The Vinca minor genome highlights conserved evolutionary traits in monoterpene indole alkaloid synthesis

Abstract

Vinca minor, also known as the lesser periwinkle, is a well-known species from the Apocynaceae family, native to central and southern Europe. This plant synthesizes monoterpene indole alkaloids, which are a class of specialized metabolites displaying a wide range of bioactive- and pharmacologically important properties. Within the almost 50 monoterpene indole alkaloids it produces, V. minor mainly accumulates vincamine, which is commercially used as a nootropic. Using a combination of Oxford Nanopore Technologies long read- and Illumina short-read sequencing, a 679,098 Mb V. minor genome was assembled into 296 scaffolds with an N50 scaffold length of 6 Mb, and encoding 29,624 genes. These genes were functionally annotated and used in a comparative genomic analysis to establish gene families and to investigate gene family expansion and contraction across the phylogenetic tree. Furthermore, homology-based monoterpene indole alkaloid gene predictions together with a metabolic analysis across 4 different V. minor tissue types guided the identification of candidate monoterpene indole alkaloid genes. These candidates were finally used to identify monoterpene indole alkaloid gene clusters, which combined with synteny analysis allowed for the discovery of a functionally validated vincadifformine-16-hydroxylase, reinforcing the potential of this dataset for monoterpene indole alkaloids gene discovery. It is expected that access to these resources will facilitate the elucidation of unknown monoterpene indole alkaloid biosynthetic routes with the potential of transferring these pathways to heterologous expression systems for large-scale monoterpene indole alkaloid production.

Keywords: Vinca minor; lesser periwinkle; genome; alkaloids; vincadifformine

Introduction

Monoterpenoid indole alkaloids (MIAs) are a class of specialized (also known as secondary) metabolites produced by plants from the Gentianale order families, including Gentianaceae, Apocynaceae, Loganiaceae, and Rutaceae but also by a few Nymphaeaceae from the Cornale order (O'Connor and Maresh 2006). MIAs display a broad spectrum of bioactive properties and thus belong to the plethora of chemical arsenals that plants evolved to cope with environmental pressure and notably abiotic and biotic attacks (Dugé de Bernonville et al. 2015). These biological activities make MIAs attractive pharmaceuticals of high economic value, as illustrated by ajmaline from Rauwolfia species (Apocynaceae) prescribed for the treatment of arrhythmia as well as vinblastine and vincristine from Catharanthus roseus (Apocynaceae) that are potent anticancer compounds (O’Connor and Maresh 2006). The lesser periwinkle, Vinca minor, also belongs to the Apocynaceae family and accumulates more than 50 MIAs (Proksa and Grossmann 1991; D’Amelio Sr et al. 2012; Hasa et al. 2013; Vrabec et al. 2022) including aspidosperma-type MIAs such as vincadifformine and eburnamine-type MIAs such as vincamine which is commercially used as a vasodilator together with its semisynthetic derivative vinpocetine (Vas and Gulyás 2005) (Fig. 1, a–c).

MIAs result from long and complex biosynthetic pathways sharing high levels of similarity among species (Fig. 1a). With only a few exceptions, all MIAs originate from strictosidine that is produced by the condensation of secolloganin and tryptamine, the monoterpenes, and indole MIA precursors, respectively. The subsequent strictosidine decorations direct MIA synthesis toward the almost 3,000 described MIAs. To date, most of our knowledge on these manifold reactions arises from C. roseus (Courdavault et al. 2014; Kulagina et al. 2022). This notably led to the characterization of the biosynthetic routes towards major MIAs such as catharanthine and tabersonine (Tatsis et al. 2017; Caputi et al. 2018; Qu et al. 2019). Interestingly, tabersonine can also be...
diverted to the famous vindoline through a 7-step pathway initiated by a couple of tabersonine 16-hydroxylases (T16H; Besseau et al. 2013) and an associated 16-hydroxytabersonine 16-O-methyltransferase (16OMT) (Levac et al. 2008). In V. minor, while vincamine is highly accumulated (Proksa and Grossmann 1991), its biosynthesis remains mostly unknown. However, it has been suggested that this compound derives from vincadifformine whose synthesis relies on 2 vincadifformine synthases (VS) as described in C. roseus (Williams et al. 2019; Caputi et al. 2020). Interestingly, a P450 from C. roseus has been proposed to catalyze a cyclization of tabersonine/vincadifformine leading to the synthesis of an eburnamine-vincamine skeleton product but no ortholog from the lesser periwinkle has been identified to date (Kellner et al. 2015a).

