PCP
PLANT & CELL PHYSIOLOGY

Special Focus Issue – Regular Paper

# CathaCyc, a Metabolic Pathway Database Built from *Catharanthus roseus* RNA-Seq Data

Alex Van Moerkercke[1,2,6], Michele Fabris[1,2,3,6], Jacob Pollier[1,2,6], Gino J.E. Baart[1,2,3], Stephane Rombauts[1,2], Ghulam Hasnain[4], Heiko Rischer[5], Johan Memelink[4], Kirsi-Marja Oksman-Caldentey[5] and Alain Goossens[1,2,*]

[1]Department of Plant Systems Biology, VIB, B-9052 Gent, Belgium
[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Gent, Belgium
[3]Department of Biology, Laboratory of Protistology and Aquatic Ecology, Ghent University, B-9000 Gent, Belgium
[4]Institute of Biology, Leiden University, 2300 RA Leiden, P.O. Box 9505, The Netherlands
[5]VTT Technical Research Centre of Finland, FIN-02044 VTT, Espoo, Finland
[6]These authors contributed equally to this work.
*Corresponding author: E-mail, algoo@psb.vib-ugent.be; Fax, +32-9-3313809.

(Received December 14, 2012; Accepted March 4, 2013)

**The medicinal plant Madagascar periwinkle (*Catharanthus roseus*) synthesizes numerous terpenoid indole alkaloids (TIAs), such as the anticancer drugs vinblastine and vincristine. The TIA pathway operates in a complex metabolic network that steers plant growth and survival. Pathway databases and metabolic networks reconstructed from 'omics' sequence data can help to discover missing enzymes, study metabolic pathway evolution and, ultimately, engineer metabolic pathways. To date, such databases have mainly been built for model plant species with sequenced genomes. Although genome sequence data are not available for most medicinal plant species, next-generation sequencing is now extensively employed to create comprehensive medicinal plant transcriptome sequence resources. Here we report on the construction of CathaCyc, a detailed metabolic pathway database, from *C. roseus* RNA-Seq data sets. CathaCyc (version 1.0) contains 390 pathways with 1,347 assigned enzymes and spans primary and secondary metabolism. Curation of the pathways linked with the synthesis of TIAs and triterpenoids, their primary metabolic precursors, and their elicitors, the jasmonate hormones, demonstrated that RNA-Seq resources are suitable for the construction of pathway databases. CathaCyc is accessible online (http://www.cathacyc.org) and offers a range of tools for the visualization and analysis of metabolic networks and 'omics' data. Overlay with expression data from publicly available RNA-Seq resources demonstrated that two well-characterized *C. roseus* terpenoid pathways, those of TIAs and triterpenoids, are subject to distinct regulation by both developmental and environmental cues. We anticipate that databases such as CathaCyc will become key to the study and exploitation of the metabolism of medicinal plants.**

**Keywords:** BioCyc • *Catharanthus roseus* • Deep sequencing • Madagascar periwinkle • Metabolism • Transcriptome.

**Abbreviations:** EST, expressed sequence tag; FL, full-length; FL-ORF, full-length open reading frame; FPKM, fragments per kilobase of transcript per million fragments mapped; JA, jasmonate; MeJA, methyl jasmonate; MVA, mevalonate; ORF, open reading frame; PDB, pathway database; STR, strictosidine synthase; TIA, terpenoid indole alkaloid.

The raw RNA-Seq reads reported in this paper have been submitted to the NCBI Short Read Archive under accession numbers SRA064076 (*C. roseus* plants) and SRA064724 (*C. roseus* cells).

## Introduction

The medicinal plant *Catharanthus roseus* (Madagascar periwinkle) synthesizes >150 different terpenoid indole alkaloids (TIAs), including the pharmaceutically important molecules ajmalicine and serpentine. In addition, it is the sole source of the commercial anticancer TIA compounds vinblastine and vincristine (van der Heijden et al. 2004, Verma et al. 2012). All TIA compounds are synthesized in a highly branched and complex pathway from the central compound strictosidine, a condensation product of the monoterpenoid compound secologanin and the indole compound tryptamine (Facchini and De Luca 2008). The biosynthesis of TIAs occurs in a jasmonate (JA)-responsive manner and involves at least three different cell types and several intracellular compartments (St-Pierre et al. 1999, Facchini and St-Pierre 2005, Guirimand et al. 2011). Because TIAs accumulate in very low amounts, they are difficult to extract, leading to a high commercial production cost. Efforts to increase or alter the production of TIAs in *C. roseus* plant, cell

*Plant Cell Physiol.* 54(5): 673–685 (2013) doi:10.1093/pcp/pct039, available online at www.pcp.oxfordjournals.org

culture or hairy root systems have only been partly successful, due to the complex cellular organization of this pathway and our fragmented knowledge of the enzymes acting in the different branches (Hughes et al. 2004, Zhou et al. 2009). Indeed, most enzymes involved in the production of TIA compounds from strictosidine onwards have not been identified, with the exception of the six-step conversion of tabersonine to vindoline (Loyola-Vargas et al. 2007, Facchini and De Luca 2008). *Catharanthus roseus* serves as one of the model systems of choice for TIA production, but it also produces several other classes of natural products, including phenolics and triterpenoids (Mustafa and Verpoorte 2007, Ferreres et al. 2011, Huang et al. 2012, Yu et al. 2013).

Inherent to the species-specific character of plant secondary metabolism, most of these molecules and/or pathways are absent in model systems such as *Arabidopsis thaliana* (Arabidopsis). Therefore, further exploration of secondary metabolism in medicinal plants will be needed to develop metabolic engineering strategies to alter the production of these compounds in plants and other systems. Many of the enzymes involved in secondary metabolic pathways in medicinal plants such as *C. roseus* await discovery and characterization. Combined with our limited understanding of the regulation of these pathways, this impedes the development of efficient metabolic engineering strategies to increase yields of the high-value compounds they produce. As a consequence, the metabolic potential of medicinal plants in general is far from being fully explored. Therefore, categorization and characterization of all relevant pathways in these plants would benefit medicinal compound discovery and metabolic engineering. This can be streamlined by constructing metabolic databases from annotated sequence information.

Pathway databases (PDBs) constructed from annotated genomes are available for model systems such as *Escherichia coli* (EcoCyc), *Saccharomyces cerevisiae* (YeastCyc), Arabidopsis (AraCyc), *Oryza sativa* (RiceCyc), *Populus trichocarpa* (PoplarCyc), *Chlamydomonas reinhardtii* (ChlamyCyc), *Homo sapiens* (HumanCyc) and >100 bacteria (Karp et al. 1997, Mueller et al. 2003, Romero et al. 2005, Zhang et al. 2005, Jaiswal et al. 2006, May et al. 2009, Zhang et al. 2010). Recently, a PDB of the model diatom *Phaeodactylum tricornutum* (DiatomCyc) has been constructed, leading to the identification of novel glycolytic pathways in eukaryotes (Fabris et al. 2012). Furthermore, a PDB of *Medicago truncatula* (MedicCyc) was constructed from the draft genome sequence, complemented with expressed sequence tag (EST) sequence information (Urbanczyk-Wochniak and Sumner 2007). The Solanaceae community also created metabolic PDBs (SolCyc) for some of its members (LycoCyc, PetCyc, CoffeaCyc, CapCyc and PotatoCyc) from EST collections (http://solcyc.solgenomics.net/). Finally, the MetaCyc PDB aims to collect every experimentally determined biochemical pathway for small molecule metabolism (Krieger et al. 2004, Caspi et al. 2012) and PlantCyc aims to catalog all plant-specific molecules, enzymes and pathways (Zhang et al. 2010). To date, PlantCyc (http://www.plantcyc.org) comprises 879 pathways and 3,455 compounds. The latter seems particularly low given the large number of compounds present in plants. For instance, >12,000 alkaloid compounds, of which 2,000 are TIAs, have already been identified in the plant kingdom (Ziegler and Facchini 2008). Many of these compounds and pathways are not represented in the current databases. Although some literature-curated pathways from *C. roseus* have been included in PlantCyc and MetaCyc (e.g. vinblastine biosynthesis), a comprehensive overview or database of *C. roseus* metabolism has not been constructed to date.

Even though the genome of *C. roseus* and other medicinal plants has not been sequenced yet, for many of them large EST or RNA-Seq collections are available (He et al. 2011, Wenping et al. 2011, Desgagné-Penix et al. 2012). Here, we show that such resources have great potential, not only for gene discovery but also for the establishment of metabolic PDBs. First, we have conducted an elaborate RNA sequencing experiment of the *C. roseus* transcriptome, spanning different organs and growth conditions. From this, we have reconstructed the *C. roseus* metabolic map, emphasizing important compound classes, using AraCyc and MetaCyc as initial templates in conjunction with the Pathway Tools prediction software (Karp et al. 2010). Optimization of the PDB by manual curation resulted in the first metabolic PDB of *C. roseus*, called CathaCyc.

## Results

### Illumina HiSeq2000 RNA sequencing

RNA-Seq is the most powerful transcript profiling method available to date and, unlike microarray technology, is applicable to species without existing genomic sequence (Wang et al. 2009), including *C. roseus*. Therefore, we designed an Illumina HiSeq2000-based RNA sequencing strategy such that both de novo sequence assembly and transcript counting was possible. The former facilitates PDB assembly, cloning of full-length open reading frames (FL-ORFs) for gene discovery projects, and proteomics analysis, amongst others. The latter enables comparative mining of gene expression, which in turn allows selection of candidate genes for gene discovery programs, for instance to fill the current gaps in TIA biosynthesis. Here, we focus on the generation and use of our RNA-Seq data set for the assembly of the first *C. roseus* PDB, CathaCyc. The *C. roseus* explant material that has been used encompasses suspension cells and shoots treated or not with methyl jasmonate (MeJA).

