Genome and transcriptome guided gene discovery in plant secondary metabolism

Franziska Kellner

This thesis is submitted in fulfilment of the requirements of the degree of Doctor of Philosophy at the University of East Anglia

Department of Biological Chemistry

John Innes Centre

Norwich

June 2016

Abstract

Plants produce a wide range of complex secondary metabolites that have many applications, for example as pharmaceutical agents. Gene discovery and the elucidation of these unique biosynthetic pathways is challenging since many of the enzymatic transformations are unprecedented. In *Catharanthus roseus*, the sole producer of the valuable anti-cancer compounds vinblastine and vincristine, the biosynthetic pathway for these alkaloids is highly complex and crucial steps are still unknown. Recently, the tight transcriptional co-regulation of the early part of this pathway enabled discovery of some of the central enzymatic steps by analysing the gene co-expression patterns and testing potential candidates using virus induced gene silencing. Additionally, it has become apparent that some plant secondary metabolite pathways exhibit physical clustering of pathway related genes in the genome. This thesis highlights how both strategies of gene discovery can be applied for the targeted discovery of genes for missing steps in biosynthesis of non-model plants. Co-expression analysis to identify candidates and subsequent testing of these candidates using virus induced gene silencing has led to the discovery and subsequent characterisation of the enzyme tabersonine 3-oxygenase (T3O), a key oxidation step in vindoline biosynthesis. This thesis furthermore reports the first *C. roseus* whole genome sequence. Additionally a BAC library was obtained and selected BACs sequenced. Analysis of the combined sequencing data established that gene clustering does indeed occur for alkaloid biosynthesis in *C. roseus* and yielded a new set of candidates for so far unknown pathway enzymes. Selected candidates have been tested by silencing or expression and results are discussed. The sequence information provides a valuable resource for the wider community, available as a searchable, publically available database (http://medicinalplantgenomics.msu.edu/). The work presented in this thesis highlights how next generation sequence data can be exploited to elucidate complex secondary metabolic pathways.

Acknowledgements

I want to thank Prof. Sarah O'Connor for giving me the chance to come here and have an amazing four years. Finishing this work has been harder than I anticipated, but it was worth it. I thank you for your guidance, the freedom you gave me to develop, for never giving up on me and for your unconditional support. I cannot thank you enough!

Thanks to all the people along the way, Fernando and Hajo, Richard for being the best lab neighbour and for keeping me sane. Thanks to all the O'Connor's past, present and future, you are lucky!

Special thanks to Dr. Lionel Hill and Dr. Brande Wulff for excellent support and supervision. Thanks for productive and successful collaboration to Prof. Vincent Courdavault and his group at the University of Tours (France) and Prof. Robin Buell and her team at the Michigan State University (USA).

Thanks to everyone here at JIC/TSL/TGAC. From greenhouse to media kitchen, from computing to metabolomics, it was a pleasure and a privilege to work here!

Thanks to my friends and family for your patience and support.

For Eddie for being the greatest!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

This thesis uses standard abbreviations for nucleic acids (one letter code) and amino acids (one and three letter codes). Standard SI units are also employed unless otherwise stated.

16OMT	16-hydroxytabersonine O-methyltransferase
7DLGT	7-deoxyloganetic acid glucosyltransferase
7DLH	7-deoxy loganic acid hydroxylase
ACN	acetonitrile
AO	amyrin oxidase
AS	amyrin synthase
BAC	bacterial artificial chromosome
BBE	berberine bridge enzyme
BCA	bicinchoninic acid
BLAST	basic local alignment search tool
bp	base pair
BSA	bovine serum albumin
C4H	cinnamate 4-hydroxylase
cDNA	complementary DNA
ChlH	photoporphyrin IX magnesium chelatase subunit H
CrOMT	4'-O-methyltransferase
CS	chalcone synthase
D4H	desacetoxyvindoline 4-hydroxylase
DAT	deacetylvindoline acetyltransferase
DIBOA	2,4-dihydroxy-2H-1,4-benzoxazin-3(4H)-one
DIMBOA	2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one
DMAPP	dimethylallylpyrophosphate
DMSO	dimethyl sulfoxide
dNTPs	deoxynucleotide triphosphates
dsRNA	double stranded ribonucleic acid
DTT	dithiothreitol
DXR	1-deoxy-D-xyulose 5-phosphate reductoisomerase
DXS	1-deoxy-D-xyulose 5-phosphate synthase
EDTA	ethylenediaminetetraacetic acid
EtOH	ethanol
EV	empty vector
F3'5'H	flavonoid 3', 5'-hydroxylase
FPKM	fragments per kilobase of transcript per million reads mapped
G8H	geraniol-8-hydroxylase
Glc	glucose

GPP	geranyl pyrophosphate
GRO	8-hydroxygeraniol reductoisomerase
IO	iridoid oxidase
IPAP	internal phloem associated parenchyma
IPP	isopentyl pyrophosphate
ISY	iridoid synthase
IT-ToF	Ion trap time of flight
kDa	kilodalton
LAMT	loganic acid methyltransferase
LB	lysogeny broth
LC-MS	liquid chromatography mass spectrometry
MAT	minovincinine 19-hydroxy-O-acetyltransferase
MATE	multidrug and toxic compound extrusion
MeJA	methyl jasmonate
MeOH	methanol
MEP	2-C-methyl-D-erythritol 4-phosphate
MES	2-(N-morpholino)ethanesulfonic acid
MeV	multiexperiment viewer
MIA	monoterpene indole alkaloid
MPGR	medicinal plant genomics resource
MVA	mevalonate
NADPH	nicotinamide adenine dinucleotide phosphate (reduced)
NCBI	national centre for biotechnology information
NMR	Nuclear Magnetic Resonance
NMT	16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase
OSC	oxidosqualene cyclase
P450	cytochrome P450
P450red	cytochrome P450 reductase
PCR	polymerase chain reaction
PTGS	post transcriptional gene silencing
RNAi	ribonucleic acid interference
RO	reticulin oxidase
SGD	strictosidine β-glucosidase
siRNA	small interfering ribonucleic acid
SLS	secologanin synthase
ssRNA	single stranded ribonucleic acid
STR	strictosidine synthase
T16H	tabersonine 16-hydroxylase
T19H	tabersonine 19-hydroxylase
T3O	tabersonine 3-oxygenase
T3R	tabersonine 3-reductase
TDC	tryptophan decarboxylase
THCGT	tetrahydroxy chalcone 2'glycosyltransferase
TIC	total ion chromatogram
TLC	thin layer chromatography
TRV	tobacco rattle virus

USER	uracil specific excision reagent
VIGS	virus induced gene silencing
WGS	whole genome sequence
WT	wild type

1 Introduction

Natural products, also called secondary or specialised metabolites, unlike primary metabolites, do not directly contribute to essential growth and development of the organism. Instead these molecules have other key functions such as mediating the interaction between plants and their environment [1]. Additionally, many plant specialised metabolites have served humans for centuries as traditional medicines and still pose a unique resource for today's pharmaceutical industry (Figure 1). The staggering amount of not yet investigated plants and compounds promises a diverse and largely untapped space of discovery [2].

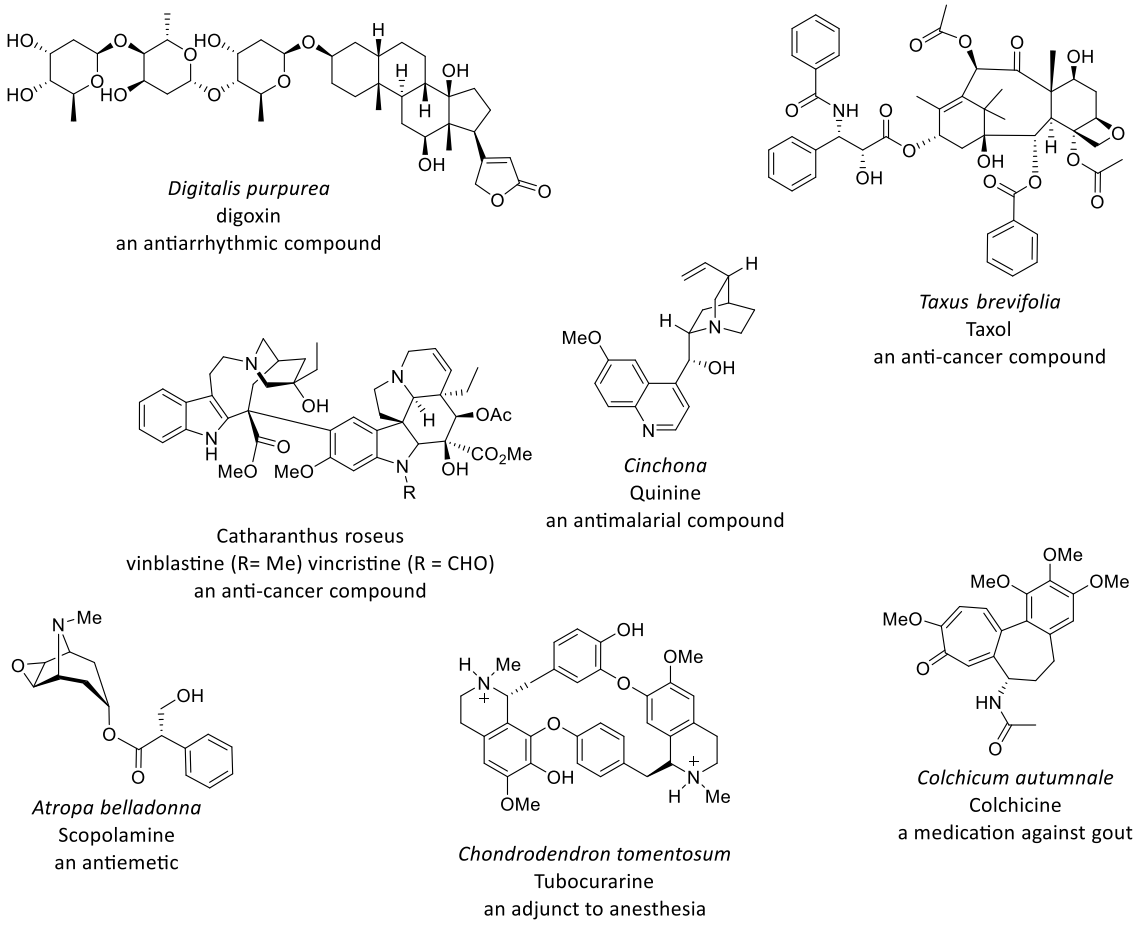


Figure 1 Selected plant derived bioactive specialised metabolites and medicinal application

Specialised metabolites are often structurally complex (Figure 1) making chemical synthesis for large-scale commercial production challenging. Efforts to alleviate supply issues in an economical way has led to research into alternative production systems including plant cell

cultures [3,4] and heterologous expression in microbial hosts [5]. The latter, however, requires a detailed understanding of the biosynthesis of the desired compound in the native plant. Most importantly, this approach requires the knowledge of which genes encode the biosynthetic enzymes responsible for the production of the target compounds.

Discovery of enzymes involved in a particular secondary metabolic pathway has in the past exploited many different approaches. Recently, the combination of metabolomic data with gene expression data has been shown to pose a powerful strategy for gene discovery and pathway elucidation [6]. This approach combines the knowledge of occurrence and distribution of target compounds in different plant tissues, whole plants or related plant species, with gene expression information. Correlating the presence of genes with the presence of the metabolite of interest can provide clues to identify candidate biosynthetic genes. A second widely used approach examines the genomic organisation of secondary metabolite biosynthesis. This strategy is based on the observation that in some plant secondary metabolite pathways the responsible genes are physically co-localised into a genomic cluster [7]. Notably, both of these approaches benefit greatly from increased availability of plant transcriptome and genome sequencing data due to decreasing sequencing costs [8].

To facilitate compilation of a short list of biosynthetic gene candidates from these data, the anticipated pathway chemistry is used to predict the enzyme class of the missing steps in metabolite biosynthesis. Yet since many plant-sourced biochemicals arise only from a narrow phylogenetic space, and involve unprecedented chemical transformations, making these predictions can be challenging. Investigation of the biosynthesis of a single metabolite or a specific group of plant-derived compounds therefore requires a detailed study of the individual metabolomic, transcriptomic and genomic data. Ultimately, progress in the dissection of the complex network of secondary metabolite biosynthesis helps not only the understanding of the production of a specific compound, but also delivers new tools and knowledge for discovery of new pathways and the manipulation of existing ones.

1.1 Monoterpene indole alkaloids of the medicinal plant *Catharanthus roseus*

1.1.1 Plant natural products

Plants produce a wide range of natural compounds with a multitude of different functions. While primary metabolites are essential for the survival of plants, secondary, also called specialised metabolites or phytochemicals, are non-essential. Nevertheless in many cases

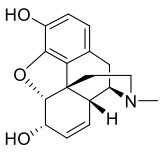
these natural products have an important function for plant survival and fitness by for example contributing to plant reproduction [9] and plant protection [10].

According to their biosynthetic origin, plant natural products can be grouped into classes, such as terpenoids, alkaloids and phenylpropanoids and other phenolic compounds. All natural products are derived from building blocks provided by primary metabolism such as acetyl coenzyme A, shikimic acid, mevalonic acid, amino acids and 1-deoxyxylulose 5-phosphate.

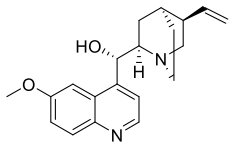
1.1.2 Alkaloids

More than 12000 of the over 50000 known natural products belong to the group of alkaloids and evidence for their use as medicine by humans reaches back more than 4000 years [11].

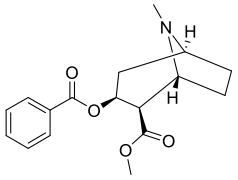
Most alkaloids are biosynthesised via complex enzymatic pathways often comprising 20 or more steps. Alkaloids are a heterogeneous class of secondary metabolites containing a basic nitrogen, and are usually derived from amino acids. These metabolites can be grouped into subclasses such as indole or quinoline alkaloids that are derived from tryptophan, benzyl isoquinoline alkaloids that are derived from tyrosine, piperidine alkaloids that are derived from lysine, pyrrolidine alkaloids that are derived from ornithine, phenylethylamine alkaloids that are derived from phenylalanine and imidazole alkaloids that are derived from histidine [12] (Figure 2).



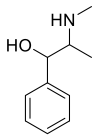
morphine
benzoisoquinoline alkaloid derived from tyrosine
analgesic



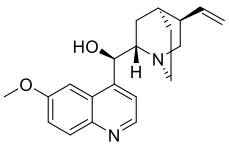
quinidine
quinoline alkaloid derived from tryptophan
antiarrhythmic



cocaine
tropane alkaloid derived from ornithine and arginine
local anaesthetic



ephedrine
phenylethylamine derived from phenylalanine
antiasthmatic



quinine
quinoline alkaloidderived from tryptophan
antimalarial

Figure 2 Selected plant derived alkaloids of different subclasses and medical application

Alkaloids frequently possess pronounced biological activities that play a role in, for example, defending plants against herbivores [13]. Additionally, many alkaloids are well known for their

poisonous and/ or medicinal attributes. The benzylisoquinoline alkaloid morphine can be used as an opiate type pain medication [14] while cocaine is a strong stimulant but can be used as a local anaesthetic [15]. Quinidine, a stereoisomer of the well-known antimalarial drug quinine, is an antiarrhythmic agent [16]. Ephedrine is structurally related to the amphetamines and known in Chinese traditional medicine since BC [17].

1.1.3 Monoterpene indole alkaloids

Approximately 3000 different monoterpene indole alkaloids (MIAs) can be found in Nature dispersed over six plant families, mainly the Apocynaceae, Loganiaceae, and Rubiaceae but also in the Icacinaceae, Nyssaceae, Alangiaceae familes. All MIAs originate from a convergent pathway that utilises tryptamine, provided by the amino acid tryptophan (shikimic acid pathway) and secologanin, a monoterpene moiety derived from geranyl-pyrophosphate (GPP) via the 2-*C*-methyl-D-erythritol 4-phosphate (MEP) pathway[18]. Many MIAs have been exploited for their potent pharmacological activities (Figure 3).

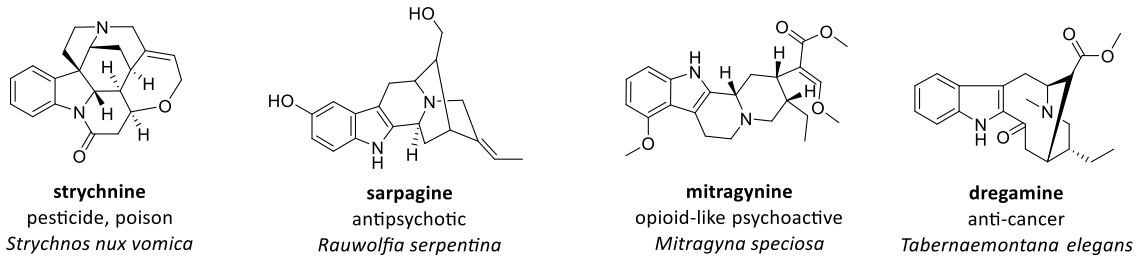


Figure 3 A selection of monoterpene indole alkaloids and their medicinal application

1.1.4 Catharanthus roseus

Catharanthus roseus or Madagascar periwinkle, a major producer of MIAs, belongs to the genus *Catharanthus* that comprises eight species of herbaceous perennial plants. Of these seven are endemic to the Island of Madagascar. *C. roseus* has long been used in traditional folk medicine.

C. roseus has glossy, dark green and oval leaves. The wild plant has a pale pink flower with a purple “eye”. It is a popular plant due to its moderate to high drought tolerance and ornamental value. Intensive breeding efforts have yielded a huge variety of different flower colours and growth shape (Figure 4).

Figure 4 Flowers and leaves of three *C. roseus* varieties

The *C. roseus* varieties "SunStorm Apricot" (left), "SantaFe" (middle) and "Little Bright Eyes" (right).

C. roseus displays a rich specialised metabolism, and is particularly well known for its alkaloid profile. *C. roseus* produces MIAs, which, like all MIAs, are produced from a central intermediate, strictosidine, that is derived from both tryptophan and the monoterpene secologanin (Figure 5) [18]. Many *C. roseus* derived alkaloids, such as ajmalicine, an antiarrhythmic, and serpentine, an antihypertensive, have been widely used as prescription drugs in the past [19]. However, intense, longstanding scientific interest in *C. roseus* is driven primarily by the discovery of the anti-cancer properties of its bisindole alkaloid metabolites vinblastine (Figure 1) and its oxidised congener vincristine. These compounds are bisindole alkaloids which form in *C. roseus* as a result of the dimerisation of the two MIAs catharanthine and vindoline. Vinblastine and vincristine were originally discovered when screening traditional medicinal herbs for their therapeutic properties in the early 1960s [19]. These alkaloids bind to microtubules and consequently arrest the cell cycle, which makes them powerful chemotherapeutic drugs, and are primarily used against Hodgkin's lymphoma and breast cancer. Semi-synthesis using vinblastine as a starting precursor has yielded vinorelbine [20] and vindesine [21], which are also now used in the clinic as anti-cancer agents. Further chemical modification of vinorelbine yield anti-cancer compounds such as vinflunine [22,23].

The demand for vinblastine and vincristine is still satisfied by extraction from *C. roseus* plants. However, the extraction of vinblastine and vincristine, or their precursors catharanthine and vindoline, from *C. roseus* plants is a laborious and costly task. Yet the complex chemical synthesis is even less economically viable, which explains the ongoing interest to either increase alkaloid production in *C. roseus* itself or to develop an alternative production system. Collectively, these efforts have resulted in *C. roseus* being one of today's best studied medicinal plants.

1.1.5 Strictosidine biosynthesis in Catharanthus roseus

More than 130 different MIAs have been reported to be produced by *C. roseus* [19]. All MIAs are derived from the central precursor molecule strictosidine that is formed as a condensation product of the terpenoid moiety secologanin (Figure 5A) and the indole moiety tryptamine (Figure 5B), via a Pictet-Spengler condensation (Figure 5C) [18,24]. Investigation into strictosidine biosynthesis in *C. roseus* over the past decades has led to the elucidation of all reactions, as well as establishing the order of all enzymatic steps involved in its production. This has consequently enabled the production of the early part of alkaloid biosynthesis, strictosidine biosynthesis, in *S. cerevisiae* (yeast) as an alternative host [25].

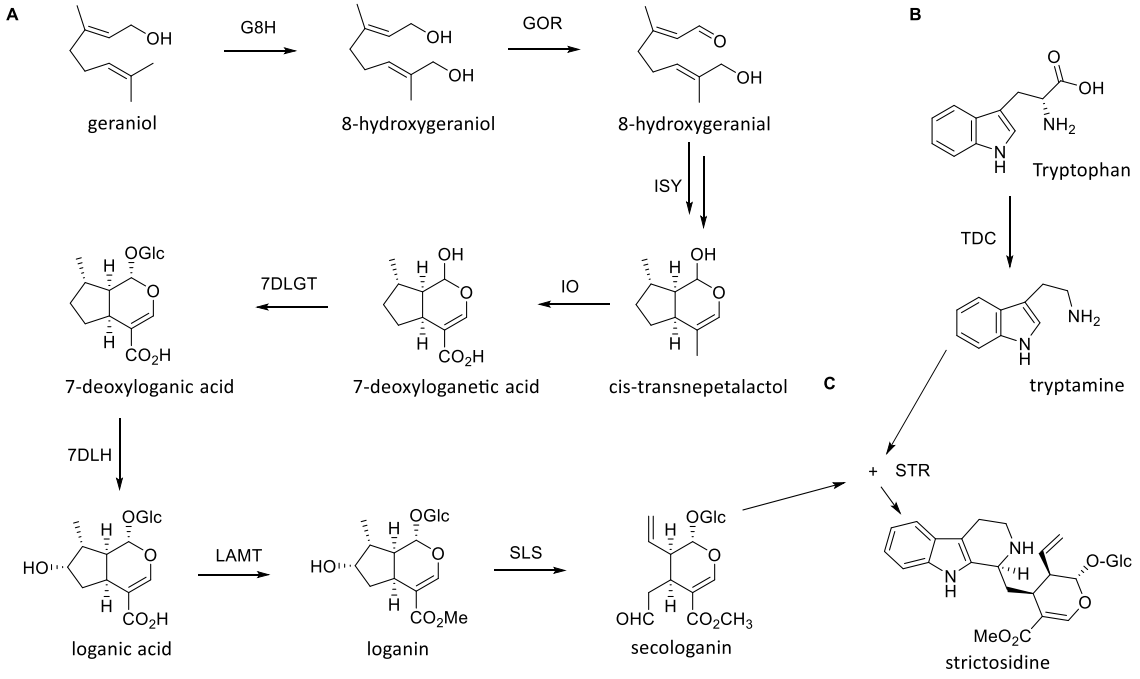


Figure 5 Biosynthesis of strictosidine precursor secologanin and tryptamine in *C. roseus*

A: The biosynthesis of secologanin in *C. roseus* proceeds from MEP pathway precursors IPP and DMAP that form GPP, with its first committed step being the formation of geraniol. The nine-step pathway comprises an oxidation by the enzymes geraniol 8 hydroxylase (G8H), an oxidation by 8-hydroxygeraniol oxidoreductase (GOR), cyclisation by iridoid synthase (ISY), another oxidation by iridoid oxidase (IO), a glycosylation by 7-deoxyloganetic acid glucosyltransferase (7DLGT), a hydroxylation by 7-deoxyloganic acid hydroxylase (7DLH) and subsequent methylation by loganic acid methyltransferase (LAMT) and a final oxidation by secologanin synthase (SLS). **B**: Tryptamine is produced via the decarboxylation of tryptophan by the enzyme tryptophan synthase (TDC). **C:** Strictosidine is formed by the action of strictosidine synthase (STR).

Secologanin production proceeds from the isoprenoids isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), which form geranyl-pyrophosphate (GPP). IPP and DMAP can be produced by two pathways the cytosolic mevalonate (MVA) and the non-mevalonate or plastidic methyl-D-erythritol 4-phosphate (MEP) pathway. [1-^{13}C]-glucose feeding of *C. roseus* cell suspension cultures has established that in *C. roseus* the MEP pathway is predominantly responsible for secologanin production [26].

GPP is then converted to geraniol by geraniol synthase [27]. Geraniol is hydroxylated by the cytochrome P450 geraniol-8 hydroxylase (G8H) to 8-hydroxygeraniol [28]. In combination with the action of 8-hydroxygeraniol oxidoreductase (GOR) the linear dialdehyde 8-oxogeranial is formed [29]. 8-oxogeranial is cyclised by iridoid synthase (ISY) [30]. Iridoid oxidase (IO), a second cytochrome P450 involved in secologanin production, oxidises cis-trans-nepetalactol to 7-deoxyloganetic acid [29,31]. This product is subsequently glycosylated by 7-deoxyloganetic acid glucosyltransferase (7DLGT) [32] and further hydroxylated by a third cytochrome P450, 7-deoxyloganic acid hydroxylase (7DLH) [32,33] and methylated by loganic acid methyltransferase (LAMT) to form loganin [34]. In a final step performed by the fourth cytochrome P450 involved in this pathway loganin is oxidatively cleaved by secologanin synthase (SLS) to yield the monoterpene secologanin [35] (Figure 5A).

In a parallel pathway tryptamine is produced. To yield tryptamine the amino acid tryptophan is decarboxylated by the pyridoxal phosphate dependent enzyme, tryptophan decarboxylase (TDC) [36] (Figure 5B). The condensation of secologanin and tryptamine to form the central metabolic intermediate, strictosidine is the first committed step in the formation of MIAs and represents a key branch point in this pathway [18].

1.1.6 Fate of deglycosylated strictosidine

The deglycosylation of strictosidine by strictosidine β-D-glycosidase (SGD) is the next step in alkaloid biosynthesis in *C. roseus* and other monoterpene indole alkaloid producing plants. Deglycosylated strictosidine rearranges to 4,21-dehydrogeissoschizine, which forms the structural basis for the diverse range of scaffolds occurring in this plant. In *C. roseus* the highly reactive intermediate is transformed by an unknown number of subsequent reactions that lead to the molecular rearrangements necessary for the formation of the three distinct molecular backbones or skeletons, representing the three different classes of *C. roseus*

alkaloids, namely the iboga, the corynanthe and the aspidosperma alkaloids (Figure 6). None of the biosynthetic genes responsible for the transformation of strictosidine aglycone to these scaffolds was known at the beginning of this work, though a biosynthetic gene responsible for the biosynthesis of the corynanthe skeleton has been identified in the last year [37].

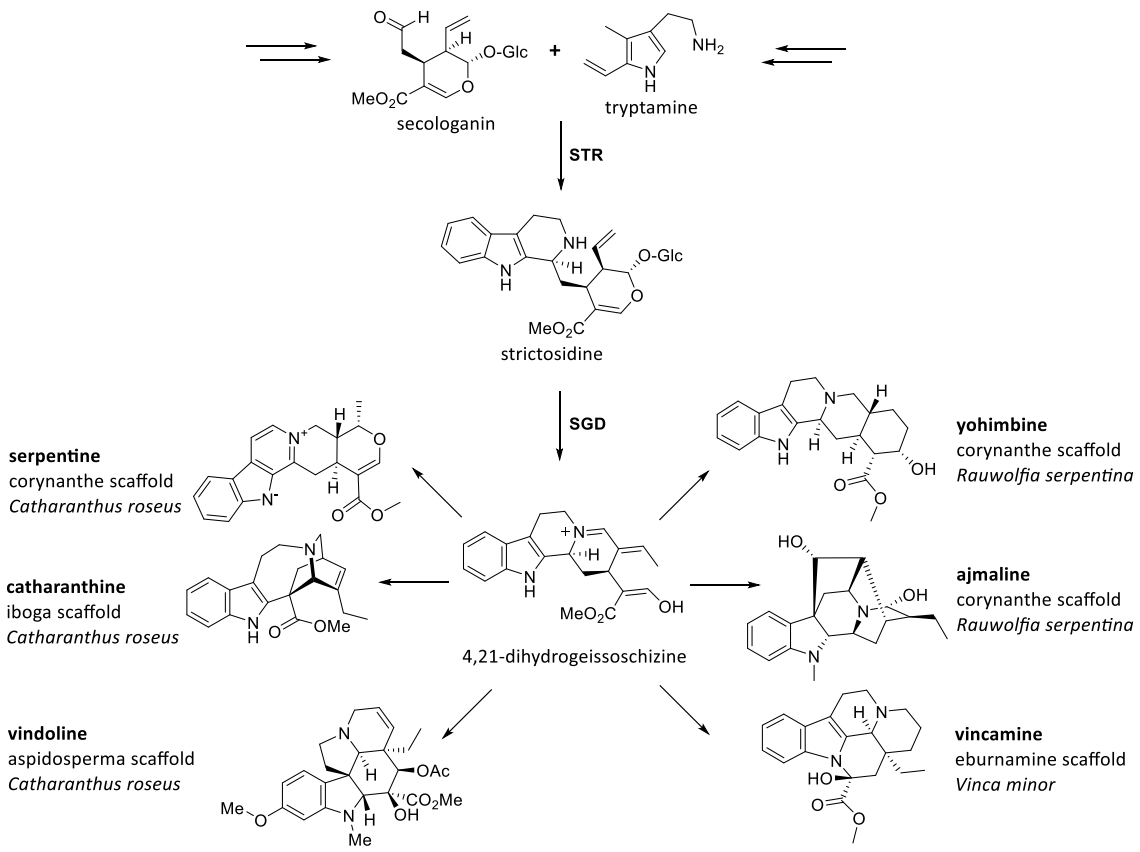


Figure 6 Examples of strictosidine derived scaffolds in *C. roseus* and related plants

Secologanin and tryptamine are condensed by strictosidine synthase (STR) to form strictosidine. Strictosidine is deglycosylated by strictosidine β-glucosidase (SGD) and rearranges to 4,21-dehydrogeissoschizine which forms the structural basis for a different of alkaloid backbones in multiple plant species. Strictosidine derived MIA scaffolds in *C. roseus* are represented by serpentine (corynanthe scaffold), catharanthine (iboga scaffold) and vindoline (aspidosperma scaffold), while *Rauwolfia serpentine* contains for example the MIA yohimbine (also corynanthe scaffold) and ajmaline (ajmalan scaffold). *Vinca minor*, a close relative of *C. roseus*, contains the MIA vincamine an eburnamine scaffold.

1.1.7 Vindoline biosynthesis in Catharanthus roseus

Vindoline is one of the precursors for vinblastine production, and is widely used in semi-synthesis of vinblastine by chemical coupling with catharanthine. Vindoline is derived from the simpler precursor tabersonine, which is produced from strictosidine aglycone in a pathway uncharacterised to date. Tabersonine is then derivatised to vindoline through a series of tailoring reactions. During the course of this work a vindoline biosynthesis gene was discovered, so vindoline biosynthesis and the genes and enzymes involved will be introduced here in more detail, presenting the state of knowledge at the start of this work.

The biogenic origin of the alkaloid vindoline from the precursor tabersonine and their shared terpenoid origin has been a focus of investigation since the 1960s, beginning with feeding studies using isotopically labeled substrates and the isolation and characterisation of intermediates and end products [38,39]. Early research into the pathway leading from tabersonine to vindoline established a set of necessary reactions and a conclusive order of those steps [40].

After initial hydroxylation of tabersonine by tabersonine-16-hydroxylase (T16H) [41,42] , the resulting 16-hydroxytabersonine is O-methylated by 16-hydroxytabersonine 16-O-methyltransferase (16OMT) [34] to form 16-methoxytabersonine. The next reaction, which yields the product 16-methoxy-2-3-dihydro-3-hydroxytabersonine, is a formal hydration of 16-methoxytabersonine. The gene or genes responsible for this transformation were unknown at the start of this thesis. Subsequently, an N-methyltransferase (NMT) [43] catalyses methylation of the indoline nitrogen. Desacetoxyvindoline, the NMT product is further hydroxylated by desacetoxyvindoline 4-hydroxylase (D4H) [44] to form deacetylvindoline and in a final step, deacetylvindoline is O-acetylated by deacetylvindoline acetyltransferase (DAT) [45] to yield vindoline (Figure 7).

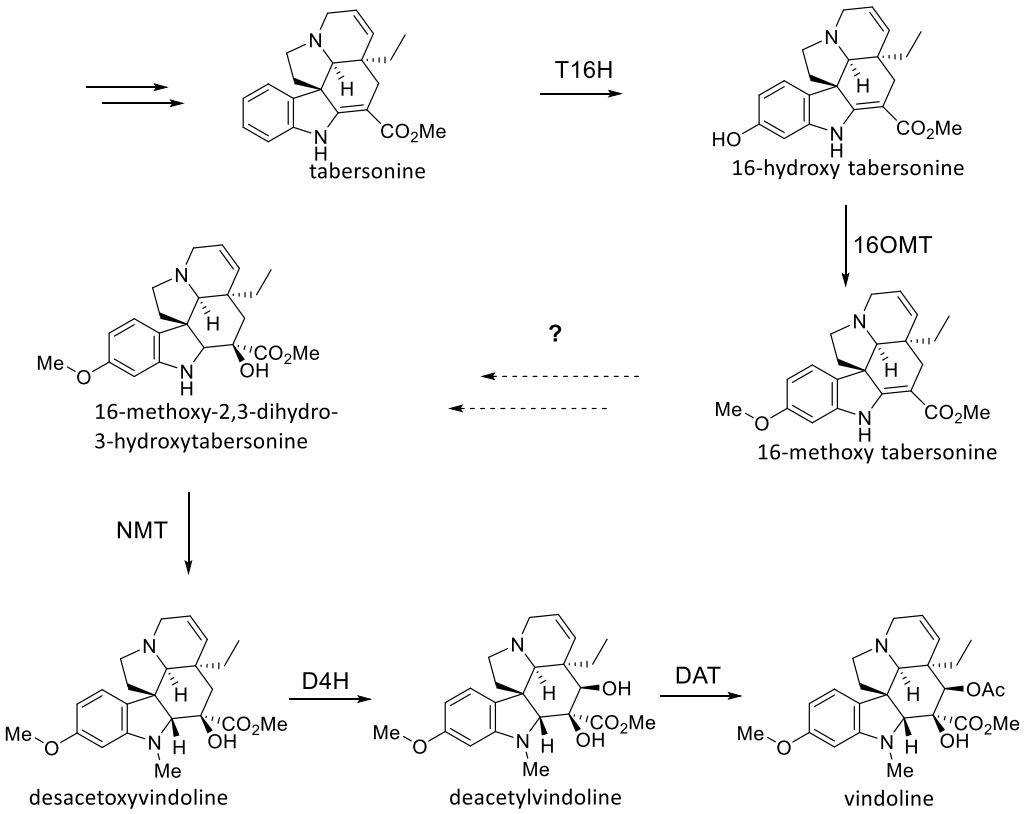


Figure 7 Vindoline biosynthesis enzymes known at the start of this work

Starting from the substrate tabersonine, the steps of vindoline biosynthesis known at the start of this work were the oxidation of tabersonine by tabersonine 16-hydroxylase (T16H) and methylation of the same position by 16-hydroxytabersonine O-methyltransferase (16OMT). After one or possibly two unknown steps the biosynthesis proceeds with the methylation by 16-hydroxy-2,3-dihydro-3-hydroxytabersonine N methyltransferase (NMT), hydroxylation by desacetoxyvindoline 4-hydroxylase (D4H) and ends with a final acetylation by deacetylvindoline 4-O-acetyltransferase (DAT).

1.1.8 Localisation of alkaloid biosynthesis in Catharanthus roseus

The localisation of MIA biosynthesis has been studied in *C. roseus* with three cell types known to be involved in this pathway. The very early steps of seco-iridoid production are localised to the IPAP cells and most likely the intermediate loganic acid is imported into the epidermis (Figure 8) [29]. In the epidermis, tabersonine, catharanthine and other alkaloids are produced. A large proportion of catharanthine is actively exported by an ABC transporter out of the epidermis to a waxy outer layer of the leaf [46]. Other end products of the alkaloid biosynthesis of *C. roseus* such as serpentine and vindoline are stored in the vacuole of the mesophyll. The application of carborundum abrasion to *C. roseus* leaves, yielding leaf epidermis enriched extracts, demonstrated an accumulation of tabersonine and 16-methoxytabersonine in this

fraction suggesting an epidermal location of the two initial steps in vindoline biosynthesis [47]. While transcripts of NMT are primarily found in the epidermis [48] the two last steps carried out by *D4H* and *DAT* are localised to the specialised mesophyll cells the idioblasts and the laticifer making desacetoxyvindoline the most likely transported intermediate [49] (Figure 8). Consequently, all current models of alkaloid biosynthesis in *C. roseus* involve a complex network of biosynthetic intermediate transport. This is true for the early part of the pathway, strictosidine production, as well as later parts like the vindoline biosynthesis.

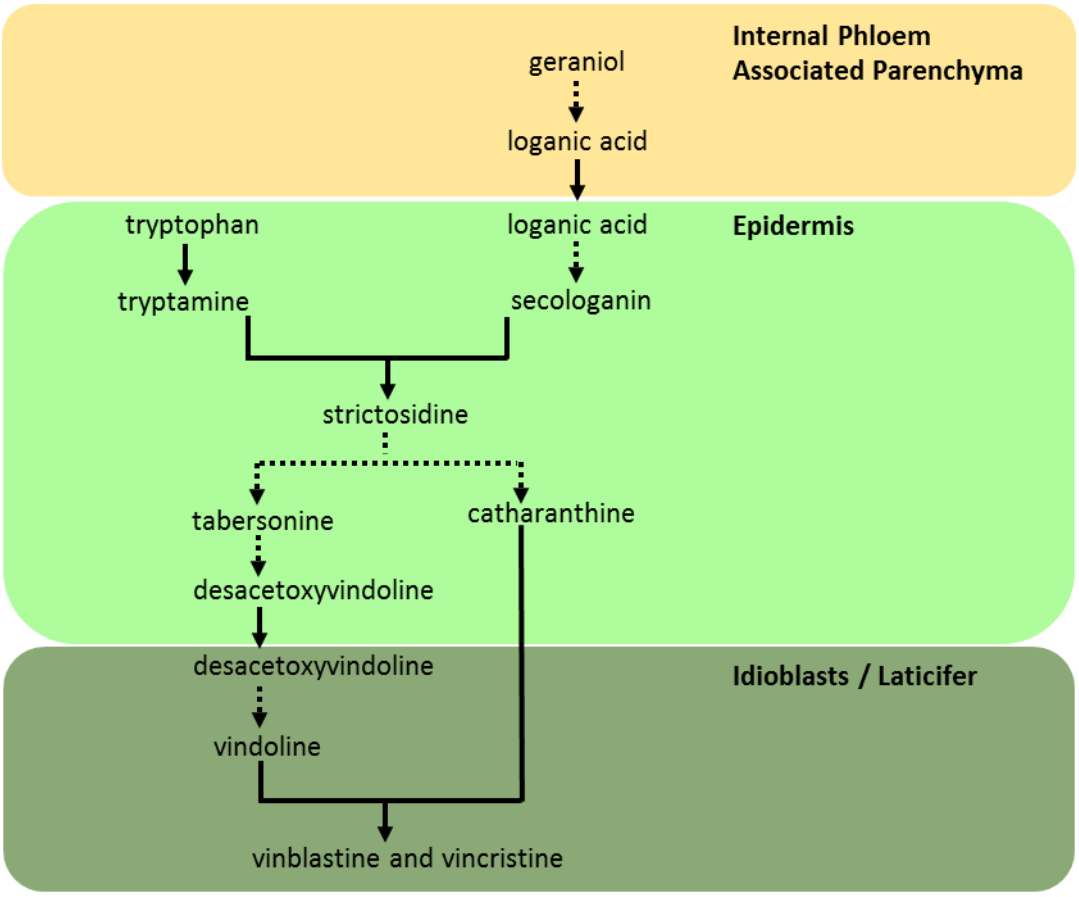


Figure 8 Localisation of alkaloid biosynthesis in *C. roseus* in three cell types

Localisation of alkaloid biosynthesis in *C. roseus* leaf, modified after [48,50]. Three cell types are involved in alkaloid biosynthesis. Early steps of iridoid biosynthesis are localised to the internal phloem associated parenchyma while later steps are localised to the epidermis. The epidermis is also the main site of tryptamine biosynthesis and the early steps of vindoline biosynthesis. Only the last two steps of vindoline biosynthesis are located at specialised cells at the leaf body mesophyll (laticifer and idioblast cells). Furthermore the medicinally important bisindole alkaloids vinblastine and vincristine are predominantly found in this cell type. Solid arrows represent single steps while dotted arrows indicate multiple enzymatic steps.

1.1.9 Function of alkaloids in Catharanthus roseus

Studies into the native biological function of *C. roseus* alkaloids are hindered by the lack of its natural aggressors. Alkaloids such as catharanthine and vindoline, which account for the largest amount of total alkaloid content in *C. roseus,* are produced by the plant as part of normal development. Yet under certain stress conditions, mimicked by application of methyl jasmonate and other elicitors, production of these alkaloids increases up to 3 fold [51,52]. This observation has led to different hypotheses regarding the function of alkaloids in *C. roseus* defence against pathogen and herbivore attack. It has been suggested that glucosyl intermediates of alkaloid biosynthesis, such as strictosidine, are inactive alkaloids that are stored in the vacuole to be reactivated by formation of the aglycone, which is a highly reactive and therefore toxic compound. Such storage of inactive compounds, that can be enzymatically activated by deglycosylation upon pathogen attack, are well known in plants for example from the glucosinolates of the Brassica species [53]. In *C. roseus* the resulting aglycone is a highly reactive dialdehyde that causes protein crosslinking and precipitation. Therefore, it was suggested that it might serve as a defence system or may reduce the nutritional value of the plant for the attacking herbivore [54]. In other studies, *C. roseus* extracts and their antimicrobial and antifeedant potential has been investigated [55] and it has been found that deglycosylated strictosidine was active against several microorganisms [56]. Catharanthine, which accumulates at the *C. roseus* leaf surface in concentrations of 14 µg to 23 µg per cm^2, was able to inhibit *in vitro* the growth of fungal zoospores (*Phytopthora parasitica*) at comparable concentrations. However commercially available *Phytopthora* resistant and unresistant *C. roseus* varieties contain similar levels of catharanthine suggesting a more complex mechanism of *in vivo* disease resistance [57].

1.1.10 Summary of MIA biosynthesis in Catharanthus roseus

More than six decades of research into *C. roseus* have helped to draw a picture of alkaloid biosynthesis including many of the individual pathway enzymes and corresponding genes (Figure 9).

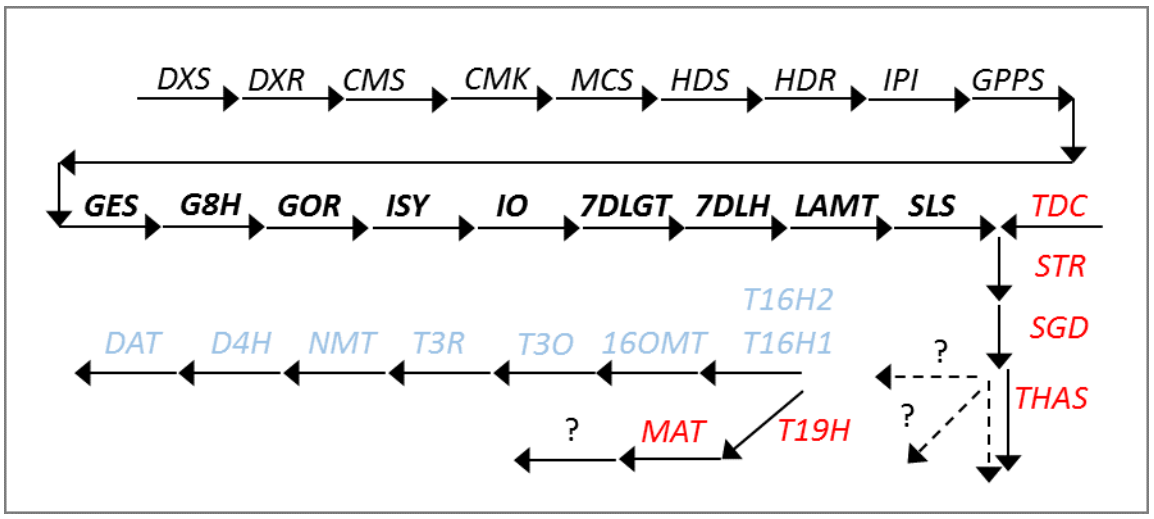


Figure 9 Characterised genes of alkaloid biosynthesis in *C. roseus*

Genes of the monoterpene indole alkaloid pathway in *C. roseus* published as of May 2016. Missing genes are represented by question marks. Dashed arrows indicate unknown numbers of enzymatic steps. Black top row arrows represent methylerythritol phosphate (MEP) pathway, the upstream terpene biosynthesis genes: DXS, 1-deoxy-D-xylulose 5-phosphate synthase 2; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; CMS, 4-diphosphocytidyl-methylerythritol 2-phosphate synthase; CMK, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MCS, 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, GCPE protein; HDR, 1-hydroxy-2-methyl-butenyl 4-diphosphate reductase; IPI, plastid isopentenyl pyrophosphate, dimethylallyl pyrophosphate isomerase; GPPS, geranyl pyrophosphate synthase and GES, plastid geraniol synthase. Bold arrows represent iridoid biosynthesis genes: G8H, geraniol 8-hydroxylase; GOR, 8-hydroxygeraniol oxidoreductase; ISY, iridoid synthase; IO, iridoid oxidase (CYP76A26); 7DLGT, UDP-glucose iridoid glucosyltransferase; 7DLH, 7-deoxyloganic acid 7-hydroxylase; LAMT, loganic acid methyltransferase and SLS, secologanin synthase. Red arrows represent downstream alkaloid biosynthesis genes: TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGD, strictosidine β-glucosidase; THAS, tetrahydroalstonine synthase; T19H, tabersonine/lochnericine 19-hydroxylase (CYP71BJ1) and MAT, minovincinine 19-hydroxy-O-acetyltransferase. Blue arrow represent vindoline biosynthesis pathway: T16H1, tabersonine 16-hydroxylase 1 (CYP71D12); T16H2, tabersonine 16-hydroxylase 2 (CYP71D351); 16OMT, 16-hydroxytabersonine O-methyltransferase; T3O 16-methoxytabersone 3-oxigenase; T3R, 16-methoxytabersonine 3-reductase NMT, 16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase; D4H, desacetoxyvindoline-4-hydroxylase and DAT, deacetylvindoline 4-O-acetyltransferase.

1.2 Available sequence resources for *Catharanthus roseus*

The following section summarises the sequence resources available at the start of this work. Since the discovery of the medicinal value of *C. roseus,* interest in the elucidation of the MIA pathways has led to the application of available sequencing technologies to this model medicinal plant. The result has been insights not only into alkaloid biosynthesis but also into areas such as the transcriptional regulation and transport of secondary metabolism.

1.2.1 Catharanthus roseus expressed sequence tags

Expressed sequence tags (ESTs) are generated by sequencing randomly selected clones from a cDNA library. The resulting sequences are relatively short with only 500 to 800 bp [58]. The NCBI EST database reports 20,168 ESTs for *Catharanthus roseus*. The first ever reported EST datasets for *C. roseus,* two cDNA libraries of leaf base and root tip tissue [59], did not succeed in capturing the majority of known genes of MIA production in *C. roseus*. However, in 2008 the identification of the gene encoding loganic acid methyl transferase (LAMT), a major pathway step in the MIA pathway of *C. roseus,* was reported. This discovery utilised an EST dataset generated using leaf epidermis tissue [50]. Since many steps of MIA biosynthesis are known to be localised to this tissue, the epidermis of *C. roseus* has been of special interest for investigation into alkaloid biosynthesis. It was shown that it is indeed the preferred site for the production of precursors of vindoline biosynthesis [34,47] as well as the site of predominant expression of other MIA genes such as *TDC*, *STR* and *SLS* [35,60].

The same EST set helped identify 16-hydroxytabersonine O-methyltransferase (*16OMT*), the second enzymatic step in vindoline biosynthesis in *C. roseus* [34] after several previous attempts to clone *16OMT* from *C. roseus* using homology based methods or protein purification had failed [61,62]. Ultimately the discovery of further pathway genes such as that of a transporter involved in catharanthine transport [46] and that of 7-deoxyloganic acid hydroxylase (*7DLH*) [33], a key step in the seco-iridoid pathway, were guided by the knowledge of the high expression in epidermis tissue.

The results of the analysis of the same EST dataset also facilitated the cloning of a gene that, when functionally expressed in yeast, results in the preferential accumulation of α-amyrin over β-amyrin. Considering its predominant expression in the leaf epidermal cells, combined with the knowledge that the main triterpene from *C. roseus,* ursolic acid, is derived from amyrin and

is exclusively secreted to the cuticular wax layer, strongly suggests that this gene is amyrin synthase [63,64] . The first of the two publications additionally reports the discovery of amyrin oxidase (AO), a multifunctional C-28 oxidase that catalyses the subsequent successive oxidations leading to the production of ursolic acid from α-amyrin, again exploiting the mentioned EST data to search for candidate genes with high expression in leaf epidermis of *C. roseus* [63].

1.2.2 Catharanthus roseus transcriptomic data

All transcriptomic data available for *C. roseus* at the start of this work was *de novo* assembled, as no reference genome for *C. roseus* was available. The NCBI sequence read archive (SRA) contains 65 entries for *C. roseus* as of May 2016.

The transcriptomic data set used in this thesis is a robust set of annotated transcript assemblies of *C. roseus*. It is publicly available at http://medicinalplantgenomics.msu.edu/. The development of this Medicinal Plant Genomics Resource (MPGR) was facilitated using isolated RNA from a broad range of cell types (Table 1), from which cDNA libraries were constructed and then sequenced using the Illumina sequencing platform with single end reads and a read length of 36 bp [65].

Table 1 Tissues and treatments contained in MPGR transcriptomic dataset

This set is derived from 23 different *C. roseus* tissues and/ or treatments [65]. Suspension cultures were treated with 0.3 mg/ml yeast extract (YE) for different amounts of time. Sterile seedlings and hairy roots were incubated with 100 µM methyl jasmonate (MeJA) for different amounts of time. The *TDC*i hairy roots are lines with silenced tryptophan decarboxylase expression [66]. The RebH hairy root line expresses a bacterial halogenase [67]. In bold are recalculated transcriptomes on the basis of the *C. roseus* genome [68].

Number	Tissue/ experimental condition
1	Flowers
2	Suspension culture YE 6 h
3	Suspension culture YE 12 h
4	Suspension culture YE 24 h
5	**Sterile seedlings MeJA 0 h**
6	**Sterile seedlings MeJA 6 h**
7	**Sterile seedlings MeJA 12 h**
8	Sterile seedlings MeJA 24 h
9	Sterile seedlings MeJA 5 d
10	Sterile seedlings MeJA 12 d
11	Sterile seedlings (control)
12	Suspension culture (control)
13	**Mature leaf**
14	**Immature leaf**
15	**Stem**
16	**Root**
17	Wild type hairy roots
18	Wild type hairy roots MeJA 0 h
19	Wild type hairy roots MeJA 24 h
20	TDCi hairy root
21	TDCi hairy root MeJA 0 h
22	TDCi hairy root MeJA 24 h
23	RebH hairy root

Two additional sources are also available. Firstly, CathaCyc (http://www.cathacyc.org/) provides a platform on which *C. roseus* transcriptomic data is linked to metabolic pathways and offers tools for their visualisation and analysis [69]. It comprises two newly generated transcriptomic datasets as well as the complete *C. roseus* transcriptomic data from the MPGR consortium mentioned above. Recalculating the FPKM values, the obtained expression data is comparable between both applied datasets. This data is easily accessible for individual genes. Secondly, the Canadian based Phytometasyn (http://www.phytometasyn.org/) provides a

further source with 74 plant transcriptome assemblies of which 9 belong to *Catharanthus* [70]. Here different members of the *Catharanthus* genus such as *C. longifolius* and *C. ovalis* are reported.

After the *C. roseus* draft genome assembly became available raw data of 8 tissues of the original MPGR data set (Table 1) was aligned to the genome and used to recalculate expression (FPKM) for those 8 transcriptomes [68].

1.2.3 Catharanthus roseus genomic data

In 2014 the genome of coffee (*Coffea canephora)* was published [71]. Belonging to the same plant order as *C. roseus*, the Gentianales, coffee also contains alkaloids. Yet the coffee alkaloids are produced by substantially different biosynthetic pathways and homologies to *C. roseus* are very unlikely. *C. roseus* belongs to the plant family Apocynaceae. Only one other genomic resource for plants of the Apocynaceae family is available, a 0.5 x genome of the common milkweed [72]. Sequence data for the *C. roseus* plastid genome is available [73], but no nuclear genomic data for *C. roseus* was available at the start of this thesis.

1.3 Co-expression analysis for gene discovery in plant secondary metabolism

1.3.1 General aspects of co-expression

Gene co-expression describes the similarity of gene expression patterns under various conditions. Functionally related genes, such as the genes belonging to a metabolic pathway, often display co-expression [74] making the "guilt by association" principle a powerful tool in gene discovery [75]. Many examples for co-expression of specific pathways in primary [74,76] and especially in secondary metabolism exist such as anthocyanin biosynthesis in *A. thaliana* [76], or the production of monoterpenes in *A. thaliana* flowers [77].

1.3.2 Aspects of co-expression analysis specific to Catharanthus roseus

Global co-expression analyses are possible for model plants such as *A. thaliana* where hundreds [65] of datasets are available and can be computationally mined [78]. For *C. roseus* far less resources are available. However, a dataset composed of 23 tissues and treatments was established in 2011 [65] and careful analysis has led to a number of gene discoveries in alkaloid biosynthesis in *C. roseus* [30,79] as this pathway displays strong co-expression at least in its upstream region [49].

However, this dataset [65] has its limitations. Out of the 23 tissues of this set, 11 are obtained from tissue cultures such as hairy root cultures and suspension cultures. This is due to the fact that a lot of early research focused on *C. roseus* cell and tissue cultures for both practical and commercial reasons [3]. Both systems are relatively easy to maintain and can be scaled up to demands. The biggest advantage is however the possibility to genetically modify these systems. Successful examples are the overexpression of transcription factor ORCA2 in cell suspension cultures to study its effect on alkaloid production [80] or the silencing of tryptamine biosynthesis in hairy roots, a tissue that is also included in the transcriptomic dataset [65], to explore the potential of producing non-natural alkaloids by feeding alternative substrates [66]. Nevertheless, limitations of cell and tissue cultures are apparent. Apart from their instability and fluctuation in alkaloid quantity over time [81] most suspension cultures produce only the MIAs ajmalicine and serpentine. Hairy roots additionally produce tabersonine and catharanthine, but both systems lack the ability to generate the full spectrum of different alkaloids produced by whole plants. Most importantly the medicinally valuable alkaloids vinblastine and vincristine cannot be obtained from *C. roseus* cell cultures or hairy roots because of their lack of vindoline production. This has an important implication: co-expression analysis of transcriptome data obtained from tissues with inconsistent vindoline production are consequently an unreliable source for gene discovery in vindoline biosynthesis.

1.4 Virus Induced Gene Silencing

While techniques such as co-expression analysis provide candidates, further investigation of a possible involvement of these identified candidates in alkaloid biosynthesis requires additional experimental tools. Classical functional genomic studies for the rapid analysis of enzymes rely on loss of function techniques such as the generation of knockout mutants. For many non-model organisms, such as *C. roseus*, a reliable protocol for stable genetic transformation is lacking. Alternatively Virus Induced Gene Silencing (VIGS), a transient posttranscriptional gene silencing technique, provides a system to asses gene function *in planta.*

Virus induced gene silencing (VIGS) utilises the RNAi mechanism for the sequence specific down regulation of protein expression. Infection of plants with a viral vector, which additionally carries part of a host plant gene, results in sequence specific post transcriptional gene silencing (PTGS) [82]. The technique exploits the fact that plants defend themselves against viral infection by targeting the viral-genome for sequence-specific degradation. Double-

stranded RNA (dsRNA) is produced during viral replication by RNA-polymerase from the single stranded viral RNA. The dsRNA is recognised and cleaved by DICER, specific enzymes that are part of the plants defence response. These enzymes degrade the dsRNA to form 21 to 30 nucleotide long small interference RNA (siRNA) [83]. The siRNA is subsequently incorporated into the RNA-induced silencing complex (RISC) which uses them as template to recognise complementary messenger RNA (mRNA). Argonaute proteins, which form a part of the multiprotein RISC complex, cleave the identified specific mRNA [84]. Both virus and plant mRNAs targeted in this manner are degraded, consequently leading to reduced accumulation of specific transcripts, reduced translation and ultimately to a knockdown phenotype of the targeted host gene [85].

One of the most widely applied VIGS techniques employs *Agrobacterium tumefaciens* to infect the host plant with Tobacco Rattle Virus (TRV) [86]. Main advantages of applying TRV, apart from its large host range, are its limited visual symptoms on the infected host plant and the ability of TRV to infect cells at an early time point in their development and differentiation and subsequently to start gene silencing before large amounts of the target protein are produced by the cell. Additionally, unlike most other plant viruses, TRV is able to penetrate growing points in plants and is therefore able to not only spread by transport but also by cell to cell distribution via cell division [87].

TRV is a bipartite positive sense RNA virus that consists of RNA1 and RNA2. RNA1 encodes several genes including a replicase and proteins involved in viral movement that allow long-distance movement of RNA1 in the plant and systemic infection without the requirement of RNA2 [88]. RNA2 includes a viral coat protein which can encapsulate both RNA1 and RNA2. Most importantly RNA2 was engineered to contain a multiple cloning site to allow insertion of target gene sequence from the host plant [89]. The TRV RNA1 and RNA2 were each cloned between cauliflower mosaic virus (CaMV) 35S promoter and a nopaline synthase (NOS) terminator of a T-DNA vector. This binary shuttle vector is able to replicate in both *E. coli* and in *A. tumefaciens* [89]. The resulting plasmids pTRV1 and pTRV2 have been used for successful gene silencing in a diverse range of plants such as tomato [90], opium poppy [91] and barley [92].

To infect plants, *A. tumefaciens* cultures carrying either pTRV1 or pTRV2 are mixed in equal amounts and used for infiltration of a host plant (Figure 10). Once the T-DNA has been integrated into the host plant genome and transcribed by the host plant's RNA polymerase, formation of the systemic virus occurs and the antiviral RNAi silencing pathway is induced.

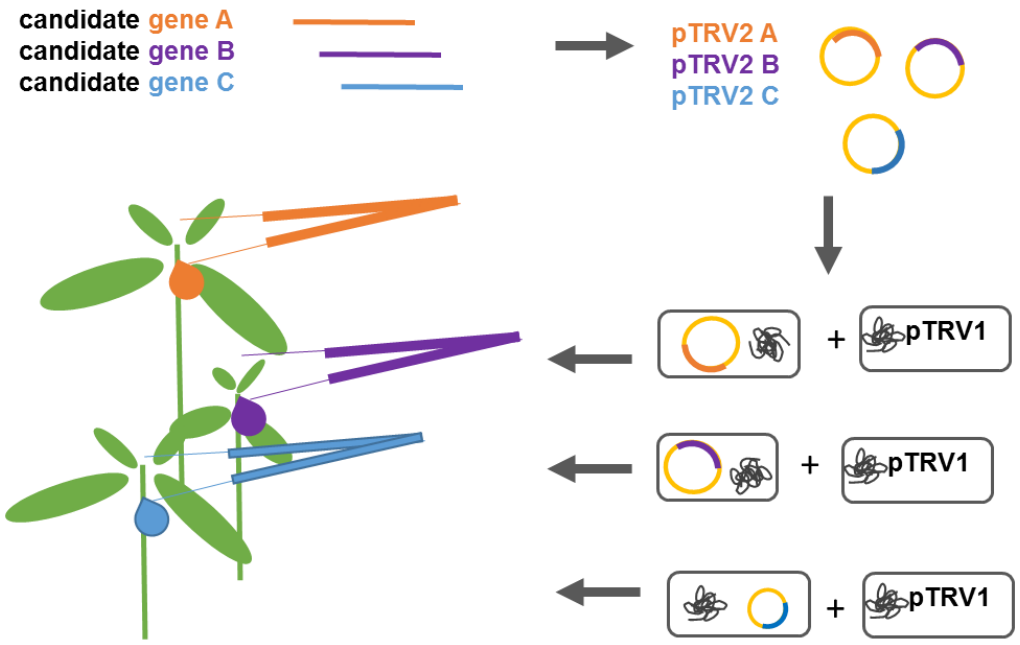


Figure 10 Schematic overview of agrobacterium mediated TRV gene silencing in plants
Short segments (usually 200-400 bp) of candidate genes are cloned into the pTRV2 vector. Competent *A. tumefaciens* was transformed with the pTRV2 vector and mixed with a second *A. tumefaciens* containing the pTRV1 plasmid. Plants are infected with the mixed Agrobacterium cultures by pinching (forceps or syringe). Other methods include drenching [93] or biolistic methods [94]. Figure modified after [95].

1.5 Gene clustering in plant secondary metabolism

The term "operon" [96] defines a group of genes that are functionally related, co-regulated and often transcribed as a single messenger RNA (mRNA). Operons were thought to be restricted to prokaryotic organisms with only a few exceptions, such as the biotin synthesis pathway of yeast [97] or the major histocompatibility complex in humans [98]. However it was shown that for example filamentous fungi harbour a rich and diverse range of gene clusters that produce secondary metabolites, many of which have antimicrobial and antibiotic function [99].

The recent discovery of a relatively modest number of secondary metabolite gene clusters in plants has stirred great attention and has been discussed [100,101] and reviewed [7,102–104] to a great extent. Plant metabolic gene clusters are now known for various plants and a variety of different secondary metabolites, with terpenoid clusters being the most abundant (Table 2). Much of the excitement generated by the discovery of plant gene clusters is due to the

consideration that the discovery of such gene clustering in plants could enable efficient elucidation of complex specialised biosynthetic pathways and circumvent the current laborious process of plant gene discovery.

Table 2 Examples of known plant secondary metabolite cluster

Table modified after [105,106].

plant	class/ secondary metabolite	reported
maize (*Zea mays* ssp.)	benzoxazinones (bx cluster, DIBOA, DIMBOA)	1997
cassava (*Manihot escolenta*)	cyanogenic glucosides/ linamarin and lotaustralin	2011
sorghum (*Sorghum bicolor*)	cyanogenic glucosides/ dhurrin	2011
lotus (Lotus japonicus)	cyanogenic glucosides/ linamarin and lotaustralin	2011
poppy (*Papaver somniferum*)	alkaloid/ noscapine	2012
tomato (*Solanum lycopersicum*)	alkaloid/ sterodial glycoalkaloids	2013
potato (*Solanum tuberosum*)	alkaloid/ sterodial glycoalkaloids	2013
various *Solanum* ssp.	monoterpene	2013
rice (*Oryza sativa*)	diterpene/ phytocassane (chromosome 2)	2009
rice (*Oryza sativa*)	diterpene/ momilactone (chromosome 4)	2004
oat (*Avena* ssp)	triterpene/avenacin	2004
lotus (Lotus japonicus)	triterpene/ lupeol	2013
thale cress (Arabidopsis thaliana)	triterpene/thalianol	2008
thale cress (Arabidopsis thaliana)	triterpene/marneral	2011
various *Euphorbiaceae* ssp.	diterpenoid	2014
ricinus (*Ricinus communis*)	diterpenoid	2014
barley (*Hordeum vulgare*) [107]	beta-diketone	2016

1.5.1 A definition of plant specialised metabolism gene cluster

According to a recent review, a gene cluster is defined as: “a set of two or more non-homologous genes encoding enzymes from the same pathway (to be distinguished from the gene clusters resulting from tandem duplication and those consisting of homologous genes)” [105]. A more detailed definition cannot be made as the plant gene clusters identified to date vary greatly in many aspects such as cluster size, the actual number of genes, the number of genes unrelated to the pathway contained within the cluster, the co-regulation of the genes, and the order of genes in respect to the order in the pathway for which they are responsible. A brief overview over a few selected gene clusters in secondary metabolism discovered to date is given in the following section.

One of the two known rice gene clusters, the chromosome 2 cluster, includes genes that are involved in the production of not one but two different phytoalexins, phytocassanes and oryzalides/ oryzadiones [108]. These genes are differently regulated, an exception to the typically co-regulated genes of plant gene cluster. The rice chromosome 2 cluster contains four closely

related members of the cytochrome P450 CYP76M subfamily (CYP76M5-8). These four homologous P450s have a similar but not identical substrate scope [109] providing a metabolic plasticity that might enable positive selection pressure and play a role in the further evolution of this cluster. The second rice specialised metabolite cluster is involved in the production of momilactone and is located on chromosome 4. It contains a putative dehydrogenase gene (*AK103462*) and two functionally unknown P450 genes (*CYP99A2* and *CYP99A3*) together with *OsKS4* and *OsCyc1*, both diterpene cyclase genes. Together these genes form a chitin oligosaccharide elicitor- and UV-inducible gene cluster involved in momilactone biosynthesis that spans a distance of 168 kb [110].

The first ever secondary metabolite cluster to be discovered in plants is much larger. It encodes the 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) biosynthesis in maize [111]. While six of the steps of this pathway (Bx1 to 5 and Bx8) are located within less than 300 kb on maize chromosome 4, the sixth step in DIMBOA biosynthesis, Bx6 is located in 2 Mb distance while the seventh step Bx7 is located in as far as 15 Mb distance and further genes identified to be belonging to the same pathway are not even on the same chromosome [105].

The 10-gene cluster for the production of noscapine, identified in *Papaver somniferum* (opium poppy) [112] is one of the few so far identified clusters implicated in the biosynthesis of alkaloids. Containing three O-methyltransferases, four cytochrome P450 enzymes, one acetyltransferase, one carboxylesterase and a short-chain dehydrogenase/reductase it comprises five different enzyme classes on a genomic region spanning only 221 kb.

1.5.2 Origin of plant specialised metabolite clusters

The known cluster can give insight into evolution of plant secondary metabolism. For example, the biosynthesis of several terpenes in *Solanum lycopersicum,* a gene cluster on chromosome 8 is responsible. Notably, similar clusters could be detected at the same position on chromosome 8 in several closely related members of the *Solanum* plant family. This syntheny thus provides insight into the possible evolution of this conserved cluster by duplication and divergence from an original set of genes [113]. Another particularity can be seen in the DIMBOA cluster of maize. It originates most likely from an ancestral monocot as wheat and rye for example are able to synthesis DIMBOA yet the gene cluster responsible is split in two parts in the genome, most likely due to a genomic translocation events [7].

Convergent or "repeated" evolution is a frequent phenomenon in plant secondary metabolism [114] and can for example be observed in the biosynthesis of cyanogenic glucosides in *Lotus japonicus*, the Poaceae sorghum (*Sorghum bicolor*) and in the Euphorbiaceae cassava (*Manihot esculenta*). In all three plants, which are only distantly related, cyanogenic glucoside biosynthesis has evolved and the responsible genes are clustered. While in *L. japonicus* and *M. esculenta* the cyanogenic glucosides linamarin and lotaustralin are biosynthesised from L-valine and L-isoleucine, in *S. bicolor* biosynthesis of a different class of cyanogenic glucoside starts with the precursor L-tyrosine. Although the biosynthesis in all three cases requires three steps and in all three plants species the first and third step is carried out by a similar enzyme (a Cyp79 and a UGT85), two facts suggest a convergent evolution. First, the gene responsible for the second step in the biosynthesis has been recruited from two different P450 families (CYP71 and CYP736) and second, the order of the genes in the individual cluster is very different and does not suggest a common inheritance but rather an independent occurrence [115].

1.5.3 Gene clustering supports pathway elucidation

The important crop plant oat (*Avena ssp.*) produces the defence compound avenacin, a saponin triterpene. Originally a mutant population was screened for saponin deficient (SAD) plants and it was shown that nine out of ten SAD mutants co-segregate, suggesting a close physical location of the responsible genes in the genome [116]. Over the following years much could be revealed regarding the location, but also identity and function, of genes involved in avenacin biosynthesis. Oat β-amyrin synthase (*AsbAS1)* encodes for the first committed step in this biosynthetic pathway. Genetic and biochemical analysis indicated that *AsbAS1* is located in the linkage group D of the diploid oat genome together with at least three other genes required for distinct biochemical processes in avenacin synthesis, forming a gene cluster that had not arisen from simple gene duplication of a common ancestor [117]. More precise information on the exact location of the genes involved in this cluster gene could be obtained by the addition of BAC data revealing that the first two steps of avenacin biosynthesis, the 2,3-oxidosqualene cyclisation by β-amyrin synthase and the β-amyrin oxidase, are located within 70 kb on the oat genome [118] [119]. A further methyltransferase gene, responsible for the SAD 9 mutant could also be linked to this cluster [120] and ultimately the cluster known to date comprises five genes, representing five of the original ten sad mutants (SAD 1, 2, 7, 9 and 10) whose identity and function in avenacin biosynthesis could be uncovered. These discoveries highlight how gene clusters streamline the functional characterisation of plant metabolic genes.