Overall, only a few MIA biosynthetic genes have been identified and characterized in V. minor including a picrinine N-methyltransferase, (VmPiNMT; Levac et al. 2016), a vincamine/vincadifformine ATP binding cassette transporter (VmTPT2/ VmABCG1; Demessie et al. 2017), and a vincadifformine 16-O-methyltransferase (Vm16OMT; Stander et al. 2020).

While almost all MIA biosynthetic gene identifications have resulted from transcriptomics and gene co-expression analyses to date (Dugé de Béronville et al. 2020), the ever-growing access to plant genome sequences opens new perspectives towards MIA pathway elucidation (Stander et al. 2022). To date, 7 nuclear genomes of MIA producing species have been sequenced thus providing new insights into the specific MIA synthesis of each plant species (Table 1). Here, we report the genome assembly, annotation, and analysis of V. minor combined with the identification and functional validation of vincadifformine 16-hydroxylase. Such a genomic resource will potentially pave the way for future MIA biosynthetic gene identification in this prominent medicinal plant species.

**Materials and methods**

**Sample collection**

Wild-growing V. minor plants were collected in Tours, France (472101.1400 N 042008.200 E) for direct DNA extraction and MIA quantification.

**Chemicals**

Vincadifformine was purchased from Biosynth Carbosynth (UK). 16-hydroxyvinca-difformine was produced by a yeast strain expressing C. roseus T16H2 via 16-hydroxylation of vincadifformine (Stander et al. 2020).

**DNA extraction and sequencing**

Nuclei were first isolated from young leaves following the procedure described in Workman et al. (2018). High-molecular weight DNA was extracted from V. minor leaves using the Nanobind HMW DNA Extraction Circulomics kit (Circulomics Inc., Baltimore, MD, USA) as per manufacturer’s instructions. For Illumina sequencing, a sequencing library was constructed using the TruSeq DNA PCR-Free Library Prep kit (Illumina, San Diego, USA) and sequenced in paired-end mode (2 x 100 bp) by Eurofins Genomics (Les Ulis, France) using Illumina Novaseq 6000 technology. Oxford Nanopore Technologies (ONT) library construction and sequencing were performed at Future Genomics Technologies (Leiden, The Netherlands). Library was constructed from approximately 1 μg of high-molecular weight DNA using Ligation sequencing kit (SQK-LSK109, Oxford Nanopore Technologies Ltd) and sequenced on a Nanopore PromethION flowcell (FLO-PRO002, Oxford Nanopore Technologies Ltd) with the guppy version 3.2.6 high-accuracy basecaller.
The estimated genome length and heterozygosity were determined by calculating the k-mer frequencies (k = 21) from Illumina short-reads with Jellyfish (v.2.3.0, Marciais and Kingsford 2011) using the following parameters: jellyfish count -c -m 21 -s 1000000000 -t 10. The resulting k-mer count histogram was imported into Genomescope (Vurture et al. 2017).

De novo genome assembly

The V. minor genome was assembled by Future Genomics Technologies (Leiden, The Netherlands). Firstly, the ONT reads were assembled into contigs with the Flye assembler (v.2.5, Kolimogorov et al. 2019) using the following parameters: --narrow --estimated_genome_size 1400M --iterations 2. Contigs containing redundant sequences were removed with purge_haplotigs (commit 981bee4) using the following parameters: -l 15 -m 50 -h 70. Next, 2 rounds of polishing were performed with Illumina paired-end reads using pilon (v.1.23, Walker et al. 2014) which resulted in a final set of 296 contigs.

