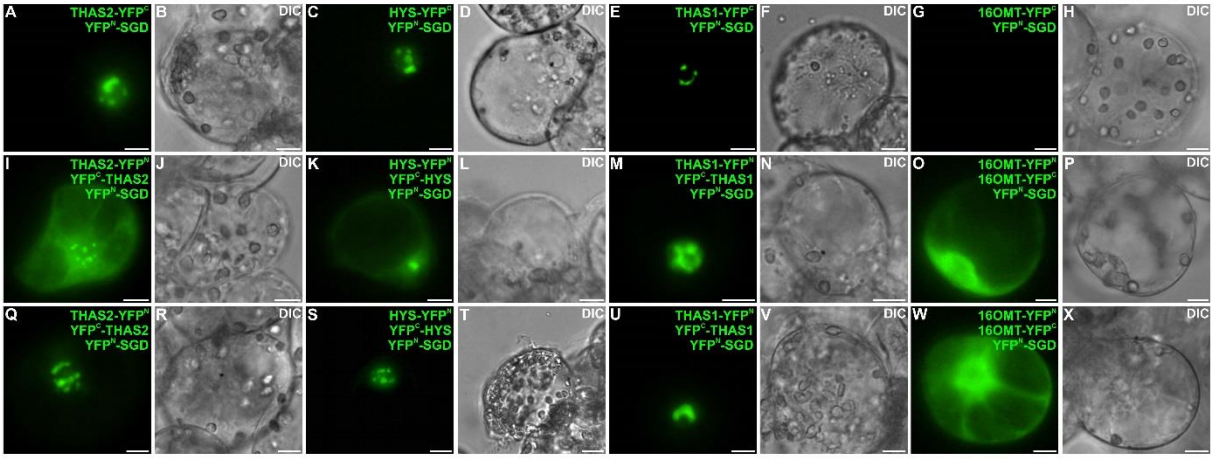
B. Ajmalicine biosynthesis mechanism in HYS




A
(1) ¹H-NMR [21α-²H]-tetrahydroalstonine
H-21β
H-20
(2) ¹H-NMR tetrahydroalstonine
H-21α
H-21β
H-20
3.1
3.0
2.9
2.8
2.7
2.6
2.5
1.7
1.6
ppm
B
(1) ¹H-NMR [21α-²H]-ajmalicine
H-21β
H-20
(2) ¹H-NMR ajmalicine
H-21α
H-21β
H-20
3.0
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
ppm
C
(1) ¹H-NMR [21α-²H]-mayumbine
H-21β
H-20
(2) ¹H-NMR mayumbine
H-21β
H-21α
H-20
3.2
3.1
3.0
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
ppm


A
loop1
loop2
Tyr
B
THAS1
THAS1 loop1 swap
THAS1 loop1+2 swap
THAS1 loop2 swap
8.00
9.00
10.00
11.00
12.00
13.00
14.00
Time
C
HYS
HYS loop1+2 swap
HYS loop2 swap
HYS loop1 swap
8.00
9.00
10.00
11.00
12.00
13.00
14.00
Time


A
THAS2-YFP
B
nucleus-CFP
C
Merge
D
DIC
E
HYS-YFP
F
nucleus-CFP
F
Merge
H
DIC
I
YFP
J
nucleus-CFP
K
Merge
L
DIC


A THAS2-YFP^{C} YFP^{N}-SGD
B DIC
C HYS-YFP^{C} YFP^{N}-SGD
D DIC
E THAS1-YFP^{C} YFP^{N}-SGD
F DIC
G 16OMT-YFP^{C} YFP^{N}-SGD
H DIC
I THAS2-YFP^{N} YFP^{C}-THAS2 YFP^{N}-SGD
J DIC
K HYS-YFP^{N} YFP^{C}-HYS YFP^{N}-SGD
L DIC
M THAS1-YFP^{N} YFP^{C}-THAS1 YFP^{N}-SGD
N DIC
O 16OMT-YFP^{N} 16OMT-YFP^{C} YFP^{N}-SGD
P DIC
Q THAS2-YFP^{N} YFP^{C}-THAS2 YFP^{N}-SGD
R DIC
S HYS-YFP^{N} YFP^{C}-HYS YFP^{N}-SGD
T DIC
U THAS1-YFP^{N} YFP^{C}-THAS1 YFP^{N}-SGD
V DIC
W 16OMT-YFP^{N} 16OMT-YFP^{C} YFP^{N}-SGD
X DIC

Partie 4: Ingénierie métabolique de la voie de biosynthèse de la vindoline dans les levures

Article 6: Bioconversion of tabersonine to vindoline in yeast

La pervenche de Madagascar constitue une vaste source de composés valorisables parmi lesquels figurent la vinblastine et la vincristine largement utilisés dans les traitements de chimiothérapie anticancéreuse. Ces alcaloïdes sont produits dans le commerce par la condensation chimique de deux précurseurs extraits des feuilles de *C. roseus* que sont la vindoline et la catharanthine. L'intérêt pharmaceutique de ces alcaloïdes dimères, leur faible abondance, et leur coût de production sont à l'origine de nombreux travaux de recherche visant à améliorer le taux de production de ces molécules. Le marché mondial pour la vincristine et la vinblastine se situe autour de 300 millions de dollars US, avec une valeur au détail se situant autour de 20 000 dollars US le gramme. Comme nous l'avons spécifié précédemment dans ce manuscrit, tous les AIM identifiés à ce jour proviennent d'un précurseur commun, la strictosidine, qui subit une série de conversions enzymatiques menant à la synthèse des différentes familles d'AIM, que sont les *aspidosperma* (vindoline), les *igoba* (catharanthine) et les *corynanthe* (alcaloïdes de type Hétéroyohimbines: tétrahydroalstonine). Bien que la synthèse d'alcaloïdes de type *corynanthe* nécessite assez peu d'étapes enzymatiques impliquant notamment des déshydrogénases/réductases, la production de la catharanthine et de la vindoline met en jeu des voies plus complexes. A l'heure actuelle, à partir de l'aglycone de strictosidine, les étapes enzymatiques impliquées dans la biosynthèse de la catharanthine et dans celle de la tabersonine ne sont pas caractérisées (figure 12).

Toutefois, à partir de la tabersonine, les sept étapes enzymatiques menant à la vindoline, sont toutes élucidées à ce jour. Au début de cette thèse, il restait deux étapes à identifier menant de la 16-méthoxytabersonine à la 16-méthoxy-2,3-dihydro-3-hydroxytabersonine (figure 12). Nous avons participé, notamment sur le plan de la localisation subcellulaire, à la caractérisation du cytochrome P450 T3O catalysant la première étape (Kellner et al., 2015). La seconde étape a également été élucidée au cours de ce travail de thèse et est catalysée par une déshydrogénase nommée T3R. Cette enzyme a aussi été caractérisée par un autre groupe en 2015 et a fait l'objet d'une publication (Qu et al., 2015) dans laquelle les auteurs ont transféré la totalité de la voie de la tabersonine à la vindoline dans la levure. Ils ont montré qu'en présence de tabersonine, la levure est capable de

synthétiser de la vindoline mais aussi de la vindorosine (composé analogue à la vindoline mais dépourvu de son groupement méthoxyle -OCH3).

Aussi, l'article qui suit, présente dans un premier temps l'identification de l'enzyme T3R. Dans un second temps, il expose les stratégies pour favoriser et optimiser la production de vindoline à partir de tabersonine dans les levures transformées.

Bioconversion of tabersonine to vindoline in yeast

Emilien Foureau*, Stephanie Brown*, Thomas Dugé de Bernonville*, Franziska Kellner*, Marc Clastre, Arnaud Lanoue, Luisa, Céline Melin, Emeline Marais, Audrey Oudin, Nicolas Papon, Sarah E. O'Connor#, Vincent Courdavault#

[1]Université François-Rabelais de Tours, EA2106 "Biomolécules et Biotechnologies Végétales", Tours, France.
[2]Department of Biological Chemistry, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, United Kingdom.
[3]Universidad de Antioquia, Laboratorio de Biotecnología, Sede de Investigación Universitaria, Colombia.

* These authors contribute equally to this work
# Corresponding authors:
Vincent Courdavault (Université François-Rabelais de Tours - EA2106 "Biomolécules et Biotechnologies Végétales", UFR Sciences et Techniques, 37200, Tours, France. Phone: +33 247 36 70 23; Fax: +33 247 27 66 60. e-mail: vincent.courdavault@univ-tours.fr)
Sarah E. O'Connor (Department of Biological Chemistry, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, United Kingdom. Phone: +44 (0)1603 450334. e-mail: Sarah.O'Connor@jic.ac.uk)

Key-words: Vindoline, Tabersonine, Catharanthus roseus, bioconversion, 16-methoxy-2,3-dihydro-3-hydroxytabersonine.

Communications to the Editor should not exceed 8 double-spaced pages of text (not including references) and should contain no more than 20 references and 4 figures and/or tables.

Communications to the Editor should not be divided into sections except for Materials and Methods (including Computational Methods). A short Abstract (preferably less than 200 words), a brief introduction (not divided into its own section), Acknowledgments (optional), and References should be included. Communications to the Editor are subjected to the same review process as Articles, and they should not constitute preliminary investigations.

Introduction

Madagascar periwinkle (*Catharanthus roseus*) constitutes the unique source of highly valuable compounds used in chemotherapy treatments including vinblastine, vincristine and their derivatives vinflunine and vinorelbine, which all belong to the monoterpene indole alkaloid (MIA) class of specialized metabolites. *In planta*, these MIAs are the product of a complex biosynthetic pathway containing at least 30 enzymatic steps whose characterization is still in progress (Courdavault et al. 2014). All MIAs originate from strictosidine, the common MIA biosynthetic intermediate, which undergoes a series of enzymatic conversions leading to the synthesis of the different MIA subfamilies such as *Aspidosperma* (vindoline), *Igoba* (catharanthine) or *Corynanthe* (tetrahydroalstonine) (Figure 1). While the synthesis of *Corynanthe* alkaloids such as tetrahydroalstonine only requires two enzymatic steps following strictosidine formation (Stravinides et al., 2015), production of catharanthine and vindoline involve more complex pathways that have not yet been fully elucidated. To date, vinblastine and related compounds are commercially produced through the chemical condensation of vindoline and catharanthine, both of which are extracted from leaves of *C. roseus*. Therefore, the elucidation of the biosynthetic pathways of these compounds and their transfer into heterologous organisms based on the current synthetic biology approaches could have a profound impact on the supply of vinblastine, vincristine and other related anti-cancer compounds.

While the formation of catharanthine remains enigmatic, significant progress on the elucidation of vindoline biosynthesis has been made over the last several years. *In folio*, vindoline is synthesized via a seven-step conversion of tabersonine, a major downstream derivative of strictosidine (**Figure 1**). Tabersonine is successively hydroxylated by tabersonine 16-hydroxylase (T16H2, Besseau et al., 2013) and methylated by 16-

hydroxytabersonine-16-*O*-methyltransferase (16OMT, Levac et al., 2008) to form 16-methoxytabersonine. This latter is then subjected to a formal hydration first performed by 16-methoxytabersonine 3-oxygenase (16T3O, Kellner et al., 2015) catalyzing an epoxide formation that is subsequently reduced by the recently identified tabersonine 3-reductase, belonging to the medium chain reductase/dehydrogenase family (MDR; Qu et al., 2015). Finally, the resulting 16-methoxy-2,3-dihydro-3-hydroxytabersonine is *N*-methylated by 16-methoxy-2,3-dihydro-3-hydroxytabersonine *N*-methyltransferase (NMT, Liscombe et al., 2010), hydroxylated by desacetoxyvindoline-4-hydroxylase (D4H; Vazquez-Flota et al., 1997) and *O*-acetylated by deacetylvindoline-4-O-acetyltransferase (DAT; St-Pierre et al., 1998) to yield vindoline, the prevalent MIA accumulated in *C. roseus* leaves. By-passing the first hydroxylation catalyzed by T16H2 leads to the formation of vindorosine (an unmethoxylated form of vindoline) that can also be accumulated in large mounts in *C. roseus* leaf but that cannot be used to synthesize vinblastine or vincristine (Magnota et al., 2006; Besseau et al., 2013).

The recent reconstruction of the strictosidine biosynthetic pathway in *Saccharomyces cerevisiae* via the transfer of no less than 21 genes, opens new perspectives towards the engineering of the MIA production in heterologous organisms (Brown et al., 2015). An analogous approach has been also described to convert tabersonine to vindoline by transferring the seven genes of the vindoline biosynthetic pathway in yeast (Qu et al., 2015). Such strategy seems promising to produce large quantities of vindoline since tabersonine can readily be obtained in huge amounts from seeds of *Voacanga africana* for instance (Koroch et al., 2009). However, it requires the elaboration of high titer vindoline producing strains with a high vindoline/vindorosine ratio to ensure an efficient sourcing of vinblastine precursor. This last point remains critical and constitutes one of the main pitfalls of the previously elaborated yeast strain that accumulated more vindorosine that vindoline (Qu et al., 2015). In this work,

we describe the transcriptomics analysis leading to the identification of T3R and we compare the production of vindoline/vindorosine in reconstituted yeast strains expressing the vindoline biosynthetic genes integrated in the yeast genome or from auto-replicative plasmids leading to optimized vindoline synthesis.

When we initiated this work, the characterization of T3R had not been reported, prompting us to undergo the identification of candidate genes for this missing enzyme of the vindoline biosynthetic pathway. Since genes of plant metabolic pathways frequently display similar tissue-specific expression patterns, we clustered all transcripts from *C. roseus* transcriptomes according to their expression profiles. The abundance of each sequence was then estimated in 39 samples generated in diverse experimental conditions, and an optimal value of 21 gene clusters was retained for a model-based clustering initiated by a *k*-means approach (Dugé de Bernonville et al., 2015). Within each cluster, transcripts with homologies to reference MIA genes were identified. Remarkably, except for 16OMT transcripts suffering from a lower reconstruction quality, T16H2, 16T3O, NMT, D4H and DAT transcripts were regrouped in the same gene cluster (cluster 18, **Figure 2A**). Genes from this cluster (1175 sequences) were marked by a higher expression in leaves and seedlings in accordance with vindoline accumulation. To identify the missing enzyme between 16T3O and NMT, we next analyzed the 350 transcripts of cluster 18 displaying the best correlations (Pearson correlation coefficient) with T16H2, 16T3O, NMT D4H and DAT. Plotting intersections of correlated genes with a Venn diagram revealed that 7 transcripts have a high correlation coefficient with all the 5 vindoline biosynthetic genes (**Figure 2B**). Among these transcripts, we found a transcript (SRR342017|TR8859|c1_g4_i3|len=1797, ADH13068) with a significant identity with dehydrogenases that were credible candidates for the uncharacterized reduction following epoxidation of 16-methoxytabersonine (**Supplemental Figure 1**). To test the activity of the corresponding protein, we conducted a functional assay in yeast

(*Saccharomyces cerevisiae*) by the successive co-expression of T16H2, 16OMT, 16T3O, ADH13068, NMT, D4H and DAT using auto-replicative plasmids. The yeast strains harbouring each combination of enzymes were fed with tabersonine (100 µM) and cultivated for 6h to 48h before metabolic analyses. We first noted that tabersonine was easily internalized by yeast cultures allowing its progressive conversion into the expected compounds (regarding masses and retention times). This is illustrated with the sequential expression of T16H2, T16H2 and 16OMT (**Figure 3**) that leads to the formation of 16-hydroxytabersonine and 16-methoxytabersonine, respectively. Interestingly, in all the tested strains, almost all the enzymatic reaction products were excreted in the culture medium since only traces of these metabolites were detected in the intracellular fraction of the yeast (data not shown). Moreover, efficient substrate conversions (70 to 96%) were observed for the first four tested enzymes (**Supplemental Table I**). Notably, when ADH13068 was expressed in combination with T16H2, 16OMT and 16T3O, the 16T3O product (m/z 383) was converted into a compound with a mass corresponding to 16-methoxy-2,3-dihydro-3-hydroxytabersonine (m/z 385, **Figure 3**), suggesting that a reduction reaction was catalyzed by ADH13068. Furthermore, the specificity of this reaction was confirmed by testing the expression of a dehydrogenase from a distinct gene cluster (ADH7661), which failed to convert the 16T3O product (**Supplemental Figure 2** –Arnaud). No endogenous yeast dehydrogenases were able to catalyze this reaction as demonstrated by the accumulation of the 383 m/z compound in the yeast strain expressing T16H2, 16OMT, 16T3O (**Figure 3**). Moreover, albeit no formal compound identification could be done in absence of disposable standard, we observed that NMT was able to methylate the product of ADH13068 according to the resulting compound mass (m/z 399) similar to desacetoxyvindoline. This methylation occurs at a lower rate than reactions catalyzed by the upstream enzymes probably due to a lower efficiency of NMT in these experimental conditions (**Supplemental Table I**). Finally, as

expected, expression of D4H and D4H, DAT led to the accumulation of deacetylvindoline and vindoline, respectively. In addition, we also noted that 16T3O, ADH13068, NMT, D4H and DAT were able to convert tabersonine into vindorosine (an unmethoxylated form of vindoline) suggesting that ADH13068 was also capable to reduce the epoxide on this tabersonine derivative (**Supplemental Figure 3**). Taken all together, these results provided evidences that ADH13068 corresponds to the missing dehydrogenase of the vindoline biosynthetic pathway that was deposited under Genbank accession number KR063270. Following the publication of the work of Qu and collaborators that characterized T3R within the periwinkle epidermome, we noted that ADH13068 and T3R (KP122966) were similar albeit distinct transcriptomic analytic tools/resources have been used for their identification. Additional analyses revealed that