1.6 Aim of the work presented in this thesis

The highly complex secondary metabolite pathway of *C. roseus,* leading to the production of valuable anti-cancer compounds, has been the focus of research for decades. Nevertheless, not all enzymatic steps involved in the biosynthesis of the complex alkaloids have been discovered. Full understanding of the entire pathway would enable the production of these compounds in alternative hosts as well as the targeted enzymatic modification of individual compounds, allowing for the rapid expansion of the existing plethora of natural products from *C. roseus*.

The recent acceleration of sequencing approaches to obtain plant genomic data has driven the discovery of a number of metabolic gene clusters involved in a variety of secondary metabolite pathways in a range of plants. In parallel, the availability of transcriptomic datasets has allowed the application of co-expression analyses to identify new genes.

This thesis aims at combining co-expression analysis, which has aided discovery of genes involved in plant secondary metabolism in many plant species, with the investigation of potential gene clustering of the *C. roseus* MIA pathway. The objective is to discover and characterise *in vitro* missing steps in the alkaloid biosynthesis in *C. roseus* as well as to functionally validate the involvement of genes *in planta* in this secondary metabolic pathway.

Co-expression analysis has been used to successfully identify a missing step in vindoline biosynthesis using existing transcriptomic resources [65]. The second part of this thesis describes the generation of newly established genomic resources, selected BACs and the first draft genome, for *C. roseus*. It describes the bioinformatics analysis of the obtained data and subsequent confirmation using PCR from genomic DNA.

Chapter II validates the *in planta* role of two functionally identical alkaloid pathway genes in *C. roseus* leaves applying VIGS. It presents and discusses the results of experimental data obtained from determination of silencing success applying qRT PCR on VIGS tissue, as well as comparing LC-MS measurements of leaf extracts of silenced and unsilenced control tissues.

Chapter III reports the discovery and characterisation of a missing step in vindoline biosynthesis. Coexpression analysis results in gene candidates and VIGS is applied to establish their involvement in vindoline biosynthesis. A positive candidate is identified and characterised using a yeast expression system. The chapter discusses the implications of an observed *in-vivo* rearrangement of the product of this step.

Chapter IV reports the newly establishment and analysis of genomic resources including genomic data and BAC data. The genomic context of alkaloid biosynthesis is investigated and further research is presented. The extent of observed gene clustering in MIA biosynthesis in *C. roseus* considering the quality and quantity of the available data is discussed.

Chapter V presents an overview of all VIGS experiments that were discovered by either co-expression analyses or genome mining and highlights some of the most promising outcomes. It also presents an approach to improve the existing VIGS protocol.

The overarching theme of this research is to utilise existing transcriptomic and newly developed genomic resources in the task of gene discovery in secondary metabolism in the medicinal model plant *C. roseus*.

2 Investigating the *in planta* role of two *C. roseus* enzymes with identical *in vitro* function

Part of this work was initiated as a collaboration with the group of Vincent Courdavault, University of Tours, France. VIGS constructs were designed, VIGS experiments and metabolite analysis as well as qRT PCR of the target gene and pathway related genes were performed by Franziska Kellner. The results of this work have been published in [42]. The full publication can be found at the end of this thesis.

2.1 Introduction

The medicinal plant *C. roseus*, known for its production of the two valuable anti-cancer compounds vinblastine and vincristine, has one of the most extensively studied secondary metabolite pathways. More than 130 different monoterpene indole alkaloids (MIAs) have been reported for *C. roseus* [19]. Despite decades of research that have shed light on the complex biosynthesis of MIAs, not all enzymatic steps are known.

For many non-model organisms, such as *C. roseus*, a reliable protocol for stable genetic transformation is lacking. Therefore to validate or functionally characterise steps of alkaloid biosynthesis alternative tools have been developed such as Virus Induced Gene Silencing (VIGS). VIGS is a transient posttranscriptional gene silencing technique that provides a system to assess gene function *in planta.*

2.1.1 Virus Induced Gene Silencing in Catharanthus roseus

VIGS in *C. roseus* was first established by Dr. Dave Liscombe in the O'Connor group by adapting the gene silencing technique that had been shown to be functional in various plants [121] . Tobacco rattle virus (TRV) was chosen, as it was known that *C. roseus* is susceptible to infection with this virus. To assay successful silencing in *C. roseus*, young plants were pinched with *A. tumefaciens* transformed with a plasmid harbouring a segment of the *C. roseus photoporphyrin IX magnesium chelatase subunit H* (*ChlH*) (Figure 11).

Silencing of this chlorophyll biosynthetic gene should result in a lack of chlorophyll production, and consequently a yellowing of the leaf. Silencing of *ChlH* had previously been used in *Nicotiana benthamiana* as visual marker [122]. As seen in Figure 11, the yellowing, which indicates successful gene silencing, is not restricted to the leaf tissue. Although not reported for *C. roseus,* specific gene silencing for flower or pericarp is possible.

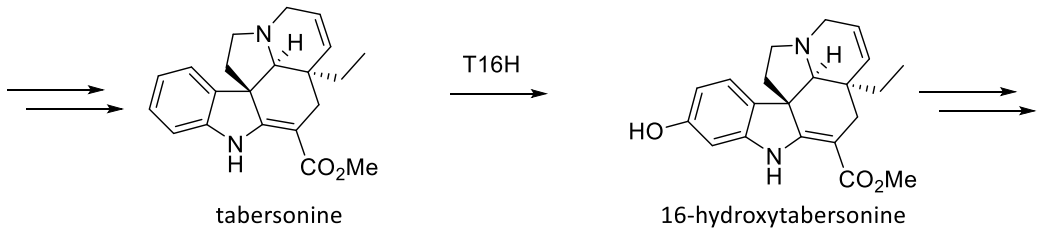
Figure 11 *C. roseus* photoporphyrin IX magnesium chelatase subunit H (ChIH) silencing
An 11-week old *C. roseus* plant three weeks after inoculation with *A. tumefaciens* harbouring the VIGS vector for silencing *C. roseus ChlH*. The visible yellowing affects the leaves above the pinching site as well as stem, sepals and pericarp. However, the silencing effect does not spread into tissue beneath the pinching site.

2.1.2 *Optimisation of VIGS in* Catharanthus roseus

Virus induced gene silencing in *C. roseus* has been streamlined by the introduction of the uracil excision-based cloning (USER cloning) method [123] to the established VIGS pTRV2 vector. This ligation-independent cloning technique utilises PCR primers that contain a single deoxyuridine residue near the 5' end. The subsequent treatment of the PCR product with Uracil-Specific Excision Reagent (USER) enzyme leads to the generation of 8 nucleotide 3' overhangs. These overhangs complement the ends of a USER compatible linearised vector allowing for transformation of *E. coli* without prior ligation. For this purpose a new cassette had been introduced into the pTRV2 vector that facilitates rapid cloning with the newly created pTRV2u vector [30].

2.2 Results and Discussion

The enzyme tabersonine 16-hydroxylase (T16H), a cytochrome P450, synthesises the first committed step of vindoline biosynthesis in *C. roseus*, hydroxylating the biosynthetic intermediate tabersonine at the 16 position of the indole ring (Figure 12).



tabersonine 16-hydroxytabersonine

Figure 12 Tabersonine 16-hydroxylase reaction catalysed by T16H

Given the importance of vindoline in the production of the anti-cancer agent vinblastine, there has been great interest in elucidating all of the biosynthetic enzymes that convert tabersonine to vindoline. In 1999 a gene encoding T16H was successfully cloned from a light and nutritional downshift induced *C. roseus* suspension culture line CP3a by the group of Schroeder [41]. It was discovered after previous sucrose density gradient centrifugation revealed a subcellular localisation at the endoplasmic reticulum suggesting it to be a cytochrome P450 [124]. Tabersonine 16-hydroxylase (T16H) was able to catalyse the conversion of tabersonine to 16-methoxytabersonine when co-expressed with the partner P450 reductase (P450red) as a translational fusion in *E. coli*.

This gene was shown to be expressed in the epidermis of *C. roseus* leaves using RT PCR of isolated epidermal cells [47]. The presence of *T16H* in an epidermis enriched EST data set [50] was consistent with these results. However, when the group of Vincent Courdavault (University of Tours, France) attempted to demonstrate the localisation of *T16H* in *C. roseus* leaves using RNA in situ hybridisation, the gene could not be detected. More careful analysis led instead to the discovery of a *T16H* like gene (*T16H2*) in *C. roseus* leaves with high sequence similarity. The two *T16H* genes, *T16H1* (the gene identified in 1999) and *T16H2,* share 86.5% nucleotide identity. A *C. roseus* expression data set available at the time [65] is not precise enough to distinguish between both isoforms. Therefore, the expression profiles of *T16H1* and *T16H2* are not accurately represented in this dataset, making it challenging to understand the function of these two isoforms (Figure 13).

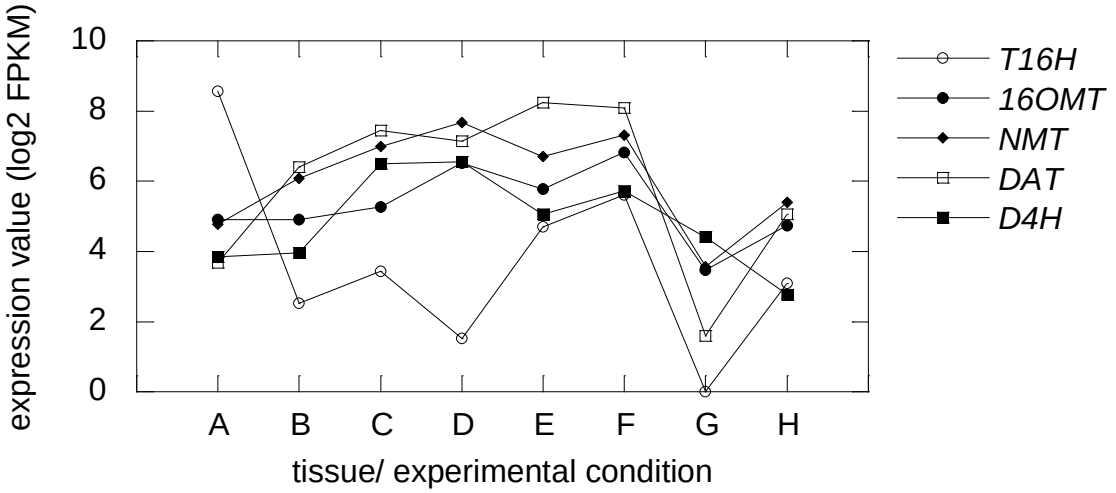


Figure 13 Expression of *T16H* and other vindoline biosynthesis genes

Expression values (log_2FPKM) for the contig that represents both *T16H* versions and all other at the time known vindoline biosynthesis genes in the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root) according to the data obtained from (http://medicinalplantgenomics.msu.edu/).

2.2.1 Mapping transcriptomic raw data to both T16H individual ORFs

Read mapping is a method used in bioinformatics by which short raw reads from transcriptome sequencing are aligned to a reference sequence such as the open reading frame (ORF) of a gene. The different amount of matching reads for each reference sequence represents their individual abundance and can be used to calculate gene expression values.

As mentioned above the expression of *T16H1* and *T16H2* was not individually represented in the transcriptomic data set available at the time [65]. In order to investigate the individual expression of each of the highly similar versions of T16H mapping of raw reads of three transcriptomic raw data sets was performed. According to data obtained from mapping the raw reads, the expression of *T16H2* and *T16H1* differs quite significantly in different tissues. For the mapping two sources were used. The raw data of two tissues of the *C. roseus* transcriptomic data set made available by the Medicinal Plant Genomics Resource (MPGR) [65] and one tissue (leaf) from a second study [70] (Phytometasyn, http://www.phytometasyn.ca/).

The quality and amount of raw data is different for both studies with the first one being composed of only single reads with 35 bp read length and considerably less coverage. However the data conclusively supports that *T16H1* is expressed at negligible levels in leaf tissues of *C. roseus* while *T16H1* is clearly the predominantly expressed version in flowers and suspension culture (Table 3).

Table 3 Distribution of mapped reads for *T16H1* and *T16H2* in different *C. roseus* tissues

Tissue	Source	Number of raw reads	Reads mapped to T16H1	Reads mapped to T16H2
Flower	SRX047002	31,645,190	15259	6333
Suspension Culture	SRX047006	26,234,705	725	198
Leaf	SRX096098	292,873,732	337	65339

To investigate the function of the newly found *T16H* gene in *C. roseus* leaf tissue, VIGS experiments were conducted as described in the following section. Along with the enzymology work performed in the Courdavault Group, these data demonstrate that both genes are functional tabersonine 16-hydorxylases, yet that they have undergone sub-functionalisation, leading to differential spatial localisation of vindoline biosynthesis [42].

2.2.2 T16H2 VIGS construct

To investigate the function of the newly found *T16H* gene in *C. roseus* leaves, VIGS experiments were conducted as described in the following sections. Even though all evidence suggested that the previously reported T16H1 [41] was not present in leaves the VIGS plasmid for *T16H2* silencing was constructed to minimize potential cross silencing. Due to the high sequence similarity it was impossible to choose a region that did not contain any identical regions between the two *T16H* genes. Ultimately, the final VIGS construct contains a gene fragment that incorporates part of the untranslated region (UTR) to maximise sequence differences and has only a single 23 bp long region that is identical between *T16H2* and *T16H1* (Figure 14).

```
T16H2           1 -------------------------TG--ATTGG--AAACTTG----CCAG--TGAT---G   23
                                            TG  ATT G   AACTT     CCA   TGAT
T16H1        1321 CATTTGCTTTGGCTAATATAGAATTGCCATTAGCACAACTTTTGTTCCATTTTGATTGGC 1380

T16H2          24 AAACAAATATTGATAAATTAGACATGACGGAGAGTAGAGGGGTAACAGTTAGAAGAGAAG   83
                  AA CAAATA TGA AAATTA A ATGA  GAGAGTAGAGGGGTAACAGTTAG  GAGAAG
T16H1        1381 AATCAAATACTGAGAAATTAAATATGAAAGAGAGTAGAGGGGTAACAGTTAGGCGAGAAG 1440

T16H2          84 ATGATTTGTGTCTGATTCCATTTCCTTATTCTGCTTCTTCTCTCAAAGGTAAATATTAGA  143
                  ATGATTTGT T TGA TCCA TT  TT TTCT C TCTTCTC      G AAAT TT  A
T16H1        1441 ATGATTTGTATTTGACTCCAGTTAATTTTTCTTCCTCTTCTCCTGCTTGAAAATCTT-TA 1499

T16H2         144 TGGAGACAACATCACCAAATAATTGAAGACCAAGA----------TTGTAGGTCATGAAT  193
                    GA ACAA      CAA   AT G AG CCA GA          TT T   TC T   T
T16H1        1500 CCGAAACAA----TTCAAGACATGGTAGTCCATGAAGGATTTTTTTTTTTTTCTTTCTT 1555

T16H2         194 AG-TTGTCGGTCAAGAA----ATGGTGGCAAAAATTGCTACGGAAGATTTCGTTGCTAAT  248
                     TT TC   C AGAA    ATG  GG   AA TT  T  GGA    TT GTT  T
T16H1        1556 CCTTTTTCTTACGAGAAGTCCATGAAGGATTAAGTT--TGTGGA----TTTGTTAGTTTG 1609

T16H2         249 CCAT------------------------------------------------------  252
                  CC T
T16H1        1610 CCTTTGGTAACAAGTTTGTGGATTTGTTTGTATCTAAATGAATAAAGCAATGATTGCCTC 1669
```

Figure 14 Sequence identity between T16H2_VIGS construct and *T16H1* gene

The nucleic acid sequence alignment of *T16H1* and T16H2 show the single 23 bp long section (highlighted in grey) that is identical between the gene segment used for silencing *T16H2* and the corresponding *T16H1* gene segment. The forward and reverse primer for cloning the VIGS construct are in bold.

The designed VIGS construct was used in silencing experiments on 8 week old *C. roseus* seedlings. As a control, tissue harvested from plants infected with an empty vector (EV) silencing construct was used. Silenced tissue was harvested three weeks after infection, milled and each sample divided into two fractions, one for RNA extraction and subsequent qRT PCR and the second fraction to be prepared for metabolite analysis with LC-MS.

2.2.3 T16H2 silencing success determined by qRT PCR

RNA was extracted from 6 samples of *T16H2*_VIGS tissue and an equal amount of RNA from EV control tissues was converted to cDNA and *T16H2* gene expression was measured. An approximately 80% decrease of transcripts in the T16H2 silenced tissues relative to the EV controls confirmed the efficient knockdown of the T16H2 gene by VIGS (Figure 15).

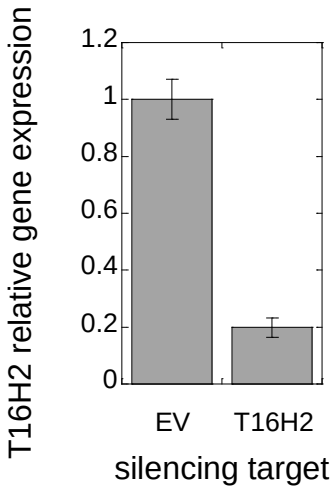


Figure 15 Silencing effect on gene expression of *T16H2*

Relative expression of *T16H2* gene in *T16H2* silenced tissues compared to EV tissue (n=6). Expression was normalised to expression of CrRPS9. Data presented is the mean ± SEM.

Repeated efforts to detect the presence of *T16H1* in the silenced or empty vector control tissues gave only negative results. If any *T16H1* expression is present in leaves of *C. roseus* it is below the detection limit of the qRT PCR protocol.

2.2.4 T16H2 silencing influences leaf alkaloid content

The alkaloid content of *C. roseus* has been investigated and reported several times over the past years [19,125]. One study examines the leaves of 64 modern commercial *C. roseus* varieties [126], reporting a large spectrum of total alkaloid concentration ranging from approximately 5 mg to 0.5 mg per g dry weight [126]. Although over 130 different MIAs are reported for *C. roseus* only a few accumulate to significant levels in *C. roseus* tissues [19]. "SunStorm Apricot" (Figure 4), the variety used predominantly for this thesis, accumulates 1.9 mg of catharanthine per g dry leaf tissue and 0.8 mg per g dry leaf tissue of vindoline and reaches an average total alkaloid content of 3 mg per g leaf tissue [126].

Liquid Chromatography Mass Spectrometry (LC-MS) measurements of leaf tissue methanol extracts are used to examine the effect of silencing a specific gene on leaf alkaloid content. As a first analysis of the obtained data, the peak area of the known metabolites secologanin, strictosidine, serpentine, catharanthine, tabersonine, vindoline and vindorosine (Figure 16) are investigated.

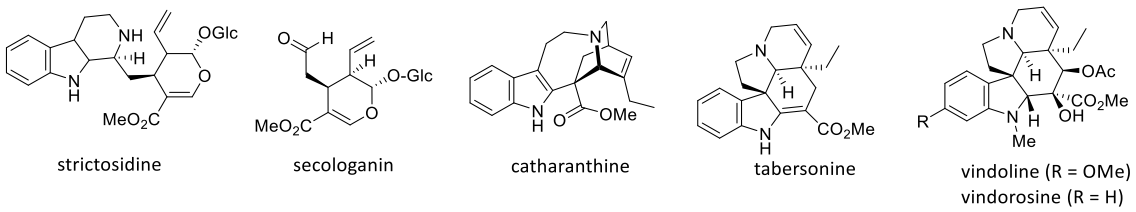


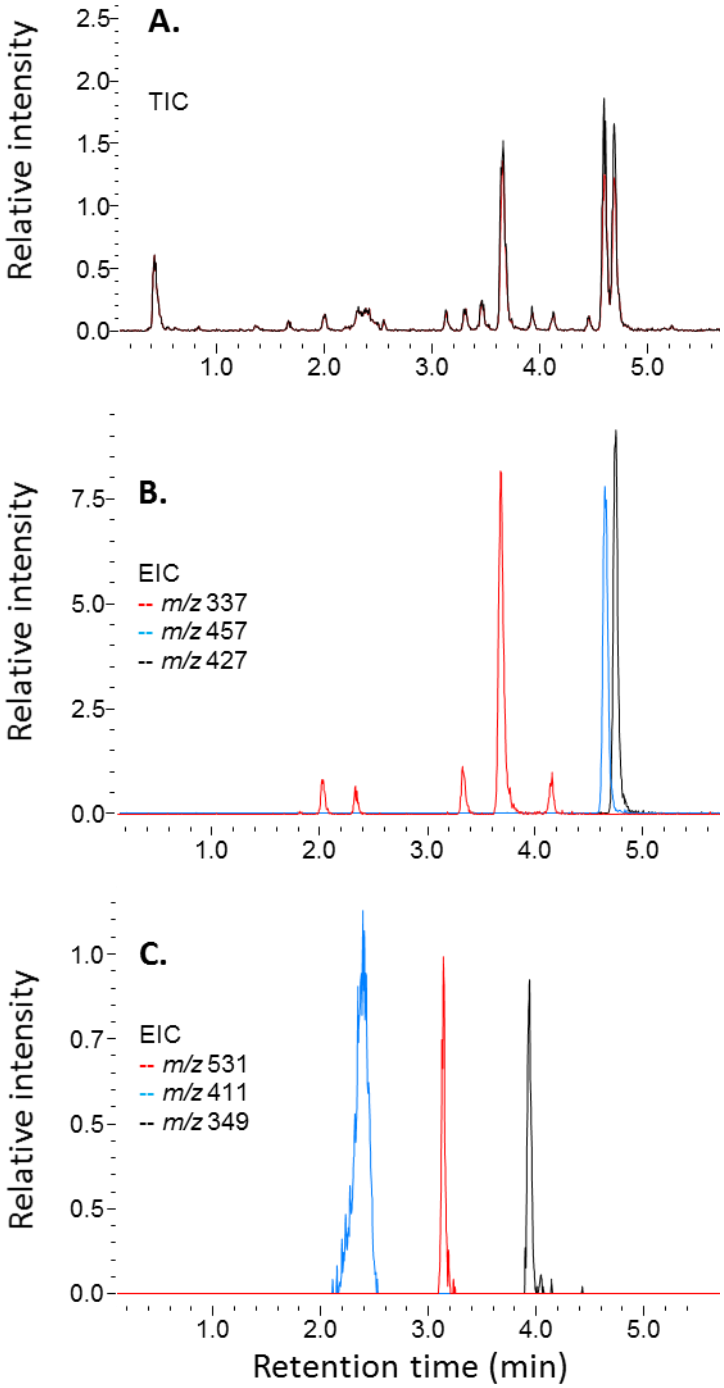
Figure 16 Chemical structures of metabolites investigated in silenced leaf tissue

Generally samples of between six and eight individually silenced plants are measured per treatment. LC-MS analysis of all VIGS experiments used either a Surveyor high performance liquid chromatography (HPLC) system attached to a DecaXPplus ion trap MS (Thermo-Finnigan) or a Nexera/Prominence UHPLC IT-ToF (Shimadzu). All compounds of interest ionise comparably on both machines (Dr. Lionel Hill, Dr. Richard Payne, personal communication).

LC-MS data was obtained as total ion chromatogram (TIC) (Figure 17 A). Peak area was calculated from the extracted ion chromatogram (EIC) for each of the metabolites of interest (Figure 17 B) using the appropriate software for each instrument.

Figure 17 Representative chromatogram of young leaf tissue of *C. roseus* methanol extract

A: Representative total ion chromatogram (TIC) of a *C. roseus* "SunStorm Apricot" methanol extract of young leaf tissue obtained on the IT-TOF mass analyser (Shimadzu). **B**: Red: Extracted ion chromatogram (EIC) for *m/z* 337 representing catharanthine and tabersonine at time 3.9 and 4.3 respectively, blue and black: EIC for vindoline and vindorosine with *m/z* 457 and *m/z* 427 respectively. **C:** Red: Extracted ion chromatogram (EIC) for strictosidine with *m/z* 531, blue: EIC for secologanin (Na^+ adduct) with *m/z* 411 and black: EIC for serpentine with *m/z* 349.



For the Thermo-Finnigan ion-trap instrument peak areas were calculated using the Thermo, Xcalibur Roadmap 2.2 SP1.48 software. For the IT-ToF mass spectrometer (Shimadzu) peak areas were calculated using the Shimadzu LC-MS solution software version 3.80.409.

The extracted peak areas are normalised by sample tissue weight and caffeine as internal standard. Subsequently statistical analyses was used to compare the average normalised peak areas of metabolites in silenced to those of empty vector control tissues to detect significantly changes in alkaloid accumulation.

2.2.4.1 T16H2 silencing alters vindoline and vindorosine accumulation significantly

The hydroxylation of tabersonine at the C16 position by T16H is the first enzymatic step in the multiple step biosynthesis of vindoline from tabersonine in *C. roseus* leaves (Chapter 1, Figure 7).

The silenced tissues of two independent VIGS experiments were analysed by LC-MS using the Thermo-Finnigan ion-trap instrument. Already during initial inspection of the data it became strikingly apparent that the vindoline peak in *T16H2* silenced plants was reduced while at the same time the peak for a closely related alkaloid, vindorosine was increased in comparison to EV control samples. Vindorosine is a naturally occurring vindoline derivative that lacks the hydroxyl moiety at carbon 16. In EV control tissues vindorosine is consistently found at lower levels than vindoline, but in *TH162* silenced tissues, this ratio is reversed (Figure 18).

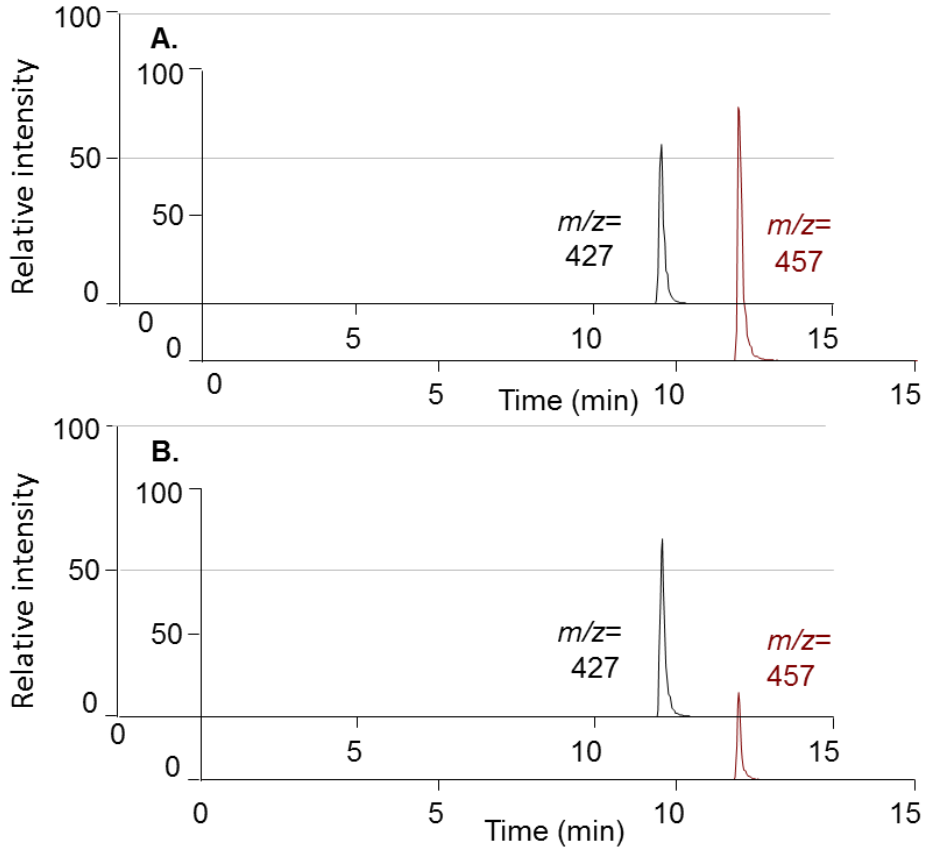


Figure 18 Comparison of extracted ion chromatograms of *T16H2* and EV silenced tissues

Extracted ion chromatograms obtained from LC-MS data generated on the Thermo-Finnigan ion-trap instrument for vindorosine (*m/z* 427) and vindoline (*m/z* 457) **A**: Representative example of an EV silenced plant. The vindoline peak, usually the largest peak in the chromatogram, exceeds the vindorosine peak. **B**: Representative example of a *T16H2* silenced plant. The area of the vindorosine peak exceeds the vindoline peak.

Peak areas were calculated from vindoline and vindorosine EIC using the Xcalibur software. Obtained data was normalised by internal standard and sample fresh weight. Statistical analysis was performed and confirmed the observed change in vindoline and vindorosine accumulation (Figure 19). Although a strong decrease of vindoline was observed, levels of tabersonine, the substrate of *T16H*, were not significantly altered.

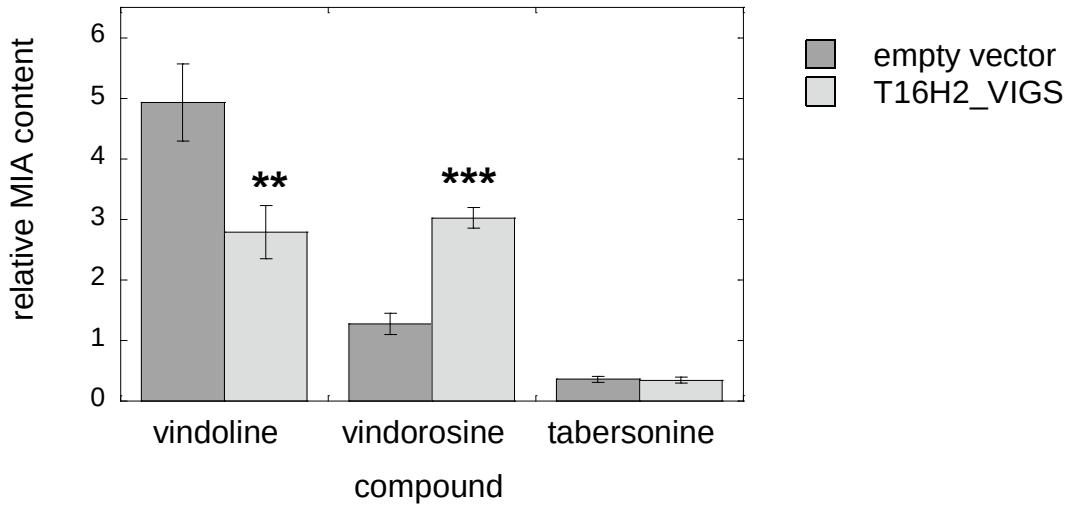


Figure 19 *T16H2* silencing effect on leaf vindoline, vindorosine and tabersonine content

Relative MIA content as mean peak area normalised by tissue weight and internal standard for vindoline, vindorosine and tabersonine in *T16H2* silenced plants in comparison to EV control plants (n=8). Error bars represent ± SEM. P-value was calculated using Student's *t* test (**p<0.01, ***p<0.001).

2.2.4.2 T16H2 silencing causes changes in accumulation of possible intermediates of vindoline and vindorosine biosynthesis

The data obtained from statistical analysis of extracted peak areas of the major metabolite clearly showed a significant decrease in the level of accumulating vindoline in favour of an equally strong increase of a new mass thought to be vindorosine, the de-methoxylated analogue of vindoline. The levels of other known alkaloids that can be identified using available standards such as secologanin, strictosidine and catharanthine, were unaffected by silencing as no significant changes could be detected.

The order of reactions in vindoline biosynthesis is well established. First the C16 position is hydroxylated by T16H and the same position is subsequently methoxylated. Although the next step (or steps) were unknown at the time, it had been shown that the known downstream enzymes of vindoline biosynthesis *NMT*, *D4H* and *DAT* were able to accept the 16-methoxylated and the demethoxylated tabersonine derivatives [121], suggesting the observed change in vindoline and vindorosine levels are due to a shift in metabolic flux caused by the silencing of *T16H2* (Figure 20).

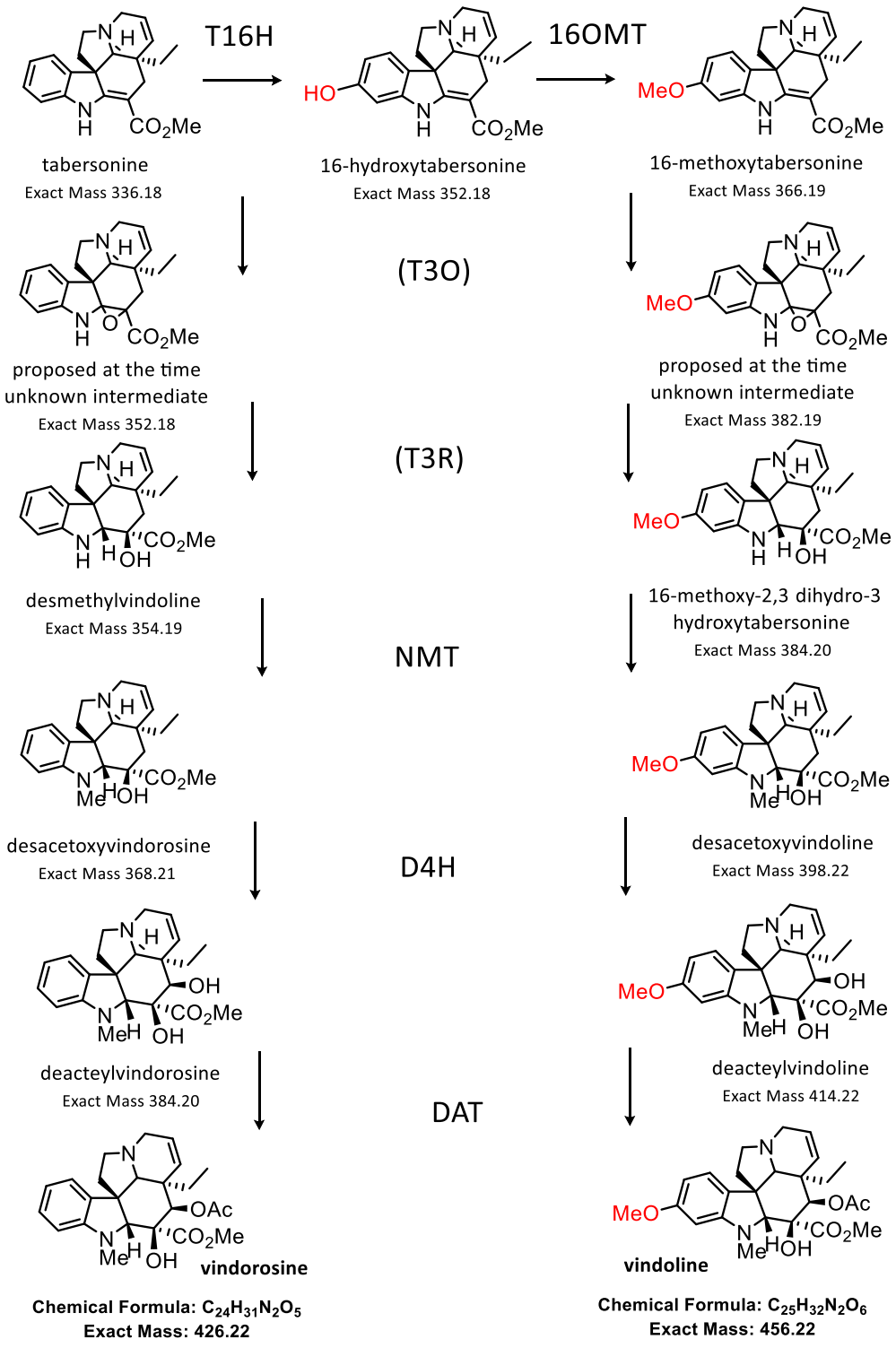


Figure 20 Vindoline and proposed vindorosine biosynthesis in *C. roseus*

Vindoline and vindorosine biosynthesis in *C. roseus* proceed from the joint substrate tabersonine. Enzymatic steps in brackets and the respective intermediates were not known at the time. At the time known genes *NMT*, *D4H* and *DAT* had been shown to accept the 16-methoxy and the unmethoxylated tabersonine derivatives [121,127].

To investigate in more detail the metabolic changes caused by silencing *T16H2,* automated peak extracted was performed on silenced and unsilenced samples (n=8) using the software XCMS on all obtained metabolite data raw files. All extracted peak areas were normalised by sample fresh weight. Interestingly, comparing the peak mean for each treatment, out of 616 extracted peaks the highest increase was that of the vindoline peak and the largest decrease of a extracted peak was for that of vindorosine, confirming the previous observations.

Vindoline and vindorosine biosynthesis proceed via several intermediate steps. None of the intermediates accumulates in *C. roseus* leaves at high levels and for none of them standards were available at the time. Two of the possible intermediates share the same calculated monoisotopic mass (Table 4).

Table 4 Intermediates of vindoline and vindorosine biosynthesis

Intermediates of vindoline and vindorosine biosynthesis with exact masses. Masses in bold occur in the vindoline as well as in the vindorosine biosynthesis.

vindoline biosynthesis intermediates	**monoisotopic mass**
16-hydroxytabersonine	352.1860
16-methoxytabersonine	366.1943
16-methoxy-2,3-dihydro-3-hydroxytabersonine	**384. 2122**
desacetoxyvindoline	398. 2278
deacetylvindoline	414. 2227
vindoline	456. 2227
vindorosine biosynthesis intermediates	
2,3-dihydro-3 hydroxytabersonine	354.1943
desacetoxyvindorosine	368. 2173
deacetylvindorosine	**384. 2122**
vindorosine	426. 2227

The dataset generated by applying the automated peak extraction was mined for peaks of masses corresponding to the intermediates in vindoline and vindorosine biosynthesis (Figure 20, Table 4). For all intermediates at least one corresponding peak could be identified (Table 5).

Table 5 Peaks corresponding to vindoline and vindorosine biosynthesis intermediates

Peaks extracted automatically using the XCMS software out of a LC-MS total ion chromatogram data set of EV and T16H silenced leaf tissue extracts, were investigated for the increase or decrease of peaks of masses that correspond to intermediates in the biosynthesis of vindoline and vindorosine. Data presented includes the *m/z*, Student's ttest (p-value), mean of corresponding peak area for *T16H2* silenced and EV silenced plants (n=8) and the calculated increase or decrease of the mean peak area. Grey underlined are peaks with significant changes (p-value <0.05). Bold is vindoline identified according to an available standard.

masses corresponding to vindoline biosynthesis					masses corresponding to vindorosine biosynthesis				
m/z	ttest	mean T16H2	mean EV	increase/ decrease	*m/z*	ttest	mean T16H2	mean EV	increase/ decrease
353.2108	0.01510	0.056	0.164	⇩-0.108	355.2383	0.01334	0.059	0.031	⇧0.0278
353.2233	0.16093	0.376	0.290	⇧0.0864	355.0168	0.71545	0.113	0.106	⇧0.0068
353.2476	0.08736	0.216	0.333	⇩-0.117	355.2612	0.23687	0.045	0.059	⇩-0.014
353.2497	0.26053	0.116	0.094	⇧0.0211	355.3352	0.01750	0.010	0.021	⇩-0.011
367.3066	0.77367	0.436	0.414	⇧0.0222	369.1218	0.61415	0.025	0.023	⇧0.0018
367.219	0.87304	0.060	0.058	⇧0.002	369.3115	0.58344	0.167	0.143	⇧0.0237
385.2615	0.02374	0.077	0.046	⇧0.0311	385.2615	0.02374	0.077	0.046	⇧0.0311
399.3247	0.03159	0.026	0.083	⇩-0.057	427.2513	0.01513	4.932	2.791	⇧2.1409
415.2446	0.00012	0.013	0.030	⇩-0.017					
457.218	**0.00001**	**1.275**	**3.025**	⇩**-1.749**					

To further investigate this two samples of silenced and control tissue extracts were subjected to an accurate mass analysis using a Surveyor LC system attached to an LTQ Orbitrap (Thermo). For the calculated exact masses of six of the ten possible intermediates, peaks were detected, with corresponding masses within a range of 3 ppm (Table 6). Four of those peaks correspond to intermediates significantly decreased in vindoline biosynthesis or significantly increased in vindorosine suggesting a redirection of flux from vindoline to vindorosine production in silenced leaves.

Table 6 Accurate mass analysis with Surveyor LC system attached to an LTQ Orbitrap

Accurate mass was determined of methanol extracts of T16H silenced and not silenced EV control tissues using a Surveyor LC system attached to an LTQ Orbitrap (Thermo). For all possible intermediates of vindoline and vindorosine biosynthesis exact masses were calculated and compared to peaks of accurate mass measured samples. For six masses a single corresponding peak could be detected.

	Formula	Calculated (m/z, [M+H])	Found (m/z, [M+H])	Error (ppm)
16-hydroxytabersonine	C21H24N2O3	353.1860	353.1850	2.74
desacetoxyvindorosine	C22H28N2O3	369.2173	369.2176	0.90
deacetylvindorosine	C22H28N2O4	385.2122	385.2120	0.48
vindorosine	C24H30N2O5	427.2227	427.2227	0.11
desacetoxyvindoline	C23H30N2O4	399.2278	399.2280	0.42
deacetylvindoline	C23H30N2O5	415.2227	415.2229	0.36

2.2.5 Effect of T16H2 silencing on other alkaloid biosynthesis genes

The silencing of *T16H2* gene expression was successful and led to a significant change in metabolite profile, clearly identifying its role in vindoline biosynthesis in leaves of *C. roseus*. Repeated attempts to quantify *T16H1* expression failed. If it is expressed at all, expression of this gene remained below the qRT PCR detection limit in both empty vector and *T16H2* silenced tissues.

The changes observed in metabolic phenotype of *T16H* silenced tissues can be explained by *T16H* silencing alone and do not suggests significant reduction in expression of other MIA pathway genes. Therefore it can be assumed that the redirection of flux does not lead to a possible feedback inhibition of the pathway. Instead expression of three exemplary chosen alkaloid biosynthesis genes both upstream (*G8H*) and downstream (*D4H* and *DAT*) of *T16H2* is increased in *T16H2* silenced tissues (Figure 21) as if to compensate for the reduced vindoline accumulation.

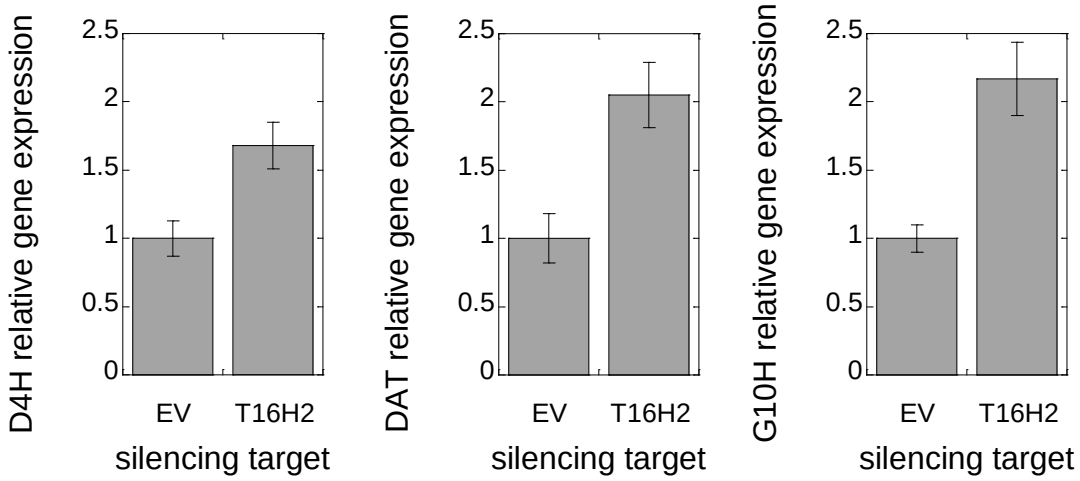


Figure 21 Gene expression of alkaloid biosynthesis genes in *T16H2* silenced tissues

Relative expression of alkaloid biosynthesis genes *G8H D4H* and *DAT* in *T16H2* silenced tissues compared to EV tissue (n=4). Expression was normalised to expression of CrRPS9. Data presented is the mean ± SEM.

2.3 Conclusion

A protein was discovered in *C. roseus* leaves with 82% identity in the deduced amino acid sequence to the known vindoline biosynthesis gene tabersonine 16-hydroxylase (T16H). The original *C. roseus T16H* (*T16H1*) gene was discovered in 1999. It had been cloned from a *C. roseus* suspension culture [41]. Rigorous *in vivo* characterisation identified the newly found gene to be a second gene encoding a functional *T16H (T16H2*).

In the scope of this thesis the *in planta* function and relevance of the *T16H2* gene in *C. roseus* leaves was investigated. It could be confirmed that indeed this new *T16H* is responsible for vindoline biosynthesis in *C. roseus* leaves, the primary site of vindoline and subsequently dimeric MIA production in *C. roseus* [42].

Successful silencing of the expression of the newly identified *T16H2* to approximately 20% of the level in empty vector control tissue, led to significant changes in metabolite profile. Most strikingly the level of vindoline is reduced by almost 60% as a result of silencing and instead, levels of the des-methoxy vindoline analogue, vindorosine are increased by a similar amount suggesting a redirection of metabolic flux. Vindorosine had been isolated in trace amounts from *C. roseus* for the first time in the early 1960s [128] and later research into vindoline/ vindorosine production in C*. roseus* linked the biosynthesis of both alkaloids.

It had been established that the downstream vindoline biosynthesis genes *NMT*, *D4H* and *DAT* are able to accept both the methoxylated and the unmethoxylated precursor as substrate [121,127] (Figure 23). The significant reduction of vindoline biosynthetic intermediates and increase of vindorosine intermediates support this hypotheis, but since authentic standards remain unavailable these results have to be interpreted not without caution. Despite this, the combined evidence of our results strongly suggest that *T16H2* and not *T16H1* is indeed the hydroxylase responsible for the first step in vindoline biosynthesis in leaves and represents a crucial control point determining the composition of accumulating alkaloids in *C. roseus* leaves and the fate of tabersonine.

An individual expression profile for the two *T16H* versions was not visible in the original MPRG expression data set [65]. Data resulting from the mapping of raw reads of two different publicly available transcriptomic studies [65,70] strongly suggests a preferential expression of *T16H2* in leaves, while *T16H1* is the predominantly expressed version in tissues like flower. The mapping confirms that *T16H1* is the dominating version expressed in suspension cultures (Table 3). This tissue specific expression of both *T16H* versions is in accordance with our own qRT PCR results and the qRT PCR results obtained by our collaborators both *T16H* genes in different tissues [42].

In retrospect, it is interesting to note, that in the publication that first reports the discovery of the *T16H1* gene [41], the possible existence of a second closely related gene is mentioned. Performing a genomic DNA Southern blot in this study showed hybridisation bands that led to the hypothesis of possibly at least two closely related genes existing. Yet given the much stronger expression of *T16H1* over *T16H2* in the cell suspension culture tissue, it is not surprising that *T16H1* rather than *T16H2* was cloned.

After the *C. roseus* genome was sequenced, as described in Chapter 4 we discovered that *T16H1* and *T16H2* are part of a gene pair located in close proximity next to each other [68]. This suggests a recent duplication event and will be discussed further at the corresponding chapter of this thesis.

The biosynthesis of vindoline and vindoline-related alkaloids such as vindorosine in *C. roseus* is highly complex with over 20 enzymatic steps, involving at least three different cell types, as well as several different subcellular localisations. The discovery of *T16H2*, a second functional tabersonine 16-hydroxylase, adds a novel dimension of complexity to the already complicated network of alkaloid production in *C. roseus*. It is the first example of two genes in *C. roseus*

that, despite catalysing the same enzymatic reaction, are transcriptionally regulated in a way that determines their individually different tissue dependent relevance and function. A clear distinction has to be made to genes with a simple redundant function. The herein presented evidence clearly shows how *C. roseus* uses sub-functionalisation of secondary metabolic genes. Unfortunately there is no literature available that would investigate if vindorosine and vindoline have different functions in for example protection of the plant from different types of herbivores. Therefore, it can only be speculated if the differing ratios of vindoline or vindorosine production holds any evolutionary advantage.

The deconvolution of the two *T16H* genes has implications for example for co-expression analysis. A set of genes co-expressed with either one of the two *T16H* versions is clearly going to be very different from a set of genes co-regulated with the combined expression of the two *T16H* genes. It raises the question if there are more genes in *C. roseus* alkaloid biosynthesis for which close paralogues with district tissue dependent expression exist.

3 Discovery of an enzymatic step in vindoline biosynthesis

Some aspects of this work were conducted in close collaboration with members of the O'Connor Lab Dr. Nat Sherden and Dr. Stephanie Brown. Results are published in (146)**. The publication can be found at the end of this document.**

Shotgun sequencing of the transcriptome (RNA-seq) allows the detailed study of gene expression within a certain tissue and/ or under a certain experimental condition. It allows us to observe and compare the individual expression profiles of genes in co-expression analyses [129]. Functionally related genes, such as the genes belonging to a certain metabolic pathway, often display a similar expression pattern in both primary [74,130] and secondary [76,131] metabolism, making co-expression analysis a powerful tool for gene discovery [75].

The search for a missing enzymatic step in vindoline biosynthesis that is presented in this chapter utilises co-expression analyses of a large *C. roseus* transcriptomic data set [65] to identify genes that are co-ordinately regulated with known genes of vindoline biosynthesis. The physiological function of the identified candidate genes was tested *in planta* utilising VIGS, a reverse genetics method for targeted transient gene silencing and resulted in identification of a promising gene candidate for the missing 16-methoxytabersonine 3-oxygenase step in vindoline biosynthesis. Subsequently 16-methoxytabersonine 3-oxygenase (T3O) was characterised *in vitro* and in a yeast expression system in collaboration with Dr. Nat Sherden and Dr. Stephanie Brown.

3.1 Introduction

Vindoline is one of the two necessary precursors for the formation of the valuable anti-cancer compounds vinblastine and vindorosine, uniquely produced by and laboriously extracted from *C. roseus*. Vindoline biosynthesis in *C. roseus* has been studied extensively, with the discovery of enzymatic activity for one of the pathway steps first reported almost 30 years ago [127].

Despite these efforts the full elucidation of vindoline biosynthesis was still lacking one crucial step at the start of this work. The order of reactions that leads to vindoline production has long been established [43,132] and the known enzymes of the vindoline pathway at the start of this work, the genes encoding those enzymes and the localisation of the enzyme within the leaf of *C. roseus* will be introduced briefly (Figure 22): Vindoline biosynthesis is initiated by two

enzymes predominantly expressed in the epidermis of leaves [47]. Tabersonine 16-hydroxylase (T16H) [41] and 16-hydroxytabersonine O-methyltransferase (16OMT) [34] hydroxylate and subsequently methylate the C16 position of tabersonine to yield 16-methoxytabersonine. The gene or genes responsible for the subsequent hydration reaction were unknown at the start of this work. After hydration, the next enzymatic step is the methylation of the indoline nitrogen catalysed by 16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase (NMT) [127]. The gene encoding NMT was discovered in 2010 [43]. All literature reporting the localisation of NMT in *C. roseus* predates this publication and therefore these localisation experiments relied on detecting protein activity by assaying the substrate with different cellular fractions obtained by ultracentrifugation. These studies report a mesophyll and not epidermis location for NMT [34,47] suggesting that either the epidermis or the mesophyll could be the location of the unknown hydration enzymatic step(s). However later studies confirmed a predominant expression of *NMT* in the epidermis using qRT PCR [48], which suggests that the hydration reaction also takes place in the epidermis. Finally the last two steps of vindoline biosynthesis, the hydroxylation of desacetoxyvindoline by desacetoxyvindoline 4-hydroxylase (D4H) [44] to form deacetylvindoline followed by O-acetylation by deacetylvindoline acetyltransferase (DAT) [45] to yield vindoline, are located in specialised mesophyll cells, the idioblasts and laticifers [60]. The subcellular localisation of vindoline/vindorosine biosynthesis has also been assessed. Utilising a combination of GFP imaging, bimolecular fluorescence complementation assays and yeast two-hybrid analysis showed that T16H, a cytochrome P450 enzyme, is localised to the endoplasmic reticulum (ER), while 16OMT homodimerises in the cytoplasm [132]. D4H and DAT are as well cytosolic [133] while NMT is associated with peroxisome membrane (Dave Liscombe and Vincent Courdavault, personal communication) (Figure 22).

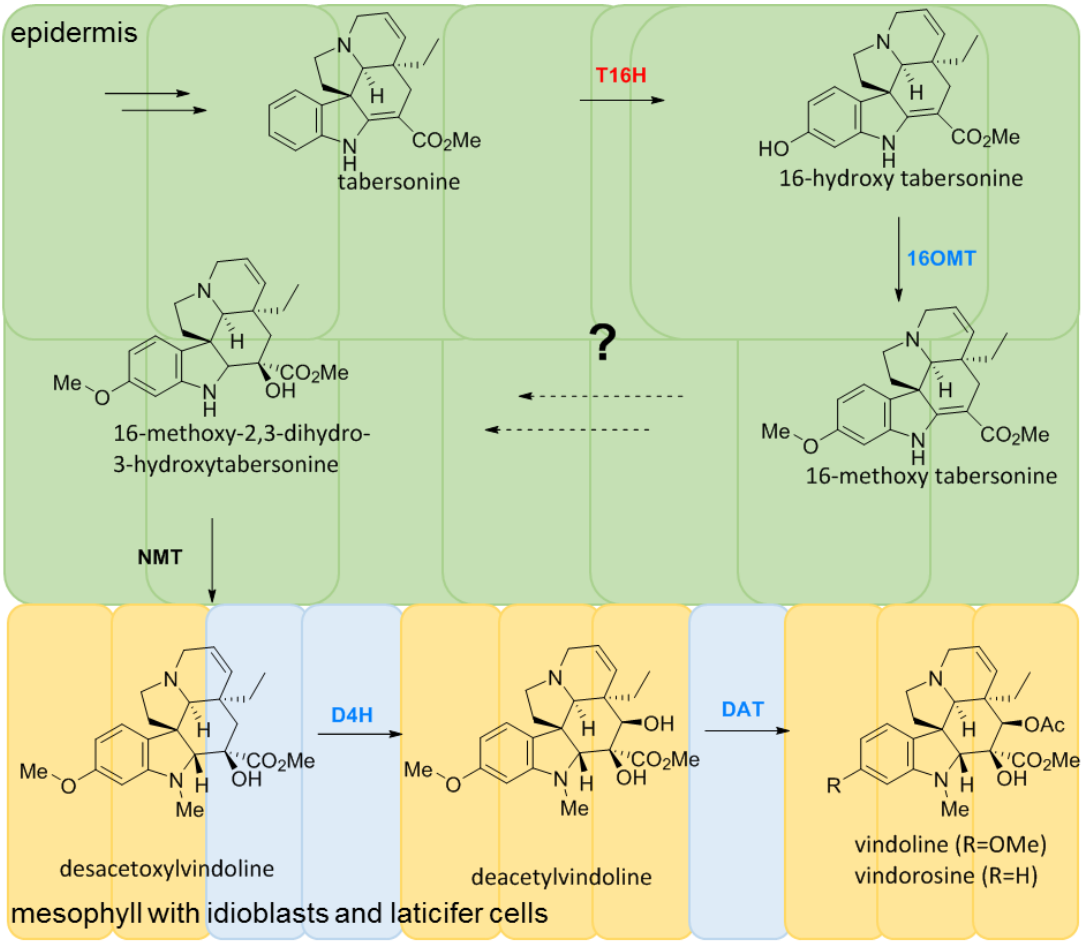


Figure 22 Known steps in vindoline biosynthesis with cellular and subcellular location

At the start of this thesis known vindoline biosynthesis genes had been identified as: *T16H*; tabersonine 16-hydroxylase, *16OMT*; 16-hydroxytabersonine O-methyltransferase, *NMT*; 16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase, *D4H*; desacetoxyvindoline-4-hydroxylase and *DAT*; deacetylvindoline 4-O-acetyltransferase. For genes in red enzyme location is the ER, in blue enzyme location is cytosolic and black depicts a localisation of the enzyme with the peroxisome membrane.

Catharanthine and vindoline are the two most abundant alkaloids in *C. roseus* [126]. The work reported in Chapter 2 conclusively describes that the *C. roseus* variety "SunStorm Apricot", predominantly employed in this thesis, contains a significant amount of an additional alkaloid, vindorosine. Vindorosine is a des-methoxy vindoline analogue and was first isolated in trace amounts from *C. roseus* in 1963 [128]. The biosynthesis of both compounds is closely linked. Silencing the first step in vindoline biosynthesis redirects the flux from vindoline towards the production of vindorosine in *C. roseus* "SunStorm Apricot" [42]. These observations are in accordance with previous publications where it had been demonstrated that the downstream

vindoline biosynthesis enzymes NMT, D4H and DAT are able to accept both the methoxylated and the unmethoxylated precursor as substrate [121,127] leading ultimately to the production of either vindoline or vindorosine (Figure 23).

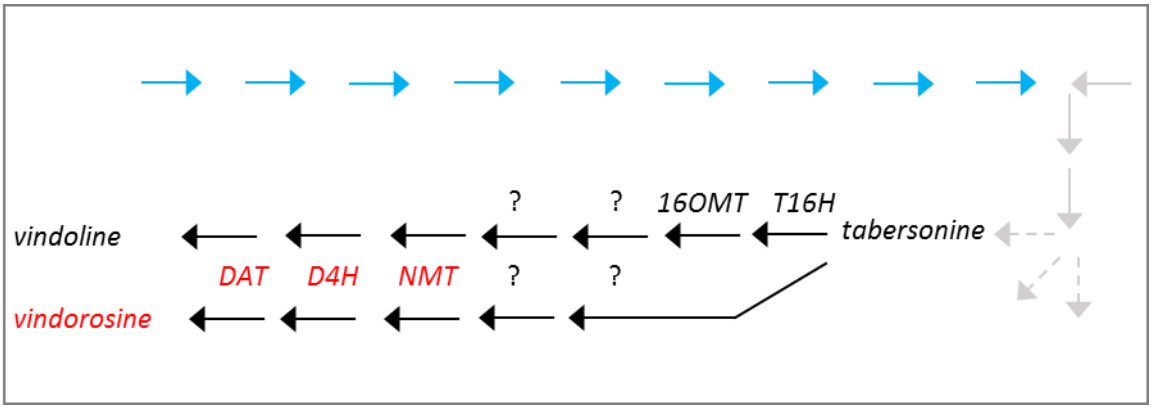


Figure 23 Vindoline/ vindorosine biosynthesis

At the start of this thesis known genes in the synthesis of vindoline and vindorosine in *C. roseus*. Arrows represent single enzymatic steps. Dashed arrows represent unknown number of enzymatic steps. Question marks represent unknown enzymatic steps. Blue (top row) arrows represent iridoid biosynthesis genes. Grey arrows represent downstream alkaloid biosynthesis genes. Black arrows represent then known vindoline/ vindorosine biosynthesis pathway genes. Genes in black are exclusive in vindoline biosynthesis: T16H, tabersonine 16-hydroxylase 1 (CYP71D12) and 16OMT, 16-hydroxytabersonine O-methyltransferase. Genes in red are active in both vindoline and vindorosine pathway: NMT, 16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase; D4H, desacetoxyvindoline-4-hydroxylase and DAT, deacetylvindoline 4-O-acetyltransferas.

3.1.1 Proposed reaction for the missing step in vindoline biosynthesis

The product of the enzymatic reaction carried out by the enzyme 16OMT– and the substrate for the missing step in vindoline biosynthesis– is 16-methoxy tabersonine. NMT, carrying out the next known enzymatic step after this missing step, accepts only the 2,3-dihydration product 16-methoxy-2,3-dihydro-3-hydroxytabersonine and not tabersonine [43]. However, the exact reaction mechanism yielding 16-methoxy-2,3-dihydro-3-hydroxytabersonine remained speculative.

Chemical studies investigating the total synthesis of vindoline all report the occurrence of an intermediate between 16-methoxy tabersonine and 16-methoxy-2,3-dihydro-3-hydroxytabersonine [134–137] although no such intermediate has been identified in plants. This intermediate suggests that the net hydration reaction most likely proceeds stepwise via two

reactions. As shown in Figure 24, it can be hypothesised that first an epoxide is formed from the C2, C3 alkene of 16-methoxy tabersonine. This epoxide spontaneously opens to the imine alcohol and a reduction of the imine leads to the formation of the final product, 16-methoxy-2,3-dihydro-3-hydroxytabersonine.

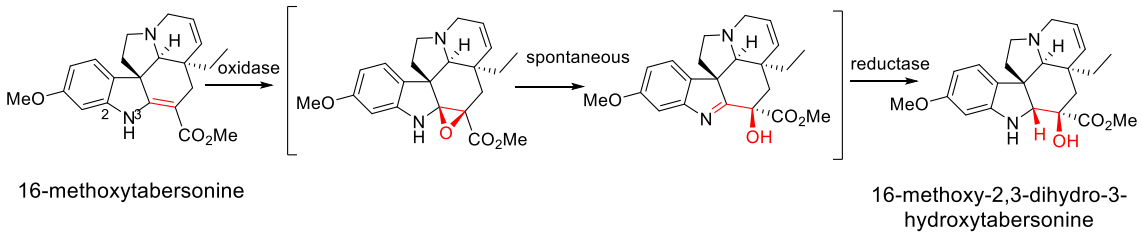


Figure 24 Proposed reaction for missing step in vindoline biosynthesis

The proposed two step reaction from 16-methoxy tabersonine to 16-methoxy-2,3-dihydro-3-hydroxytabersonine.

On the basis of these chemical studies, it was assumed that the biosynthetic process of the hydration likely requires an oxidising enzyme to first catalyse the formation of the epoxide/imine alcohol. For the following necessary reduction of the resulting imine alcohol, a second enzyme was proposed to be involved. To start our search, an oxidising enzyme, in particular an epoxidase, became the focus of the search for suitable gene candidates for the missing step in vindoline biosynthesis. Since epoxidation reactions can be catalysed by cytochrome P450s [138], special attention was paid to candidates annotated as a P450.

3.1.2 Strategies for identifying candidates for missing step in vindoline biosynthesis

Extensive research over the past decades had focused on finding the remaining unknown steps in *C. roseus* alkaloid biosynthesis. Co-expression analysis has proven to be a powerful gene discovery tool in plant systems. Based on the observation that genes belonging to a single specialised metabolic pathway are often expressed in a similar pattern [74], co-expression analysis is employed to identify gene candidates for unknown enzymatic steps in this pathway, in an approach sometimes called “guilt-by-association” [75].

In *C. roseus* the expression of MIA pathway genes strongly responds to MeJA elicitation [60] and the approach of “guilt-by-association” has been employed successfully by comparing elicited

and non-elicited tissues. For example the close co-regulation of a majority of the alkaloid biosynthesis genes in *C. roseus* [49] had previously enabled the discovery of genes such as iridoid synthase [30] and tabersonine 19-hydroxylase [79].

The strategy for the search of the missing steps in vindoline biosynthesis was led by two major hypotheses. First, the missing gene/s in vindoline biosynthesis would be co-regulated with the already known vindoline biosynthesis genes. Second, it is most likely that an oxidising enzyme, such as a cytochrome P450, is responsible for this undiscovered enzymatic step and that a subsequent reduction or reducing enzyme might also be required.

In consequence the candidate selection combined both hypotheses. Co-expression analysis of the available *C. roseus* gene expression data for various physiological and experimental conditions would provide a candidate list with genes whose expression profile is similar to that of known vindoline biosynthetic genes. Functional characterisation focused on candidates that belong to the cytochrome P450 enzyme class, proposed to be the most likely enzyme class to catalyse formation of an epoxide (Figure 24).

3.2 Results and Discussion

3.2.1 Transcriptomic data for co-expression analysis

For the co-expression analysis a dataset comprising expression values (log_2FPKM) of 33,287 contigs over 23 different *C. roseus* tissues and/ or treatments was employed [65]. This dataset is available at (http://medicinalplantgenomics.msu.edu/).

Since vindoline/vindorosine are not reproducibly produced in either suspension cultures or hairy roots [81], those tissues were excluded from further analysis. Consequently all co-expression analysis used up to eight *C. roseus* tissues and experimental conditions. For most analyses six tissues were employed: flower, mature leaf, young leaf, non-elicited seedling, seedling elicited with methyl jasmonate for 5 days and seedling elicited with methyl jasmonate for 12 days.

3.2.2 Initial data examination

The upstream region of alkaloid biosynthesis, the seco-iridoid pathway (Chapter 1, Figure 5), has been shown to be tightly co-regulated [49], which has facilitated the discovery of genes such as iridoid synthase [30] . An initial examination of the expression of known vindoline biosynthetic genes revealed that the very close co-expression is not fully maintained (Figure 25 A) in comparison to genes involved in the upstream seco-iridoid production (Figure 25 B).

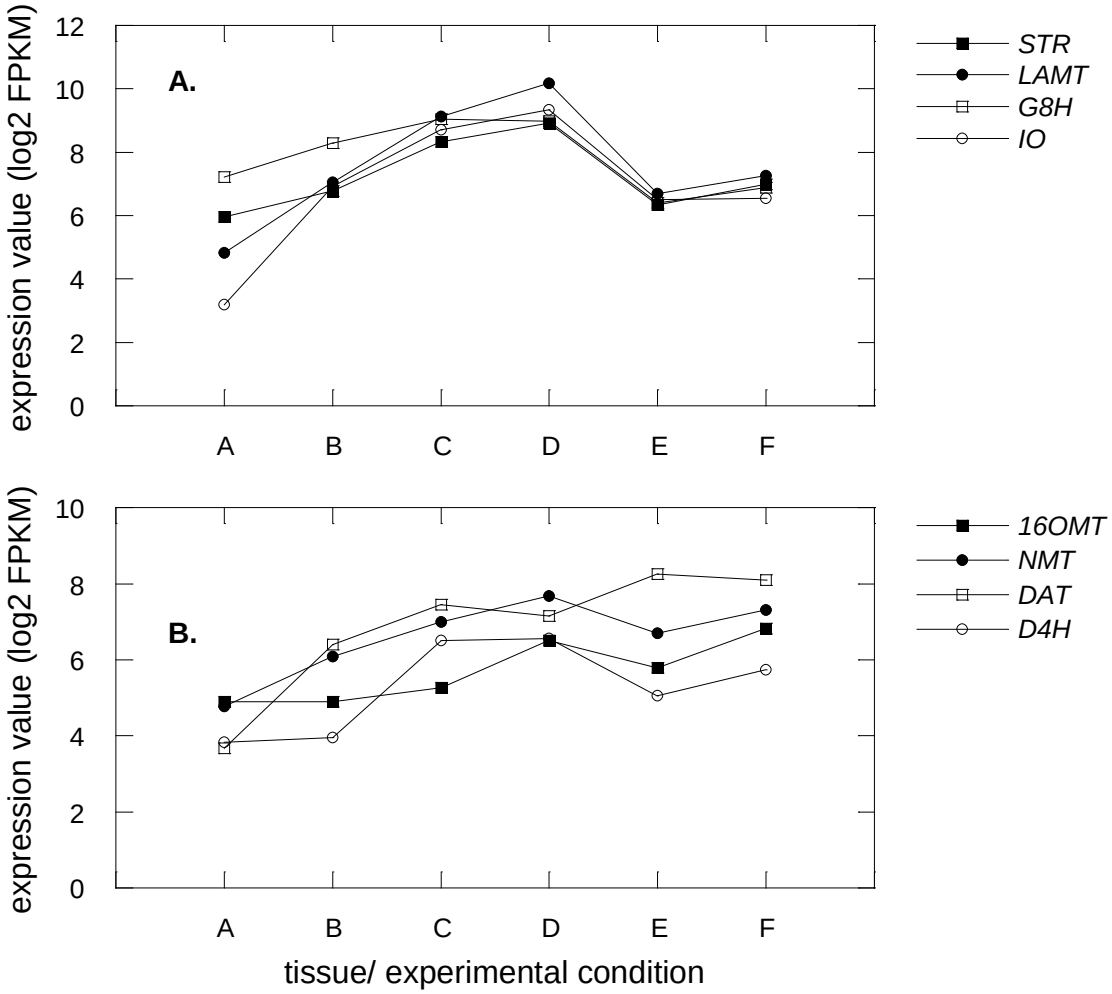


Figure 25 Expression of vindoline biosynthesis genes and seco-iridoid biosynthesis genes

Expression values (as log_2FPKM) of *C. roseus* genes. Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf) and F (young leaf). **A:** A selection of seco-iridoid pathway genes are displayed. **B:** A selection of vindoline biosynthesis genes are displayed.