Gene model prediction and functional annotation

RNA-seq based gene model prediction was performed using 8 RNA-seq samples (accession number PRJEB40906) including samples from young leaves, old leaves, and adventitious roots exposed to high light and low light conditions (Stander et al. 2020). The RNA-seq reads were individually aligned to the genome using HiSAT2 (v.2.2.1, Kim et al. 2015) and the resulting RNA-seq alignments were assembled into transcripts using StringTie (v.2.1.7, Pertea et al. 2015). Next, the 8 individual transcriptomes were merged using stringtie -merge into a non-redundant set of representative transcripts (v2.1.7, Pertea et al. 2015). The consensus transcriptome was then functionally annotated with the Trinotate pipeline (v3.0.1, Marciais and Kingsford 2017) that combines BlastX (v.2.6.0-1, Camacho et al. 2009) and BlastP (v.2.6.0-1, Camacho et al. 2009) results from TransDecoder (v.5.5.0, Haas et al. 2013) predicted ORFs against the Uniprot database, and hmmscan (v.3.1b2, Finn et al. 2011) against the PFAM database (https://pfam.xfam.org/).

Assembly statistics were obtained using the stat bash program of the BBMap tool (v.38.94, Bushnell 2014). Genome assembly quality was also evaluate using the k-mer based method meryl (v.1.3, Koren et al. 2018). Base level consensus quality score (QV) was estimated from k-mer survival rate. Genome coverage by illumina short-reads was estimated to gain insight into the level of continuity of the final assembly. Reads were firstly aligned using bwa-mem2 (v.2.2.1, Vasimuddin et al. 2019). Resulting SAM file was sorted using samtools sort (v.1.16.1, Li et al. 2009) and the contig median coverage was calculated from samtools coverage. To assess the genome and gene annotation completeness, Benchmarking Universal Single-Copy Orthologs (BUSCO v.5.2.2, Simão et al. 2015) was applied with default setting using a plant-specific database of 2,326 single-copy orthologs. Number of annotated genes and associated statistics were obtained using agat_sp_statistics perl script from AGAT package (v.0.8.0, Dainat et al. 2022). Illumina RNA-seq reads used to perform gene annotation were mapped against the annotated genome using the splice-aware aligner STAR (v.2.7.10a, Dobin et al. 2012) and evaluating the classification consistency (evaluate). Long-terminal repeat (LTR) are annotated using LTR-finder (v.1.07, Xu and Wang 2007) and LTRHarvest included in GenomeTools (v.1.5.10, Ellinghaus et al. 2018). Terminal inverted repeat (TIR) are annotated using Generic repeat finder (v.1.0, Shi and Liang 2019) and TIR-learner (v.2.5, Su et al. 2019). Helitrons are annotated using HelitronScanner (v.1.1, Xiong et al. 2014).TIR candidates of less than 80 bp as well as LTR and Helitrons candidates of less than 100 bp are considered as tandem repeats and short sequences. False LTR discoveries are further filtered using LTR_retriever (v.2.9.0, Ou and Jiang 2018). Reported TIR candidates shorter than 600 bp are classified as MITEs. To remove false positives TIR and Helitrons candidates, several advanced filters are included in the EDTA pipeline (full-length copy numbers, simple sequence repeats). The obtained TE library is used to mask the genome and the remaining unmasked portion is scanned by the RepeatModeler (v.2.0.1, Flynn et al. 2020) to identify non-LTR retrotransposons and unclassified TE missed by structure-based TE identification tools. Finally, gene-related sequences have been removed using provided CDS. Transposable enrichment in biosynthetic gene clusters (BGCs) was performed by comparing TEs proportion in BGCs to their proportion in the corresponding contig by an exact Poisson test with the poisson.test function implemented in stat package (v. 4.1.1) in R (v. 4.1.1, R Core Team 2020).

Table 1. Main features of nuclear genomes from MIA producing plant species.