In total about 44.2 × $10^9$ bases were sequenced. The cell and shoot samples were run in separate batches and initially assembled as separate RNA-Seq sets (**Supplementary Table S1**). Of the cell-derived libraries, a total of 141,031,789 reads were sequenced corresponding to 28,206,357,800 bases. De novo assembly was performed with the VELVET transcriptome assembler (Zerbino and Birney 2008) and its module OASES (Schulz et al. 2012), which generated a total of 31,015 contigs with an average length of 840 nucleotides (nt). The contig maximum length was 8,252 nt. Of the shoot-derived libraries, a total of 80,127,936 reads were sequenced corresponding to 16,025,587,200 bases. A total of 36,363 contigs was generated from this set with an average length of 1,084 nt. The contig maximum length was 11,904 nt.

Pilot BLAST screens revealed that all known TIA genes could be retrieved in either the cell or shoot contig collections, or both (**Supplementary Table S2**). Importantly, all of them were nearly (minimum 98%) or 100% identical to the sequences in the National Center for Biotechnology Information (NCBI) database and in the assemblies from both explant sets. Furthermore, for 25 out of the 36 known TIA genes, the full-length (FL) sequence could be retrieved. This analysis supports the quality and potential utility of our sequence data set for gene discovery and the construction of CathaCyc.

## Assembly and annotation of the 'reference transcriptome'

To construct a reference transcriptome for *C. roseus*, the generated raw RNA-Seq reads of the suspension cell and plant samples were assembled into two distinct sets of contigs by the sequencing service provider. However, to serve as a basis for expression analysis, an exhaustive reference unigene set, representing the whole *C. roseus* transcriptome, was needed.

To inspect the provided contigs and to ensure completeness of the sequence set, we made de novo assemblies from a subset of the publicly available data set from the Medicinal Plant Genomics Resource (MPGR) consortium (http://medicinalplantgenomics.msu.edu/), that comprises RNA-Seq data from >20 different *C. roseus* tissues and cultures grown in different conditions (Góngora-Castillo et al. 2012), and compared these with our assemblies. This evaluation did not yield any longer transcripts. Furthermore, allelic differences made it more difficult to build a strict unigene and would render future downstream analyses more complex. It was therefore decided to restrict and generate the reference *C. roseus* unigene set by combining and joining both our sets only. This assembly resulted in a unigene set of 31,450, in the majority FL, transcripts, designated with the prefix Caros, on which we predicted ORFs. Due to indels in the assemblies of some transcripts, more than one ORF was returned for a number of transcripts. For highly relevant *C. roseus* pathways, such as the TIA and triterpenoid pathways, as well as their precursor and the MeJA biosynthesis pathways, we edited and curated all sequences that showed indels or frameshifts.

We assigned a functional description to each Caros transcript through a guilt-by-homology approach based on BLASTX analysis (see the Materials and Methods). Subsequently, the annotation of the predicted translated ORFs was further refined by using an established orthology-based method (Fabris et al. 2012), in which translated genomic sequences of 17 annotated organisms were used as reference (see the Materials and Methods). Thereby, gene functions and gene–reaction associations were transferred from the reference genomes to the query *C. roseus* transcriptome by means of a score-driven semi-automated pipeline (Fabris et al. 2012).

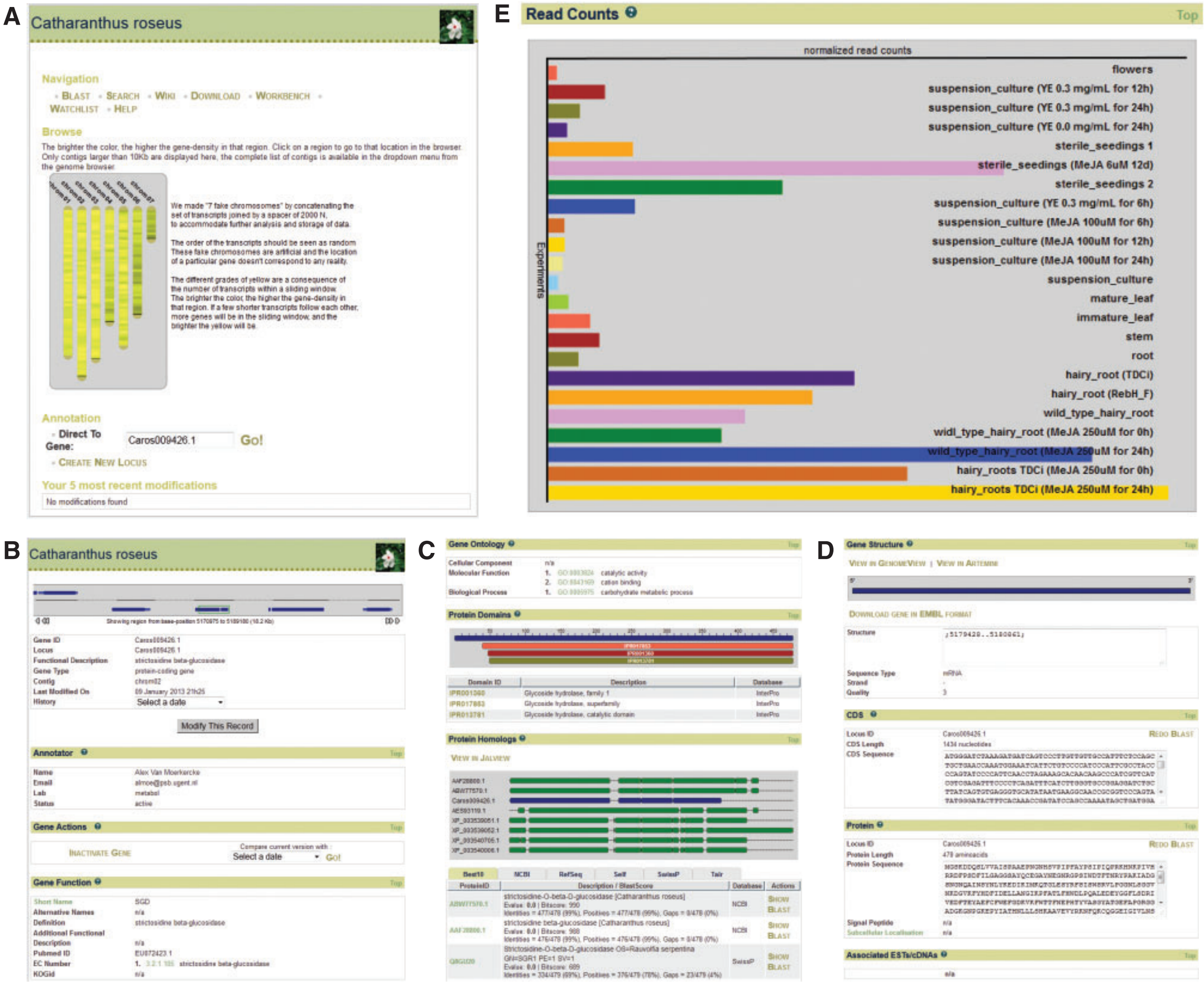
To accommodate analyses and storage in the ORCAE database with web interface (Sterck et al. 2012), we built 'fake chromosomes' by concatenating the set of Caros transcripts joined by a spacer of 2,000 N. This resulted in seven chromosomes, of which the first six contain 5,000 transcripts each. This platform, accessible at http://bioinformatics.psb.ugent.be/orcae/, offers the possibility to edit and curate the predicted Caros ORFs, append functional annotations where needed and display pre-computed analyses such as protein domains, BLAST alignments and expression data [e.g. FPKM (fragments per kilobase of transcript per million fragments mapped) values] depicted as bar diagrams. Besides displaying gene-related data, blast and search options are at the disposal of the user (see **Fig. 1** for an example).

To assess the quality of our reference transcriptome, we verified the presence and completeness of publicly known *C. roseus* sequences. To this end, we downloaded the 406 *C. roseus* protein sequences that were publicly available at NCBI (at October 24, 2012) and filtered out the 168 unique FL entries. A TBLASTN search with these 168 protein sequences against our reference transcriptome showed that 164/168 (98%) of the sequences were represented in our assembled transcriptome (i.e. the 'fake chromosomes'). Of the four proteins that were not present in our data set, only one was present in the MPGR data set. Verification of the sequences revealed that the FL sequence of 133 of the 164 proteins (81%) was present in our data set (**Supplementary Table S3**).

## Creation and manual curation of CathaCyc v1.0

The information obtained from the orthology-based functional annotation of transcripts was imported into Pathway Tools 16 (Karp et al. 2010, Paley et al. 2012) and used as input for Pathologic (Karp et al. 2010, Paley et al. 2012) for the raw construction of CathaCyc. The resulting draft metabolic network lacked several important pathways common to all plants and, at the same time, included many unnecessary ones, specific to prokaryotes, for instance. Therefore, extensive manual curation was carried out by importing missing pathways from the MetaCyc (Caspi et al. 2010, Caspi et al. 2012), AraCyc (Zhang et al. 2005) and PlantCyc (Zhang et al. 2010) databases and by using the literature as a reference to add missing gene–reaction associations, as well as to remove false-predicted pathways. After this refining procedure, the database still contained 910 pathway gaps, related to reactions, which lacked the correct link to a gene. By using the implemented BLAST-based Pathway Hole Filler algorithm, we could find gene connections for 498 pathway holes (see the Materials and Methods).