In order to evaluate the potential of bio-engineered yeasts as an alternative source of vindoline, we next compared the titer and ratio of vindoline/vindorosine synthesis in yeast strains expressing the seven genes of the pathway either from autoreplicative plasmids or from genome integrated constructs. This last type of strain, exhibiting a greater genetic stability, was generated through homologous recombination of the vindoline biosynthetic genes under the control of strong constitutive promoters according to Brown et al. (2015) (**Supplemental Figure 4**). Such strategy usually led to the creation of yeasts with a greater genetic stability. By contrast, yeast strain bearing autoreplicative plasmids are potentially subjected to plasmid losses along growth but possesses an increased number of transgene copies due to plasmid multiplication allowing to expect a higher production titer. Our "autoreplicative yeast strain" (AYS) was almost similar to that described by Qu et al. (2015) since it used inducible plasmids to express the vindoline biosynthetic genes but differed in cytochrome P450 reductases (CPRs). In our initial conditions of production (125 µM tabersonine), we measured that AYS produced around 0.31 mg of vindoline per mg of yeast

dry weight, which is in the same order of magnitude than the production observed by Qu et al. (2015) reaching a 0.6 mg.g dw^{-1} 12 h^{-1} rate with 225 µM of tabersonine as starting substrate. Increasing the tabersonine amount did not result in…

However, in AYS we noted that vindorosine production was around 0.045 mg.g dw^{-1} (7:1 vindoline/vindorosine ratio) which seemed substantially lower than the evaluated. We subsequently fed these yeasts with increasing concentration of tabersonine (20, 125 and 225µM) of tabersonine that corresponds to the successive integration of definitive evidence was obtained by reconstituting the entire vindoline pathway…

Figure 1. The biosynthesis of vindoline from tabersonine. Tabersonine is converted into vindoline through a series of 7 reactions catalyzed by tabersonine 16-hydroxylase 2 (T16H2), 16-hydoxytabersonine-16-O-methyltransferase (16OMT), 16-methoxytabersonine 3-oxygenase (16T3O), uncharacterized deshydrogenase, N-methyltransferase (NMT), desacetoxyvindoline-4-hydroxylase (D4H) and deacetylvindoline-4-O-acetyltransferase (DAT).

Figure 2. Clustering of genes related to the vindoline biosynthetic pathway according to their expression levels A) Top: Average log2 fold changes in cluster n°18 in the 39 experimental conditions; bottom: log2 fold changes of target MIA genes (T16H2, 16T3O, NMT, D4H and DAT) found in cluster 18. B) Venn diagram of best co-expressed genes with known MIA genes in cluster n°18. Co-expressed gene lists were constructed by taking the 350 top ranking (Pearson correlation coefficient) transcripts for each MIA gene.

Figure 3. ADH13068 catalyzes the missing step of the vindoline biosynthetic pathway. LC-MS results using selected ion monitoring of the reaction compounds released in the culture medium of yeast expressing no heterologous enzymes; T16H2; T16H2 and 16OMT; T16H2, 16OMT and 16T3O; T16H2, 16OMT, 16T3O and ADH13068; or T16H2, 16OMT, 16T3O, ADH13068 and NMT (tabersonine, m/z 337, compound 1; 16-hydroxytabersonine,

m/z 353; compound 2; 16-methoxytabersonine, m/z 387; compound 3; 16-methoxytabersonine epoxy, m/z 383, compound 4; 16-methoxy-2,3-dihydro-3-hydroxytabersonine, m/z 385, compound 5; desacetoxyvindoline, m/z 399, compound 6).

Supplemental Table I: Time course analyses of the tabersonine conversion in yeasts gradually expressing the five first genes of the vindoline biosynthetic pathway via auto-replicative plasmids. Following induction of protein expression, 100 µM of tabersonine were added to each yeast strain and supernatants were collected after 16, 24 and 34 hours. The amounts of the detected products are expressed as 100 µM equivalent tabersonine.

Supplemental Table II: Primers used in this study.

Supplemental Figure 2. ADH7661 fails to catalyze the reduction of 16-methoxytabersonine epoxy. LC-MS results using selected ion monitoring of the reaction compounds released in the culture medium of yeast expressing no heterologous enzymes; T16H2; T16H2 and 16OMT; T16H2, 16OMT and 16T3O or T16H2, 16OMT, 16T3O and ADH7661 (tabersonine, m/z 337, compound 1; 16-hydroxytabersonine, m/z 353; compound 2; 16-methoxytabersonine, m/z 387; compound 3; 16-methoxytabersonine epoxy, m/z 383, compound 4).

Supplemental Figure 3. 16T3O, ADH13068 and NMT directly convert tabersonine to produce desactoxyvindorosine leading to the synthesis of vindorosine. LC-MS results using selected ion monitoring of the reaction compounds released in the culture medium of yeast expressing no heterologous enzymes; 16T3O; 16T3O and ADH13068 or 16T3O, ADH13068 and NMT (tabersonine, m/z 337, compound 1; epoxy tabersonine, m/z 353, compound 4'; desmethylvindoline, m/z 355, compound 5'; desacetoxyvindorosine, m/z 369, compound 6').

Material and Methods

Transcriptome analysis

Transcript abundance was estimated after mapping reads from 39 SRR accessions to representative sequences of the CDF97 transcriptome. Reads were aligned with Bowtie2 and expression levels determined with RSEM using the default parameters, as described in Dugé de Bernonville et al 2015. The resulting FPKM matrix was next processed in R to cluster transcripts according to their expression profiles with the package 'MBseq' (Si et al 2014). This package adapts a non-parametric model to a *k*-means based clustering to optimize the grouping of genes. Several *k* values were selected after examination of the distribution of total within-cluster sum of squares. The optimal number of cluster is expected to be the lowest value minimizing the total within-cluster sum of squares. Protein sequences were predicted from transcripts in CDF97 with Transdecoder and were annotated using BLASTP against the UniprotS database and HMMER against the PFAM database. Results were integrated with Trinotate into a MySQL database. Annotations were performed in parallel with the HpcGridRunner perl script on the CCSC-Artemis computing grid (CNRS, Orléans).

Functional assay in yeast

The functional assay was conducted by expressing sequentially T16H2, 16OMT, 16T3O, ADH13068 and NMT in the *S. cerevisiae* WAT11 strain harbouring the *Arabidopsis thaliana* P450 reductase (Pompon et al., 1996). The codding sequence of each gene was amplified using specific primers and cloned in pYeDP60, pESC-Leu, pESC-HIS, pESC-TRP (Agilent Technologies) as described in Supplemental Table 1. The resulting plasmids were gradually introduced in yeasts to construct strains expressing no enzyme (empty vectors), T16H2, T16H2-16OMT, T16H2-16OMT-16T3O, T16H2-16OMT-16T3O-ADH13068 or T16H2-16OMT-16T3O-ADH13068-NMT. Each strain was grown in an appropriated droup-out liquid media (4 ml) at 30°C for 24h prior harvesting by centrifugation and re-suspension

of yeast pellets in 10 ml of induction medium (YPGal). Yeasts (200 µl) were grown for 6 hours before addition of tabersonine (100 µM) and samples were recovered 16, 24 or 48 hours later. Cells and culture media were separated by centrifugation (8, 000 g) and 50 µl of supernatant were recovered and mixed with 150 µl of methanol. Samples were centrifuged again (15, 000 g) and supernatants were analysed by UPLC-MS as described in Besseau et al., 2013.

Acknowledgments

We gratefully acknowledge support from the "Région Centre" (France, ABISAL grant)

References

Besseau S, Kellner F, Lanoue A, Thamm AM, Salim V, Schneider B, Geu-Flores F, Höfer R, Guirimand G, Guihur A, Oudin A, Glevarec G, Foureau E, Papon N, Clastre M, Giglioli-Guivarc'h N, St-Pierre B, Werck-Reichhart D, Burlat V, De Luca V, O'Connor SE, Courdavault V. 2013. A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in *Catharanthus roseus*. Plant Physiol 163:1792-1803.

Brown S, Clastre M, Courdavault V, O'Connor SE. 20115. De novo production of the plant-derived alkaloid strictosidine in yeast. Proc Natl Acad Sci U S A 112:3205-3210.

Courdavault V, Papon N, Clastre M, Giglioli-Guivarc'h N, St-Pierre B, Burlat V. 2014. A look inside an alkaloid multisite plant: the *Catharanthus* logistics. Curr Opin Plant Biol 19:43-50.

Dugé de Bernonville T, Foureau E, Parage C., Lanoue A, Clastre M, Londono MAA, Oudin A, Houillé B, Papon N, Besseau S, Glévarec G, Atehortùa L, Giglioli-Guivarc'h N, St-Pierre B, De Luca V, O'Connor SE, Courdavault V. 2015. Characterization of a second secologanin

synthase isoform producing both secologanin and secoxyloganin allows enhanced de novo assembly of a *Catharanthus roseus* transcriptome. BMC Genomics

Kellner F, Geu-Flores F, Sherden NH, Brown S, Foureau E, Courdavault V, O'Connor SE. 2015. Discovery of a P450-catalyzed step in vindoline biosynthesis: a link between the aspidosperma and eburnamine alkaloids. Chem Commun DOI: 10.1039/C5CC01309G.

Koroch AR, Juliani HR, Kulakowski D, Arthur H, Asante-Dartey J, Simon JE (2009) *Voacanga africana*: Chemistry, Quality and Pharmacological Activity. In: ACS Symposium Series, Vol. 1021 Chapter 20, pp 363–380

Levac D, Murata J, Kim WS, De Luca V. Application of carborundum abrasion for investigating the leaf epidermis: molecular cloning of Catharanthus roseus 16-hydroxytabersonine-16-O-methyltransferase. Plant J 53:225-236.

Liscombe DK, Usera AR, O'Connor SE. 2010. Homolog of tocopherol C methyltransferases catalyzes N methylation in anticancer alkaloid biosynthesis. Proc Natl Acad Sci U S A 107:18793-18798.

Pompon D, Louerat B, Bronine A, Urban P. 1996. Yeast expression of animal and plant P450s in optimized redox environments. Methods Enzymol 272:51–64.

Si Y, Liu P, Li P, Brutnell TP. 2014. Model-based clustering for RNA-seq data. Bioinformatics 30:197–205.

Stavrinides A, Tatsis EC, Foureau E, Caputi L, Kellner F, Courdavault V, O'Connor SE. 2015. Unlocking the Diversity of Alkaloids in *Catharanthus roseus*: Nuclear Localization Suggests Metabolic Channeling in Secondary Metabolism. Chem Biol 22:336-341.

St-Pierre B, Laflamme P, Alarco AM, De Luca V. 1998/ The terminal *O*-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer. Plant J 14:703-713.

Vazquez-Flota F, De Carolis E, Alarco AM, De Luca V. 1997. Molecular cloning and characterization of desacetoxyvindoline-4-hydroxylase, a 2-oxoglutarate dependent-dioxygenase involved in the biosynthesis of vindoline in *Catharanthus roseus* (L.) G. Don. Plant Mol Biol 34:935-948.

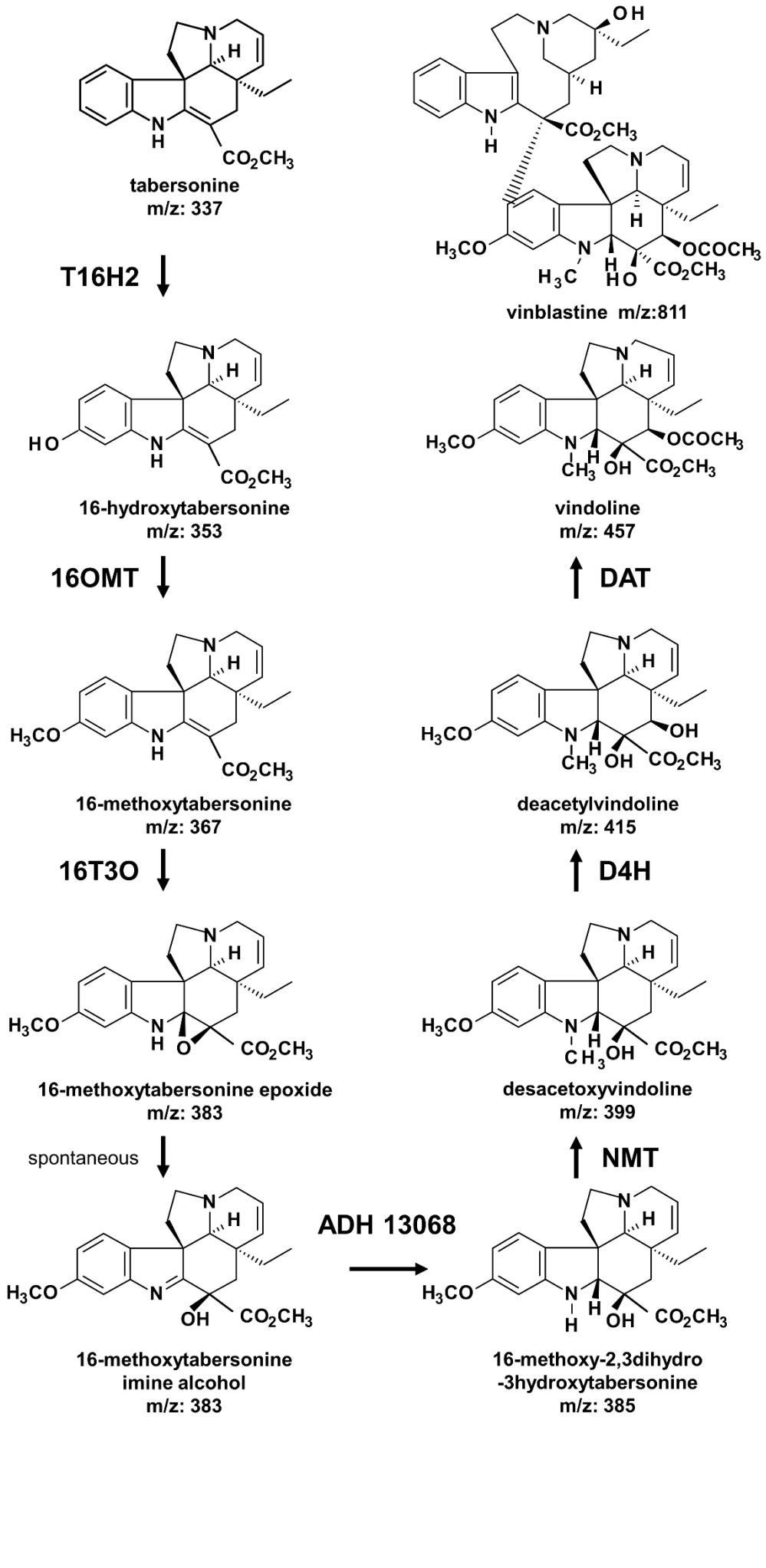

tabersonine
m/z: 337
T16H2
vinblastine m/z:811
16-hydroxytabersonine
m/z: 353
vindoline
m/z: 457
16OMT
DAT
16-methoxytabersonine
m/z: 367
deacetylvindoline
m/z: 415
16T3O
D4H
16-methoxytabersonine epoxide
m/z: 383
desacetoxyvindoline
m/z: 399
spontaneous
NMT
ADH 13068
16-methoxytabersonine
imine alcohol
m/z: 383
16-methoxy-2,3dihydro
-3hydroxytabersonine
m/z: 385

Figure 1. The biosynthesis of vindoline from tabersonine. Tabersonine is converted into vindoline through a series of 7 reactions catalyzed by tabersonine 16-hydroxylase 2 (T16H2), 16-hydoxytabersonine-16-O-methyltransferase (16OMT), 16-methoxytabersonine 3-oxygenase (16T3O), uncharacterized deshydrogenase, N-methyltransferase (NMT), desacetoxyvindoline-4-hydroxylase (D4H) and deacetylvindoline-4-O-acetyltransferase (DAT).