<table>
<thead>
<tr>
<th>Plant name</th>
<th>Family</th>
<th>Assembly size</th>
<th>No. of scaffolds/pseudo-chromosomes</th>
<th>N50 scaffold (Mb)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitrayna speciosa</td>
<td>Rubiaceae</td>
<td>1.1 Gb</td>
<td>17,031</td>
<td>1</td>
<td>Brose et al. (2021)</td>
</tr>
<tr>
<td>Gelsemium sempervirens</td>
<td>Gelsemiaceae</td>
<td>244 Mb</td>
<td>3,352</td>
<td>0.41</td>
<td>Franke et al. (2019)</td>
</tr>
<tr>
<td>Catharanthus roseus V2</td>
<td>Apocynaceae</td>
<td>541 Mb</td>
<td>2,090</td>
<td>2.57</td>
<td>Franke et al. (2019)</td>
</tr>
<tr>
<td>Campotheca acuminate V3</td>
<td>Cornales</td>
<td>414.95 Mb</td>
<td>21</td>
<td>18.28</td>
<td>Kang et al. (2021)</td>
</tr>
<tr>
<td>Ophiophriza pumila</td>
<td>Rubiaceae</td>
<td>439.90 Mb</td>
<td>11</td>
<td>42.83</td>
<td>Rai et al. (2021)</td>
</tr>
<tr>
<td>Rhazya stricta</td>
<td>Apocynaceae</td>
<td>274.0 Mb</td>
<td>980</td>
<td>5.5 Mb</td>
<td>Sabir et al. (2016)</td>
</tr>
<tr>
<td>Neolamarckia cadamba</td>
<td>Rubiaceae</td>
<td>744.5 Mb</td>
<td>22</td>
<td>29.20</td>
<td>Zhao et al. (2022)</td>
</tr>
</tbody>
</table>

GO enrichment

GO term enrichment analysis based on the GO classification was performed by comparing the relative occurrence of a GO term in increased orthogroups gene list to its relative occurrence in the genome by a Fisher’s Exact test (2-sided) with the R function topGO (v.2.44.0, Alexia and Rahnenfuehrer 2022). A Benjamini-Hochberg adjusted P-value of 0.05 was used as the cut-off for significance.
criteria. Enriched GO terms were grouped based on semantic similarities using rvgo package (Rel method, cut-off 0.7, v.1.4.0, Sayols 2020). Enriched terms graphs were performed using ggplot2 (v.3.3.5, Wickham 2011).

**Whole-genome duplication analysis**
The DupPipe pipeline (Barker et al. 2010) was used to infer WGD events in V. minor, Arabidopsis thaliana (Lamesch et al. 2012), C. roseus (Franke et al. 2019), Mitrayoga speciosa (Brose et al. 2021), Solanum lycopersicum (Hosmani et al. 2019), Campylotheca acuminata (Kang et al. 2021), Calotropis gigantea (Hoopes et al. 2018), Gelsemium sempervirens (Franke et al. 2019), Ophiopilissa pumila (Rai et al. 2021), and Papaver somniferum (Guo et al. 2018). For each dataset, duplicated gene pairs were identified using a discontinuous MegaBLAST (Ma et al. 2002, Zhang et al. 2004) which selected sequences that shared at least 40% sequence similarity over 300 bp. The open reading frame for each gene pair was established by comparison to the NCBI’s plant RefSeqs protein database (2021). The maximum-likelihood phylogenetic tree was obtained from Genome-wide synteny analysis

**Orthology analysis and phylogenetic tree reconstruction**
Gene families were constructed by comparing the protein sequences of V. minor with 9 other plant species: M. speciosa (Brose et al. 2021), G. sempervirens (Franke et al. 2019), C. roseus (Franke et al. 2019), C. gigantea (Hoopes et al. 2018), C. acuminata (Kang et al. 2021), O. pumila (Rai et al. 2021), S. lycopersicum (Hosmani et al. 2019), A. thaliana (Lamesch et al. 2012), and P. somniferum (Guo et al. 2018). For each species, protein sequences with a length of < 30 amino acids were filtered out and protein sequences were clustered using CD-HIT (v.4.7, Fu et al. 2012) to select the longest representative sequence for each cluster. These representative sequences were then used as input for OrthoFinder (v.2.5.4, Emms and Kelly 2019) using the following parameters: -S diamond -t 100 -M msa -A muscle -T taxml-ng. A maximum-likelihood phylogenetic tree was obtained from OrthoFinder using the 799 single-copy orthologs. Orthogroup gain and expansion were determined across the phylogenetic tree using Cafe5 (v.4.2.1, Mendes et al. 2020).