3.2.3 Co-expression analysis of vindoline biosynthesis

Numerous software tools for gene expression analysis are available. Data in this work were primarily analysed using the MultiExperiment Viewer (MeV v.4.8) software. Co-expression analysis employs different clustering methods with a broad range of clustering algorithms being available. They apply either similarity or distance measure and can be further combined with additional parameters [139]. For this work several clustering methods were tested. Generally, the most productive method was hierarchical clustering (average linkage distance) applying different distance metrics such as Pearson correlation similar to the co-expression

experiments described in [140]. Known genes involved in the vindoline biosynthesis pathway were marked and it was observed to what extent different distance metrics and different tissue combinations produced a cluster of those genes. The goal was to identify a cluster that would contain all or most known vindoline biosynthesis genes and a minimal number of other uncharacterised co-regulated genes. However the outcome of these analyses were of limited success as vindoline biosynthesis itself is not very tightly coregulated (Figure 25). Candidates identified from those experiments and resulting VIGS experiments are reported in Chapter 5.

The set of VIGS candidates that ultimately led to the discovery of the missing step in vindoline biosynthesis originates from filtering the original large dataset to reduce the number of contigs. Filtering was performed by applying different thresholds for minimum expression value in certain tissues as well as minimum induction values by methyl jasmonate. Ultimately those contigs with an expression value (log_2FPKM) lower than 1.0 in flower, mature leaf, young leaf, and three seedling tissues were eliminated, since vindoline is predicted to be synthesised in these tissues. Since methyl jasmonate is a known elicitor of MIA biosynthesis [141], expression values of contigs in non-elicited seedlings were compared to the corresponding expression value in seedlings after 12 days of elicitation with methyl jasmonate. Only contigs with a ratio greater than 1.0 (elicited/non-elicited) were retained. Together, these steps reduced the total number of contigs from 33,248 to 6538. All previously identified vindoline biosynthesis pathway genes (*T16H*, *16OMT*, *NMT*, *D4H and DAT*) were still included in this data set. These 6538 contigs were named Subset_A, and screened for contigs annotated as cytochrome P450. This search resulted in 38 cytochrome P450 candidates (Figure 7).

Transcript ID	log_2FPKM young leaf tissue	non elicitated/ elicitated seedling	VIGS candidate name	revised gene annotation
cra_locus_7369_iso_2_len_1702_ver_3	2.09	2.38		
cra_locus_1753_iso_2_len_2512_ver_3	4.75	1.84	C44	
cra_locus_24716_iso_1_len_1549_ver_3	2.82	1.69		
cra_locus_4293_iso_2_len_2076_ver_3	5.69	1.63	C47	
cra_locus_5338_iso_3_len_1820_ver_3	3.22	1.52		
cra_locus_15539_iso_2_len_1747_ver_3	4.06	1.46		
cra_locus_9393_iso_1_len_941_ver_3	4.79	1.33		SLS (KF309242.1)
cra_locus_3243_iso_4_len_2217_ver_3	5.49	1.32		7DLH (KF415115.1)
cra_locus_5201_iso_5_len_1638_ver_3	4.16	1.3		
cra_locus_589_iso_4_len_2044_ver_3	4.34	1.28		
cra_locus_605_iso_4_len_1576_ver_3	6.51	1.26		IO (KF302066.1)
cra_locus_15915_iso_3_len_2758_ver_3	1.43	1.24		
cra_locus_7549_iso_4_len_1960_ver_3	4.18	1.23	**C20**	
cra_locus_11931_iso_1_len_427_ver_3	6.77	1.22		SLS like
cra_locus_1071_iso_4_len_2072_ver_3	5.97	1.18		
cra_locus_2442_iso_7_len_1193_ver_3	7.19	1.18	C30	
cra_locus_18082_iso_1_len_750_ver_3	1.51	1.14		
cra_locus_9461_iso_1_len_2170_ver_3	5.33	1.13		
cra_locus_671_iso_4_len_1679_ver_3	6.04	1.13		
cra_locus_1656_iso_6_len_2093_ver_3	3.45	1.13		
cra_locus_4585_iso_4_len_2055_ver_3	4.44	1.12		
cra_locus_3614_iso_5_len_2267_ver_3	4.66	1.11		
cra_locus_1227_iso_6_len_1706_ver_3	5	1.1	C48	
cra_locus_17827_iso_1_len_1524_ver_3	3.63	1.1		
cra_locus_3274_iso_2_len_1806_ver_3	2.3	1.1		
cra_locus_6313_iso_4_len_2019_ver_3	1.38	1.09		
cra_locus_4244_iso_1_len_2009_ver_3	3.77	1.08		
cra_locus_11820_iso_4_len_2037_ver_3	2.95	1.05		
cra_locus_6425_iso_2_len_1992_ver_3	2.74	1.05		
cra_locus_12789_iso_2_len_1965_ver_3	5.18	1.05		
cra_locus_47945_iso_1_len_897_ver_3	5.06	1.04		
cra_locus_14905_iso_1_len_1191_ver_3	4.04	1.04		
cra_locus_828_iso_9_len_1821_ver_3	6.35	1.03		
cra_locus_2317_iso_3_len_1999_ver_3	4.82	1.02		
cra_locus_1131_iso_1_len_1777_ver_3	5.12	1.02		
cra_locus_17344_iso_1_len_1088_ver_3	5.5	1.01		
cra_locus_11715_iso_1_len_2413_ver_3	4.08	1.01		
cra_locus_14103_iso_6_len_1909_ver_3	3.54	1		

Table 7 Transcripts annotated as cytochrome P450 contained in Subset_A

Filtering expression data of *C. roseus* as described in the text above yielded a subset of 6538 contigs that contained 38 contigs annotated as cytochrome P450. These cytochrome P450 genes were sorted by their individual response to elicitation by methyl jasmonate calculated as the ratio of expression in sterile seedlings against sterile seedling treated with methyl jasmonate for 12 days. A BLASTn search of the nucleotide sequence of each of the individual 38 contigs revealed that contigs cra_locus_9393, cra_locus_11931, cra_locus_3243, cra_locus_605 correspond to the known pathway genes *SLS*, *7DLH* and *IO*, which are involved in secologanin biosynthesis.

The nucleotide sequences of all 38 contigs were inspected manually and it was revealed that they contained four already known pathway genes: *7DLH*, *IO* and two partial fragments of *SLS*. This is unsurprising, considering the criteria that led to the selection of this set. As the genes encoding 7DLH [33] and IO [29]were discovered after the publication of the transcriptomic data set [65] those genes were still only annotated as cytochrome P450. The known pathway gene T16H is also a cytochrome P450 and had been manually removed prior to the extraction of the cytochrome P450 candidate genes. Of the remaining 34 contigs, five candidates (C30 (cra_locus_2442), C44 (cra_locus_1753), C47 (cra_locus_4293), C48 (cra_locus_1227) and C20 (in bold in Table 7) corresponds to *T3O*) (cra_locus_7549)) were chosen for immediate further analysis based on their high response to methyl jasmonate or tight co-expression with various MIA biosynthetic genes, with the expectation that if no positive results were obtained, additional candidates from the list would be tested.

3.2.4 Virus induced gene silencing of selected cytochrome P450 candidates

To determine whether any of the candidates, (C30 (cra_locus_2442), C44 (cra_locus_1753), C47 (cra_locus_4293), C48 (cra_locus_1227) and C20 (cra_locus_7549), were involved in vindoline biosynthesis, we used VIGS to assess the metabolic phenotype that results from silencing of these genes. The expected outcome of successful silencing of an enzyme directly involved in a secondary metabolite pathway is the accumulation of precursors and/or direct substrate to this enzymatic step as well as the reduction of products of this reaction. The intermediates in vindoline biosynthesis are known to accumulate only in very small amounts in *C. roseus* leaf tissue [42,121] making their detection difficult, however both vindoline and the upstream precursor tabersonine are clear peaks that can easily be identified using available standards.

VIGS experiments were conducted as described in Chapter 2. Briefly, competent *Agrobacterium* was transformed with a plasmid harbouring a 300-400 bp long section of the gene of interest. *Agrobacterium* was then used to infect young *C. roseus* plants. Leaf tissue of plants silenced with this construct were harvested and prepared for metabolic analysis.

Of the five tested candidates, one (candidate C20) gave the desired metabolic phenotype, implicating this gene in vindoline biosynthesis. Consequently, C20 was named 16-

methoxytabersonine 3-oxygenase (*T3O*). The results of silencing of the remaining four candidates suggested that they played no role in MIA biosynthesis and are discussed in Chapter 5.

3.2.4.1 Initial inspection of silencing results for C20, C30, C44, C47 and C48

Initial inspection of the total ion chromatograms (TIC) of LC-MS traces of tissue extracts from C30, C44, C47 and C48 showed no major changes in metabolite composition or content. However a first inspection of the total ion chromatograms (TIC) of LC-MS traces of C20/ *T3O* tissue extracts showed the clear emergence of a significantly increased peak compared to traces of EV control tissue (Figure 26).

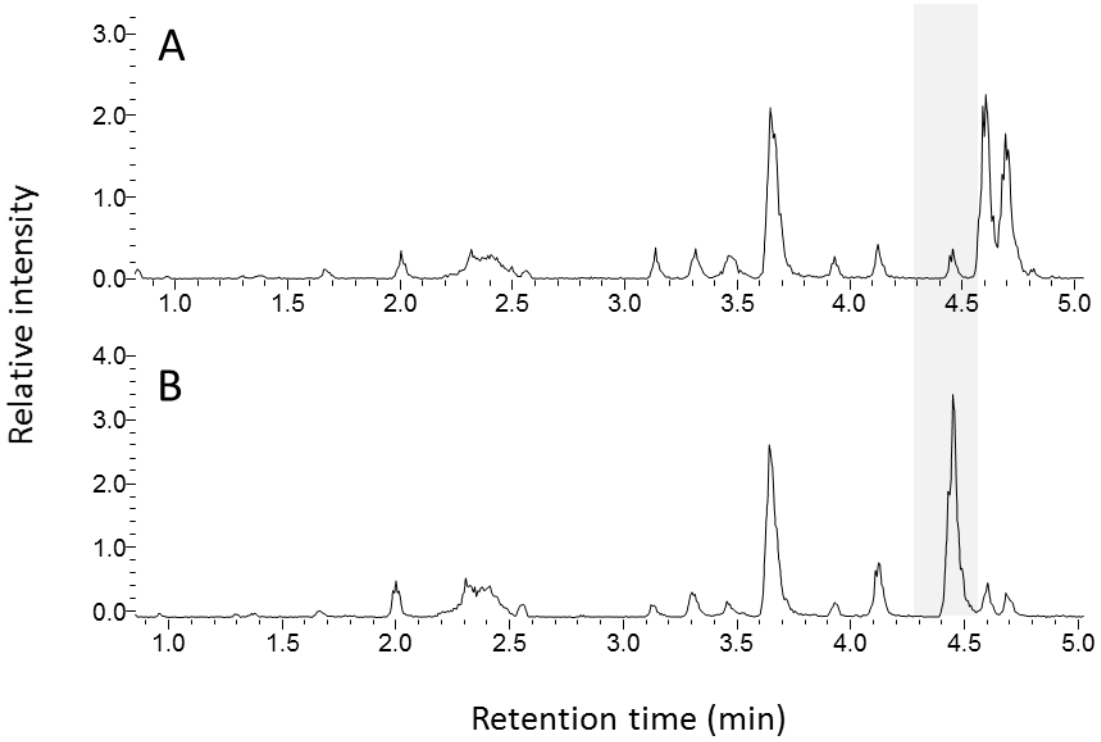


Figure 26 Comparison of TIC of an EV silenced and a *T3O* silenced leaf extract

The comparison of total ion chromatograms of two samples. **A**: Representative total ion chromatogram of an EV silenced leaf extract. **B**: Representative total ion chromatogram of a *T3O* silenced leaf extract with a strong increase of a peak with a mass of *m/z* 367, which corresponds to predicted substrate of T3O, 16-methoxytabersonine.

3.2.4.2 Metabolite analysis of silencing results for C20, C30, C44, C47 and C48

Analyses of six major accumulating alkaloids by manual extraction of peak area and normalisation of data by weight and internal standard revealed that silencing of the candidates C30, C44, C47 and C48 caused no major changes in metabolites (Chapter 5, Table 28).

However the peak that increases in intensity upon *T3O* silencing has a mass of *m/z* 367, which corresponds to the mass of the predicted substrate of T3O, 16-methoxytabersonine, suggesting that this gene likely encodes the missing step in vindoline biosynthesis (Figure 27).

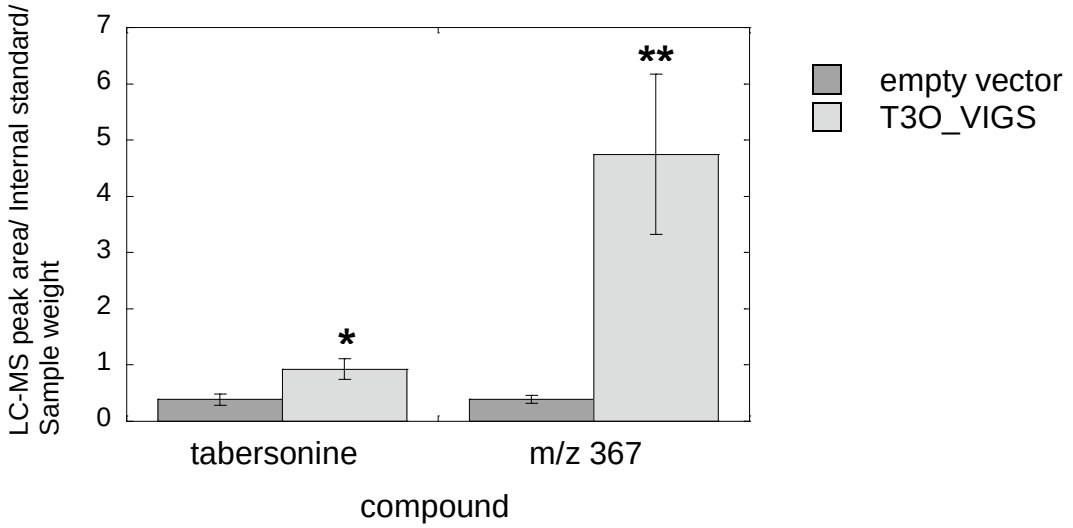


Figure 27 *T3O* silencing effect on leaf *m/z* 367 and tabersonine content

Relative MIA content as mean peak area for *m/z* 367 and tabersonine in *T3O* silenced plants in comparison to EV control plants (n=8). Error bars represent mean ± SEM. P-value was calculated using Student's *t* test (*p<0.05, **p<0.005).

Further strong evidence for the involvement of *T3O* in vindoline biosynthesis was provided by the significant decrease of vindoline (Figure 28). Additionally vindorosine, the des-methoxylated analog of vindoline, was also significantly reduced supporting the assumption that *T3O* is involved in both pathways (Figure 23). Three independent VIGS experiments were conducted for this candidate, with all resulting in similar significant metabolite changes.

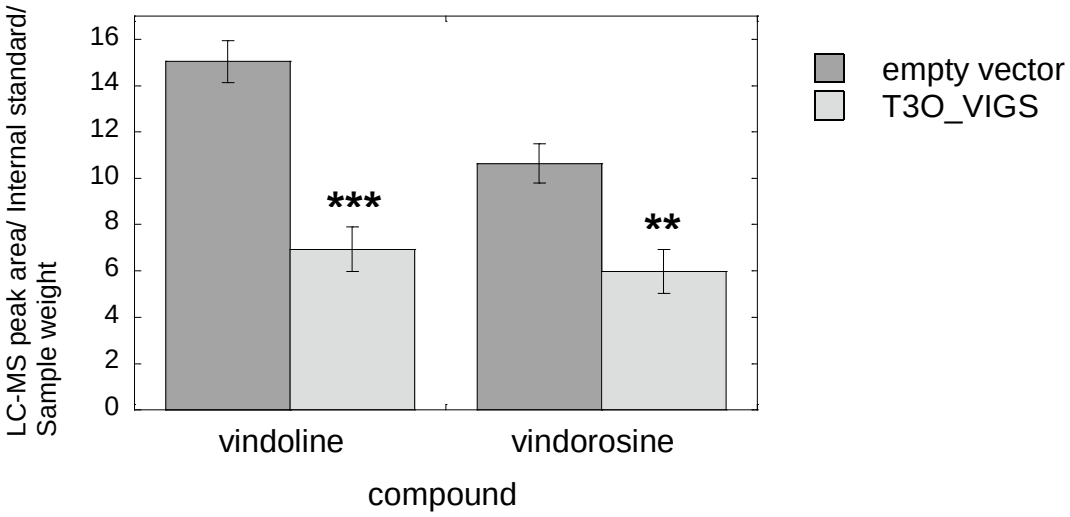


Figure 28 *T3O* silencing effect on leaf vindoline and vindorosine content

Relative MIA content as mean peak area for vindoline and vindorosine in *T3O* silenced plants in comparison to EV control plants (n=8). Error bars represent mean ± SEM. P-value was calculated using Student's *t* test (**p<0.005, ***p<0.00005).

3.2.5 Transcriptional down-regulation of candidate C20/ T3O

To confirm that the observed strong metabolic phenotype is in fact caused by successful *T3O* gene silencing, the expression of *T3O* was measured with qRT PCR in silenced tissues. In comparison to EV control tissues, *T3O* silencing led to a significant change in *T3O* gene expression (Figure 29). The transcripts of *T3O* were 70% decreased compared to EV control silenced tissues.

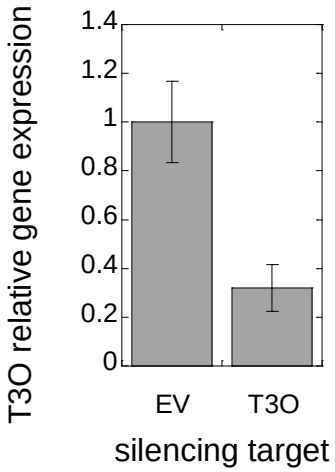


Figure 29 Gene expression of *T3O* in silenced tissue

Relative expression of *T3O* in tissue silenced for candidate *T3O* in comparison to tissue silenced with EV control (n=6). Expression was normalised to CrRPS9 expression. Data presented is the mean ± SEM.

3.2.6 In vitro assays confirm T3O role in vindoline biosynthesis

The successful significant downregulation of *T3O* expression and the resulting strong changes in vindoline concentration suggest an involvement of *T3O* in alkaloid biosynthesis in *C. roseus*. To further investigate the function of this candidate we proceeded to characterise this enzyme in a yeast expression system.

Numerous expression systems such as insect cells [142] and *E. coli* [108] have been reported for successful plant P450 expression. However, previous research in P450 enzymes conducted by the O'Connor lab [79] and other labs working on *C. roseus* P450 characterisation [29,42] has most successfully used the *Saccharomyces cerevisiae* strain WAT11, a yeast strain harbouring the *A. thaliana* P450 reductase [143], as expression system for plant P450 enzymes. Therefore, the heterologous expression of T3O was performed in this system.

T3O was cloned into the yeast expression plasmid pXP218 [144] and competent WAT11 yeast cells were transformed with this pXP218 plasmid. WAT11 transformed with an empty plasmid pXP218 served as empty vector control. The yeast culture was grown as 100 ml liquid culture in selective media. Protein expression was induced and yeast microsomes harbouring the P450 were isolated as described previously [145]. This was done for the yeast strain WAT11 with the *T3O* harbouring plasmid (WAT11+pT3O), as well as a yeast strain containing the EV plasmid (WAT11+pEV) that served as a negative control.

Both microsomal fractions were assayed in the presence of NADPH and the crude, aqueous fraction of a leaf extract of a *T3O* silenced tissue as substrate. After one hour of incubation, all of the accumulated mass of *m/z* 367 observed in the silenced tissues was consumed by the WAT11+pT3O microsomes, and a new peak with an *m/z* of 383 appeared as the product (Figure 30A). The mass of this product corresponds to the mass of the proposed epoxide/imine alcohol intermediate in vindoline biosynthesis (Figure 24). It was therefore concluded that the newly identified gene *T3O* does catalyse only the first step of the two step reaction from 16-methoxytabersonine to 16-methoxy-2,3-dihydro-3-hydroxytabersonine. No conversion of *m/z* 367 into *m/z* 383 by the WAT11+pEV microsomes even after 12 hours was observed (Figure 30B).

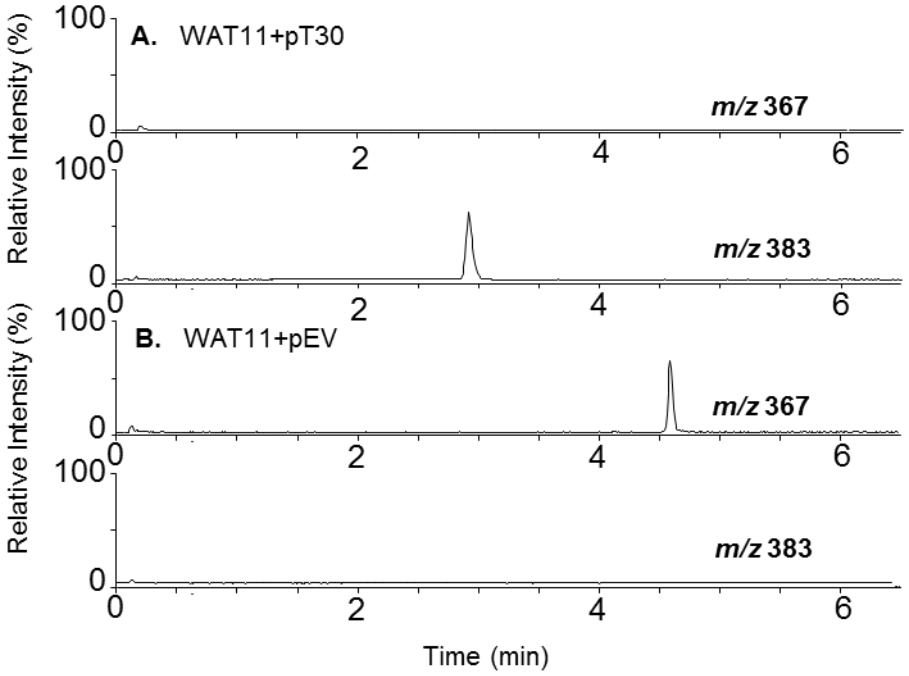


Figure 30 Microsomal assays using *T3O* silenced tissue extracts as substrate

Assays using *T3O* silenced leaf extract and yeast microsomal fractions in presence of NADPH. **A**: The peak of mass *m/z* 367 is consumed in the WAT11+pT3O microsomes containing assays. Instead a new peak appears with a mass *m/z* of 383. **B**: No conversion of *m/z* 367 by the WAT11+pEV microsomes containing assay into *m/z* 383 even after 12 hours was observed.

It had been speculated that the missing step in vindoline biosynthesis would be able to accept both tabersonine and 16-methoxytabersonine, as all subsequent genes in vindoline/ vindorosine biosynthesis are able to accept methoxylated and unmethoxylated

substrates [121,127] (Figure 23). Since tabersonine is not present in high quantities in the VIGS extracts, commercially available tabersonine was used in microsomal assays at 5 µM final concentration, an amount sufficient for LC-MS measurements. Theses assays confirmed that tabersonine is also an accepted substrate of T3O (Figure 31).

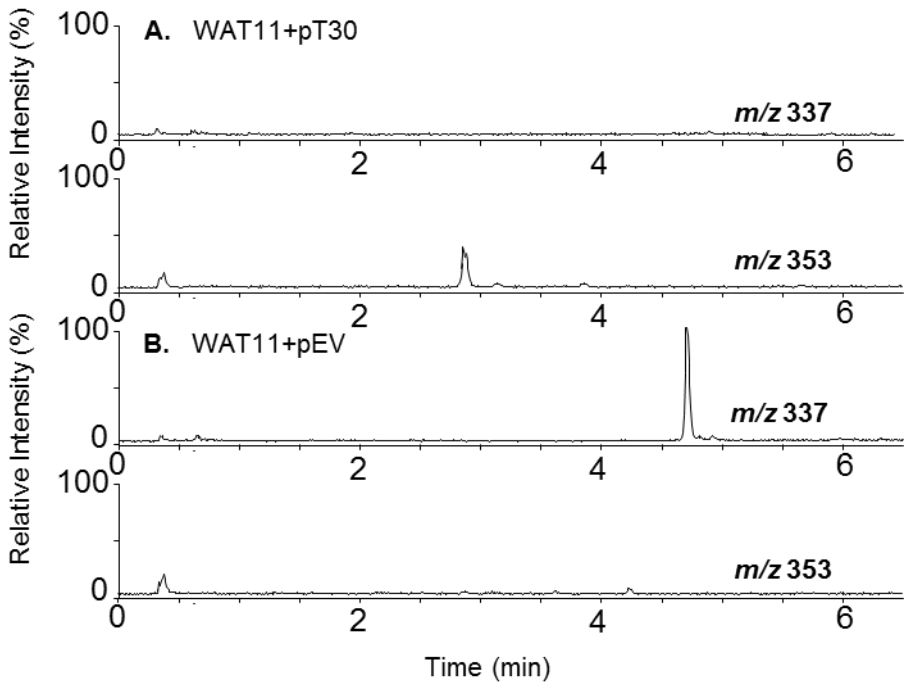


Figure 31 Microsomal assays using tabersonine as substrate

Assays using tabersonine and yeast microsomal fractions in presence of NADPH. **A**: The peak of tabersonine disappears in assays with WAT11+pT3O microsomes. Instead a new peak appears with an *m/z* of 353. **B**: No conversion of tabersonine by the WAT11+pEV even after 12 hours was observed.

3.2.7 Characterisation of substrate and product of T3O *reaction*

The combined results of the gene silencing experiments and the microsomal assays clearly suggested that *T3O* is involved in vindoline biosynthesis in *C. roseus*. We wanted to further confirm the reaction carried out by T3O by characterising the structure of the enzymatic product. Therefore, sufficient amounts of both substrate 16-methoxytabersonine and product needed to be available for subsequent structural investigations such as nuclear magnetic resonance (NMR).

Neither the T3O natural substrate in vindoline biosynthesis, 16-methoxytabersonine, nor the product of the T3O reaction were commercially available. In experiments performed in parallel with this work, a yeast strain (Strain A) containing all previously identified vindoline

biosynthesis genes (*T16H*, *16OMT*, *NMT*, *D4H*, *DAT*) (Figure 23) integrated into its genome including *CPR*, the P450 reductase from *C. roseus* [146] was engineered by Stephanie Brown (O'Connor lab).

Strain A was used to produce 16-methoxytabersonine from tabersonine by the action of T16H and 16OMT. Feeding of tabersonine to a liquid culture of Strain A had shown that this results in *m/z* 367 accumulation, corresponding to the product of T16H and 16OMT 16-methoxytabersonine. No further downstream vindoline biosynthesis intermediates accumulate in the culture, since the enzyme that acts after 16OMT is absent.

To produce 16-methoxytabersonine on a milligram scale, a liquid culture of Strain A was supplemented with 120 µM tabersonine and the formation of the 16-methoxytabersonine production was monitored over time by LC-MS. After all tabersonine was consumed, the product was extracted from the supernatant of the culture with ethyl acetate and analysed by NMR to confirm its structure by comparing it to reported literature values [147].

Strain A was also used to produce the T3O product. The *T3O* gene was introduced into Strain A by transforming competent cells of Strain A with the previously constructed *T30* containing pXP218 plasmid to create strain_A+pT3O. A corresponding negative control was created by transforming in the respective empty plasmid into strain A to create strain_A+pEV. Strain A with the empty vector control plasmid displayed predictably the same product as Strain A without any plasmid forming only 16-methoxytabersonine as identified by co-elution with the now available standard (Figure 32B and C). However, strain_A+pT3O led to the production of both *T3O* products 16-methoxy-2,3-dihydro-3-hydroxytabersonine and 2,3-dihydro-3-hydroxytabersonine (Figure 32A). Under these conditions, the ratio of products favoured the des-methoxy product derived directly from tabersonine (*m/z* 353) as opposed to the product derived from 16-methoxytabersonine (*m/z* 383), which is on pathway to vindoline.

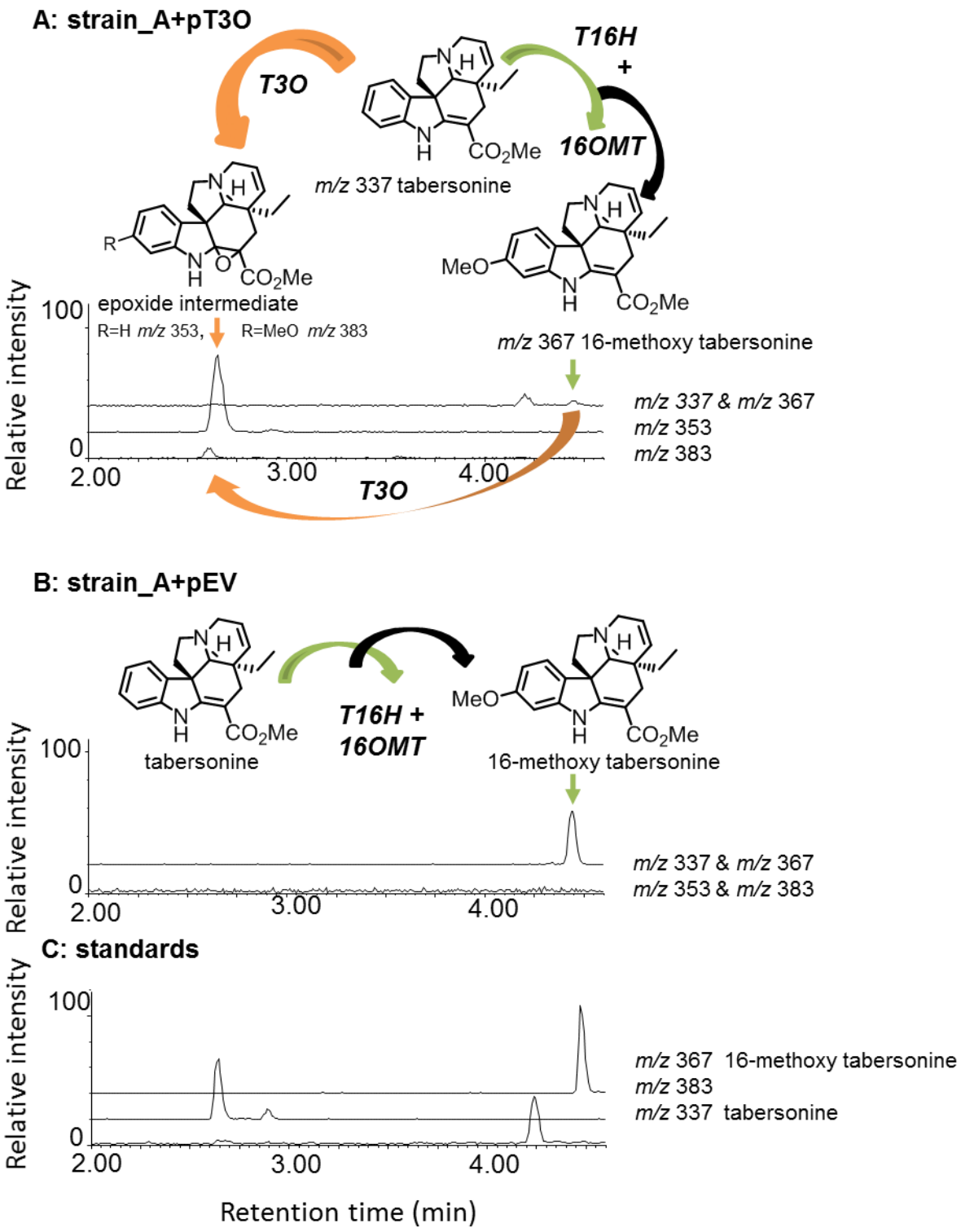


Figure 32 Products of yeast strain A, with and without *T3O*

Yeast strain A transformed with a plasmid containing *T3O* or an EV plasmid was supplemented with tabersonine and extracted ion chromatograms of expected products were monitored after 24 hours. Of the extracted ion chromatograms three are co-eluting with known standards, these are *m/z* 337 for tabersonine, *m/z* 367 for 16-methoxytabersonine and *m/z* 383 for the subsequently characterised T3O

product. **A**: Extracted ion chromatograms of the tabersonine supplemented strain_A+pT3O reveals that after 24 hours the substrate tabersonine (*m/z* 337) is almost fully consumed. Three new peaks emerge: *m/z* 367, which corresponds to the 16-methoxylated substrate of T3O as well as *m/z* 383 and *m/z* 353, likely representing the methoxylated and non-methoxylated products of T3O. **B**: Extracted ion chromatograms of the tabersonine supplemented strain_A+pEV show that after 24 hours culture the substrate tabersonine (*m/z* 337) has been completely consumed and instead 16-methoxylated tabersonine (*m/z* 367) is accumulating. No *m/z* 383 and *m/z* 353 are produced in a strain lacking the *T3O* gene. **C**: Extracted ion chromatograms of standards for *m/z* 337 (tabersonine), *m/z* 367 (16-methoxytabersonine) and *m/z* 383 for the later characterised T3O product.

To improve the yield of the vindoline pathway T3O product 16-methoxy-2,3-dihydro-3-hydroxytabersonine for structural characterisation, an alternative production method was used. A large scale liquid culture of the previously employed yeast WAT11 transformed with the *T3O* harbouring plasmid (WAT11+pT3O) was used to convert 16-methoxytabersonine via a single step reaction into the T3O product. As with 16-methoxytabersonine, the product was extracted from the media, purified and characterised as described below.

3.2.8 Structural characterisation of the T3O *product in vindoline biosynthesis*

The obtained T3O enzymatic product was analysed and characterised by IR, HRMS and $^{1}H/^{13}C^{15}N$ NMR by Nathaniel Sherden (O'Connor lab). It had been anticipated that either the epoxide or, if the epoxide spontaneously opens, the imine alcohol, would be observed. However, the NMR data did not correspond to either of these products. After an extensive review of the literature, Dr. Sherden recognised that 2-imine,3-alcohols on the tabersonine framework are able to undergo an acid catalysed rearrangement [148–151]. Gratifyingly the NMR resonances corresponded to literature reports of this rearranged product. This implies that, although not observable, the epoxide and imine alcohol forms initially but in absence of the subsequent enzyme that reduces the imine alcohol, the product rearranges to this new scaffold (Figure 33).

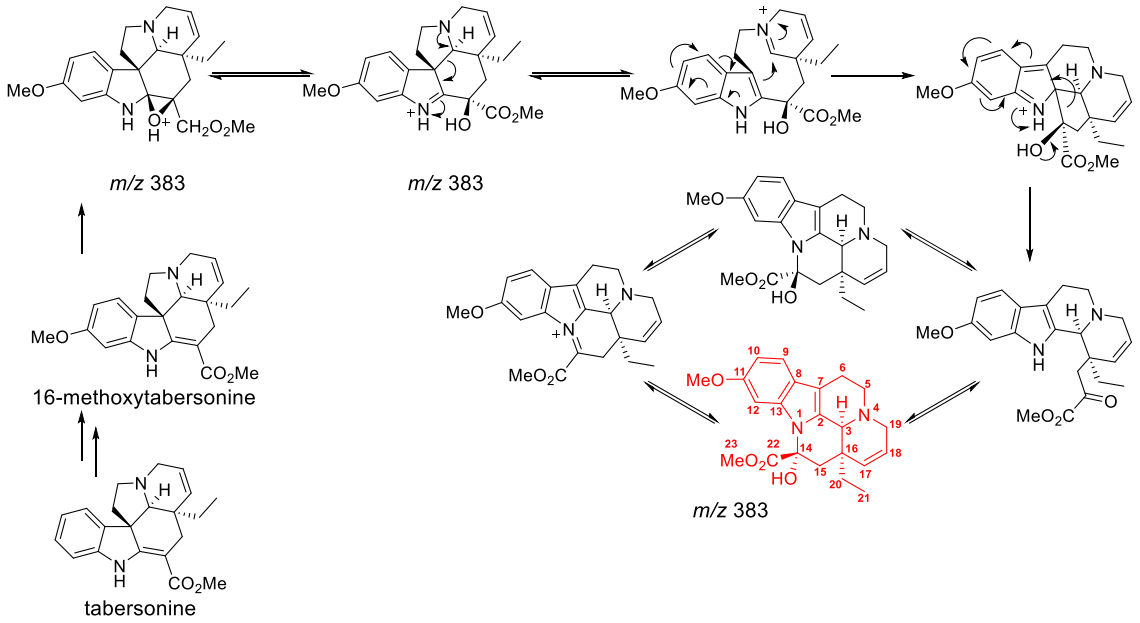


Figure 33 T3O reaction and observed rearrangement of reaction product

Acid catalysed rearrangements of the 2-imine 3-alkohol, in absence of the subsequent reductase, turn the initially by the action of T3O formed epoxide and into the eburnamine-vincamine skeleton. Proposal by Dr. Nat Sherden according to previously reported data [150].

This rearrangement had first been proposed for the biosynthesis of the pharmaceutically important vincamine from the tabersonine/ aspidosperma skeleton [150]. Vincamine, an alkaloid from the *C. roseus* close relative *Vinca minor*, is used to synthesise vinpocetine (Figure 34) first in the early 1960s. Vinpocetine, available under the trade name Cavinton since 1978, is a drug used for treatment of conditions such as cerebrovascular disorders, chronic stroke and degenerative senile cerebral dysfunction (Alzheimers) [152].

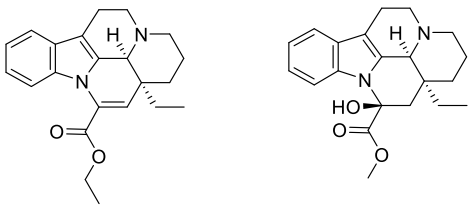


Figure 34 Chemical structures of vincamine and vinpocetine

The structural similarity between the rearranged intermediate and vincamine suggest that perhaps in *V. major* the reductase is missing completely and a homolog of *T3O* is involved in vincamine biosynthesis. These results also experimentally establish a long standing hypothesis, namely the biosynthetic relationship of the aspidosperma (tabersonine, vindoline)- and eburnamine (vincamine, vincopetine)-type alkaloids.

3.3 Conclusion

Vindoline is a *C. roseus*-derived natural product that is used in the semi-synthesis of valuable anti-cancer drugs [153]. The order of steps in vindoline biosynthesis from the precursor tabersonine is not arbitrary but follows a strict sequence of reactions. This sequence is largely dictated by the substrate specificity of the enzymes involved. At the start of this thesis five [34,41,43–45] out of the proposed seven [134–137] enzymes as well as the encoding genes were known (Figure 22, Figure 23).

In this work, a cytochrome P450 was discovered amongst a number of candidate genes that is required in the biosynthesis of vindoline and also the alternative shunt product vindorosine. It oxidises 16-methoxy tabersonine and tabersonine and was named 16-methoxytabersonine 3-oxidase or *T3O*.

The successful candidate was selected following two major leads:

Firstly we hypothesised that it was a cytochrome P450 enzyme, as most chemical studies suggested the missing step to involve an epoxidation [134–137], a reaction often carried out by P450 enzymes [138].

Secondly special attention was paid to the gene expression of the candidates in different *C. roseus* tissues and experimental conditions. Candidates were selected that showed a strong resemblance to the expression profile of known alkaloid biosynthesis genes in *C. roseus* as the alkaloid biosynthesis genes of *C. roseus* are known to be co-regulated [49].

VIGS experiments of five selected P450 candidates revealed that significantly reduced gene expression of cra_locus_7549 (candidate C20) by successful transient gene downregulation, leads to the significant decrease in levels of vindoline in leaf tissue of *C. roseus*. This decrease is accompanied by a significant increase in a compound with a mass and retention time identical to 16-methoxytabersonine, the expected substrate of the silenced *T3O* gene. The shunt product 16-des-methoxy-vindoline (vindorosine) is also significantly reduced while tabersonine is increased, suggesting that the candidate gene is involved in both the vindoline and vindorosine pathways.

Enzymatic assays with the successful candidate showed that T3O is able to enzymatically convert both tabersonine and 16-methoxytabersonie to their respective oxidised products as evidenced by the appearance of a compound with the expected mass on an LC-MS assay. Several yeast strains were employed to produce milligram quantities of the substrate and product of *T3O* in vindoline biosynthesis for structural elucidation. Structural studies of the product, however, revealed that in the absence of the subsequent reduction, the reaction product undergoes rearrangement. This rearrangement leads to formation of a product structurally similar to the alkaloid vincamine from the plant *V. major*, a close *C. roseus* relative. Based on extensive work reported in the literature, we can infer that the expected epoxide/imine alcohol was formed by T3O, but rearranged to the new product during purification. This discovery also opens up opportunities of gene discovery in vincamine biosynthesis following hypothesis that a *T3O* homolog in absence of a subsequent reductase could be involved in vincamine biosynthesis.

The discovery of *T3O* closes the gap in vindoline biosynthesis and will ultimately aid the understanding and possible reconstitution of the complete vindoline pathway. It furthermore provides evidence for a long-standing hypothesis about the biosynthetic relationship of the aspidosperma- and eburnamine-type alkaloids and might aid the discovery of vincamine biosynthesis in *V. major.*

The function of the remaining cytochrome P450 candidates (Table 7) still remains to be determined. Obviously the expression pattern of the four candidates that had been subjected to VIGS, especially their strong increase in expression after induction with methyl jasmonate suggests a potentially involved in stress response and/or plant defence. As it has recently been shown that beginning an investigation with as little as a single pathogen induced upregulated P450 in combination with metabolomics and co-expression analysis can lead to exciting discovery like that of the previously unknown 4-hydroxyindole-3-carbonyl nitrile pathway even in a well-studied organism such as *Arabidopsis thaliana* [154].

After it had become apparent that a second enzymatic step would be needed to complete the vindoline biosynthesis pathway, the same Subset_A from which the *T3O* gene had been successfully identified, was mined for contigs with a reductase annotation. They were tested using VIGS experiments and the results are presented in Chapter 5.

In parallel to this work and the resulting publication [146] a second publication reported the discovery of not only *T3O* but also of the subsequent enzymatic step carried out by *T3R* the tabersonine 3-reductase [48]. This discovery completed the search for missing genes and consequently for the first time reports the complete reconstitution of the tabersonine to vindoline pathway in the alternative host *Sacchomyces cerevisiae* [48].

4 Gene clustering in alkaloid biosynthesis in *C. roseus*

The major findings of analysing the *C. roseus* draft genome and the individual BAC assemblies for gene clustering in alkaloid biosynthesis have been published [68].

Biosynthesis of the valuable anti-cancer compounds vinblastine and vincristine by the plant *C. roseus* has been under investigation ever since the discovery of their chemical structure and anti-cancer activity in the mid twentieth century [19]. Precursors of vinblastine and vincristine, which are classified as bisindole alkaloids, are vindoline and catharanthine, produced primarily in above ground tissues. The approximately 30 step biosynthetic pathway, although studied for decades, is still not fully resolved [18], despite the many efforts to discover the enzymes responsible for alkaloid biosynthesis that have been made over the years. These efforts have included exploitation of major sequencing technique developments such as obtaining *C. roseus* EST data [59] and RNA seq data [65]. Recent discoveries, such as the 10 gene cluster in alkaloid noscapine synthesis in opium poppy [112], gave evidence that the genes of some plant secondary metabolic pathways are located in close proximity to each other on the genome. This raised the question whether such gene clustering could also be observed in *C. roseus,* a discovery that would greatly progress the elucidation of alkaloid biosynthesis in this model medicinal plant. To answer this question a draft genome for *C. roseus* in combination with sequence of BAC constructs was obtained, and the organisation of the monoterpene indole alkaloid biosynthetic genes was analysed.

4.1 Introduction

4.1.1 Bacterial artificial chromosomes

The ability to clone sections of eukaryotic DNA larger than 100 kb is crucial for many biological applications. With the introduction of the yeast artificial chromosomes (YACs) in the late 1980s, cloning and maintenance of DNA fragments of up to 1000 kb became feasible [155]. Apparent disadvantages of this system, such as high level of chimerism, insert instability and contamination of cloned insert DNA with yeast DNA [156], led to research into alternative systems and the development of a bacterial (*E. coli*) based system [157]. The low copy fertility (F-factor) plasmid of *E. coli* was modified by the addition of a multiple cloning site and universal promoters together with the addition of a chloramphenicol resistance gene to facilitate negative selection of non-transformed bacteria, resulting in the original bacterial artificial

chromosome (BAC) plasmid pBAC108L [158]. Since then, the original plasmid pBAC108L has been modified and optimised to suit different purposes, leading to the development of alternative plasmids such as the commercial pIndigo-BAC5-HindIII vector (Epicentre, USA) used in this study. This vector is strictly regulated to a single copy per cell and has been widely used for construction of plant BAC libraries.

A BAC genomic library combines a set of recombinant clones that contain all of the DNA sequence present in an individual organism. To construct such a library, high molecular weight DNA is digested with the appropriate restriction nuclease, in this case HindIII, and the resulting DNA fragments are cloned into the linearised pIndigo-BAC5-HindIII vector. Competent *E. coli* was transformed with the resulting plasmids (Figure 35).

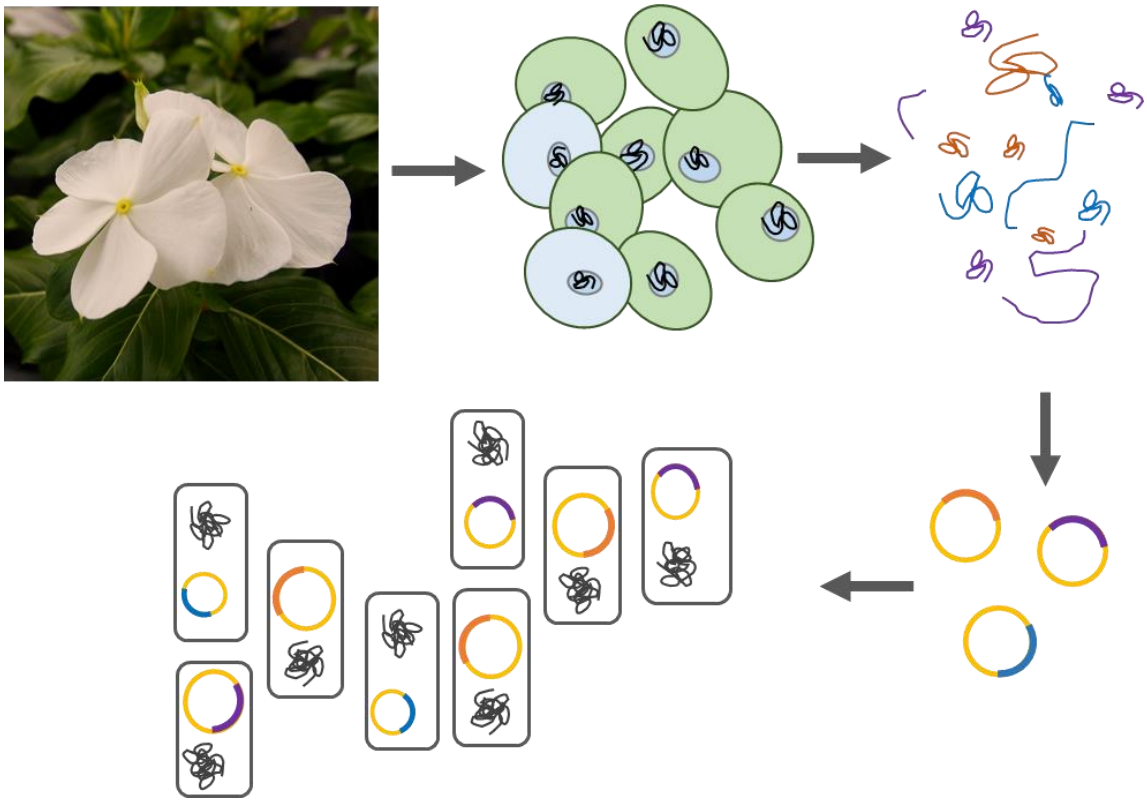


Figure 35 Schematic overview of BAC library construction

DNA is extracted from the cells and digested with HindIII, a restriction enzyme. DNA fragments are cloned into a vector and *E. coli* transformed with this vector.

The resulting library can either be sequenced fully and used to facilitate genomic analysis such as large-scale physical mapping [159] and genomic sequencing [160]. Alternatively, the clones can be sequenced sequentially using the known sequence of one BAC to screen the library for the overlapping/adjacent BAC in a technique called chromosome walking that can be exploited to sequence entire chromosomes or a specific genomic region of interest [161].

4.1.2 Plant whole genome sequencing

The number of available plant genomes has been growing steadily since the first successful plant genome project, the *Arabidopsis thaliana* genome, was completed in 2001 [160]. In 2012 alone, 13 new plant genomes were published, by the end of 2013 more than 50 different plant genomes were available, and with this trajectory likely to continue, many more are expected to be published over the coming years [162]. Plant genome size is variable with *C. roseus* having a moderate size genome in comparison with other plants (Figure 36).

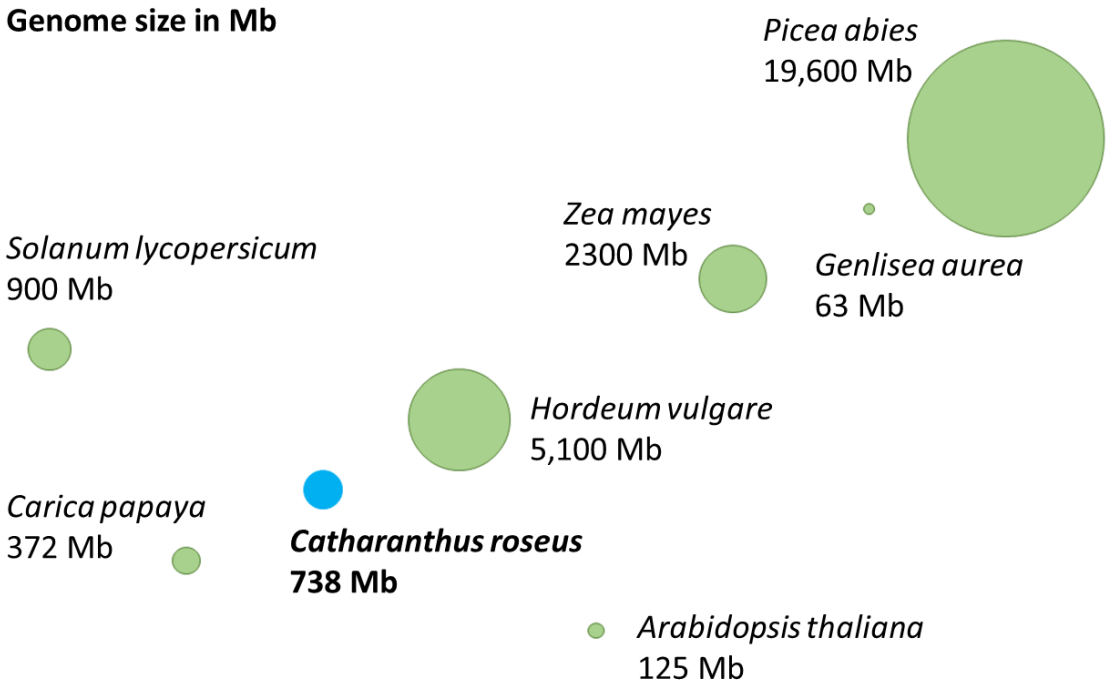


Figure 36 Genome size of different plants

Genome size for Genlisea aurea [163], *Picea abies* [164], *Zea mayes* [165], *Solanum lycopersicum* [166], *Hordeum vulgare* [167], *Carica papaya* [168] and *Arabidopsis thaliana* [160] in comparison to *C. roseus* [68,169].

Over the past decade sequencing costs have steadily declined while the speed in which data is generated has increased massively due to the development of high throughput sequencing [8]. Most of the recently published plant genomes have been sequenced using the Illumina technique that increasingly replaces the previously applied Sanger sequencing. Illumina paired-end sequencing, which was used to generate the genomic sequence for *C. roseus,* creates a read-pair for every sequenced DNA template. Although the full sequence of the DNA fragment is not known, the distance between the two ends is known (Figure 37). Sequencing is followed by assembly, the computational reconstruction of an original long sequence from short sequenced reads. First, individual short reads are assembled into longer continuous sections, so called contigs. With paired-end sequencing data this process is continued using the additional position information to join contigs into even longer scaffolds (Figure 37). Without an existing reference genome this process is called *de-novo* assembly.

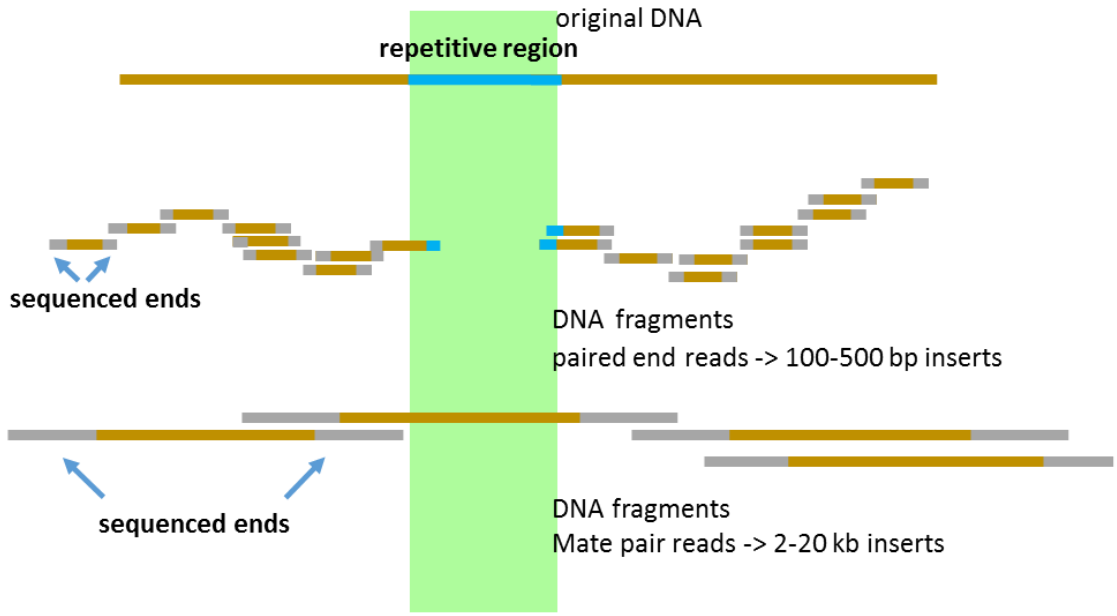


Figure 37 Paired-end sequencing and Mate-pairs

Paired-end sequencing results in both ends of a DNA fragment being sequenced creating two individual reads for each piece of DNA. Although the full sequence of the DNA fragment is not known the distance between the two ends is known and can be used to increase the assembly quality especially across repetitive regions that are generally difficult to sequence and assemble. Mate pair sequencing provides information for even longer DNA fragments improving assembly scaffold length strongly.

The quality of an assembly depends on many factors comprising, for example, the quality of the original template DNA, the quality, length and amount of created reads as well as the assembly process and the software used. It is difficult to define the success of a sequencing project or the quality of an assembly without a relevant standard for comparison [170]. However commonly used metrics are physical coverage which refers to how often the sum of all reads or read-pairs could cover the whole genome. This is different from the sequencing coverage or sequencing depth which refers to the average number of reads for a specific locus. As the most common way to describe the quality of an assembly the N50 value for contigs or scaffolds is used [162]. The N50 value refers to the size (in kb) above which half of the total length of the sequence set can be found. A higher N50 therefore is indicative of a higher percentage of longer contigs which ultimately determines the explanatory capacity of the assembly with respect to gaining insight into the genomic organisation of the organism sequenced. A low N50 can be caused by highly repetitive sequences, polymorphisms or simply low quality or missing sequencing data [170].

A typical N50 value for a *de novo* Illumina sequencing and assembly project, according to similar studies, was expected to be close to the value of 12.5 kb, reported for *Cucumis sativus* [171]. Other comparable genome sequencing projects, such as the sequencing of *Cannabis sativa,* relied on including further data obtained from, for example, mate-pair libraries to increase the N50 from 2 kb (only Illumina data) to 12 kb and reach an even higher N50 of 16 kb after further addition of large-insert 454 data [172]. It has to be noted that the N50 can be significantly higher for scaffolds containing genes, as the higher complexity of theses sequences enables better assembly and therefore for example in the *Cannabis sativa* study the N50 is 24.9 kb for gene containing scaffolds [172].

Apart from Sanger sequencing and Illumina, other sequencing platforms have been developed such as the commercial Affimetrix, the SoLiD and the Roche 454 platform as well as the relatively recently established PacBio sequencing. Since each sequencing technique has limitations, indeed most plant genome projects have used hybrid approaches creating an assembly by combining data from more than one of the available platforms [173].

4.1.3 Discovery of gene clusters in plant secondary metabolism

The possible clustering of genes involved in a specific biosynthetic pathway in plant secondary metabolism is a relatively recent observation. Today clustering has only been shown for a small number of plant secondary metabolic pathways. On the other hand the actual number of fully elucidated plant secondary metabolite pathways is, with less than 50, relatively small and therefore a prediction as to whether to expect more clustered pathways, or whether clustering remains the exception to the rule, is speculative [174].

Even in the absence of a genome sequence, trait guided discoveries of plant secondary metabolite cluster have been reported. It was for example possible to link a clear phenotypic trait, such as the accumulation of noscapine in *Papaver somniferum* [112] or the resistance to a herbivore in *Zea mays* [111], to a certain region of the genome by mapping and to consequently determine the multi-genic region responsible for the trait as a cluster of biosynthesis pathway genes. In both cases BACs of the trait identified regions were obtained and facilitated the discovery of the cluster of secondary metabolite pathway genes.

The availability of a genome sequence enables the targeted and detailed analyses of neighbouring regions of known pathways genes and has led to for example the discovery of gene clustering in triterpene biosynthesis in *Lotus japonicus* [175]. Triterpene biosynthesis had previously been shown to cluster for example from the work with *Avena strigose* (oat) [116–118]. The strategy for identification of triterpene gene clustering in *Lotus japonicus* was to search for "signature" enzymes [103], oxidosqualene cyclase (OSC) enzymes that form the skeleton structure of triterpenes, in the available genomic data. Consequently the genomic areas containing the OSCs was mined for potential gene clustering, resulting in the identification and characterisation of a triterpene gene cluster [175]. The mining of publicly available sequencing data for combinations of certain genes in close proximity has led for example to the discovery of the triterpene thalianol [176] and marneral [177] cluster in *Arabidopsis thaliana*. In conclusion the discovery of gene secondary metabolite clusters in plant genome data strongly depends on the amount and quality of the genomic data available, most importantly on the length of contiguous sequences available in the assembly.

4.1.4 Aim of the project

The goal of this sequencing project was to gain insight into the genomic organisation of alkaloid biosynthesis in *C. roseus* and to investigate possible gene clustering of the monoterpene indole alkaloid secondary metabolite pathway. Despite recent advancements in gene discovery in alkaloid biosynthesis in *C. roseus,* that have for example enabled the reconstitution of the seco-iridoid part of the pathway in yeast [25] and *Nicotiana benthamiana* [29], crucial steps that lead to the *C. roseus* typical alkaloid classes, and therefore to the precursors of valuable anti-cancer compounds vinblastine and vincristine, still remain unknown.

Reported plant gene clusters vary in size between 35 and several hundred kb [174]. The expected N50 scaffold size of an assembly based on Illumina HiSeq sequencing raw data of a single TruSeq genomic DNA library (as used in this case) is in the lower part of this range. Therefore to increase the chances of observing gene clustering in alkaloid biosynthesis it was decided to additionally obtain BACs, which cover a region of up to 150 kb surrounding a target gene, for selected alkaloid pathway genes. A BAC library was created and then screened for BACs containing the desired pathway genes. The sequence information obtained would also serve as a confirmation of the results of the assembly of the draft genome as the coverage of the BAC sequencing is much higher and a better assembly of this data is expected. It would most importantly enable us to investigate a relatively large neighbouring genomic regions of the chosen alkaloid pathway gene.

Ample transcriptomic data is available for *C. roseus* and mining this data, especially using co-expression analysis, has led to a great number of discoveries of so far unknown pathway components, regulators and transporters. The data is however limited in precision and completeness for example by not being precise enough to distinguish between the individual expression of close homologs. With the availability of a reference genome all existing transcriptomic data could be re-evaluated increasing the precision and explanatory power of co-expression analysis.

C. roseus is self-pollinating and diploid (2n = 16). The genome size was determined by flow cytometry and revealed a C-value of 1C = 0.76 pg, corresponding to 738 Mbp [169]. Overall, this inbred plant, with a moderate sized genome, is an excellent candidate for whole genome sequencing. In summary, obtaining one of the first medicinal plant genomes would not only

provide a valuable source for further pathway discovery in *C. roseus*, but provides a valuable resource for comparative plant genomics especially in the light of highly specialised and complex secondary metabolism.

4.2 Results and Discussion

C. roseus genomic DNA was extracted and sent to The Genome Analysis Centre (TGAC, Norwich, UK) were it was sequenced and assembled. In parallel *C. roseus* plant material was sent to Bio S&T Inc. (Montreal, Canada) where DNA was extracted and the BAC library was constructed and later the screening of the BAC library for specific targets was conducted.

4.2.1 Catharanthus roseus whole genome sequencing and assembly statistics

Genomic DNA of *C. roseus* "SunStorm Apricot" was extracted yielding 10.8 µg high molecular weight DNA. From this material a single TruSeq library was constructed with a 398 bp fragment size and sequenced on one lane Illumina HiSeq at The Genome Analysis Centre (TGAC, Norwich, UK), yielding 33 Gb of data that was subsequently assembled using ABySS [178] and applying a k-mer size of 71. This assembly consisted of 78 766 scaffolds larger than 200 bp and half of the assembly is covered by scaffolds larger than 26,256 bp (N50). The sum of all scaffolds larger than 200 bp is 5235 Gb. The longest scaffold in the assembly is 250 200 bp. The raw reads of the draft genome are available at NCBI SRA BioProject number PRJNA252611. The assembly (*C. roseus* SunStorm Apricot v.1.0) has been deposited under the accession JQHZ00000000 at the NCBI Whole-Genome Shotgun Sequence database. A database that is publicly available and searchable of the assembly *(C. roseus* SunStorm Apricot v.1.0) is located at The Medicinal Plant Genomics Resource (http://medicinalplantgenomics.msu.edu/).

4.2.2 *Catharanthus roseus* BAC library construction, candidates and screening

4.2.2.1 Construction of BAC library

The BAC library was constructed from plant material of the same *C. roseus* "SunStorm Apricot" plants used to obtain the material for genome sequencing to ensure maximal comparability. At Bio S&T Inc. (Montreal, Canada) the pINDIGOBAC-5 vector in *E. coli* strain DH10B was used to create the BAC library. The obtained library was pooled in a 96 well plate. Each of the wells contained approximately 500 clones. The BAC library had an approximately 10 fold coverage and an average insert size of 155 kb (Areti Karadimos, personal communication).

4.2.2.2 *Candidate genes for BAC library screening*

The obtained BAC library was to be screened for individual BACs that contain known genes of *C. roseus* alkaloid biosynthesis to increase the chance of discovery of possible gene clustering. Initially three genes, *ISY, T16H2* and *SGD*, were chosen as targets for screening the BAC library (Figure 38). Each of these candidates represent a key step in alkaloid biosynthesis: *ISY* for the monoterpene seco-iridoid pathway, *T16H* as the first committed enzyme of the vindoline pathway, and *SGD* which is the origin of the complex and poorly understood scaffold formation that results in the three classes of alkaloids, the iboga, the corynanthe and the aspidosperma, found in *C. roseus*. The reaction product of the deglycosylation reaction catalysed by SGD formulates an even greater diversity, as this intermediate enters multiple pathways to both indole and quinoline alkaloid types in the three plant families Apocynaceae, Rubiaceae, and Loganiaceae, though still none of the genes responsible for the formation of those diverse alkaloid types is known today [18]. In summary this makes *SGD* the most interesting gene for which clustering could potentially lead to discoveries that are applicable to not only *C. roseus* but other strictosidine derivative producing plants.

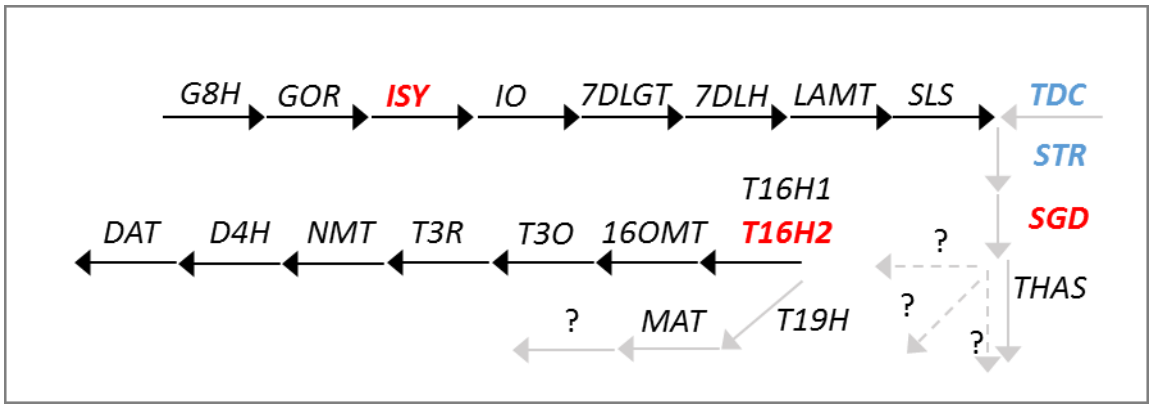


Figure 38 Genes of alkaloid biosynthesis including candidate genes for BAC library screening

Monoterpene indole alkaloid pathway in *C. roseus*. Missing genes are represented by question marks. Dashed arrows indicate unknown numbers of enzymatic steps. Genes for which a BAC was obtained in the initial screen are in red bold. Genes for which a BAC was obtained in the second screen are in blue bold. Black (top row) arrows represent iridoid biosynthesis genes: G8H, geraniol 8-hydroxylase; GOR, 8-hydroxygeraniol oxidoreductase; ISY, iridoid synthase; IO, iridoid oxidase (CYP76A26); 7DLGT, UDP-glucose iridoid glucosyltransferase; 7DLH, 7-deoxyloganic acid 7-hydroxylase; LAMT, loganic acid methyltransferase and SLS, secologanin synthase. Grey arrows represent downstream alkaloid biosynthesis genes: TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGD, strictosidine β-glucosidase; THAS, tetrahydroalstonine synthase; T19H, tabersonine/lochnericine 19-hydroxylase (CYP71BJ1) and MAT, minovincinine 19-hydroxy-O-acetyltransferase. Black arrows (middle row)

represent vindoline biosynthesis pathway: T16H1, tabersonine 16-hydroxylase 1 (CYP71D12); T16H2, tabersonine 16-hydroxylase 2 (CYP71D351); 16OMT, 16-hydroxytabersonine O-methyltransferase; T3O 16-methoxytabersone 3-oxigenase; T3R, 16-methoxytabersonine 3-reductase NMT, 16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase; D4H, desacetoxyvindoline-4-hydroxylase and DAT, deacetylvindoline 4-O-acetyltransferase.

Suitable primers were designed and tested for each candidate gene, with the aim to obtain a gene specific band of 200 to 500 bp when amplifying from genomic DNA. The suitable primers were send to Bio S&T Inc. (Montreal, Canada) to screen the *C. roseus* BAC library. For each candidate gene a BAC was identified and extracted from the library. Both an individual culture, as LB stab, and a small amount of extracted plasmid were received for each extracted BAC.

The initial set of candidates *ISY*, *T16H2* and *SGD* were chosen without any prior knowledge of the genomic context of alkaloid biosynthesis in *C. roseus*. Analysis of the draft *C. roseus* genome, as described below, revealed that the two pathway genes *TDC* and *STR* are located in close proximity on the same scaffold (Figure 46). This scaffold has a size of only 29 490 bp and it was decided that a BAC of this region would not only strengthen the evidence pointing to the existence of this cluster but, with an expected size of up to 150 kb, a BAC could additionally reveal possible more genes involved in this cluster. Only one further gene, *MATE*, a transport protein, is found on the *STR*/*TDC* genome scaffold as shown in (Figure 46). The genes are located on the scaffold in the following order: *MATE, TDC* and STR. To maximise the extension of the known region into the direction of the *MATE* gene, Bio S&T Inc. (Montreal, Canada) first screened the BAC library for positive clones of the gene *MATE.* Then, all positive identified BACs for *MATE* would, in a second round, be screened for the existence of the gene *STR* and a single BAC that contained *MATE* but not *STR* would be chosen to obtain a BAC with maximum extension into the genomic region beyond the *MATE* gene. For this purpose two further primer pairs specific for the genes *MATE* and *STR* were designed and sent to Bio S&T Inc. from which positive identified clones were sent back.