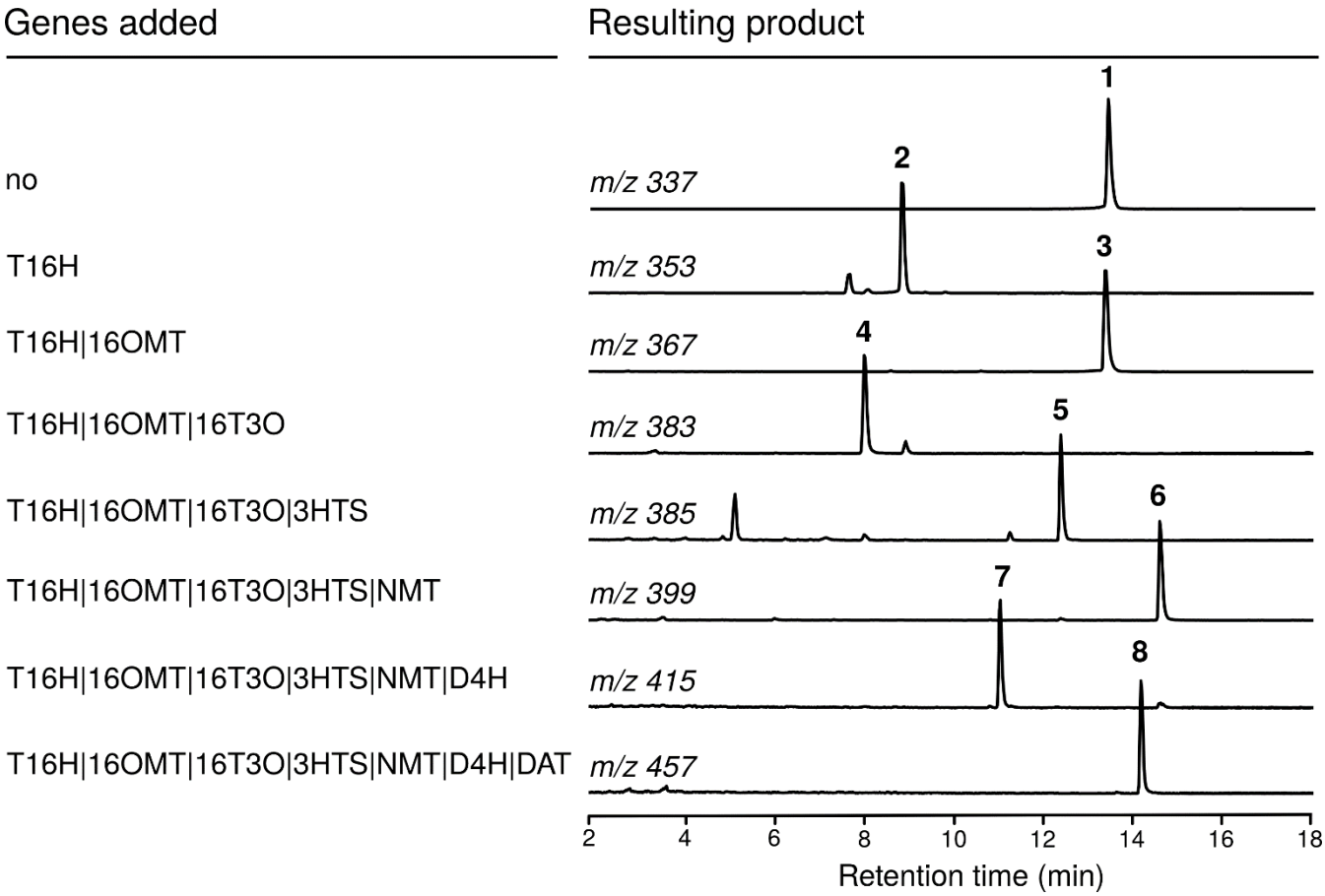


Figure 2. Clustering of genes related to the vindoline biosynthetic pathway according to their expression levels A) Top: Average log2 fold changes in cluster n°18 in the 39 experimental conditions; bottom: log2 fold changes of target MIA genes (T16H2, 16T3O, NMT, D4H and DAT) found in cluster 18. B) Venn diagram of best co-expressed genes with known MIA genes in cluster n°18. Co-expressed gene lists were constructed by taking the 350 top ranking (Pearson correlation coefficient) transcripts for each MIA gene.

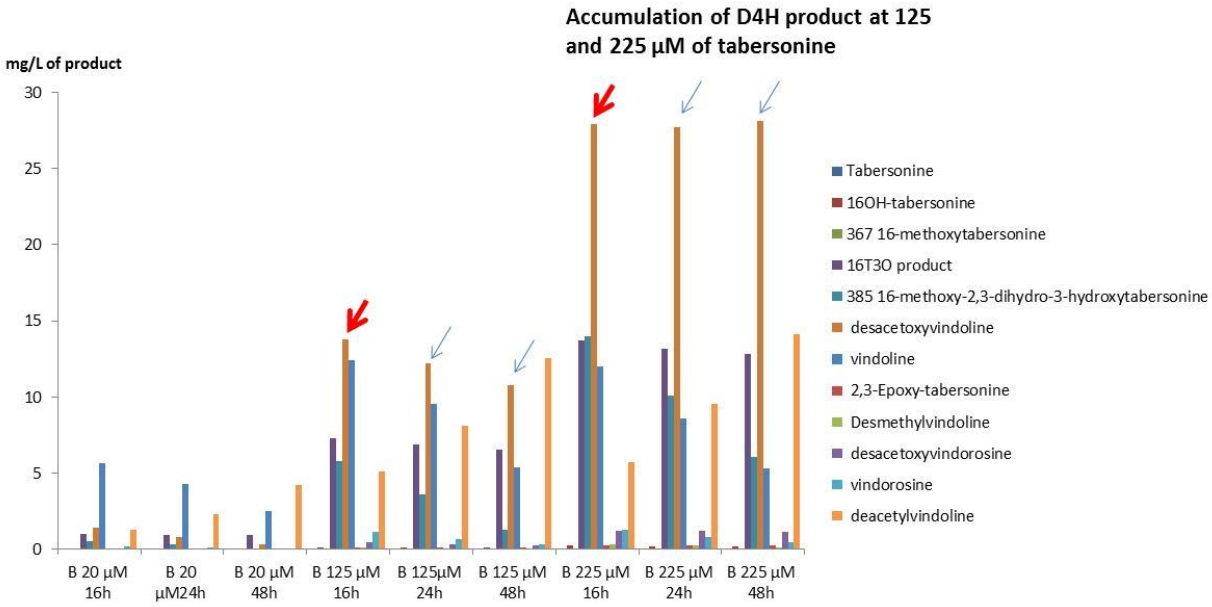
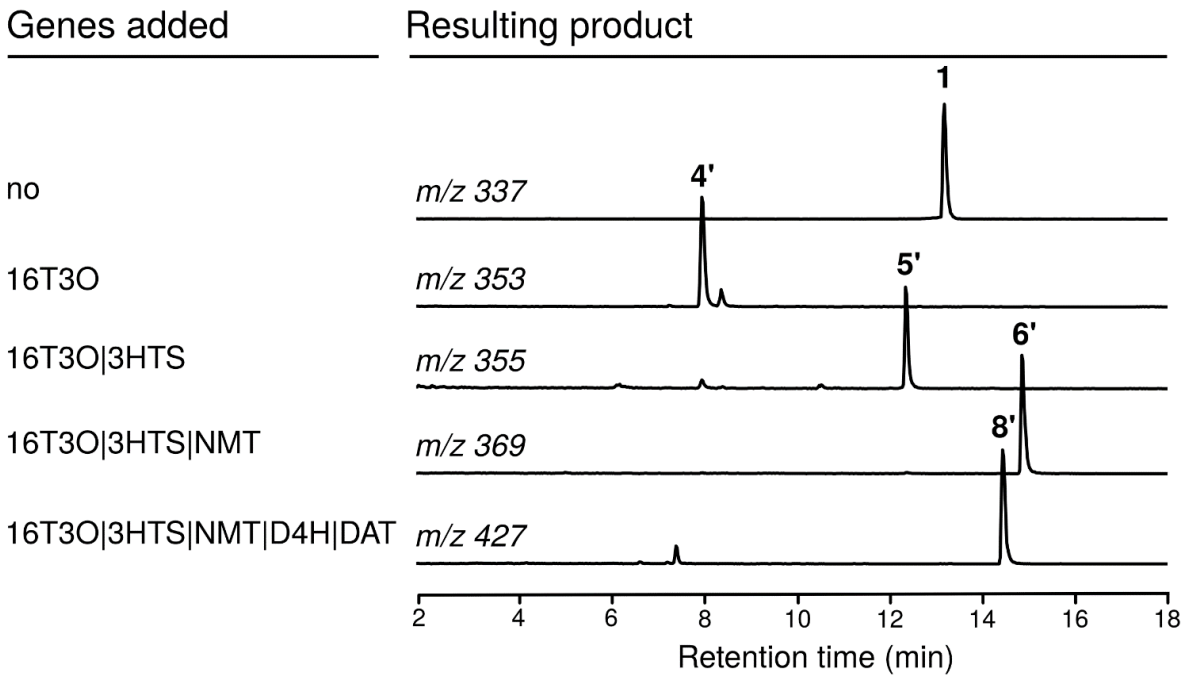
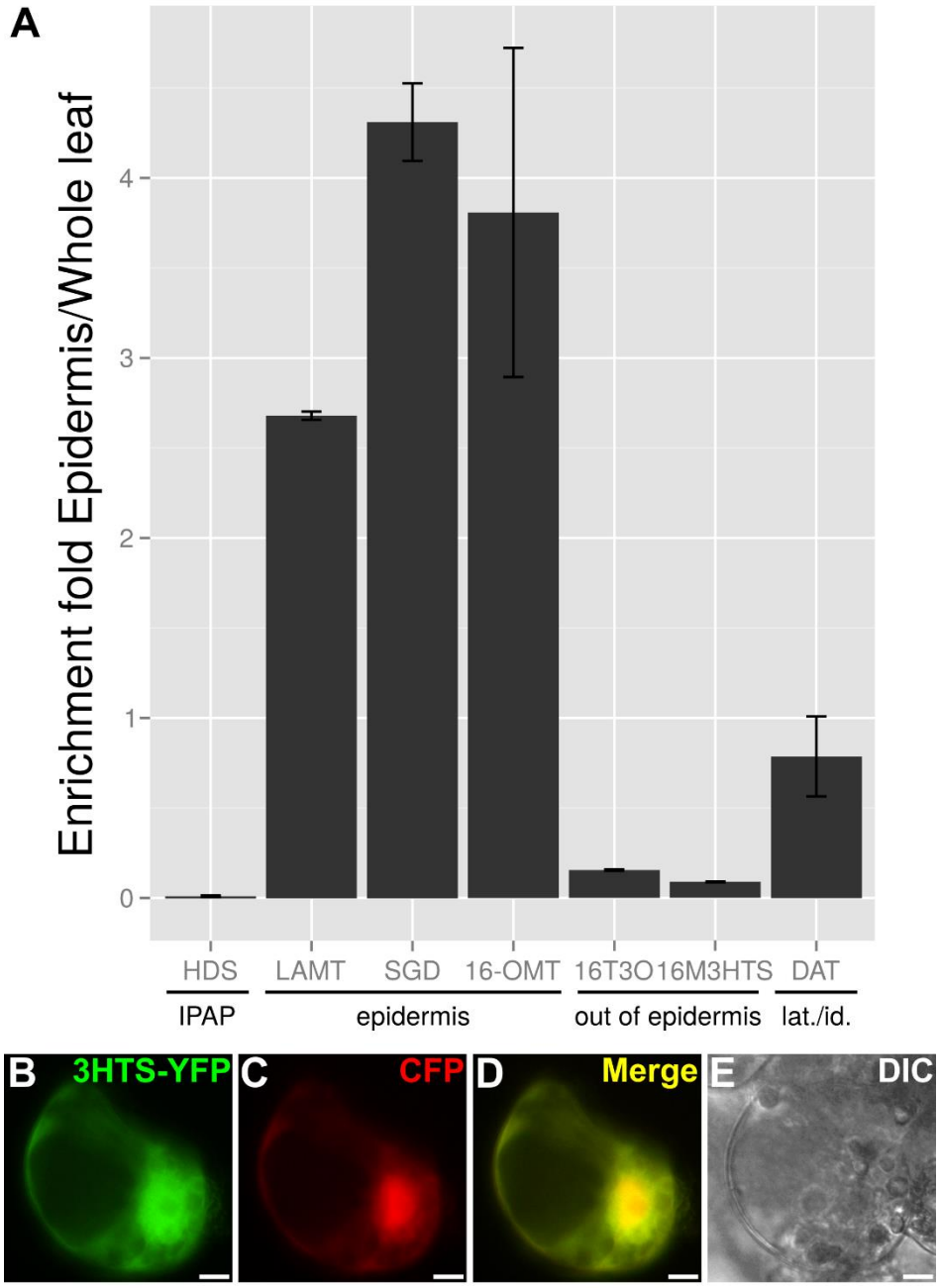


Figure 3. Converting tabersonine to vindoline and vindorosine using yeast strain with self-replicative plasmids. Yeast strain was fed with 20, 125 and 225 µM of tabersonine. Levels production of vindoline and vindorosine was analyze by UPLC-MS at 16, 24, and 48 hour intervals.



Supplemental Figure 3. 16T3O, ADH13068 and NMT directly convert tabersonine to produce desactoxyvindorosine leading to the synthesis of vindorosine. LC-MS results using selected ion monitoring of the reaction compounds released in the culture medium of yeast expressing no heterologous enzymes; 16T3O; 16T3O and ADH13068 or 16T3O, ADH13068 and NMT (tabersonine, m/z 337, compound 1; epoxy tabersonine, m/z 353, compound 4'; desmethylvindoline, m/z 355, compound 5'; desacetoxyvindorosine, m/z 369, compound 6').



Supplemental Figure 4. Gene levels expression in tissu. And subcellular localization of 16M3HTS.

Conclusion et Perspectives

Conclusion et Perspectives

La production de molécules d'intérêts d'origine végétale à partir de cultures cellulaires a toujours été une alternative prometteuse à l'extraction à grande échelle classiquement faite à partir des plantes. Leur faible taux de biosynthèse *in planta*, a conduit à la recherche de stratégies alternatives de production et à l'étude de leurs mécanismes de régulation afin d'optimiser la production de ces AIM. Après la synthèse chimique jugée trop chère, de multiples tentatives de production dans des cultures d'hairy roots et de cellules indifférenciées ont vu le jour. A l'exception du taxol produit dans des cultures cellulaires d'if (Jacrot et *al*., 1983; Kolewe et *al*., 2008), et dans une moindre mesure les alcaloïdes du pavot (Frick et *al*., 2007 ; Larkin et *al*., 2007), très peu d'autres cas ont été couronnés de succès y compris celui des alcaloïdes indoliques monoterpéniques. En effet, l'utilisation de ce type de stratégie chez *Catharanthus roseus* à un niveau industriel s'est avéré être limitée, par une vitesse et des taux de production insuffisants, ainsi qu'un faible nombre d'alcaloïdes synthétisés, se limitant à une faible minorité essentiellement composée d'AIM monomériques comme l'ajmalicine, la serpentine ou la tabersonine (Meijer et *al*., 1993 ; Hallard, 2000). Cette difficulté de production, semble être liée à une compartimentation complexe de la voie de biosynthèse des AIM chez *Catharanthus roseus* impliquant notamment une trentaine d'étapes enzymatiques réparties au sein de plus de 5 compartiments subcellulaires et 3 compartiments tissulaires distincts (Courdavault et *al*., 2014). Inhérente au caractère espèce-spécifique du métabolisme secondaire des plantes, la plupart de ces molécules sont absentes dans les systèmes modèles tels que celui d'*Arabidopsis thaliana* et leur voie de biosynthèse sont encore très mal caractérisée dans des espèces homologues produisant des AIM. Toutefois, de récents travaux de reconstitution de voies de biosynthèse hétérologue chez les levures de type *S. cerevisiae* ont ouvert la voie vers de nouvelles méthodes de production utilisant des approches du métabolisme engineering pour produire des molécules d'intérêt comme l'acide artémisinique, précurseur de l'artémisinine utilisé dans le traitement du paludisme ou encore la strictosidine, précurseur universel commun de tous les AIM (Ro et *al*., 2006 ; Paddon et *al*., 2013 ; Brown et al., 2015).