Currently, CathaCyc v1.0 consists of 1,802 reactions organized in 390 pathways, 1,347 enzymes and 1,322 compounds, which is similar to other plant PDBs (see www.plantcyc.org for an overview). CathaCyc v1.0 is accessible at www.cathacyc.org through an intuitive and user-friendly web interface based on the MetaCyc family format. Users can explore *C. roseus* metabolism at different levels, by browsing specific information pages



Catharanthus roseus

Navigation

- BLAST - SEARCH - WIKI - DOWNLOAD - WORKBENCH - WATCHLIST - HELP

Browse

The brighter the color, the higher the gene-density in that region. Click on a region to go to that location in the browser. Only contigs larger than 10Kb are displayed here, the complete list of contigs is available in the dropdown menu from the genome browser.

We made "7 fake chromosomes" by concatenating the set of transcripts joined by a spacer of 2000 N, to accommodate further analysis and storage of data.

The order of the transcripts should be seen as random. These fake chromosomes are artificial and the location of a particular gene doesn't correspond to any reality.

The different grades of yellow are a consequence of the number of transcripts within a sliding window. The brighter the color, the higher the gene-density in that region. If a few shorter transcripts follow each other, more genes will be in the sliding window, and the brighter the yellow will be.

Annotation

- Direct To Gene: Caros009426.1 Go!
- CREATE NEW LOCUS

Your 5 most recent modifications

No modifications found

Catharanthus roseus

Showing region from base-position 5170875 to 5189100 (18.2 Kb)

| | |
|---|---|
| Gene ID | Caros009426.1 |
| Locus | Caros009426.1 |
| Functional Description | strictosidine beta-glucosidase |
| Gene Type | protein-coding gene |
| Contig | chrom02 |
| Last Modified On | 09 January 2013 21h25 |
| History | Select a date |

Modify This Record

Annotator

| | |
|---|---|
| Name | Alex Van Moerkercke |
| Email | almoe@psb.ugent.nl |
| Lab | metabol |
| Status | active |

Gene Actions

INACTIVATE GENE Compare current version with : Select a date Go!

Gene Function

| | |
|---|---|
| Short Name | SGD |
| Alternative Names | n/a |
| Definition | strictosidine beta-glucosidase |
| Additional Functional Description | n/a |
| Pubmed ID | EU072423.1 |
| EC Number | 1. 3.2.1.105 strictosidine beta-glucosidase |
| KOGid | n/a |

Gene Ontology

| | | | |
|---|---|---|---|
| Cellular Component | n/a | | |
| Molecular Function | 1. | GO:0003824 | catalytic activity |
| | 2. | GO:0043169 | cation binding |
| Biological Process | 1. | GO:0005975 | carbohydrate metabolic process |

Protein Domains

| Domain ID | Description | Database |
|---|---|---|
| IPR001360 | Glycoside hydrolase, family 1 | InterPro |
| IPR017853 | Glycoside hydrolase, superfamily | InterPro |
| IPR013781 | Glycoside hydrolase, catalytic domain | InterPro |

Protein Homologs

VIEW IN JALVIEW

Best10 | NCBI | RefSeq | Self | SwissP | Tair

| ProteinID | Description / BlastScore | Database | Actions |
|---|---|---|---|
| ABW77570.1 | strictosidine-O-beta-D-glucosidase [Catharanthus roseus]<br>Evalue: 0.0 \| Bitscore: 990<br>Identities = 477/478 (99%), Positives = 477/478 (99%), Gaps = 0/478 (0%) | NCBI | SHOW<br>BLAST |
| AAF28800.1 | strictosidine beta-glucosidase [Catharanthus roseus]<br>Evalue: 0.0 \| Bitscore: 988<br>Identities = 476/478 (99%), Positives = 476/478 (99%), Gaps = 0/478 (0%) | NCBI | SHOW<br>BLAST |
| Q8GU20 | Strictosidine-O-beta-D-glucosidase OS=Rauvolfia serpentina GN=SGR1 PE=1 SV=1<br>Evalue: 0.0 \| Bitscore: 689<br>Identities = 334/479 (69%), Positives = 376/479 (78%), Gaps = 23/479 (4%) | SwissP | SHOW<br>BLAST |

Gene Structure

VIEW IN GENOMEVIEW | VIEW IN ARTEMIS

DOWNLOAD GENE IN EMBL FORMAT

| | |
|---|---|
| Structure | ;5179428..5180861; |
| Sequence Type | mRNA |
| Strand | - |
| Quality | 3 |

CDS

| | | |
|---|---|---|
| Locus ID | Caros009426.1 | REDO BLAST |
| CDS Length | 1434 nucleotides | |
| CDS Sequence | [illegible] | |

Protein

| | | |
|---|---|---|
| Locus ID | Caros009426.1 | REDO BLAST |
| Protein Length | 478 aminoacids | |
| Protein Sequence | [illegible] | |
| Signal Peptide | n/a | |
| Subcellular Localisation | n/a | |

Associated ESTs/cDNAs

n/a

Read Counts

normalized read counts

Experiments
flowers
suspension_culture (YE 0.3 mg/mL for 12h)
suspension_culture (YE 0.3 mg/mL for 24h)
suspension_culture (YE 0.0 mg/mL for 24h)
sterile_seedings 1
sterile_seedings (MeJA 6uM 12d)
sterile_seedings 2
suspension_culture (YE 0.3 mg/mL for 6h)
suspension_culture (MeJA 100uM for 6h)
suspension_culture (MeJA 100uM for 12h)
suspension_culture (MeJA 100uM for 24h)
suspension_culture
mature_leaf
immature_leaf
stem
root
hairy_root (TDCi)
hairy_root (RebH_F)
wild_type_hairy_root
widl_type_hairy_root (MeJA 250uM for 0h)
wild_type_hairy_root (MeJA 250uM for 24h)
hairy_roots TDCi (MeJA 250uM for 0h)
hairy_roots TDCi (MeJA 250uM for 24h)
Top

**Fig. 1** The *C. roseus* 'fake chromosomes' in the ORCAE database. (A) The ORCAE home page (http://bioinformatics.psb.ugent.be/orcae/) for the *C. roseus* transcriptome with visual representations of the 'fake chromosomes' and a search box for genes. (B) Caros009426.1, corresponding to SGD, has been given as an example. Clicking 'Go!' leads to a gene-specific webpage that displays gene IDs and functional annotation. (C) Gene ontologies and homologous genes from NCBI. (D) sequence data. (E) Atlas of expression data from publicly available *C. roseus* RNA-Seq analyses.

that provide graphical overviews of the whole metabolic network, single pathways and single reactions (see **Fig. 2** for an example). Furthermore, specific pages are available for genes, proteins and metabolites, complete with links to external online databases and to a dedicated gene expression database (http://bioinformatics.psb.ugent.be/orcae/) (see **Fig. 1** for an example). By using the set of tools provided by the website, users can query CathaCyc through keywords and searches by specific parameters. Furthermore, CathaCyc allows comparative analysis between external PDBs and the upload and graphical visualization of high-throughput experimental data.

## Manually curated *C. roseus* pathways

CathaCyc significantly covers the central metabolism of *C. roseus*; however, here we focused particularly on its secondary metabolism. We extensively curated the reconstruction of biochemical pathways that are relevant for the biotechnological and pharmaceutical exploitation of this species, i.e. the biosynthesis of TIAs and their precursors, as well as of the triterpenoids, another reasonably well studied class of *C. roseus* metabolites.

The pathways that produce the precursors for TIA biosynthesis, the chorismate, 2-*C*-methyl-D-erythritol 4-phosphate (MEP), mevalonate (MVA), indole and tryptamine pathways (El-Sayed and Verpoorte 2007), have been well described in multiple organisms and are consequently present in PDBs such as MetaCyc, PlantCyc and AraCyc. Our RNA-Seq data contained all transcripts of the above-mentioned pathways, and all enzymes they encode have therefore been included in CathaCyc v1.0. For 1-deoxy-D-xylulose-5-phosphate synthase (DXS), acetoacetyl-CoA thiolase (AACT), anthranilate phosphoribosyltransferase and D-3-phosphoglycerate dehydrogenase, multiple Caros gene copies have been annotated to the corresponding reactions.

The monoterpenoid compound secologanin is produced from geranyl pyrophosphate (GPP) involving at least eight enzymatic steps, of which five have been characterized in *C. roseus* (El-Sayed and Verpoorte 2007, Geu-Flores et al. 2012, Verma et al. 2012, Simkin et al., 2013) (**Fig. 2B**). Since only three were included in MetaCyc or PlantCyc, we have added the recently discovered genes encoding geraniol synthase (GES) (Simkin et al., 2013) and iridoid synthase (IS) (Geu-Flores et al. 2012) in CathaCyc v1.0 (**Fig. 3A, B**). Secologanin is further condensed with tryptamine by strictosidine synthase (STR), resulting in the production of strictosidine (**Fig. 2C**), which is further metabolized by strictosidine β-glucosidase (SGD) (Geerlings et al. 2000) to yield strictosidine aglycone, the precursor of all TIAs in *C. roseus* (Verma et al. 2012). Although this reaction is present in MetaCyc and PlantCyc, the *C. roseus* SGD (Geerlings et al., 2000) had not been enclosed within either of these databases. Therefore, we have included it in CathaCyc v1.0 (**Fig. 3C**). Furthermore, we found distinct transcripts encoding three isoforms of STR (**Fig. 2C**).