The results of the sequencing of the first *SGD* BAC and the subsequent assembly suggested that *SGD* was not located on this BAC, so it was decided to repeat this screen with an optimised primer pair. As at this time point the *C. roseus* draft genome was available, a second improved *SGD* specific primer pair was designed and the BAC library screened again for *SGD* and sequenced.

4.2.2.3 BAC extraction, sequencing and assembly

DNA was extracted for all candidates in house, except for the second SGD_BAC that was sent directly to Robin Buell's lab (MSU, East Lansing, USA). All BAC DNA was individually sequenced on a MiSeq with a 150 bp or 250 bp fragment size, after construction of a single TruSeq DNA library at Robin Buell's lab (MSU, East Lansing, USA).

Sequencing data was assembled after reducing the raw read number to only the first 50000 reads using the MIRA[179] assembler. The assembly was analysed regarding scaffold number, scaffold size and the combined length of unique scaffolds in comparison to expected insert size of the individual BAC (Table 8).

Table 8 BAC assembly results

Assembly of BAC sequencing data including number of scaffolds (cut-off value for scaffolds is 1 kb), number of scaffolds larger or smaller 2000 bp, sum of all scaffolds larger 2000 bp and largest scaffold of the assembly in comparison to experimentally determined expected insert size of individual BAC.

Target	No. of scaffolds	No. scaffolds > 2000 bp	No. scaffolds < 2000 bp	Largest scaffold	Sum of all scaffolds > 2000 bp	Size of BAC insert
SGD	2	2	0	93 247 bp	105 551 bp	105 000 bp
T16H2	6	4	2	59 670 bp	113 751 bp	125 000 bp
ISY	17	7	10	29 781 bp	79 090 bp	125 000 bp
STR	18	18	0	15 240 bp	78 250 bp	97 000 bp
MATE	22	13	9	16 340 bp	85 000 bp	90 000 bp
SGD	44	24	20	19 891 bp	99 178 bp	100 00 bp

4.2.3 Representation of alkaloid genes in draft genome and BAC assemblies

No nuclear genomic sequence for *C. roseus* or any other alkaloid producing plant was available at the beginning of this thesis. The analysis described below was aimed to determine the genomic organisation of the known MIA biosynthetic genes.

Encouraged by the high number of large scaffolds present in the genome, we investigated the completeness of the assembly by searching for all known MIA biosynthesis pathway genes (Figure 38). All published sequences of genes involved in the production of alkaloids in *C.*

roseus were searched manually and could be identified in the draft unannotated genome (Table 9). Manual inspection revealed that for *7DLH* the exons are distributed on two rather than one scaffold but those two scaffolds overlap and can be joined to form one continuous sequence. With the exception of *SGD* and *T3O* all pathway genes could be unambiguously assigned to a single scaffold revealing the individual exon intron structure for each known pathway gene and making the investigation of co-localised genes, at least to some extent, possible.

Table 9 *C. roseus* scaffolds of all alkaloid pathway genes known today

For all 25 known (to date) alkaloid pathway genes of *C. roseus* the gene abbreviation, full gene name, scaffold number and size, as well as the identifier (ID) of gene locus on individual contig are given. The genes *SGD* and *T3O* are not available (n.a.) in full length in the assembly.

Gene	Full Gene Name	Locus ID	Scaffold No.	Scaffold size (bp)
G8H	geraniol 8-hydroxylase	CRO_015220	3061275	29,875
GOR	8-hydroxygeraniol oxidoreductase	CRO_033830	3068899	39,235
ISY	iridoid synthase	CRO_033829	3047130	46,594
IO	iridoid oxidase (CYP76A26)	CRO_006321	3064716	28,101
7DLGT	UDP-glucose iridoid glucosyltransferase	CRO_027006	3050165	46,244
7DLH	7-deoxyloganic acid 7-hydroxylase	CRO_014616 & CRO_005949	2952847 & 3058188	5,196 & 40,675
LAMT	loganic acid methyltransferase	CRO_028497	3052816	52,328
SLS_1	secologanin synthase-like protein 1	CRO_009297	3045390	49,729
SLS_2	secologanin synthase/ cytochrome P-450	CRO_032842	3047714	65,404
SLS_3	secologanin synthase-like protein 2	CRO_012140	3071001	48,935
SLS_4	secologanin synthase-like protein 3	CRO_024556	3063455	76,903
TDC	tryptophan decarboxylase	CRO_006098	3045674	29,490
STR	strictosidine synthase	CRO_006099	3045674	29,490
SGD	strictosidine β-glucosidase	n.a.	n.a.	n.a.
T16H1	tabersonine 16-hydroxylase CYP71D12	CRO_017448	3064268	32,093
T16H2	T16H-like protein	CRO_017447	3064268	32,093
16OMT	16-hydroxytabersonine O-methyltransferase	CRO_004356	3051716	19,234
T3O	tabersonine 3-hydroxylase	n.a.	n.a.	n.a.
T3R	tabersonine 3-reductase	CRO_021541	3066548	34,599
NMT	16-hydroxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase	CRO_033266	3065019	10,334
D4H	desacetoxyvindoline-4-hydroxylase	CRO_012504	2969470	16,035
DAT	deacetylvindoline 4-O-acetyltransferase	CRO_020280	3060125	26,177
THAS	tetrahydroalstonine synthase	CRO_024553	3063455	76,903
T19H	tabersonine/lochnericine 19-hydroxylase	CRO_021082	2964965	34,685
MAT	minovincinine 19-hydroxy-O-acetyltransferase	CRO_005213	3067490	69,370

4.2.4 Gene content of pathway gene genomic scaffolds and BAC assemblies

For initial analysis, the unannotated draft genome was examined. Scaffolds identified to contain known pathway genes were examined for further gene content using BLASTn search against the *C. roseus* transcriptome available at the time [65] to mine for the presence and identity of neighbouring genes. From those transcripts, candidates that could potentially be missing steps in *C. roseus* alkaloid biosynthesis have been chosen and investigated either in the scope of this thesis or by other members of the O'Connor group. For example, a gene annotated as an alcohol dehydrogenase that is located in proximity to SLS was functionally characterised as an MIA enzyme involved in heteroyohimbine biosynthesis [37].

After this first extended analysis of the assembly and in preparation of the publication of this data [68] the genome was annotated and the raw reads of 8 out of the 23 tissues contained in the original transcriptomic dataset [65] were used to calculate FPKM values for this revised transcriptomic dataset by Dr. Jeongwoon Kim (Robin Buell's Lab, MSU, East Lansing, USA). Recently, in preparation of this thesis, the original analysis was repeated using the newly obtained annotation of the genome and the recalculated transcriptomic data as it is of much higher quality. In the original transcriptomic data, individual transcripts were often not full length and required individual manual inspection. In general, the obtained results of both analyses are the same. Furthermore the original mining was extended to pathway genes that had been discovered after the initial analysis, such as the vindoline pathway genes T3O [48,146] and T3R [48].

In total, 25 known pathway genes were employed in this analysis. The detailed investigation of their genomic context in combination with the BAC assembly data yielded 96 individual transcripts that are located in close proximity to the previously identified 25 *C. roseus* alkaloid pathway genes. Individual results are presented below with the results of the parallel analysis of the content of the BAC assemblies. If possible, the order of BAC scaffolds was manually determined using the BAC ends, overlapping regions of individual scaffolds, transcripts spanning more than one scaffold, and by BLASTn search against the draft genome scaffolds. To confirm individual findings, PCR from genomic DNA was conducted. The analysis presented below begins with the seco-iridoid pathway genes and follows the natural order of enzymatic steps in *C. roseus* alkaloid biosynthesis.

4.2.5 Geraniol-8 hydroxylase and 8-hydroxygeraniol oxidoreductase

The hydroxylation of geraniol by geraniol-8 hydroxylase (G8H) [28] represents the first committed step of the seco-iridoid pathway. This cytochrome P450 enzyme belongs to the CYP76 family and geraniol hydroxylation activity has been shown for several members of this family, for example, from *Arabidopsis thaliana* [180]. The product of G8H, 8-oxogeraniol, is converted by 8-hydroxygeraniol oxidoreductase (GOR) [29], which is a medium chain alcohol dehydrogenase, to the linear dialdehyde 8-oxogeranial (Figure 39).

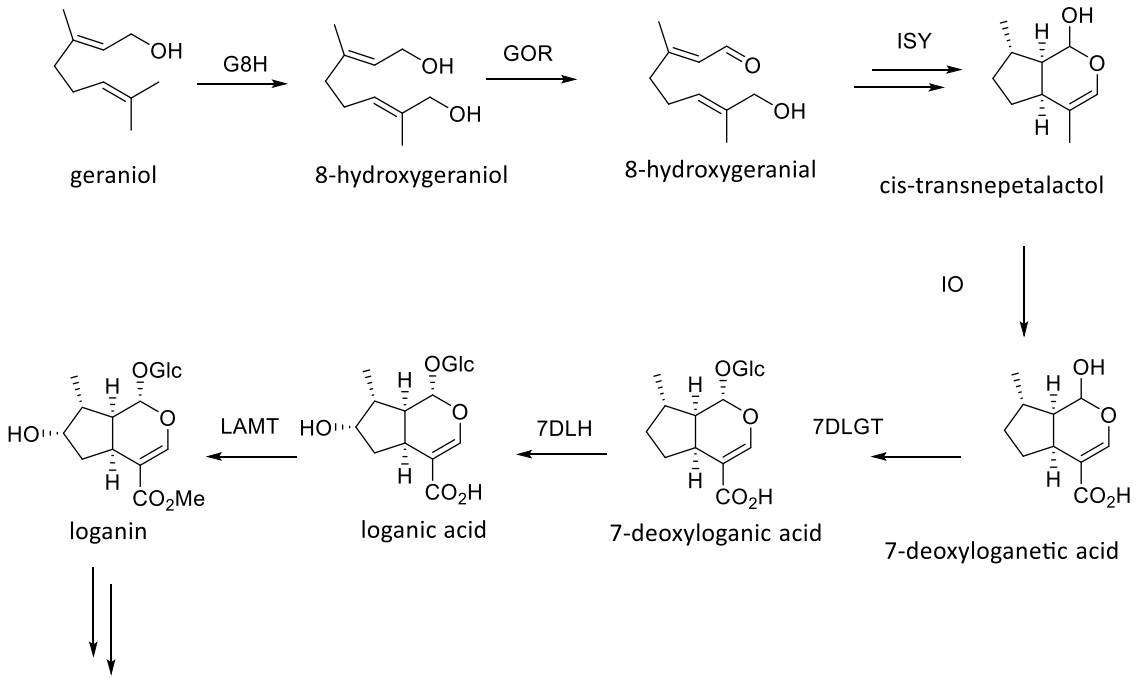


Figure 39 Monoterpene indole alkaloid pathway in *C. roseus (Part I)*

Seco-iridoid section of MIA pathway from geraniol to loganin: G8H, geraniol 8-hydroxylase; GOR, 8-hydroxygeraniol oxidoreductase; ISY, iridoid synthase; IO, iridoid oxidase; 7DLGT, UDP-glucose iridoid glucosyltransferase; 7DLH, 7-deoxyloganic acid 7-hydroxylase.

The *G8H* scaffold contains only one further transcript, not a known alkaloid pathway gene and the *GOR* scaffold contains six further transcripts none of which is a known alkaloid pathway gene (Table 10). Based on the annotation, it seems unlikely that any of these genes adjacent to *G8H* and *GOR* play a role in the MIA pathway.

Table 10 Scaffold containing alkaloid biosynthesis gene geraniol 8-hydroxylase

Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway gene in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
G8H	3061275	29,875	CRO_015221	DUF1649 domain containing protein
			CRO_015220	**geraniol 8-hydroxylase**
GOR	3068899	39,235	CRO_033830	**8-hydroxygeraniol oxidoreductase**
			CRO_004421	aspartate aminotransferase
			CRO_004419	Glycosyl hydrolase family protein
			CRO_004420	Glycosyl hydrolase family protein
			CRO_004422	hypothetical protein
			CRO_004424	RNA polymerase I-associated factor PAF67
			CRO_004423	Tetraspanin family protein

4.2.6 Iridoid synthase

The product of GOR, 8-oxogeranial, is converted by iridoid synthase (ISY) to a mixture of cis-trans-iridodial and cis-trans-nepetalactol [30] (Figure 39). For the *ISY* gene, the draft genome data as well as the corresponding BAC (ISY_BAC) were analysed. The genome scaffold containing ISY contains six further transcripts (Table 11).

Table 11 Genome scaffold containing alkaloid biosynthesis gene iridoid synthase

Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway gene in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
ISY	3047130	46,594	CRO_025458	alpha/beta-Hydrolase protein
			CRO_025459	cytochrome B5 isoform B
			CRO_025460	glycine-rich protein
			CRO_033829	**iridoid synthase**
			CRO_025461	iridoid synthase paralog
			CRO_025462	Protein of unknown function (DUF620)
			CRO_025463	RNA binding (pumilio)

The combined ISY_BAC assembly was only approximately 79,000 bp, smaller than the previously experimentally predicted insert size of 125,000 bp for this BAC (Table 8). However in combination with the draft genome scaffolds, the size of the covered region could be increased by manual curation of the assembly to about 160,000 bp (Table 12). Importantly the draft genome scaffold for *ISY* and the ISY_BAC scaffolds are close to identical. This accordance was observed for all BAC data and strengthens the validity of both assemblies. A number of interesting observations were made from the annotations of the genes on these scaffolds, and are discussed further below.

Table 12 ISY_BAC assembly scaffolds

Size of ISY_BAC scaffolds larger 2000 bp, corresponding scaffolds with size in bp, transcripts contained and transcripts functional annotation. In bold is the known pathway gene *ISY* (CRO_025460/ JX974564.1). Its neighbouring gene CRO_025461 represents a duplication of the former.

BAC Scaffold/size in bp	**Corresponding Scaffold/size in bp**	**Transcripts contained**	**Functional annotation**
c3/29781	3070152/ 76482	CRO_015511	adenine phosphoribosyl transferase
		CRO_015512	Family of unknown function (DUF566)
		CRO_015513	Glycosyl hydrolase family 35 protein
		CRO_015508	LOB domain-containing protein
		CRO_015509	nucleic acid binding
		CRO_015510	ROOT HAIR DEFECTIVE 6-LIKE
		CRO_015507	salt tolerance zinc finger
c1/27294	3068236/ 44765	CRO_029922	ankyrin repeat family protein
		CRO_029919	casein kinase alpha
		CRO_029921	Peroxisomal membrane 22 kDa (Mpv17/PMP22)
		CRO_029920	Syntaxin/t-SNARE family protein
		CRO_029918	TBP-associated factor
c2/22015	3047130/ 46594	CRO_025458	alpha/beta-Hydrolases superfamily
		CRO_025459	cytochrome B5 isoform B
		CRO_025460	**iridoid synthase**
		CRO_025461	iridoid synthase paralog
		CRO_025462	Protein of unknown function (DUF620)
		CRO_025463	RNA binding (pumilio)
c5/7939	3068236/ 44765	already contained in c1	
c6/6467	3068236/ 44765	already contained in c1	
c7/4986	3047130/ 46594	already contained in c2	
c4/4041	3068236/ 44765	already contained in c1	

According to the analysis of the genomic data, ISY is accompanied by a second gene CRO_025461, an iridoid synthase paralog. The *C. roseus* genome contains six members of the progesterone 5 β-reductase (P5βR) family homologous to the actual iridoid synthase gene (Table 13). Four out of the six members of this family have been shown to accept 8-oxogeranial as substrate in vitro, though their function *in planta* remains unclear [181].

Table 13 Sequence similarity of *C. roseus* progesterone 5 β-reductase enzymes

Sequence similarity in % based on nucleotide alignment of six *C. roseus* P5βR gene family members. CRO_025460, the actual *ISY* gene, corresponds to KJ873886.1 and is highlighted in red. CRO_025461 corresponds to KJ873885.1. In bold are the genes that have been shown to be able to *in vitro* convert 8-oxogeranial, the original *ISY* substrate [181].

	KJ873884.1	**KJ873885.1**	**KJ873883.1**	**KJ873886.1**	**KJ873882.1**	KJ873887.1
KJ873884.1		69.3	81.3	64	83.6	56.9
KJ873885.1	69.3		68	65.6	70.6	58.2
KJ873883.1	81.3	68		62.7	86.5	54.6
KJ873886.1	64	65.6	62.7		64.5	59.4
KJ873882.1	83.6	70.6	86.5	64.5		56.8
KJ873887.1	56.9	58.2	54.6	59.4	56.8	

The two co-localised and functionally identical genes iridoid synthase and CRO_025461 are additionally very tightly co-ex pressed (Figure 40). The physiological function and involvement of iridoid synthase in *C. roseus* alkaloid biosynthesis had been confirmed previously with targeted virus induced gene silencing (VIGS) [30]. However VIGS experiments silencing the gene corresponding to contig CRO_025461 cannot support an involvement of CRO_025461 in alkaloid biosynthesis [181].

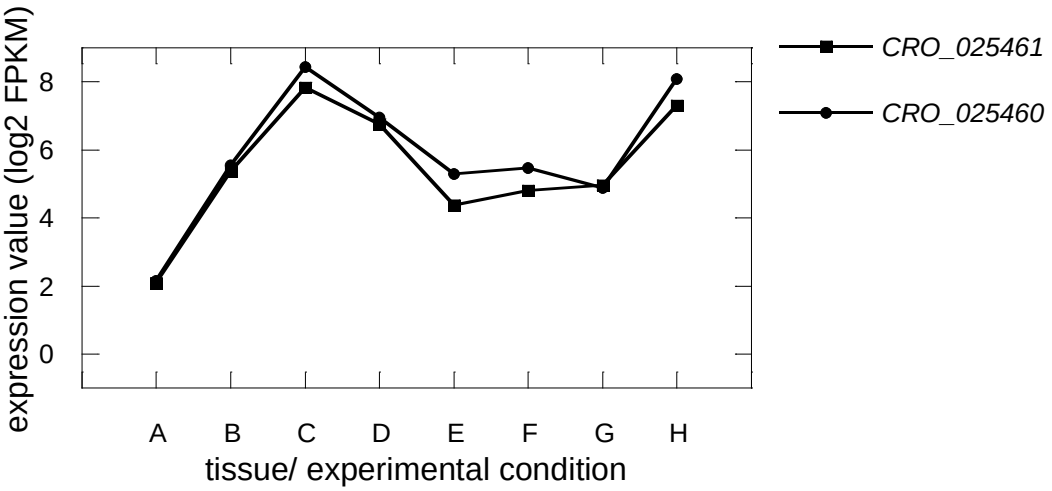


Figure 40 Gene expression of iridoid synthase and iridoid synthase paralog

Expression values (as log_2FPKM) of *ISY* (CRO_025460) and the *ISY* paralogue (CRO_02461). Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

Contig CRO_025459, a second gene localised in close proximity to the iridoid synthase gene, is annotated as cytochrome B5. The alkaloid biosynthetic pathway in *C. roseus* includes seven enzymatic steps that are catalysed by cytochrome P450 enzymes (G8H, IO, 7-DLH, SLS, T16H, T19H and T3O). The catalysis of these P450s relies on electron shuttling from cytochrome P450 reductase (CPR). It has been shown that this can be aided by additional action from cytochrome B5 [182]. However, the involvement of this enzyme in the alkaloid biosynthetic pathway has not been experimentally investigated.

It is not clear is any of the other genes located in the ISY_BAC covered region are used in the MIA pathway, though these genes are under consideration for future experimental analyses.

4.2.7 Iridoid oxidase, 7-deoxyloganetic acid glucosyltransferase, 7-deoxyloganic acid hydroxylase and loganic acid methyltransferase

A second cytochrome P450 in the seco-iridoid pathway in *C. roseus*, iridoid oxidase (*IO*) [31], oxidises cis-trans-nepetalactol to 7-deoxyloganetic acid. The 7-deoxyloganetic acid is subsequently glycosylated by 7-deoxyloganetic acid glucosyltransferase (*7DLGT*), [47]. The reaction product is hydroxylated by the third cytochrome P450 of the seco-iridoid pathway in

C. roseus, 7-deoxyloganic acid hydroxylase (*7DLH*) [33] and methylated by loganic acid methyltransferase (*LAMT*) to form loganin [34] (Figure 39).

The *IO* scaffold contains only one further transcript. The scaffold containing *7DLGT* contains five further transcripts. Two overlapping scaffolds each contain a part of *7DLH.* No further transcript is located on the first of the two scaffolds, while five further transcripts are located on the second scaffold. Finally the *LAMT* scaffold contains only one further transcript (Table 14). None of these transcripts is a known alkaloid pathway gene.

Table 14 Scaffold containing alkaloid biosynthesis gene *IO*, *7DLGT*, *7DLH* and *LAMT*
Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway gene in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
IO	3064716	28,101	CRO_006322	AGAMOUS-like
			CRO_006321	**iridoid oxidase (CYP76A26)**
7DLGT	3050165	46,244	CRO_027003	UDP-glucosyl transferase
			CRO_027004	UDP-glucosyl transferase
			CRO_027001	UDP-glucosyl transferase
			CRO_027005	UDP-glucosyl transferase
			CRO_027006	**UDP-glucose iridoid glucosyltransferase**
			CRO_027002	heat shock protein 18.2
7DLH	2952847	5,196	CRO_014616	**7-deoxyloganic acid 7-hydroxylase**
7DLH	3058188	40,675	CRO_005949	**7-deoxyloganic acid 7-hydroxylase**
			CRO_005951	glyoxalase II
			CRO_005950	Ribosomal protein L18ae family
			CRO_029081	hypothetical protein
			CRO_029082	hypothetical protein
			CRO_029080	Plasma-membrane choline transporter
LAMT	3052816	52,328	CRO_028498	hypothetical protein
			CRO_028497	**loganic acid methyltransferase**

Four of the transcripts located close to the pathway gene *7DLGT* are as well annotated as UDP-glucosyl transferases (UGT). These proteins belong to the large and diverse family of glucosyl transferases. Many (155) *C. roseus* transcripts are annotated as UGT, a value that corresponds to that of approximately 120 UGTs found in *Arabidopsis thaliana* [183] and 137 UGTs in flax [184]. There is no evidence that suggests that those UGTs of close physical proximity are involved in alkaloid biosynthesis in *C. roseus*.

4.2.8 Secologanin synthase and tetrahydroalstonine synthase

In the final step of the seco-iridoid section of MIA biosynthesis in *C. roseus* loganin is oxidatively cleaved by the fourth cytochrome P450 of this pathway, secologanin synthase [35] to yield the iridoid terpene secologanin.

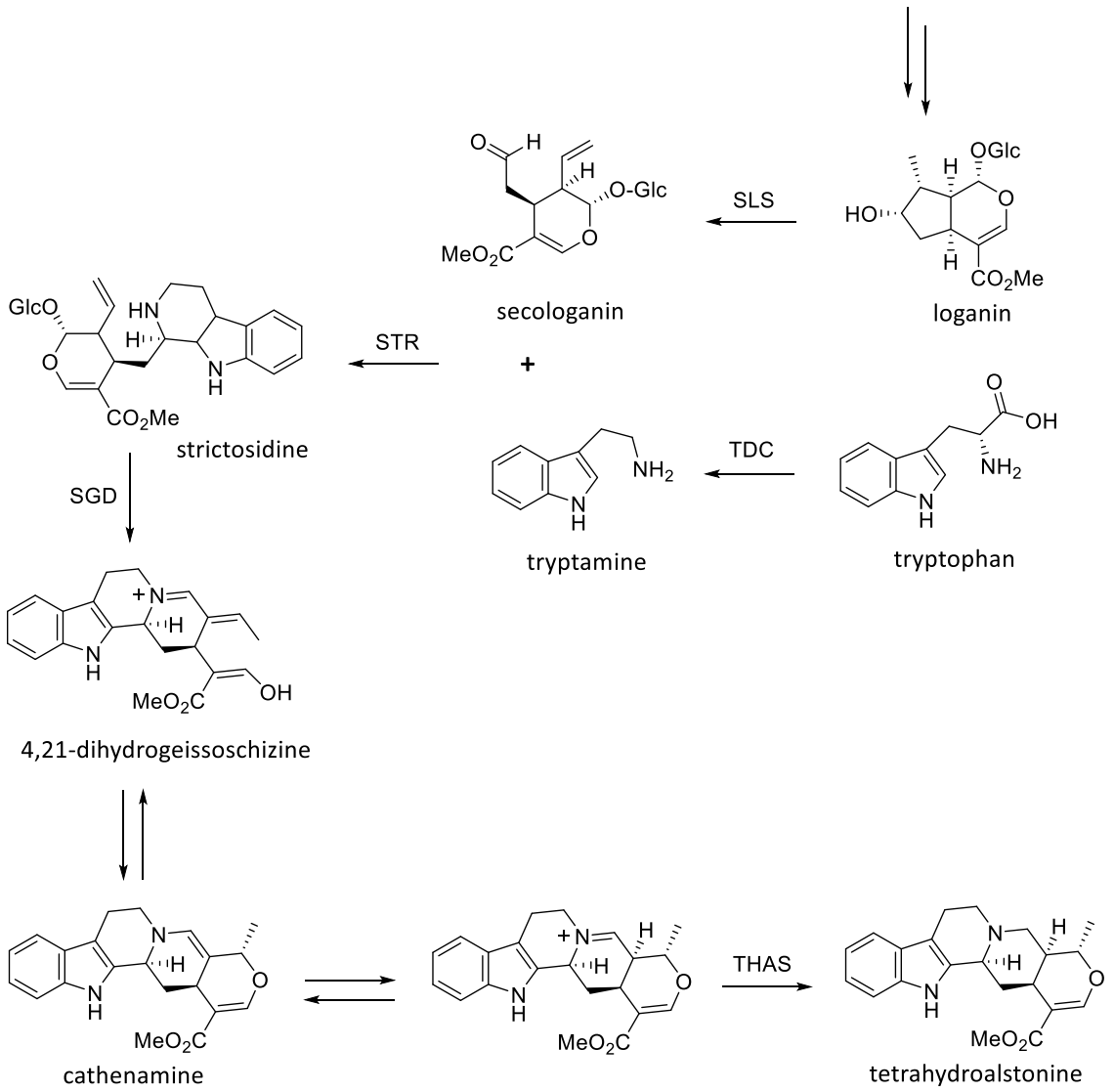


Figure 41 Monoterpene indole alkaloid pathway in *C. roseus (Part II)*

Section of the MIA pathway in *C. roseus*: SLS, secologanin synthase; TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGD, strictosidine β-glucosidase and THAS, tetrahydroalstonine synthase.

The initial analysis of the draft genome identified four paralogs of this gene. Each could be linked to an individual scaffold, placing each gene in a district genomic context (Table 15).

Table 15 Scaffolds containing alkaloid biosynthesis gene secologanin synthase

Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway gene in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
SLS_1	3045390	49,729	CRO_009299	hypothetical protein
			CRO_009298	hypothetical protein
			CRO_009297	**secologanin synthase-like protein**
SLS_2	3047714	65,404	CRO_032843	hypothetical protein
			CRO_032844	hypothetical protein
			CRO_032845	hypothetical protein
			CRO_032842	**secologanin synthase/ cytochrome (CYP72)**
SLS_3	3071001	48,935	CRO_012141	hypothetical protein
			CRO_012139	MuDR family transposase
			CRO_012142	hypothetical protein
			CRO_012140	**secologanin synthase-like protein**
SLS_4 /THAS	3063455	76,903	CRO_024555	hypothetical protein
			CRO_024559	hypothetical protein
			CRO_024558	hypothetical protein
			CRO_024554	hypothetical protein
			CRO_024552	reticuline oxidase like-protein (RO)
			CRO_024557	hypothetical protein
			CRO_024556	**secologanin synthase-like protein**
			CRO_024553	**tetrahydroalstonine synthase**

The exon intron structure of all four paralogs is identical but the size and sequence of the introns varies (Figure 42). To confirm this initial observation, primers were designed for the individual SLS versions for PCR from genomic DNA. Sequencing of the four PCR products confirmed the differences between the four SLS paralogs in the genic and the non-genic regions, strengthening the correctness of the genome assembly.

Figure 42 Exon intron structure of all four secologanin synthase paralog genes

Schematic representation of exon intron structure of *SLS* paralog genes. Black sections represent exons. Gene open reading frame size is 1584 bp (*SLS*_1 and 4) or 1575 bp (*SLS*_2 and 3).

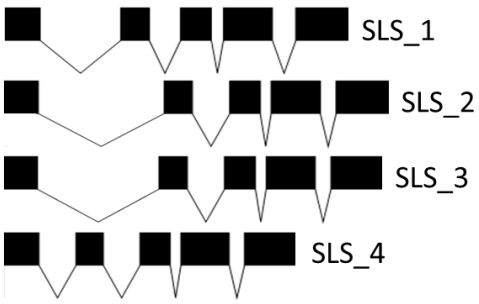


Table 16 Nucleotide sequence similarity of the four secologanin synthase paralog genes

Nucleotide alignment of all four SLS sequences cloned from *C. roseus* "SunStorm Apricot" display their sequence identity in %.

	SLS_1	SLS_2	SLS_3	SLS_4
SLS_1		94.2	94.2	93.9
SLS_2	94.2		97.7	97.2
SLS_3	94.2	97.7		96.6
SLS_4	93.9	97.2	96.6	

All four *SLS* paralogs were cloned successfully from cDNA. The *SLS* paralogs are between 94 and 98% identical in nucleotide sequence (Table 16). This high sequence similarity had made it impossible to assemble the four paralogs individually in the original transcriptomic data set [65] leading to various partial *SLS* transcripts but no full *SLS* transcript [49]. In the new improved transcriptomic dataset obtained using the genome assembly as reference [68] , the individual expression profiles of these genes can be resolved (Figure 43). An involvement of all four versions of *SLS* in *C. roseus* alkaloid biosynthesis is assumed based on their high sequence similarity and the fact that all four versions are expressed in the relevant tissues.

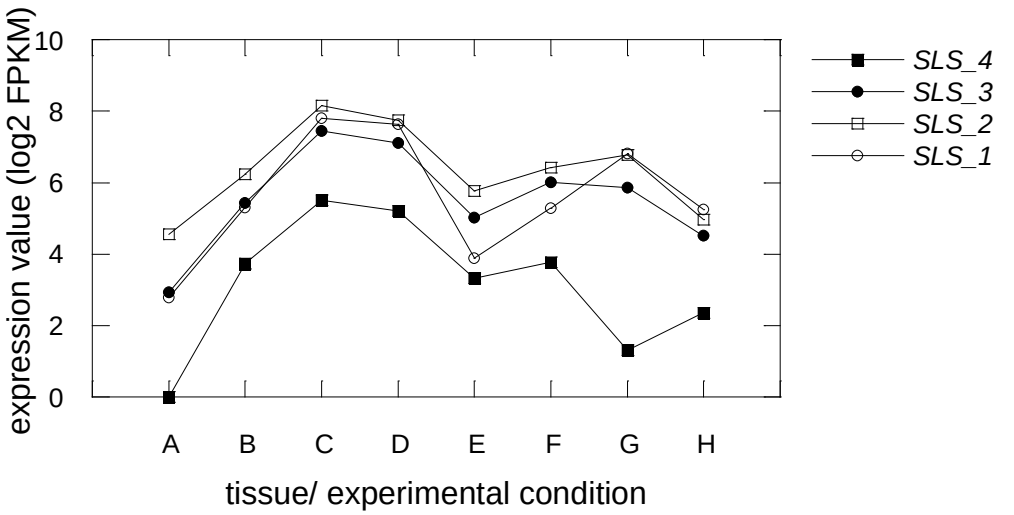


Figure 43 Gene expression of all four secologanin synthase paralog genes

Expression values (as log_2FPKM) of all four *SLS* paralogs. Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

4.2.8.1 Further content of SLS_4 scaffold – tetrahydroalstonine synthase (THAS)

After establishing the consistency of the assembly data, a closer inspection of all four *SLS* scaffold revealed that the SLS_4 scaffold contains two further transcripts of interest, the annotation of which suggests that they may have the potential to be involved in the unknown steps of alkaloid biosynthesis (Table 15). The contigs CRO_024553 (cra_locus_1974_iso_3 in the original transcriptomic dataset [65]) was annotated as "sinapyl alcohol dehydrogenase". Despite its relatively low expression in mature and young leaf it displays a strong upregulation after methyl jasmonate induction and was therefore suspected to be involved in alkaloid biosynthesis in *C. roseus*.

After observing the evidence of gene clustering for this candidate, it was subsequently tested for its potential involvement in alkaloid biosynthesis. The activity of this enzyme was tested by Anna Stavrinides and Dr. Evangelos Tatsis, who were working on elucidation of enzymatic steps following the deglycosylation of strictosidine. Those experiments included targeted gene silencing as well as expression of this candidate in *E. coli* and testing it with the strictosidine

aglycone, the *SGD* reaction product. It was shown that this enzyme can accept the aglycone and convert it to the heteroyohimbine-type alkaloid tetrahydroalstonine (Figure 41). It was consequently named tetrahydroalstonine synthase. It represents the first ever report of cloning and characterisation of a gene involved at the crucial branch point in monoterpene alkaloid diversity not only in *C. roseus* but in all monoterpene alkaloid producing plants. *THAS* has been further characterised and the results have been published [37].

The term collinearity in plant genomics describes the conserved order of genes that can be observed between different plant species [185]. In plant gene clustering, collinearity refers to the order of genes belonging to a cluster representing the exact order of reactions in the pathway. It is however rather the exception then the rule that plant gene cluster display a strong collinearity [174]. Therefore the fact that THAS is an enzyme not acting directly on the SLS product (Figure 41) is not unexpected.

4.2.8.2 Further content of SLS_4 scaffold – reticuline oxidase (RO)

The second contig of potential interest, CRO_024552, is annotated as a reticuline oxidase like-protein. Reticuline oxidase, or (S)-reticuline oxygen oxidoreductase, is a methylene-bridge-forming enzyme, also called berberine bridge enzyme, first characterised from *Eschscholzia californica* (california poppy) where it is essential to the formation of benzophenanthridine alkaloids that are involved in plant defence against pathogenic attacks [186]. However *RO* is not expressed in leaves and therefore silencing of this gene using VIGS in *C. roseus* is not possible (Figure 44). Furthermore, it is only minimally induced by methyl jasmonate, a fact that makes its involvement in alkaloid biosynthesis less likely, as all so far reported genes are strongly methyl jasmonate induced. The function of RO remains unknown.

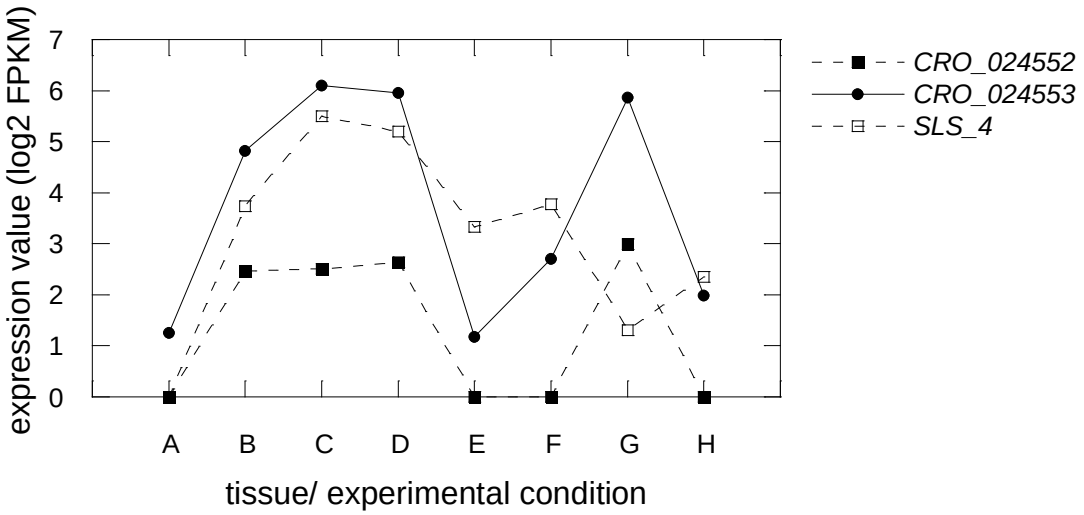


Figure 44 Gene expression of *SLS*, *THAS* and *RO* in *C. roseus*

Expression values (as log_2FPKM) of SLS_4 paralogue as well as *THAS* (CRO_024553) and *RO* (CRO_024552). Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

4.2.9 Tryptophan decarboxylase & strictosidine synthase

In parallel to the reaction of loganin to secologanin catalysed by secologanin synthase (SLS), the amino acid tryptophan is decarboxylated by the pyridoxal phosphate dependent enzyme, tryptophan decarboxylase (TDC) to form tryptamine [187] (Figure 41). Both products, secologanin and tryptamine are enzymatically joined in a Pictet-Spengler condensation reaction catalysed by strictosidine synthase (STR), the first committed step in the creation of monoterpene indole alkaloids by formation of the central metabolic intermediate strictosidine (Figure 41) [188,189].

4.2.9.1 Genome scaffold for tryptophan decarboxylase

The *TDC* genome scaffold contains two further transcripts, one representing the pathway gene *STR* and one further transcript annotated as an efflux family protein of the Multi-antimicrobial extrusion protein type (MATE) (Table 17).

Table 17 Scaffold containing alkaloid biosynthesis gene tryptophan decarboxylase

Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway gene in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
TDC / STR	3045674	29,490	CRO_006099	**strictosidine synthase**
			CRO_006098	**tryptophan decarboxylase**
			CRO_006097	MATE efflux family protein

To ensure that this observation of pathway gene clustering is not an artefact of the assembly, PCR using genomic DNA as template was conducted. The forward primer for this PCR was located in the coding region of the *TDC* gene while the reverse primer was located in the coding region of the *STR* gene. Both primers amplified only one single target as was determined by BLAST search against the genome assembly. The expected band size was 6849 bp. A band of approximately this size was amplified using PCR on *C. roseus* genomic DNA (Figure 45).

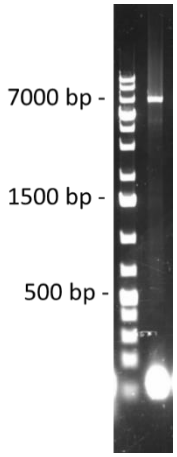


Figure 45 Products obtained by amplifying *TDC* and *STR* from genomic DNA

A band of approximately 7000 bp could be obtained using a *TDC* and *STR* specific primer pair on *C. roseus* genomic DNA, suggesting the close proximity of these genes in the *C. roseus* genome.

Mapping the full open reading frames (ORFs) of all three genes to the genomic scaffold revealed their exact intron exon structure, direction of the ORFs and the distance between the individual genes. The genes *TDC* and *STR* are located only approximately 6000 bp apart with no other transcript mapping to this section (Figure 46).

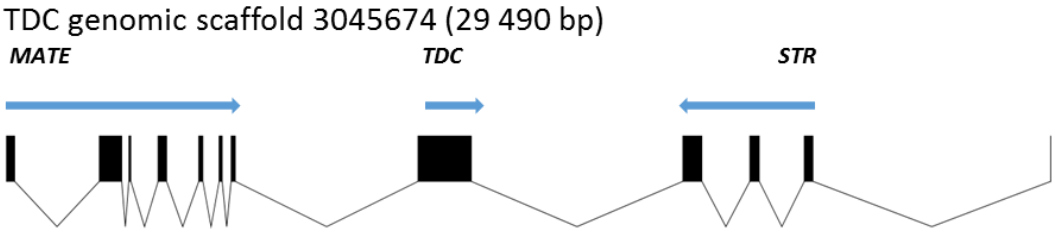


Figure 46 Representation of scaffold 3045674 with genes *MATE, TDC* and *STR*

Schematic representation of location of the three open reading frames for the gene *MATE*, *TDC* and *STR*. Black sections represent exons.

The TDC/ STR genomic scaffold contains a third transcript. CRO_006097 is annotated as a Mate efflux family protein (MATE), a transport protein [190]. Members of this family have been implicated in various transport processes in plants and interestingly also in the transport of secondary metabolites [191]. The expression profile of these three co-localised genes displays remarkably strong co-regulation (Figure 47) making MATE an excellent candidate for further characterisation. This characterisation was done by Dr. Richard Payne from the O'Connor lab (John Innes Centre, Norwich, UK), with his work strongly suggesting that MATE is responsible for transporting secologanin, the co-substrate of STR, from the cytosol into the vacuole.

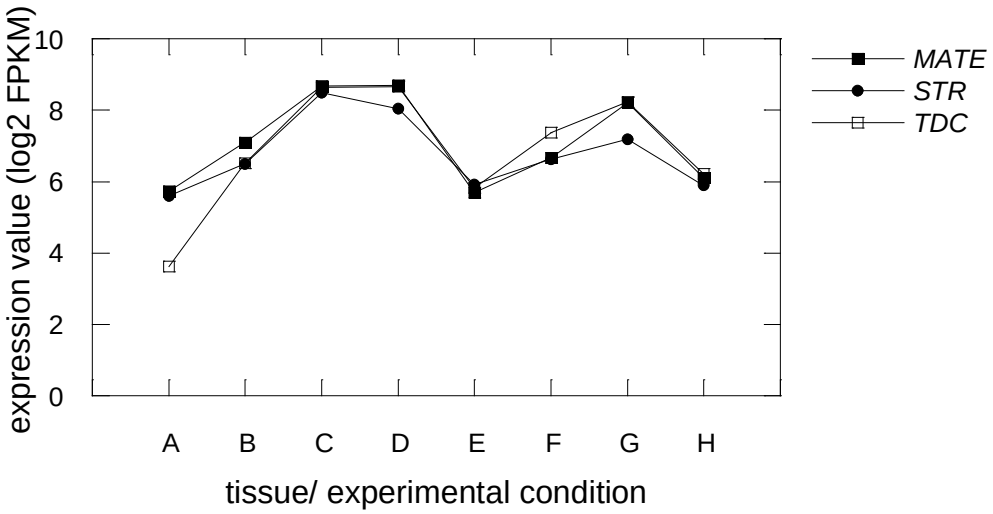


Figure 47 Expression profile of *TDC*, *STR* and *MATE*

Expression values (as log_2FPKM) of *TDC*, *STR* and *MATE*. Gene expression is displayed in the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

4.2.9.2 STR_BAC and MATE_BAC

Originally we instructed Bio S&T Inc. (Montreal, Canada) to extract a BAC that contained *MATE,* but not *STR,* to ensure that we could extend the BAC sequence data in the direction of *MATE*. This approach however failed and both BACs obtained still contained both *STR* and *MATE* (Table 18) The content of the STR_BAC and the MATE_BAC assemblies is almost identical. Only the overlap of a further genomic scaffold with one of the MATE_BAC scaffolds allows the extension of the sequence. MATE_BAC scaffold c3 overlaps with genomic scaffold 3064339, a scaffold that is 57 729 bp long and allows to connect seven further transcripts to the existing gene cluster (Table 18).

Table 18 STR_BAC and MATE_BAC assembly scaffolds

Size of STR_BAC and MATE_BAC scaffolds larger 2000 bp, corresponding genomic scaffolds with size in bp, transcripts contained and transcript functional annotation. Bold are the known pathway genes strictosidine synthase and tryptophan decarboxylase.

STR_BAC Scaffold/size in bp	**Corresponding genomic scaffold/size in bp**	**Transcripts contained**	**Functional annotation**
c1/15240	3045674/29490	CRO_006097	MATE efflux family protein
		CRO_006098	**tryptophan decarboxylase**
		CRO_006099	**strictosidine synthase**
c2/5461		already contained in c1	
c3/3698		already contained in c1	
MATE_BAC Scaffold/size in bp	**Corresponding genomic scaffold/size in bp**	**Transcripts contained**	**Functional annotation**
c4/16340	3045674/29490	CRO_006097	MATE efflux family protein
		CRO_006098	**tryptophan decarboxylase**
		CRO_006099	**strictosidine synthase**
c3/14187	3064339/57729	CRO_006603	cytochrome P450, family 71, subfamily B
		CRO_006604	F-box family protein
		CRO_006605	Galactosyltransferase family protein
		CRO_006606	alpha carbonic anhydrase
		CRO_006607	Ubiquitin-like superfamily protein
		CRO_006608	aconitase
		CRO_006609	Polynucleotidyl transferase
c2/6507		already contained in c4	
c1/6003		already contained in c4	

A surprisingly low number of transcripts that can be associated with both BACs is likely due to the large proportion of highly repetitive sequences. A dotplot graph [192] was used to investigate sequence similarity as similarity matrix. Horizontal and vertical edges representing the sequences to be compared with individual "dots" being coloured if they are identical in the two sequences. The structure and complexity of a sequence can be investigated by comparing it in a dotplot to itself. A dotplot then visualises regions of low complexity such as highly repetitive and low complexity sequences that are particularly challenging to assemble [193] as more or less fully coloured regions. The dotplot of the combined assembly of the MATE_BAC (85,000 bp, Table 8) displays a high proportion of highly repetitive sequences (Figure 48). However, the MATE_BAC contains a transcript annotated as a cytochrome P450 enzyme, transcript CRO_006603, which belongs to the CYP71 family. The expression values for this P450 is however very different from that of the pathway genes *STR* and *TDC*. This P450 is only expressed in seedling and stem tissue (Figure 49). Therefore, no VIGS experiment was conducted and the function of this gene remains unknown.

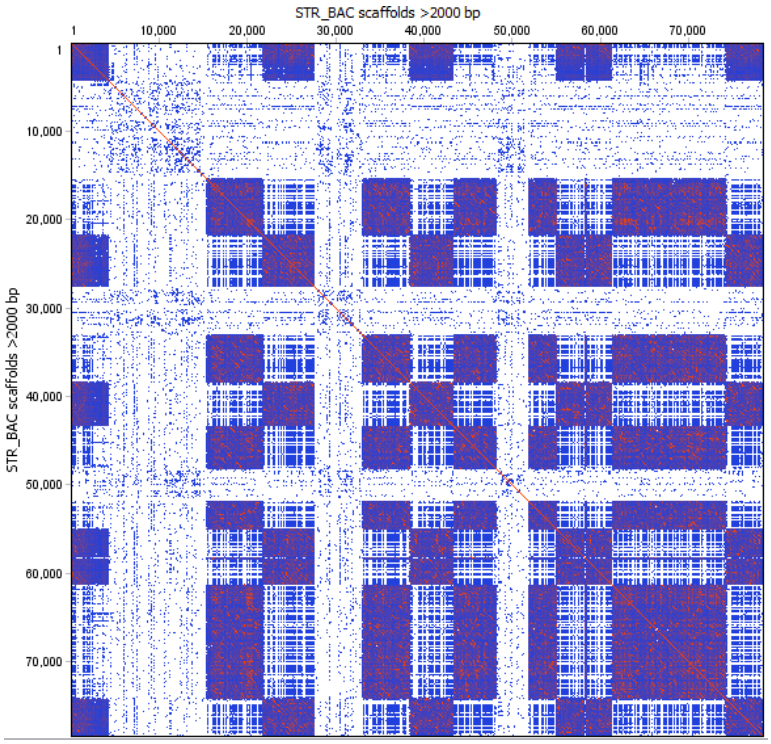


Figure 48 Dotplot (self) of all scaffolds > 2000 bp of the STR_BAC

The assembly of the STR_BAC results in a total number of 13 scaffolds with highly repetitive content.

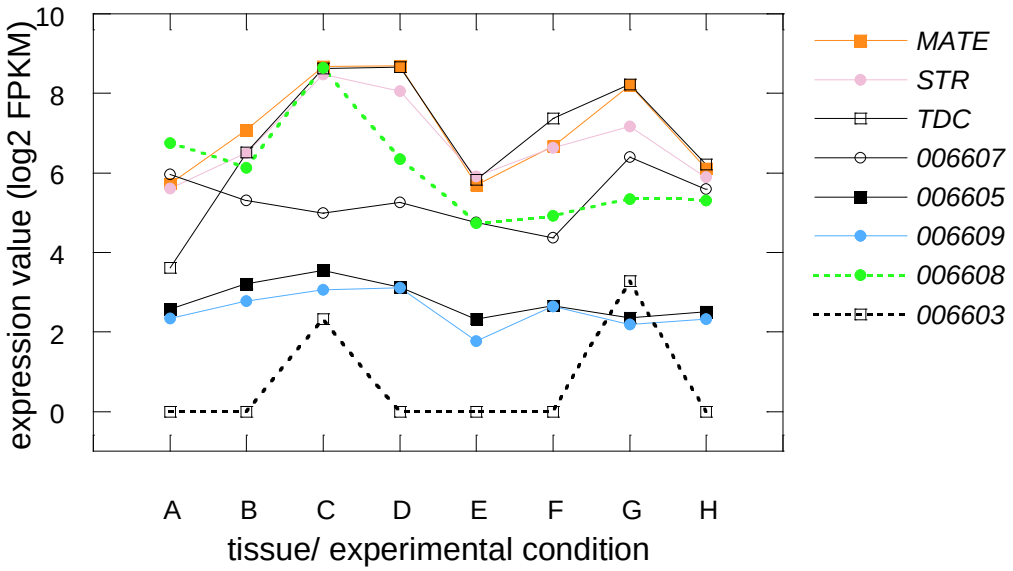


Figure 49 Expression of transcripts of MATE_BAC scaffold

Expression values (as log_2FPKM) of transcripts contained in scaffold 3064339 in comparison to *TDC*, *STR* and *MATE*. Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

Other genes in close proximity include a contig annotated as aconitase. Aconitases are essential iron-sulphur proteins facilitating the isomerisation of isocitrate from citrate indicating their function in primary metabolism pathways. Additionally aconitase genes from *Arabidopsis* were shown to be involved in transcriptional regulating resistance to oxidative stress [194]. Although aconitase (CRO_006608) is co-expressed with the alkaloid pathway genes *TDC* and *STR* (Figure 49) there is no other evidence suggesting an involvement of this gene in alkaloid biosynthesis in *C. roseus*. Given the known biochemistry of this alkaloid pathway, a role for aconitase activity does not seem likely.

4.2.10 Strictosidine β-glucosidase

The deglycosylation of strictosidine by strictosidine β-glucosidase (SGD) [195] results in formation of a reactive di-aldehyde species that subsequently rearranges to 4,21-dehydrogeissoschizine, an intermediate from which nearly all structural scaffolds of monoterpene indole alkaloids are derived [18] (Figure 41).

4.2.10.1 SGD scaffold

Multiple scaffolds can be associated with the *SGD* gene sequence, none of which represent the full gene, which contains 13 exons (Table 19).

Table 19 Scaffolds containing alkaloid biosynthesis gene strictosidine β-glucosidase

Scaffold number, size and contained transcripts ID and annotation. Neither any of the genomic nor the transcripts represent the full open reading frame of SGD.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
SGD	2954985	834	CRO_015796 CRO_028721 CRO_031126 CRO_005478	**strictosidine β-glucosidase**
	3022133	734		
	3042873	2880		
	3043349	2919		
	3042874	5337		
	3038592	3868		
	3038594	5212		
	3062364	4,646		

Three contigs resembled the N-terminus of the gene, four the C-terminus of the gene, and one contig the middle region. The gene was manually reconstructed from these 8 scaffolds and consists of 13 exons and 12 introns. As in the case of the *SLS* genes, individual parts of the *SGD* gene were cloned from genomic DNA and sequenced to confirm the manual reconstruction. *SGD* spans more than 10 kb on the genome (Figure 50). The exon intron organisation is not unusual and can be found in other plant genomes such as rice where a comparison of all members of the glucosyl hydrolase family 1 β-glucosidases revealed several types of gene structures with the most common gene pattern being the one in which there are 13 exons separated by 12 introns [196]. Manual examination by Robin Buell and co-workers at Michigan State University (MSU, East Lansing, USA) had suggested that the misassembled *SGD* represented a collapsed genome assembly possibly representing a minimum of two close *SGD* paralogs. Later comparison with the Mate pair data confirmed the existence of two partial *SGD* genes surrounding the actual full length SGD gene in the *C. roseus* genome.

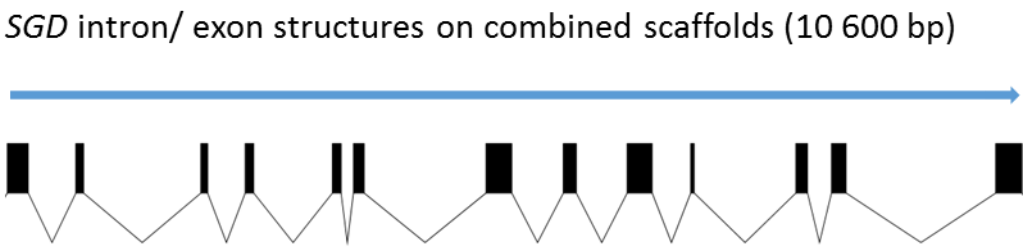


Figure 50 Reconstructed exon intron structure of the *SGD* gene

The alkaloid biosynthesis gene *SGD* is represented partially by eight different scaffolds. Combined the exon/ intron structure becomes visible and reveals a total length of the gene in its genomic context of approximately 10000 bp.

4.2.10.2 SGD_BAC

The assembly of the first SGD_BAC resulted in only two scaffolds with a combined length of 105000 bp, identical to the experimentally predicted insert size. Unfortunately the assembly of this SGD_BAC does not contain the actual *SGD* screening target sequence, in fact it contains almost no transcripts at all (Table 20). Attempts to map all obtained reads from the sequencing of the first SGD_BAC automatically onto the sequence of *SGD* did not result in any matches, suggesting that failure to detect *SGD* in the assembly was not due to a problem with the assembly.

Table 20 First SGD_BAC assembly scaffolds

Scaffolds resulting from the assembly of the SGD_BAC sequencing data, corresponding genomic scaffold and contained transcripts with functional annotation.

BAC Scaffold/size in bp	Corresponding genomic scaffold/size in bp	Transcripts contained	Functional annotation
c1/93247 bp	3057594/43594	CRO_022807	hypothetical protein
	2971237/8169	-	
	3062333/20761	CRO_013231	hypothetical protein
	3051495/16542	-	
	2959308/6578	-	
c2/12304	2981604/43644	CRO_000238	MuDR family transposase

As a consequence of the failure to locate *SGD* on the sequenced SGD_BAC, an improved primer pair was designed to screen the BAC library for a second time. The new knowledge of its exact genomic sequence helped identify an optimised primer pair for the second screening of the

BAC library. This time the *SGD* gene was part of the assembly and could, in full length, be localised on one scaffold of the SGD_BAC assembly (Table 21) confirming the previously manually reconstructed genomic structure (Figure 50). At this point, the genes surrounding *SGD* could be identified.

Table 21 Second SGD_BAC assembly scaffolds

Scaffolds larger 2000 bp, corresponding genomic scaffolds (in bp), transcripts contained and transcript functional annotation.

BAC Scaffold/size in bp	**Corresponding genomic Scaffold/size in bp**	**Transcripts contained**	**Functional annotation**
c1/19897	2954369/20065	-	
c4/14177	3038756/4822	-	
c2/10658	2954985/834, 3022133/734, 3042873/2880, 3043349/2919, 3042874/5337, 3038592/3868, 3038594/5212, 3062364/4646	CRO_015796, CRO_028721, CRO_031126, CRO_005478	*SGD*/ strictosidine beta glucosidase
c5/6829	3039922/3323	-	
	3110662/1692	-	
c7/6246	3067087/8810	-	
	3039593/6049	-	
c9/6202	3047206/8855	-	
c8/5969	contained in c7		
c3/5287	3041229/2400	-	
	3041230/2400	-	
c25/4071	3066832/2982	-	
	3066655/35788	CRO_011777	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase
		CRO_011778	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase
		CRO_011779	cytochrome P450, family 71, subfamily B, polypeptide
		CRO_011781	hypothetical protein
		CRO_011780	Peptidase C78, ubiquitin fold modifier-specific
c6/3659	3038311/3332	-	
	3038311/3301	-	
	295449/1670	-	
c29/3566	contained in c25		
c11/3204	3059350/4493	CRO_026512	oligopeptide transporter
c10/2930	3039593/6049	-	
c27/2271	contained in c11		
c14/2217	contained in c11		
c12/2063	contained in c25		

The second SGD_BAC contains a transcript annotated as oligopeptide transporter. These transport family enzymes are implicated in various biological functions with demonstrated substrates in plants being metal-chelates, small peptides and glutathione [197]. Transport of secondary metabolites is essential for complex pathways like the alkaloid pathway in *C. roseus* [198] of which only two transporter have been characterised so far [46,199]. However the contig CRO_026512 is only 600 bp long which is in stark contrast to the length described for the oligopeptide transporters from other plants such as *A. thaliana* [200] suggesting this is a partial gene/ pseudogene. Additionally CRO_026512 is not expressed in leaf tissues making it an unsuitable candidate for gene silencing (Figure 51).

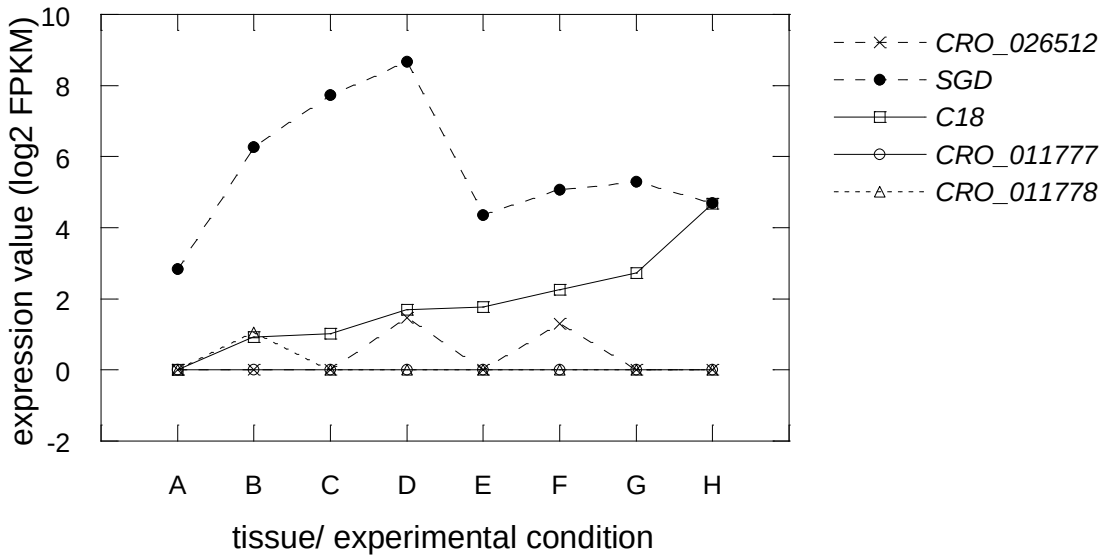


Figure 51 Expression profile of *SGD* and co-localised genes

Expression values (as log_2FPKM) of *SGD* and co-localised transcripts *CRO*_026512 (partial transporter gene), CRO_011777 and CRO_011778 (2-oxoglutarate (2OG) and Fe (II)-dependent oxygenase) and CRO_011779 (P450 gene, VIGS candidate C18). Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

The low expression values are also the reason two further transcripts on the scaffold, CRO_011777 and CRO_011778, have not been subjected to further silencing experiments despite the fact that 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenases can be implicated in many examples of plant specialised metabolism [201]. CRO_011779 is annotated as P450 enzyme of the CYP71 family. Although again it displays low expression values, it was subjected to two VIGS experiments. The results of these experiments are detailed in Chapter 5.

4.2.11 Tabersonine 16-hydroxylase & 16-hydroxytabersonine O*-methyl-transferase*

The first committed step of the seven enzymatic reactions comprising vindoline biosynthesis from the intermediate tabersonine in *C. roseus* is catalysed by the cytochrome P450 tabersonine 16-hydroxylase (T16H). This enzyme hydroxylates tabersonine at the 16 position to 16-hydroxytabersonine [41] (Figure 52). It has been established as part of this thesis (Chapter 2) that two *T16H* genes, *T16H1* and *T16H2*, exist in *C. roseus* and that both are involved in vindoline biosynthesis in a tissue dependant manner [42]. In the next enzymatic step the newly installed hydroxyl group of 16-hydroxytabersonine is methylated by 16-hydroxytabersonine *O*-methyltransferase (16OMT) [34].

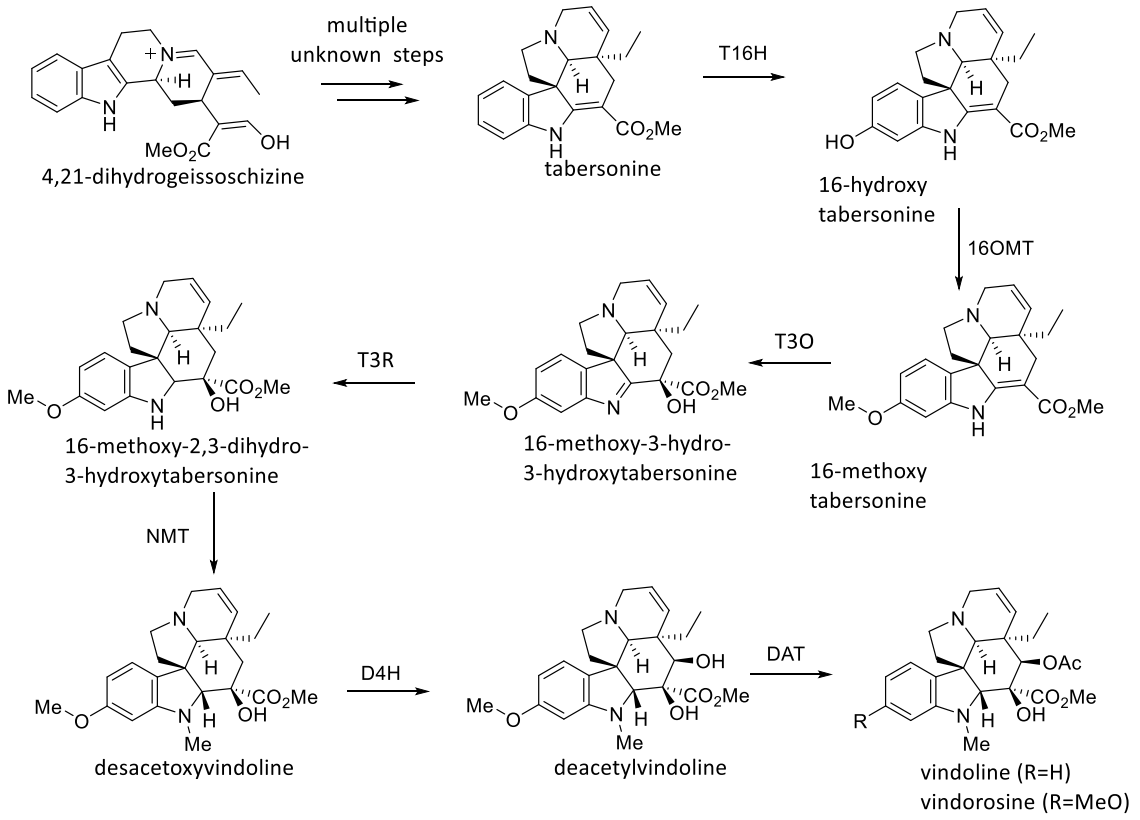


Figure 52 Monoterpene indole alkaloid pathway in *C. roseus (Part III)*

Downstream section of the MIA pathway in *C. roseus:* T16H2, tabersonine 16-hydroxylase 2 (CYP71D351); 16OMT, 16-hydroxytabersonine O-methyltransferase; T3O 16-methoxytabersone 3-oxigenase; T3R, 16-methoxytabersonine 3-reductase NMT, 16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase; D4H, desacetoxyvindoline-4-hydroxylase; DAT, deacetylvindoline 4-O-acetyltransferase.

4.2.11.1 T16H and 16OMT scaffold

The genomic scaffold for *T16H* contains both transcripts *T16H1* and *T16H2* along with two additional transcripts. The genomic scaffold for 16OMT contains one additional transcript (Table 22).

Table 22 Scaffolds containing alkaloid biosynthesis genes *T16H* and *16OMT*

Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway genes in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
T16H	3064268	32,093	CRO_017450	DUF4283 domain containing protein
			CRO_017447	**T16H-like protein**
			CRO_017449	cytochrome P450, family 71 B
			CRO_017448	**tabersonine 16-hydroxylase CYP71D12**
16OMT	3051716	19,234	CRO_004355	cytochrome P450, family 71 B
			CRO_004356	**16-hydroxytabersonine O-methyltransferase**

4.2.11.2 T16H2_BAC scaffolds

As already discovered from the T16H scaffold, the alkaloid pathway genes *T16H1* and *T16H2* represent a reversed tandem duplicated gene pair of a member of the CYP71 protein family. Both genes have been shown to carry out the same enzymatic reaction [42]. The alkaloid pathway gene *16OMT* acts immediately after the T16H reaction and is part of the T16H2_BAC assembly (Table 23).

Table 23 T16H2_BAC assembly scaffolds

Size of T16H2_BAC scaffolds larger 1000 bp, corresponding genomic scaffolds with size in bp, transcripts contained and transcript functional annotation. In bold are the known pathway genes *16OMT* (CRO_004356), *T16H1* (CRO_ 017448) and *T16H2* (CRO_017447). The transcript CRO_017449 is not contained on the BAC scaffold c3 but only on the genomic scaffold that overlaps partly with c3.

BAC Scaffold/size in bp	**Corresponding genomic scaffold/size in bp**	**Transcripts contained**	**Functional annotation**
c1/59670	2997859/42721	CRO_025929	cell wall protein precursor, putative
		CRO_025930	Protein phosphatase 2C family protein
		CRO_025927	Putative uncharacterised protein ycf15
		CRO_025928	photosystem II family protein
		CRO_025931	cytochrome P450, family 71, subfamily B
	3051716/19234	CRO_004356	**16-hydroxytabersonine O-methyltransferase**
		CRO_004355	cytochrome P450, family 71, subfamily B
c3/24101	3064268/32093	CRO_017449	cytochrome P450, family 71, subfamily B
		CRO_017450	DUF4283 domain containing protein
		CRO_017447	**T16H-like protein**
		CRO_017448	**tabersonine 16-hydroxylase CYP71D12**
c4/16202	3043234/7991	CRO_002269	Peroxisomal membrane 22 kDa family protein
		CRO_002270	signal peptide peptidase
	2999338/4703	CRO_017854	FAR1-related sequence
c2/13778	3030070/17788	CRO_015028	Auxin-responsive GH3 family protein
		CRO_015029	Pentatricopeptide repeat (PPR-like) superfamily
c7/1969	3041661/3032	no transcripts contained	
c5/1235	3051716/19234	already contained in c1	

The assembly results reflect the previously experimentally predicted insert size of this BAC. The largest T16H2_BAC scaffold c1 contains the known pathway gene *16OMT*, while the second largest T16H2_BAC scaffold c3 contains the known pathway genes *T16H1* and *T16H2* (Figure 53). It was not possible to conclusively order the scaffolds of this BAC manually but given the fact that *16OMT* is located on a 56 000 bp long scaffold and the given total length of the T16H2_BAC of 125 000 bp, we conclude that *16OMT* and *T16H1* and *T16H2* cannot be more then approximately 60 000 bp distant.

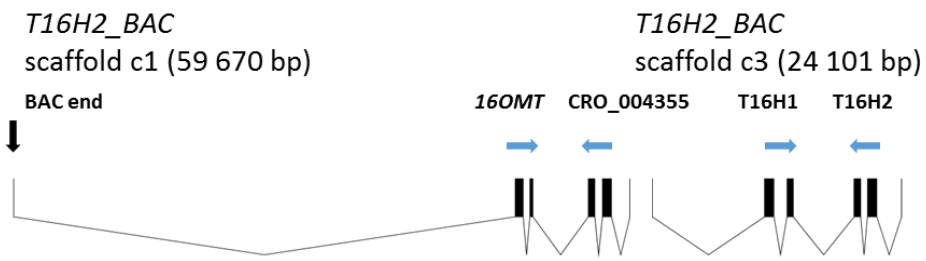


Figure 53 T16H2_BAC scaffolds c1 and c3

The T16H2_BAC scaffold c1 with the transcripts CRO_004356 (pathway gene *16OMT*) and CRO_004355, a second CYP71 gene, and the T16H2_BAC scaffold c3 with the reversed tandem duplicated transcripts CRO_017447 (pathway gene *T16H2*) and CRO_017448 (pathway gene *T16H1*). Blue arrow indicate direction of gene ORF. The BAC scaffolds c1 contains one of the two BAC ends indicated by a black arrow.