Mes travaux de thèse s'inscrivent dans cette dynamique et traitent de la caractérisation d'étapes de la voie de biosynthèse des AIM de *C. roseus* en vue d'obtenir une meilleure connaissance des enzymes associées aux étapes métaboliques de la voie. D'identifier certains

verrous métaboliques comme la présence d'isoenzymes spécifiques impliquées dans des étapes métaboliques, la distribution subcellulaire particulière de certaines enzymes de la voie ainsi que la formation de complexes entre partenaires protéiques qui sont autant d'étapes limitantes pour la production des AIM réalisée par bioingénierie.

Caractérisation d'étapes de la voie de biosynthèse des AIM

Les isoformes

La voie de biosynthèse des AIM est partagée au sein de trois types cellulaires différents que sont les IPAP, les épidermes et les cellules spécialisées comme les idioblastes et les laticifères. S'ajoute à cela, une distribution des enzymes de la voie dans plusieurs organites d'une même cellule (Courdavault et *al.*, 2014). Cette répartition complexe entre plusieurs types cellulaires et plusieurs compartiments subcellulaires distincts suscitent l'existence de multiples étapes de transports constituants des points limitants pour la biosynthèse. Cette complexité, se retrouve aussi dans la diversité des enzymes impliquées dans les étapes métaboliques de cette voie de biosynthèse avec parfois la présence d'isoformes impliquant des notions de spécificité d'enzyme.

Les AIM dérivent de la condensation de deux précurseurs, la tryptamine et la sécologanine respectivement synthétisés dans la voie des indoles et la voie des monoterpènes sécoiridoïdes (MTSI). Cette dernière est composée de 9 étapes réactionnelles aboutissant sur la formation de la loganine, convertie en sécologanine (précurseurs monoterpénique des AIM) par la sécologanine synthase (SLS) (Irmler et *al.*, 2000). La découverte d'une seconde isoforme de sécologanine synthase, nommée SLS2 chez *C. roseus*, pose donc la question d'une possible spécificité d'une isoforme dans la biosynthèse des AIM. Pour répondre à cette question nous nous sommes assurés en premier lieu de la validité de cette isoforme en comparant dans un premier temps la séquence du cytochrome P450 correspondant, avec la première isoforme de SLS identifiée (renommée SLS1) puis dans un second temps en comparant leur activité catalytique vis-à-vis de la loganine. Le gène codant l'isoforme SLS2 a montré une homologie de 97% avec la séquence nucléotidique du gène codant SLS. Des tests enzymatiques ont ensuite confirmés que SLS2 avait une activité catalytique comparable à celle de SLS1 dans la synthèse de la sécologanine à partir de loganine. Puis dans une seconde

partie, en mesurant l'abondance des produits des gènes *sls* et *sls2* nous permettant d'établir que SLS1 et SLS2 contribuent de façon concomitante à la synthèse de la sécologanine dans les racines de *C. roseus* tandis que SLS2 est l'isoforme prédominante de biosynthèse des AIM dans la partie aérienne de la plante. Dans les travaux d'Irmler et *al*., 2000 rapportant l'identification de la première isoforme de SLS (SLS1), l'enzyme avait été isolée de microsomes issus de culture cellulaire de *Catharanthus roseus*. La non différenciation cellulaire et tissulaire des cellules utilisées dans cette analyse est vraisemblablement la raison pour laquelle ils n'ont pas trouvé de spécificité à cette enzyme. La présence d'isoformes catalysant des étapes réactionnelles de la voie des monoterpènes sécoiridoïdes (MTSI) avait déjà été rapporté dans les travaux de Munkert et *al*., 2015, en montrant que l'iridoïde synthase (IS) catalysant l'étape de reduction du 8-oxogeranial en iridodial, appartient à une famille de six membres de progestérone 5β-réductase (P5βR: P5βR1 à P5βR6) dont certains sont capables de réduire le 8-oxogeranial et principalement le P5βR4 qui est impliqué dans la voie des MTSI.

L'implication d'isoformes de cytochrome P450 dans la biosynthèse des AIM, figure aussi dans des étapes de synthèse d'AIM finaux comme en témoigne les travaux de l'équipe auquel j'ai été associé. La première étape de conversion de la tabersonine en 16 hydroxytabersonine, dans la voie de biosynthèse de la vindoline, est initiée par une paire de cytochrome P450 (T16H1 et T16H2) possédant 82% d'identité protéique. Comme pour les isoformes de sécologanine synthase, nous avons montré l'existence d'une seconde isoforme de T16H (nommée T16H2, CYP71D351) qui catalyse l'hydroxylation de la tabersonine. Les gènes codant ces isoformes (T16H1 et T16H2) sont respectivement exprimés dans les fleurs et les jeunes feuilles impliquant une spécificité différente de ces isoformes. Des approches de génétique inverse de type VIGS (Virus Induce Gene Silencing) nous ont permis de montrer que l'isoforme nouvellement identifiée T16H2 était l'isoforme majoritairement impliquée dans la biosynthèse des AIM dans les jeunes feuilles (Besseau et *al*., 2013). L'existence d'isoformes de cytochrome P450 et l'expression différentielle de leur gène dans les organes de la plante a donc posé la question d'une éventuelle spécificité des cytochromes P450 dans la formation d'isoenzymes et du rôle relatif de chacune d'entre elles dans la biosynthèse des AIM.

Cette notion d'isoformes, se retrouve aussi au sein de la famille des NADPH cytochromes P450 réductases (CPR). Bien que ces enzymes ne soient pas directement

impliquées dans la production des AIM, elles participent au transfert des électrons vers les cytochromes P450 pour catalyser des réactions d'oxygénation. L'étude des CPR de *Catharanthus roseus* a montré l'existence de trois CPR homologues respectivement CPR1, CPR2 et CPR3 possédant des motifs conservés au sein des CPR, comme le domaine de liaison au NADPH, les domaines FMN et FAD impliqués dans le transfert des électrons. Une analyse globale de co-expression des gènes de *C. roseus* avec chaque CPR, utilisant le calcul du coefficient de corrélation de Pearson (PCC) a cependant montré malgré une forte homologie de séquence ces CPR entre elles, une spécificité d'action vis-à-vis de différents partenaires. CPR1 a majoritairement montré une corrélation d'expression avec des cytochromes P450 impliqués dans le métabolisme primaire comme la synthèse d'hormones et 25% provenant du métabolisme secondaire. Tandis que CPR2 corrèle avec l'expression de P450 associées au métabolisme secondaire dans la voie des phénylpropanoïdes notamment. CPR3 quant à elle montre peu de corrélation avec des P450 suggérant ainsi que cette protéine est impliquée dans la réduction d'autres protéines.

L'étude des alcools déshydrogénases (ADH), enzymes impliquées dans les étapes de biosynthèse des alcaloïdes de type hétéroyohimbine a aussi révélé la présence d'isoformes d'enzymes parmi lesquelles figures les tétrahydroalstonine synthase (THAS1 et THAS2), enzymes recherchée depuis le début des années 1980 (Hemscheidt et Zenk, 1985), ainsi qu'une hétéroyohimbine synthase fabriquant des molécules dont le squelette moléculaire est peu différent de la tétrahydroalstonine.

Fort de ses constatations les isoformes d'enzymes semblent être une entité récurrente ancrée au sein de la voie de biosynthèse des AIM. D'autant plus que d'anciens travaux de l'équipe ont également montré la présence d'isoformes d'enzymes de *C. roseus* impliquées dans des voies situées en amont synthétisant des composés terpéniques spécifiques alimentant la voie des AIM. La réaction d'isomérisation de l'IPP et son isomère le DMAPP dans la voie du MEP est catalysée par l'isopentényl diphosphate isomérase (IDI) qui possède quatre isoformes provenant respectivement de mécanismes de transcription différentielle de deux gènes *CrIDI1* et *CrIDI2* (une forme courte et une forme longue pour chaque gène) (Guirimand et *al*., 2012). Ces isoformes d'IDI introduisent un autre niveau de complexité qui est celui de la compartimentation intracellulaire au sein d'une même cellule. En effet dans ces travaux, l'équipe a pu montrer que par des mécanismes de transcription différentielle, la longue isoforme de *CrIDI1* était adressée dans 3 compartiments subcellulaires distincts, les

mitochondries, les plastes et les péroxisomes (en accord avec la présence en N-terminal d'un double peptide d'adressage aux mitochondries et aux plastes de 77 acides aminés ainsi qu'une séquence d'adressage aux péroxisomes en C-terminal d'une) tandis que la courte isoforme de *CrIDI1*, ne possédant pas de peptide bi-fonctionnel en N-terminal était adressée aux péroxisomes.

Cette compartimentation subcellulaire des isoformes est d'autant plus pertinente que dans une logique de régulation une telle spécificité pourrait expliquer le côté limitant dans le métabolisme AIM de *C. roseus*. Les isoformes constituent dorénavant une notion importante qui pose la question de spécificité d'une isoforme sur une autre et du choix de la bonne isoforme à utiliser dans des approches de métabolisme engineering.

Impact de la localisation subcellulaire

Dans cette première partie nous avons vu la complexité des systèmes enzymatiques intervenant sous la forme d'isoformes, démultipliant ainsi les acteurs potentiels dans la voie de biosynthèse des AIM, soulevant de multiples questions concernant la spécificité des enzymes de la voie. Un tel niveau de complexité est à prendre en compte dans la mise en place d'une bioproduction dans des systèmes hétérologues.

Dans ce contexte il est important de noter que la voie de biosynthèse des AIM possède une compartimentation intracellulaire complexe introduisant de multiples organites subcellulaires comme le plaste, hébergeant les principales voies de biosynthèses de précurseurs des terpènes (voie MEP et MVA), le cytosol et le réticulum endoplasmique hébergeant les étapes de la voie des monoterpènes sécoiridoïdes ainsi que la voie de la vindoline ou encore la vacuole, lieu principal de la formation du précurseur des AIM (la strictosidine) et d'accumulation des AIM dimères comme la vincristine et la vinblastine. Cette compartimentation complexe nécessite, de nombreuses étapes de transports entre les organites impliquant des transporteurs pas encore identifiés et des acteurs nécessaires pour orienter les flux de synthèse d'un organite à un autre qui sont autant de verrous dans une logique de bio-production.

La récente découverte de la voie de biosynthèse des alcaloïdes de type hétéroyohimbine a permis de conforter le rôle du noyau dans la biosynthèse des AIM jusqu'à

présent connu uniquement comme étant le lieu de la déglucosylation de la strictosidine en strictosidine aglycone par la SGD (Guirimand *et al.*, 2010). Dans nos travaux nous avons identifié une THAS rénommée THAS1 (tétrahydroalstonine synthase), et l'expression de cette dernière fusionnée à des protéines fluorescentes (YFP) dans des cellules de *C. roseus* nous a permis de montrer une localisation nucléaire de cette enzyme en accord avec la présence d'une NLS (Nuclear Localization Sequence) de type V. Des tests d'interaction protéique par BiFC (Bimolecular Fluorescence Complementation) avec la SGD ont ensuite montré la formation de complexes multimériques en forme de faucilles, de haut poids moléculaire, impliquant des multimères de THAS et de SGD dans le noyau.

Le rôle prépondérant du noyau dans la biosynthèse des AIM de type hétéroyohimbine s'est ensuite renforcé avec la découverte d'isoformes de THAS (THAS2, HYS) localisées au noyau grâce à la présence d'une NLS de type V pour HYS et par diffusion passive pour THAS2. Des essais en BiFC ont aussi confirmé leur interaction avec la SGD en formant des complexes multimériques au sein du noyau dont la forme est comparable à une faucille. Plusieurs hypothèses peuvent être formulées pour expliquer ces multimérisation. De récents travaux de l'équipe ont proposé le concept d'une « bombe à retardement », attribuant un rôle de défense de la SGD lors d'une agression de la plante. La rupture des membranes tonoplastiques et nucléaires liée à l'attaque d'un l'herbivore provoquerait une mise en contact massive de strictosidine et de SGD et la formation de strictosidine aglycone, produit hautement toxique pour l'insecte car il possède un pouvoir important de réticulation des protéines (Barleben *et al.*, 2007 ; Guirimand *et al.*, 2010). En condition normale de non stress, la synthèse de strictosidine et l'expression de SGD se maintiendraient à un niveau basal impliquant néanmoins une faible production de strictosidine aglycone. A ce stade, et pour pallier à la toxicité de ce dialdéhyde nous avons envisagé l'hypothèse que des enzymes situées à proximité de son site de production pourrait être impliquées dans la métabolisation rapide de ce composé actif, formant un métabolon nucléaire avec SGD (Guirimand *et al.*, 2010). Les interactions multiples de THAS1, THAS2 et HYS avec SGD renforceraient l'hypothèse d'un mécanisme d'évolution développé par les plantes qui accumule la strictosidine pour prendre en charge la réactivité de la strictosidine aglycone généré par SGD.

La formation des complexes multimériques avec SGD peut aussi s'expliquer par un recrutement des isoformes de THAS pour favoriser les flux métaboliques en faveur de la

production de tétrahydroalstonine et orienter ainsi la voie vers une production d'ajmalicine. La formation de complexes protéiques avec SGD chez *C. roseus*

Cette capacité de former des complexes associant des multimères de protéines a déjà été rapporté dans les travaux de Kim et *al*., 2000. En effet la glucosidase de l'avoine possède la capacité de former des multimères de sa sous unité monomérique AS-Glu1assemblées selon un empilement de cercle formant des fibrilles de différentes tailles. La capacité de former des multimères a été très discutée et semble être lié à un regroupement des enzymes pour optimiser la métabolisation du substrat. Il en est de même pour la forme de fibrille associée à un empilement en cercle des glucosidases qui permet un meilleur transite du substrat dans le tube de la fibrille qui par rapprochement spatiale est plus rapidement métabolisé. Cette notion abordée est celle du clustering d'enzyme discutée dans les travaux de Castellana et *al*., 2014 dans lesquels ils décrivent la stratégie de regroupement des enzymes appelée « clustering » formant des agglomérats compacts pour accélérer la métabolisation des substrats et le rendement de la réaction.

Valorisation et production d'AIM par ingénierie métabolique

Caractérisation des enzymes

La production de molécules d'intérêt d'origine végétale utilisant des approches tirées du métabolisme engineering, comme la reconstitution de voies de biosynthèse hétérologues chez des hôtes cellulaires, nécessite au préalable une connaissance complète de l'organisation des voies de biosynthèse.

La plupart des voies de biosynthèse de métabolites spécialisées présente une compartimentation intracellulaire complexe, reposant sur un adressage des enzymes dans des organites distincts. Chez *C. roseus* la répartition de la voie de biosynthèse des AIM dans de nombreux compartiments cellulaires, implique en corollaire, des transporteurs transmembranaires acheminant les intermédiaires métaboliques dans les organites cellulaires spécifiques influençant les flux de biosynthèse (Courdavault et *al*., 2014). L'acheminement de ces enzymes dans leurs compartiments subcellulaires respectifs se fait via la présence de motifs d'adressages présents sur leurs séquences protéiques. Toutefois certaines séquences

spécifiques d'adressage des plantes telles que les séquences d'adressages aux plastes, ne sont pas reconnues chez les levures et les bactéries, entrainant ainsi des biais de localisation lors de leur transfert. A titre d'exemple, les séquences d'adressages aux plastes présentes chez certaines enzymes des plantes sont supprimées avant leur expression dans des systèmes hétérologues. De tels phénomènes ont été illustrés aussi par l'expression de la strictosidine synthase (STR) en levure, qui a montré une sécrétion massive dans le milieu de culture au lieu d'avoir une localisation vacuolaire due à la promiscuité des séquences d'adressages à la vacuole avec la voie de sécrétion (Geerlings et *al*., 2001). Dans ce cas précis il convient alors, dans une logique de transfert de ces voies chez des hôtes hétérologues, de remplacer ces séquences d'adressages par d'autres, adaptées à l'organisme hôte.