For the vindoline branch of the TIA pathway, which starts from tabersonine, one transcript, corresponding to the putative enzyme catalyzing the conversion of 16-methoxytabersonine to 3-hydroxy-16-methoxy 2,3-dihydrotabersonine, was not included in CathaCyc v1.0, since the pathway has not been further characterized at the gene level so far. Likewise, the conversion of α-3′-4′-anhydrovinblastine to vinblastine is not included in CathaCyc. In contrast, peroxidase 1, which catalyzes the condensation of catharanthine with vindoline (Costa et al. 2008) and is not included in MetaCyc, is presented in CathaCyc (**Fig. 3D**).

The pentacyclic triterpenoids ursolic acid and oleanolic acid have been found in considerable amounts in the cuticular wax layer of *C. roseus* leaves (Murata et al. 2008). Synthesis of these triterpenoids depends on the same precursor pathways as sterol synthesis, i.e. the MVA pathway, which delivers the isopentenyl pyrophosphate necessary for synthesis of 2,3-oxidosqualene, the common substrate for the oxidosqualene cyclases that catalyze the first committed steps towards the different branching triterpenoid pathways (Pollier et al. 2011). As for the TIA precursor pathways, all of these triterpenoid precursor pathways have been well described and included in the PDBs. Similarly, our RNA-Seq data contained all the corresponding transcripts, and all enzymes they encode have therefore been included in CathaCyc v1.0. Recently the genes encoding amyrin synthase (AAS) and amyrin C-28 oxidase (CYP716AL1/AO) have been isolated (Huang et al. 2012, Yu et al. 2013). AAS was characterized as a novel multifunctional oxidosqualene cyclase producing α- and β-amyrin, whereas AO was a multifunctional C-28 oxidase converting α-amyrin and β-amyrin to ursolic and oleanolic acid, respectively. Both genes are present in our RNA-Seq collection and have been added to CathaCyc v1.0 (**Fig. 4**).

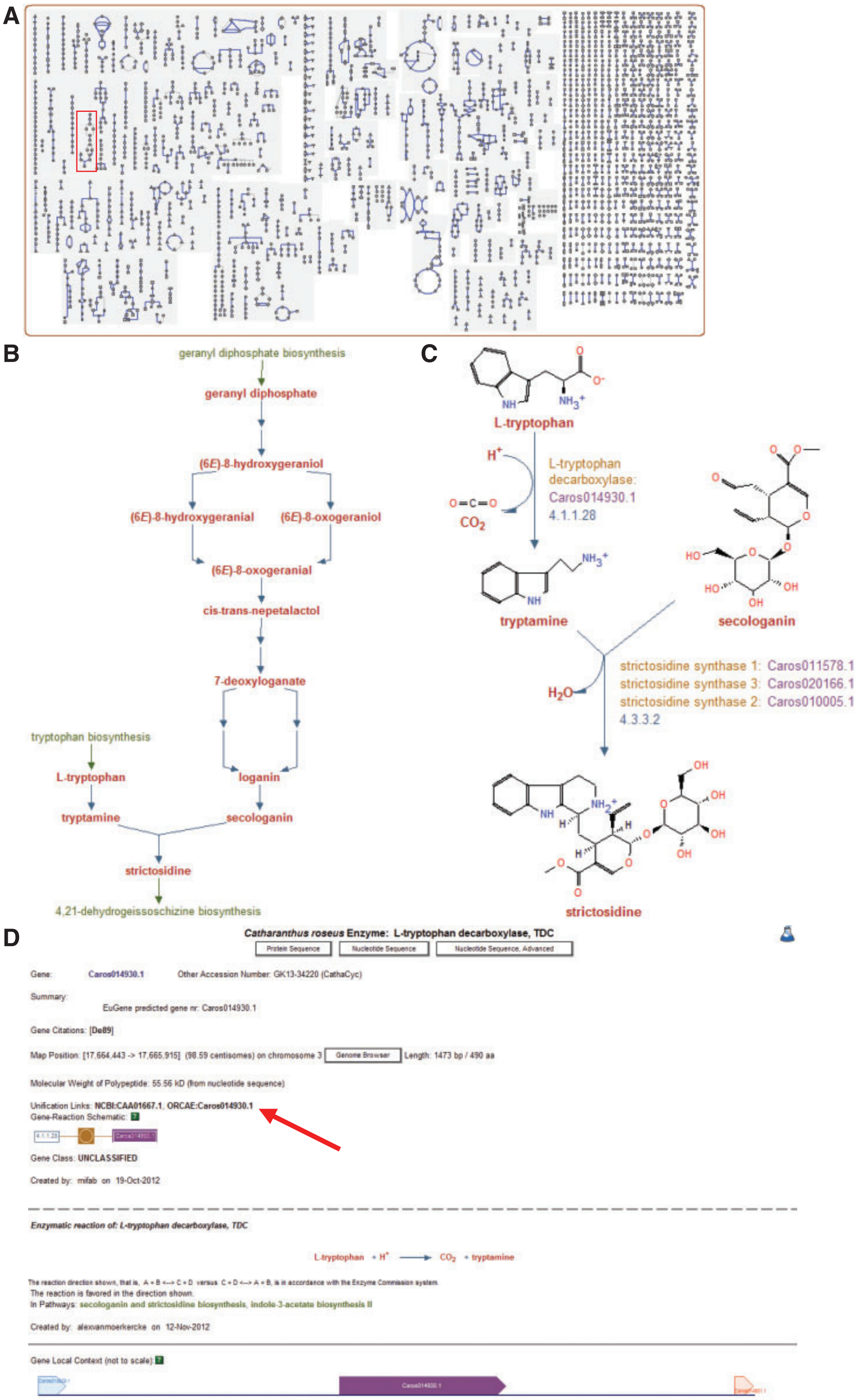
## The JA pathway

Furthermore, we curated some pathways involved in the metabolism of plant hormones, in particular in the biosynthesis of JAs, one of the main drivers of TIA synthesis in *C. roseus* and plant secondary metabolism in general (Rischer et al. 2006, De Geyter et al. 2012). The substrate for the synthesis of JAs is α-linolenic acid, which is released by lipase activity on chloroplast membranes. JAs comprise jasmonic acid, MeJA, JA–amino acid conjugates, such as (+)-7-iso-jasmonoyl-L-Ile (JA-Ile) that is now accepted as the endogenous bioactive JA, and further JA metabolites (Wasternack 2007, Pauwels and Goossens 2011). The corresponding biosynthetic pathways have been thoroughly characterized in *A. thaliana* mainly. For all known Arabidopsis JA and JA-Ile synthesis genes, the corresponding ortholog or homologs were detected in the CathaCyc database (**Supplementary Fig. S1A, B**). We also encountered two homologs of the gene encoding *S*-adenosyl-L-methionine:jasmonic acid carboxyl methyltransferase (JMT), that catalyzes the conversion of JA to MeJA. Since the corresponding pathway was missing from the MetaCyc databases, we created it in CathaCyc (**Supplementary Fig. S1C**). No candidate ortholog for the Arabidopsis gene shown to be involved in hydroxyjasmonate sulfate biosynthesis was found in the *C. roseus* transcriptome.

Overall, the level of redundancy in the JA synthesis genes from *C. roseus* was similar to that of Arabidopsis. For instance, for allene oxide synthase (AOS), 3-oxo-2(2′-[Z]-pentenyl)cyclopentane-1-octanoic acid CoA ligase (OPCL1) and the JA-Ile synthetase JASMONATE RESISTANT 1 (JAR1), only one hit was encountered in the CathaCyc database, whereas for all other JA synthesis genes two or more copies were detected (**Supplementary Fig. S1**), as is also the case in the Arabidopsis genome.