4.2.11.3 CYP71 from T16H2_BAC: CRO_025931, CRO_017449 and CRO_004355

The gene content of the T16H2_BAC assembly was evaluated for possible candidates for unknown steps in alkaloid biosynthesis. Three transcripts are annotated as CYP71, the same cytochrome P450 family as *T16H*. Closer inspection of the CYP71 annotated transcripts, CRO_025931 and CRO_017449, revealed that they are only partial genes with CRO_025931 containing only 612 bp and CRO_017449 containing only 441 bp. CRO_025931 represents most likely a partial duplication of *T16H2* as they still share 93.3% identity and was therefore not included in any further investigations. For CRO_017449, the picture is less clear as the 441 bp can be found in at least four other transcripts of various length, all of which have very low expression values in the alkaloid producing relevant tissues such as flower, leaf, seedling, stem and root. This contig was therefore also not included in further investigations. The third CYP71, a full length P450 gene, found on the T16H2_BAC, is CRO_004355 (Figure 53). This gene however is negligibly expressed in the *C. roseus* tissues leaf, stem, flower, seedling or root, making it a very unlikely candidate for alkaloid biosynthesis.

According to functional annotation, the CYP71 family is the largest family of cytochrome P450s in *C. roseus* accounting for more than 70 of the 255 as P450 enzyme annotated transcripts. This is in accordance with the numbers reported for other land plant genomes where CYP71 also represents the largest P450 family [202]. However, how many of the transcripts annotated as CYP71 in *C. roseus* are actually expressed, full length, functional genes and not partial duplications, pseudogenes or assembly fragments remains to be investigated. Additionally

CRO_017449 is not part of the actual T16H2_BAC but can only be found on the scaffold that exceeds the T16H2_BAC scaffold c3 (Table 23). CRO_017449 was used as template to design a VIGS construct and the construct was used to infiltrate young *C. roseus* plants in a VIGS experiment. However, this did not lead to any significant perturbations in metabolism (Chapter 5, Table 28).

4.2.12 Tabersonine 3-oxygenase, tabersonine 3-reductase, N-methyltransferase and desacetoxyvindoline 4-hydroxylase

The 16-methoxytabersonine product of 16OMT is further hydroxylated and reduced by the concerted action of tabersonine 3-oxygenase (T3O) [146] and tabersonine 3-reductase (T3R)[203]. At the start of this thesis both genes were unknown. The reaction product of T3O and T3R, 16-methoxy 2,3-dihydro-3-hydroxytabersonine, is the substrate for N-methylation by the N-methyltransferase (NMT) [127] and further hydroxylation of the resulting desacetoxyvindoline by desacetoxyvindoline 4-hydroxylase (D4H) (Figure 52) [133].

The *T3O* gene is, like *SGD* not represented by a single contig in the new transcriptome [68] but by five different partial transcripts. On genome level none of the scaffolds contains a full *T3O* gene and none of the scaffolds that contain part of *T3O* contain any other transcripts.

The *T3R* gene is located on a single scaffold containing one further transcript. *NMT* is located on a single scaffold containing one further transcript. *D4H* is located on a single scaffold containing no further transcripts (Table 24). No other known or potential alkaloid biosynthesis gene is located on any of the scaffolds.

Table 24 Scaffold containing alkaloid biosynthesis genes *T3R, NMT* and *D4H*
Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway gene in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
T3R			CRO_021542	hypothetical protein
			CRO_021541	**tabersonine 3-reductase**
NMT	3065019	10,334	CRO_033267	hypothetical protein
			CRO_033266	**16-hydroxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase**
D4H	2969470	16,035	CRO_012504	**desacetoxyvindoline-4-hydroxylase**

4.2.13 Minovincinine 19-hydroxy O-acetyltransferase, deacetylvindoline acetyl-transferase and tabersonine 19-hydroxylase

Tabersonine is a central intermediate in vindoline biosynthesis, which is derived in a seven step pathway in aerial tissues [48]. In roots a different set of genes is responsible for the formation of several additional tabersonine derived alkaloids (Figure 54).

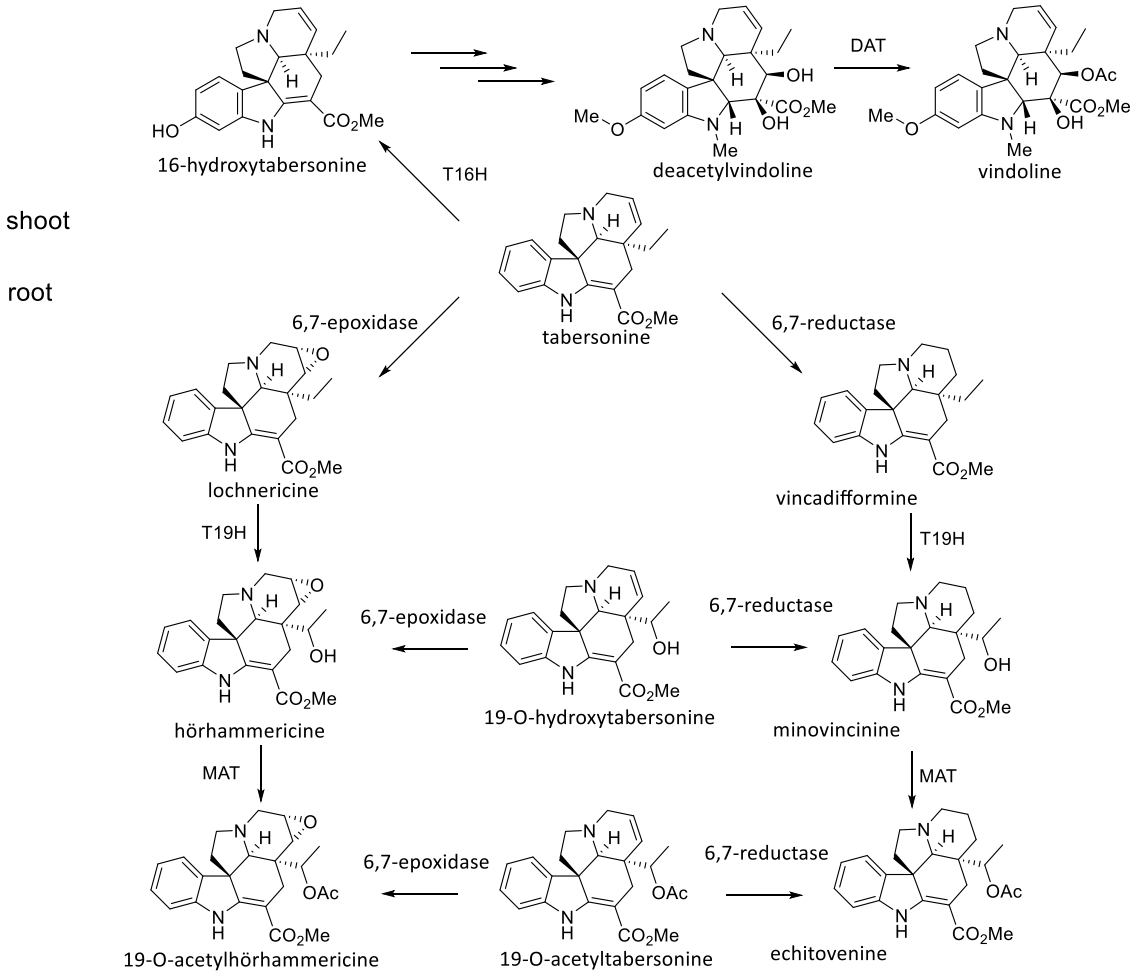


Figure 54 Monoterpene indole alkaloid pathway in *C. roseus (Part IV)*

Section of the MIA pathway in C. roseus DAT, deacetylvindoline 4-O-acetyltransferase; T19H, tabersonine/lochnericine 19-hydroxylase (CYP71BJ1) and MAT, minovincinine 19-hydroxy-O-acetyltransferase

In hairy root cultures of *C. roseus*, the alkaloids hörhammericine and lochnericine have been detected and as in vindoline biosynthesis, their production can be strongly increased by addition of elicitors such as jasmonic acid (JA) [204]. Acting on tabersonine directly is tabersonine 19-hydroxylase (T19H), a cytochrome P450 enzyme that hydroxylates tabersonine in the 19 position yielding 19-hydroxytabersonine. Additionally, it was shown that T19H can convert

lochnericine into hörhammericine [79]. While in the last step of vindoline biosynthesis deacetylvindoline is acetylated by deacetylvindoline acetyltransferase (DAT) [45] in the root specific alkaloid biosynthesis a highly similar gene encodes the enzyme minovincinine 19-hydroxy-O-acetyltransferase (MAT) which conducts the reaction from minovincinine to echitovenine, similar to the one of DAT, in both cases transferring a acetyl group to a tabersonine derived scaffold. It was however shown that MAT was unable to accept either tabersonine directly or any other tabersonine-like substrates that were not hydroxylated at the 19 position [205]. Deacetylvindoline, the natural substrate of DAT and hörhammericine was converted by MAT albeit at a much lower turnover rates. In contrast DAT is specific to deacetylvindoline and accepts no other substrate [205].

The scaffold for *DAT* contains two further O-acetyltransferase genes and the *MAT* scaffold contains three further O-acetyltransferase genes. *T19H* is located on a single scaffold that contains a second gene annotated as cytochrome P450 (Table 25).

Table 25 Scaffold containing alkaloid biosynthesis genes *MAT, DAT* and *T19H*
Scaffold number, size and contained transcripts ID and annotation. Alkaloid pathway gene in bold.

Gene	Scaffold	Scaffold size (bp)	Transcript ID	Transcript annotation
DAT	3060125	26,177	CRO_020282	HXXXD-type acyl-transferase family protein
			CRO_020283	hypothetical protein
			CRO_020281	HXXXD-type acyl-transferase family protein
			CRO_020280	**deacetylvindoline 4-O-acetyltransferase**
			CRO_028722	RNA helicase family protein
MAT	3067490	69,370	CRO_005217	beta-1,3-glucanase
			CRO_005218	Kunitz family trypsin and protease inhibitor
			CRO_005212	Kunitz family trypsin and protease inhibitor
			CRO_005211	HXXXD-type acyl-transferase family protein
			CRO_005216	Kunitz family trypsin and protease inhibitor
			CRO_005214	Kunitz family trypsin and protease inhibitor
			CRO_005210	HXXXD-type acyl-transferase family protein
			CRO_005215	HXXXD-type acyl-transferase family protein
			CRO_005213	**minovincinine 19-hydroxy-O-acetyltransferase**
T19H	**2964965**	**34,685**	CRO_021078	DEAD box RNA helicase (PRH75)
			CRO_021079	Peptidase family M48 family protein
			CRO_021080	Tetratricopeptide repeat (TPR
			CRO_021081	cytochrome P450, family 704 A
			CRO_021082	**tabersonine/lochnericine 19-hydroxylase**
			CRO_021083	hypothetical protein

4.2.13.1 Content of DAT and MAT scaffolds

The *C. roseus* transcriptome contains 84 acetyl coenzyme A-dependent O-acetyltransferases (HXXXD acyl-transferase family proteins). Minovincinine 19-hydroxy-O-acetyltransferase [206] and deacetylvindoline 4-O-acetyltransferase [45] share sequence identity on a nucleic acid level and no other *C. roseus* gene has a higher similarity to either gene, suggesting the evolution from a common ancestor. The only HXXXD acyl-transferase family protein on either of these scaffolds that is expressed in leaf is the *DAT* gene (Figure 55) (Figure 56). The function of the other acyl-transferases has not yet been elucidated.

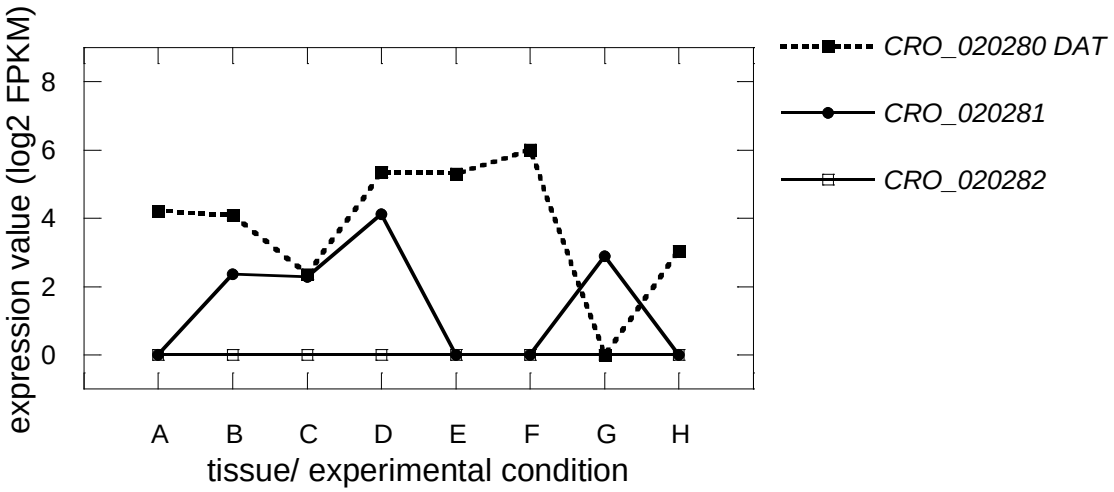


Figure 55 Expression of O-acetyltransferase genes located on the *DAT* scaffold

Expression values (as log_2FPKM) of *DAT* and co-localised transcripts annotated as HXXXD acyl-transferase family proteins. Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

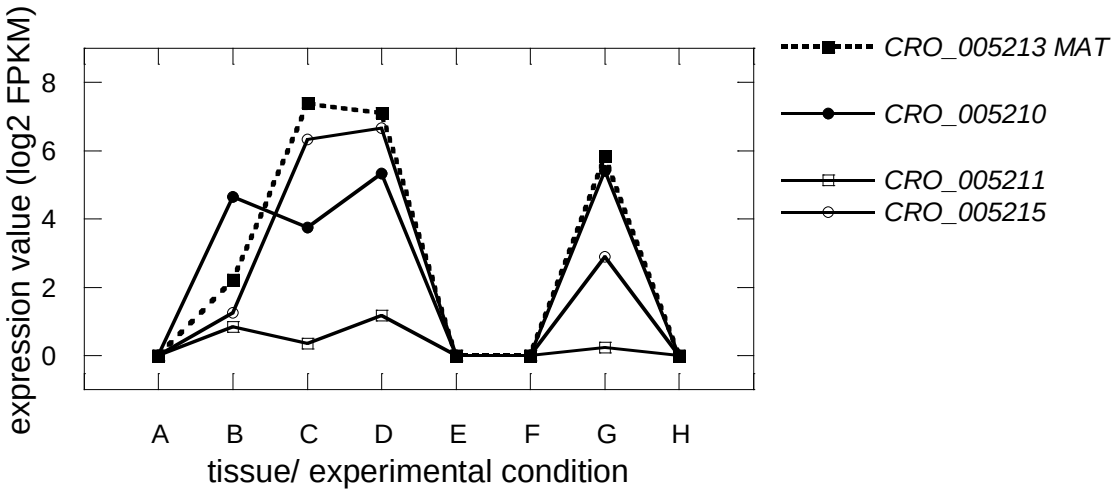


Figure 56 Expression of O-acetyltransferase genes located on the *MAT* scaffold

Expression values (as log_2FPKM) of *MAT* and co-localised transcripts annotated as HXXXD acyl-transferase family proteins. Gene expression for displayed in the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

4.2.13.2 Content of T19H scaffold

The biosythesis of root specific tabersonine derived alkaloids is missing one enzymatic step (Figure 54) the epoxidation of tabersonine to yield lochnericine. Biochemical characterisation using labeled ^{14}C-tabersonine had suggested a cytochrome P450 to be involved in this reaction [205] making CRO_021081, a transcript located in close proximity to *T19H* a compelling candidate for this reaction despite its overall expression being relatively low Figure 57.

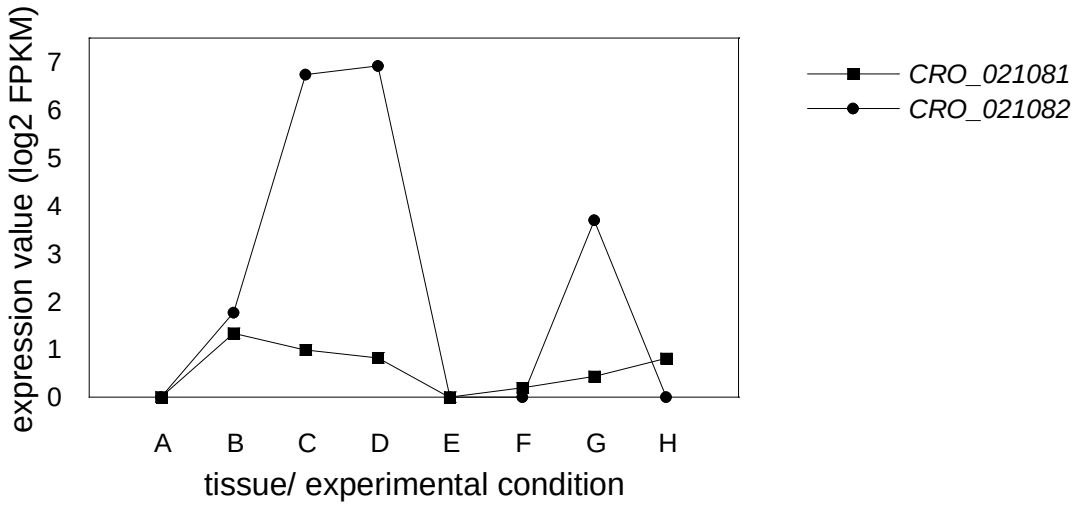


Figure 57 Expression of T19H and neighbouring P450 transcript

Expression values (as log_2FPKM) of CRO_021082/T19H and co-localised P450 transcript CRO_021081. Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

4.2.13.3 Expression of candidate for epoxidation in root specific alkaloid biosythesis

Testing of the physiological function of CRO_021081 *in planta* using silencing was omitted since this candidate was predicted to be involved in a pathway localised to roots which cannot be targeted by VIGS. Instead, the activity of this gene was assessed in vitro. Primers were designed and the full length gene encoded by CRO_021081 was cloned from *C. roseus* "Sunstorm Apricot" stem and root cDNA into the cloning vector pJET and subsequently into the yeast expression vector pXP218. Sequencing confirmed a 1527 bp ORF for this gene, referred to as candidate gene *C72*. The nucleotide sequence of *C72* was aligned to the *T19H* nucleotide sequence. Interestingly both P450 genes share only 52.2% sequence identity on nucleotide level and therefore do not represent a recent gene duplication event.

Competent cells of yeast strain WAT11, which harbours a P450 reductase [143], were transformed with pXP218 containing the candidate gene *C72*. WAT11 transformed to contain an empty plasmid served as empty vector control. Liquid cultures in appropriate selective media were induced for protein expression and supplemented with 5.5 µM tabersonine. The cultures were maintained for 48 hours and subsequently methanol extracts of culture samples measured by LC-MS (Xevo, Waters). No conversion of tabersonine in comparison to EV plasmid could be detected (Figure 58)

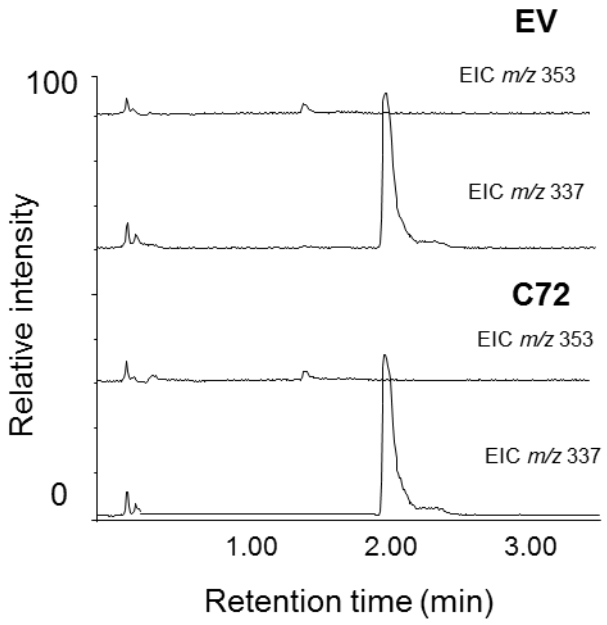


Figure 58 Yeast expressing *C72* culture supplemented with tabersonine

Extracted Ion chromatogram of LC-MS measurement (Xevo, Waters) for the masses *m/z* 337, corresponding to tabersonine and *m/z* 353, corresponding to the product of the oxidation/ epoxidation of tabersonine for yeast culture supplemented with tabersonine, with empty vector plasmid (top) or yeast culture supplemented with tabersonine, with plasmid containing *C72* (bottom).

It is possible that the missing epoxidation step is unable to accept tabersonine directly as a substrate. Instead the enzyme could require 19-hydroxytabersonine, the reaction product of the tabersonine hydroxylation by T19H [79]. To investigate this possibility *T19H* was cloned from *C. roseus* stem and root cDNA and cloned into pXP218 plasmid. Competent WAT11 was transformed with the resulting plasmid. Mixed cultures of *C72* plus *T19H* were tested. The cultures were maintained for 48 hours in the presence of tabersonine and subsequently measured by LC-MS. Previous intensive work with tabersonine derivatives in yeast cultures (Chapter 3) had shown that the products of the yeast cultures can be freely exchanged between yeast and media.

The yeast culture expressing T19H shows a new peak of a mass corresponding to that of 19-hydroxylated tabersonine (m/z 353). No such peak is detectable in the empty vector control culture (Figure 59). However the combination of C72 and T19H does not lead to the production of a further new peak suggesting no conversion of 19-hydroxytabersonine.

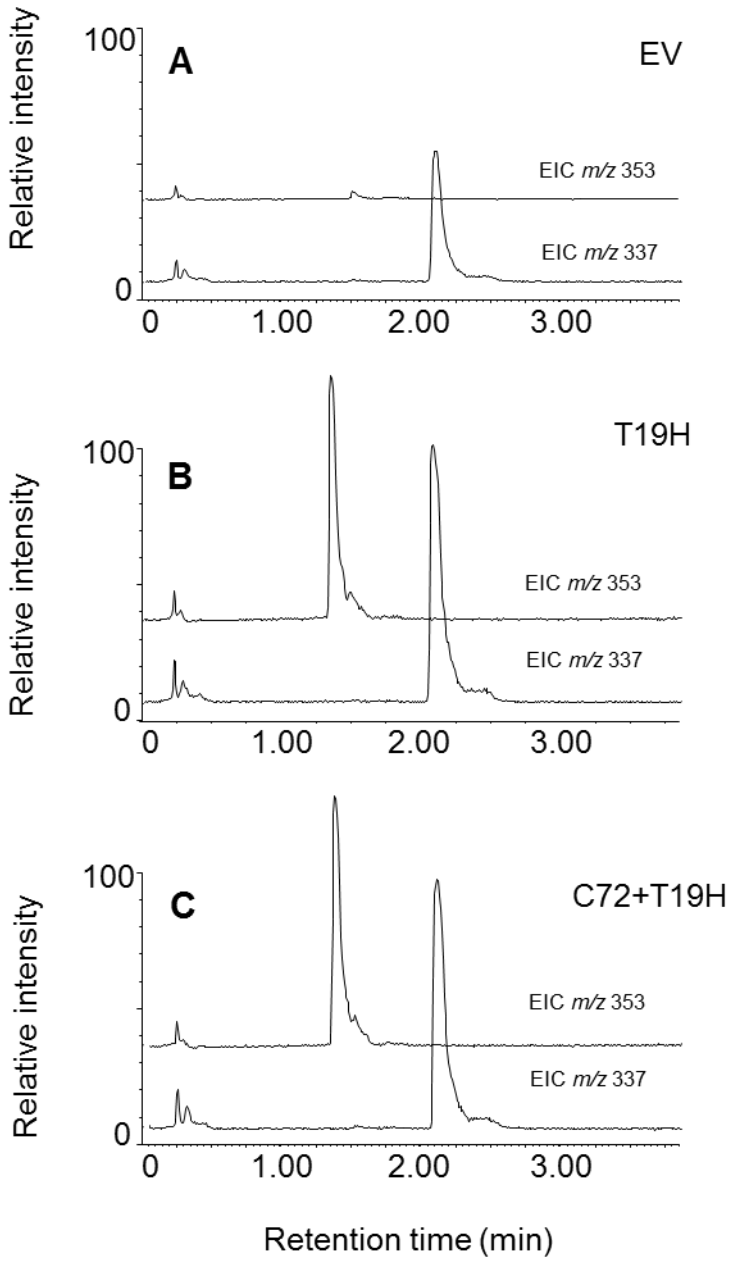


Figure 59 Extracted Ion chromatogram *C72/ T19H* cultures supplemented with tabersonine

Extracted Ion chromatogram of LC-MS measurement (Xevo, Waters) for the masses *m/z* 337, corresponding to tabersonine and *m/z* 353, corresponding to the product of the oxidation/ epoxidation of tabersonine. **A**: Yeast culture supplemented with tabersonine, with empty vector plasmid. **B**: Yeast culture supplemented with tabersonine, with plasmid containing *T19H*. **C**: Yeast culture supplemented with tabersonine, with plasmid containing *T19H* and *C72*.

In conclusion candidate *C72* was tested with two out of three possible substrates and did not show any activity. The most likely explanation is that *C72* is not a gene involved in this alkaloid pathway.

4.2.14 Clustering of other known secondary metabolite genes

Clustering of pathway genes has been observed for a range of secondary metabolites. Clustering has been reported for monoterpene [113], diterpene [108,110] and triterpene [117,120,175–177] biosynthesis genes. *C. roseus* contains the triterpenoid ursolic acid [207], which is synthesised from amyrin. Both amyrin synthase (AS) and amyrin oxidase (AO) from *C. roseus* have been characterised [63,64]. However the two genes of *C. roseus* implemented in the biosynthetic pathway of the triterpenoid ursolic acid, *AO* and *AS*, are not located on the same scaffold (Table 26).

Table 26 Scaffolds containing amyrin synthase and amyrin oxidase from *C. roseus*
Scaffold name and size in bp as well as all contained transcripts and transcript annotation. Transcripts corresponding to *AS* and *AO* are in bold.

Gene	Scaffold	Scaffold size (bp)	Transcripts	Transcript annotation
AS	3068272	45,086	CRO_032476	Concanavalin A-like lectin protein kinase family protein
			CRO_032479	Concanavalin A-like lectin protein kinase family protein
			CRO_032478	hypothetical protein
			CRO_032477	Human Cytomegalovirus UL139 protein domain containing
			CRO_032475	**camelliol C synthase**
AO	2996021	16,843	**CRO_027941**	**cytochrome P450, family 716, subfamily A, polypeptide**

Triterpenoid biosynthesis has reported to be clustered in for example oat [117,118], lotus [175] and *A. thaliana* [176,177]. The functional annotation of further genes located on those two scaffolds does not indicate that they belong any of the known typical tailoring enzymes of the triterpene backbone such as glycosyltransferases or cytochrome P450s. However the obtained scaffolds are relatively short and therefore no final conclusion can be drawn whether or not triterpene biosynthesis in *C. roseus* is clustered.

Flavonoid biosynthesis has also been investigated in *C. roseus*. For example, chalcone synthase (*CS*) and flavonoid 3', 5'-hydroxylase (*F3'5'H*) from *C. roseus* have been characterised [208]. Cinnamate 4-hydroxylase (*C4H*), the P450 enzyme that catalyses the second step of the phenylpropanoid pathway [209], tetrahydroxy chalcone 2'glycosyltransferase (*THCGT*) [210] and 4'-O-methyltransferase (*CrOMT*) have also been shown to be involved in flavonoid biosynthesis in *C. roseus* [62]. None of these genes are linked to the same scaffold (Table 27).

Table 27 Scaffolds containing known flavonoid biosynthesis genes from *C. roseus*

Scaffold name and size in bp as well as all contained transcripts and transcript annotation. Transcripts corresponding to the genes *CS, F3'5'H, C4H, THCGT* and *CrOMT* are in bold.

Gene	Scaffold	Scaffold size (bp)	Transcripts	Transcript annotation
CS	2987430	18,619	CRO_014716	hypothetical protein
			CRO_014717	**Chalcone and stilbene synthase family**
F3'5'H	2965398	18,111	**CRO_022260**	**Cytochrome P450 superfamily protein**
C4H	2986973	98,188	CRO_000895	hypothetical protein
			CRO_000894	**cinnamate-4-hydroxylase**
THCGT	3066087	75,392	CRO_029295	conserved hypothetical protein
			CRO_029296	Got1/Sft2-like vescicle transport protein
			CRO_029297	**UDP-Glycosyltransferase superfamily**
CrOMT	3053220	69,879	CRO_022616	hypothetical protein
			CRO_022617	**O-methyltransferase family protein**

These results are not surprising; despite for example anthocyanin biosynthesis being transcriptionally very tightly clustered [76] there are no reports of gene clustering in flavonoid biosynthesis. In *A. thaliana* it has been demonstrated that the genes are scattered across the whole genome [211], as has also been established for carotenoid biosynthesis and the biosynthesis of glucosinolates.

4.3 Conclusion

C. roseus produces a great variety of structurally complex secondary metabolites, some of which with high medicinal value, but not all steps of the biosynthesis have been elucidated. Investigating the biosynthetic pathway of a taxonomically restricted class of specialised metabolites such as the ones produced by *C. roseus* is challenging. Simple homology models to available genomic data is of limited applicability because genes of specialised metabolism may evolve more quickly than those of primary metabolism and may not exist in a common reference organism such as *A. thaliana* [212].

The aim of the work presented in this chapter was to obtain the genome of a medicinal plant that produces a wide range of structurally diverse alkaloids. This would represent one of the first genomes of an alkaloid-producing plant. Of specific interest was to investigate if a draft genome sequence would be sufficient to detect any potential gene clustering in the *C. roseus* monoterpene alkaloid biosynthesis. In turn, the genome could serve as a source for future research by identifying candidates for still missing steps of the MIA pathway.

Whole genome sequencing of *C. roseus* genomic DNA on an Illumina HiSeq at The Genome Analysis Centre (Norwich, UK) and assembly using Abyss [178] yielded a draft genome for *C. roseus*. The sequencing output was 374,771,760,101 nucleotide paired-end reads to generate a 523 Mb assembly consisting of 79,302 scaffolds (> 200bp). This represents 71% of the previously determined genome size and is not unexpected, as the median value of coverage of predicted genome size, for all plant genomes published until 2013, has been reported to be 85%. This value is realistic due to areas in the genome that generally do not assemble such as repetitive structures, transposable elements but also functional elements of the genome such as telomeres and centromeres [162]. The corresponding scaffold N50 of the *C. roseus* genome assembly is 26.2 kb, with the size of the scaffolds containing the *C. roseus* alkaloid pathway genes ranging from 10 to 76 kb. This value is again comparable to the median N50 of all plant genomes published till 2013, sequenced with comparable Illumina techniques, that was reported as 25.9 kb [162]. The genome displayed excellent coverage, with 95.7% of all *C. roseus* ESTs mapping to the genome and 97.6% of the Core Eukaryotic Genes Mapping Approach (CEGMA) proteins mapping to the genome. All characterised MIA genes could be located in the genome as full length genes, with the exception of *SGD* and *T3O*, which had to be manually assembled.

In summary the draft genome provides a comprehensive representation of the genic regions of *C. roseus* making it a valuable resource. In the course of publishing our results [68] the v1.0 annotated draft genome was made available as a searchable database (http://medicinalplantgenomics.msu.edu/).

In parallel with genome sequencing efforts, a BAC library was created and screened for three major branch points in the monoterpene indole alkaloid pathway to increase the chances of identifying possible gene clustering in this pathway. The BAC assemblies gave additional information on genomic context of known pathway genes. In total six BACs were sequenced and assembled and the results presented in this thesis. Two BACs are published and the corresponding datasets can be found under NCBI WGS accession number ERP006960 and PRJEB7256.

A total of 25 alkaloid pathway genes were mined for their respective genomic context using the available draft genome and selected BAC assemblies. Following a minimum definition that

a gene cluster consists of two or more non-homologous genes that encode enzymes from the same pathway [105] our efforts have confirmed gene clustering in alkaloid biosynthesis in *C. roseus* in three cases (Figure 60).

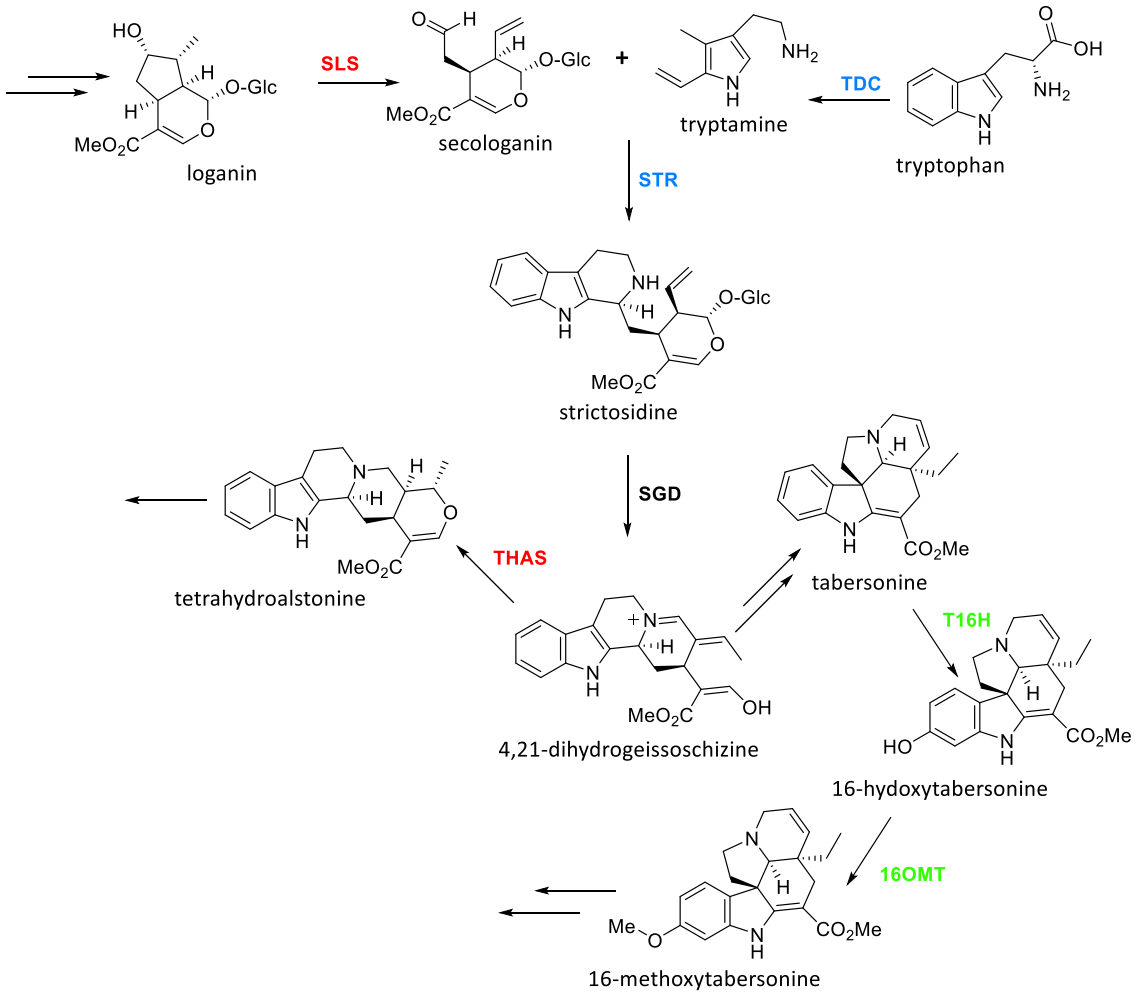


Figure 60 Detected gene clusters in monoterpene indole alkaloid biosynthesis in *C. roseus*

First gene cluster (blue): TDC, tryptophan decarboxylase; STR, strictosidine synthase and MATE transport protein (not in figure). Second gene cluster (green): T16H, tabersonine 16-hydroxylase (both T16H1 and T16H2) and 16OMT, 16-hydroxytabersonine O-methyltransferase. Third gene cluster (red) SLS, secologanin synthase, THAS, tetrahydroalstonine synthase. SGD, strictosidine β-glucosidase (not part of a detected gene cluster.)

Three clusters were observed:

1. The clustering of the pyridoxal dependent aromatic acid decarboxylase *TDC* that generates tryptamine from tryptophan and the 'Pictet-Spenglerase' *STR* that condenses the resulting

tryptamine with the iridoid secologanin to form strictosidine. The metabolic gene cluster is appears to be extended by a third gene. Silencing experiments performed by Dr. Richard Payne (manuscript submitted) strongly suggests the involvement of this MATE, a tightly co-expressed transport protein located in close physical proximity to *STR* and *TDC*, in transport of secologanin into the vacuole where *STR* is localised.

2. The two paralogs of 16-tabersonine hydroxylase *T16H1* and *T16H*, responsible for the hydroxylation of tabersonine at the 16 position [42], are located next to the 16-hydroxytabersonine-16-O-methyltransferase *16OMT* that installs a methyl group at the same position.

3. Four paralogues for the pathway gene SLS could be identified in the genome assembly. One of the SLS paralogues is co-localised with a gene that has been identified by Anna Stavrinides as tetrahydroalstonine synthase, the first gene to be reported to accept strictosidine aglycone and therefore the first to be implicated at this critical branch point of alkaloid diversity[37].

Considering that the size of most pathway gene containing scaffolds of the current first and only available *C. roseus* draft whole genome assembly was below 50000 bp, while reported gene cluster for secondary metabolites range from 35 to several hundred kb [174] the amount of clustering detected is encouraging. The analysis of the genome focused on identifying candidate genes for unknown steps in this alkaloid pathway. A total of 25 alkaloid pathway genes were mined for their respective genomic context using the available draft genome and selected BAC assemblies. Collectively, this search resulted in 90 transcripts that can be physically linked to the known alkaloid pathway genes. Of those, 31 transcripts are annotated as gene of unknown function and likely play no catalytic role in this biosynthetic pathway. Resulting from this list a candidate linked to root specific alkaloid biosynthesis did not show activity when heterologously expressed in yeast and fed with tabersonine and 19-hydroxytabersonine. A number of potentially promising biosynthetic gene candidates were targeted for further analysis by VIGS. The results of those experiments are provided in Chapter 5.

The monoterpene indole alkaloid pathway of *C. roseus* shows the presence of small clusters of genes comprising alkaloid and/or iridoid biosynthetic genes. It further displays several cases of pathways gene that are located in close distance of functional related genes, like *MAT* and *DAT* that are co-localised with other acetyltransferase genes or the five as UGT-glycosyltransferase

annotated contigs on the 7DLGT genomic scaffold. For other pathway genes paralogs have been observed, some of which have identical *in vitro* function, as it has been reported for *SLS* [213], *ISY* [181] or *T16H* [42]. Duplication of a gene, were one gene retains the original function, while the other one undergoes mutations, can lead to novel functions. This neo-functionalisation provides a metabolic plasticity that might in turn provide positive selection pressure enabling facile metabolic evolution [214,215]. Evolution of whole biosynthetic pathways depends strongly on the evolution of individual gene function. Noteworthy changes in gene regulation, as observed for the duplicated T16H1 T16H2 gene pair [42], and subsequently altered compound availability and spatial distribution can be linked to functional shifts [216]. Gene duplication and neo functionalisation has also been implicated in the recruitment of enzymes into secondary metabolic pathways. For example the *C. roseus* vindoline biosynthesis gene NMT, is suggested to have evolved from y-tocopherol C methyltransferases rather than other known N-methyl transferases, as phylogenetic analyses places *C. roseus* NMT clearly into a distinct clade with previously characterised y-tocopherol C methyltransferases for example from *Arabidopsis thaliana* [43].

Early research into a connection between co-expression of certain genes and their location in eukaryotic genomes gave statistical evidence that the frequency in which neighbouring are co-expressed is too high to be by chance [217]. In terpenoid biosynthesis, terpene synthases are responsible for the generation of scaffold diversity while P450s modify these scaffolds further, helping to create the largest class of plant-derived natural products. Available genomic and transcriptomic data for plants with specialised metabolism is still limited in comparison with for example *A. thaliana*. Here global analysis of 1469 publicly available expression profile datasets for *Arabidopsis thaliana* resulted in the prediction of around 100 gene clusters [78].

The availability of further genomes of additional monoterpene indole alkaloid producing plant species will uncover the extent to which they are conserved or if there is evidence for converged evolution of these gene clusters helping to evaluated and better understand the biological role of these clusters in plant.

Screening of BAC libraries [218] requires specialised equipment, such as a colony picking robot. Furthermore, BAC vectors are challenging to handle, and along with the size of the insert, additional care must be taken while handling the plasmid as BACs are known for their instability [219]. Given the outstanding improvements that have been made with whole genome sequencing of plants, BAC screening is likely to become obsolete.

Research into *C. roseus* is driven by the strong interest in its secondary metabolism. However, the interest in *C. roseus* extends beyond alkaloid biosynthesis into many other fields. For example, *C. roseus* is being studied as a host for phytoplasma, insect transmitted bacterial pathogens that are able to live intracellularly in plants as well as insect hosts [220]. Furthermore, research has focused on *C. roseus* for the investigation of signalling and transcriptional regulation especially of secondary metabolism [221–223] and a recent publication has specifically used the newly available genome data for the investigation of regulation of seco-iridoid biosynthesis in *C. roseus* [224].

At the start of this work, no genome sequence was available for *C. roseus* or any plant of the Apocynaceae family, and this resource will undoubtedly enable progress in dissecting alkaloid biosynthesis as well as permit further research in related fields.

Physical clustering of biosynthetic genes provides a powerful strategy to rapidly identify plant secondary metabolite pathways. Yet it is still unknown to what extent clustering occurs across the plant kingdom, and whether it is widespread or is limited to specific groups of metabolites. The phenomena of clusters of two or more paralogs is common in plants, while the occurrence of non-homologous clustering is a rather new observation. The clustering of genes in fungi and bacteria metabolic pathways greatly enables their discovery and characterisation [225]. It has to be expected that a more detailed knowledge as to if and how certain plant secondary metabolic pathways are clustered would similarly enhance their detection and characterisation.

5 Gene function screening by targeted gene silencing

5.1 Introduction

Virus Induced Gene Silencing (VIGS) is a transient posttranscriptional gene silencing technique used throughout this thesis. It utilises RNA interference (RNAi), the sequence specific inhibition of protein expression and provides an alternative functional genomics tool in non-model organisms such as *C. roseus* for which no reliable protocols for the generation of knockout mutants exists.

Monoterpene indole alkaloid biosynthesis in *C. roseus* requires the formation of secologanin via the monoterpene pathway and formation of tryptamine via the indole pathway. Both compounds together form strictosidine, the central intermediate from which all monoterpene indole alkaloids are derived. From strictosidine, after deglycosylation and a series of unknown enzymatic steps, three distinct alkaloid classes are derived. Tabersonine, catharanthine and ajmalicine each represent a different alkaloid class. Tabersonine is further converted to vindoline or 19-O-acetyl horhammercine (Figure 61). Vindoline and catharanthine can be dimerised to yield the anti-cancer agent vinblastine. At the beginning of this work five out of nine monoterpene genes leading to secologanin were unknown, no gene responsible for the formation of the three different alkaloids classes after strictosidine deglycosylation was known and two steps in vindoline as well as a step in 19-O-acetyl horhammercine biosynthesis were missing (Figure 61).

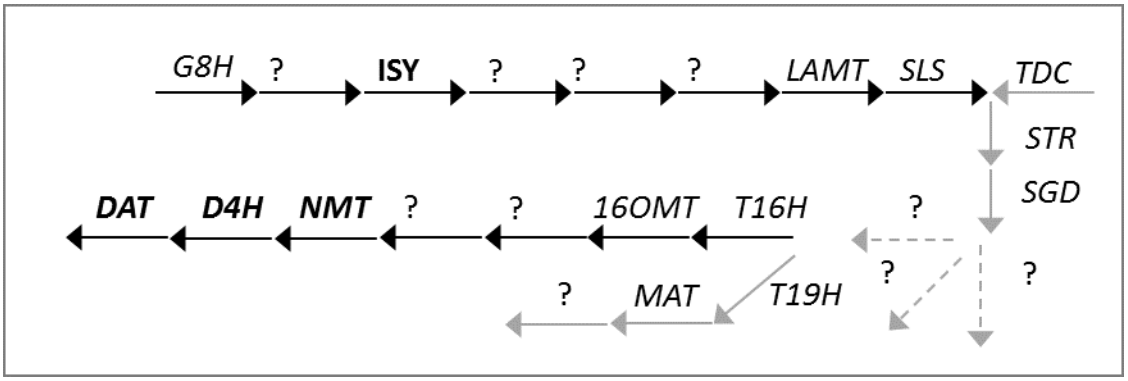


Figure 61 Known genes *C. roseus* alkaloid biosynthesis at the start of this thesis

Steps involved in the synthesis of alkaloids in *C. roseus* known at the start of this thesis. Black arrows represent single enzymatic steps. Dashed arrows represent an unknown number of enzymatic steps. Question marks represent unknown enzymatic steps. In bold are the genes of which the physiological function had been tested using the VIGS system. Black (top row) arrows represent at the time known iridoid biosynthesis genes: G8H, geraniol 8-hydroxylase; ISY, iridoid synthase; LAMT, loganic acid methyltransferase and SLS, secologanin synthase. Grey arrows represent then known downstream alkaloid biosynthesis genes: TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGD, strictosidine β-glucosidase; T19H, tabersonine/lochnericine 19-hydroxylase (CYP71BJ1) and MAT, minovincinine 19-hydroxy-O-acetyltransferase. Black arrows (middle row) represent then known vindoline biosynthesis pathway: T16H, tabersonine 16-hydroxylase 1 (CYP71D12); 16OMT, 16-hydroxytabersonine O-methyltransferase; NMT, 16-methoxy-2,3-dihydro-3-hydroxytabersonine N-methyltransferase; D4H, desacetoxyvindoline-4-hydroxylase and DAT, deacetylvindoline 4-O-acetyltransferase.

Additionally, only *DAT, D4H*, *NMT* [121] and *ISY* [30] had been subjected to gene silencing to confirm their physiological function *in planta* (Figure 61). All other known *C. roseus* alkaloid biosynthesis genes had been discovered before the VIGS system was established, or were expressed predominantly in tissues that were not accessible for silencing with the current VIGS protocol, like the root specific *T19H* [79].

5.1.1 Aim

This thesis focuses on exploring *C. roseus* alkaloid metabolism with an emphasis on gene discovery. The mining of existing transcriptomic data [65] and newly established genomic data [68] were utilised to ultimately identify candidates for missing steps in alkaloid biosynthesis in *C. roseus,* with a particular emphasis on finding the missing enzymes in vindoline biosynthesis, by testing identified candidates employing VIGS for targeted gene silencing.

As presented in Chapter 2 VIGS is suitable for the *in planta* functional characterisation of alkaloid biosynthesis genes. The work presented in Chapter 3 highlights how VIGS can additionally be utilised to test candidates for the discovery of new genes by screening for a specific change in metabolite content of silenced tissue.

The aim of this last chapter is to present how VIGS can be used to screen a large set of candidates for which a general involvement in the pathway is suspected without prior knowledge as to where in the pathway the candidate would be potentially involved. This includes unknown steps in the seco-iridoid pathway as well as further downstream where an unknown amount of unidentified enzymatic steps is responsible for the formation of three different alkaloid scaffolds from the precursor deglycosylated strictosidine (Figure 6)

A VIGS experimental pipeline was set up and optimised. Once this was established, the pipeline was maintained by continuously testing new gene candidates and over the course of this thesis 28 different genes were silenced. The VIGS pipeline was employed to screen of candidate genes for their potential involvement in alkaloid biosynthesis in *C. roseus* by investigating if their targeted expression reduction would lead to significant changes in alkaloid content and composition. At the start of this thesis much progress had been made in the discovery of the missing monoterpene biosynthesis steps (personal communication Dr. Nathanial Sherden, Dr. Fernando Geu-Flores, Dr. Wesley Glenn, O'Connor lab) therefore discovery of genes responsible for this upstream part of the pathway was of lowest priority.

In this chapter, the optimisation of the VIGS pipeline is described, and the candidates selected for screening are provided. The motivation that was used to select the particular candidate is presented. The results of some of the more important gene candidates are discussed in more detail. The findings are discussed in the greater context of gene function screening by targeted gene silencing in plant secondary metabolism.

5.2 Results and Discussion

For every VIGS candidate, data from five to eight plants was collected. Each VIGS experiment contained the same number of plants treated with an empty vector control. Plants were eight weeks old at the time of infection. At 21 days after infection the last two leaves to emerge above the pinch site were harvested from the *C. roseus* plants and milled under liquid nitrogen. Each tissue sample was split in two aliquots. From one, metabolites were extracted and analysed by LC-MS to assess potential changes in metabolite profile by comparing plants

silenced for a candidate gene to the empty vector control. The remaining tissue was used for RNA extraction for q RT PCR. Primers used for cloning can be found under Chapter 6, Table 37. Sequences of the candidate genes used for VIGS can be found under Appendix 9.4. The VIGS candidates, reason for selection of candidates and results are summarised in the following table (Table 28).

Table 28 VIGS candidates

Candidates for VIGS experiments with contig number, annotation and reason for silencing/ origin of contig. **A**: Functional *in planta* validation of known pathway genes. **B**: Co-localisation with known alkaloid pathway genes on the *C. roseus* draft genome. **C**: Co-localisation and co-expression with known alkaloid pathway gene *SGD* on the *C. roseus* genome **D**: Co-expression analysis in scope of search for *T3O* (Chapter 3). **E:** Co-expression analysis, Subset_A: P450 candidates for missing step in vindoline biosynthesis (Chapter 3). **F**: Co-expression analysis Subset_A: Reductase candidates for missing step in vindoline biosynthesis (Chapter 3).

	contig	annotation	origin of candidate	number of experiments	significant changes, (p-value <0.01)
C36	CRO_032842	secologanin synthase/ cytochrome P-450 protein	(A)	3	yes (Chapter 5.3.1.)
C38	CRO_006099	strictosidine synthase		3	yes (Chapter 5.3.1.)
C39	CRO_006098	tryptophan decarboxylase		3	yes (Chapter 5.3.1.)
C16	CRO_017449	cytochrome P450, family 71, subfamily B	(B)	2	no (Chapter 5.3.2.)
C18	CRO_011779	cytochrome P450, family 71, subfamily B	(C)	2	yes (Chapter 5.3.3.)
C2	CRO_027093	cytochrome P450, family 71, subfamily B, polyp.		1	no
C6	CRO_027095	NAD(P)-binding Rossmann-fold superfamily		1	no
C7	CRO_027094	Haloacid dehalogenase-like hydrolase (HAD)		1	no
C8	CRO_003636	cytochrome P450, family 97, subfamily B, polyp.		1	no
C9	CRO_021613	cytochrome P450, family 97, subfamily B, polyp.		1	no
C45	CRO_019099	alpha/beta-Hydrolases superfamily protein	(C) & (D)	3	yes (Chapter 5.3.4.)
C43	CRO_016518	HXXXD-type acyl-transferase family protein	(D)	2	no
C46	CRO_016488	Chalcone-flavanone isomerase family protein		3	no
C3	CRO_026945	cyclase family protein		1	no
C4	CRO_026235	cinnamyl alcohol dehydrogenase		1	no
C10	CRO_010055	cytochrome P450, family 721, subfamily A, polyp.		1	no
C12	CRO_019049	cytochrome P450, family 78, subfamily A, polyp.		1	no
C30	partial: CRO_002621, CRO_015047, CRO_023156, CRO_031187	cytochrome P450, family 71, subfamily B, polyp.	(E)	1	no
C44	CRO_013485 & CRO_023307	cytochrome P450, family 76, subfamily C, polyp.		1	no
C47	CRO_003361	cytochrome P450, family 76, subfamily C, polyp.		2	no
C48	CRO_027850	cytochrome P450, family 71, subfamily B, polyp.		1	no
C41	CRO_016423	methylenetetrahydrofolate reductase	(F)	1	no
C14	CRO_021541 (T3R)	elicitor-activated gene 3-2		1	no
C53	CRO T030221	12-oxophytodienoate reductase		1	no
C32	CRO T002544	NAD(P)-linked oxidoreductase superfamily		1	no
C15	CRO_012109	NAD(P)-linked oxidoreductase superfamily		2	no
C13	CRO_008301	elicitor-activated gene 3-2		1	no
C5	CRO_001335	NAD(P)-binding Rossmann-fold superfamily		1	no

5.2.1 VIGS for gene function validation in planta

Three VIGS candidates were chosen to establish an efficient VIGS pipeline. These are known alkaloid biosynthesis genes acting at the central section of alkaloid biosynthesis in *C. roseus*. Tryptophan decarboxylase (TDC) [36] catalyses the conversion of tryptophan to tryptamine, loganin is cleaved by secologanin synthase (SLS) [35] to yield the iridoid terpene secologanin, after which strictosidine synthase (STR) [189] condenses secologanin and tryptamine to form the central metabolic intermediate, strictosidine [18] (Figure 62). None of these genes had ever been silenced in planta previously.

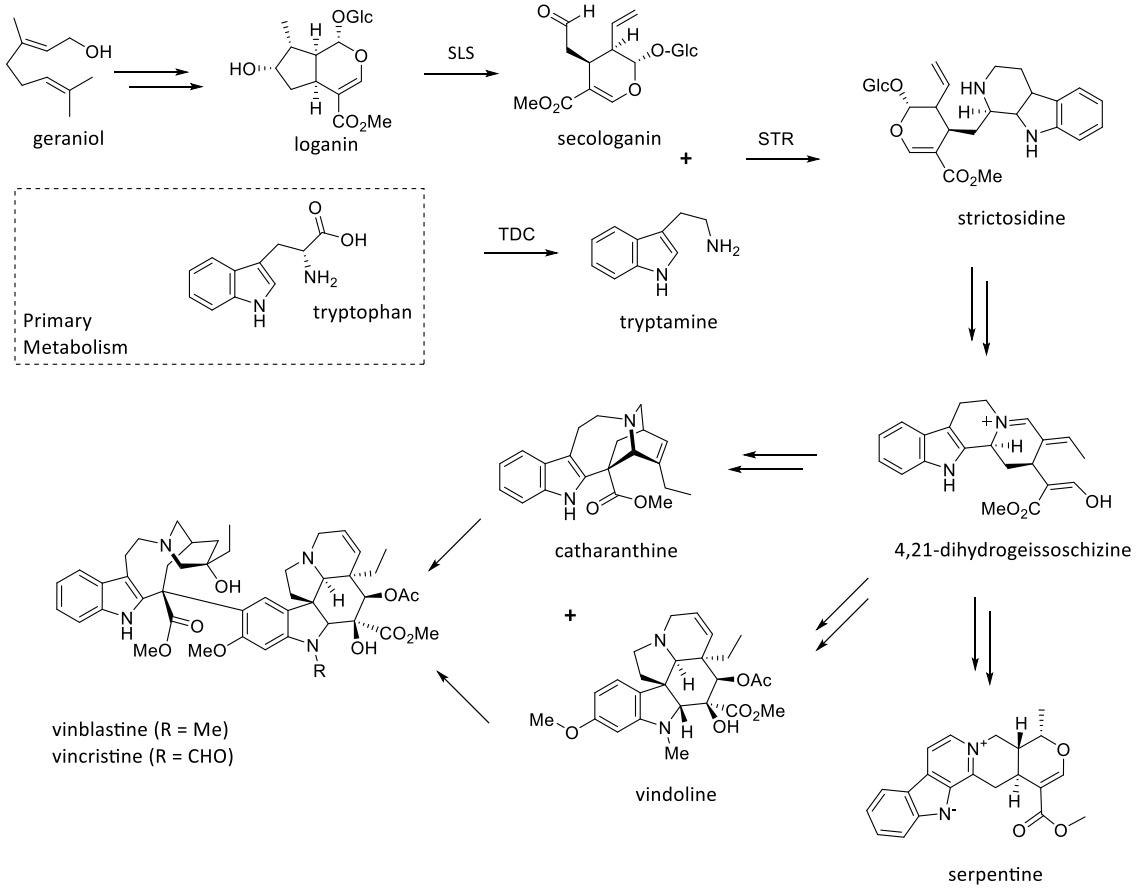


Figure 62 Products and substrates of central enzymes in *C. roseus* alkaloid biosynthesis

Secologanin is a glycosylated iridoid terpene derived from geraniol by a nine-step biosynthetic pathway. The final enzyme in this seco-iridoid pathway is the cytochrome P450 SLS (secologanin synthase). Tryptamine is synthesised from the decarboxylation of tryptophan by the enzyme TDC (tryptophan decarboxylase). Secologanin and tryptamine are condensed by STR (strictosidine synthase) to form strictosidine. After deglycosylation of strictosidine and rearrangement to 4,21-dehydrogeissoschizine. *C. roseus* can produce bisindole alkaloids such as vinblastine and vincristine, as a result of the condensation of catharanthine, an iboga alkaloid, and vindoline, an aspidosperma alkaloid.

For all three genes a section of the respective gene was cloned into pTRV2, the plasmid was used to transform *A. tumefaciens* strain GV3101 and eight week old plants were pinched in three subsequent experiments. It had been established in the course of this thesis that four almost identical paralogs for SLS exist that share between 94 and 98% sequence identity on nucleotide level and are all expressed in *C. roseus* leaves. This high similarity ensures that any VIGS construct ultimately targets the expression of all four SLS paralogs simultaneously.

Successful silencing was confirmed by qRT PCR for all three genes. Expression was normalised to the expression of a known *C. roseus* housekeeping gene[30,121], the 40S ribosomal protein 9 (*Rps9*). A 43%, 72% and 66% decreased expression for *TDC*, *SLS* and *STR* respectively could be observed (Figure 63).

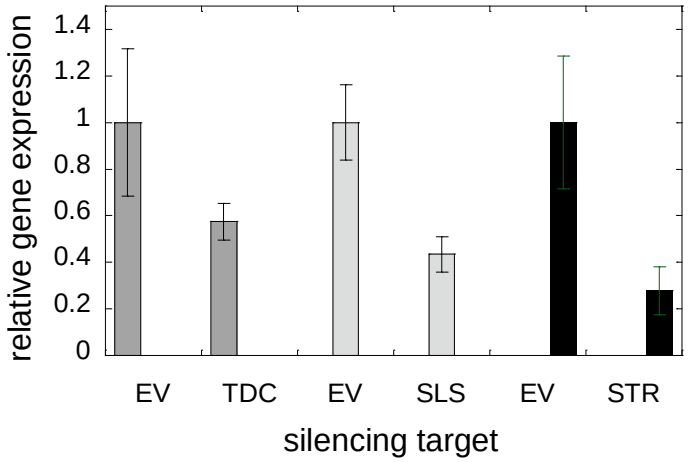


Figure 63 Silencing of *TDC*, *SLS* and *STR*

Relative expression of TCD, SLS and STR in silenced tissues compared to EV tissue (n=6). Expression was normalised to expression of CrRPS9. Data presented is the mean ± SEM. Efficiency for qRT PCR primer is 110% (STR), 106.8% (SLS) and 104.3% (TDC).

VIGS reduces the transcription but the effect is a “knock-down” and not a “knock-out” effect. The results are within the expected range of in literature reported values for silencing of *C. roseus* genes that varies from 45% [30] to 80% reduction [32]. Interestingly the downregulation of all paralogs of *SLS* was successful. The primer pair for the qRT PCR experiments was designed to ensure the transcriptional abundance of all four paralogs would be detected equally.

For the known metabolites secologanin, strictosidine, catharanthine, tabersonine, vindoline and vindorosine (Figure 16) the peak area was manually extracted from the LC-MS data, normalised and the mean was compared to the mean of the empty vector control tissue. Three subsequent silencing experiments were conducted. Over this period the protocol was optimised in a way that would ensure identical treatment for all samples by randomisation at all crucial steps and minimal time delay in sample preparation to ensure maximum homogeneity of the results. Only data that has a p-value of <0.05 is considered. All experiments included eight plants per treatment.

5.2.1.1 TDC silencing results

Although silencing of *TDC* results in the smallest decrease in expression, it results in one of the strongest phenotypes when comparing the results of the three genes. Of the investigated six metabolites secologanin, strictosidine, catharanthine, tabersonine, vindoline and vindorosine three to four are significantly changed in silenced tissue of three separate VIGS experiments (Figure 64).

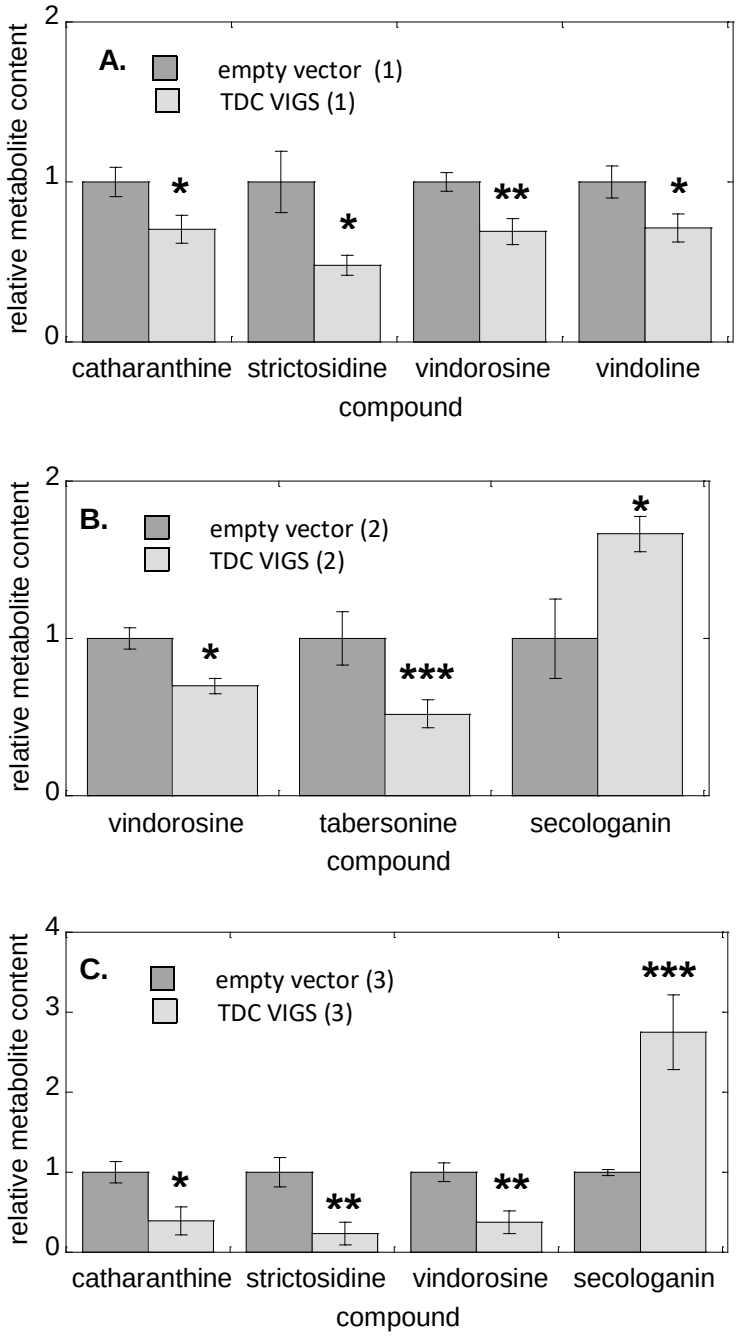


Figure 64 *TDC* silencing effect on leaf metabolite content in three *TDC* VIGS experiments

Relative metabolite content of silenced plants as mean peak area calculated from extracted ion chromatograms of corresponding compounds, normalised by fresh sample weight and internal standard, relative to that determined in EV control tissues (normalised to 1). Compounds significantly changed in *TDC* silenced plants in comparison to EV control plants (n=8). Error bars represent ± SEM. P-value was calculated using Student's *t* test (*p<0.05, **p<0.01, ***p<0.005). **A**: First experiment. **B**: Second experiment. **C**: Third experiment.

The substrate for the enzyme TDC is the amino acid tryptophan. While no accumulation of tryptophan was observed, this is expected for a substrate that can be metabolised in so many different ways in plants. Additionally, neither tryptophan nor tryptamine, the TDC substrate and enzymatic product, are detectable with the standard LC-MS method used. The formation of strictosidine, from which all downstream alkaloids originate requires the condensation of tryptamine and secologanin. In two experiments strictosidine is significantly decreased and secologanin significantly increased. This indicated that the reduced availability of tryptamine, the TDC enzymatic product, limits the formation of strictosidine. In turn this leads to the decreased accumulation of downstream alkaloids such as catharanthine, tabersonine, vindoline and vindorosine. In contrast, secologanin accumulates since not enough tryptamine is available.

5.2.1.2 SLS silencing results

Strong effects of silencing could also be observed for silencing *SLS*. Levels of secologanin, strictosidine, catharanthine, tabersonine, vindoline and vindorosine were variable among the three VIGS experiments. However, the appearance of loganin, the substrate for SLS that is normally below the detection limit, was found to be significantly increased in all *SLS* silenced tissues, which provides the strongest evidence of silencing success (Figure 65).

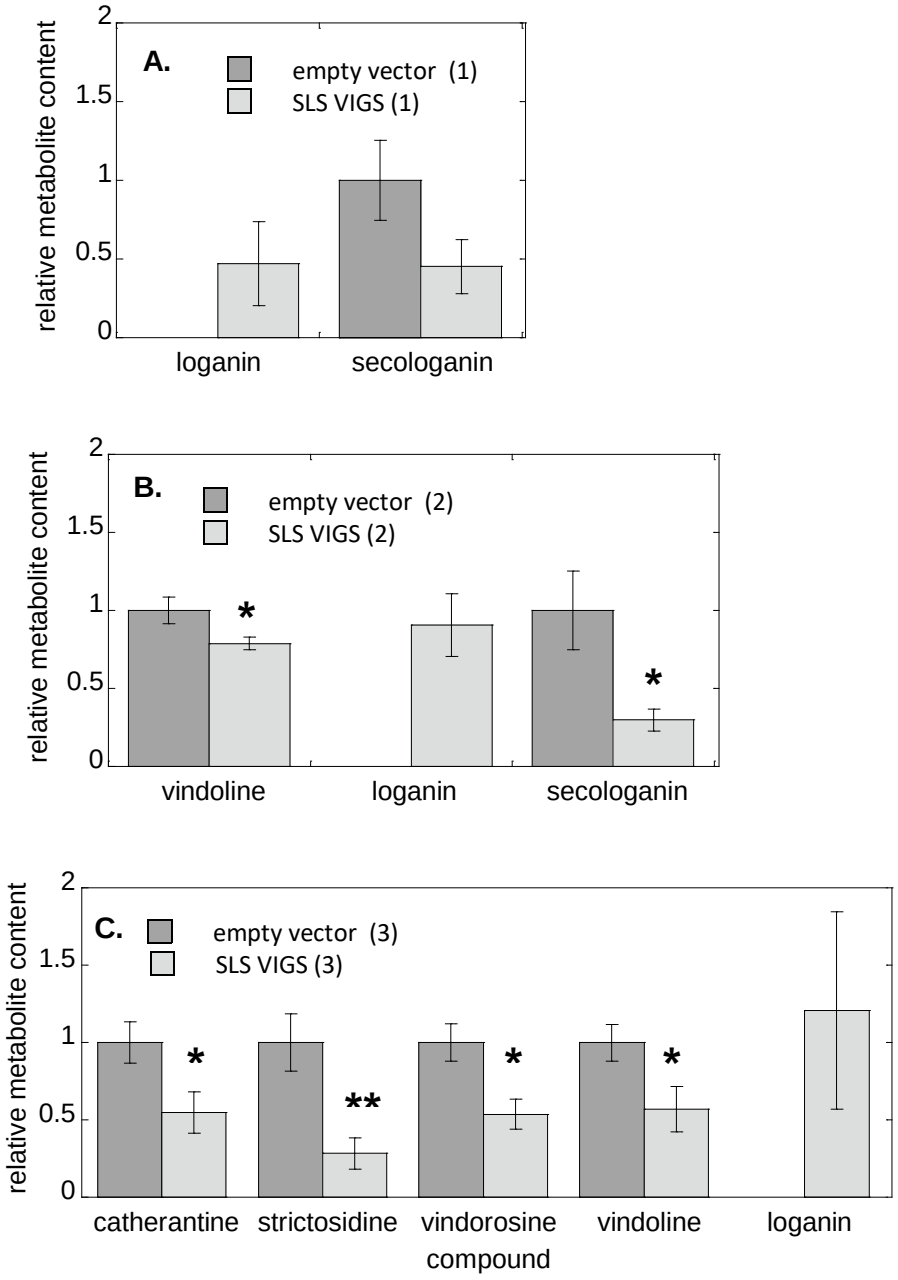


Figure 65 *SLS* silencing effect on leaf metabolites in three subsequent VIGS experiments

Relative metabolite content of silenced plants as mean peak area calculated from extracted ion chromatograms of corresponding compounds, normalised by fresh sample weight and internal standard, relative to that determined in EV control tissues (normalised to 1). Compounds in *SLS* silenced plants in comparison to empty vector control plants (n=8). The new loganin peak corresponds to the mass of m/z 413 (Na^+ adduct). Error bars represent $\pm$ SEM. P-value was calculated using Student's *t* test (*p<0.05, **p<0.01). **A**: First experiment. Although the decrease in secologanin is not significant here it was later found significant using a second LC-MS measurement on a different instrument (Xevo, Waters) applying a loganin and secologanin specific MRM method. **B**: Second experiment. **C**: Third experiment.