D'autre part, cette production dans des usines cellulaires requière une parfaite maitrise du fonctionnement de chaque enzyme de la voie de biosynthèse chez la plante. A ce titre, un exemple a été illustré dans les travaux de Brown et *al*., 2015, avec l'utilisation d'une farnesyl diphosphate synthase mutée mFPS144 de poulet (Fernandez et *al*., 2000). Dans leur travaux, Fernandez et *al*., 2000 ont montré que la mutation N144W au niveau du site catalytique de l'enzyme entraine une synthèse préférentielle du GPP sur le FPP d'ordinaire non produit chez *S. cerevisiae*. Cette modification génique est mise à profit dans la levure pour orienter la voie du mévalonate endogène à la levure vers la production de GPP, précurseur essentiel dans la voie des monoterpènes sécoiridoïdes pour produire la strictosidine (Brown et *al*., 2015).

Expression de voie de biosynthèse hétérologue chez les levures

Une fois s'être assuré de la bonne localisation des enzymes dans leur compartiments subcellulaires respectifs dans les organismes hôtes et de leur caractérisation on peut commencer le travail de transfert des voies de biosynthèse.

La voie de la vindoline, constitue une partie finale de la voie de biosynthèse des AIM permettant la production de la vindoline qui par condensation avec la catharanthine donnera la vinblastine puis la vincristine. Cette voie de biosynthèse peut s'avérer limitante car du flux de vindoline peut dépendre celui des AIM. Celle-ci est constituée de 7 étapes enzymatiques initiées à partir de la tabersonine principalement accumulée dans les feuilles et possède une localisation subcellulaire partagée entre le réticulum endoplasmique et le cytosol ainsi qu'une distribution tissulaire complexe: avec au moins les deux premières étapes catalysées par la T16H et la 16OMT dans les épidermes et les deux dernières, correspondant à la D4H et la

DAT dans les laticifères et idioblastes. Cette distribution suggère donc l'existence de transporteurs membranaires pour permettre la translocation du métabolite intermédiaire entre ces types cellulaires. La question demeure sur la nature de cet intermédiaire de même que sur ces modalités d'acheminement jusqu'aux laticifères/idioblastes qui peut emprunter la voie apoplasmique ou symplasmique. Dans tous les cas, cette distribution peut limiter le flux métabolique par une limitation de l'accès des dernières enzymes de la voie à leur substrat. Fort de ces informations, le projet d'ingénierie métabolique de ces voies métaboliques dans un système hétérologue de levure repose sur la capacité de « reconstruire » ces chaînes dans un seul est unique système cellulaire afin de favoriser la vitesse d'enchainement des étapes.

Quand j'ai initié le travail d'expression de ces deux voies de biosynthèse dans la levure *Saccaromyces cerevisiae*, l'enzyme catalysant l'étape de réduction du 16-méthoxytabersonine imine alcool en 16-méthoxy-2,3dihydro-3hydroxytabersonine dans la voie de la vindoline n'était pas encore connue et la minovincinine-19-Oacetyltransferase (MAT) était inactive chez les levures. Pour cela, des analyses transcriptomiques visant à identifier des gènes candidats corrélés avec les gènes de la voie de la vindoline nous ont permis d'établir des coefficient de corrélation d'expression génique entre les 7 gènes de cette voie et l'ensemble des autres gènes pour identifier un candidat correspondant à une alcool déshydrogénase. La caractérisation de cette enzyme dans un système de levure s'est effectuée en parallèle des travaux de l'équipe du professeur De Luca qui a montré que cette alcool déshydrogénase identifiée renommée T3R (tabersonine-3-réductase) était capable de réduire la 16-méthoxytabersonine imine alcool en 16-méthoxy-2,3dihydro-3hydroxytabersonine, permettant *in fine*, après transfert des 7 gènes dans la levure, la production de la vindoline.

Optimisation des flux de synthèse

L'étude réalisée sur la caractérisation de la voie de la vindoline, et le transfert de l'ensemble de ses 7 gènes (de T16H2 à DAT) chez une souche *Saccaromyces cerevisiae* a montré une production de vindoline similaire avec l'étude réalisée par l'équipe du professeur De Luca. Nous obtenons une production de 3,1 mg/L contre 2,7 mg/L de vindoline à un temps de 16h. Qu et *al.*, 2015 ont aussi testé la production de la vindoline dans une souche de levure *S. cerevisiae* exprimant uniquement la CPR endogène de levure sans avoir co-exprimé au préalable la CPR de *C.roseus*. Après 12h d'incubation, ils obtiennent une production 2 à 3 fois moins importante (1,1 mg/L) que dans une souche sur-exprimant la CPR de *C.roseus (*2,7

mg/L) ou dans une souche WAT11 sur exprimant la CPR d'*Arabidopsis thaliana* (3,1 mg/L) dans notre expérience. Cette notion est d'autant plus importante que parmi les 7 gènes transféré dans la levure 2 d'entre eux codent des cytochromes P450 (T16H2 et T3O) utilisant des électrons provenant de NADPH cytochromes P450 réductases (CPR) pour catalyser des réactions d'oxygénations. Le nombre, le taux d'expression ainsi que la spécificité des interactions entre les CPR et les P450 de plantes semble être un facteur déterminant dans la production de la vindoline.

Dans notre étude, nous avons utilisés deux méthodes de transfert des gènes de la voie de la vindoline. Dans un premier temps, nous avons utilisé une souche *S. cerevisiae* (WAT11 exprimant la CPR d'*Arabidopsis thaliana* en plus de la CPR endogène de levure) auto-réplicative dans laquelle les plasmides utilisés permettant l'expression des gènes de la voie, sont répliqués de nombreuses fois pour obtenir de nombreuses copies des gènes de la voie. Dans la deuxième nous avons intégré tous les gènes un à un dans le génome de cette même levure. La production de vindoline par ces deux méthodes montre une production de 0,60 mg/L dans la souche intégrative contre 3,2 mg/L dans une souche auto-réplicative. Cette production semble être influencée par la méthode utilisée pour exprimer les gènes au sein de la levure. Une souche auto-réplicative est plus efficace sur les rendements de production vraisemblablement lié au fait que l'expression de chaque gène de la voie est plus forte lorsqu'ils sont sous la dépendance de promoteur inductible au galactose comme le promoteur GAL1 utilisé dans tous nos plasmides d'expression

La voie de la vindoline est une voie de biosynthèse imbriquée dans laquelle, à partir de la tabersonine, les enzymes forment successivement la vindoline et de la vindorosine. En effet, la tabersonine peut également être converties en vindorosine (16-desméthylvindoline) lorsque les deux premières étapes de la voie de la vindoline sont contournées. Cette dernière est une voie homologue de la vindoline dans laquelle T3O, T3R, NMT, D4H et DAT se succèdent pour former la vindorosine. La présence de ce métabolisme imbriquée nécessite de trouver des alternatives permettant d'optimiser la synthèse en faveur de la vindoline. Dans notre étude et pour forcer le flux de synthèse en faveur de la vindoline nous avons dans la souche de levure intégrative comportant l'ensemble des gènes de la voie de la vindoline ajouté une copie supplémentaire d'un plasmide auto réplicatif contenant le gène de la T16H2 pour favoriser la conversion de la tabersonine en 16 hydroxytabersonine. Nous avons obtenu une production de 0,66mg/L de vindoline pour un ratio (vindoline/vindorosine) de 15 contre

3,27mg/L de vindoline et un ratio (vindoline/vindorosine) de 9 dans une souche auto-réplicative. Ces résultats suggèrent donc que l'intégration des gènes dans le génome de *S. cerevisiae* entraine une production moins importante en termes de ratio mais plus spécifique pour la vindoline. Beaucoup de facteur influençant la conversion des métabolites sont à prendre en compte pour la production de la vindoline. L'utilisation d'une méthode auto-réplicatif offre une meilleur production final de vindoline, seulement avec une accumulation des intermédiaires réactionnels importante liée à une forte expression de chaque protéine et donc une perte de production et une accumulation de vindorosine. Tandis que la méthode intégrative montre une synthèse de vindoline moins importante mais plus spécifique de la vindoline avec une meilleur utilisation des intermédiaires réactionnels. Ces systèmes de production sont dans une plus grande mesure inhérent à la vitesse de conversion enzymatique et donc aux constantes catalytique de chaque enzyme. Un travail de caractérisation de chaque enzyme est nécessaire si l'on souhaite augmenter les rendements de production comme je l'ai déjà décrit avec l'exemple de la Farnésyl diphosphate synthase de (Fernandez et *al*., 2000).

Fort de toutes ces informations, l'étude de la voie de biosynthèse des AIM chez *Catharantus roseus* et la découverte des nouvelles enzymes nous permettent de dégager certaines notions inhérentes aux voies métaboliques. Dans le métabolisme spécialisé de *C. roseus*, nous avons découvert qu'il y a un nombre limité d'enzyme qui interviennent selon un ordre précis dans des portions de voie métabolisme de la voie de biosynthèse des AIM. Cet ordre représente une combinatoire propre à une espèce de plante dans lequel chaque enzyme catalyse une réaction à la suite d'une autre. Cette combinatoire peut être retrouvée parmi les plantes médicinales produisant des AIM. A titre d'exemple chez *Rauvolfia serpentina*, d'anciens travaux de l'équipe ont participé à la caractérisation de la localisation subcellulaire d'une orthologue de STR et de SGD impliquées dans des étapes de conversion de la strictosidine en strictosidine aglycone puis en cathénamine (Guirimand et *al*., 2010). La présence de ses orthologues suggère l'existence d'un mécanisme commun avec *C. roseus* selon une même combinatoire enzymatique.

Faire de l'analogie métabolique avec d'autres plantes ou de l'analogie d'expression génique est un bon moyen pour alimenter la connaissance sur les voies métaboliques et l'utilisation des données transcriptomiques à haut débit permet de caractériser des gènes de la voie de biosynthèse des AIM. Deux stratégies non exclusives peuvent être utilisées: l'orthologie qui permet de comparer les données transcriptomiques avec d'autres espèces

produisant ou non des AIM et le regroupement de l'expression génique. La première approche requiert des données transcriptomiques (disponibles à partir des projets du MPGR et PhytoMetaSyn) d'espèces candidates avec un métabolisme des alcaloïdes spécifique et d'autres espèces ne possédant pas ce métabolisme. La comparaison d'orthologue se fait en utilisant des algorithmes servant à identifier les gènes codant pour des enzymes conservées impliquées dans le biosynthèse des précurseurs communs chez les espèces productrices d'AIM (*Camptotheca acuminata, Rauwolfia serpentina, C. roseus*), mais aussi des activités enzymatiques spécifiques dédiée à la synthèse d'alcaloïdes particuliers. Une seconde analyse d'orthologie peut encore être effectuée sur l'ensemble des gènes conservés trouvés dans les espèces produisant des AIM pour isoler les candidats de voie spécifique en comparant avec les données du transcriptome d'espèce ne produisant pas d'AIM. La seconde stratégie repose sur l'analyse du transcriptome de diverses conditions expérimentales pour construire un transcriptome de référence qui sera utilisé pour estimer l'abondance de chaque transcription dans chaque condition afin de regrouper des gènes en fonction de leurs niveaux d'expression.

De ces analogies de voie métaboliques on peut dégager plusieurs règles qui se répètent dans plusieurs modèles métaboliques. La succession des enzymes d'une voie métabolique se fait selon une combinatoire bien précise qui peut être commune ou non à un ensemble de plante. Pour autant l'étude des P450 nous a permis de vérifier certaines règles : un cytochrome P450 ne suit pas un autre P450 dans une voie métabolique. L'action de deux P450 est à chaque fois séparée par une autre enzyme comme par exemple les alcools déshydrogénases. De plus, la proximité des enzymes des voies métaboliques peuvent parfois de façon naturelle former des fusions de gènes conférant un avantage métabolique. Dans la voie de l'acide mycophénolique des levures, un cytochrome P450 est fusionné à une ADH. Cette proximité permet de réaliser deux réactions successives en empêchant la fuite du métabolite après oxygénation par le P450.

Il semble aussi que le vivant soit dans une moindre mesure capable de former à partir d'enzymes impliquées dans le métabolisme secondaire des enzymes homologues catalysant des réactions similaires dans le métabolisme secondaire des plantes impliquant la notion de promiscuité enzymatique. Un bon exemple a été illustré dans les travaux de Liscombe et *al*., 2010 dans lequel ils ont montré qu'une protéine possédant une très forte homologie de séquence avec les γ-tocopherol methyltransferases (γ-TMT) impliquées dans le métabolisme primaire de biosynthèse de la vitamine E était capable de réaliser la méthylation de substrat azoté comme la conversion de 16-methoxy-2,3-dihydro-3-hydroxytabersonine en

desacetoxyvindoline dans la voie de la vindoline. Cet exemple montre la complexité des systèmes enzymatiques associés au métabolisme secondaire de *Catharanthus roseus* étant capable, par des mécanismes évolutifs de recycler des enzymes du métabolisme primaire pour les adapter au métabolisme secondaire.

En perspectives de ces travaux de thèse, il est incontournable de continuer l'étude de caractérisation de la voie de biosynthèse des AIM de *Catharanthus roseus* car beaucoup d'étapes de biosynthèse sont encore inconnues. Notamment la voie de la catharanthine pour laquelle rien est élucidé aujourd'hui mais aussi le noyau formateur des AIM résidant dans les étapes de conversion de la strictosidine en strictosidine aglycone et ses dérivés jusqu'à la cathénamine. Les étapes suivant la biosynthèse de la strictosidine aglycone par déglucosylation de la strictosidine catalysée par la strictosidine glucosidase (SGD) génèrent différentes formes d'AIM à l'origine des divers précurseurs pour les voies arborescentes comme la voie de la vindoline initiée par la tabersonine , la voie de la catharanthine, les voies formant les alcaloïdes de type hétéroyohimbine amenant vers l'ajmalicine et la serpentine longtemps recherchées. Ces étapes de biosynthèse sont d'autant plus importantes que les précurseurs qui y sont formés sont à l'origine de la structure arborescente et de la complexité des trois grandes familles d'AIM retrouvés chez *Catharanthus roseus* que sont les alcaloïdes de type Iboga, Corynanthe et les Aspidosperma.

Dans ce contexte particulier il semble aussi très important de continuer l'étude sur l'architecture de la voie de biosynthèse des AIM car de nombreux transporteurs restent à élucider. L'architecture tissulaire du métabolisme des AIM est partagée au sein de trois types cellulaires différents que sont les IPAP, les épidermes et les cellules spécialisées comme les idioblastes. Un tel niveau de complexité implique la présence de transporteurs exprimés pour permettre le transport des métabolites d'un tissu à un autre. S'ajoute à cela une répartition dans de nombreux compartiments cellulaires, impliquant en corollaire, des transporteurs transmembranaires acheminant les intermédiaires métaboliques dans les organites cellulaires spécifiques influençant les flux de biosynthèse comme la présence de transporteurs vacuolaires permettant l'entrée et la sortie de la tryptamine et de la sécologanine condensés en strictosidine dans la vacuole puis acheminée au noyau.

Il reste encore certains verrous à lever pour assurer une production complète des AIM dans les levures. En effet, certains biais de localisation d'enzyme liés à la présence de motifs d'adressage non reconnus par les levures restent toujours sans solution. A titre d'exemple,

SGD qui possède une localisation nucléaire chez *C. roseus* est aussi vacuolaire chez les levures. De tels problèmes de localisation sont à prendre en compte dans des approches de production par ingénierie métabolique. Par ailleurs la réactivité du dialdéhyde généré après déglucosylation de la strictosidine par la SGD est un problème qui est difficilement contournable dans la levure. Les multiples tentatives de fusion de protéines réalisées avec SGD, permettant de convertir directement l'aglycone réactif après sa synthèse, son encore à leur stade de balbutiement.