## The *C. roseus* RNA-Seq atlas

Several *C. roseus* RNA-Seq libraries are already publicly available, for instance from the MPGR consortium (http://medicinalplantgenomics.msu.edu/). However, to make full use of the information they contain, researchers have to download and process the vast amounts of data, which is currently beyond the capacity of many labs. Therefore, we created a *C. roseus* RNA-Seq atlas that archives all publicly available *C. roseus* expression data derived from RNA-Seq experiments. To generate the *C. roseus* RNA-Seq atlas, the MPGR reads were mapped to the artificial genome and counted. To allow the comparison of the expression of the genes in the different samples, the counts were normalized by transcript length and the total number of fragments to FPKM values (Trapnell et al. 2010). As a part of CathaCyc/ORCAE, the *C. roseus* RNA-Seq atlas allows the retrieval of sequences of genes of interest and visualization of the expression pattern of the genes in the different experimental conditions. Currently, the *C. roseus* RNA-Seq atlas holds the expression data from the MPGR consortium that comprise gene expression profiles of different *C. roseus* plant organs



*Catharanthus roseus* Enzyme: L-tryptophan decarboxylase, TDC

Gene: Caros014930.1 Other Accession Number: GK13-34220 (CathaCyc)

Summary:
EuGene predicted gene nr: Caros014930.1

Gene Citations: [De89]

Map Position: [17,664,443 -> 17,665,915] (98.59 centisomes) on chromosome 3 Length: 1473 bp / 490 aa

Molecular Weight of Polypeptide: 55.56 kD (from nucleotide sequence)

Unification Links: NCBI:CAA01667.1, ORCAE:Caros014930.1
Gene-Reaction Schematic:

Gene Class: UNCLASSIFIED

Created by: mifab on 19-Oct-2012

*Enzymatic reaction of: L-tryptophan decarboxylase, TDC*

L-tryptophan + $H^+$ ⟶ $CO_2$ + tryptamine

The reaction direction shown, that is, A + B <--> C + D versus C + D <--> A + B, is in accordance with the Enzyme Commission system.
The reaction is favored in the direction shown.
In Pathways: secologanin and strictosidine biosynthesis, indole-3-acetate biosynthesis II

Created by: alexvanmoerkercke on 12-Nov-2012

Gene Local Context (not to scale):

**Fig. 2** Screenshots from CathaCyc illustrating the different levels. (A) Cellular overview of the complete metabolism of *C. roseus*. The 'Secologanin and strictosidine biosynthesis' pathway, shown in more detail on the information page (B), is boxed in red. (B and C) Parts of the 'Secologanin and strictosidine biosynthesis' pathway, from the least to the most detailed view. The latter provides details of the Caros genes and proteins associated with the corresponding reactions. (D) Selection and visualization of a single reaction on a gene information page that, in turn, includes links to external databases and literature references, and information relative to gene length and protein size. The genomic localization of the protein is graphically represented and the genomic coordinates are indicated. The link to ORCAE is indicated by a red arrow.

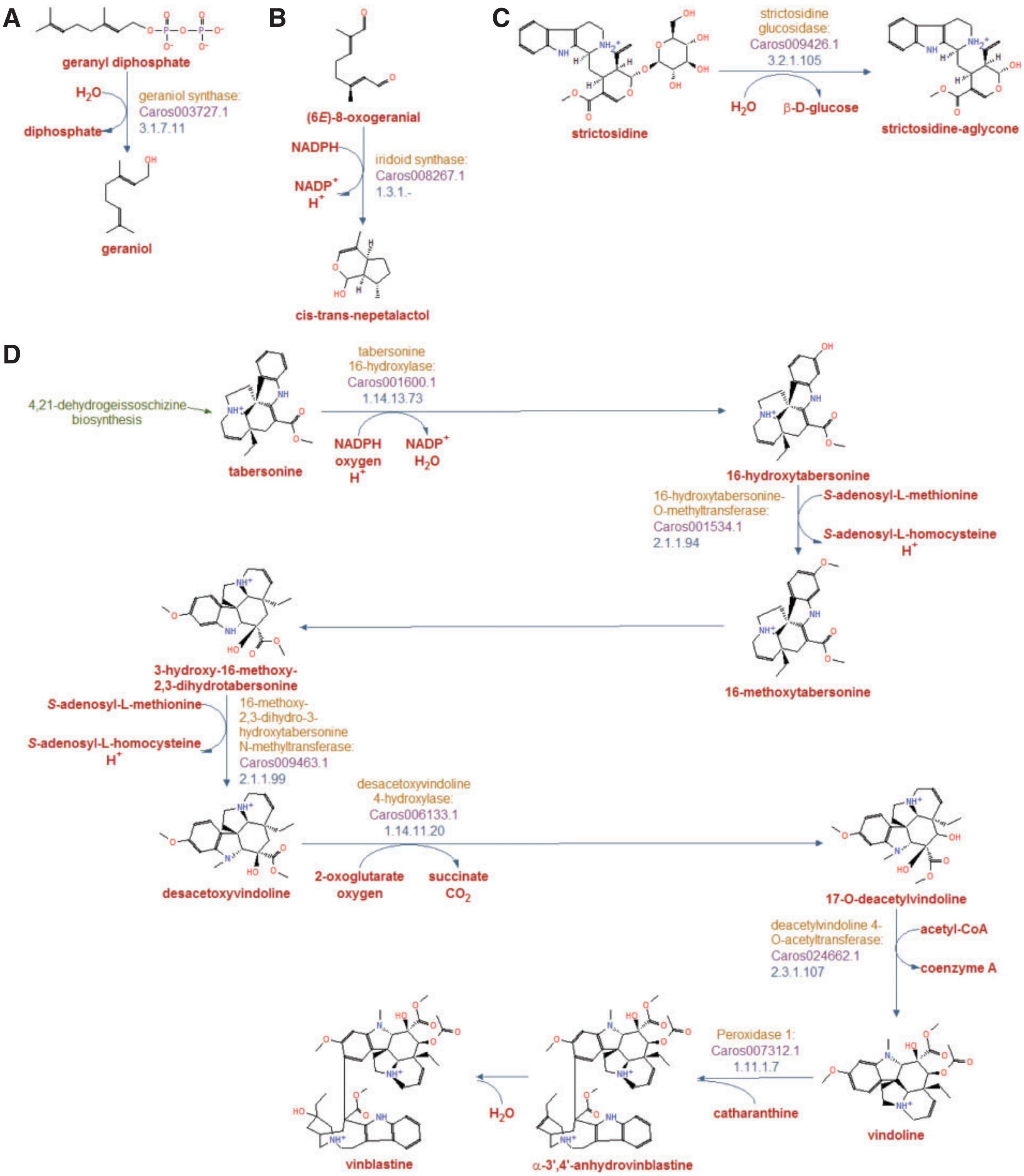


**Fig. 3** TIA pathways presented in CathaCyc. (A–C) Reactions from the 'Secologanin and strictosidine biosynthesis' pathway: (A) geraniol synthase, (B) iridoid synthase, (C) strictosidine glucosidase. (D) Vindoline and vinblastine biosynthesis.

and plant, suspension cell and hairy root cultures grown under different conditions, and totals 23 different samples (**Fig. 1E**).

## Triterpenoid and TIA biosynthesis is differentially regulated in *C. roseus* tissues

The *C. roseus* RNA-Seq atlas was used to analyze the expression of all CathaCyc genes from the curated TIA and triterpenoid pathways, as well as of their precursor and elicitor pathways. Average linkage hierarchical cluster analysis was performed with the full sample set from the MPGR collection or with the hairy root samples only (**Fig. 5**). Most of the genes were expressed in all explant types, except for deacetylvindoline-4-*O*-acetyltransferase (*DAT*), minovincinine-19-*O*-acetyltransferase (*MAT*), *AO*, *JMT1* and *JMT2* that were not expressed in non-

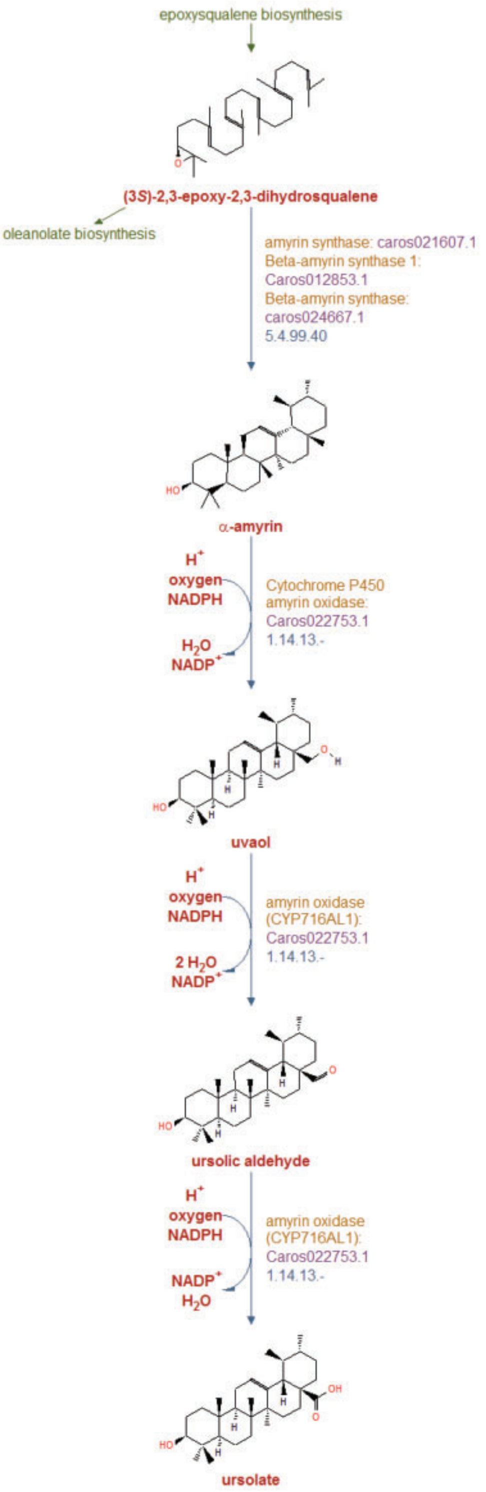


**Fig. 4** The ursolate pathway presented in CathaCyc.

elicited suspension cells (Magnotta et al. 2007), and tabersonine 16-hydroxylase (*T16H*) that was not expressed in non-elicited hairy roots (Schröder et al. 1999), respectively.