Secologanin synthase is a cytochrome P450 enzyme [35] that catalyses the last step in the seco-iridoid biosynthesis pathway, the oxidative ring cleavage converting loganin into secologanin (Figure 66).

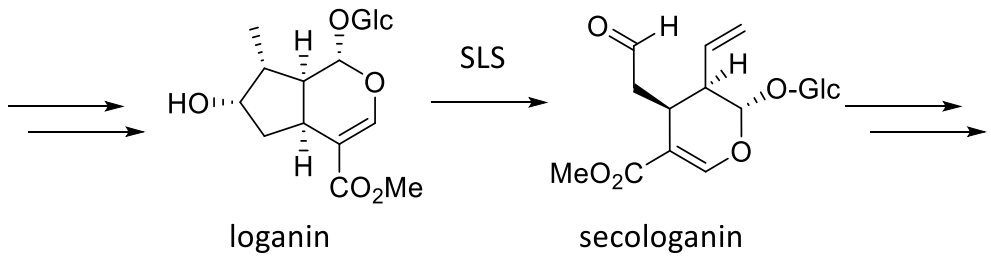


Figure 66 Secologanin synthase catalysed reaction in *C. roseus* alkaloid biosynthesis

To confirm the results for loganin accumulation and secologanin decrease the Waters Xevo TQ-S mass spectrometer (Milford, MA, USA) equipped with an electrospray (ESI) source was used and analysis was carried out by monitoring specific mass transitions for loganin and secologanin. For this purpose a method was developed using commercial standards. Flow injection of individual compounds was used to optimize the multiple reaction monitoring (MRM) automatically using the Waters Intellistart software. As expected *SLS* silenced tissue displays a strong accumulation of loganin and a reduction of secologanin (Figure 67).

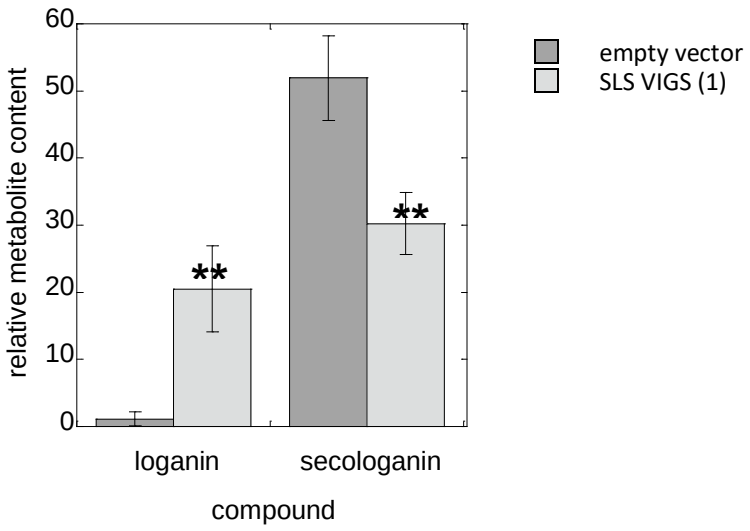


Figure 67 Loganin and secologanin in *SLS* VIGS tissues

Relative loganin and secologanin content as determined by monitoring specific mass transitions for loganin and secologanin/ MRM method. Displayed is relative metabolite content as mean peak area in *SLS* silenced plants in comparison to EV control plants (n=8). The tissues correspond to the tissues of the first VIGS experiment. Error bars represent ± SEM. P-value was calculated using Student's *t* test (**p<0.01).

VIGS results for *SLS* have been subsequently published by another group after our initial experiments. Here comparable results to the ones obtained in the course of thesis are reported, including the appearance of loganin and the significant reduction of secologanin, catharanthine and vindoline [32].

5.2.1.3 STR silencing results

The effect of silencing *STR* results in a much weaker metabolic change compared to the effects observed for *SLS* and *TDC* silencing. Again the results were variable among the experiments (Figure 68). The enzyme STR catalyses the Pictet-Spengler condensation of secologanin and tryptamine to form strictosidine, the central metabolic intermediate and key branch point in the alkaloid pathway of *C. roseus* (Figure 69). It is not clear why reduction of the expression of the *STR* gene only leads to a minimally changed metabolite phenotype but it highlights the challenges of interpreting VIGS data during functional screening.

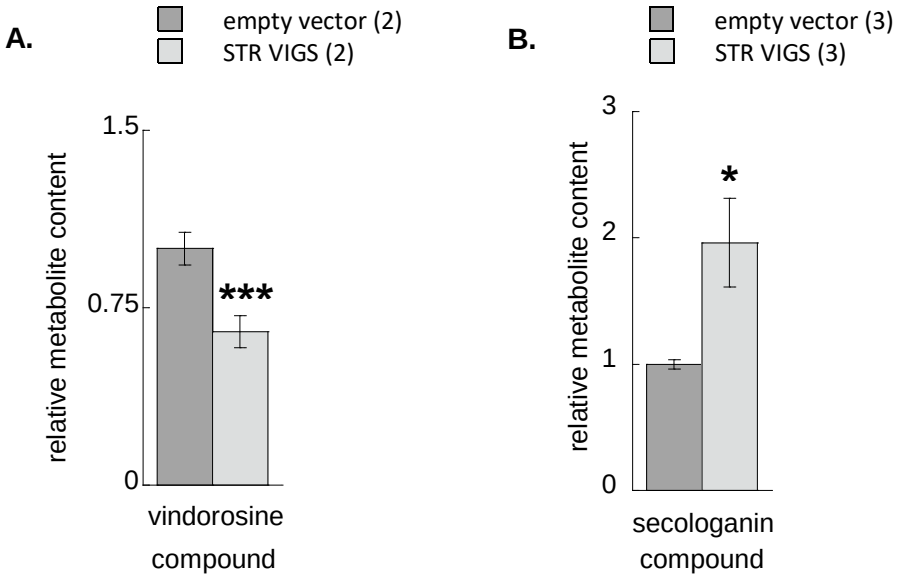


Figure 68 Significant changes in *STR* silenced leaf tissue

Relative metabolite content of silenced plants as mean peak area calculated from extracted ion chromatograms of corresponding compounds, normalised by fresh sample weight and internal standard, relative to that determined in EV control tissues (normalised to 1). Content displayed for vindorosine in the second VIGS experiment and secologanin in the third experiment of *STR* silencing in comparison to EV control plants (n=8, n=5). Error bars represent ± SEM. P-value was calculated using Student's *t* test (*p<0.05, **p<0.01, ***p<0.005). **A:** Second VIGS experiment. **B:** Third VIGS experiment. No statistically significant changes were noted in the first VIGS experiment.

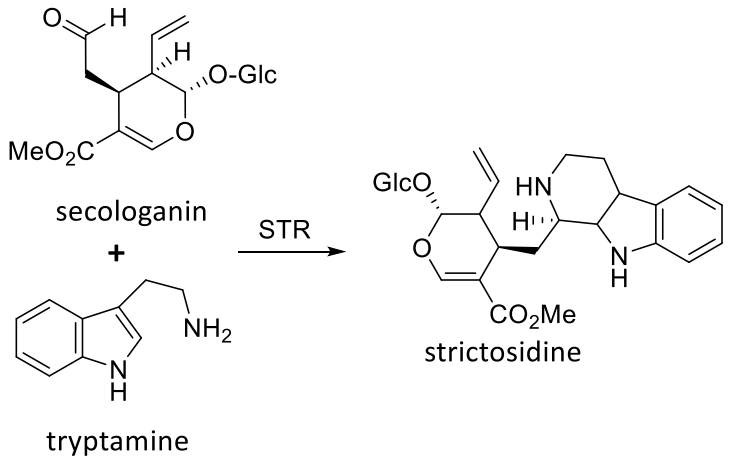


Figure 69 Strictosidine catalysed reaction in monoterpene alkaloid biosynthesis

5.2.1.4 Summary of metabolic profiles with TDC, STR and SLS silenced tissue

The three known pathway genes *TDC*[36], *STR*[189] and *SLS*[35] were silenced to validate their physiological role in *C. roseus* leaf tissue and to establish a robust VIGS system for further applications. Reduction of expression of silencing target genes could be confirmed with qRT PCR. The *TDC*, *STR* and *SLS* silenced plants show a 43%, 72% and 66% decreased expression of the respective target gene. The successful gene silencing of the target genes *TDC*, *STR* and *SLS* had the expected significant effects on alkaloid metabolism in the silenced tissues compared to empty vector treated tissues, though the results showed variability.

The successful downregulation of *TDC* gene expression expectedly caused the decrease of downstream alkaloids (Figure 64). The three *SLS* VIGS experiments show a weaker overall response to the silencing. Only the vindoline content shows reduced levels consistently (Figure 68). The interpretation of the VIGS phenotype in the *TDC* and *SLS* silencing experiments is strengthened by the observation of accumulation of the expected substrates. In the *TDC* VIGS experiments secologanin is significantly increased, likely due to the reduced availability of tryptamine to react with secologanin to form strictosidine. In *SLS* silenced tissues, loganin, the substrate for SLS, accumulates to substantial levels compared to empty vector controls. In combination with the decrease of downstream alkaloids the evidence strongly supports the physiological relevance of *SLS* and *TDC* in alkaloid metabolism in leaves of *C. roseus*. The

minimal VIGS phenotype of the *STR* silencing clearly shows some of the major limitations of the VIGS system.

In conclusion it was established that the number of plants and method used for silencing was sufficient for the observation of statistical significant changes. Previous VIGS studies in *C. roseus* had used the plant variety "Little Bright Eyes" [30,121]. The new variety "SunStorm Apricot" used in all experiments conducted in the scope of this thesis proved more homogeneous in growth and alkaloid content.

5.2.2 VIGS experiment following gene clustering analysis

The obtained *C. roseus* genome assembly was used for a detailed analysis of the genomic context of the 25 known alkaloid biosynthesis genes. Gene clustering was observed in the *C. roseus* draft genome and the resulting candidate genes have been published [68]. The most relevant candidates were investigated separately. A contig annotated as sinapyl alcohol dehydrogenase, located on one the *SLS* genomic scaffolds, has been characterised as the corynanthe biosynthetic gene *THAS* [37]. The *MATE* transport protein encoding gene located on the *TDC/ STR* genomic contig has been successfully silenced and the results suggest that this acts as a secologanin transporter (Dr. Richard Payne, submitted).

For the 25 investigated alkaloid pathway genes 90 transcripts could be identified as co-located, meaning that they can be linked to the known alkaloid pathway genes by their physical proximity on the *C. roseus* draft genome. All 90 contigs were investigated manually. In a first step, a BLASTn search was conducted against the NCBI database to confirm the annotation. Furthermore it was investigated if the contig represents a full open reading frame or if the contig is likely a partial/ pseudogene. The confirmed annotation and the expression profile for each contig served as the basis for which contig would be relevant for testing with VIGS.

5.2.2.1 Contigs with cytochrome P450 annotation

Specific attention was paid to contigs annotated as members of the cytochrome P450 gene family because eight of the known monoterpene indole alkaloid biosynthesis enzymes in *C. roseus* are cytochrome P450 enzymes (Figure 70) including the *T3O* gene reported in this thesis (Chapter 3) and the functionally identical (Chapter 2) and physically clustered (Chapter 5) *T16H1* and *T16H2.*

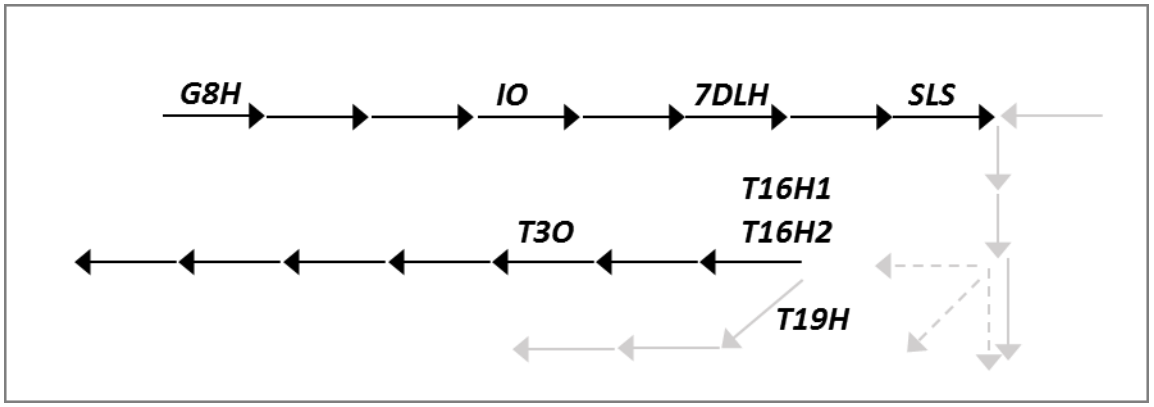


Figure 70 Cytochrome P450 known in MIA biosynthesis in *C. roseus*

P450s of the monoterpene indole alkaloid pathway in *C. roseus*. Dashed arrows indicate unknown numbers of enzymatic steps. Black top row arrows represent iridoid biosynthesis genes: G8H, geraniol 8-hydroxylase; IO, iridoid oxidase and SLS, secologanin synthase. Grey arrows represent downstream alkaloid biosynthesis genes: T19H, tabersonine/lochnericine 19-hydroxylase Black arrow middle row represent vindoline biosynthesis pathway: T16H1, tabersonine 16-hydroxylase 1; T16H2, tabersonine 16-hydroxylase 2 and T3O 16-methoxytabersone 3-oxygenase.

It is believed that indeed P450 enzymes and their huge diversification is one the major driving forces in phytochemical diversity with members of this large super family of monooxygenase are present in most domains of life such as plants, animals, fungi, bacteria and even viruses [226]. P450s typically catalyse a variety of hydroxylation reactions but there is ample evidence for unusual P450 reactions such as oxidative rearrangement of carbon skeletons or C-C bond cleavage, specifically in plant secondary metabolism [227]. The details of the mechanism underlying these complex reactions are often not very well understood. Additionally P450s are often implicated in plant chemical defence [228].

Six of the 90 co-located contigs are annotated as cytochrome P450 (CRO_021081, CRO_006603, CRO_011779, CRO_025931, CRO_017449 and CRO_004355).

CRO_021081 is located next to T19H, a gene involved in the root specific alkaloid biosynthesis[79]. CRO_021081 or candidate C72, has been investigated in this work using a yeast expression system as the candidate is not expressed in leaf tissue and was, according to its genomic location next to a root specific alkaloid pathway gene, implicated in root specific alkaloid production (Chapter 4).

CRO_006603, found on the STR_BAC, has been eliminated from VIGS experiments because of its very low gene expression. CRO_011779 was identified on the SGD_BAC and the results for this candidate (C18) are presented in the context the SGD_BAC and SGD MATE-pair assembly below (Chapter 5.3.3).

Three P450s, CRO_025931, CRO_004355 and CRO_017449, were found in the context of the T16H2_BAC assembly. All three contigs are annotated as members of the cytochrome P450 CYP71 family, the same family as the know alkaloid pathway genes *T16H1* and *T16H2* [42]. CRO_025931 is a partial (only 612 bp) duplication of *T16H2,* sharing 93.3% identity. It represents most likely a pseudogene with no full open reading frame. CRO_004355 is expressed at extremely low levels in the *C. roseus* tissues leaf, stem, flower, seedling or root, making it a very unlikely candidate for alkaloid biosynthesis. CRO_017449 is only 441 bp long and again does not represent a full gene. The 441 bp of this transcript can be found in three other transcripts (Table 29) suggesting either an assembly fragment or a pseudogene.

Table 29 Nucleotide similarity in % between CRO_017449 and related transcripts

	CRO_T017449
CRO T014272	100%
CRO T014273	98.60%
CRO T017449	100%
CRO T022496	100%

Despite the fact that CRO_017449 (named C16), and all transcripts relating to C16 (Table 29), have relatively low expression values in the alkaloid producing relevant tissues such as flower, leaf, seedling, stem and root (Figure 71), a 208 bp long section was cloned into the VIGS vector and 8 plants infiltrated with *A. tumefaciens* harbouring the resulting C16 construct.

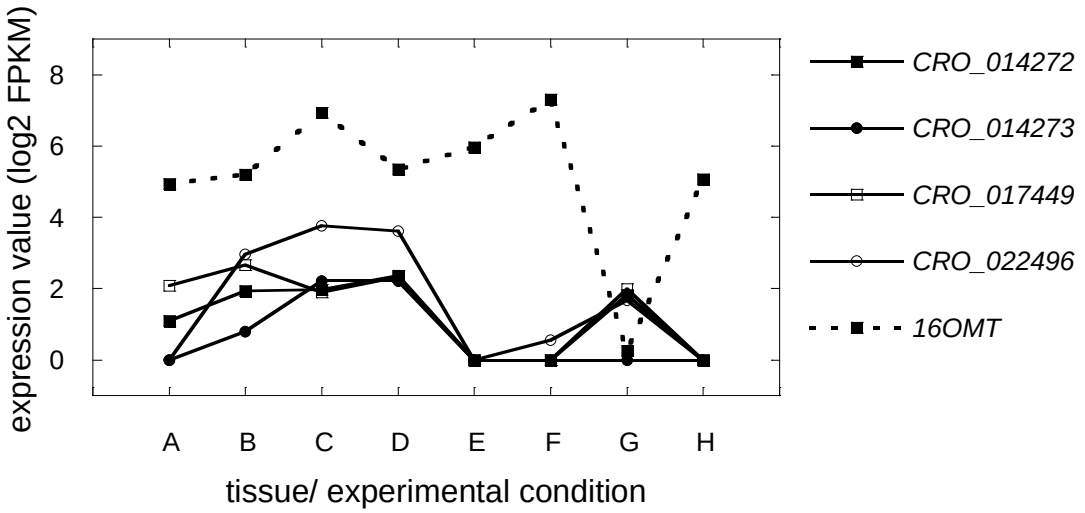


Figure 71 Expression (log_2FPKM) of CRO_017449/ C16 and related transcripts

Expression values (FPKM) for eight tissues and/or treatments for the contig CRO_017449 and its related contigs in comparison to the known pathway genes *16OMT* to which CRO_017449 is co-located in the *C. roseus* draft genome. Gene expression is displayed for the tissues: A (flower), B, C and D (sterile seedlings treated with MeJA 0, 6 and 12 h), E (mature leaf), F (young leaf), G (stem) and H (root).

The LC-MS measurements of candidate C16 silenced tissues in comparison to the empty vector did not show any new peaks or significant changes of known metabolite peaks (Table 28). The VIGS experiment was not repeated and unless additional evidence is uncovered, this gene is assumed not to play a role in MIA biosynthesis.

5.2.2.2 Expression values of co-localised contigs

Of the 90 contigs identified to be located in close proximity on the draft *C. roseus* genome to known alkaloid pathway genes, we considered only contigs that can be assessed using VIGS, and therefore excluded all transcripts that display an expression in young leaves of less than 1 (log_2FPKM), reducing the number from 90 to only 32 (**Error! Not a valid bookmark self-reference.**). The remaining 32 contigs include for example the *ISY* paralog for which VIGS had been reported [181]. So far only a single contig (CRO_011779/ C16/ Chapter 5.3.2.1) has been tested in the VIGS pipeline as reported above, as the annotation of the other candidates do not suggest an involvement in alkaloid biosynthesis.

Table 30 Annotation and expression values for 32 contigs

Functional annotation of 32 contigs that are located in physical proximity to known alkaloid pathway genes with an expression value of >1 log_2FPKM in young leaves. Highlighted in yellow is VIGS candidate C18.

		mature leaf	young leaf
CRO_026512	oligopeptide transporter	0.000	1.295
CRO_029921	Syntaxin/t-SNARE family protein	0.249	1.395
CRO_025030	expansin A1	0.875	3.544
CRO_009298	hypothetical protein	1.140	1.864
CRO_006609	Polynucleotidyl transferase, ribonuclease H	1.769	2.635
CRO_011779	cytochrome P450, family 71, subfamily B	1.771	2.263
CRO_025462	iridoid synthase paralog	2.205	3.189
CRO_006605	Galactosyltransferase family protein	2.321	2.665
CRO_005951	glyoxalase II	2.380	3.422
CRO_015508	LOB domain-containing protein	2.519	1.998
CRO_015513	Glycosyl hydrolase family 35 protein	2.629	4.781
CRO_029918	TBP-associated factor	2.748	2.833
CRO_015512	Family of unknown function (DUF566)	2.846	3.040
CRO_027004	UDP-glucosyl transferase	2.893	3.764
CRO_015509	nucleic acid binding	3.046	3.383
CRO_015507	salt tolerance zinc finger	3.247	4.901
CRO_029922	ankyrin repeat family protein	3.571	4.108
CRO_011780	Peptidase C78, ubiquitin fold modifier-specific	3.578	4.006
CRO_021079	Peptidase family M48 family protein	3.672	3.779
CRO_027001	UDP-glucosyl transferase	3.698	4.388
CRO_029919	casein kinase alpha	3.796	4.416
CRO_015221	DUF1649 domain containing protein	3.819	3.820
CRO_029920	Peroxisomal membrane 22 kDa (Mpv17/PMP22)	4.340	4.227
CRO_025459	cytochrome B5 isoform B	4.356	5.437
CRO_025460	glycine-rich protein	4.701	5.473
CRO_006608	aconitase	4.731	4.929
CRO_006607	Ubiquitin-like superfamily protein	4.753	4.373
CRO_005950	Ribosomal protein L18ae family	4.812	5.835
CRO_004424	RNA polymerase I-associated factor PAF67	4.901	5.211
CRO_025458	alpha/beta-Hydrolases superfamily protein	4.928	2.782
CRO_015511	adenine phosphoribosyl transferase	5.859	6.386
CRO_021078	DEAD box RNA helicase (PRH75)	7.136	7.545

5.2.3 VIGS experiments following analysis of SGD genomic context

A *C. roseus* Mate pair library was obtained and in combination with the previously generated sequencing data [68] a new improved assembly was created by Robin Buell and co-workers at Michigan State University (MSU, East Lansing, USA) (data not publicly available).

The analysis of this unpublished assembly was restricted to the *SGD* gene primarily because this data became available only at a late point of this work. *SGD* is the only gene that failed to be assembled into a single contig in the initial assembly based on the Illumina sequencing data

[68]. The deglycosylation of strictosidine is a crucial step in the biosynthesis of alkaloids in *C. roseus* as the resulting aglycone undergoes a sequence of unknown enzymatic reactions that are the basis for downstream differentiation of the pathway yielding the different alkaloid classes [18] (Chapter 1, Figure 6). Therefore, this is a potentially interesting place to search for gene clustering.

The *SGD* containing scaffold resulting from the Mate pair enhanced assembly has a length of 1,026,926 bp with 88 individual transcripts mapping to this sequence. The *SGD* gene is located between position 231,468 and 294,861. Closer inspection of this genomic context revealed that the assembly results in three *SGD* versions in this area of which only one represents the full open reading frame of the *SGD* gene.

For further analysis, all transcripts between position 1 and 800903 bp, a total of 68 transcripts, were considered. For those 68 transcripts the expression values in the tissues flower, sterile seedling (control, plus methyl jasmonate for 5 days or 12 days), mature and immature leaf, stem and root were investigated. A total of 35 transcripts were not followed up as they were either not expressed at all in those tissues or expressed with a value of less than 1 $\log_2$FPKM in young leaf. The remaining 33 transcripts were subjected to hierarchical clustering applying Pearson correlation (average linkage distance) in the MultiExperiment Viewer (MeV v.4.8) (Figure 72). The results were analysed and a set of VIGS candidates defined.

Figure 72 Hierarchical clustering of 33 transcriptome contigs related to *SGD*

Expression values for 33 *C. roseus* transcriptome contigs. Tissues are A (flower), B (sterile seedling control), C (sterile seedling plus methyl jasmonate for 5 days), D (sterile seedling plus methyl jasmonate for 12 days), E (mature leaf), F (immature leaf), G (stem) and H (root). Orange arrow highlights candidate C45 and blue arrow the two partial *SGD*.

As described above, of the 90 contigs that were found to be co-localised with one of the 25 known pathway genes in the older version of the *C. roseus* genome, candidate C18 (Figure 72, CRO_011779) was identified based on its physical closeness to *SGD* and its annotation as cytochrome P450 enzyme. C18 was also observed on the *SGD* contig of the Mate pair enhanced assembly. Further candidates chosen based on a combination of their annotation, actual physical distance to *SGD* and/or their co-expression with *SGD* were C2 (CRO_027093), C6 (CRO_027095), C7 (CRO_027094), C8 (CRO_003636) and C9 (CRO_021613) (Figure 72), while C45, selected for co-localisation with SGD, had already been chosen as part of a list of co-regulated genes.

VIGS experiments were conducted. However, only silencing C18 (Figure 73) and C45 (Figure 74) led to significant changes.

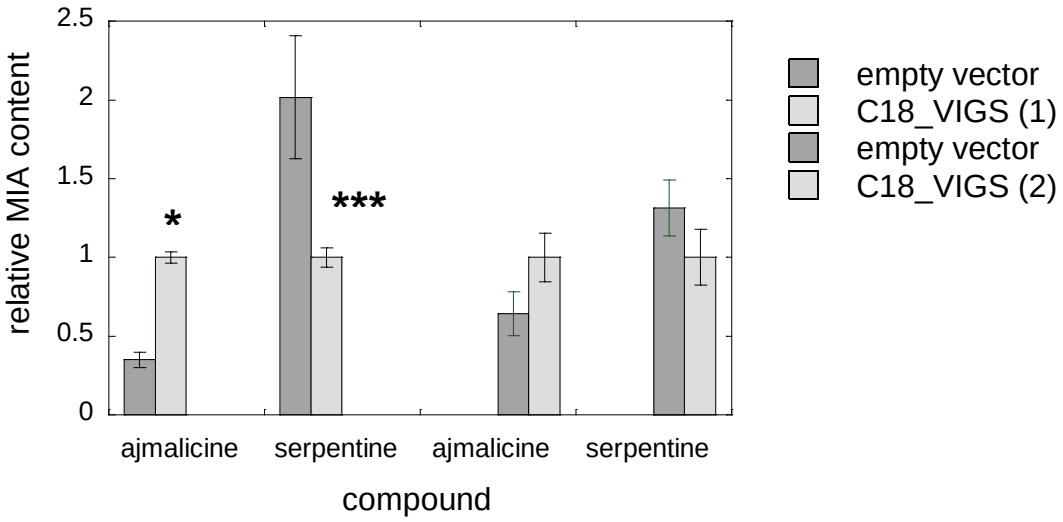


Figure 73 C18 silencing effect on leaf metabolites

Normalised peak area of compounds ajmalicine and serpentine of plants of two independent (n=5 and n=8) C18 silencing experiments in comparison to EV control plants. Error bars represent ± SEM. P-value was calculated using Student's *t* test (*p<0.05, ***p<0.00001).

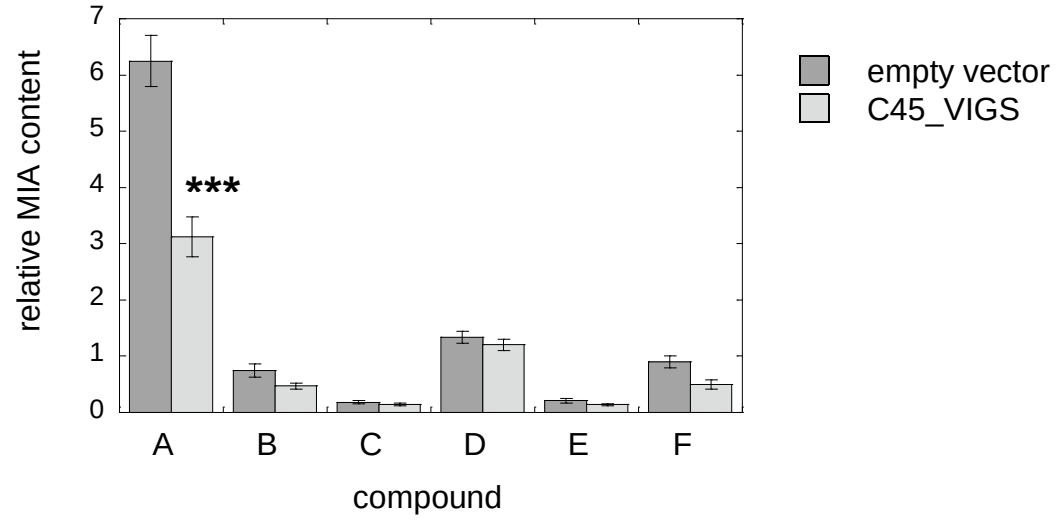
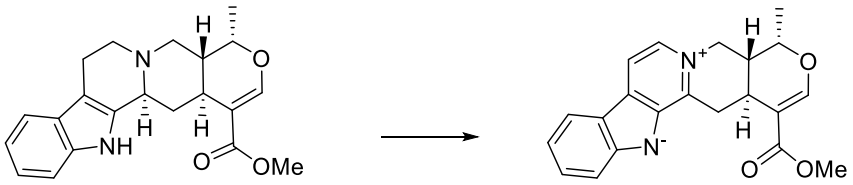


Figure 74 C45 silencing effect on leaf metabolites

Normalised peak area of (n=6) plants infected with C45 silencing construct in comparison to EV control plants for compounds: A (catharanthine), B (tabersonine), C (strictosidine), D (vindorosine), E (vindoline), F (secologanin). Error bars represent ± SEM. P-value was calculated using Student's *t* test (***p<0.0005).

C18

The first silencing of C18 displayed two major changes, the significant increase of ajmalice and the significant decrease of serpentine. However in a second experiment, although the same general trend in the data is visible, the data is not significant (Figure 73). No other metabolite in influenced significantly. Ajmalicine is the precursor of serpentine in *C. roseus* biosythesis and the enzyme/s catalysing this reaction are unknown [18]. However, this is an oxidatve transformation and could be catalysed by a P450.



ajmalicine serpentine

Figure 75 Reaction ajmalicine to serpentine

C18 is only expressed at relatively low levels (Figure 51). A BLASTn search against the *C. roseus* transcriptome [68] suggests that the observed change is not caused by cross-silencing. The biochemical function of this gene will be subjected to further analysis in future studies.

C45

One of the 33 *SGD* co-localised contigs (Figure 72), candidate *C45,* had also been identified as co-expresssion analysis of transcriptomic data [65]. This makes *C45* the only candidate that has been discovered twice, via co-expression analysis and as a co-localised gene with a known pathway gene. *SGD* contains two contigs, CRO_019099 and CRO_027160 that are both annotated as alpha/beta-Hydrolases superfamily protein (Figure 72). C45 can be associated with both of these contigs as closer investigation revealed that the only difference between both contigs is that CRO_019099 is about 200 bp longer than CRO_027160 while the total 948 bp of CRO_027160 are 100 % identical to CRO_019099. Both contigs are immediately adjacent to each other on a single genomic scaffold of the *C. roseus* mate pair enhanced genome assembly (data not shown).

VIGS experiments with this candidate reproducibly led to an approximately 50% reduction of catharanthine production. Unfortunately a BLASTn search of the 307 bp long section of this

gene used for silencing revealed that the *C. roseus* transcriptome contains several contigs that share sufficient sequence similarity to result in cross silencing. Future efforts will focus on identifying the enzymes that generate the putative substrate for this hydrolase (stemmadinine), which will then allow biochemical assay with each of the highly similar hydrolase genes for activity.

5.2.4 VIGS experiments of candidates resulting from co-expression analysis

Transcriptomic data [65] was analysed for gene co-expression in a search for a missing step in vindoline biosynthesis. In the course of this thesis 11 candidates displaying a strong co-expression with known vindoline biosynthesis genes were selected and tested using VIGS, including the newly discovered *T3O* gene described in Chapter 3 [146] (Table 28).

T3O must work in partnership with a reductase (Figure 33). The discovery of this gene, *T3R,* was reported during the course of this thesis [48]. This gene was present in our list of candidates that co-expressed with vindoline biosynthetic enzymes, so we also silenced the *T3R* gene. The published VIGS results [48] were compared to the *T3R* VIGS results obtained here. The published VIGS results for the *T3R* gene reports information on the levels of four compounds [48]. No change in vindoline concentration is observed; secondly, a compound with a mass of *m/z* 398.22, corresponding to desacetoxyvindoline (the T3R product) is reduced; thirdly, the alkaloid catharanthine is increased; finally three new compound peaks with a mass of *m/z* 383, corresponding to the T3R substrate, were detected. No values for the significance of those changes were provided in the published report.

The data obtained for *T3R* silencing in the scope of this thesis are comparable. No significant decrease in either vindoline or vindorosine can be detected in *T3R* silenced tissue (Figure 76). Additionally neither the levels of tabersonine or 16-methoxytabersonine are changed (Figure 77).

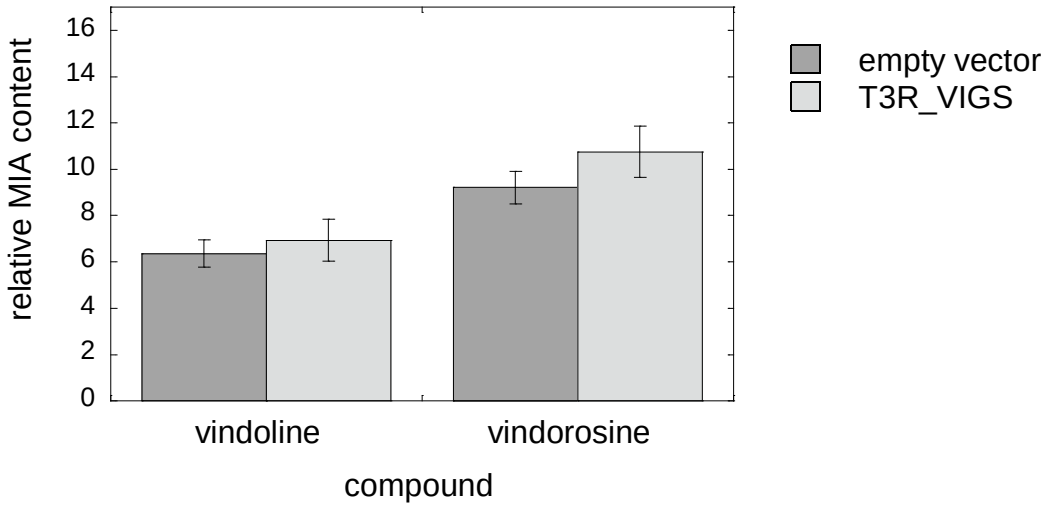


Figure 76 *T3R* silencing effect on vindoline and vindorosine

Relative MIA content as mean peak area for vindoline and vindorosine in *T3R* silenced plants (n=8) in comparison to EV control plants (n=6). Error bars represent mean ± SEM. P-value was calculated using Student's *t* test. No significant differences.

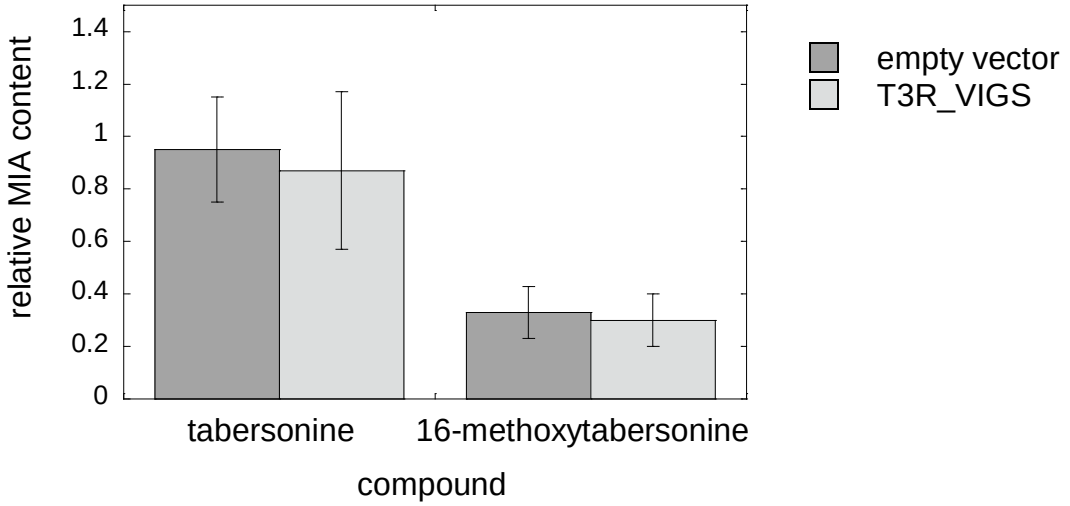


Figure 77 *T3R* silencing effect on tabersonine and 16-methoxytabersonine

Relative MIA content as mean peak area for vindoline and vindorosine in *T3R* silenced plants (n=8) in comparison to EV control plants (n=6). Error bars represent mean ± SEM. P-value was calculated using Student's *t* test. No significant differences.

Applying the software for automated peak extraction on the metabolic data of eight samples for *T3R* silencing and six samples of EV control yielded peak areas for 791 different peaks. After data normalisation a calculation of the t-test revealed that 15 of the extracted peaks were significantly changed (p-value < 0.01) between *T3R* silenced and EV control tissues (Table 31). However applying the Benjamini-Hochberg method for False Discovery Rate (FDR) to correct the p-value for the occurrence of false positives [229] leads to none of the observed changes being significant at all in this dataset. This work highlights how a VIGS screen can lead to false negatives.

Table 31 Significant changed peaks of *T3R* silenced tissues prior to FDR correction

Mean peak area for *T3R* silenced in comparison to EV control tissues, elution time of peak and accurate mass of peak. Data is sorted by p-value starting with the lowest. In bold are masses that would correspond to the substrates of T3R in vindoline (*m/z* 383) or vindorosine (*m/z* 353) biosynthesis respective.

m/z	elution time	average T3R	average EV	p-value
353.1849	2.323	160168	12014	0.00020
383.1971	2.66	79524	0	0.00162
384.1962	2.35	34634	0	0.00174
260.0718	2.339	24899	17202	0.00209
354.1918	2.558	4523	0	0.00458
383.1834	2.331	137652	5632	0.00463
506.223	4.628	9697	12418	0.00517
383.1899	3.487	7914	2334	0.00615
427.0987	2.331	41886	32915	0.00714
336.1735	1.411	6292	2649	0.00814
195.0648	2.327	9967	3686	0.00837
443.2171	4.722	6658	1199	0.00986
512.2015	3.197	4216	882	0.01003

5.2.5 Infiltration on three weeks old Catharanthus roseus seedlings

VIGS in *C. roseus* can be optimised, for example in a recently report using a biolistic-mediated method by which the virus encoding plasmids are directly transferred into the plant tissue and *Agrobacterium*-mediated inoculation can be circumvented [94]. Another alternative to pinching the stem is whole seedling infiltration instead of classical pinching. This could lead to silencing of root tissue, a part of the plant that is not silenced with the pinching method [93].

The currently employed VIGS system has two major limitations. First, no silencing of tissue below the pinching site can be achieved, therefore root or stem specific genes and metabolites cannot be investigated using gene silencing. Second, the current method requires eight week old plants, which makes the VIGS experiment time consuming, inflexible and low-throughput. To address these issues, we attempted to develop a VIGS method in which the entire plant would be affected by the virus and also would be considerably less time consuming. Whole plant infiltration as well as infiltration of seedlings and cuttings has proven successful for silencing in other systems [230] and was attempted for *C. roseus* here.

5.2.5.1 Magnesium chelatase silencing of whole seedlings of Catharanthus roseus

After applying the syringe press method, in which whole three week old *C. roseus* seedlings are fully submerged with *A. tumefaciens* harbouring the pTRVu_*ChlH* plasmid and pressure is applied to effectively infiltrate the whole seedling, the typical yellowing that results from successful silencing of the *photoporphyrin IX magnesium chelatase subunit H* gene occurred as early as one week after infiltration (Figure 78). Of initially 17 infiltrated plants only 11 survived, and of those 9 showed yellowing after one week. This was repeated twice with similar results.

Figure 78 Successful silencing with syringe-press method in *C. roseus* 3 weeks old seedlings

5.2.5.2 Metabolic phenotype of silenced whole seedlings of Catharanthus roseus

As an initial test and proof of concept, plants were infiltrated with available VIGS vectors harbouring MIA genes. To investigate the root specific changes the samples were separated into two fractions, leaf and root, that were investigated separately. Although for each VIGS construct 6 -8 individual seedlings were harvested those samples were mixed resulting in a mixed leaf fraction and mixed root fraction for each VIGS vector and for the empty vector control.

First the *T16H2* plasmid was used for infiltration. Classical *T16H2* silencing, using the standard VIGS protocol as conducted in the work presented in Chapter 2, leads to a significant reduction of vindoline and accumulation of the substrate tabersonine and the alternative shunt products vindorosine. Similar results can be observed for the leaf fraction of the whole seedling *T16H2* silencing experiment in comparison to the leaf fraction infiltrated with an empty vector (Figure 79).

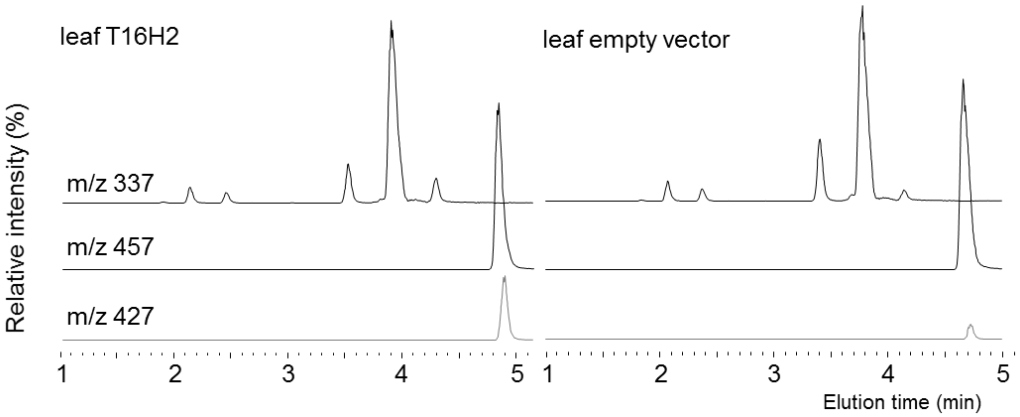


Figure 79 LC-MS profile of *T16H2* VIGS and empty vector control seedling leaf tissue
Extracted ion chromatogram for mixed sample (n=6) leaf tissue from *T16H2* silenced whole seedling with increased peak area for tabersonine (*m/z* 337 at elution time 4.35) and vindorosine (*m/z* 427) compared to empty vector treated tissue.

The obtained results are similar to results from the classic VIGS method. *T16H2* silencing has expectedly no influence on stem or root tissue metabolite composition or content (data not shown) as these tissues do not express *T16H* and do not produce vindoline or vindorosine. Therefore it cannot conclusively be determine if the silencing was successful in those tissues.

The *SLS* VIGS construct was tested using this method. Successful silencing of *SLS* with the classical VIGS method leads to significant decrease of various downstream products but most importantly to the emergence of a peak (*m/z* 413) that is below the detection limit in EV control samples (Chapter 5.3.1., Figure 67). Although the results from this first initial test for whole seedling infiltration are comparable to previous VIGS results with the same VIGS vector the results for root tissue are less clear (Figure 80).

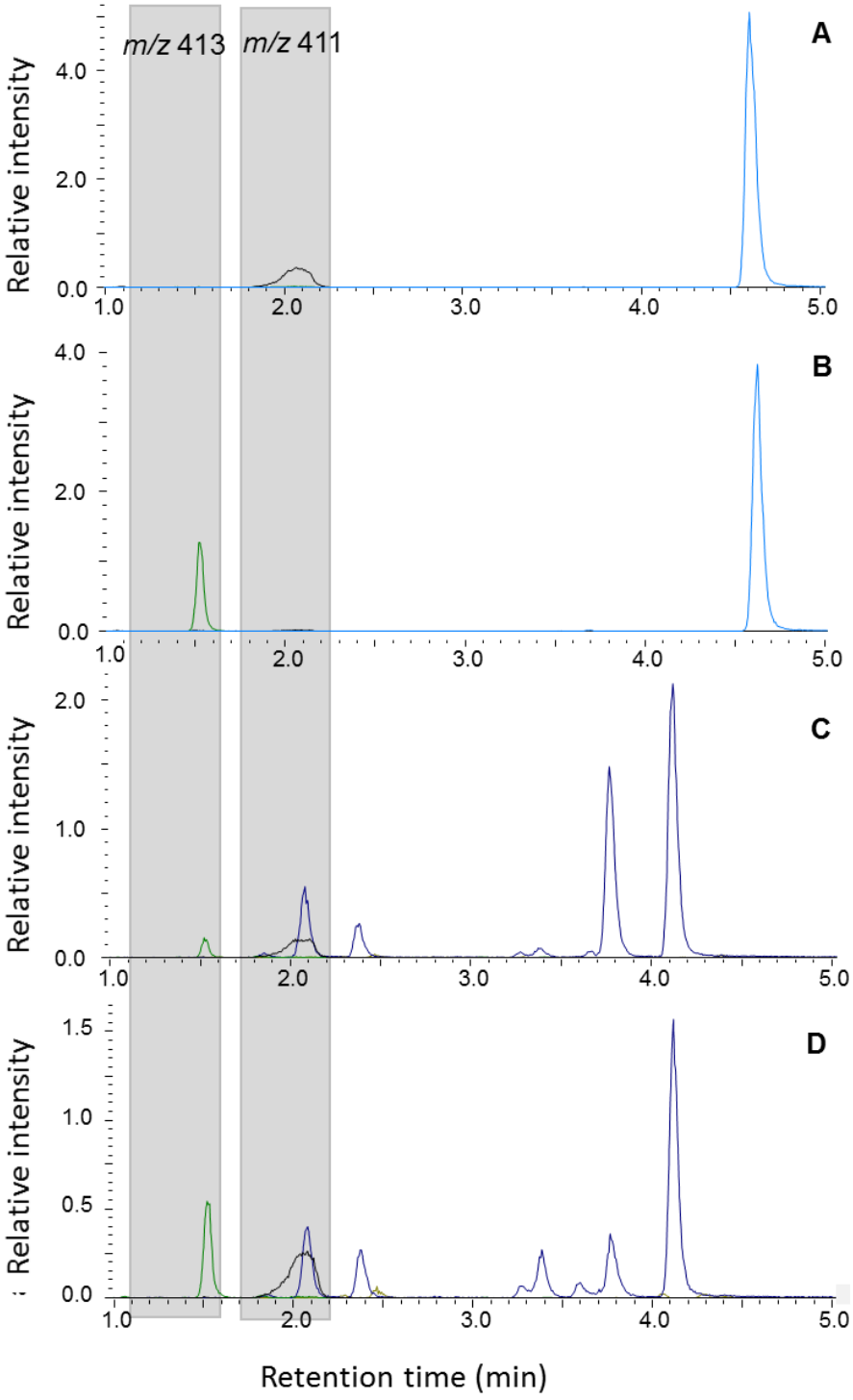


Figure 80 LC-MS profile of *SLS* VIGS and empty vector control seedling leaf and root tissue

Extracted ion chromatogram for mixed sample (n=6) leaf tissue from *SLS* silenced whole seedling and root tissue. **A:** Leaf tissue empty vector. No peak for loganin (*m/z* 413) is observed. Black peak is secologanin (*m/z* 411) in comparison to vindoline (light blue peak). **B:** Leaf tissue *SLS* silencing. New peak for loganin (*m/z* 413) is observed. Black peak is secologanin (*m/z* 411) in comparison to vindoline (light

blue peak). **C:** Root tissue empty vector. Small peak for loganin (*m/z* 413). Black peak is secologanin (*m/z* 411) in comparison to catharanthine as roots contain no vindoline (*m/z* 337 at elution time 4.35, dark blue peak). **D:** Root tissue *SLS* silencing. Larger peak for or loganin (*m/z* 413). Black peak is secologanin (*m/z* 411) in comparison to catharanthine as roots contain no vindoline (*m/z* 337 at elution time 4.35, dark blue peak).

Although a potential increase in loganin accumulation can be observed, the nature of the experiment, measuring a mixed sample of six individually silenced seedlings, allows no statistical analysis and so it cannot be determined if the observed larger peak for loganin is a result of an possibly increased variability of root tissue in general or indeed a silencing result.

In general there are less examples for successful silencing of roots then there are for other tissues reported in the literature. In *C. roseus* it was shown that the here reported syringe-press method for the infiltration of whole seedlings [93] led to accumulation of both pTRV1 and pTRV2 in root tissue, a prerequisite for virus assembly. However the only gene silenced is in this study is *phytoene desaturase* (*PDS*) a gene involved in carotenoid biosynthesis that, if silenced, causes sever albinism and is used as a visual marker for silencing success similar to *magnesium chelatase* [231], an approach that is however unsuitable for roots. Additionally the qRT PCR results presented in [93] do not provide a strong evidence of successful silencing in roots of *C. roseus*. This method has therefore to be tested further before it can be concluded if silencing of root specific genes is feasible as a tool to investigate the alkaloid production in *C. roseus* roots.

5.3 Conclusion

Since its establishment in 2011, VIGS in *C. roseus* has facilitated the discovery and investigation into physiological function of several alkaloid pathway genes. In 2012 it facilitated the discovery of *ISY* [30]. In 2013 the role of *T16H2* was investigated using VIGS [42]. VIGS played an essential role in the discovery of a ATP-binding cassette transporter implicated in the transport of the *C. roseus* alkaloid catharanthine [46] and the discovery of *7DLGT* [32] and *7DLH* [31,33]. In 2015 VIGS helped clarify the role of THAS [37] the first *C. roseus* gene ever to be characterised to catalyse a reaction using the deglycosylated strictosidine intermediate. Testing candidate genes with the VIGS technique led to the discovery of *T3R* [48] and *T3O* [48,146], which completed the search for vindoline biosynthesis enzymes after decades of research.

In this chapter we outline the efficiency and limitations of the VIGS technique by silencing the three previously identified MIA biosynthetic genes *SLS* [35], *TDC* [36] and *STR* [189], that had been biochemically validated, but not subjected to *in planta* silencing. This initial investigation would test the robustness of the established VIGS procedure and additionally give insight into the role of the three tested known alkaloid biosynthesis genes in young leaf of *C. roseus*. These experiments on the whole suggest that the VIGS method produces reproducible, accurate results, but also highlights that false negatives can occur in this experimental technique.

Additionally, this VIGS system was used in a general candidate screen. Based on the co-expression data and genome data, 28 candidates were silenced *in planta* using VIGS. The vast majority of these candidates yielded negative results when silenced. However, two genes emerged from this screen that produced metabolic changes upon silencing for which follow up experiments are being planned.

Many steps of alkaloid biosynthesis in *C. roseus* are co-regulated and this co-regulation is reflected in the co-ordinate expression of genes in available expression data. Moreover not just alkaloid biosynthesis is co-regulated, the production of its precursors is most likely controlled in a similar way. Tryptophan biosynthesis for example is strictly co-regulated in *A. thaliana* [140]. The rational for choosing a VIGS candidate is in many cases its co-expression with known alkaloid pathway genes. The observed co-expression in *C. roseus* alkaloid biosynthesis is caused largely by the coordinate upregulation of alkaloid biosynthesis as reaction to the elicitation with for example methyl jasmonate [232,233]. Therefore it is very possible that a chosen VIGS candidate gene, co-expressed with a known alkaloid biosynthesis genes, might be an essential part of the methyl jasmonate mediated defence and general stress response [234]. Silencing such a gene can have unpredictable results as signalling pathways [235] as well as individual secondary metabolite pathways [236] are linked. In consequence a VIGS result for such a gene could result in a phenotype with significant changes of alkaloid profile, even if the gene is not actually part of alkaloid biosynthesis. Therefore VIGS results must be interpreted with caution. In principle, the successful silencing of a single step in a multistep pathway results in the reduction of the product of this step and the accumulation of the substrate. Without either, the interpretation becomes far more difficult. Still, in conclusion, as VIGS candidates are most of the time highly expressed and methyl jasmonate inducible genes with an annotation that suggests involvement in metabolism of secondary metabolites, the high number of VIGS

results without any significant changes is rather encouraging, as it speaks for the robustness of the system.

While false positive results will possibly eventually be discovered during follow up experiments such as the heterologous expression, candidates that lead to false negative VIGS results are more in danger of simply being overlooked and not followed up. Over the course of testing 28 candidates we have observed two genes, *T3R* [48] and *STR* [189], for which the function has *in vitro* been characterised, that produces no or only few statistically significant changes in metabolite profile upon silencing. This shows some limitations of VIGS as a system for gene function screening. Nevertheless, despite these caveats, VIGS remains the most powerful method to date to rapidly screen the function of putative metabolic gene candidates.

Although heterologous expression of candidate genes in alternative hosts and testing their function by feeding various substrates is a valid method for gene discovery and characterisation, this approach lacks the explanatory power of a VIGS experiment, as a VIGS experiment is additionally able to demonstrate the *in planta* function of a gene. Currently, a single VIGS experiment requires at least a three month period. Efforts were made to perform VIGS using a new infiltration method that could shorten this time and also potentially silence a greater range of tissues compared to the standard pinching method. Testing the efficiency of this infiltration method using *ChlH* silencing as visual marker of silencing success in leaves led to 91% of plants showing yellowing. This is in accordance with previous finding for *C. roseus* [121]. Testing the known pathway genes *T16H* [41] and *SLS* [35] leads to the expected and previously observed phenotype (Chapter 2, Chapter 5.3.1.) in leaf tissue, with the results being less clear in root tissue. This method effectively cuts the experimental time frame in half from currently 11 weeks to 6 weeks with comparable results for leaf tissue. Of greater interest however is whether silencing of a root expressed gene is possible and would result in a metabolic phenotype and measurable expression reduction that would validate the involvement of this gene in alkaloid biosynthesis in *C. roseus*.

6 Discussion

Alkaloids are complex secondary metabolites that often display potent bioactivity. Their biosynthesis in plants typically proceeds via intricate multistep enzymatic pathways. According to the literature more than 130 different alkaloids of monoterpene origin are produced by the plant *C. roseus* [19]. All are derived from the central precursor molecule strictosidine that is formed as a condensation product of the terpenoid moiety secologanin and the indole moiety tryptamine, via a Pictet-Spengler condensation (Chapter 1, Figure 5) [18,24]. The first gene successfully identified for strictosidine biosynthesis in *C. roseus* was *TDC* (tryptophan decarboxylase) reported almost three decades ago in 1989 [36], *STR* (strictosidine synthase) was reported in 1992 [237], *SLS* (secologanin synthase) was reported in 2000 [35], followed by *G8H* (geraniol 8-hydroxylase) in 2001 [28] and *LAMT* (loganic acid methyl transferase) in 2008 [34]. The remaining five steps were published over the course of just two years in 2012 and 2013 [29–32] leading to the complete elucidation of strictosidine biosynthesis in *C. roseus* and the knowledge being used to engineer this early part of alkaloid biosynthesis in *S. cerevisiae* [25] and *N. benthamiana* [29]. The great acceleration in gene discovery in *C. roseus* was aided by two factors: the increased availability of comprehensive transcriptomic data and the observation that the genes of this pathway are tightly co-expressed [49].

In *C. roseus* as well as other MIA producing plants strictosidine is deglycosylated by strictosidine β-D-glycosidase (SGD) and rearranges to 4,21-dehydrogeissoschizine that forms the structural basis for the diverse range of alkaloid scaffolds (Chapter 1, Figure 6). In *C. roseus* the highly reactive 4,21-dehydrogeissoschizine intermediate is transformed by an unknown number of subsequent reactions that lead to the molecular rearrangements necessary for the formation of the three distinct molecular backbones or skeletons, representing the three different classes of *C. roseus* alkaloids, namely the iboga, the corynanthe and the aspidosperma alkaloids. At the start of this work none of the enzymes involved in this complex transformation was known.

Pathway elucidation of alkaloid production in *C. roseus* has been a focus of attention ever since the discovery in the 1960s of the anti-cancer properties of its compounds vinblastine and vincristine [18]. These bisindole alkaloids form as a result of the coupling of the two main *C. roseus* monoterpene alkaloids vindoline (aspidosperma alkaloid) and catharanthine (iboga alkaloid). While catharanthine is a direct product of the reactions following the deglycosylation of strictosidine, vindoline originates from tabersonine and is the product of a set of seven

further enzymatic reactions, referred to as vindoline biosynthesis. Of those seven steps only five were known at the start of this thesis (Chapter 1, Figure 7).

The first of two key aims of this thesis was to close this gap in vindoline biosynthesis by following the successful strategy applied in strictosidine biosynthesis gene discovery: the mining of transcriptomic data for co-expressed genes. The identified gene candidates were tested using VIGS. Biochemical analysis confirmed the activity observed in planta.

The second aim of this thesis was to investigate gene clustering in *C. roseus*. It is a recent observation that in some plant secondary metabolic pathways the genes are located in close proximity on the genome. As a result of the increasing availability of plant genome information, this has now been reported for a variety of species and different classes of specialised metabolites [174]. The successful identification of gene clustering can enhance pathway discovery [177,198,238,239]. This is of particular interest for genes that might be difficult to discover with the co-expression approach as not all genes of a pathway are necessarily closely co-regulated. To investigate if any gene clustering in *C. roseus* can be observed, a draft genome sequence was obtained, as well as several BACs containing known genes of key steps in alkaloid biosynthesis in *C. roseus*. As with the co-expression approach identified candidates were predominantly tested using VIGS.

An important consideration is that most studies that successfully applied co-expression analysis for gene discovery in *C. roseus* [31,32,79] had previously determined the most probable enzyme class of the missing step and therefore substantially limited otherwise lengthy candidate lists. However, limiting the search to a specific enzyme class can also be obstructive. The highly complex chemistry of plant secondary metabolite pathways has often evolved in only a small phylogenetic space, with enzymes catalysing unprecedented reactions, making an accurate judgment of the probable enzyme class impossible. The discovery of *C. roseus* iridoid synthase [30], representing a new class of enzymes, from a transcript that displayed highest sequence similarity to progesterone-5β- reductase (P5bR), was aided by the fact that this transcript was very tightly co-expressed with known genes and that chemical understanding predicted a NADPH requiring enzyme for this step. Gene clustering, if observed, can therefore aid further gene discoveries by combining information of gene co-expression with gene clustering/ gene co-localisation.

6.1 Gene silencing and gene clustering

The VIGS system is essential to the work presented in this thesis. No reliable protocol for the stable genetic transformation of *C. roseus* exists. Gene candidates for alkaloid biosynthesis have successfully been tested for example in alternative systems such as *N. benthamiana* [29] and yeast [79]. However such alternative systems always rely on the availability of the correct substrates, a difficult task as many intermediates of *C. roseus* are not commercially available or structurally so complex and/ or unstable that chemical synthesis is not always an option. Ultimately for the validation of *in planta* function of alkaloid pathway genes in *C. roseus* only VIGS is currently available. VIGS experiments are detailed in Chapter 2, Chapter 3 and Chapter 5.

Chapter 2 describes the successful *in planta* functional validation of a tabersonine 16-hydroxylase (*T16H*). In 1999 a publication reported a cytochrome P450 (*T16H1*) that had been cloned from a light induced *C. roseus* cell suspension culture and was characterised to catalyse the first step in vindoline biosynthesis, the 16-hydroxylation of tabersonine [41]. In 2012 our collaborators (Vincent Courdavault, University Tours, France) had cloned a second version of this gene (*T16H2*) from *C. roseus* leaf tissue and *in vitro* characterisation revealed it had identical function to the previously discovered gene. Successful targeted gene silencing of this gene in *C. roseus* leaf tissue using VIGS was confirmed with q RT PCR and led repeatedly to a significant decrease of vindoline, the end product of the vindoline pathway. Close investigation revealed that reduction of vindoline levels was accompanied with the coinciding increase of vindorosine (des-methoxy-vindoline). Untargeted peak extraction of the obtained mass spectrometry data revealed several peaks with exact masses corresponding to intermediates of vindoline or vindorosine biosynthesis were significantly decreased or increased respectively, further strengthening the redirection of flux to the un-methoxylated product vindorosine as a result of the targeted gene silencing and therewith ultimately validating the essential function of *T16H* in *C. roseus* leaves. This highlights the explanatory power of VIGS as a tool for functional characterisation of known enzymes.

Two independent approaches, remapping transcriptomic raw reads to the two *T16H* open reading frames, conducted as part of this thesis, and q RT PCR of both *T16H* versions on different *C. roseus* tissues in the lab of Vincent Courdavault came to the same conclusion. Both versions of this enzyme are active in *C. roseus* but the predominance of one paralog is tissue dependent [42]. This observation has many implications especially for further pathway discovery.

Notably, the two functional and co-ordinately regulated versions of *T16H* are unsuitable candidates for co-expression analysis as none of the individual *T16H* expression profiles matches the expression profiles of downstream vindoline biosynthetic genes.

It can further be speculated that *T16H* is potentially not the only enzyme in the pathway for which functionally identical paralogs exist. Indeed this was later reported additionally for *ISY* [181] and *SLS* [68,213] and the genome sequence data presented in Chapter 5 shown for both *T16H* and *ISY* that the functional paralogs belong to recently duplicated gene pairs [68].

In Chapter 3 VIGS was successfully employed to screen a number of candidates for missing steps in vindoline biosynthesis, leading to the discovery of tabersonine 3-oxygenase (T3O) [146]. The selection of candidates was based two major assumptions: firstly, that the missing genes would co-express with known vindoline pathway genes (with the exception of *T16H*), and secondly that one of the genes would be annotated as a cytochrome P450 enzyme, since the proposed chemistry entailed formation of an epoxide, a reaction known to be catalysed by P450s (Chapter 3, Figure 24). The successful candidate could unambiguously be identified using VIGS, since silencing in this case significantly reduced both products of this pathway, vindoline and vindorosine. VIGS also resulted in the dramatic accumulation of a peak with a mass corresponding to the substrate of T3O. The identity of this peak could be further verified by consumption in assays with heterologously expressed *T3O* from yeast. The structural characterisation of the T3O reaction product carried out by Dr. Nathaniel Sherden concludes that in the absence of a partnering reductase the unstable intermediate does first form an epoxide, but then readily rearranges to a product with a new scaffold structurally similar to the alkaloid vincamine from the plant *V. major*, a close *C. roseus* relative. This is in accordance with previous publications that report that 2-imine,3-alcohols on the tabersonine framework are able to undergo acid catalysed rearrangements [148–151] and leads us to speculate on the possibly of a similar biosynthetic pathway for vincamine in *V. major*.

While Chapter 2 and 3 highlight successful examples of how VIGS can be used for the validation of *in planta* function of known pathway genes as well as for a screening tool for unknown genes, Chapter 5 shows limitations to this system. Here the targeted silencing of three further known pathway genes, the previously identified MIA biosynthetic genes *SLS* [35], *TDC* [36] and *STR* [189], clearly shows that measurable downregulation of expression is not necessarily reflected in a strong and unambiguously interpretable VIGS phenotype. When both *SLS* and *TDC* are silenced, the respective substrates, loganin and secologanin, accumulate while

downstream alkaloids are significantly reduced. The result for *SLS* is especially interesting as it shows that gene silencing results in silencing of all four functionally redundant genes due to their high sequence identity. However, 60% reduction of *STR* expression leads to only minimal changes in alkaloid content. These results would have possibly been overlooked without prior knowledge of the *in vitro* confirmed function of this particular gene. In three subsequent experiments no significant accumulation of an *STR* substrate peak could be observed although the accumulation of secologanin, one of the two precursors for the *STR* catalysed reaction, has been observed previously in the *TDC* silencing experiments, showing that in principle secologanin accumulation in plant tissue is possible.

A second example for a weak and therefore difficult to interpret VIGS phenotype is observed with the reductase *T3R* in vindoline biosynthesis. The VIGS phenotype observed allows for only a tentative interpretation of the *in planta* function of this gene only in combination with the *in vitro* results for this enzyme, which are required in this case to clearly demonstrate its involvement in vindoline biosynthesis [48]. Interestingly our VIGS results match the results published for this gene by another group [48] despite the fact that they use a different *C. roseus* variety, which points to the general reproducibility of the VIGS system.

Why *STR* or *T3R* silencing do not lead to the expected metabolic changes is not easy to answer. For both genes the reduction of expression was measured to be around 60%. One possibility is that the remaining levels of enzyme are efficient enough to carry out the reactions without an alteration of the metabolite flux. The existence of additional enzymes that can carry out the same reaction seems unlikely in these two cases. The reaction catalysed by *STR* is highly specific and while STR homologs are found in the plant, they are known to be inactive (O'Connor, personal communication). Obtaining knockouts for *STR*, perhaps in cell suspension or hairy root culture which are amenable to stable transformation [240], may clarify these points. This option however only applies to genes of the upstream part of alkaloid production as the metabolites produced in roots are significantly different in the downstream pathway.

In the course of this thesis, a variety of P450 enzymes were functionally assayed by VIGS with no measurable positive results. While this could be due to the negative results sometimes exhibited in this silencing system, it must be noted that *C. roseus* produces far more than the major alkaloids such as catharanthine or tabersonine for which standards exist. Some publications state up to 130 different alkaloids [19] and new publications continuously report the isolation of more new compounds isolated from plants such as *C. roseus* with better extraction

methods applying mass spectrometry with higher sensitivity. Alkaloids, like many plant secondary metabolites, play essential roles in interaction with the environment [11] for example to attract pollinators [9] or most importantly as a means by which plants defend themselves against microorganisms and herbivores[10]. From an ecological point of view it appears logical not to "limit" this arsenal to a few major compounds as this holds the risk of potential predators overcoming a single defence. Therefore the possibility exists that some of the unsuccessfully tested genes might conduct interesting chemistry yet the metabolite profiling used in the VIGS system is simply not sensitive enough to detect changes in these minor products. However, using VIGS to screen for enzymes that conduct novel chemistry is an enormous challenge, as it requires the prediction of the structure of the substrates and products from the metabolomic analysis of VIGS tissue. Without a clear, statistically significant emergence of a new peak and/or reduction of a product, these results are not interpretable in the absence of other experimental evidence.

Chapter 4 reports the analysis of a newly established resource, the *C. roseus* draft genome available as a searchable database (http://medicinalplantgenomics.msu.edu/). Together with six BAC sequences this data was used to investigate the genomic context of *C. roseus* biosynthesis with regards to gene clustering. Most importantly this analysis provides a list of 90 transcripts that are in close proximity to the 25 known alkaloid pathway genes, a number that is challenging to be tested within the VIGS system. Therefore a combination of functional annotation and co-expression profiles were employed to pick initial candidates from this list. Despite all reported limitations and constraints, two candidates show tentative results. Preliminary VIGS results for a P450 candidate suggest a potential involvement of this gene in the biosynthesis of serpentine. A second candidate, selected primarily from the co-expression analysis conducted in Chapter 3 but also because of co-localisation with *SGD,* when subjected to VIGS, repeatedly results in catharanthine reduction. These candidates are the focus of further experiments by other group members.

The analysis of gene clustering in the draft *C. roseus* genome revealed several small clusters consisting of three genes. *TDC* and *STR* are co-localised with a MATE transport protein and all three genes display very tight co-expression. The co-localisation of *SLS* with a genes annotated as a sinapyl-alcohol dehydrogenase has led to the discovery of *THAS*, the first enzyme reported to be involved in the formation of the different typical alkaloid scaffolds after the deglycosylation of strictosidine [37]. The pair of *T16H* paralogs, which is a gene duplication, is

located next to *16OMT* representing the first and second steps in vindoline biosynthesis. The duplication of a gene foremost removes the evolutionary constraint that is associated with the need to retain the original gene function. The common fate of duplicated gene pairs is very often the emergence of deleterious mutations leading to non-functionalisation of one of the copies[241]. It is also possible that both gene copies are maintained with their redundant function. Over time, changes in the regulatory or coding sequence can lead to functional divergence such as sub-functionalisation, where the original gene function is partitioned into the two copies, as is observed for *T16H* [42]. Alternatively, the occurrence of an adaptive mutation in one of the duplicated genes that can lead to novel functions, so called neo-functionalisation. It has to be noted that a strong driver of enzyme recruitment and functional divergence is a change in transcription and/or translational level of the enzyme, as such changes might lead to the physiological relevance of a previous latent existing enzyme function [242].