Bibliographie

A

Almagro, L., Fernández-Pérez, F., Pedreño, M. A. (2015) Indole Alkaloids from Catharanthus roseus: Bioproduction and Their Effect on Human Health. Molecules., 20, 2973-3000.

Argos, P., Landy, A., Abremski, K., Egan, J. B., Haggard-Ljungquist, E., Hoess, R. H., Kahn., M. L., Narayana.,S. V., Pierson 3rd, L. S. (1986) The integrase family of site-specific recombinases: regional similarities and global diversity. The Journal of the European molecular Biology Organization., 5, 433-440.

Asada, K., Salim, V., Masada-Atsumi, S., Edmunds, E., Nagatoshi, M., Terasaka, K., Mizukami, H., De Luca, V. (2013) A 7-Deoxyloganetic Acid Glucosyltransferase Contributes a Key Step in Secologanin Biosynthesis in Madagascar Periwinkle. The Plant Cell., 25, 4123–4134.

Ashihara, H., Sano, H., Crozier, A. (2008) Caffeine and related purine alkaloids: biosynthesis, catabolism, function and genetic engineering. Phytochemistry., 69, 841-856.

Aslam, J., Khan, S. H., Siddiqui, Z. H., Fatima, Z., Maqsood, M., Bhat, M. A., Nasim, S.A., Ilah, A., Khan, S.A., Mujib, A., Sharma, M., P. (2010) Catharanthus roseus (L.) G. Don. An important drug: it's applications and production. Pharmacie Globale International Journal of Comprehensive Pharmacy., 4, 1-16.

Avila, J., Ulloa, L., González, J., Moreno, F., Díaz-Nido, J. (1994) Phosphorylation of microtubule-associated proteins by protein kinase CK2 in neuritogenesis. Cellular and Molecular Biology Research., 40, 573–579.

B

Balint, G. A. (2001) Artemisinin and its derivatives: an important new class of antimalarial agents. Pharmacology and Therapeutics., 90, 261-265.

Balsevich, J. and Bishop, G. (1989) Distribution of Catharanthine, Vindoline and 3′,4′-Anhydrovinblastine in the Aerial Parts of some Catharanthus roseus Plants and the Significance Thereof in Relation to Alkaloid Production in Cultured Cells. In Prim Second Metabolism Plant Cell Culture II., 149–153.

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., Werck-Reichhart, D. (2011) Cytochromes P450. The Arabidopsis Book/American Society of Plant Biologists., 9.

Barleben, L., Panjikar, S., Ruppert, M., Koepke, J., Stöckigt, J. (2007) Molecular architecture of strictosidine glucosidase: the gateway to the biosynthesis of the monoterpenoid indole alkaloid family. The Plant Cell., 19, 2886–2889.

Bellmunt, J., Fougeray, R., Rosenberg, J. E., Von der Maase, H., Schutz, F. A., Salhi, Y., Choueiri, T. K. (2013) Long-term survival results of a randomized phase III trial of vinflunine plus best supportive care versus best supportive care alone in advanced urothelial carcinoma patients after failure of platinum-based chemotherapy. Annals of Oncology., 24, 1466-1472.

Besseau, S., Kellner, F., Lanoue, A., Thamm, A.M.K., Salim, V., Schneider, B., Geu-Flores, F., Hofer, R., Guirimand, G., Guihur, A., Oudin, A., Glévarec, G., Foureau, E., Papon, N., Clastre, M., Giglioli-Guivarc'h, N., St-Pierre, B., Werck-Reichhart, D., Burlat, V., De Luca, V., O'Connor, S.E., Courdavault, V. (2013) A Pair of Tabersonine 16-Hydroxylases Initiates the Synthesis of Vindoline in an Organ-Dependent Manner in Catharanthus roseus. Plant Physiology., 163, 1792-1803.

Blom, T. J. M., Sierra, M., Vliet, T. B., Franke-van Dijk, M. E. I., Koning, P., Iren, F., Verpoorte, R., Libbenga, K. R. (1991) Uptake and accumulation of ajmalicine into isolated vacuoles of cultured cells of Catharanthus roseus (L.) G. Don. and its conversion into serpentine. Planta., 183, 170–177.

Bouvier, F., Rahier, A., Camara, B. (2005) Biogenesis, molecular regulation and function of plant isoprenoids. Progress in Lipid Research., 44, 357-429.

Brown, G. D. (2010) The biosynthesis of artemisinin (Qinghaosu) and the phytochemistry of Artemisia annua L.(Qinghao). Molecules., 15, 7603-7698.

Burch-Smith, T. M., Anderson, GJ. C., Martin, G. B., Dinesh-Kumar, S. P. (2004) Applications and advantages of virus-induced gene silencing for gene function studies. The Plant Journal., 39, 734–746.

Burlat, V., Oudin, A., Courtois, M., Rideau, M., St-Pierre, B. (2004) Co-expression of three MEP pathway genes and geraniol 10-hydroxylase in internal phloem parenchyma of Catharanthus roseus implicates multicellular translocation of intermediates during the biosynthesis of monoterpene indole alkaloids and isoprenoid-derive. The Plant Journal., 38, 131–141.

C

Carqueijeiro, I., Masini, E., Foureau, E., Sepúlveda, L. J., Marais, E., Lanoue, A., Besseau, S., Papon, N., Clastre, M., Dugé de Bernonville, T., Glévarec, G., Atehortùa, L., Oudin, A., Courdavault, V. (2015) Virus-induced gene silencing in Catharanthus roseus by biolistic inoculation of tobacco rattle virus vectors. Plant Biology., 17, 1242-1246.

Castellana, M., Wilson, M. Z., Xu, Y., Joshi, P., Cristea, I. M., Rabinowitz, J. D., Gitai, Z., Wingreen, N. S. (2014) Enzyme clustering accelerates processing of intermediates through metabolic channeling. Nature Biotechnology., 32, 1011-1018.

Chahed, K., Oudin, A., Guivarc'h, N., Hamdi, S., Chénieux, J. C., Rideau, M., Clastre, M. (2000) 1-Deoxy-D-xylulose 5-phosphate synthase from periwinkle: cDNA identification and induced gene expression in terpenoid indole alkaloid-producing cells. Plant Physiology and Biochemistry., 38, 559-566.

Collu, G., Unver, N., Peltenburg-Looman, A.M.G., van der Heijden, R., Verpoorte, R., Memelink, J. (2001) Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. FEBS Letters., 508, 215-220.

Conn, E. E. (1981) Secondary plant products. P. K. Stumpf (Ed.), Academic Press., 7.

Contin, A., van der Heijden, R., Lefeber, A. W., Verpoorte, R. (1998) The iridoid glucoside secologanin is derived from the novel triose phosphate/pyruvate pathway in a Catharanthus roseus cell culture. FEBS Letters., 434, 413–416.

Cordier, H., Karst, F., Bergès, T. (1999) Heterologous expression in Saccharomyces cerevisiae of an Arabidopsis thaliana cDNA encoding mevalonate diphosphate decarboxylase. Plant Molecular Biology., 39, 953–967.

Courdavault, V., Papon, N., Clastre, M., Giglioli-Guivarc'h, N., St-Pierre, B., Burlat, V. (2014) A look inside an alkaloid multisite plant: the Catharanthus logistics. Current Opinion in Plant Biology., 19, 43-50.

Crea, R., Kraszewski, A., Hirose, T., Itakura, K. (1978). Chemical synthesis of genes for human insulin. Proceedings of the National Academy of Sciences., 75, 5765-5769.

D

De Carolis, E. and De Luca, V. (1993) Purification, characterization, and kinetic analysis of a 2-oxoglutarate-dependent dioxygenase involved in vindoline biosynthesis from Catharanthus roseus. The Journal of Biological Chemistry., 268, 5504–5511.

De Luca, V., Balsevich, J., Kurz, W. G. W. (1985) Acetyl Coenzyme A: Deacetylvindoline O-Acetyltransferase, A Novel Enzyme from Catharanthus. Journal of Plant Physiology., 121, 417–428.

De Luca, V., Balsevich, J., Tyler, R. T., Kurz, W. G. W. (1987) Characterization of a novel N-methyltransferase (NMT) from Catharanthus roseus plants. Plant Cell Reports., 6, 458–461.

De Luca, V. and Cutler, A. J. (1987) Subcellular Localization of Enzymes Involved in Indole Alkaloid Biosynthesis in Catharanthus roseus. Plant Physiology., 85, 1099–1102.

De Luca, V., Brisson, N., Balsevich, J., Kurz, W. G. W. (1989) Regulation of Vindoline Biosynthesis in Catharanthus roseus: Molecular Cloning of the First and Last Steps in Biosynthesis. Primary and Secondary Metabolism of Plant Cell Culture II., 18, 154–161.

Deng, C. X. (2012) The Use of Cre–loxP Technology and Inducible Systems to Generate Mouse Models of Cancer. In Genetically Engineered Mice for Cancer Research., 17-36.

Dethier, M., and De Luca, V. (1993) Partial purification of an N-methyltransferase involved in vindoline biosynthesis in Catharanthus roseus. Phytochemistry., 32, 673-678.

Deus, B. and Zenk, M. H. (1982) Exploitation of plant cells for the production of natural compounds. Biotechnology and Bioengineering., 24, 1965-1974.

De Waal, A., Meijer, H., Verpoorte, R. (1995) Strictosidine synthase from Catharanthus roseus: purification and characterization of multiple forms. Biochemistry Journal., 306, 571–580.

Donald, K. A., Hampton, R. Y., Fritz, I. B. (1997) Effects of overproduction of the catalytic domain of 3-hydroxy-3-methylglutaryl coenzyme A reductase on squalene synthesis in Saccharomyces cerevisiae. Applied and Environmental Microbiology., 63, 3341-3344.

Drea, S., Hileman, L. C., de Martino, G., Irish, V. (2007) Functional analyses of genetic pathways controlling petal specification in poppy. Development (Cambridge, England)., 134, 4157–4166.

Duffin, J. (2000) Poisoning the spindle: serendipity and discovery of the anti-tumor properties of the Vinca alkaloids. Canadian Bulletin of Medical History., 17, 155–192.

De Bernonville, T. D., Clastre, M., Besseau, S., Oudin, A., Burlat, V., Glévarec, G., Lanoue, A., Papon, N., Giglioli-Guivarc'h, N., St-Pierre, B., Courdavault, V. (2015) Phytochemical genomics of the Madagascar periwinkle: Unravelling the last twists of the alkaloid engine. Phytochemistry., 113, 9-23.

Dutta, A., Batra, J., Pandey-Rai, S., Singh, D., Kumar, S., Sen, J. (2005) Expression of terpenoid indole alkaloid biosynthetic pathway genes corresponds to accumulation of related alkaloids in Catharanthus roseus (L.) G. Don. Planta., 220, 376–383.

E

El-Sayed, M. and Verpoorte, R. (2007) Catharanthus terpenoid indole alkaloids: biosynthesis and regulation. Phytochemistry Reviews., 6, 277–305.

Elson, C. E., and Yu, S. G. (1994) The chemoprevention of cancer by mevalonate-derived constituents of fruits and vegetables. The Journal of Nutrition., 124, 607.

F

Facchini, P. J. (2001) Alkaloid Biosynthesis in plants: Biochemistry, Cell Biology, Molecular Regulation, and Metabolic Engineering Applications. Annual Review of Plant Physiology and Plant Molecular Biology., 52, 29–66.

Fang, F., Salmon, K., Shen, M. W., Aeling, K. A., Ito, E., Irwin, B., Sandmeyer, S. (2011) A vector set for systematic metabolic engineering in Saccharomyces cerevisiae. Yeast., 28, 123-136.

Farnsworth, N. R., Akerele, O., Bingel, A. S., Soejarto, D. D., Guo, Z. (1985) Medicinal plants in therapy. Bulletin of the World Health Organization., 63, 965-981.

Favali, M. A., Musetti, R., Benvenuti, S., Bianchi, A., Pressacco, L. (2004) Catharanthus roseus L. plants and explants infected with phytoplasmas: alkaloid production and structural observations. Protoplasma., 223, 45–51.

Finch, S. A. E., and Stier, A. (1991) Rotational diffusion of homo-and hetero-oligomers of cytochrome P-450, the functional significance of cooperativity and the membrane structure. Frontiers in Biotransformation., 5, 34-70.

Forster, A. C., and Church, G. M. (2006) Towards synthesis of a minimal cell. Molecular Systems Biology., 2, 45.

Friedman, M., Henika, P. R., Levin, C. E., Mandrell, R. E. (2004) Antibacterial activities of plant essential oils and their components against Escherichia coli O157: H7 and Salmonella enterica in apple juice. Journal of Agricultural and Food Chemistry., 52, 6042-6048.

Frick, S., Kramell, R., Kutchan, T. M. (2007) Metabolic engineering with a morphine biosynthetic P450 in opium poppy surpasses breeding. Metabolic engineering., 9, 169-176.

G

Geerlings, A. (2000) Molecular Cloning and Analysis of Strictosidine beta -D-Glucosidase, an Enzyme in Terpenoid Indole Alkaloid Biosynthesis in Catharanthus roseus. Journal of Biology Chemistry., 275, 3051–3056.

Geerlings, A., Redondo, F., Contin, A., Memelink, J., van Der Heijden, R., Verpoorte, R. (2001) Biotransformation of tryptamine and secologanin into plant terpenoid indole alkaloids by transgenic yeast. Applied Microbiology and Biotechnology., 56, 420-424.

Gerasimenko, I., Sheludko, Y., Ma, X., Stöckigt, J. (2002) Heterologous expression of a Rauvolfia cDNA encoding strictosidine glucosidase, a biosynthetic key to over 2000 monoterpenoid indole alkaloids. European Journal of Biochemistry., 269, 2204–2213.

Geu-Flores, F., Sherden, N. H., Courdavault, V., Burlat, V., Glenn, W. S., Wu, C., Nims, E., Cui, Y., O'Connor, S. E. (2012) An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. Nature., 492, 138–142.

Giddings, LA., Liscombe, DK., Hamilton, JP., Childs, KL., Dellapenna, D., Buell, CR., O'Connor SE. (2011) A stereoselective hydroxylation step of alkaloid biosynthesis by a unique cytochrome P450 in Catharanthus roseus. The Journal of Biological Chemistry., 286,16751-1677.

Gigant, B., Wang, C., Ravelli, R. B., Roussi, F., Steinmetz, M. O., Curmi, P. A., Knossow, M. (2005) Structural basis for the regulation of tubulin by vinblastine. Nature., 435, 519-522.

Giglioli-Guivarc'h, N. (2013) New Light on Alkaloid Biosynthesis and Future Prospects. Academic Press. 68.

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Venter, J. C. (2006) Essential genes of a minimal bacterium. Proceedings of the National Academy of Sciences., 103, 425-430.

Góngora-Castillo, E., Childs, K.L., Fedewa, G., Hamilton, J.P., Liscombe, D.K., Magallanes-Lundback, M., Mandadi, K.K., Nims, E., Runguphan, W., Vaillancourt, B., Varbanova-Herde, M., DellaPenna, D., McKnight, T.D., O'Connor, S., Buell, C.R. (2012) Development of Transcriptomic Resources for Interrogating the Biosynthesis of Monoterpene Indole Alkaloids in Medicinal Plant Species., PloS one, 7, e52506.

Gregory, R. K., Smith, I. E. (2000) Vinorelbine–a clinical review. British Journal of Cancer., 82, 1907-1913.

Grellier, P., Sinou, V., Garreau-de Loubresse, N., Bylen, E., Boulard, Y., Schrevel, J. (1999) Selective and reversible effects of vinca alkaloids on Trypanosoma cruzi epimastigote forms: blockage of cytokinesis without inhibition of the organelle duplication. Cell Motility and the Cytoskeleton., 42, 36-47.

Guéritte, F., Bac, N. V., Langlois, Y., Potier, P. (1980) Biosynthesis of antitumour alkaloids from Catharanthus roseus. Conversion of 20′-deoxyleurosidine into vinblastine. Journal of the Chemical Society Chemical Communications., 10, 452-453.

Guirimand, G., Burlat, V., Oudin, A., Lanoue, A., St-Pierre, B., Courdavault, V. (2009) Optimization of the transient transformation of Catharanthus roseus cells by particle bombardment and its application to the subcellular localization of hydroxymethylbutenyl 4-diphosphate synthase and geraniol 10-hydroxylase. Plant Cell Reports., 28, 1215-1234.