When using the full MPGR sample set, distinct regulation between the TIA and triterpenoid genes was apparent. Genes from both pathways clustered together in a group with lower expression in suspension cells. However, within this group, most TIA genes assembled together in a subcluster with relatively high expression in hairy root cultures (**Fig. 5A**), whereas nearly all triterpenoid genes, including *AAS*, *AO*, all those of the oxidosqualene precursor pathway and most of those of the MVA pathway, assembled in a subcluster that showed higher expression in seedlings and/or other whole-plant organs (**Fig. 5B**). The latter is in agreement with the reported accumulation profile of triterpenoid compounds in planta (Murata et al. 2008).

Assessing the expression within the hairy root sample set only revealed another marked differential regulation between both pathways. In agreement with the existing literature, expression of all TIA pathway genes was MeJA inducible (**Fig. 5C**). The only exception was the *N*-methyltransferase *NMT* (Liscombe et al. 2010), which was repressed following MeJA treatment. Conversely, the expression of *AAS*, *AO* and some genes encoding rate-limiting enzymes from the triterpenoid precursor pathways, such as 3-hydroxy-3-methylglutaryl-CoA reductase (*HMGR*) and squalene epoxidase (*SQE*), was repressed in hairy roots by MeJA treatment (**Fig. 5D**). The latter observation is remarkable considering that triterpenoid biosynthesis has been reported to be JA inducible in many different plant species (Pauwels et al. 2009, Yendo et al. 2010).

Interestingly, BLAST searches of the CathaCyc sequences disclosed the presence of clear homologs of two of the *Rauvolfia serpentina* genes that encode enzymes catalyzing steps in the conversion of strictosidine to vinorine, i.e. Caros023628.1 for vinorine synthase (VS; Bayer et al. 2004) and Caros017545.1 for polyneuridine aldehyde esterase (PNAE; Dogru et al. 2000), respectively. *Catharanthus roseus* is not known to produce vinorine or derivatives thereof, but expression analysis shows co-regulation of both the *VS* and *PNAE* homolog with the known TIA genes (**Fig. 5A, C**), suggesting that both genes might be involved in TIA synthesis in *C. roseus* as well.

## Discussion

*Catharanthus roseus* is the single biological source of the anti-cancer compounds vinblastine and vincristine (El-Sayed and Verpoorte 2007). Efforts to increase the production of these compounds have been moderately successful, partly because a significant number of steps of the pathways that produce these compounds have remained uncharacterized in *C. roseus*. The identification of missing steps can be simplified using a metabolic database of *C. roseus* that catalogs its metabolic potential based on annotated sequence data. In addition, metabolic databases allow for the visualization and interpretation of large-scale omics data. An increasing number of plant metabolic databases is being constructed (Urbanczyk-Wochniak and Sumner 2007, May et al. 2009, Mazourek et al. 2009, Zhang et al. 2010, Fabris et al. 2012).

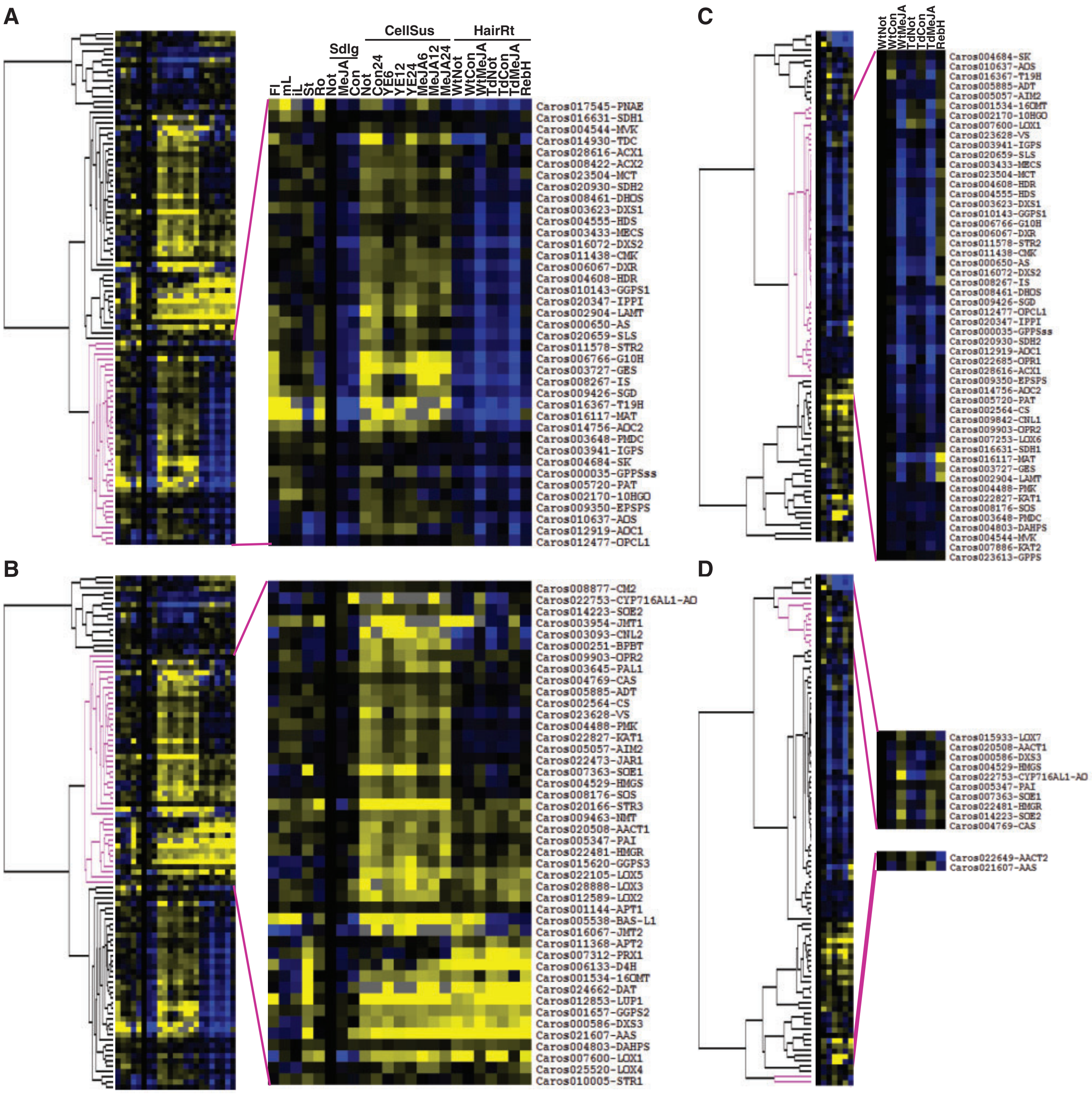


**Fig. 5** Average linkage hierarchical clustering of secondary metabolism-related transcriptomes from *C. roseus*. (A) Clustering of the TIA genes based on all MPGR expression data. (B) Clustering of the triterpenoid genes based on all MPGR expression data. (C) Clustering of the TIA genes based on hairy root expression data. (D) Clustering of the triterpenoid genes based on hairy root expression data. Each panel shows the full cluster and a subcluster on the left and right, respectively. Blue and yellow boxes reflect transcriptional activation and repression, respectively, relative to the expression level of the non-treated seedlings (Not, for A and B) and non-treated wild-type hairy roots (WtNot, for C and D). Gray boxes correspond to no expression for a particular gene in a particular explant. Average linkage hierarchical cluster analysis was performed with the CLUSTER and TREEVIEW software (Eisen et al, 1998) and using the log10-transformed values of the normalized FPKM values as input for CLUSTER. Caros numbers and enzyme names are indicated at the right (for enzyme abbreviations, see **Supplementary Table S4**). The explants and/or treatments and the time points (in hours) are indicated at the top. Abbreviations: Fl, flower; mL, mature leaves; iL, immature leaves; St, stem; Ro, root; Sdlg, seedling; CellSus, suspension cells; HairRt, hairy roots; Not, non-treated; MeJA, MeJA elicited; Con, mock-treated; YE, yeast-elicitor treated; Wt, wildtype; Td, TDCi hairy roots; RebH, RebH_F hairy roots (see http://medicinalplantgenomics.msu.edu/ for sample details).

These species-specific databases are generally created from reference databases such as MetaCyc, which contains many non-plant-specific pathways, and therefore require extensive literature-based curation (Zhang et al. 2010). In addition, highly significant pathways from medicinal plants are mostly absent or only partially included in these reference databases. For this reason, there is a need to create metabolic databases of medicinal plants, which are characterized by their highly specialized metabolism.

The genomes of many pharmaceutically important medicinal plants like *C. roseus* have not been sequenced. The use of transcriptomes rather than genomes to reconstruct pathway databases can circumvent the necessity for fully assembled and annotated genomes. By doing so, PDBs reconstructed from large EST collections have been built for a limited number of plant species such as *M. truncatula* (Urbanczyk-Wochniak and Sumner 2007) and some Solanaceae (SolCyc; SGN). To our knowledge, the use of RNA-Seq data for the creation of plant PDBs has not been demonstrated.