Although some common features do exist between plant gene clusters identified to date, the relatively small number of clusters precludes development of a global definition and general rules for prediction of when clusters might be observed. Predictions as to how many more clusters might be found in plants is difficult. Much has been speculated about the significance and function of clustering. Several of the today known plant secondary gene clusters are responsible for the production of metabolites that have shown to confer resistance and therefore a selective advantage [111,116]. It has been argued that organisation of biosynthetic pathways in clusters is driven by both the selective advantage and the beneficial role the pathway [101]. However the basis of gene clustering might be rooted in many different evolutionary reasons.

Extensive analyses in *Arabidopsis thaliana* revealed that in general genes from a specific pathway are more likely co-expressed then genes that are not part the same pathway, but also that only a small number of genes have many co-expression partners while most genes have less than 10 partners [130]. Interestingly, in this study, genes that have fewer coexpression partners are often multi-copy genes, suggestive again of the evolutionary aspects that gene duplication and subsequent release of regulatory constraint play in the expansion of existing networks [130]. Such an accumulation of duplicated genes can be observed for alkaloid biosynthetic enzymes in the *C. roseus* draft genome. For some of these like ISY an *in vitro* function redundancy has been shown, though VIGS failed to provide evidence for an

involvement of one of the paralogs in planta[181] while the other paralog is clearly implicated in alkaloid biosynthesis [30]. Others, like *T16H1* and *T16H2* could be examples of "evolution in progress", where the differential regulation might, in the future, evolve into the development of an alternative pathway chemistry or the loss of the vindoline biosynthesis.

Over the last decade many newly established plant biology techniques have been applied to *C. roseus* to decipher alkaloid biosynthesis. Some of these techniques and tools have played an active role in gene discovery while others supplied candidate lists for further research. Examples include EST sequencing of specific alkaloid biosynthesis relevant tissues such as the leaf epidermis [50], proteome mining of *C. roseus* hairy root cultures [233] or suspension cultures [243], establishment of VIGS for *C. roseus* [121], the creation of larger transcriptomic resources facilitating co-expression analysis over multiple tissues and/or conditions [65], or for various members of the *Catharanthus* genus [70]. Amongst those, the combination of co-expression analysis with knowledge of gene location have proven to be the most successful approach for identification of gene candidates and along with experimental validation by VIGS have led together to the successful engineering of the early part of *C. roseus* in *N. benthamiana* in 2014 [29] and S. *cerevisiae* in 2015 [25] and that of vindoline biosynthesis in *S. cerevisiae* in 2015 [48].

6.2 Future directions

The availability of transcriptomic and genomic data is crucial and has allowed, over the past decade, a rapid acceleration of discoveries in plant metabolism pathways [244]. Automated genome mining for secondary metabolite clusters in plants, similar to existing solutions for discovery of microbial secondary metabolite gene clusters [245,246], may facilitate the more rapid application of the increasingly number of plant genomes that are publically available. With the establishment of more genomic data this might be feasible for plant gene clusters on a larger scale and for a more diverse range of gene clusters [176,177,238,239]. This might enable the discovery of more of the shared features of these gene clusters that can in turn facilitate development of new strategies to identify these clusters.

Experimental techniques for in planta screening, such as large scale VIGS experiments and the subsequently necessary metabolomics, are still a limiting factor and efforts have to be undertaken to streamline these methods as it is currently too time and labour intensive for high-throughput functional characterisation.

6.3 Outlook

Undoubtedly, the vast majority of plant specialised metabolism remains an untapped resource [247]. Sequencing is no longer a limiting factor with steadily declining costs, increasing precision and fast turnaround times [248]. Systems biology promises quick advances in gene and pathway discovery in plant specialised metabolism by combining –omics techniques such as genomics with metabolomics and transcriptomics [249]. It has already has been shown that the information provided with the *C. roseus* draft genome accelerated discoveries related to secondary metabolism in plants, and it for example also facilitated the research into the transcriptional regulation of strictosidine biosynthesis [224]. Future genome sequences of related species will shed light on the possible conserved evolution of the observed gene cluster. Better sequencing and improved assembly might eventually be able to answer the question if and how much more gene clustering can be observed in *C. roseus*.

7 Materials and Methods

Unless otherwise stated, all chemicals were purchased from Sigma-Aldrich and all restriction enzymes from New England BioLabs.

7.1 Plant material

All plant material used in this thesis was obtained from *Catharanthus roseus* variety "SunStorm Apricot" (Syngenta) and all VIGS experiments were conducted on this variety. The variety "Little Bright Eyes" was occasionally used when noted to compare alkaloid content between these two cultivars. All plants were grown in a walk in growth chamber at 25oC under 12 h days using the John Innes compost mix No. 2 (peat based).

7.2 General methods for molecular biology

7.2.1 Primers and sequencing

All primers were supplied by Integrated DNA Technologies (IDT), except for primers used for USER cloning that were purchased from Sigma-Aldrich. Unless otherwise stated, all primers were designed using the OligoAnalyser 3.1 (https://www.idtdna.com/calc/analyzer) and optimised for a melting temperature of 55^{o}C to 58^{o}C.

Sequencing was carried out by SourceBioscience (Cambridge) using plasmid DNA or purified PCR product as template.

7.2.2 PCR and DNA purification

Generally, each PCR reaction contained 1 µM of each primer, 1 mM dNTP mix, a suitable polymerase, template, and the manufacturers supplied buffer, as well as up to 3% DMSO and up to 3 mM $MgSO_4$. The total PCR volume was adjusted to 20, 25 or 50 µl with water depending on the required amount of PCR product. Template amount was variable depending on the nature of the template. The amount of template cDNA was usually 1 µg and never more than 10% of total PCR volume. Plasmids as templates were used at 1 pg to 1 ng per reaction. For amplification from genomic DNA, 1 ng to 1 µg was used. For colony PCR, the template

consisted of 1 µl fresh bacterial or yeast culture or 1 µl of a single colony picked with a sterile toothpick and then diluted into 10 µl of water.

For construction of VIGS plasmids, the X7 polymerase produced in house by Dr. Fernando Geu-Flores following a published protocol [250]. This polymerase is able to incorporate the Uracil Specific Excision Reagent (USER) compatible primers in combination with 5x HF Buffer (NEB, cat. No. B0518S) and up to 3% DMSO.

Colony PCR to confirm cloning success was generally performed with Taq polymerase (GoTaq, Promega, cat. No. M5001), while standard amplification from cDNA or plasmid template was conducted with Pfu Ultra II polymerase (Agilent, cat. No. 600670) or KOD Hot Start polymerase (MerckMillipore, cat. No. 71086) and the PCR reaction mix adjusted to manufacturer's instructions.

All PCR reactions were carried out on a Veriti 96 Well Thermal Cycler (Applied Biosystems). For a general PCR, the stepwise gradient started with a melting period of 95°C for 2 minutes, annealing period of 55°C for 20 seconds and extension period of 72°C. The extension time was adjusted to the expected length of PCR product and the speed of the polymerase used. Standard PCR reactions were done with 30 cycles, colony PCR with up to 33 cycles. All PCR reactions were carried out with a final extension of 10 minutes at 72°C.

For evaluation of the obtained band size, 2 µl of the PCR product were run at 100 V on a 1% (w/v) agarose gel using GelRed™ (Biotium cat. No. 41010) to stain the DNA. Gels were run in TAE buffer (40 mM Tris, 20 mM acetic acid, and 1 mM EDTA). Bands were visualised using the SYNgene bioimager. For gel extraction, the entire PCR reaction was run on a gel, excised and purified using a Nucleospin Gel and PCR clean-up kit (Machery-Nagel cat. No. 740609.50) in accordance with the manufacturer's instructions. Normal PCR purification was done with the same kit. For the final elution from the column, 40 µl nuclease free H_2O was used and the DNA was spectrophotometrically quantified on a nanodrop (NanoDrop, ND-1000).

7.2.3 cDNA synthesis

cDNA from *C. roseus* variety "SunStorm Apricot" used as PCR template was synthesised using the SuperScript® III First-Strand Synthesis System (cat. No. 18080-051). Unless otherwise stated the procedure followed the manufacturer's instructions using total RNA as template,

Oligo(dT)$_{20}$ as primer and the Superscript III reverse transcriptase (cat. No. 18080-044). The total RNA was obtained following the procedure described in section 7.3 below. Typically, RNA from young leaf tissue was used. However for cloning of specific genes that are not expressed in young leaves, other tissues were used for RNA extraction such as root tissue, flower tissue and stem as well as methyl jasmonate induced seedlings. *C. roseus* seedlings were grown in petri dishes on water soaked filter paper. The seeds were surface sterilised before germination using 70% ethanol for 30 seconds and subsequently soaked for 2 minutes in 5% natriumhypochlorite with 2% Tween 20 as detergent, while vortexed repeatedly. The seeds were rinsed three times with sterile water. Seeds were kept in the dark at room temperature and transferred to the growth chamber after seven days (germination had occurred), where they were kept for further two weeks. Methyl jasmonate was applied to the seedlings by spraying with 500 µM solution and induced seedlings were harvested 1 day after application. All tissue samples for RNA extraction were immediately snap frozen in liquid nitrogen after collection and kept at -80^{o}C till use.

7.2.4 Culture conditions and glycerol stocks

E. coli cultures were cultured at 37^{o}C, *A. tumefaciens* and *S. cerevisiae* were kept at 28^{o}C. Liquid cultures were grown at appropriate temperatures and under constant shaking (200 rpm). For long term storage at -80^{o}C, glycerol stocks were made by adding 500 µl of a saturated culture to 500 µl of 50% glycerol (*E. coli*), to 500 µl of 80% glycerol (*A. tumefaciens*) or to 500 µl of 30% glycerol (*S. cerevisiae*).

7.2.5 Growth media and selective media

Growth media used in this thesis are summarised in Table 32. For selection of *E. coli* and *A. tumefaciens* appropriate antibiotics summarised in Table 33 were added to the media after autoclaving.

Table 32 Media composition

Media	Content
LB	1% (w/v) tryptone
	0.5% (w/v) yeast extract
	1% (w/v) NaCl
SOC	2% (w/v) tryptone
	0.5% (w/v) yeast extract
	10 mM NaCl
	2.5 mM KCl
	10 mM $MgCl_2$
	10 mM $MgSO_4$
	20 mM glucose
2xYT	16 g bacto tryptone/ l
	10 g bacto yeast extract/ l
	5 g NaCl/ l
YPD	1% yeast extract
	2% peptone
	2% glucose
SC-Ura (yeast selection media)	0.67% yeast nitrogen base (without amino acids, with NH_4SO_4)
	dropout media without uracil (Formedium)
	2% glucose

Table 33 Antibiotics used for selection

Antibiotic	Final concentration (µg/ml)
carbenicillin	100
kanamycin	50
rifampicin	100
gentamycin	50
chloramphenicol	150

7.2.6 Plasmid extraction

Plasmid extraction for *E. coli* cultures for sequencing or for transformation into new hosts was done with a Miniprep kit (Qiagen, cat. No. 27104) in accordance with the manufacturer's instructions**.** DNA was eluted from the silica membrane using 40 µl water and stored at -20^{o}C for future use.

7.2.7 Competent cells

For *E. coli,* commercially available chemically competent cells were used. For plasmid propagation either One Shot TOP10 (Invitogen, cat. No. 4040-10) or strain HST08, Stellar Competent Cells (TaKaRa, Clontech, cat. No. 636763) were used.

For *A. tumefaciens* GV3101 electro competent cells were generated using the following protocol: Starting from a single colony of *A. tumefaciens* strain GV3101 a 5 ml LB overnight culture, supplemented with the antibiotics Gentamycin and Rifampicin was used to inoculate 1 litre of LB media and grown at 28^{o}C under constant shaking at 200 rpm to an OD of ~5x10^{7} cells/ml. All subsequent steps were conducted at 4^{o}C. The culture was centrifuged at 3000 g, the supernatant discarded, and the cell pellet washed with 500 ml ice cold 10% glycerol. This wash step was repeated twice reducing the resuspension volume first to 250 ml 10% glycerol and next to 5 ml 10% glycerol. The final cell pellet was resuspended in 1 ml sterile ice cold 1 M sorbitol, and dispensed in 200 µl aliquots for long term storage at -80^{o}C.

For *S. cerevisiae* chemically competent cells were generated fresh every time a transformation was planned. A single colony was grown overnight at 28^{o}C in 2 ml 2x YPD media. The culture was then pelleted by centrifugation and resuspended in 25 ml fresh 2x YPD media and then incubated at 28°C under constant shaking at 200 rpm for at least four hours. All subsequent steps were performed on ice or at 4°C. The culture was centrifuged at 3000 g for 10 min and the supernatant discarded. The pellet was washed twice with 250 ml ice cold water, aliquoted according to demand and used for transformation immediately.

7.2.8 Transformation protocols

For transformation into *E. coli* a maximum amount of 200 ng plasmid was mixed with 25 to 50 µl chemically competent *E. coli* cells on ice. After heat-shock, 45 seconds in a 42°C water bath, the cells were immediately placed on ice and 150 µl SOC medium was added. Following transformation the mixture was kept at 37°C for 1 hour, before being spread on agar plates containing the appropriate antibiotic, and grown overnight at 37°C. Colony PCR confirmed transformation success and positive colonies were selected for overnight cultures.

For transformation of *A. tumefaciens,* up to 300 ng of plasmid DNA were mixed with 50 µl of electro competent *A. tumefaciens* GV3101 in a 2 mm electroporation cuvette and placed on ice for 5 minutes. Using the Biorad micropulser (165-2100) the mixture was electroporated at 2.2 kV and immediately after electroporation 200 µl SOC media were added to the cuvette and

the complete mixture transferred into a fresh tube and incubated at 28°C for 2 hours. This mixture was then spread onto LB agar plates containing rifampicin, gentamycin and the plasmid specific antibiotic selection marker kanamycin. The plates were incubated for two days at 28°C until individual colonies became visible. Colony PCR identified transformation success and positive colonies were selected for overnight cultures.

A method modified after Giez (2007) [251] was employed for yeast transformation. Briefly, 100 to 300 ng of plasmid DNA were mixed with 34 µl water, 36 µl 100 mM lithium acetate and 240 µl 40% PEG 3350. The mix was vortexed and kept on ice. The yeast cells (45 µl of chemically competent *S. cerevisiae* prepared as described above) were mixed with 50 µl of a 2 mg/ml solution of single stranded carrier DNA (UltraPure salmon sperm DNA solution, ThermoFischer, cat. No. 15632011) and both mixes combined were vortexed again and incubated at 30°C for 80 minutes. Over the course of the incubation, cells were vortexed briefly three times. The mixture was centrifuged at 3000 g for 15 seconds, the supernatant removed and the cells resuspended in 200 µl TE buffer. This solution was spread onto SC-Ura agarose plates for selection, with colonies appearing after two days of growth at 28°C. Colony PCR confirmed transformation success and positive colonies were selected for further cultures.

7.2.9 Cloning and plasmids

For protein expression in *S. cerevisiae,* the yeast strain WAT11 was transformed with the pXP218 [144] expression vector. The commercial pJET1.2/blunt vector was used in *E. coli* for PCR cloning to identify the correct ORF if not previously known or for propagation prior to digestion. The pTRV2u plasmid [123] was used for preparation of the VIGS vector.

Table 34 Plasmids used and individual selection markers

plasmid	selection marker
pJET	carbenicillin
pTRV2u	kanamycin
pXP218	minus Uracil (auxotrophic marker)

7.2.9.1 Cloning into linearised pXP218 vector

Cloning into pXP218 used sticky ends produced by restriction digestion to join vector and insert in the ligation reaction in a directional manner. The vector was isolated from a 5 ml overnight culture of *E. coli*. The concentration of the vector was determined using a nanodrop spectrophotometer (NanoDrop, ND-1000). A maximum of 1 µg of vector was digested with the

restriction enzymes SpeI and XhoI in a 50 µl reaction mix containing 5 µl 10x reaction buffer and 10 units of each enzyme for 60 minutes at 37°C. Inserts were PCR products amplified using primer with the appropriate overhangs (Table 35) cut by the same restriction enzymes SpeI and XhoI. After restriction endonuclease digestion, the vector DNA was dephosphorylated using calf intestinal alkaline phosphatase (New England BioLabs, cat. No. M0290S) following the manufacture's instructions. Both the linearised vector and the digested insert were run on a 1% (w/v) agarose gel and a band of the right size was cut from the gel and purified (Machery-Nagel, cat. No. 740609.50) and the extraction product resuspended in 25 µl water. The vector and insert were mixed in a ratio of 1 to 3 for the cloning reaction. Cloning was carried out using the T4 DNA ligase (Invitrogen, cat. No. 15224-017) or, alternatively, T4 DNA ligase (Roche, cat. No. 10 481 220 001) according to the manufacturer's instructions. Briefly, for the T4 DNA ligase (Invitrogen, cat. No. 15224-017), 4 µl T4 buffer were mixed with vector and insert, not exceeding 0.1 µg total DNA and 0.1 µl T4 ligase. After adding water to a total volume of 20 µl, the reaction was allowed to proceed for 1 hour at room temperature prior to transformation of *E. coli* as described above.

Table 35 Primers used for cloning genes into the pXP218 yeast expression vector
Cloning/restriction sites are in bold.

name		sequence
CRO_021081	fw	**TTATTACTACTAGT**ATGGATGCTTTGCTTAATCCTGT
CRO_021081	rev	**AATCTAATCTCGAG**TTAAGCACCTGAAGAATTTGGCCTCA
T19H	fw	**TTATTACTACTAGT**ATGTTGTCTTCATTGAAAGATTTCTTCG
T19H	rev	**ATCCATGCCTCGAG**CTAAAAAATGGTAACCGGAGTTGC
MAT	fw	**TTATTACTACTAGT**ATGGACTCAATAACAATGGTTGAAACC
MAT	rev	**ATCCATGCCTCGAG**TTAATTAGAAGCAAATGAAAGGAGC
T3O	fw	**TTATTACTACTAGT**ATGGAGTTTCATGAATCTTCTCCCTTC
T3O	rev	**AATCTAATCTCGAG**TCATGCATAGGACGTAGCGATTAAA

7.2.9.2 Cloning into pJET1.2/blunt

The commercial vector pJET1.2 was used to amplify blunt end PCR products for sequencing and subcloning. Constructs in pJET1.2 were generated using the CloneJET PCR Cloning Kit (ThermoFischer Scientific, cat. No. 1231). Briefly, about 100 ng of PCR product (amount varies with length of PCR product) were mixed with 10 µl reaction buffer, 1 µl linearised cloning vector, 1 µl T4 DNA ligase, 1 µl PCR product and brought to a volume of 20 µl. After five

minutes at room temperature the mix was put on ice and then *E. coli* were transformed as described above.

Table 36 Primers used for cloning genes into the pJET1.2/blunt vector

name		sequence
CRO_021081	fw	ATGGATGCTTTGCTTAATCCTGT
CRO_021081	rev	TTAAGCACCTGAAGAATTTGGCCTCA
T19H	fw	ATGTTGTCTTCATTGAAAGATTTCTTCG
T19H	rev	CTAAAAAATGGTAACCGGAGTTGC
MAT	fw	ATGGACTCAATAACAATGGTTGAAACC
MAT	rev	TTAATTAGAAGCAAATGAAAGGAGC
T3O	fw	ATGGAGTTTCATGAATCTTCTCCCTTC
T3O	rev	TCATGCATAGGACGTAGCGATTAAA

7.2.9.3 Cloning into the pTRV2u vector

For all VIGS experiments a modified version of the original pTRV2 plasmid [90] was utilised. This modification introduced a USER cassette into the multiple cloning site of the original plasmid [30] resulting in the pTRV2u vector that could be used for USER cloning[123]. USER cloning is a ligation-independent cloning technique that utilises PCR primer that contain a single deoxyuridine residue. The empty pTRV2u plasmid was linearised using the restriction enzyme AsiSI and overhangs created using the nicking enzyme Nt.BbvC1. *E. coli* containing the empty vector plasmid were grown up from a stock and plasmids extracted. A maximum of 1 µg of plasmid was digested in a 50 µl reaction mix containing 5 µl 10x reaction buffer and 10 units of each enzyme. Reaction was allowed to proceed overnight at 37°C. The linearised plasmid was subsequently gel purified and could be stored at -20°C until further use. For the preparation of the pTRV2 plasmid with inserts for VIGS, a 200 to 500 bp long section of the gene of interest was amplified with the deoxyuridine residue containing primer. The PCR product was purified from agarose gel and subsequently treated with USER enzyme in a mixture containing the PCR product and the linearised vector. This procedure leads to the generation of 8 nucleotide 3' overhangs, which can complement the ends of a USER compatible linearised vector [123] and allows for transformation of *E. coli* without prior ligation. The reaction had a volume of 10 µl and was kept at 37°C for 40 minutes. Subsequently it was cooled on ice before the reaction mixture was used to transform 40 µl TOP10 *E. coli* by heat shock. SOC media (200 µl) was then added to the *E. coli* and they were incubated at 37°C for 1 hour, spread onto kanamycin selective agar plates and incubated at 37°C overnight. Colony PCR identified transformation

success and positive colonies were selected for overnight cultures. Plasmids were extracted from a positive colony. Correct inserts were confirmed by sequencing. *Agrobacterium tumefaciens* strain GV3101 was transformed as described above.

Table 37 Primers used for cloning into the pTRV2u vector
Cloning sites are in red.

C2	fw	GGCGCGAUGCT GAATCATCCTCATCAACAATG
C2	rev	GGTTGCGAUGATCTCTTCCGATTGCCCAAC
C3	fw	GGCGCGAUACTATAATTTCTCCGAGCTGAAAT
C3	rev	GGTTGCGAUCAGACAAGTAATCAATTCCAACAA
C4	fw	GGTTGCGAUAATGTTGGTACATGCATTCTG
C4	rev	GGCGCGAUGAAGCTAGTTGTTGTTGGTGC
C5	fw	GGCGCGAUAAC TCC GGC CAT TTG GTA TC
C5	rev	GGTTGCGAUATCTTTAATAGAAAGTGGCAAATGGTGC
C6	fw	GGCGCGAUACTTCATATCAGATGGCTCACC
C6	rev	GGTTGCGAUGCAGGTAAGCAGCTGCAAA
C7	fw	GGCGCGAUTTC TTC CTC CAC TTT CTG CTC TC
C7	rev	GGTTGCGAUAATCTACATCCCAGGTGACAC
C8	fw	GGCGCGAUCAG AAG TTT GAC GTG GAG C
C8	rev	GGTTGCGAUCTTCAACCAAATTCCCTGAGC
C9	fw	GGCGCGAUCCTTTGTCGTTGTATCAGATCC
C9	rev	GGTTGCGAUGTACCATATACAGCCTTAATAACAGG
C10	fw	GGCGCGAUTTCAGTGTCATGATGTAATGCAGTTAAG
C10	rev	GGTTGCGAUAAGAATGGCAAAGTAAAGCACGAG
C12	fw	GGCGCGAUAAT CCA ATA GAA CAT CCA TAA AGT CTC CAG
C12	rev	GGTTGCGAUTCAATGGAGAGAAAAGGTGAGGT
C13	fw	GGCGCGAUTTGGTCTAGGTGGACTTGGT
C13	rev	GGTTGCGAUTGTTCTGCTGCAAAATCCAGCAT
C14	fw	GGCGCGAUAGTTAGAGTTGGTATAGTTGGTC
C14	rev	GGTTGCGAUGTTATGTTCTGCTGCAAAA
C15	fw	GGCGCGAUTGC CAC CGT TGG AAG AAC
c15	rev	GGTTGCGAUCATGGCCTTCCATGTACCTTGTA
C16	fw	GGCGCGAUGAC AAG ATT TGT AAT AAG TTG ATG GAA TCA ACT
C16	rev	GGTTGCGAUCCATTTGGTGCTGGACGAAG
C18	fw	GGCGCGAUATT TAT GCA AAC TCA TGG CCT AAA CTT TGC
C18	rev	GGTTGCGAUAACTTCTTTGCTTTGGGCCTCAATG
C20	fw	GGCGCGAUTGAGCAAATTGATGACGTCATGCA
C20	rev	GGTTGCGAUAAGGGAATATCAAGGTCAAGATCACCAC
C30	fw	GGCGCGAUACCAGGTCCAAGAACACTACC
C30	rev	GGTTGCGAUGTAATCTCCATACTGAGCAAAGGTAACAC

C32	fw	GGCGCGAUGTGGTGTATTGGCAAACTTCCAG
C32	rev	GGTTGCGAUCTTTCACAAAGTAGTTTCTTGGCAATACA
C36	fw	GGCGCGAUTGGTCCAAACACAGAAGAATCATTTC
C36	rev	GGTTGCGAUGATGATGAACCTTAGCATATCTGTGATT
C38	fw	GGC GCG AUC TTC GCC TAC GCA TCT CCC TTC TGG A
C38	rev	GGT TGC GAU TTG TAT TCA TGA TTT CTT CCA CAC CTT
C39	fw	GGC GCG AUA TGG AAT CGA ACG AGT TGA CTC ACT
c39	rev	GGT TGC GAU TCG GTA CCA CAA TTT CGA ATC TGG
C41	fw	GGC GCG AUATGAAGGTGATAGAGAAGATCAACGAT
C41	rev	GGTTGCGAUCAGGCAAATCCACCTTCCACCT
C43	fw	GGCGCGAUAGC TCG AGT TCA TTC CCA ACT CTC A
C43	rev	GGTTGCGAUCATAGGCGGAAAATGATGGGATCC
C44	fw	GGCGCGAUAAAGAACCTCCAGCGAGCA
C44	rev	GGTTGCGAUCGAGTCAAGCTGTGCTACTGA
C45	fw	GGC GCG AUACA CTT CGT AGC CGA AAC CAA AGT TA
C45	rev	GGT TGC GAU GTTGTGAACAATATTAGCACCAGGACTG
C46	fw	GGCGCGAUAGC TCG AGT TCA TTC CCA ACT CTC A
C46	rev	GGTTGCGAUCATAGGCGGAAAATGATGGGATCC
C47	fw	GGCGCGAUGCT TCC CTA CTT GCA AGC AAT AGT TA
C47	rev	GGTTGCGAUCCGCCAACTGGACAGTCATG
C48	fw	GGCGCGAUGCC TAG CTT CAA GAA CAG GAT GC
C48	rev	GGTTGCGAUATCCTAAGTATTGGTCCGAACCTGA
C53	fw	GGC GCG AUTACCCAATGACACCTGGTAT
C53	rev	GGTTGCGAUGGTTAATCCCTTGTCTGTCG
T16H2	fw	GGCGCGAUTGATTGGAAACTTGCCAGTGATGA
T16H2	rev	GGTTGCGAUATGGATTAGCAACGAAATCTTCCGTA

Table 38 Primers used for colony PCR and sequencing

plasmid	name	sequence
pJET	pJET fw	CGACTCACTATAGGGAGAGCGGC
pJET	pJET rev	AAGAACATCGATTTTCCATGGCAG
pXP218	pXP218 fw	GGACCTAGACTTCAGGTTGTCTAACTC
pXP218	pXP218 rev	TTTACAGATCATCAAGGAAGTAATTATCTAC
pTRVu	pTRV2u fw	GATGGACATTGTTACTCAAGGAAGC
pTRVu	pTRV2u rev	GTAGTTTAATGTCTTCGGGACATGC

7.3 Virus induced gene silencing (VIGS)

Construction of VIGS plasmids and cloning are described in section 7.2. Subsequent RNA extraction and qRT PCR experiments are described in section 7.3. VIGS experiments were

conducted using eight week old *C. roseus* plants [121]. To silence whole plants including tissues such as roots of *C. roseus* an existing protocol was modified protocol [93].

7.3.1 VIGS on Catharanthus roseus*: eight week old plants pinching method*

Eight week old plants were used for VIGS of *C. roseus*. Briefly, 5 ml cultures from *A. tumefaciens* glycerol stocks harbouring the pTRV1, the pTRV2_EV and the pTRV2 constructs containing the insert of choice were grown overnight. Each experiment included pTRV2_ChlH as a visual marker of silencing success, visible as yellowing of green tissue. The overnight cultures were centrifuged and resuspended in *A. tumefaciens* infiltration buffer (10 mM MES, 10 mM $MgCl_2$, 100 µM acetosyringone, pH 5.5) to an optical density of 0.7 at 600 nm. The cultures were incubated for an additional two hours at room temperature. Prior to pinching using forceps just below the top leaf pair of each plant *A. tumefaciens* carrying pTRV2 constructs were mixed with an equal volume of the *A. tumefaciens* pTRV1 culture. Plants were kept for 21 days in the growth chamber, until the seedlings infiltrated with the pTRV-ChlH vector displayed substantial yellowing of leaves. The last leaf pair to emerge above the inoculation site was harvested from each plant after the plants were visually inspected for any other physiological changes in colour or shape. About 30-50 mg tissue were harvested from each plant and immediately frozen in liquid nitrogen. Per construct up to eight plants were harvested. The tissues were homogenised using a cryo bead mill (Cryomill, Retsch, Germany). Each sample was subsequently divided into two equal batches of milled tissue, one for RNA and one for metabolite extraction to obtain maximum comparability between gene expression levels and metabolic phenotype of each plant sample.

7.3.2 VIGS on Catharanthus roseus*: young seedling syringe-press method*

For an alternative VIGS method, three to four week old seedlings were used which were grown until the point where the first pair of true leaves had not fully emerged above the cotyledons. After carefully extracting these seedlings from the soil, the roots were washed with water to remove excess soil. The plants were placed in a 50 ml syringe with its front opening sealed off. The syringe was filled with the mixture of pTRV1 and pTRV2 plasmid of interest containing *A. tumefaciens* that was prepared as described above and the syringe plunger was pressed tree times firmly into the syringe creating pressure for up to 15 seconds. The plants were washed in water before replanting into soil.

7.3.3 Sample preparation for mass spectrometry

For each sample between 10-25 mg of fresh harvested, frozen in liquid nitrogen and ground tissue was weighed, collected into 200 µl methanol containing 40 µM caffeine as an internal standard and incubated at 57°C for 2 hours. After a 30 minute centrifugation step at 5000 g, an aliquot of the supernatant, mixed with an equal volume of water, was analysed by LC-MS.

7.4 General methods for analytical chemistry/ mass spectrometry

This section describes the instruments and methods used to analyse alkaloid content and composition of plant tissues as well as enzyme assays.

7.4.1 Analysis of VIGS samples

Liquid Chromatography Mass Spectrometry (LC-MS) to analyse all VIGS experiments used either a Surveyor high performance liquid chromatography (HPLC) system attached to a DecaXPplus ion trap MS (Thermo-Finnigan) or a Nexera/Prominence UHPLC IT-TOF. All compounds of interest ionise comparably on both machines and do not interfere with the general result of individual VIGS experiments (Dr. Lionel Hill, Dr. Richard Payne, personal communication). For all experiments caffeine was used as an internal standard and peak areas were normalised to the fresh weight of the sample.

All analyses were performed in positive ion mode. The injection volume was 2 µL for all samples.

<u>Surveyor HPLC system attached to a DecaXPplus ion trap MS (Thermo-Finnigan)</u>

VIGS samples analysed on the Thermo-Finnigan instrument with Deca XP ion trap detector were separated using a Phenomenex Luna C18 column (100 x 2.00 mm, 3 µm) with a binary solvent system consisting of acetonitrile (ACN) and 0.1% formic acid in water (12% ACN for 1 minute, 12 to 25% in 7 minutes, 25 to 50% in 4 minutes, 50 to 100% in 2 minutes, 100% for 7 minutes, 100 to 12% in 2 minutes, 12% for 6 minutes). Flow rate was set to 0.5 ml per minute.

Metabolites were detected by positive mode electrospray ionisation (ESI). MS spectra were collected from *m/z* 100-1000. Spray chamber conditions were 350°C capillary temperature, 50 units sheath gas, 5 units aux gas, and a spray voltage of 3.8 kV using a steel needle kit.

Nexera/Prominence UHPLC IT-ToF (Shimadzu)

VIGS samples that were analysed on the Shimadzu LC-MS system attached to an ion-trap ToF mass spectrometer were separated with a Phenomenex Kinetex C18 100Å (100 × 2.10 mm, 2.6 µm) column using a binary solvent system consisting of acetonitrile (ACN) and 0.1% formic acid in water (10 to 25% ACN in 5 minutes, 25 to 100% ACN in 1 minute, 100% ACN for 1.5 minutes, 100 to 10% in 0.5 minutes, 10% for 2 minutes). Flow rate was set to 0.5 ml per minute.

Metabolites were detected by positive mode electrospray ionisation (ESI). Full MS spectra were collected from *m/z* 100-1000 Spray chamber conditions were 250°C curved desorbation line temperature, 300°C heat block, 1.5 l/min nebulizer gas, and drying gas "on". The instrument was calibrated immediately before use, using sodium trifluoroacetate as a mass standard according to the manufacturer's instructions.

7.4.2 Analysis of enzymatic assays and yeast cultures

Liquid Chromatography Mass Spectrometry (LC-MS) analysis of all enzyme assays and yeast cultures used a Waters Xevo TQ-S mass spectrometer (Milford, MA, USA) equipped with an electrospray (ESI) source. Peaks were separated with a BEH Shield RP18 (50 x 2.1 mm; 1.7 µm) column (Waters) using a binary solvent system consisting of acetonitrile (ACN) and 0.1% formic acid in water (12 to 25 % ACN in 50 seconds, 25 to 50% ACN in 180 seconds, 50 to 100% ACN in 130 seconds, 100% for 130 seconds, 100 to 10% in 30 seconds, 10% for 2 minutes).

Metabolites were detected by positive mode electrospray ionisation (ESI). Full MS spectra were collected from *m/z* 100-800. Capillary voltage was 2.5 kV in positive mode, the source was kept at 150°C, desolvation temperature was 500°C, cone gas flow was 50 litre per hour, and desolvation gas flow 900 litre per hour. Flow rate was set to 0.6 ml per minute.

For loganin and secologanin analysis was carried out by monitoring specific mass transitions. Methods were developed using commercial standards. Flow injection of individual compounds was used to optimize the multiple reaction monitoring (MRM) automatically using the Waters Intellistart software. Loganin and secologanin were monitored using the following parameters: secologanin: *m/z* 227.1>107.1 (Cone 20, Collision 18), loganin: *m/z* 229.1>81.1 (Cone 14, Collision 20), *m/z* 229.1>109.1 (Cone 12, Collision 12). Data was processed using MassLynx 4.1 and TargetLynx software (Waters).

7.4.3 Accurate mass measurements

Accurate mass measurements were done at the JIC mass spectrometry facility by Lionel Hill. Tissue samples were run on a Surveyor LC system attached to an LTQ Orbitrap (Thermo). Separation was performed on a 100 × 2 mm 3 µm Luna C18 column (Phenomenex) using a 30 minute gradient of acetonitrile versus 0.1% formic acid in water (0 to 12% in 1 minute, 12 to 25% in 7 minutes, 25 to 50% in 4 minutes, 50 to 100% in 2 minutes, 100% for 7 minutes, 100 to 12% in 2 minutes, 100% for 2 minutes). Flow rate was set to 0.5 ml per minute.

The instrument was set up to collect full scan FT-MS data at from *m/z* 100-1000, and data-dependent MS2 of the most abundant ions at an isolation width of *m/z* 3.0 and collision energy of 35%. The instrument was calibrated according to the manufacturer's instructions at the start of the sequence.

7.4.4 Data extraction and analysis

Mass spectrometry data were analysed using extracted ion chromatogram and individual peak area was calculated using the appropriate software for each instrument. For the Thermo-Finnigan instrument peak areas were calculated with the Thermo, Xcalibur Roadmap 2.2 SP1.48 software. For the Shimadzu LC-MS-IT-TOF mass spectrometer peak areas were calculated using the Shimadzu LC-MS solution software version 3.80.409. For the Waters Xevo TQ-S mass spectrometer the Waters, MassLynx V4.1 SNC 714 software was used.

Extraction of all peaks from the mass spectrometry data was performed with the software tool XCMS. The raw data files obtained by mass spectrometric measurements from the Thermo-Finnigan instrument and Shimadzu-IT-TOF mass spectrometer data were converted to mzXML files, the required file format for XCMS analysis. File conversion was done with a free online version of readW (http://sourceforge.net/ projects/sashimi/files/). Peak extraction was done using R (version 2.8.0) with the additional packages XCMS, http://www.bioconductor.org/. XCMS [252] enables to visualise changes in specific metabolites by identifying the relative metabolite ion intensities after correcting for potential retention time shifts by automated calculation of a nonlinear retention time correction profile. The data obtained in this fashion were normalised by weight of tissue sample. This method can also be applied for comparing a treatment with a control. The mean peak area for each treatment was calculated from the normalised data, the fold change between VIGS-gene of interest and VIGS-EV controls, as well as a t-test were calculated between the two sets.

The raw data files obtained by mass spectrometric measurements from the Shimadzu-IT-TOF mass spectrometer were used directly in the Profiling Solution (Shimadzu) software for automated peak extraction. Here the software standard parameters were used. Only alteration was to increase the ion intensity to 50000 counts to focus on larger, more reliable peaks. Only data obtained during the gradient from 10 to 25% ACN were considered. As with all other metabolic data the so obtained peak areas were normalised by weight of tissue sample and used for further statistical analysis.

7.4.5 Statistical analysis

Significant differences in peak area were calculated in Microsoft Excel using the Students's T-test. Only p-values of less than 0.01 were considered significant. For automated peak extraction additionally Benjamini-Hochberg false discovery rate was calculated [229].

7.5 Quantitative Real Time Polymerase Chain Reaction (qRT PCR)

To investigate the altered expression of genes in silenced tissue qRT PCR was conducted on six to eight individual samples per treatment.

7.5.1 RNA extraction

Material for RNA extraction was either milled (RETSCH mill, Germany) or ground under liquid nitrogen. RNA extraction was carried out using the Qiagen RNeasy plant mini kit (cat. No. 74904), including the on-column DNase digestion using the Qiagen RNase-Free DNase set (cat. No. 79254). The amount of tissue used was < 100 mg per sample. For each VIGS candidate, tissue samples from 6 to 8 plants were selected for RNA extraction. As control, the same amounts of empty vector control sample were used. RNA quality was assessed on a 1% agarose gel and the concentration measured on a nanodrop spectrophotometer (NanoDrop, ND-1000).

7.5.2 cDNA synthesis for qRT PCR

For qRT PCR the obtained RNA was used to synthesise cDNA for each replicate using the Biorad iScript cDNA synthesis kit (cat. No. 170-8891). Briefly, the 5x iScript reaction mix, iScript reverse transcriptase, nuclease free H_2O and 1 µg RNA from each VIGS sample were mixed to yield a total volume of 20 µl. The mixture was incubated for 5 minutes at 25°C, 30 minutes at 42°C, and a final 5 minutes at 85°C, resulting in the formation of the cDNA which was used for qPCR.

7.5.3 Primers for qRT PCR

Primers for qRT PCR were designed to specifically amplify a fragment of approximately 100 bp of the gene of interest. Only primers that successfully and specifically amplified a DNA fragment of expected size in a standard PCR using *C. roseus* cDNA were considered further.

Table 39 Primers used for qRT PCR of target genes

STR q RT PCR fw	CGCAGATGGTTCCTTTGTTG
STR q RT PCR rev	GGCAGTGCAGAGTTCTTAGT
TDC q RT PCR fw	TATCCGGTCCTTAGCGAAGT
TDC q RT PCR rev	GCCACTTGAAGCTGAGGAAT
SLS q RT PCR fw	GGCCAAAACTTTGCACTTCT
SLS q RT PCR rev	GGAGCATGAACATAGGATGG
T3O q RT PCR fw	GGCAACTCCCAGATGGTTCTACT
T3O q RT PCR rev	TCATGCATAGGACGTAGCGATTAAATGAA
T16H2 q RT PCR fw	ATCAACTCACAGTGGCAGTC
T16H2 q RT PCR rev	GACTTGAGGACTTGTGATTGGC
G8H q RT PCR fw	CATTTATTAGGCGACCAACC
G8H q RT PCR rev	GAACTTCTTTCGCCATTGTT
DAT q RT PCR fw	GGTTTCAATTTATTTCTCACGTAC
DAT q RT PCR rev	AACTATCAGAAAGGTAAGCATCGA
D4H q RT PCR fw	ATAGTTAATCATGGGATTCCACAAGATGTT
D4H q RT PCR rev	GTTCATGAAACTTACGAACTCCATCTAC

Suitable qRT PCR primers for each target gene were tested for their efficiency to amplify the gene in a linear relationship for a given cDNA concentration range. For this purpose cDNA from the VIGS-EV and the relevant target VIGS-pTRV2 constructs was pooled and a serial dilution of the pooled cDNA was generated at concentrations of 1:5, 1:10, 1:20, 1:80, 1:320. A qRT PCR reaction for each primer pair was performed in technical triplicate at each concentration step and the Cq plotted against log[cDNA]. Only primer pairs that gave a linear regression (R^2) value of 0.99 were used in this study. For target genes with lower expression the serial dilution was changed to 1:2, 1:4, 1:8, 1:16, 1:32. Only primers with an efficiency value between 98-110% were used in this study and the individual efficiency values were incorporated in the calculation of normalised expression.

7.5.4 qRT PCR experiments

To assess the level of silencing of different target genes in silenced and unsilenced tissue, material from six to eight independently silenced plants was compared to six to eight plants treated with the empty vector control plasmid. The qRT PCR reaction was performed in technical duplicate. Generally, the cDNA concentration was optimised according to the results of the primer testing to ensure that the Cq value for each gene analysed in the experiment would be between 20-30 cycles in the actual qRT PCR reaction.

All melting curves generated in each experiment as well as the initial primer testing were inspected visually to ensure that the primer pair was specific in its amplification with only one amplified target in the cDNA. A single transition in the melting curve indicated a specific binding of the primer to only the desired target.

Table 40 Reference gene primer for qRT PCR

qRT PCR-Rbps9_fwd	TTGAGCCGTATCAGAAATGC
qRT PCR-Rbps9_rev	CCCTCATCAAGCAGACCATA
qRT PCR-EF1a_fwd	TCAGGAGGCTCTTCCTGGTGA
qRT PCR-EF1a_rev	AGCTCCCTTGGCAGGGTCAT

The relative quantification of gene expression was calculated using the expression values of a reference gene for normalisation of the data. The reference gene used in all experiments was the *40S Ribosomal protein 9* (*Rps9*) or *Elongation Initiation Factor 4a* (*EIF4a*). These reference genes have previously been established for their suitability for use in VIGS qRT PCR experiments in *C. roseus* [30,121].

All qRT PCR measurements were done using the CFX96 touch Real-Time PCR system (BioRad) and the SYBR Green I dye, a cyanine dye that preferentially binds to double stranded DNA. Each reaction was performed in a total volume of 25 µl consisting of a normalised concentration of cDNA, 0.2 mM of appropriate forward and reverse primer and the SsoAdvanced SYBR Green Supermix (BioRad, cat. No. 1725271) containing dNTPs, Sso7d fusion polymerase, $MgCl_2$, SYBR® Green I and ROX normalisation dyes. The qRT PCR reaction was initiated by a denaturation step at 95°C for 10 min followed by 40 cycles at 95°C for 15 s and 60°C for 1 minute. The resulting data were analysed using CFX software.

7.6 General methods for bioinformatics

7.6.1 Geneious

Primer design, identification of restriction sites and analysis of obtained sequencing data was performed with the Geneious software, version 7.0.6 [253]. This software was also used for BLAST search[254] against the various datasets utilised in this thesis. Unless otherwise stated, this was done using the default setting for the blastn or Megablast search functions. For an alignment of two or more sequences, unless otherwise stated, the default setting of the Geneious software was used which is a Smith-Waterman local alignment with a cost matrix of 65% similarity (5.0/-4.0). For mapping of two or more sequences, unless otherwise stated, the default setting of the Geneious software was used.

7.6.2 Mapping individual reads and visualisation

For mapping raw sequencing, data files were downloaded from NCBI. Using the JIC cluster, raw reads were mapped onto reference sequences using Bwa [255] version 0.7.12, and default parameters. Results were sorted using samtools [256] version 0.1.19. Mapping results were visualised and manually inspected using the Tablet version 1.15.09.01, James Hutton Institute [257]. Data sources employed for mapping are given in Table 41.

Table 41 Sequencing data employed in mapping

SRR122239	SRX047002: Catharanthus roseus Flowers RNA-Seq (CRA_AA)
SRR122251	SRX047016: Catharanthus roseus Mature Leaf RNA-Seq (CRA_AM)
SRR122252	SRX047017: Catharanthus roseus Immature Leaf RNA-Seq (CRA_AN)
SRR122253	SRX047018: Catharanthus roseus Stem RNA-Seq (CRA_AO)
SRR122254	SRX047019: Catharanthus roseus Root RNA-Seq (CRA_AP)
SRR122255	SRX047020: Catharanthus roseus Hairy Root (TDCi) RNA-Seq (CRA_AQ)
SRR122243	SRX047007: Catharanthus roseus Sterile Seedlings RNA-Seq (CRA_AE)
SRR122244	SRX047009: Catharanthus roseus Sterile Seedlings RNA-Seq (CRA_AF)
SRR122245	SRX047010: Catharanthus roseus Sterile Seedlings RNA-Seq (CRA_AG)
SRR122257	SRX047022: Catharanthus roseus Wild Type Hairy Root RNA-Seq (CRA_AS)

7.6.3 Co-expression analysis

For the analysis of co-expression in a large transcriptomic dataset comprising 23 different tissues and/or treatments [65] , the MeV (MultiExperiment Viewer, version 4.9.0) software was employed to perform hierarchical clustering [25868]. The standard settings for hierarchical clustering, the Pearson correlation with average linkage clustering, was used. Prior to co-expression analysis, the larger datasets were filtered to reduce the number of contigs in the

individual study. Different hierarchical clustering dendrograms were generated using different statistical distance metrics for the clustering algorithm. The position of known pathway genes in those dendrograms were analysed and co-expression analysis applying different combinations of tissues and/or treatments, depending on scientific question, repeated.

7.7 Protein expression in *Saccharomyces cerevisiae*

Enzyme expression in *S. cerevisiae* was performed using the yeast strain WAT11. This yeast strain harbours the *A. thaliana* cytochrome P450 reductase [143] making WAT11 particularly suitable for the expression of plant cytochrome P450 enzymes. For enzyme expression, WAT11 was transformed [251] with the plasmid pXP218 [144] containing the enzyme of interest or an empty plasmid as control (Table 42).

Table 42 WAT11 strains harbouring *C. roseus* genes.

strain name	gene included
WAT11+pEV	empty vector
WAT11+pT3O	*T3O* JN613016.1
WAT11+pT19H	*T19H* AF253415.1
WAT11+pMAT	*MAT* HQ901597.1
WAT11+C16	CRO_021081 (sequence in appendix)

General information on methods such as preparation of competent yeast cells and yeast transformation, cloning of genes into the yeast expression vector, primers, media and growth conditions are given above. All LC-MS analyses described in this section are conducted using an AQUITY UPLC with a Xevo TQ-S mass spectrometer equipped with a BEH Shield RP18 1.7 µm column (Waters) as described in Section 7.4.2.

For the characterisation of the cytochrome P450 *T3O* in Chapter 3 a second yeast strain was employed. This strain had been engineered [146] using the commercial parent strain BY4741 (EUROSCARF, Frankfurt, Germany) to constitutively express known vindoline biosynthesis genes *T16H*, *16OMT*, *NMT*, *D4H* and *DAT* [34,41,43–45] integrated into its genome. Additionally it contains the *C. roseus* cytochrome P450 reductase [259]. The engineering was done by Dr. Stephanie Brown (O'Connor lab). This yeast strain (strain A) was additionally transformed to

contain a *T3O* expressing plasmid or alternatively an empty plasmid (strain_A+pT3O or strain_A+pEV).

7.7.1 Large scale liquid cultures and product extraction

For the production of 16-methoxytabersonine, the expected substrate of the T3O catalysed reaction, a 500 ml YPD media culture of strain A was started from a single colony and grown to an OD of 0.5 under constant shaking at 30 °C. At that point the culture was supplemented with tabersonine (AvaChem Scientific) to a final concentration of 120 µM. Conversion of tabersonine was monitored every 24 hours by LC-MS. After 48 hours no tabersonine could be detected and the enzymatic product was extracted from the liquid fraction using 3 x 400 ml ethyl acetate (EtOAc). The combined EtOAc fractions were dried with Na_2SO_4 and concentrated to a yellow oil. As judged by NMR the product eluted as the salicylic acid salt. It was therefore re-dissolved in a small volume of EtOAc and filtered through K_2CO_3. The final amount of product recovered was 3.9 mg.

To obtain the uncharacterised product of the T3O catalysed reaction, a 500 ml culture of strain A supplemented with 120 mM tabersonine was grown and monitored as described above. In parallel, from a single colony of the WAT11+pT3O yeast, a 10 ml culture was grown to an OD of 1 in the appropriate selective media SC-Ura supplemented with 2% (w/v) glucose. After 24 hours the cells were spun down at 3000 g for 3 minutes, the pellet was washed with sterile water and spun down again. The pellet was used to start a 500 ml culture of SC-Ura supplemented for induction with 1% (w/v) galactose and 1% (w/v) glucose and grown to an OD of 0.5. At full conversion of tabersonine to 16-methoxytabersonine in the culture of strain A, both cultures were spun down and the supernatant (media fraction) of the strain A culture was added to the pelleted WAT11+pT3O cells. This culture was allowed to continue to grow under constant shaking at 30 °C. It was monitored by LC-MS till almost full conversion of the 16-methoxytabersonine was observed. Subsequently the T3O enzymatic product was extracted using 3 x 400ml EtOAc. The combined EtOAc fractions were dried with Na_2SO_4 and concentrated to a yellow oil. The oil was subjected to silica column chromatography. It was first eluted with EtOAc. Elution was continued with 10% MeOH in CH_2Cl_2. As judged by LC-MS, the product eluted in the last EtOAc fractions and the early 10% MeOH fractions. Consequently these fractions were combined, concentrated, and subjected to preparative HPLC using a C18 column and an increasing gradient of acetonitrile in 0.1% formic acid/water. All resulting product-containing fractions were pooled and concentrated to give 10 mg product.

7.7.2 Structural elucidation of products obtained from yeast culture

The product of the first culture was confirmed to be 16-methoxytabersonine by comparing the obtained nuclear magnetic resonance (NMR) spectra, as ^{1}H-^{13}C 2D heteronuclear single-quantum correlation (HSQC) experiments, to reported literature values [147]. Data were recorded on a Bruker Avance III 400 NMR spectrometer, operating at 400 MHz for ^{1}H and 100 MHz for ^{13}C respectively.

The T3O reaction product was analysed and characterised using data obtained from infrared spectroscopy (IR), high resolution mass spectroscopy (HRMS) and $^{1}H/^{13}C/^{15}N$ NMR by Dr. Nathaniel Sherdan [146].

7.7.3 Yeast microsomal extraction

For yeast microsomal preparation a protocol modified after ref. [145] was used. The buffer TE (50 mM Tris, pH 7.4, 1 mM EDTA), TEK (0.1 M KCl in TE), TES-B (0.6 M sorbitol in TE) and TEG (20% glycerol in TE and a complete protease inhibitor tablet) were prepared prior to microsomal extraction. The yeast strain WAT11 expressing the desired cytochrome P450 enzyme was grown up from a single colony in a 10 ml overnight culture in the selective yeast media SC-Ura supplemented with 2% (w/v) glucose. This pre-culture was spun down, the supernatant discarded and the pellet washed twice with water before resuspension in 200 ml of selective induction media SC-Ura plus 1% (w/v) galactose and 1 % (w/v) raffinose, to induce production of recombinant genes. The culture was grown for 48 hours before it was centrifuged for 5 minutes at 2000 g. The supernatant was discarded and the pellet resuspended in 5 ml buffer TEK and left for 5 minutes at room temperature.

All subsequent steps were carried out on ice or at 4°C. After a second centrifugation step at 2000 g for 5 minutes the pellet was resuspended in 2 ml cold buffer TES-B. The cells were lysed using a cell disruptor applying 25 KPSI twice. The lysate was centrifuged for 10 minutes at 11 000 g to separate un-lysed cells and cell debris (pellet) from the microsomal fraction (supernatant). The supernatant was collected and spun down 180 minutes at 35 000 to 40 000 g. The resulting pellet was carefully resuspended in a small volume (150–300 µl) TEG by gently stirring it with a pipette tip. After 10 minutes on ice, aliquots of the resuspended pellet were transferred to fresh 1.5-mL tubes and stored at –80°C for further use.

7.7.4 Bradford assay to determine protein concentration

The total protein concentration of the yeast microsomal fraction was assessed using the Bradford method [260]. Microsomal protein content was determined using Bradford reagent (SIGMA, cat. No. B6916). A standard curve for absorbance relative to protein concentration was generated using known concentrations of bovine serum albumin (BSA) in the range of 62.5 µg/ml to 1 mg/ml, with each concentration measured in triplicate at 595 nm.

7.7.5 Yeast microsomal enzymatic assays

Assays consisted of substrate, 55 µg microsomal protein extraction, 1 mM NADPH, 4 mM DTT and 100 mM phosphate buffer pH 7.0 in a total volume of 100 µl. The assays were stopped by adding 1 volume of methanol. Substrates such as commercially available tabersonine, ajmalicine and serpentine were used. Additionally, 5 µl aqueous leaf extract prepared as described below were also used as a substrate.

Aqueous leaf extracts were obtained by incubating approximately 100 mg of leaf tissue in which the T3O gene had been silenced with VIGS tissue with 2 ml methanol for 60 minutes at 57°C. The methanol extract (500 µl) was applied onto a C18 column (Kinetex, 100 x 2.1 mm, C18, 1.7 µm), eluted with 500 µl of water and assayed with extracted microsomes.

7.8 Methods for whole genome sequencing and BAC library

Sequencing was carried out at The Genome Analysis Centre (Norwich, UK, http://www.tgac.ac.uk/main-icons/platforms/sequencing-platforms/) after constructing a single TruSeq genomic DNA library that was sequenced on an Illumina HiSeq and required a minimum of 5 µg high quality genomic DNA. Genomic DNA extraction is described under 7.8.1. For the construction of a mate pair library further material was obtained but only a crude DNA extraction was performed and can be found under 7.8.2, as the following steps were conducted at our collaborator Robin Buell's Lab at the Department of Plant Biology of the Michigan State University (MSU, East Lansing, USA).

7.8.1 Genomic DNA extraction for sequencing

Material for whole genome sequencing was obtained from purified nuclei of four individual plants. Young leaves of 12 week old plants of *C. roseus* "SunStorm Apricot" was harvested, using primarily young leaves. Plants were grown in a growth chamber at 25°C with 12 h light/12 h dark as previously described. Prior to tissue harvest plants were kept in the dark for two days to reduce levels of metabolites such as carbohydrates or polyphenols contained in

tissue [261]. The material combined was 27 g of young leaf tissue. A modified protocol was used [218,262,263].

7.8.1.1 Isolation of nuclei, DNA extraction and dialysis

The tissue was ground to a fine powder under liquid nitrogen. Prior to extraction 500 ml of TKE buffer with 0.1 M Tris, 1 M KCL, 0.1 M EDTA (the pH was not adjusted) and subsequently 1000 ml of a modified sucrose-based extraction buffer (SEB) 10% v/v TKE, 500 mM sucrose, 4 mM spermidine, 1 mM spermine tetrahydrochloride, 1.2 g/L PEG 8000 and 0.2% v/v ß-mercaptoethanol (BME) was prepared fresh. The SEB buffer was prepared fresh and adjusted to pH 9.5 with HCl. The SEB with BME mixture had to be stored on ice.

SEB buffer with 10% v/v Triton X-100 (15 ml per gram of leaf tissue) was added to the powdered material and gently stirred. The mixture was kept on ice for further 10 minutes. Using a funnel the mixture was filtered through two layers of Miracloth (Merck Millipore, cat. No. 475855-1R) squeezing gently to obtain more nuclei. 1/20 volume of SEB + BME/Triton was added to the filtered mixture and left on ice for further 10 minutes with 20 seconds of gentle swirling every 2 minutes. The mixture was transferred into 250 ml polypropylene centrifuge tubes and centrifuged at 650 g for 20 minutes at 4°C. The supernatant was pipetted off without disturbing the pellet. A small amount of SEB + BME (<1 ml) was added to the pellet and the nuclei were resuspended by gently shaking. More SEB + BME was added to a total volume of 20 ml. This wash step was repeated twice. From a 20% (w/v) aqueous solution, SDS was added to a final concentration of 2% (w/v) to lyse the nuclei. The content of the tube was mixed by gentle inversion. The tube containing the nuclear lysate was heated in a water bath to 60°C for 10 minutes and then left to cool to room temperature. Sodium perchlorate was added to a final concentration of 1 M. The lysate was spun down in a swinging bucket rotor at 500 g for 20 min at 10°C to pellet the starch grains. The supernatant was transferred to a new 15 ml polypropylene tube using a 1000 µl plastic pipette tip from which the bottom third had been cut off to minimize shearing of the DNA. To extract the DNA the nucleic acid solution was mixed with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1). To minimize shearing of DNA while mixing the organic and aqueous phases, a test tube rocker (VariMix, Thermo Scientific) was used at 18 cycles per minute for 15 minutes. The mixture was centrifuged at 3000 g in a swinging bucket rotor for 10 minutes. The upper aqueous phase was transferred into a new sterile polypropylene tube to perform a second extraction as described

above. The so obtained aqueous phase was dialysed for 16 hours into 5 l of 10 mM Tris-HCl, 1 mM disodium EDTA (TE) buffer (pH 7.0) at 4°C under constant gentle stirring using.

7.8.1.2 RNase & proteinase digestions, extraction and precipitation

The dialysed DNA sample was treated with RNase A/T1 mix (Thermo Fischer Scientific, cat. No. EN 0551) to remove RNA from the DNA sample. According to manufacturer's instructions 20 µl of RNase A/T1 mix were used for each ml of DNA sample and incubated at 37°C for 60 minutes. Next the DNA sample was supplemented with 150 µg/ml Proteinase K (Roche, cat. No. 03115887001) and incubate at 37°C for 60 minutes to inactivate proteins such as RNases and DNases. Two more phenol/chloroform/isoamyl alcohol extractions followed as described above. The aqueous phase was transferred to a clean tube to determine volume and one-tenth volume of 3 M sodium acetate (pH 5.2) was added. The solutions were mixed by inverting the tube several times. Two volumes of ethanol were added and mixed thoroughly. The precipitated DNA was collected by centrifuging the mixture again at 20 000 g and 4°C for 30 minutes. The supernatant was discarded, 1 ml 70% ethanol added and the centrifugation repeated. The supernatant was discarded and the remaining pellet allowed to air dry for 2 hours at room temperature. The pellet was resuspended in 100 µl water. The DNA concentration was quantified (Qubit, Thermo Fischer Scientific) to be 10.8 ng/µl. Therefore the extraction of 27 g young leaf *C. roseus* tissue resulted in 10.8 µg high molecular weight DNA. The DNA quality was checked for evidence of shearing or RNA contamination. For this purpose 200 ng of genomic DNA were run on a 0.6% agarose gel.

7.8.2 Material for mate pair library

Prior to extraction, 200 ml of 2 x CTAB buffer were prepared using 4 g cetyltrimethylammonium bromide (CTAB), 5 M NaCl, 1 M Tris-HCl (pH 8.0) and 0.5 M EDTA. Additionally 1% (w/v) of polyvinylpyrrolidinone (PVP) and 0.5% (w/v) activated charcoal were added [264]. Obtained from the same set of plants as described above, 5 g of young leaves were harvested and ground under liquid nitrogen. CTAB buffer (10 ml) was added after all liquid nitrogen had evaporated. The mixture was incubated in a water bath at 65°C for 1 hour and mixed by inverting every 10 minutes. After cooling to room temperature the mixture was shaken using a test tube rocker (VariMix, Thermo Scientific) at 18 cycles per minute for 10 minutes. An equal volume of chloroform/ isoamyl alcohol (24:1) was added. The sample was spun down for 10 minutes at 2000 g and the supernatant was transferred to a fresh tube. An equal volume of isopropanol was added and the centrifugation repeated. The DNA formed a

pellet that was extracted and transferred to a tube containing 70% ethanol. The mixture was inverted several times to wash the DNA and spun down for 15 seconds. The ethanol was removed and the pellet was left to dry at room temperature for 2 hours. The dried pellet was resuspended in TE buffer and sent to the Buel laboratory for sequencing.

7.8.3 Test PCRs to verify C. roseus *assembly*

After receiving the assembled draft genome of *C. roseus* the scaffolds were inspected manually for clustering of known pathway genes. To verify the most important of these findings, PCR from genomic DNA was employed to further strengthen the evidence for the existence of the clustering evident in the draft genome of *C. roseus*. Primer employed for this purpose are given in Table 43.

Table 43 Primers for verifying observed clustering of pathway genes

Target gene	Primer name		Target band size
TDC	TDC fw	TAC TTG GAT GGA ATC GAA CG	6849 bp
STR	STR rev	GAA TTC TGA TGG CCA TTT TT	

7.8.4 Methods specific for BAC library

BAC library construction was carried out at Bio S&T Inc. (Montreal, Canada). Briefly, for library construction the pIndigo-BAC5-HindIII vector (Epicentre Technologies) was employed. High molecular weight DNA was cloned into the HindIII site of this plasmid. The obtained clones were maintained in the *E. coli* strain DH10B. The library was pooled in a 96-well plate with each well containing approximately 500 independent primary clones. The library was confirmed to have 10x coverage and an average insert size of 155 kb (personal communication, Areti Karadimos). The company Bio S&T required a small amount of genomic DNA to serve as a positive control when screening the BAC library. This was taken from the remaining material of the genomic DNA extraction described above.

7.8.5 Plant material

Plant material was harvested from the same plants as were used for the whole genome sequencing material to ensure consistency of obtained sequence data. As described before plants were kept in the dark for two days prior to tissue harvest. Young leaves (25 g) were immediately frozen in liquid nitrogen and sent to Bio S&T Inc. (Montreal, Canada).

7.8.6 Gene specific primer for BAC library screening

Screening the BAC library required gene specific primers that bind exclusively to the target gene and produce a band of a size between 200 to 500 bp. For the identification of suitable primers that would amplify specific bands of the desired length the same genomic DNA prepared for the whole genome sequencing project served as template.

For the initial screen for three pathway genes the assembly of the whole genome sequence was not yet available, therefore the exact intron exon structure of the target genes was unknown and many primer pairs had to be tested. Especially for *SGD* this required substantial optimisation. The final band for *SGD* was only 167 bp long as no primer pair resulting in a longer band could be obtained.

Table 44 Primers for screening the *C. roseus* BAC library

Primer sequences and expected band size for screening of the *C. roseus* BAC library. Three gene were targeted in the initial screen (*SGD*, *T16H2* and *ISY*), a second screen was targeted at the genes *STR* and *MATE*, while the failure to retrieve *SGD* in the first round resulted in an additional screening for this gene with a second primer pair.

Target	Primer	Sequence	Band size
SGD	BAC_SGD fw	GGA CGA CTT TCA ATG AAC CAC ATA C	167 bp
	BAC_SGD rev	CCT ATA TAC TTC CAC AGC AGC TTT GT	
T16H2	BAC_T16H fw	TGA TTG GAA ACT TGC AGT GAT GA	265 bp
	BAC_T16H rev	TAC AAT GGA TTA GCA ACG AAA TCT TCC GTA GC	
ISY	BAC_ISY fw	TTG AGT ACA TCC AAT GTG ATG TCT CA	422 bp
	BAC_ISY rev	TGT ACT GAC AAT GTT CAT. CAT. ACT ACA TGG	
MATE	BAC_MATE fw	GCC AAG TGT CCT TAG TCC TAT CAC TAT ATA G	377bp
	BAC_MATE rev	TCT AAC CCT AAC AGG CAT. TGT TAT TAT GTT TG	
STR	BAC_STR fw	TTG AAT GGC ACT TCT TGC ACA CT	150 bp
	BAC_STR rev	ACC ATT GTG TGG GAG GAC ATA TGA TA	
SGD	BAC_SGD (2) fw	ATG GTG TGA ATG TAA AAG GAT TCT TTG TT	362 bp
	BAC_SGD (2) rev	TTG CTT ATT GAA GTA CCT TAA AGA GCG GT	

7.8.7 Testing BAC integrity

An *E. coli* LB stab of each of the by Bio S&T Inc. positively identified BAC clone was received together with 20 µl of a plasmid DNA of the same clone as positive control. Before large scale BAC plasmid extraction, the integrity of the received BACs was tested by colony PCR by restriction digestion with NotI.

For each clone the LB stab material was transferred to a selective plate (LB agar with 12.5 mg/l chloramphenicol) and grown overnight at 37°C to obtain single colonies. From those colonies 10 ml overnight liquid cultures were prepared (LB with 12.5 mg/l chloramphenicol). The overnight cultures were tested by PCR to verify the right insert using the same gene specific primers as for the screening of the BAC library (Table 44). Glycerol stocks were made of positive cultures.

Starting either from glycerol stocks with a fresh 10 ml overnight culture or from the overnight culture directly, a protocol modified after [265] was employed. Briefly after spinning down the culture (3000 rpm, 10 minutes at 4°C) the obtained pellet was resuspended in 200 µl resuspension buffer P1 (Qiagen, cat. No. 19051, without added RNAse) and transferred to a 1.5 ml Eppendorf tube. The cells were lysed with 400 µl of Buffer P2 (Qiagen, cat. No. 19052) and neutralised by adding 300 µl of buffer P3 (Qiagen, cat. No. 19053). After centrifuging (3000 rpm, 10 minutes at 4°C), the supernatant was transferred to a new Eppendorf tube and to precipitate the DNA, 600 µl of 100% isopropanol were added, mixed by inversion and subsequently centrifuged at high speed (13,000 rpm, 10 minutes at 4°C). The supernatant was removed and the pellet was washed with 500 µl 70% EtOH and centrifuged again. After removal of the supernatant the pellet was dried using a Savant SC210A speed vac concentrator for 15 minutes at medium setting. The pellet was resuspended in 50 µl LB buffer and left at room temperature for one hour. RNA was removed by adding 2 µl 10 mg/ml RNAase (RNAase A, Qiagen, cat. No. 19101) for one hour at 37°C. The DNA concentration of each extraction was measured on a nanodrop spectrophotometer (NanoDrop, ND 1000).

It was found that the yield for each BAC and each individual extraction was highly variable and the obtained material amount often insufficient for downstream applications. Additionally glycerol stocks of previously positive colonies did not maintain the plasmid consistently. Therefore the small scale extraction was optimised by starting two to three 10 ml overnight cultures from the same initial colony. To avoid loss of plasmid, enough material from the 10 ml cultures was kept to start one or two 500 ml cultures per construct, depending on yield of individual candidates, for the large scale BAC plasmid extraction.

7.8.8 Large scale BAC plasmid extraction for sequencing

Once it was confirmed the BAC culture contained the expected BAC and that the integrity of the BAC had not been compromised, one or two 500 ml cultures for each candidate were started and grown overnight for 12 to 16 hours. For extraction of the plasmid, the Large-

Construction Kit (Qiagen, cat. No. 12462) was used according to manufacturer's instructions. After pelleting the bacteria and alkaline lysis of the pellet, the DNA was precipitated with isopropanol and washed with 70% ethanol before being treated with ATP-dependent exonuclease to minimise contamination with genomic DNA and nicked BAC plasmids. Instead of the supplied Qiagen-tip 20 columns, the Qiagen-tip 500 columns (Qiagen, cat. No. 10262) were used for efficient binding, washing and subsequent elution of BAC DNA in this large scale application. The eluted DNA was precipitated a second time with isopropanol, centrifuged and the resulting pellet washed with 70% ethanol, dried and finally resuspended in TE buffer and concentration of DNA was determined using a nanodrop spectrophotometer (NanoDrop, ND 1000).

7.8.9 BAC sequencing and assembly

BAC sequencing was conducted by the group of Robin Buell from Michigan State University (MSU, East Lansing, USA). Briefly, from the obtained BAC DNA a single TruSeq DNA library was constructed and sequenced on a MiSeq (150-nt or 250-nt paired end reads).

After sequencing data was received, the BAC assemblies were performed using MIRA, version 4.0rc4 (http://sourceforge.net/projects/mira-assembler). The total amount of sequencing data was reduced for optimal assembly and only the first 50 000 read pairs of each dataset were used for further processing. Each read was aligned to the vector sequence (EU140754.1) and the genome sequence of *E. coli* strand K12, sub-strand DH10B (CP000948.1). Only reads with a BLASTN hit [266] with an e-value less than 1E-10 to the vector within 1 kb distance to the restriction site and reads without any hit to *E. coli* or vector were used in the assembly. Remaining vector parts were clipped from the assembled scaffolds. The assembly was repeated using the first 100 000 read pairs but this did not change the assembly result significant and was there for discarded. The assembly data for the *T16H* BAC and the *SGD* BAC are deposited to NCBI under ERX556143 and ERX556142 as part of the publication of the *C. roseus* genome [68].