Guirimand, G., Courdavault, V., Lanoue, A., Mahroug, S., Guihur, A., Blanc, N., Giglioli-Guivarc'h, N., St-Pierre, B., Burlat, V. (2010) Strictosidine activation in Apocynaceae: towards a "nuclear time bomb"? BMC Plant Biology., 10, 182.

Guirimand, G. (2011) Organisation cellulaire et subcellulaire de la voie de biosynthèse des alcaloïdes indoliques monoterpéniques de Catharantus roseus. PhD Thesis, Tours.

Guirimand, G., Guihur, A., Ginis, O., Poutrain, P., Héricourt, F., Oudin, A., Lanoue, A., St-Pierre, B., Burlat, V., Courdavault, V. (2011a) The subcellular organization of strictosidine biosynthesis in Catharanthus roseus epidermis highlights several trans-tonoplast translocations of intermediate metabolites. FEBS Journal., 278, 749–63.

Guirimand, G., Guihur, A., Phillips, M.A., Oudin, A., Glevarec, G., Mahroug, S., Melin, C., Papon, N., Clastre, M., Giglioli-Guivarc'h, N., St-Pierre, B., Rodriguez-Concepcion, M., Burlat, V., Courdavault, V. (2012) Triple subcellular targeting of isopentenyl diphosphate isomerases encoded by a single gene. Plant Signal Behaviour., 7, 1495-1497.

H

Hagel, J. M., & Facchini, P. J. (2013) Benzylisoquinoline alkaloid metabolism–a century of discovery and a brave new world. Plant and Cell Physiology., 54.,647-672.

Hallard, D. (2000) Transgenic plant cells for the production of indole alkaloids. PhD Thesis, Leiden 2000.

Han, M., Heppel, S. C., Su, T., Bogs, J., Zu, Y., An, Z., Rausch, T. (2013) Enzyme inhibitor studies reveal complex control of methyl-D-erythritol 4-phosphate (MEP) pathway enzyme expression in Catharanthus roseus. PloS one., 8, e62467.

Hasemann, C. A., Kurumbail, R. G., Boddupalli, S. S., Peterson, J. A., Deisenhofer, J. (1995) Structure and function of cytochromes P450: a comparative analysis of three crystal structures. Structure, 3, 41-62.

Harborne, J. B. (1999) Classes and functions of secondary products from plants. Chemicals from Plants., 1-25.

Harborne, J. B., and Baxter, H. (1999) The handbook of natural flavonoids. John Wiley & Sons, Volume 1 and Volume 2.

Hemscheidt, T. and Zenk, M. H. (1985) Partial purification and characterization of a NADPH dependent tetrahydroalstonine synthase from Catharanthus roseus cell suspension cultures. Plant Cell Reports., 4, 216 – 219.

Hileman, L. C., Drea, S., de Martino, G., Litt, A., Irish, V. F. (2005) Virus-induced gene silencing is an effective tool for assaying gene function in the basal eudicot species Papaver somniferum (opium poppy). The Plant Journal., 44, 334–341.

Hong, S.-B., Hughes, E. H., Shanks, J. V, San, K.-Y. & Gibson, S. I. (2003) Role of the non-mevalonate pathway in indole alkaloid production by Catharanthus roseus hairy roots. Biotechnology Progress., 19, 1105–1108.

I

Irmler, S., Schröder, G., St-Pierre, B., Crouch, N. P., Hotze, M., Schmidt, J., Strack, D., Matern, U., Schröder, J. (2000) Indole alkaloid biosynthesis in Catharanthus roseus: new enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. Plant Journal., 24, 797–804.

J

Jacrot, M., Riondel, J., Picot, F., Leroux, D., Mouriquand, C., Beriel, H., Potier, P. (1983) Action du taxol vis-a-vis de tumeurs humaines transplantees sur des souris athymiques. Comptes rendus des séances de l'Académie des sciences. Série 3, Sciences de la Vie., 297, 597-600.

Jensen, K., and Møller, B. L. (2010) Plant NADPH-cytochrome P450 oxidoreductases. Phytochemistry., 71, 132-141.

Jordan, A., Hadfield, J. A., Lawrence, N. J., McGown, A. T. (1998) Tubulin as a target for anticancer drugs: agents which interact with the mitotic spindle. Medicinal research reviews., 18, 259-296.

Jörnvall, H., Hedlund, J., Bergman, T., Oppermann, U., Persson, B. (2010) Superfamilies SDR and MDR: from early ancestry to present forms. Emergence of three lines, a Zn-metalloenzyme, and distinct variabilities. Biochemical and Biophysical Research Communications., 396, 125-130.

Jörnvall, H., Landreh, M., Östberg, L. J. (2015) Alcohol dehydrogenase, SDR and MDR structural stages, present update and altered era. Chemico-biological interactions, 234, 75-79.

K

Kavanagh, K. L., Jörnvall, H., Persson, B., Oppermann, U. (2008) Medium-and short-chain dehydrogenase/reductase gene and protein families. Cellular and Molecular Life Sciences., 65, 3895-3906.

Kellner, F., Geu-Flores, F., Sherden, N. H., Brown, S., Foureau, E., Courdavault, V., O'Connor, S. E. (2015) Discovery of a P450-catalyzed step in vindoline biosynthesis: A link between the aspidosperma and eburnamine alkaloids. Chemical Communications., 51, 7626-7628.

Képès, F. (2011) Biologie synthétique et intégrative. Comptes Rendus Chimie., 14, 420-423.

Kim, Y. W., Kang, K. S., Kim, S. Y., Kim, I. S. (2000) Formation of fibrillar multimers of oat β-glucosidase isoenzymes is mediated by the As-Glu1 monomer. Journal of Molecular Biology., 303, 831-842.

Knight, T. (2003) Idempotent vector design for standard assembly of biobricks. Massachusetts Institute of Technology Synthetic Biology Working Group Technical Report 0.

Kolewe, M. E., Gaurav, V., Roberts, S. C. (2008) Pharmaceutically active natural product synthesis and supply via plant cell culture technology. Molecular Pharmaceutics., 5, 243-256.

Keasling, J. D. (2012) Synthetic biology and the development of tools for metabolic engineering. Metabolic Engineering., 14, 189-195.

Kurata, A., Kurihara, T., Kamachi, H., Esaki, N. (2005) 2-Haloacrylate reductase, a novel enzyme of the medium chain dehydrogenase/reductase superfamily that catalyzes the reduction of a carbon-carbon double bond of unsaturated organohalogen compounds. Journal of Biological Chemistry., 280, 20286-20291.

Kutchan, T. M., Hampp, N., Lottspeich, F., Beyreuther, K., Zenk, M. H. (1988) The cDNA clone for strictosidine synthase from Rauvolfia serpentina DNA sequence determination and expression in Escherichia coli. FEBS Letters., 237, 40–44.

Kwon, D. H., Kim, M. D., Lee, T. H., Oh, Y. J., Ryu, Y. W., Seo, J. H. (2006) Elevation of glucose 6-phosphate dehydrogenase activity increases xylitol production in recombinant Saccharomyces cerevisiae. Journal of molecular catalysis B: Enzymatic., 43, 86-89

L

Laflamme, P., St-Pierre, B., De Luca, V. (2001) Molecular and biochemical analysis of a Madagascar periwinkle root-specific minovincinine-19-hydroxy-O-acetyltransferase. Plant Physiology., 125, 189-198.

Larkin, P., and Harrigan, G. G. (2007) Opportunities and surprises in crops modified by transgenic technology: metabolic engineering of benzylisoquinoline alkaloid, gossypol and lysine biosynthetic pathways. Metabolomics., 3, 371-382.

Lenihan, J. R., Tsuruta, H., Diola, D., Renninger, N. S., Regentin, R. (2008) Developing an industrial artemisinic acid fermentation process to support the cost-effective production of antimalarial artemisinin-based combination therapies. Biotechnology progress., 24, 1026-1032.

Levac, D., Murata, J., Kim, W. S., De Luca, V. (2008) Application of carborundum abrasion for investigating the leaf epidermis: molecular cloning of Catharanthus roseus 16-hydroxytabersonine-16-O-methyltransferase. Plant Journal., 53, 225–36.

Lichtenthaler, H.K., Schwender, J., Disch, A., Rohmer, M. (1997) Biosynthesis of isoprenoids in higher plant chloroplasts proceeds via a mevalonate-independent pathway. FEBS Letters., 400, 271-274.

Liscombe, D. K., Usera, A. R., O'Connor, S. E. (2010) Homolog of tocopherol C methyltransferases catalyzes N methylation in anticancer alkaloid biosynthesis. Proceedings of the National Academy of Sciences., 107, 18793–18798.

Liu, Y., Schiff, M., Marathe, R., Dinesh-Kumar, S. P. (2002) Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. The Plant Journal., 30, 415–429.

Look, S. A., Fenical, W., Jacobs, R. S., Clardy, J. (1986) The pseudopterosins: anti-inflammatory and analgesic natural products from the sea whip Pseudopterogorgia elisabethae. Proceedings of the National Academy of Sciences., 83, 6238-6240.

Lu, R., Martin-Hernandez, A. M., Peart, J. R., Malcuit, I., Baulcombe, D. C. (2003) Virus-induced gene silencing in plants. Methods., 30, 296–303.

Luijendijk, T. J. C., Meijden, E., Verpoorte, R. (1996) Involvement of strictosidine as a defensive chemical in Catharanthus roseus. Journal of Chemical Ecology., 22, 1355–1366.

Luijendijk, T. J. C., Stevens, L. H., Verpoorte, R. (1998) Purification and characterisation of strictosidine β-d-glucosidase from Catharanthus roseus cell suspension cultures. Plant Physiology and Biochemistry., 36, 419–425.

M

Maeda, H., Dudareva, N. (2012) The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants. Annual Review of Plant Biology., 63, 73-105.

Mahroug, S., Courdavault, V., Thiersault, M., St-Pierre, B., Burlat, V. (2006) Epidermis is a pivotal site of at least four secondary metabolic pathways in Catharanthus roseus aerial organs. Planta., 223, 1191-1200.

Maresh, J. J., Giddings, L.-A., Friedrich, A., Loris, E. A., Panjikar, S., Trout, B. L., Stöckigt, J., Peters, B., O'Connor, S. E. (2008) Strictosidine synthase: mechanism of a Pictet-Spengler catalyzing enzyme. Journal of the American Chemical Society., 130, 710–723.

Meijer, A. H., Lopes Cardoso, M. I., Voskuilen, J. T., de Waal, A., Verpoorte, R., Hoge, J. H. (1993b) Isolation and characterization of a cDNA clone from Catharanthus roseus encoding NADPH:cytochrome P-450 reductase, an enzyme essential for reactions catalysed by cytochrome P-450 mono-oxygenases in plants. Plant Journal., 4, 47–60.

Meshnick, S. R., Taylor, T. E., Kamchonwongpaisan, S. (1996) Artemisinin and the antimalarial endoperoxides: from herbal remedy to targeted chemotherapy. Microbiological Reviews., 60, 301-315.

Miettinen, K., Dong, L., Navrot, N., Schneider, T., Burlat, V., Pollier, J., Woittiez, L., van der Krol, S., Lugan, R., Ilc, T., Verpoorte, R., Oksman-Caldentey, K.M., Martinoia, E., Bouwmeester, H., Goossens, A., Memelink, J., Werck-Reichhart, D. (2014) The seco-iridoid pathway from Catharanthus roseus. Nature Communication., 5, 1-11.

Mizutani, M., and Sato, F. (2011) Unusual P450 reactions in plant secondary metabolism. Archives of Biochemistry and Biophysics., 507, 194-203.

Mostofa, M., Choudhury, M. E., Hossain, M. A., Islam, M. Z., Islam, M. S., Sumon, M. H. (2007) Antidiabetic effects of Catharanthus roseus, Azadirachta indica, Allium sativum and glimepride in experimentally diabetic induced rat. Bangladesh Journal of Veterinary Medicine., 5, 99-102.

Mujib, A., Ilah, A., Aslam, J., Fatima, S., Siddiqui, Z. H., Maqsood, M. (2012) Catharanthus roseus alkaloids: application of biotechnology for improving yield. Plant Growth Regulation., 68, 111-127.

Munkert, J., Pollier, J., Miettinen, K., Van Moerkercke, A., Payne, R., Müller-Uri, F., Burlat, V., O'connor, S.E., Memelink, J., Kreis, W., Goossens, A. (2015) Iridoid synthase activity is common among the plant progesterone 5β-reductase family. Molecular Plant., 8, 136-152.

Munro, A. W., Girvan, H.M., Mason, A. E., Dunford, A. J., McLean, K. J. (2013) What makes a P450 tick? Trends in Biochemical Sciences., 38, 140-150.

Murata, J. and Luca, V. De. (2005) Localization of tabersonine 16-hydroxylase and 16-OH tabersonine-16-O-methyltransferase to leaf epidermal cells defines them as a major site of precursor biosynthesis in the vindoline pathway in Catharanthus roseus. Plant Journal., 44, 581–594.

Murata, J., Roepke, J., Gordon, H., De Luca, V. (2008) The leaf epidermome of Catharanthus roseus reveals its biochemical specialization. Plant Cell., 20, 524–542.

N

Naaranlahti, T., Auriola, S., Lapinjoki, S. P. (1991) Growth-related dimerization of vindoline and catharanthine in Catharanthus roseus and effect of wounding on the process. Phytochemistry., 30, 1451–1453.

Nammi, S., Boini, M. K., Lodagala, S. D., Behara, R. B. S. (2003) The juice of fresh leaves of Catharanthus roseus Linn. reduces blood glucose in normal and alloxan diabetic rabbits. BMC Complementary and Alternative Medicine., 3, 4-10.

Nelson, D. R., Zeldin, D. C., Hoffman, S. M. G., Maltais, L. J., Wain, H. M., Nebert, D. W. (2004b) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics and Genomics., 14, 1-18.

Nelson, D. R. (2011) Progress in tracing the evolutionary paths of cytochrome P450. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics., 1814, 14-18.

Newman, D. J., and Cragg, G. M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. Journal of Natural Products., 75, 311-335.

Nielsen, M. L., De Jongh, W. A., Meijer, S. L., Nielsen, J., Mortensen, U. H. (2007) Transient marker system for iterative gene targeting of a prototrophic fungus. Applied and Environmental Microbiology., 73, 7240-7245.

O

De Montellano, P. R. O. (2005) Cytochrome P450: structure, mechanism, and biochemistry. 2nd edition, Plenum, New York, 1995.

Oudin, A., Mahroug, S., Courdavault, V., Hervouet, N., Zelwer, C., Rodríguez-Concepción, M., St-Pierre, B., Burlat, V. (2007) Spatial distribution and hormonal regulation of gene products from methyl erythritol phosphate and monoterpene-secoiridoid pathways in Catharanthus roseus. Plant Molecular Biology., 65, 13–30.

P

Paddon, C. J., and Keasling, J. D. (2014) Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. Nature Reviews Microbiology., 12, 355-367.

Pasquali, G., Goddijn, O. J. M., Waal, A., Verpoorte, R., Schilperoort, R. A., Hoge, J. H. C., Memelink, J. (1992) Coordinated regulation of two indole alkaloid biosynthetic genes from Catharanthus roseus by auxin and elicitors. Plant Molecular Biology., 18, 1121–1131.

Pennings, E. J., van den Bosch, R. A., van der Heijden, R., Stevens, L. H., Duine, J. A., Verpoorte, R. (1989) Assay of strictosidine synthase from plant cell cultures by high-performance liquid chromatography. Analytical Biochemistry., 176, 412–415.

Phillips, MA., D'Auria, JC., Gershenzon, J., Pichersky, E. (2008b) The Arabidopsis thaliana type I isopentenyl diphosphate isomerases are targeted to multiple subcellular compartments and have overlapping functions in isoprenoid biosynthesis. Plant Cell., 20, 677-696.

Q

Qu, Y., Easson, M. L., Froese, J., Simionescu, R., Hudlicky, T., De Luca, V. (2015) Completion of the seven-step pathway from tabersonine to the anticancer drug precursor vindoline and its assembly in yeast. Proceedings of the National Academy of Sciences., 112, 6224-6229.

R

Rai, A., Smita, S.S., Singh, A.K., Shanker, K., Nagegowda, D.A. (2013) Heteromeric and Homomeric Geranyl Diphosphate Synthases from Catharanthus roseus and Their Role in Monoterpene Indole Alkaloid Biosynthesis. Molecular Plant., 6, 1531-1549.