We assembled and annotated reads of an RNA-Seq experiment of the *C. roseus* transcriptome spanning different plant and cell material. With this information, we created CathaCyc v1.0, which significantly covers the central metabolism of *C. roseus*. Here, however, we focused particularly on its secondary metabolism. We extensively curated the reconstruction of biochemical pathways that are relevant for the biotechnological and pharmaceutical exploitation of this species, i.e. the biosynthesis of TIAs and their precursors, as well as of oleanane- and ursane-type triterpenoids. Furthermore, we also concentrated on the reconstruction of pathways involved in the metabolism of plant hormones, in particular in the biosynthesis of JAs, one of the main drivers of TIA synthesis in *C. roseus* and plant secondary metabolism in general (De Geyter et al. 2012). Starting from the transcriptome data, we were able to identify transcripts for all known steps in these pathways, illustrating the quality of the RNA-Seq data set, as well as its utility to reconstruct a metabolic PDB.

As with most metabolic network reconstructions, CathaCyc v1.0 still contains some pathway holes. Many of these holes might correspond to reactions with an incomplete or missing EC number, which is due to the absence of sufficient knowledge to link them accurately to a specific enzyme, or to reactions for which the predicted candidate genes show a confidence score that was lower than the minimal threshold. Some pathway holes, however, might also originate from the discrepancy between the pathway templates stored in MetaCyc that are often inferred from model organisms, and the real pathway organization that occurs in *C. roseus*, which can be specifically rearranged or slightly modified.

Next-generation sequencing technologies demand increasing efforts to assemble and annotate sequencing projects. The enzymes annotated in CathaCyc v1.0 are linked to the online genome annotation web interface ORCAE (Sterck et al. 2012), which allows viewing and editing of the initial automatic predictions. Users are thus able to contribute to the completeness and accuracy of the annotation. In addition, it enables the integration and visualization of large-scale omics data. As an example of the latter, we have mapped the publicly accessible RNA-Seq data of the MPGR consortium, spanning 23 different tissues and treatments (Góngora-Castillo et al., 2012), to our annotated *C. roseus* transcriptome. By doing so, expression of the genes of individual pathways can be viewed in the different conditions. Analysis of the expression of the genes from the curated CathaCyc v1.0 pathways demonstrated that the two well-characterized *C. roseus* terpenoid pathways, i.e. of TIAs and triterpenoids, are subject to distinct regulation by both developmental and environmental cues. Furthermore, BLAST searches of the CathaCyc sequences identified clear homologs of *R. serpentina* genes that encode enzymes catalyzing steps in the conversion of strictosidine to vinorine and that show co-regulation with the known *C. roseus* TIA genes. Since *C. roseus* is not known to produce vinorine or derivatives thereof, these genes might correspond to missing steps in the TIA pathway. These lead findings warrant further exploitation and underscore the value of CathaCyc for the study and exploitation of the metabolism of the Madagascar periwinkle.

## Materials and Methods

### Illumina HiSeq2000 RNA sequencing

The *Catharanthus roseus* cell line was treated with 10 μM MeJA (Bedoukian Research Inc.) for 24 h. The control treatment was with the solvent dimethylsulfoxide (DMSO) at 0.1% final concentration. The cell line and RNA extraction method were described in Pauw et al. (2004).

*Catharanthus roseus* plants were germinated from seeds originally received from the Botanical Garden Wuerzburg (Germany). One clone was maintained and vegetatively propagated in a growth chamber in pots containing a substrate mixture [soil (Karkea Ruukutusseos):vermiculite 50:50] at 27°C, 16 h photoperiod, 110 μmol $s^{-1}$ $m^{-2}$ and 66.4% humidity. For the elicitation experiment, shoots (approximately 2 g FW) from flowering plants were cut with sterile scissors and individually placed in cups containing 20 ml of water. The shoots were left to recover from the cutting for 10 d in the growth chamber. Evaporated water was replaced during this time. Then the cups with the shoots were individually placed in zip-lock bags. The elicitation was started by adding MeJA solution to a final concentration of 1 mM or an equivalent amount of the solvent DMSO (800 μl) as a control. The bags were tightly closed and incubated under the same conditions for 0, 6 and 24 h, respectively. Five replicates per treatment were used. Finally, samples were shock-frozen in liquid nitrogen and stored at –20°C until extraction. Total RNA from *C. roseus* shoots was prepared with the RNeasy Mini Kit (Qiagen).

RNA samples were sent to Fasteris Life Sciences SA for mRNA purification, cDNA library construction and Illumina HiSeq2000-based RNA sequencing with the Solexa technology. For both the cell and plant samples, non-normalized and duplex-specific nuclease (DSN)-normalized libraries were prepared and sequenced. In brief, processing included DNase treatment, transcript purification, transcript breaking, double-stranded cDNA synthesis using random primers and RNase H, ends repair, 3′ A addition, ligation of adaptors, gel purification to isolate fragments of an insert size of 150–250 bp, PCR amplification to generate the DNA Colonies Template Library, and library purification. A subset of the libraries was additionally normalized with the DSN protocol. The Hi-Seq 2000 instrument and the TruSeq$^{TM}$ SBS v5 kit were used for paired-end (2 × 100 bp) and indexed sequencing of the RNA-Seq libraries.

## Assembly and annotation of the 'reference transcriptome'

To construct a reference transcriptome for *C. roseus*, the generated raw RNA-Seq reads were assembled de novo by the sequencing service provider using VELVET (v1.1.04) (Zerbino and Birney 2008) and its module OASES (v0.1.21) (Schulz et al. 2012) into two sets of contigs, corresponding to the cell and plant samples, respectively. For VELVET, the pair-read insert average size was set at 200 bases with an SD of 10%. No kmer average coverage and coverage threshold for nodes were used because of the high dynamic range of gene expression. Validation mapping indicated that the highest representativity was found for the assembly of hash 87, which was selected for the complete mapping for both sets of contigs.

The *C. roseus* unigene set was created by combining and joining both sets using the CAP3 software with default parameters (Huang and Madan 1999). On this unigene set we predicted ORFs using the FrameDP (Gouzy et al. 2009) software. Each transcript was assigned a functional description through a guilt-by-homology approach. Therefore, BLASTX was run with the transcript sequences and was preferred to BLASTP with the predicted ORFs as the occasional indels would not affect hits too much for BLASTX. The BLASTX (v2.2.24) parameters were: -evalue 0.001 -num_descriptions 100 -num_alignments 100 -seg yes. The protein blast database was the non-redundant protein database supplemented with tomato proteins including the human readable descriptions. The BLASTX output was then parsed using perl programs and hits scored according to their bitscore and number of occurring descriptions using a comparable choice of words. To accommodate further analysis and storage of data, we made seven 'fake chromosomes' by concatenating the set of transcripts joined by a spacer of 2,000 N. These data were all loaded in the database with the web interface of the ORCAE (Sterck et al. 2012) platform. The integrated data from ORCAE were extracted and appropriately formatted into GenBank files to upload the data in a BioCyc system dedicated to *C. roseus*.

## Creation of CathaCyc (orthology prediction and database reconstruction)

Translated annotated genomic sequences were downloaded from PLAZA [http://bioinformatics.psb.ugent.be/plaza/, April 2012; Van Bel et al. (2012)] for *M. truncatula*, *Zea mays*, *Ricinus communis*, *P. trichocarpa*, *Vitis vinifera*, *Malus domestica*, *Brachypodium distachyon*, *Physcomitrella patens*, *Carica papaya*, *Glycine max* and *O. sativa*, or from the KEGG database [http://www.genome.jp/kegg/, February 2010; Kanehisa and Goto (2000)] for *E. coli*, *S. cerevisiae*, *H. sapiens*, *P. tricornutum*, *C. reinhardtii* and *A. thaliana*, respectively, and were used as the reference data set for the orthology prediction. Orthology prediction, semi-automated functional annotation, BioCyc database construction and manual curation have been done as described previously (Fabris et al. 2012). Pathway-gap filling was performed with the Pathway Hole Filler utility of Pathologic (Caspi et al. 2012) in the user-driven mode. Available gene–reaction associations of *C. roseus* were used as a training data set. We confirmed the correctness of a random group of 20 candidate genes (with a confidence score of $\geq$0.1) by manual BLAST, InterProscan (Hunter et al. 2009) and Inparanoid (O'Brien et al. 2005) searches. Therefore, the minimal score for candidate genes to be automatically selected as Hole Fillers was set to 0.1, and any gene with a lower score was excluded.

## Mapping of the transcriptome data

The FASTQ files containing the read sequences and quality scores of the MPGR consortium were extracted from the NCBI Short Read Archive accessions (accession No. SRA030483) using the NCBI SRA Toolkit version 2.1.7. Processing of the extracted reads, mapping of the reads on the 'artificial genome' with TOPHAT version 2.0.3 (Trapnell et al. 2009) and counting of the uniquely mapped reads and calculation of the FPKM values with CUFFLINKS version 1.3.0 (Trapnell et al. 2010) were performed using default parameters as described (Pollier et al. 2013).

## Supplementary data

Supplementary data are available at PCP online.

## Funding

This work was supported by the European Union Seventh Framework Programme FP7/2007-2013 [222716 - SMARTCELL]; the European Cooperation in Science and Technology Action [FA1006 - PlantEngine]; the Agency for Innovation by Science and Technology in Flanders ['Strategisch Basisonderzoek' 80031 - SUNLIGHT]; the European Molecular Biology Organization [Long-Term Fellowship to A.V.M.] and European Commission support from Marie Curie Actions [EMBOCOFUND2010, GA-2010-267154 to A.V.M].

## Acknowledgments

We thank Annick Bleys for help in preparing the manuscript, and Mari Lämsä, Robin Vanden Bossche, Thomas Van Parys and Dany Cuyt for technical assistance.