8 References

1. Osbourn, A. E. & Lanzotti, V. *Plant-derived Natural Products*. (Springer US, 2009).
2. Balunas, M. J. & Kinghorn, A. D. Drug discovery from medicinal plants. *Life Sci.* **78,** 431–441 (2005).
3. Van Der Heijden, R., Verpoorte, R. & Ten Hoopen, H. J. G. Cell and tissue cultures of Catharanthus roseus (L.) G. Don: a literature survey. *Plant Cell. Tissue Organ Cult.* **18,** 231–280 (1989).
4. Wilson, S. A. & Roberts, S. C. Recent advances towards development and commercialization of plant cell culture processes for the synthesis of biomolecules. *Plant Biotechnol. J.* **10,** 249–68 (2012).
5. Mora-Pale, M., Sanchez-Rodriguez, S. P., Linhardt, R. J., Dordick, J. S. & Koffas, M. A. G. Metabolic engineering and in vitro biosynthesis of phytochemicals and non-natural analogues. *Plant Sci.* **210,** 10–24 (2013).
6. Sumner, L. W., Lei, Z., Nikolau, B. J. & Saito, K. Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. *Nat. Prod. Rep.* **32,** 212–29 (2015).
7. Nützmann, H.-W. & Osbourn, A. Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **26,** 91–99 (2014).
8. Schatz, M. C., Witkowski, J. & McCombie, W. R. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* **13,** 243 (2012).
9. Harborne, J. B. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2001).
10. Wink, M. Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theor. Appl. Genet.* **75,** 225–233 (1988).
11. Wink, M. in *Alkaloids* 265–300 (Springer US, 1998).
12. Dewick, P. M. *Medicinal Natural Products: A Biosynthetic Approach: Third Edition.* (John Wiley & Sons, Ltd, 2009).
13. Fester, K. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2010).
14. Lipp, J. Possible mechanisms of morphine analgesia. *Clin. Neuropharmacol.* **14,** 131–47 (1991).
15. Ruetsch, Y. A., Böni, T. & Borgeat, A. From cocaine to ropivacaine: the history of local anesthetic drugs. *Curr. Top. Med. Chem.* **1,** 175–82 (2001).
16. Achan, J. *et al.* Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. *Malar. J.* **10,** 144–152 (2011).
17. Chen, K. K. & Schmidt, C. F. The action and clinical use of ephedrine, an alkaloid isolated from the Chinese drug ma huang; historical document. *Ann. Allergy* **17,** 605–18
18. O'Connor, S. E. & Maresh, J. J. Chemistry and biology of monoterpene indole alkaloid biosynthesis. *Nat. Prod. Rep.* **23,** 532–47 (2006).

19. van Der Heijden, R., Jacobs, D. I., Snoeijer, W., Hallard, D. & Verpoorte, R. The Catharanthus alkaloids: pharmacognosy and biotechnology. *Curr. Med. Chem.* **11,** 607–28 (2004).

20. Potier, P. In: History of the discovery of navelbine. Navelbine (vinorelbine). *Updat. New Trends* **3,** 3–8 (1990).

21. Barnett, C. J. *et al.* Structure-activity relationships of dimeric Catharanthus alkaloids. 1. Deacetylvinblastine amide (vindesine) sulfate. *J. Med. Chem.* **21,** 88–96 (1978).

22. Jacquesy, J. C. Reactivity of Vinca alkaloids in superacid. An access to vinflunine, a novel anticancer agent. *J. Fluor. Chem.* **127,** 1484–1487 (2006).

23. Mamtani, R. & Vaughn, D. J. Vinflunine in the treatment of advanced bladder cancer. *Expert Rev. Anticancer Ther.* **11,** 13–20 (2011).

24. Nagakura, N., Rüffer, M. & Zenk, M. H. The biosynthesis of monoterpenoid indole alkaloids from strictosidine. *J. Chem. Soc., Perkin Trans. 1* 2308–2312 (1979).

25. 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. U. S. A.* **112,** 3205–10 (2015).

26. Contin, A., van der Heijden, R., Lefeber, A. W. . & Verpoorte, R. The iridoid glucoside secologanin is derived from the novel triose phosphate/pyruvate pathway in a Catharanthus roseus cell culture. *FEBS Lett.* **434,** 413–416 (1998).

27. Simkin, A. J. *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* **85,** 36–43 (2013).

28. Collu, G. *et al.* Geraniol 10-hydroxylase, a cytochrome P450 enzyme involved in terpenoid indole alkaloid biosynthesis. *FEBS Lett.* **508,** 215–20 (2001).

29. Miettinen, K. *et al.* The seco-iridoid pathway from Catharanthus roseus. *Nat. Commun.* **5,** 3606 (2014).

30. Geu-Flores, F. *et al.* An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. *Nature* **492,** 138–42 (2012).

31. 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* **101,** 23–31 (2014).

32. Asada, K. *et al.* A 7-deoxyloganetic acid glucosyltransferase contributes a key step in secologanin biosynthesis in Madagascar periwinkle. *Plant Cell* **25,** 4123–34 (2013).

33. 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.* **76,** 754–765 (2013).

34. Levac, D., Murata, J., Kim, W. S. & 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 (2008).

35. Irmler, S. *et al.* Indole alkaloid biosynthesis in Catharanthus roseus: New enzyme activities and identification of cytochrome P450 CYP72A1 as secologanin synthase. *Plant J.* **24,** 797–804 (2000).

36. 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. U. S. A.* **86,** 2582–2586 (1989).

37. 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).

38. Money, T., Wright, I. & McCapra, F. Biosynthesis of indole alkaloids. Vindoline. *J. Am. Chem. Soc.* 4144–4150 (1968).

39. Battersby, A. R., Burnett, A. R. & Parsons, P. G. Alkaloid biosynthesis. Part XV. Partial synthesis and isolation of vincoside and isovincoside: biosynthesis of the three major classes of indole alkaloids from vincoside. *J. Chem. Soc. C Org.* **8,** 1193 (1969).

40. Balsevich, J., De Luca, V. & Kurz, W. G. W. Altered alkaloid pattern in dark grown seedlings of Catharanthus roseus. The isolation and characterization of 4-desacetoxyvindoline: A novel indole alkaloid and proposed precursor of vindoline. *Heterocycles* **24,** 2415–2421 (1986).

41. Schröder, G., Unterbusch, E. & Kaltenbach, M. Light-induced cytochrome P450-dependent enzyme in indole alkaloid biosynthesis: tabersonine 16-hydroxylase. *FEBS Lett.* **458,** 97–102 (1999).

42. Besseau, S. *et al.* A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in Catharanthus roseus. *Plant Physiol.* **163,** 1792–803 (2013).

43. Liscombe, D. K., Usera, A. R. & O'Connor, S. E. Homolog of tocopherol C methyltransferases catalyzes N methylation in anticancer alkaloid biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 18793–18798 (2010).

44. Vazquez-Flota, F., De Carolis, E., Alarco, A. M. & 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.* **34,** 935–48 (1997).

45. St-Pierre, B., Laflamme, P., Alarco, a M. & 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.* **14,** 703–13 (1998).

46. Yu, F. & De Luca, V. ATP-binding cassette transporter controls leaf surface secretion of anticancer drug components in Catharanthus roseus. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15830–5 (2013).

47. Murata, J. & De Luca, V. 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 J.* **44,** 581–94 (2005).

48. 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.* **112,** 6224–6229 (2015).

49. Courdavault, V. *et al.* A look inside an alkaloid multisite plant: the Catharanthus logistics. *Curr. Opin. Plant Biol.* **19C,** 43–50 (2014).

50. Murata, J., Roepke, J., Gordon, H. & De Luca, V. The leaf epidermome of Catharanthus roseus reveals its biochemical specialization. *Plant Cell* **20,** 524–542 (2008).

51. Lee-Parsons, C. W. T. & Royce, A. J. Precursor limitations in methyl jasmonate-induced Catharanthus roseus cell cultures. *Plant Cell Rep.* **25,** 607–12 (2006).

52. He, L., Yang, L., Tan, R., Zhao, S. & Hu, Z. Enhancement of vindoline production in suspension culture of the Catharanthus roseus cell line C20hi by light and methyl jasmonate elicitation. *Anal. Sci.* **27,** 1243–8 (2011).

53. Bones, A. M. & Rossiter, J. T. The enzymic and chemically induced decomposition of glucosinolates. *Phytochemistry* **67,** 1053–67 (2006).

54. Guirimand, G. *et al.* Strictosidine activation in Apocynaceae: towards a 'nuclear time bomb' ? *BMC Plant Biol.* **10,** 182–192 (2010).

55. Chaojun, L. *et al.* Four Botanical Extracts are Toxic to the Hispine Beetle, Brontispa longissima, in Laboratory and Semi—field Trials. *J. Insect Sci.* **12,** 1–8 (2012).

56. Luijendijk, T. J., van der Meijden, E. & Verpoorte, R. Involvement of strictosidine as a defensive chemical in Catharanthus roseus. *J. Chem. Ecol.* **22,** 1355–66 (1996).

57. Roepke, J. *et al.* Vinca drug components accumulate exclusively in leaf exudates of Madagascar periwinkle. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 15287–15292 (2010).

58. Parkinson, J. & Blaxter, M. Expressed sequence tags: an overview. *Methods Mol. Biol.* **533,** 1–12 (2009).

59. Murata, J., Bienzle, D., Brandle, J. E., Sensen, C. W. & De Luca, V. Expressed sequence tags from Madagascar periwinkle (Catharanthus roseus). *FEBS Lett.* **580,** 4501–4507 (2006).

60. 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* **11,** 887–900 (1999).

61. Cacace, S. *et al.* A flavonol O-methyltransferase from Catharanthus roseus performing two sequential methylations. *Phytochemistry* **62,** 127–37 (2003).

62. Schröder, G. *et al.* Flavonoid methylation: A novel 4'-O-methyltransferase from Catharanthus roseus, and evidence that partially methylated flavanones are substrates of four different flavonoid dioxygenases. *Phytochemistry* **65,** 1085–1094 (2004).

63. Huang, L. *et al.* Molecular characterization of the pentacyclic triterpenoid biosynthetic pathway in Catharanthus roseus. *Planta* **236,** 1571–81 (2012).

64. Yu, F. *et al.* Functional characterization of amyrin synthase involved in ursolic acid biosynthesis in Catharanthus roseus leaf epidermis. *Phytochemistry* **91,** 122–127 (2013).

65. Góngora-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).

66. Runguphan, W., Maresh, J. J. & O'Connor, S. E. Silencing of tryptamine biosynthesis for production of nonnatural alkaloids in plant culture. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 13673–13678 (2009).

67. Glenn, W. S., Nims, E. & O'Connor, S. E. Reengineering a Tryptophan Halogenase To Preferentially Chlorinate a Direct Alkaloid Precursor. *J. Am. Chem. Soc.* **133,** 19346–19349 (2011).

68. Kellner, F. *et al.* Genome-guided investigation of plant natural product biosynthesis. *Plant J.* **82,** 680–692 (2015).

69. Van Moerkercke, A. *et al.* CathaCyc, a metabolic pathway database built from catharanthus roseus RNA-seq data. *Plant Cell Physiol.* **54,** 673–685 (2013).

70. Xiao, M. *et al.* Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *J. Biotechnol.* **166,** 122–34 (2013).

71. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345,** 1181–4 (2014).

72. Straub, S. C. *et al.* Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. *BMC Genomics* **12,** 211 (2011).

73. Ku, C., Chung, W.-C., Chen, L.-L. & Kuo, C.-H. The Complete Plastid Genome Sequence of Madagascar Periwinkle Catharanthus roseus (L.) G. Don: Plastid Genome Evolution, Molecular Marker Identification, and Phylogenetic Implications in Asterids. *PLoS One* **8,** e68518 (2013).

74. Movahedi, S., Van Bel, M., Heyndrickx, K. S. & Vandepoele, K. Comparative co-expression analysis in plant biology. *Plant. Cell Environ.* **35,** 1787–98 (2012).

75. Wolfe, C. J., Kohane, I. S. & Butte, A. J. Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks. *BMC Bioinformatics* **6,** 227 (2005).

76. Vanderauwera, S. *et al.* Genome-Wide Analysis of Hydrogen Peroxide-Regulated Gene Expression in Arabidopsis Reveals a High Light-Induced Transcriptional Cluster Involved in Anthocyanin Biosynthesis. *Plant Physiol.* **139,** 806–821 (2005).

77. Ginglinger, J.-F. *et al.* Gene Coexpression Analysis Reveals Complex Metabolism of the Monoterpene Alcohol Linalool in Arabidopsis Flowers. *Plant Cell* **25,** 4640–4657 (2013).

78. Wada, M. *et al.* Prediction of operon-like gene clusters in the Arabidopsis thaliana genome based on co-expression analysis of neighboring genes. *Gene* **503,** 56–64 (2012).

79. Giddings, L. A. *et al.* A stereoselective hydroxylation step of alkaloid biosynthesis by a unique cytochrome P450 in Catharanthus roseus. *J. Biol. Chem.* **286,** 16751–16757 (2011).

80. Liu, D. *et al.* Enhanced accumulation of catharanthine and vindoline in Catharanthus roseus hairy roots by overexpression of transcriptional factor ORCA2. *J. Biotechnol.* **10,** 3260–3268 (2011).

81. Morris, P., Rudge, K., Cresswell, R. & Fowler, M. W. Regulation of product synthesis in cell cultures of Catharanthus roseus. V. Long-term maintenance of cells on a production medium. *Plant Cell. Tissue Organ Cult.* **17,** 79–90 (1989).

82. Baulcombe, D. C. Fast forward genetics based on virus-induced gene silencing. *Curr. Opin. Plant Biol.* **2,** 109–13 (1999).

83. Donaire, L. *et al.* Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs. *J. Virol.* **82,** 5167–77 (2008).

84. Baumberger, N. & Baulcombe, D. C. Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 11928–33 (2005).

85. Ramegowda, V., Mysore, K. S. & Senthil-Kumar, M. Virus-induced gene silencing is a versatile tool for unraveling the functional relevance of multiple abiotic-stress-responsive genes in crop plants. *Front. Plant Sci.* **5,** 323 (2014).

86. Scholthof, K. B. G. *et al.* Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology* **12,** 938–954 (2011).

87. Ratcliff, F., Martin-Hernandez, A. M. & Baulcombe, D. C. Technical Advance. Tobacco rattle virus as a vector for analysis of gene function by silencing. *Plant J.* **25,** 237–45 (2001).

88. Swanson, M., Barker, H. & MacFarlane, S. A. Rapid vascular movement of tobraviruses does not require coat protein: evidence from mutated and wild-type viruses. *Ann. Appl. Biol.* **141,** 259–266 (2002).

89. Liu, Y., Schiff, M., Marathe, R. & Dinesh-Kumar, S. P. Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. *Plant J.* **30,** 415–29 (2002).

90. Liu, Y., Schiff, M. & Dinesh-Kumar, S. P. Virus-induced gene silencing in tomato. *Plant J.* **31,** 777–86 (2002).

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

92. Brueggeman, R. *et al.* The stem rust resistance gene Rpg5 encodes a protein with nucleotide-binding-site, leucine-rich, and protein kinase domains. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 14970–5 (2008).

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

94. Carqueijeiro, I. *et al.* Virus-induced gene silencing in *Catharanthus roseus* by biolistic inoculation of tobacco rattle virus vectors. *Plant Biol.* **17,** 1242–1246 (2015).

95. Becker, A. & Lange, M. VIGS – genomics goes functional. *Trends Plant Sci.* **15,** 1–4 (2010).

96. Jacob, F., Perrin, D., Sanches, C. & Monod, J. Operon: a group of genes with the expression coordinated by an operator. *Comptes rendus Hebd. des séances l'Académie des Sci.* **250,** 1727–9 (1960).

97. Hall, C. & Dietrich, F. S. The Reacquisition of Biotin Prototrophy in Saccharomyces cerevisiae Involved Horizontal Gene Transfer, Gene Duplication and Gene Clustering. *Genetics* **177,** 2293–2307 (2007).

98. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5,** 889–899 (2004).

99. Keller, N. P. Translating biosynthetic gene clusters into fungal armor and weaponry. *Nat. Chem. Biol.* **11,** 671–7 (2015).

100. Chu, H. Y., Wegel, E. & Osbourn, A. From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J.* **66,** 66–79 (2011).

101. Takos, A. M. & Rook, F. Why biosynthetic genes for chemical defense compounds cluster. *Trends Plant Sci.* **17,** 383–388 (2012).

102. Osbourn, A. E. & Field, B. Operons. *Cell. Mol. Life Sci.* **66,** 3755–75 (2009).

103. Osbourn, A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet.* **26,** 449–57 (2010).

104. Osbourn, A. Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology. *PLANT Physiol.* **154,** 531–535 (2010).

105. Boycheva, S., Daviet, L., Wolfender, J.-L. & Fitzpatrick, T. B. The rise of operon-like gene clusters in plants. *Trends Plant Sci.* **19,** 447–459 (2014).

106. Fernie, A. R. & Tohge, T. Location, location, location--no more! The unravelling of chromatin remodeling regulatory aspects of plant metabolic gene clusters. *New Phytol.* **205,** 458–60 (2015).

107. Schneider, L. M. *et al.* The Cer-cqu gene cluster determines three key players in a β-diketone synthase polyketide pathway synthesizing aliphatics in epicuticular waxes. *J. Exp. Bot.* **67,** 2715–2730 (2016).

108. Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B. & Peters, R. J. CYP76M7 Is an ent-Cassadiene C11 -Hydroxylase Defining a Second Multifunctional Diterpenoid Biosynthetic Gene Cluster in Rice. *Plant Cell* **21,** 3315–3325 (2009).

109. Wang, Q. *et al.* Characterization of CYP76M5-8 indicates metabolic plasticity within a plant biosynthetic gene cluster. *J. Biol. Chem.* **287,** 6159–68 (2012).

110. Wilderman, P. R., Xu, M., Jin, Y., Coates, R. M. & Peters, R. J. Identification of syn-pimara-7,15-diene synthase reveals functional clustering of terpene synthases involved in rice phytoalexin/allelochemical biosynthesis. *Plant Physiol.* **135,** 2098–105 (2004).

111. Frey, M. Analysis of a Chemical Plant Defense Mechanism in Grasses. *Science (80-.).* **277,** 696–699 (1997).

112. Winzer, T. *et al.* A Papaver somniferum 10-Gene Cluster for Synthesis of the Anticancer Alkaloid Noscapine. *Science (80-.).* **336,** 1704–1708 (2012).

113. Matsuba, Y. *et al.* Evolution of a complex locus for terpene biosynthesis in solanum. *Plant Cell* **25,** 2022–36 (2013).

114. Pichersky, E. & Lewinsohn, E. Convergent Evolution in Plant Specialized Metabolism. *Annu. Rev. Plant Biol.* **62,** 549–566 (2011).

115. Takos, A. M. *et al.* Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in Lotus japonicus and suggests the repeated evolution of this chemical defence pathway. *Plant J.* **68,** 273–86 (2011).

116. Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J. & Osbourn, a E. Compromised disease resistance in saponin-deficient plants. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 12923–12928 (1999).

117. Qi, X. *et al.* A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 8233–8 (2004).

118. Qi, X. *et al.* A different function for a member of an ancient and highly conserved cytochrome P450 family: from essential sterols to plant defense. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 18848–53 (2006).

119. Geisler, K. *et al.* Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E3360-7 (2013).

120. Mugford, S. T. *et al.* Modularity of plant metabolic gene clusters: a trio of linked genes that are collectively required for acylation of triterpenes in oat. *Plant Cell* **25,** 1078–92 (2013).

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

122. Hiriart, J.-B., Lehto, K., Tyystjärvi, E., Junttila, T. & Aro, E.-M. Suppression of a key gene involved in chlorophyll biosynthesis by means of virus-inducing gene silencing. *Plant Mol. Biol.* **50,** 213–24 (2002).

123. Geu-Flores, F., Nour-Eldin, H. H., Nielsen, M. T. & Halkier, B. A. USER fusion: a rapid and efficient method for simultaneous fusion and cloning of multiple PCR products. *Nucleic Acids Res.* **35,** e55 (2007).

124. 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.* **109,** 131–139 (1995).

125. Magnotta, M., Murata, J., Chen, J. & De Luca, V. Identification of a low vindoline accumulating cultivar of Catharanthus roseus (L.) G. Don by alkaloid and enzymatic profiling. *Phytochemistry* **67,** 1758–64 (2006).

126. Chung, I.-M. *et al.* Screening 64 cultivars Catharanthus roseus for the production of vindoline, catharanthine, and serpentine. *Biotechnol. Prog.* **27,** 937–43 (2011).

127. De Luca, V., Balsevich, J., Tyler, R. T. & Kurz, W. G. Characterization of a novel N-methyltransferase (NMT) from Catharanthus roseus plants : Detection of NMT and other enzymes of the indole alkaloid biosynthetic pathway in different cell suspension culture systems. *Plant Cell Rep.* **6,** 458–61 (1987).

128. Moza, B. K. & Trojánek, J. On alkaloids. VIII. Structure of vindorosine. *Collect. Czechoslov. Chem. Commun.* **28,** 1427–1433 (1963).

129. Wolf, J. B. W. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol. Ecol. Resour.* **13,** 559–572 (2013).

130. Wei, H. *et al.* Transcriptional Coordination of the Metabolic Network in Arabidopsis. *PLANT Physiol.* **142,** 762–774 (2006).

131. Gholami, A., De Geyter, N., Pollier, J., Goormachtig, S. & Goossens, A. Natural product biosynthesis in Medicago species. *Nat. Prod. Rep.* **31,** 356–380 (2014).

132. Guirimand, G. *et al.* Spatial organization of the vindoline biosynthetic pathway in Catharanthus roseus. *J. Plant Physiol.* **168,** 549–557 (2011).

133. De Luca, V. & Cutler, A. J. Subcellular Localization of Enzymes Involved in Indole Alkaloid Biosynthesis in Catharanthus roseus. *Plant Physiol.* **85,** 1099–1102 (1987).

134. Kuboyama, T., Yokoshima, S., Tokuyama, H. & Fukuyama, T. Stereocontrolled total synthesis of (+)-vincristine. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 11966–70 (2004).

135. Kobayashi, S., Ueda, T. & Fukuyama, T. An efficient total synthesis of (-)-vindoline. *Synlett* **2000,** 0883–0886 (2000).

136. Danieli, B., Lesma, G., Palmisano, G. & Riva, R. Rapid access to the highly oxygenated aspidosperma alkaloids vindoline, vindorosine, and cathovaline. *J. Chem. Soc. Chem. Commun.* 909 (1984).

137. Kuehne, M. E., Podhorez, D. E., Mulamba, T. & Bornmann, W. G. Biomimetic alkaloid syntheses. 15. Enantioselective syntheses with epichlorohydrin: total syntheses of (+)-, (-)- and (.+-.)-vindoline and a synthesis of (-)-vindorosine. *J. Org. Chem.* **52,** 347–353 (1987).

138. Thibodeaux, C. J., Chang, W. C. & Liu, H. W. Enzymatic chemistry of cyclopropane, epoxide, and aziridine biosynthesis. *Chemical Reviews* **112,** 1681–1709 (2012).

139. Xu, R. & WunschII, D. Survey of Clustering Algorithms. *IEEE Trans. Neural Networks* **16,** 645–678 (2005).

140. Gachon, C. M. M., Langlois-Meurinne, M., Henry, Y. & Saindrenan, P. Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. *Plant Mol. Biol.* **58,** 229–245 (2005).

141. Aerts, R. J., Gisi, D., Carolis, E., Luca, V. & Baumann, T. W. Methyl jasmonate vapor increases the developmentally controlled synthesis of alkaloids in Catharanthus and Cinchona seedlings. *Plant J.* **5,** 635–643 (1994).

142. Han, J. *et al.* The cytochrome P450 CYP86A22 is a fatty acyl-CoA hydroxylase essential for estolide synthesis in the stigma of Petunia hybrida. *J. Biol. Chem.* **285,** 3986–3996 (2010).

143. Urban, P., Mignotte, C., Kazmaier, M., Delorme, F. & Pompon, D. Cloning, Yeast Expression, and Characterization of the Coupling of Two Distantly Related Arabidopsis thalianaNADPH-Cytochrome P450 Reductases with P450 CYP73A5. *J. Biol. Chem.* **272,** 19176–19186 (1997).

144. Fang, F. *et al.* A vector set for systematic metabolic engineering in Saccharomyces cerevisiae. *Yeast* **28,** 123–36 (2011).

145. Pompon, D., Louerat, B., Bronine, A. & Urban, P. in *Methods in enzymology* **272,** 51–64 (1996).

146. Kellner, F. *et al.* Discovery of a P450-catalyzed step in vindoline biosynthesis: a link between the aspidosperma and eburnamine type alkaloid scaffolds. *Chem. Commun.* **51,** 7626–7628 (2015).

147. Kozmin, S. A., Iwama, T., Huang, Y. & Rawal, V. H. An efficient approach to Aspidosperma alkaloids via [4 + 2] cycloadditions of aminosiloxydienes: stereocontrolled total synthesis of (+/-)-tabersonine. *J. Am. Chem. Soc.* **124,** 4628–41 (2002).

148. Danieli, B., Lesma, G., Palmisano, G. & Gabetta, B. Ozonation in alkoloid chemistry: a mild and selective conversion of vincadifformine into vincamine. *J. Chem. Soc. Chem. Commun.* 908 (1981).

149. Calabi, L., Danieli, B., Lesma, G. & Palmisano, G. Dye-sensitized photo-oxygenation of the Aspidosperma alkaloids vincadifformine and tabersonine. A new, convenient approach to vincamine. *J. Chem. Soc. Perkin Trans. 1* 1371 (1982).

150. Wenkert, E. & Wickberg, B. General Methods of Synthesis of Indole Alkaloids. IV. A Synthesis of dl-Eburnamonine 1,2. *J. Am. Chem. Soc.* **87,** 1580–1589 (1965).

151. Kutney, J. P., Beck, J. F., Nelson, V. R. & Sood, R. S. Indole alkaloid biosynthesis. VI. Eburnamie-vincamine alkaloids. *J. Am. Chem. Soc.* **93,** 255–257 (1971).

152. Vinpocetine. Monograph. *Altern. Med. Rev.* **7,** 240–3 (2002).

153. Moudi, M., Go, R., Yien, C. Y. S. & Nazre, M. Vinca alkaloids. *International Journal of Preventive Medicine* **4,** 1131–1135 (2013).

154. Rajniak, J., Barco, B., Clay, N. K. & Sattely, E. S. A new cyanogenic metabolite in Arabidopsis required for inducible pathogen defence. *Nature* **525,** 376–9 (2015).

155. Burke, D., Carle, G. & Olson, M. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science (80-.).* **236,** 806–812 (1987).

156. Anderson, C. Genome shortcut leads to problems. *Science* **259,** 1684–7 (1993).

157. O'Connor, M., Peifer, M. & Bender, W. Construction of large DNA segments in Escherichia coli. *Science* **244,** 1307–12 (1989).

158. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci.* **89,** 8794–8797 (1992).

159. Schulte, D. *et al.* BAC library resources for map-based cloning and physical map construction in barley (Hordeum vulgare L.). *BMC Genomics* **12,** 247 (2011).

160. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408,** 796–815 (2000).

161. Wang, Y. *et al.* Chromosome walking in the Petunia inflata self-incompatibility (S-) locus and gene identification in an 881-kb contig containing S2-RNase. *Plant Mol. Biol.* **54,** 727–42 (2004).

162. Michael, T. P. & Jackson, S. The First 50 Plant Genomes. *Plant Genome* **6,** 34–45 (2013).

163. Leushkin, E. V *et al.* The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. *BMC Genomics* **14,** 476 (2013).

164. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497,** 579–84 (2013).

165. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–5 (2009).

166. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 635–41 (2012).

167. International Barley Genome Sequencing Consortium *et al.* A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491,** 711–6 (2012).

168. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452,** 991–6 (2008).

169. Guimaraes, G. *et al.* Cytogenetic characterization and genome size of the medicinal plant Catharanthus roseus (L.) G. Don. *AoB Plants* **2012,** pls002-pls002 (2012).

170. Baker, M. De novo genome assembly: what every biologist should know. *Nat. Methods* **9,** 333–337 (2012).

171. Huang, S. *et al.* The genome of the cucumber, Cucumis sativus L. *Nat. Genet.* **41,** 1275–1281 (2009).

172. van Bakel, H. *et al.* The draft genome and transcriptome of Cannabis sativa. *Genome Biol.* **12,** R102 (2011).

173. Hamilton, J. P. & Robin Buell, C. Advances in plant genome sequencing. *Plant J.* **70,** 177–190 (2012).

174. Nützmann, H.-W., Huang, A. & Osbourn, A. Plant metabolic clusters - from genetics to genomics. *New Phytol.* (2016).

175. Krokida, A. *et al.* A metabolic gene cluster in Lotus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytol.* **200,** 675–90 (2013).

176. Field, B. & Osbourn, A. E. Metabolic diversification--independent assembly of operon-like gene clusters in different plants. *Science* **320,** 543–7 (2008).

177. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci.* **108,** 16116–16121 (2011).

178. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–1123 (2009).

179. Chevreux, B. Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Res.* **14,** 1147–1159 (2004).

180. Höfer, R. *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.* **20,** 221–32 (2013).

181. Munkert, J. *et al.* Iridoid Synthase Activity Is Common among the Plant Progesterone 5β-Reductase Family. *Mol. Plant* **8,** 136–52 (2015).

182. Schenkman, J. B. & Jansson, I. The many roles of cytochrome b5. *Pharmacol. Ther.* **97,** 139–152 (2003).

183. Li, Y., Baldauf, S., Lim, E.-K. & Bowles, D. J. Phylogenetic Analysis of the UDP-glycosyltransferase Multigene Family of Arabidopsis thaliana. *J. Biol. Chem.* **276,** 4338–4343 (2001).

184. Barvkar, V. T., Pardeshi, V. C., Kale, S. M., Kadoo, N. Y. & Gupta, V. S. Phylogenomic analysis of UDP glycosyltransferase 1 multigene family in Linum usitatissimum identified genes with varied expression patterns. *BMC Genomics* **13,** 175 (2012).

185. Tang, H. *et al.* Synteny and Collinearity in Plant Genomes. *Science (80-.).* **320,** 486–488 (2008).

186. Dittrich, H. & Kutchan, T. M. Molecular cloning, expression, and induction of berberine bridge enzyme, an enzyme essential to the formation of benzophenanthridine alkaloids in the response of plants to pathogenic attack. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 9969–73 (1991).

187. Pennings, E. J. M., Groen, B. W., Duine, J. A. & Verpoorte, R. Tryptophan decarboxylase from Catharanthus roseus is a pyridoxoquinoprotein. *FEBS Lett.* **255,** 97–100 (1989).

188. Treimer, J. F. & Zenk, M. H. Purification and Properties of Strictosidine Synthase, the Key Enzyme in Indole Alkaloid Formation. *Eur. J. Biochem.* **101,** 225–233 (1979).

189. Maresh, J. J. *et al.* Strictosidine Synthase: Mechanism of a Pictet−Spengler Catalyzing Enzyme †. *J. Am. Chem. Soc.* **130,** 710–723 (2008).

190. Brown, M. H., Paulsen, I. T. & Skurray, R. A. The multidrug efflux protein NorM is a prototype of a new family of transporters. *Mol. Microbiol.* **31,** 394–5 (1999).

191. Morita, M. *et al.* Vacuolar transport of nicotine is mediated by a multidrug and toxic compound extrusion (MATE) transporter in Nicotiana tabacum. *Proc. Natl. Acad. Sci.* **106,** 2447–2452 (2009).

192. Gibbs, A. J. & McIntyre, G. A. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* **16,** 1–11 (1970).

193. Claros, M. G. *et al.* Why Assembling Plant Genome Sequences Is So Challenging. *Biology (Basel).* **1,** 439–459 (2012).

194. Moeder, W., Del Pozo, O., Navarre, D. a, Martin, G. B. & Klessig, D. F. Aconitase plays a role in regulating resistance to oxidative stress and cell death in Arabidopsis and Nicotiana benthamiana. *Plant Mol. Biol.* **63,** 273–287 (2007).

195. Geerlings, A. Molecular Cloning and Analysis of Strictosidine beta -D-Glucosidase, an Enzyme in Terpenoid Indole Alkaloid Biosynthesis in Catharanthus roseus. *J. Biol. Chem.* **275,** 3051–3056 (2000).

196. Opassiri, R. *et al.* Analysis of rice glycosyl hydrolase family 1 and expression of Os4bglu12 beta-glucosidase. *BMC Plant Biol.* **6,** 33 (2006).

197. Lubkowitz, M. The Oligopeptide Transporters: A Small Gene Family with a Diverse Group of Substrates and Functions? *Mol. Plant* **4,** 407–415 (2011).

198. Ziegler, J. & Facchini, P. J. Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.* **59,** 735–69 (2008).

199. Carqueijeiro, I., Noronha, H., Duarte, P., Gerós, H. & Sottomayor, M. Vacuolar transport of the medicinal alkaloids from Catharanthus roseus is mediated by a proton-driven antiport. *Plant Physiol.* **162,** 1486–96 (2013).

200. Koh, S. *et al.* An oligopeptide transporter gene family in Arabidopsis. *Plant Physiol.* **128,** 21–29 (2002).

201. Farrow, S. C. & Facchini, P. J. Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. *Front. Plant Sci.* **5,** (2014).

202. Nelson, D. R., Ming, R., Alam, M. & Schuler, M. A. Comparison of Cytochrome P450 Genes from Six Plant Genomes. *Trop. Plant Biol.* **1,** 216–235 (2008).

203. Henry, L. K., Gutensohn, M., Thomas, S. T., Noel, J. P. & Dudareva, N. Orthologs of the archaeal isopentenyl phosphate kinase regulate terpenoid production in plants. *Proc. Natl. Acad. Sci.* 201504798 (2015).

204. Rijhwani, S. K. & Shanks, J. V. Effect of elicitor dosage and exposure time on biosynthesis of indole alkaloids by Catharanthus roseus hairy root cultures. *Biotechnol. Prog.* **14,** 442–9

205. Rodriguez, S., Compagnon, V., Crouch, N. P., St-Pierre, B. & De Luca, V. Jasmonate-induced epoxidation of tabersonine by a cytochrome P-450 in hairy root cultures of Catharanthus roseus. *Phytochemistry* **64,** 401–409 (2003).

206. Laflamme, P., St-Pierre, B. & De Luca V. Molecular and biochemical analysis of a Madagascar periwinkle root-specific minovincinine-19-hydroxy-O-acetyltransferase. *Plant Physiol.* **125,** 189–98 (2001).

207. Usia, T., Watabe, T., Kadota, S. & Tezuka, Y. Cytochrome P450 2D6 (CYP2D6) inhibitory constituents of Catharanthus roseus. *Biol. Pharm. Bull.* **28,** 1021–4 (2005).

208. Kaltenbach, M., Schröder, G., Schmelzer, E., Lutz, V. & Schröder, J. Flavonoid hydroxylase from Catharanthus roseus: cDNA, heterologous expression, enzyme properties and cell-type specific expression in plants. *Plant J.* **19,** 183–93 (1999).

209. Hotze, M., Schröder, G. & Schröder, J. 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,** 345–50 (1995).

210. Togami, J. *et al.* Isolation of cDNAs encoding tetrahydroxychalcone 2'-glucosyltransferase activity from carnation, cyclamen, and catharanthus. *Plant Biotechnol.* **28,** 231–238 (2011).

211. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126,** 485–93 (2001).

212. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized metabolism in plants. *Science* **344,** 510–3 (2014).

213. Dugé de Bernonville, T. *et al.* 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 (2015).

214. Ohno, S. Evolution by Gene Duplication. *Springer-Verlag* (1970).

215. Causier, B. *et al.* Evolution in action: Following function in duplicated floral homeotic genes. *Curr. Biol.* **15,** 1508–1512 (2005).

216. Theis, N. & Lerdau, M. The Evolution of Function in Plant Secondary Metabolites. *Int. J. Plant Sci.* **164,** S93–S102 (2003).

217. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5,** 299–310 (2004).

218. Farrar, K. & Donnison, I. S. Construction and screening of BAC libraries made from Brachypodium genomic DNA. *Nat. Protoc.* **2,** 1661–1674 (2007).

219. Song, J., Dong, F., Lilly, J. W., Stupar, R. M. & Jiang, J. Instability of bacterial artificial chromosome (BAC) clones containing tandemly repeated DNA sequences. *Genome* **44,** 463–9 (2001).

220. Su, Y.-T., Chen, J.-C. & Lin, C.-P. Phytoplasma-induced floral abnormalities in Catharanthus roseus are associated with phytoplasma accumulation and transcript repression of floral organ identity genes. *Mol. Plant. Microbe. Interact.* **24,** 1502–12 (2011).

221. van der Fits, L. ORCA3, a Jasmonate-Responsive Transcriptional Regulator of Plant Primary and Secondary Metabolism. *Science (80-.).* **289,** 295–297 (2000).

222. Zhang, H. *et al.* The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in Catharanthus roseus. *Plant J.* **67,** 61–71 (2011).

223. Pan, Q. *et al.* Overexpression of ORCA3 and G10H in Catharanthus roseus plants regulated alkaloid biosynthesis and metabolism revealed by NMR-metabolomics. *PLoS One* **7,** e43038 (2012).

224. Van Moerkercke, A. *et al.* The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in Catharanthus roseus. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 8130–5 (2015).

225. Cimermancic, P. *et al.* Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* **158,** 412–421 (2014).

226. Lamb, D. C. *et al.* The First Virally Encoded Cytochrome P450. *J. Virol.* **83,** 8266–8269 (2009).

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

228. Nelson, D. & Werck-Reichhart, D. A P450-centric view of plant evolution. *Plant J.* **66,** 194–211 (2011).

229. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57,** 289–300 (1995).

230. Tian, J. *et al.* TRV-GFP: a modified Tobacco rattle virus vector for efficient and visualizable analysis of gene function. *J. Exp. Bot.* **65,** 311–322 (2014).

231. Kumagai, M. H. *et al.* Cytoplasmic inhibition of carotenoid biosynthesis with virus-derived RNA. *Proc. Natl. Acad. Sci.* **92,** 1679–83 (1995).

232. Ruiz-May, E., Galaz-Ávalos, R. M. & Loyola-Vargas, V. M. Differential Secretion and Accumulation of Terpene Indole Alkaloids in Hairy Roots of Catharanthus roseus Treated with Methyl Jasmonate. *Mol. Biotechnol.* **41,** 278–285 (2009).

233. Ruiz-May, E. *et al.* Methyl jasmonate induces ATP biosynthesis deficiency and accumulation of proteins related to secondary metabolism in Catharanthus roseus (L.) G. hairy roots. *Plant Cell Physiol.* **52,** 1401–21 (2011).

234. Sasaki, Y. *et al.* Monitoring of methyl jasmonate-responsive genes in Arabidopsis by cDNA macroarray: self-activation of jasmonic acid biosynthesis and crosstalk with other phytohormone signaling pathways. *DNA Res.* **8,** 153–61 (2001).

235. Kang, J.-H. *et al.* The flavonoid biosynthetic enzyme chalcone isomerase modulates terpenoid production in glandular trichomes of tomato. *Plant Physiol.* **164,** 1161–74 (2014).

236. Jacobo-Velázquez, D. A., González-Agüero, M. & Cisneros-Zevallos, L. Cross-talk between signaling pathways: The link between plant secondary metabolite production and wounding stress response. *Sci. Rep.* **5,** 8608 (2015).

237. Pasquali, G. *et al.* Coordinated regulation of two indole alkaloid biosynthetic genes from Catharanthus roseus by auxin and elicitors. *Plant Mol. Biol.* **18,** 1121–31 (1992).

238. King, a. J., Brown, G. D., Gilday, a. D., Larson, T. R. & Graham, I. a. Production of Bioactive Diterpenoids in the Euphorbiaceae Depends on Evolutionarily Conserved Gene Clusters. *Plant Cell* **26,** 3286–3298 (2014).

239. Boutanaev, A. M. *et al.* Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.* **112,** E81–E88 (2015).

240. Verma, P., Mathur, A. K., Khan, S. A., Verma, N. & Sharma, A. Transgenic studies for modulating terpenoid indole alkaloids pathway in Catharanthus roseus: present status and future options. *Phytochem. Rev.* (2015).

241. Aguilar-Hernández, V. & Guzmán, P. The fate of tandemly duplicated genes assessed by the expression analysis of a group of Arabidopsis thaliana RING-H2 ubiquitin ligase genes of the ATL family. *Plant Mol. Biol.* **84,** 429–41 (2014).

242. Soria, P. S., McGary, K. L. & Rokas, A. Functional divergence for every paralog. *Mol. Biol. Evol.* **31,** 984–92 (2014).

243. Champagne, A., Rischer, H., Oksman-Caldentey, K. M. & Boutry, M. In-depth proteome mining of cultured Catharanthus roseus cells identifies candidate proteins involved in the synthesis and transport of secondary metabolites. *Proteomics* **12,** 3536–3547 (2012).

244. Kim, J. & Buell, C. R. A revolution in plant metabolism: Genome-enabled pathway discovery. *Plant Physiol.* **169,** pp.00976.2015 (2015).

245. Khaldi, N. *et al.* SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47,** 736–41 (2010).

246. Tremblay, N., Hill, P., Conway, K. R. & Boddy, C. N. in *Methods in molecular biology* **1401,** 233–252 (2016).

247. Harvey, A. L. Natural products as a screening resource. *Curr. Opin. Chem. Biol.* **11,** 480–484 (2007).

248. Rounsley, S. D. & Last, R. L. Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. *Plant J.* **61,** 922–7 (2010).

249. Schilmiller, A. L., Pichersky, E. & Last, R. L. Taming the hydra of specialized metabolism: how systems biology and comparative approaches are revolutionizing plant biochemistry. *Curr. Opin. Plant Biol.* **15,** 338–44 (2012).

250. Nørholm, M. H. H. A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnol.* **10,** 21 (2010).

251. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 31–4 (2007).

252. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78,** 779–87 (2006).

253. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–9 (2012).

254. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–402 (1997).

255. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

256. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

257. Milne, I. *et al.* Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* **14,** 193–202 (2013).

258. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14863–8 (1998).

259. Meijer, A. H. *et al.* 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 (1993).

260. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254 (1976).

261. Cseke, L. J., Kirakosyan, A., Kaufman, P. B. & Westfall, M. V. *Handbook of Molecular and Cellular Methods in Biology and Medicine, Third Edition*. (CRC Press, 2011).

262. Peterson, D. G., Pearson, W. R. & Stack, S. M. Characterization of the tomato (Lycopersicon esculentum) genome using in vitro and in situ DNA reassociation. *Genome* **41,** 346–356 (1998).

263. Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491,** 705–710 (2012).

264. Krizman, M., Jakse, J., Baricevic, D., Javornik, B. & Prosekm, M. Robust CTAB-activated charcoal protocol for plant DNA extraction. *Acta Agric. Slov.* **87,** 427–433 (2006).

265. Poulsen, T. S. Purification of BAC DNA. *Methods Mol. Biol.* **255,** 91–100 (2004).

266. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.* **7,** 203–214 (2000).

9 Appendix

9.1 Additional data for Chapter 4

Sequence for gene expressed in yeast in Chapter 4.2.13.

>CRO_021081

ATGGATGCTTTGCTTAATCCTGTTTTAATCATCACAATAGCTTTTGCTATCGGTTTCACAGTATGTTTCTT
GAAGATTATGAATCCAGGATTACCGGGGAAGAAAAGGAGGTATCATCCTGTTGTTGGCACCATCTTCA
GCCTGCTATTTAACTTCCACAGGCTGCATGATTATATGGCTGATTTGGCAAGCAAATACAAGACATATA
GACTGCTTAACTTTTTCCAGAGGGACTTGATTTATACTGCAGACCCTGCAAATGTTGAATACATTCTGAA
AACAAATTTTCCCAACTATGGCAAGGGATTGTACCACCATGGCATTTTGAAAGATCTTCTTGGCGATGG
TATTTTCACAGTGGACGGTGAAAAGTGGCGACATCAGAGGAAGACATCAAGCTATGAATTCTCCACCA
AAATACTGAGGGACTTCAGCAGTGGAGTGTTTAAAGTCATTGCAGTAAAGCTTTCTAATATAATTTCTG
AAGCTTCATCGTCAGATCAAATTATAGAAATCCAAGATTTGTTCATGAAGTGCACTCTTGACTCGGTCTT
TAAGGTTGTTCTTGGTGTTGATCTGGACAGCATGTGTGGAACTCATGAAGAAGGCACACATTTTTCCCA
GTGTTTCGATGAAGCAAGTGCGATAACCATGTTCCGGTATGCTGATGTATTCTGGAAATTGAAAAGGAT
ATTGAACATTGGATCAGAAGCAATATTGAAGAAAAATATCAAAGTGATTGATGAATATGTTTACAAGAT
TCTCAAAAGCAAAACTGAGTCACTTGATGATAGCCCTATGAGAAAGAAAGAAGACCTTTTGTCAAGATT
TCTGGAAACGAACGAAACAGATCCCAAGTATCTAAAAGATATCATCCTTAGCTTCATAATTGCTGGGAA
GGACACTACTGCCAGCACTCTTTCCTGGTTCTTCTACATGATTTGCAGCTACCCTCACCTACAAGAAAAA
ATCGCTGACGAAGTGATGAAAGCGACAAATGTCCAGGAGAAGAATTTCTCCATTAGTGAATTAGCAAA
CAGCATCACAGAAGAAGCCTTGGACAAAATGCAGTATCTTCACGCAGCCTTAACCGAAACACTCAGACT
CTATCCTTCAGTTCCTGTGGATGGTAAGGTGTGCTTTTCGGATGATATATTTCCGGATGGATACTGTGTC
AAGAAAGGAAATGTTGTATCATACGTACCGTGGTCTATGGGCAGGATGAAATTCATATGGGGAGATGA
TGCTGAAGCTTTCAGACCAGAAAGATGGCTAGATGAAAATGGCGTCTTCAAACAAGAAAGCCCATTCA
AATTCACAGCATTCCAGGCAGGGCCAAGGATCTGTTTAGGGAAGGAATTTGCTTACAGGCAGATGAAG
ATCTTCTCAGCTGTCTTGTTGGGCAGTTTCATGTTCAAGTTGAGTGATGAACAGAGAACTGTGAGCTAC
AGAACAATGCTGACGCTGCACATCAATGGAGGATTACATATTCGTGCTTTCCTGAGGCCAAATTCTTCA
GGTGCTTAA

9.2 Additional data for Chapter 5

Sequences of all VIGS candidates used for silencing.

VIGS C2:
GCTGAATCATCCTCATCAACAATGGAATGGGCAATGTCTGAAATGATAAAAAATCCAAGAATAATGAA
GAGGGCACAAAAAGAGGTTAGAGAAGTCTATAATAAAGAAGGAAATGTGGATGAATCAAAGCTTCA
TGAATTAAAATACCTTCAAGCCATTATTAAAGAAACTCTAAGACTTCATCCAAGTGTTCCACTTTTGATT
CCAAGAGAATCTCGTGAAGAATGTGAAATTAAAGGGTATAGAATACCAGCTAAATGTAGGGTTCTGG
TTAATGGTTGGGCAATCGGAAGAGAT
VIGS C3:
ACTATAATTTCTCCGAGCTGAAATTACCTGTTCATGCTGGTACCCATGTAGATGCACCAGGGCATATGT
ATGATAATTACTTTGATGCCGGATTTGATGTGGATTCTCTTGACCTAAGAGTCCTAAATGGTCCTGCAT
TGCTAATTGATGTTCCAAGGGAGAAGAACATAACTGCTGATGTCATGAAGTCCTTACACATCCCTAGG
GGAGTGAGACGAGTACTTTTTAGAACATTAAACACTGACAGGCGGCTGATGTACAAGAAGGAGTTCG
ACACAAGTTATGTGGGATTCATGAAGGATGGAGCACAGTGGCTTGTGGATAACACTGATATCAAACT
TGTTGGAATTGATTACTTGTCTG
VIGS C4:
GAAGCTAGTTGTTGTTGGTGCACCAAATAAACCAGTTGAATTAGATATTCTATTTCTTGTCATGGGAAG
GAAAATGCTCGGAACATCTTCTGTTGGGGGAGTGAAGGAGACACAAGAGATGATCGATTTTGCAGCA
AAACATGGTATAGTTGCAGATGTAGAAGTTGTGGAAATGGAGAATGTGAACAATGCAATGGAGCGG
CTTGCGAAGGGGGATGTTAGATATAGATTTGTTCTTGATATTGGGAATGCAACAGTAGCTGTCTGATA
AATAATGTTTAATGCAATTCTTTATTTAGTTGTTGCCACAAATTCACCCAGAATGCATGTACCAACATTT
VIGS C5:
AACTTCGGCCATTTGGTATCCATGTTATAAATGTAGTACCTGGAGCAGTGAGGTCTAACATAGGAAAT
TCAGCTATTGCAAATTACAGTAAGATGCCAGAATGGAAGATATACAAGCAATTTGAAGAAGCAATTCG
AGCAAGAGCGTTTCTTTCACAGGGCCTGAAAGCAACCCCGGCAGAAGAGTTTGCAAGAAAGACGGTG
AATGTTGTGCTGAAGAAGAATCCCCCAGCTTGGTTCTCTGCAGGACAGTACTCAACAGTAATGTCAAT
CATGCACCATTTGCCACTTTCTATTAAAGA
VIGS C6:
ACCATACTTCATATCAGATGGCTCACCAATCAATAGTTTTGAGTTTATTCGTCCATTACTTAAGAGCCTG
GATTACGAACTACCGAGTAGGTCAGTAGCCGTTTCCCATGCACTTGTTTTGGGAAATTTCTTTACAGCT
CTCTACTCTGTTATGTATCCATGGTTGAATCAAAAGTGGCTTCCGCAGCCCTTCATCCTTCCTGCTGAA
GTCTACAAGGTTGGCGTAACGCATTACTTTTCGTTTTTGAAAGCAAGAGAGGAGCTTGGCTACATCCC
GATGGTGAGCCCTAAAGAGGGTATGGCTGCAACCATCTCATATTGGCAAGATAAGAAGAAGAGAGA
ATTGGATGGACCTACAATATATGCATGGTTATTTGCAGTGATTGGAATGTTGCTATTATTTGCAGCTGC
TTAC
VIGS C7:
TTCTTCCTCCACTTTCTGCTCTCAAAACCCACGCAGTCCCTCTGCTTTTACCTCTGGAACTGCAAATGCA
AACAGGGGTTTTTTATCATCGAGATTATTCTCTGCTTTGTCACTGAGAAACTCGGCAAGTTTTGGTAAT
TTGAAGGGAAAATGTAAAATCGGTGGGCGTGGAATTGGAATTATGTGCTCTGCTGCTACTGTTCCGCA
GGCTCTGCTGTTTGATTGTGATGGTGTTCTTGTTGATACTGAGAAAGATGGCCATCGCATTTCTTTCAA
TGACACGTTTGCTGAAATGGAATTGGGTGTCACCTGGGATGTAGATT
VIGS C8:
CAGAAGTTTGACGTGGAGCTGAAGGGATCACCGGAGGATGTAGAACTTGTTACCGGAGCAACAATCC
ATACAAAAAATGGATTGTGGTGCAGATTGAGGAAGAGGTGAAATTCCAGGATAAAGTCCTCTGATGA
AGAAGAGGTTTGAGACATTTTGGTAGCAATTTCTTTCTAAAACTAGTGTACATAACCTGTGTAAACTAT
AGGCATGACAATGTGGGGAGAAGTAAAATTTTCAAGTGGATAAAGATTCTCTTTTAATGCTTCTCTCTC
CACATGAAACTGTGATCAAAATCCTTGTTGCCTAAGTTGTATAATACAGTTTTGCTCAGGGAATTTGGT
TGAAG

VIGS C9:
CCTTTGTCGTTGTATCAGATCCTATTGTTGCAAGATACATTCTTCGCGAAAATGCATTTTCTTATGACAA
GGGGGTTCTTGCAGAAATTCTAGAACCAATCATGGGAAAAGGGCTAATACCTGCAGACCTTGATACTT
GGAAACAAAGGAGAAGAGTCATTGCTCCTGGATTTCACGCATTATACTTGGAAGCCATGGTCAAAGTT
TTTATGGACTGTTCAGAACGAACAACATCAAAATTTGTAAGGCTTCTTGAGAAAGAAGAGTCAAAAG
GAGGGAAGACCATTGAGCTGGATCTTGAAGCAGAATTTTCCAGCTTAGCACTAGATATTATTGGGCTT
GGAGTTTTCAATTATGACTTTGGCTCTGTTACCAAAGAATCTCCTGTTATTAAGGCTGTATATGGTAC
VIGS C10:
TTCAGTGTCATGATGTAATGCAGTTAAGGMCACATAGAACTGTGTATCAGCTGGAATGTTTAGYTTTC
CCAACTTGATATCCTTTGATGTTTTCCTCGTCTGWTGGACTRCTGGTGGATAGAGCCGAAGCGTTTCG
TTCAGTATCATCGTCACAAGCTTAAAATCATTTAGGRTCTCTGCATTTGGAAGCACATTGCCCTTGCAA
ACCTGCAAAACTTCTTCTCGTGCTTTACTTTGCCATTCTTG
VIGS C12:
TCAATGGAGAGAAAAGGTGAGGTTGAGATTAAAAAAGTGTTGCATTTTGGGTCTTTAAACAATGTCAT
GATGAGTGTGTTTGGAAAATGCTATGATTTTTGTGAAAAAAGTGGAAATACTGAAGGGTTTGAACTA
GAAAGTTTGGTAAATGAAGGGTATGAGTTGTTGGGTATTTTTAACTGGAGTGATCATTTTCCTGTGTT
GGGTTGGTTGGACATGCAAGGTGTGAGGAAGAGGTGCAAGGCTTTGGTTAGTAGAGTCAATGTTTAT
GTTGAGAAAATCATAGAGGAGCACAGGCAGAAAAGGGCTGAGAATGGCGGAAGAATTTCTGAGTCT
GAGGATTATATTAATTCTGGTACTGGAGACTTTATGGATGTTCTATTGGATT
VIGS C13:
GTTGGTCTAGGTGGACTTGGTCATTTAGCTGTGAAGTTTGGCAAGGCTTTTGGGGCCAAAGTCACTGT
AATTAGTACCTCTCCAAGCAAGAAAGATGAAGCTATCAATTTTCTTGGTGCTGATGGCTTCTTGGTCAG
CGGTGATGCTGAACAAATGCAGGCTGCTGCTGGAACCTTGGATGGGATTATTGATACCGTGCCTGTT
GTCCATTCTATTGAGACCTTGCTATGGCTTTTGAAGAATCATTCCAAGCTTGTTTTGGTTGGAGCTACA
GGGGGCTCATTCGATTTGCCAATTCTTCCTTTAGCAATGGGCAGAAGAACTGTGGCTTCAAGCATTGG
TGGAAGTACGAAGGAGGCTCAAGAGATGCTGGATTTTGCAGCAGAACATAA
VIGS C14:
GAGTTAGAGTTGGTATAGTTGGTCTAGGAGCAGTTGGACATTTAGCTATTAAATTTGCAAAAGCTTTT
GGTGCTAGGGTTACTTTGATCAGTTCATCCCCTGGAAAAAAGGATGAGGCTTTTCAGAAATTTGGTGT
AGATTCTTTCTTGGTCAGCAGTAATGCAGAGGAAATGCAGGCTGCAGCTGAAACTCTGGATGGTATCC
TAGACACTGTACCAGTGGTTCACCCCCTTGAGCCACTCTTTGCTTTACTGAAACCTCTTGGGAAACTTA
TCATTATAGGTGAACCGCATAAGCCTTTTGAGGTATCCGCAATGTCCCTCATGGAGGGTGGAAAAATA
ATTAGCGCGAGTACGGGTGGAAGTATAAAGGACACACAAGAGATAGTCGATTTTGCAGCAGAACATA
AC
VIGS C15 :
CACCGTTGGAAGAACTAACCGCAATCTTCTTAGATGCAATAAAGGTTGGTTACCGGCACTTCGACACG
GCTTCGAGTTACGGCACAGAGGAGGCACTTGGTAAAGCCATAGCTGAGGCTATAAACAGCGGTTTGG
TTAAAAGTAGGGAGGAATTCTTCATCAGTTGTAAGCTGTGGATTGAAGATGCAGATCATGACCTTATC
TTGCCTGCCCTCAACCAGTCACTTCAGATTCTTGGGGTTGATTATTTGGATCTATATATGATACACATG
CCGGTGAGAGTGAGGAAAGGTGCTCCCATGTTCAATTATTCAAAAGAGGATTTCCTTCCATTTGACAT
ACAAGGTACATGGAAG
VIGS C16:
CCATTTGGTGCTGGACGAAGAATTTGCCCCGGCATGTCATTTGCCATACCTAATGTTACGTTGCCGCTG
GCACAATTATTGTTACACTTCGATTGGAAATCAGCCGATGGTAAACTCGAAGATTTGGACATGACTGA
AGGTTCGAGTTTGACAGTTAGACGAAAGAATGAACTAGAGTTGATTCCATCAACTTATTACAAATCTT
GTC
VIGS C18:
TTTATGCAAACTCATGGCCTAAACTTTGCGGATAGGCCACAGACTGTTATTGCCAAGATCATGATGTAC
AATTGTTCAGGCGTTACACTTTCAATGTACGGCGATTATTGGAGAAAATTGCGGCAAATTTATGTCAC
GGAATTACTTAACACAAAAAGTGTACAATCATTTTCTTCCATCATGGAAGAAGAGCTCATTCTTATGGT

TAAATCTATTGAGTCAGAAGTTGGGAAACCTATGGAATTGATTGAGAAAATTCGGTCATATTTATTTG
ACACACTTTGTAGATCAGCACTTGGTAAAATACATGGTAAAGGGAAGGAGACATTAATTGAGATAAG
CAGAGAAATGGTGGCACTTTCTGGAGTACAAACNCTAGAAGATATTTTTCCTTCAGTGAAGTTATTTTC
ATATTTGAATCCATTGAGGCCCAAAGCAAAGAAG
VIGS T3O:
TTGAGCAAATTGATGACGTCATGCACAAATTCAATTATTAATAAAGTAGCTTTTGGTAAAGTACGTTAT
GAACGGGAGGTGTTTATTGATCTAATTAATCAAATATTAGCATTAGCAGGCGGTTTTAAGCTGGTTGA
TCTGTTTCCGTCCTACAAGATACTTCATGTTCTTGAAGGTACAGAACGTAAGCTGTGGGAAATCCGCG
GTAAGATTGACAAGATTTTGGATAAAGTCATAGACGAGCACAGAGAAAATTCGTCAAGAACTGGAAA
GGGCAACGGTTGTAATGGCCAGGAAGATATAGTTGATATTTTACTTAGGATTGAAGAGGGTGGTGAT
CTTGACCTTGATATTC
VIGS C30:
ACCAGGTCCAAGAACACTACCCTTAATTGGAAACCTTCATCAACTCTCGGGACCTTTACCTCATCGTAC
CCTAAAAAAATTTGTCCGATAAACATGGTCCTTTGATGCACGTGAAAATGGGCGAACGTTCGGCAATTA
TAGTATCAGATGCAAGAATGGCAAAAATAGTTCTTCATAATAACGGTTTAGCCGTTGCAGATCGGTCA
VIGS C32:
TAACAGCAAGAACTTCATTGATGCTTTCCCCAGCCACCTCATTAATAGGAACTGCATCACTAATCTCTT
GCAAATCTTGTTTTGTCAGCTTTACTTTCACTGCACCAACGTTATCGTGAAGATTTTTAATCTTCGTGGT
ACCAGGAATAGGTACCACATCATCGCCCTGATGAAGAACCCAAGCAAGAGAAAGTTGAGCAGGAGTA
CATCCATGCTTTTGTGCCAAGGCTTCTATACGATAATATATTTGCTTGTTCTTTTCAAGGTTCTCCCCTGT
AAATCTGGGATGTATTGCCAAGAAACTACTTTGTGAAAGGCTTTCCGTAACAGCCTTCCCAGCAAAAA
GACC
VIGS C36:
ATGGTCCAAACACAGAAGAATCATTTCCCCTGCCTTCACTCTTGAGAAACTAAAGTCAATGTTGCCGGC
TTTTGCCATATGCTATCATGACATGTTGACGAAATGGGAGAAATTAGCAGAAAAAGAAGGATCCCAT
GAAGTTGATATCTTCCCCACATTTGATGTTTTGACAAGTGATGTAATTTCAAAGGTTGCATTTGGTAGC
ACATACGATGAAGGAGGCAAAATCTTCAGACTATTGAAAGAACTCATGGATCTCACAATTGATTGCAT
GAGAGATGTCTACATCCCAGGGTGGAGCTACTTGCCAACCAAGAGGAACAAGAGGATGAAAGAAAT
TAACAAAGAAATCACAGATATGCTAAGGTTC
VIGS C38:
CTTCGCCTACGCATCTCCCTTCTGGAACAAAGCTTTTTGTGAGAACAGCACCGATCCAGAGAAAAGAC
CATTGTGTGGGAGGACATATGATATTTCCTATGACTATAAGAACAGCCAAATGTACATTGTTGATGGC
CATTACCATCTTTGTGTGGTTGGAAAAGAAGGTGGGTATGCCACACAACTAGCCACAAGTGTGCAAG
GAGTGCCATTCAAATGGCTCTATGCAGTAACTGTTGATCAGAGAACAGGGATTGTTTATTTCACTGAT
GTTAGCTCCATACATGATGACAGTCCCGAAGGTGTGGAAGAAATCATGAATACAA
VIGS C39:
TGGAATCGAACGAGTTGACTCACTGAGTCTGAGTCCACACAAATGGCTACTCGCTTACTTAGATTGCA
CTTGCTTGTGGGTCAAGCAACCACATTTGTTACTAAGGGCACTCACTACGAATCCTGAGTATTTAAAAA
ATAAACAGAGTGATTTAGACAAAGTTGTGGACTTCAAAAATTGGCAAATCGCAACGGGACGAAAATT
TCGGTCGCTGAAACTTTGGCTCATTTTACGTAGCTATGGAGTTGTTAATTTACAGAGTCATATTCGTTC
TGACGTCGCAATGGGCAAAATGTTCGAAGAATGGGTTAGATCAGACTCCAGATTCGAAATTGTGGTA
CCGA
VIGS C41:
ATGAAGGTGATAGAGAAGATCAACGATGCCATTGCCGCTGACAAGGTGGTCTTCTCGTTCGAGTTTTT
TCCACCGAAGACTGAGGACGGAGTGGAGAATCTGTTCGAGAGAATGGACCGTATGGTAGCTCACAAT
CCTTCCTTCTGTGATATCACTTGGGGAGCTGGAGGATCCACTGCGGATCTGACTTTGGAGATCGCTAA
CAGAATGCAGAACATGGTTTGTGTGGAAACTATGATGCATCTCACTTGCACCAATATGCCTGTGGAGA
AGATCGACCACGCTCTCGATAGTATTAAATCCAATGGCATCCAGAACGTGCTCGCTCTTCGAGGTGAT
CCTCCTCATGGTCAGGATAAGTTTGTTCAGGTGGAAGGTGGATTTGCCTG
VIGS C43:

AGCTCGAGTTCATTCCCAACTCTCAGAAGCAATTAAAAATGTCGCCATAGATGAACTCAATCAATACCT
GCCATTCCAACCTTATCCCGGTGGGGAGGAAAGTGGGTTGAAAAAAGATATCCCCTTAGCTGTAAAA
ATCAGTTGTTTCGAATGTGGCGGAACAGCAATTGGGGTCTGTATTTCCCACAAGATTGCCGATGCGTT
GTCCTTGGCTACCTTCCTCAATTCATGGACCGCAACATGCCAGGAAGAGACTGATATTGTTCAACCTAA
TTTCGATCTGGGATCCCATCATTTTCCGCCTATG
VIGS C44:
TTATCAAGTAGAACAATTCTCAATGCAATTCAAGCACAAGATCATCATAAATATTCTATGGCGTGGTTA
CCGGTGTCAGCAGAATGGCATAAGCTTCGTAAGATAAGTAAGGAACAGATGTTTTCAGCAGCACAGC
TTGACTCGAGACAAAATTTTCGTCAAGAAAAATTTAAAGAGTTACGCGAATATTTGCATCTTTACAGTT
TTAATAGAGAATTTGTTAATATTGGTAAAGCTGCTTTTACAACTTCTTTGAATTTAATTTCAACTACCAT
GTTTTCAAAAGATTTTGCTACTTATGATTCAAATTCATCTCAAG
VIGS C45:
ACACTTCGTAGCCGAAACCAAAGTTATTGGGGTTTCGATTGAATATAGACTTGCCCCAGAGCACCTTTT
ACCCGCAGCTTATGAAGATTGTTGGGAAGCCCTTCAATGGGTTGCTTCTCATGTGGGTCTCGACAATT
CCGGCCTAAAGACAGCTATTGATAAAGATCCATGGATAATAAACTATGGTGATTTCGATAGACTGTAT
TTGGCGGGTGACAGTCCTGGTGCTAATATTGTTCACAAC
VIGS C46:
AAGCCATTGTCTTTGTTCGGTCACGGAATAACTGACATTGAGATTCATTTCCTTCAAATTAAGTTCACT
GCTATTGGGGTTTATCTCGATCCGGCAATCGTCACTCATTTACAGAAATGGAAGGGCAAGAAAGGTTC
TGAACTTGCCGGCGACGATGAGTTTTTTGATGCCGTTATTTCAGCTCCTGTGGAAAAGTTATTGAGAA
TTGTTGTGATTAAGGAGATAAAGGGTTCCCAATATGGAGCCCAGCTTGAAAGTGCAGTGAGGGATCG
ATTGGCAGAAGAAGATAAATATGAAGATGAAGAGGAAGAAGCACTTGAGAAAATTCTTGAATTTTTG
CAATCTAAATATTTCAGCAAAGATTCAATCCTCACTTATTATTTTCCTGCTAATTCTTCAACTGCTGAGA
T
VIGS C47:
CTATTGCTTGCAAGTAGGGAAGCCTCGAGATATCTGATTCATGAATGATTTCCTTCTCCCCGATGACTT
GATGGAGTTCAGCTCTAACTCTTTCTAATTTTTCAGGGTTGCGTAATAACTCTGCCATAGCCCATTCCA
GCGTGGTCGATGTTGTTTCAGTCCCTCCTAGAAATAAATCCAAAAGCATGTGCTTGAGCTCATTGAGG
CTGAACTCAGACTCATTATTGGCACTGTGATTGAGAAGAGCTTCTAGTAAATCATTTTTCTTTGAAGTG
TCCAATGAACCTCTCATCATCAGTCGCTCATTAATA
VIGS C48:
ATCCTAAGTATTGGTCCGAACCTGAAGAATTCAAACCCGAAAGATTTATTGATTCTCGTATTGATTATA
AAGGTAATGATTTTGAGTACATACCATTTGGTGCCGGAAGAAGGATTTGTCCTGGTATGTCATTTGCT
ATACCTAATGTGGCTTTACCATTGGCACAATTGTTGTTTCATTTTGATTGGAAAGTTGACGATGATGCT
GGGAAATTAGAAGATTTGGACATGGTTGAGCATCCTGTTCTTGAAGCTAGGCG
VIGS T16H2:
TTGGAAACTTGCCAGTGATGAAACAAATATTGATAAATTAGACATGACGGAGAGTAGAGGAGTAACA
GTTAGAAGAGAAGATGATTTGTGTCTGATTCCATTTCCTTATTCTGCTTCTTCTCTCAAAGGTAAATATT
AGATGGAGACAACATCACCAAATAATTGAAGACCAAGATTGTAGGTCATGAATAGTTGTCGGTCAAG
AAATGGTGGCAAAAATTGCGATGGAAGATTTCGTTGCT
VIGS C53:
TACCCAATGACACCTGGTATATGGACTAAAGAACAAGTTGAGGCCTGGAAACCTATTGTAGATGCAGT
TCATGCCAAAGGTGGTGTCTTCTTTTGCCAGATAGGGCATGTTGGAAGAGTTTCAAATTACAGTTATC
AGCCAAATGGGCAGGCGCCAATCTCTTCGACAGACAAGGGATTAACC

Manuscript and supplementary data for “A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in Catharanthus roseus.”

Manuscript and Supplement data for “Discovery of a P450-catalyzed step in vindoline biosynthesis: a link between the aspidosperma and eburnamine type alkaloid scaffolds.”

Manuscript and Supplement data for “Genome-guided investigation of plant natural product biosynthesis”