Renault, H., Bassard, J. E., Hamberger, B., Werck-Reichhart, D. (2014) Cytochrome P450-mediated metabolic engineering: current progress and future challenges. Current opinion in plant biology., 19, 27-34.

Riou, C., Tourte, Y., Lacroute, F., Karst, F. (1994) Isolation and characterization of a cDNA encoding Arabidopsis thaliana mevalonate kinase by genetic complementation in yeast. Gene., 148, 293–297.

Ro, D. K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. A., Eachus, R. A., Ham, T. S., Kirby, J., Chang, M. C., Withers, S. T., Shiba, Y., Sarpong, R., Kiesling, J. D. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature., 440, 940-943.

Roepke, J., Salim, V., Wu, M., Thamm, A. M. K., Murata, J., Ploss, K., Boland, W., De Luca, V. (2010) Vinca drug components accumulate exclusively in leaf exudates of Madagascar periwinkle. Proceedings of the National Academy of Sciences., 107, 15287–15292.

Rohmer, M., Knani, M., Simonin, P., Sutter, B., Sahm, H. (1993) Isoprenoid biosynthesis in bacteria: a novel pathway for the early steps leading to isopentenyl diphosphate. Biochemistry Journal., 295, 517–524.

Rouilly, V., Canton, B., Nielsen, P., Kitney, R. (2007) Registry of BioBricks models using CellML. BMC Systems Biology., 1, 1-2.

Ruckpaul, K., and Rein, H. (1984) Cytochrome P450. Berlin, Akademie-Verlag.

S

Salim, V., Yu, F., Altarejos, J., Luca, V. (2013) Virus-induced gene silencing identifies Catharanthus roseus 7-deoxyloganic acid-7-hydroxylase, a step in iridoid and monoterpene indole alkaloid biosynthesis. The Plant Journal., 76, 754-765.

Saito, K. (2013) Phytochemical genomics a new trend. Current Opinion in Plant Biology., 16, 373-380.

Sauer, B., and Henderson, N. (1988) Site-specific DNA recombination in mammalian cells by the Cre recombinase of bacteriophage P1. Proceedings of the National Academy of Sciences., 85, 5166-5170.

Schenkman, J. B., and Jansson, I. (2003) The many roles of cytochrome b 5. Pharmacology and therapeutics., 97, 139-152.

Schröder, G., Unterbusch, E., Kaltenbach, M., Schmidt, J., Strack, D., De Luca, V., Schröder, J. (1999) Light-induced cytochrome P450-dependent enzyme in indole alkaloid biosynthesis: tabersonine 16-hydroxylase11. FEBS Letters., 458, 97–102.

Schuler, M. A., and Werck-Reichhart, D. (2003) Functional genomics of P450s. Annual Review of Plant Biology., 54, 629-667.

Schwarz, D. (1991) Rotational motion and membrane topology of the microsomal cytochrome P450 systemas analyzed by saturation transfer EPR. Frontiers in Biotransformation., 5, 94-137.

Senthil-Kumar, M., and Mysore, K. S. (2011) New dimensions for VIGS in plant functional genomics. Trends in Plant Science., 16, 656–665.

Simkin, A.J., Guirimand, G., Papon, N., Courdavault, V., Thabet, I., Ginis, O., Bouzid, S., Giglioli-Guivarc'h, N., Clastre, M. (2011) Peroxisomal localisation of the final steps of the mevalonic acid pathway in planta. Planta., 234, 903-914.

Simkin, A.J., Miettinen, K., Claudel, P., Burlat, V., Guirimand, G., Courdavault, V., Papon, N., Meyer, S., Godet, S., St-Pierre, B., Giglioli-Guivarc'h, N., Fischer, M.J.C., Memelink, J., Clastre, M. (2013) Characterization of the plastidial geraniol synthase from Madagascar periwinkle which initiates the monoterpenoid branch of the alkaloid pathway in internal phloem associated parenchyma. Phytochemistry., 85, 36-43.

Sottomayor, M., López-Serrano, M., DiCosmo, F., Ros Barceló, A. (1998) Purification and characterization of α-3′,4′-anhydrovinblastine synthase (peroxidase-like) from Catharanthus roseus (L.) G. Don. FEBS Letters., 428, 299–303.

Sottomayor, M., and Barceló, A. R. (2003) Peroxidase from Catharanthus roseus (L.) G. Don and the biosynthesis of α-3′, 4′-anhydrovinblastine: a specific role for a multifunctional enzyme. Protoplasma., 222, 97-105.

Stanley Fernandez, S. M., Kellogg, B. A., Poulter, C. D. (2000) Farnesyl diphosphate synthase. Altering the catalytic site to select for geranyl diphosphate activity. Biochemistry., 39, 15316-15321.

Stavrinides, A., Tatsis, E. C., Foureau, E., Caputi, L., Kellner, F., Courdavault, V., O'Connor, S. E. (2015) Unlocking the diversity of alkaloids in Catharanthus roseus: nuclear localization suggests metabolic channeling in secondary metabolism. Chemistry and Biology., 22, 336-341.

Stephanopoulos, G. (1999). Metabolic fluxes and metabolic engineering. Metabolic Engineering., 1, 1-11.

Sternberg, N., Sauer, B., Hoess, R., Abremski, K. (1986) Bacteriophage P1 cre gene and its regulatory region: evidence for multiple promoters and for regulation by DNA methylation. Journal of Molecular Biology., 187, 197-212.

St-Pierre, B., and De Luca, V. (1995) A cytochrome P-450 monooxygenase catalyzes the first step in the conversion of tabersonine to vindoline in Catharanthus roseus. Plant Physiology., 109, 131-139.

St-Pierre, B., Laflamme, P., Alarco, A.-M., D, V., Luca, E. (1998) The terminal O-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer. Plant Journal., 14, 703–713.

St-Pierre, B. (1999) Multicellular Compartmentation of Catharanthus roseus Alkaloid Biosynthesis Predicts Intercellular Translocation of a Pathway Intermediate. Plant Cell Online., 11, 887–900.

Strommer, J. (2011) The plant ADH gene family. The Plant Journal., 66, 128-142.

Szczebara, F. M., Chandelier, C., Villeret, C., Masurel, A., Bourot, S., Duport, C., Cauet, G. (2003) Total biosynthesis of hydrocortisone from a simple carbon source in yeast. Nature Biotechnology., 21, 143-149.

T

Teoh, K. H., Polichuk, D. R., Reed, D. W., Covello, P. S. (2009) Molecular cloning of an aldehyde dehydrogenase implicated in artemisinin biosynthesis in Artemisia annua This paper is one of a selection of papers published in a Special Issue from the National Research Council of Canada-Plant Biotechnology Institute. Botany., 87, 635-642.

Tholl, D., Kish, C. M., Orlova, I., Sherman, D., Gershenzon, J., Pichersky, E., & Dudareva, N. (2004) Formation of monoterpenes in Antirrhinum majus and Clarkia breweri flowers involves heterodimeric geranyl diphosphate synthases. The Plant Cell., 16, 977-992.

Tsay, Y. H. and Robinson, G. W. (1991) Cloning and characterization of ERG8, an essential gene of Saccharomyces cerevisiae that encodes phosphomevalonate kinase. Molecular and Cellular Biology., 11, 620–631.

Tu, Y. (2011) The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. Nature medicine., 17, 1217-1220.

Tyler, VE. (1994) Herbs of choice: the therapeutic use of phytomedicinals.Pharmaceutical Product Press.,imprint of Haworth Press, Inc.

V

Van Dam, N. M., Meijden, E., Verpoorte, R. (1993) Induced responses in three alkaloid-containing plant species. Oecologia., 95, 425–430.

Van der Heijden, R., Louwe, C. L., Verhey, E. R., Harkes, P. A., Verpoorte, R. (1989) Characterization of a Suspension Culture of Tabernaemontana elegans on Growth, Nutrient Uptake, and Accumulation of Indole Alkaloids. Planta Medica., 55, 158–62.

Van der Heijden, R., Jacobs, D.I., Snoeijer, W., Hallard, D., Verpoorte, R. (2004) The Catharanthus Alkaloids: Pharmacognosy and Biotechnology. Current Medicinal Chemistry., 11, 1241-1253.

Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G.J.E., Rombauts, S., Hasnain, G., Rischer, H., Memelink, J., Oksman-Caldentey, K.M., Goossens, A. (2013) CathaCyc, a Metabolic Pathway Database Built from Catharanthus roseus RNA-Seq Data. Plant Cell Physiology., 54, 673-685.

Vazquez-Flota, F., De Carolis, E., Alarco, A. M., De Luca, V. (1997) Molecular cloning and characterization of desacetoxyvindoline-4-hydroxylase, a 2-oxoglutarate dependent-dioxygenase involved in the biosynthesis of vindoline in Catharanthus roseus (L.) G. Don. Plant Molecular Biology., 34, 935–948.

Vázquez-Flota, F., Carrillo-Pech, M., Minero-García, Y., De Lourdes Miranda-Ham, M. (2004) Alkaloid metabolism in wounded Catharanthus roseus seedlings. Plant Physiology and Biochemistry 42, 623–628.

Verpoorte, R. (1998) Exploration of nature's chemodiversity: the role of secondary metabolites as leads in drug development. Drug Discovery Today., 3, 232–238.

W

Wang, M., Roberts, D. L., Paschke, R., Shea, T. M., Masters, B. S., Kim J. J. (1997) Three-dimensional structure of NADPH-cytochrome P450 reductase: prototype for FMN- and FAD-containing enzymes. Proceedings of the National Academy of Sciences USA., 94, 8411–8416.

Weathers, P. J., Elkholy, S., Wobbe, K. K. (2006) Artemisinin: the biosynthetic pathway and its regulation in Artemisia annua, a terpenoid-rich species. In Vitro Cellular and Developmental Biology-Plant., 42, 309-317.

Werck-Reichhart, D. and Feyereisen, R. (2000) Cytochromes P450: a success story. Genome Biology., 1, 1–9.

Werck-Reichhart, D., Hehn, A., Didierjean, L. (2000) Cytochromes P450 for engineering herbicide tolerance. Trends in Plant Science., 5, 116-123.

Westekemper, P., Wieczorek, U., Gueritte, F., Langlois, N., Langlois, Y., Potier, P., Zenk, M. (1980) Radioimmunoassay for the Determination of the Indole Alkaloid Vindoline in Catharanthus. Planta Medica., 39, 24–37.

Westfall, P. J., Pitera, D. J., Lenihan, J. R., Eng, D., Woolard, F. X., Regentin, R., Fickes, S. (2012) Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. Proceedings of the National Academy of Sciences., 109, 111-118.

Wijekoon, C. P., and Facchini, P. J. (2011) Systematic knockdown of morphine pathway enzymes in opium poppy using virus-induced gene silencing. The Plant Journal., 69, 1052–1063.

Wink, M. (1999) Functions of plant secondary metabolites and their exploitation in biotechnology.

Woodward, R. B., Sondheimer, F., Taub, D., Heusler, K., McLamore, W. M. (1952) The Total Synthesis of Steroids1. Journal of the American Chemical Society., 74, 4223-4251.

World Health Organization. (2006) WHO briefing on Malaria Treatment Guidelines and artemisinin monotherapies. Geneva., WHO.

X

Xiao, M., Zhang, Y., Chen, X., Lee, E.J., Barber, C.J.S., Chakrabarty, R., Desgagné-Penix, I., Haslam, T.M., Kim, Y.B., Liu, E., MacNevin, G., Masada-Atsumi, S., Reed, D.W., Stout, J.M., Zerbe, P., Zhang, Y., Bohlmann, J., Covello, P.S., De Luca, V., Page, J.E., Ro, D.K., Martin, V.J.J., Facchini, P.J., Sensen, C.W. (2013) Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. Journal of Biotechnology., 166, 122-134.

Y

Yamazaki, Y. (2003) Camptothecin Biosynthetic Genes in Hairy Roots of Ophiorrhiza pumila: Cloning, Characterization and Differential Expression in Tissues and by Stress Compounds. Plant Cell Physiology., 44, 395–403.

Qu, Y., Easson, M. L., Froese, J., Simionescu, R., Hudlicky, T., De Luca, V. (2015) Completion of the seven-step pathway from tabersonine to the anticancer drug precursor vindoline and its assembly in yeast. Proceedings of the National Academy of Sciences., 112, 6224-6229.

Yang Y-T, Bennett GN, San K-Y. (1998) Genetic and metabolic engineering. Electron Journal of Biotechnology., 1, 20-21.

Yordanov, M., Dimitrova, P., Patkar, S., Falcocchio, S., Xoxi, E., Saso, L., Ivanovska, N. (2005) Ibogaine reduces organ colonization in murine systemic and gastrointestinal Candida albicans infections. Journal of Medical Microbiology., 54, 647-653.

Yu, F., and De Luca, V. (2013) ATP-binding cassette transporter controls leaf surface secretion of anticancer drug components in Catharanthus roseus. Proceedings of the National Academy of Sciences., 110, 15830-15835.

Z

Zenk, M. H. (1991) Chasing the enzymes of secondary metabolism: plant cell cultures as a pot of gold. Phytochemistry., 30, 3861-3863.

Zhang, Y. H. P., Myung, S., You, C., Zhu, Z., Rollin, J. A. (2011) Toward low-cost biomanufacturing through in vitro synthetic biology: bottom-up design. Journal of Materials Chemistry., 21, 18877-18886.

Ziegler, J., Facchini, P. J. (2008) Alkaloid biosynthesis: metabolism and trafficking. Annual Review Plant Biology., 59, 735–769.

Emilien FOUREAU

Elucidation de la voie de biosynthèse des alcaloïdes de *Catharanthus roseus* et ingénierie métabolique dans la levure

Résumé

Catharanthus roseus est une plante médicinale produisant divers types d'alcaloïdes indoliques monoterpéniques (AIM) d'intérêt en santé humaine. Ainsi, les AIM dimères comme la vinblastine et la vincristine sont utilisés en chimiothérapie anticancéreuse et les alcaloïdes monomères de type hétéroyohimbine présentent diverses activités pharmacologiques. La fabrication de ces molécules dans la plante est fort complexe. Elle requiert un haut niveau de compartimentation tissulaire et subcellulaire et met en jeu plus d'une trentaine d'étapes enzymatiques, dont certaines sont encore très mal connues. Dans ce contexte, l'objectif de la thèse a consisté à élucider plusieurs étapes enzymatiques de la voie de biosynthèse des AIM. Nos travaux ont permis de caractériser de nouvelles isoformes enzymatiques de la famille des cytochromes P450 ainsi que les réductases qui leur sont associées. Ils ont abouti à l'identification de nouvelles déshydrogénases et mis en évidence, *in planta*, leurs interactions avec la strictosidine synthase suggérant une biosynthèse orientée vers les divers alcaloïdes de type hétéroyohimbine. Enfin, en ayant recours à l'ingénierie métabolique, un segment de la voie de biosynthèse a été transféré dans la levure *Saccharomyces cerevisiae*, lui conférant la capacité de bio-transformer la tabersonine en vindoline, l'un des deux précurseurs finaux des alcaloïdes dimères.

Mots-clefs: *Catharanthus roseus*, alcaloïdes indoliques monoterpéniques, biosynthèse, compartimentation subcellulaire, ingénierie métabolique.

Abstract

Catharanthus roseus is a medicinal plant producing various types of monoterpene indole alkaloids (MIA) with a great interest in human health. Dimeric alkaloids such as vinblastine and vincristine are used in cancer chemotherapy and monomeric heteroyohimbine alkaloids exhibit various pharmacological activities. The production of these molecules in the plant is very complex. It requires a high level of tissular and subcellular compartmentalization and involves more than thirty enzymatic steps, some of which are largely unknown. In this context, the aim of this thesis was to elucidate several enzymatic steps of the MIA biosynthetic pathway. Our work allowed us to characterize new enzyme isoforms of cytochrome P450 and their associated reductases. They also resulted in the identification of new dehydrogenases and highlighted their interactions with the strictosidine synthase suggesting a directed biosynthesis towards various heteroyohimbine type of alkaloids. Finally, engineered yeast containing a segment of the MIA biosynthetic pathway was able to convert tabersonine into vindoline, one of the two final precursors of the dimeric alkaloids.

Keywords: *Catharanthus roseus*, monoterpene indole alkaloid, biosynthesis, subcellular compartmentalization, metabolic engineering