## References

Bayer, A., Ma, X. and Stöckigt, J. (2004) Acetyltransfer in natural product biosynthesis—functional cloning and molecular analysis of vinorine synthase. *Bioorg. Med. Chem.* 12: 2787–2795.

Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F. et al. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 38: D473–D479.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M. et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 40: D742–D753.

Costa, M.M.R., Hilliou, F., Duarte, P., Pereira, L.G., Almeida, I., Leech, M. et al. (2008) Molecular cloning and characterization of a vacuolar class III peroxidase involved in the metabolism of anticancer alkaloids in *Catharanthus roseus*. *Plant Physiol.* 146: 403–417.

De Geyter, N., Gholami, A., Goormachtig, S. and Goossens, A. (2012) Transcriptional machineries in jasmonate-elicited plant secondary metabolism. *Trends Plant Sci.* 17: 349–359.

Desgagné-Penix, I., Farrow, S.C., Cram, D., Nowak, J. and Facchini, P.J. (2012) Integration of deep transcript and targeted metabolite profiles for eight cultivars of opium poppy. *Plant Mol. Biol.* 79: 295–313.

Dogru, E., Warzecha, H., Seibel, F., Haebel, S., Lottspeich, F. and Stöckigt, J. (2000) The gene encoding polyneuridine aldehyde esterase of monoterpenoid indole alkaloid biosynthesis in plants is an ortholog of the α/α hydrolase super family. *Eur. J. Biochem.* 267: 1397–1406.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 95: 14863–14868.

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

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A. and Baart, G.J.E. (2012) The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner–Doudoroff glycolytic pathway. *Plant J.* 70: 1004–1014.

Facchini, P.J. and De Luca, V. (2008) Opium poppy and Madagascar periwinkle: model non-model systems to investigate alkaloid biosynthesis in plants. *Plant J.* 54: 763–784.

Facchini, P.J. and St-Pierre, B. (2005) Synthesis and trafficking of alkaloid biosynthetic enzymes. *Curr. Opin. Plant Biol.* 8: 657–666.

Ferreres, F., Figueiredo, R., Bettencourt, S., Carqueijeiro, I., Oliveira, J., Gil-Izquierdo, A. et al. (2011) Identification of phenolic compounds in isolated vacuoles of the medicinal plant *Catharanthus roseus* and their interaction with vacuolar class III peroxidase: an $H_2O_2$ affair? *J. Exp. Bot.* 62: 2841–2854.

Geerlings, A., Ibañez, M.M.-L., Memelink, J., van der Heijden, R. and Verpoorte, R. (2000) Molecular cloning and analysis of strictosidine β-D-glucosidase, an enzyme in terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *J. Biol. Chem.* 275: 3051–3056.

Geu-Flores, F., Sherden, N.H., Courdavault, V., Burlat, V., Glenn, W.S., Wu, C. et al. (2012) An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. *Nature* 492: 138–142.

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

Gouzy, J., Carrere, S. and Schiex, T. (2009) FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics* 25: 670–671.

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

He, M., Wang, Y., Hua, W., Zhang, Y. and Wang, Z. (2011) De novo sequencing of *Hypericum perforatum* transcriptome to identify potential genes involved in the biosynthesis of active metabolites. *PLoS One* 7: e42081.

Huang, L., Li, J., Ye, H., Li, C., Wang, H., Liu, B. et al. (2012) Molecular characterization of the pentacyclic triterpenoid biosynthetic pathway in *Catharanthus roseus*. *Planta* 236: 1571–1581.

Huang, X. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.* 9: 868–877.

Hughes, E.H., Hong, S.-B., Gibson, S.I., Shanks, J.V. and San, K.-Y. (2004) Metabolic engineering of the indole pathway in *Catharanthus roseus* hairy roots and increased accumulation of tryptamine and serpentine. *Metab. Eng.* 6: 268–276.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D. et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37: D211–D215.

Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K. et al. (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.* 34: D717–D723.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28: 27–30.

Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J. et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.* 11: 40–79.

Karp, P.D., Riley, M., Paley, S.M., Pellegrini-Toole, A. and Krummenacker, M. (1997) EcoCyc: enyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* 25: 43–50.

Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M. et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 32: D438–D442.

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

Loyola-Vargas, V.M., Galaz-Ávalos, R.M. and Kú-Cauich, R. (2007) *Catharanthus* biosynthetic enzymes: the road ahead. *Phytochem. Rev.* 6: 307–339.

Magnotta, M., Murata, J., Chen, J. and De Luca, V. (2007) Expression of deacetylvindoline-4-*O*-acetyltransferase in *Catharanthus roseus* hairy roots. *Phytochemistry* 68: 1922–1931.

May, P., Christian, J.-O., Kempa, S. and Walther, D. (2009) ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii*. *BMC Genomics* 10: 209.

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

Mueller, L.A., Zhang, P. and Rhee, S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.* 132: 453–460.

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

Mustafa, N.R. and Verpoorte, R. (2007) Phenolic compounds in *Catharanthus roseus*. *Phytochem. Rev.* 6: 243–258.

O'Brien, K.P., Remm, M. and Sonnhammer, E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33: D476–D480.

Paley, S.M., Latendresse, M. and Karp, P.D. (2012) Regulatory network operations in the Pathway Tools software. *BMC Bioinformatics* 13: 243.

Pauw, B., van Duijn, B., Kijne, J.W. and Memelink, J. (2004) Activation of the oxidative burst by yeast elicitor in *Catharanthus roseus* cells occurs independently of the activation of genes involved in alkaloid biosynthesis. *Plant Mol. Biol.* 55: 797–805.

Pauwels, L. and Goossens, A. (2011) The JAZ proteins: a crucial interface in the jasmonate signaling cascade. *Plant Cell* 23: 3089–3100.

Pauwels, L., Inzé, D. and Goossens, A. (2009) Jasmonate-inducible gene: what does it mean? *Trends Plant Sci.* 14: 87–91.

Pollier, J., Moses, T. and Goossens, A. (2011) Combinatorial biosynthesis in plants: a (p)review on its potential and future exploitation. *Nat. Prod. Rep.* 28: 1897–1916.

Pollier, J., Rombauts, S. and Goossens, A. (2013) Analysis of RNA-Seq data with TOPHAT and CUFFLINKS for genome-wide expression analysis of jasmonate-modulated plant transcriptomes. *Methods Mol. Biol.* 1011 (in press).

Rischer, H., Orešič, M., Seppänen-Laakso, T., Katajamaa, M., Lammertyn, F., Ardiles-Diaz, W. et al. (2006) Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. *Proc. Natl Acad. Sci. USA* 103: 5614–5619.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M. and Karp, P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* 6: R2.

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

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

Simkin, A.J., Miettinen, K., Claudel, P., Burlat, V., Guirimand, G., Courdavault, V. et al. (2013) Characterization of the plastidial geraniol synthase from Madagascar periwinkle which initiates the monoterpenoid branch of the alkaloid pathway in internal phloem associated parenchyma. *Phytochemistry* 85: 36–43.

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

Sterck, L., Billiau, K., Abeel, T., Rouzé, P. and Van de Peer, Y. (2012) ORCAE: online resource for community annotation of eukaryotes. *Nat. Methods* 9: 1041.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.

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

Urbanczyk-Wochniak, E. and Sumner, L.W. (2007) MedicCyc: a biochemical pathway database for *Medicago truncatula*. *Bioinformatics* 23: 1418–1423.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y. et al. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.* 158: 590–600.

van der Heijden, R., Jacobs, D.I., Snoeijer, W., Hallard, D. and Verpoorte, R. (2004) The Catharanthus alkaloids: pharmacognosy and biotechnology. *Curr. Med. Chem.* 11: 607–628.

Verma, P., Mathur, A.K., Srivastava, A. and Mathur, A. (2012) Emerging trends in research on spatial and temporal organization of terpenoid indole alkaloid pathway in *Catharanthus roseus*: a literature update. *Protoplasma* 249: 255–268.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57–63.

Wasternack, C. (2007) Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Ann. Bot.* 100: 681–697.

Wenping, H., Yuan, Z., Jie, S., Lijun, Z. and Zhezhi, W. (2011) De novo transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. *Genomics* 98: 272–279.

Yendo, A.C.A., de Costa, F., Gosmann, G. and Fett-Neto, A.G. (2010) Production of plant bioactive triterpenoid saponins: elicitation strategies and target genes to improve yields. *Mol. Biotechnol.* 46: 94–104.

Yu, F., Thamm, A.M.K., Reed, D., Villa-Ruano, N., Quesada, A.L., Gloria, E.L. et al. (2013) Functional characterization of amyrin synthase involved in ursolic acid biosynthesis in *Catharanthus roseus* leaf epidermis. *Phytochemistry* (in press).

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18: 821–829.

Zhang, P., Dreher, K., Karthikeyan, A., Chi, A., Pujar, A., Caspi, R. et al. (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.* 153: 1479–1491.

Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D. et al. (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.* 138: 27–37.

Zhou, M.-L., Shao, J.-R. and Tang, Y.-X. (2009) Production and metabolic engineering of terpenoid indole alkaloids in cell cultures of the medicinal plant *Catharanthus roseus* (L.) G. Don (Madagascar periwinkle). *Biotechnol. Appl. Biochem.* 52: 313–323.

Ziegler, J. and Facchini, P.J. (2008) Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.* 59: 735–769.