**Genome-wide synteny analysis**
A genome-wide synteny analysis between V. minor genome and the genome of its closely relative species C. roseus (Franke et al. 2019) has been performed. Both genomes were aligned using minimap2 (v.2.24, Li 2018) using the following options: -cx asm20 -cs. The resulting paf file was visualized using D-Genies (https://dgenies.toulouse.inra.fr/), Cabanettes and Klopp 2018 selecting hits with at least 80% identity and sorting contigs by size.

**Identification of co-localized MIA gene regions on the Vinca minor genome and synteny analysis**
Regions of co-localized biosynthetic genes were annotated on the V. minor genome using a personalized script that is available from https://doi.org/10.6084/m9.figshare.20749096.v1. Briefly, the script uses the genome .gff file, MIA blastn results in output format 6, and uniprot search results to anchor the putative MIA orthologs from the blastn analysis onto the genome. The script then searches 100,000 bp regions around each side of the annotation for additional genes annotated by pfam accessions: PF03171 (2OG-Fe(II) oxygenase superfamily), PF14226 (non-haem dioxygenase in morphine synthesis N-terminal), PF00891 (O-methyltransferase domain), PF08240 (Alcohol dehydrogenase GroES-like domain), PF00067 (Cytochrome P450), FF08031 (Berberine and berberine-like), and FF00201 (UDP-glucosyl transferase). Regions with more than one gene of interest are recorded as a cluster of interest.

**Syntenic regions between the P450/OMT gene cluster for V. minor (scaffold 2420), G. sempervirens (scaffold 505, Franke et al. 2019), and C. roseus (scaffold 16, Franke et al. 2019) were compared using BLASTN (v.2.6.0-1, Camacho et al. 2009) with the following parameters: blastn-outfmt 6 -task blastn -perc_identity 70 -evalue 1e-10. The resulting hits between the clusters were filtered to only include alignments with an E-value of 1e-6 and alignment length of 700 bp and alignments were visualized using the R genePlotR library (v0.8.11, Guy et al. 2010).

**Relative expression analysis**
RNA-seq reads from young leaves (YL), old leaves (OL), and adventitious roots (AR) (Stander et al. 2020, RJE040906) were pseudo-aligned onto the annotated transcripts from the V. minor genome using Salmon (v.0.14.1, Patro et al. 2017) with bias correction (-biasCorrect). Abundance estimates were established as transcripts per million (TPM) using the variational Bayesian optimized (-vbo) mode of Salmon.

**Template DNA construction for functional validation of candidate genes**
Vinca minor MSTRG.7518 candidate was selected for functional validation based on its high nucleotide sequence identity with C. roseus T16Hs, co-expression with the validated Vm16OMT, and the synteny of this gene cluster with C. roseus and G. sempervirens methoxylation clusters. Given nucleotide sequence identity with T16H2, UTR-discriminating forward primer specific to V. minor MSTRG.7518 was designed to perform a 2-step PCR. In this respect, V. minor MSTRG.7518 was first amplified from V. minor cDNA using UTR forward primer and a reverse cloning primer (2 rounds), followed by the second PCR amplification of the obtained column-purified PCR product using cloning primers to introduce SpeI restriction sites (Supplementary Table 1). Vinca minor MSTRG.7518 was further ligated into donor DNA vector pPET104 under pTEF1 promoter (Kulagina et al. 2021).

**Yeast strains**
Saccharomyces cerevisiae CEN.PK113-7D (MATa MAL2-8C, SUC2) strain was transformed with Cas9-expressing pCB2312 plasmid (Jessop-Fabre et al. 2016) for downstream CRISPR/Cas9-assisted gene integration via linearized PCR product using cloning primers to introduce SpeI restriction sites (Supplementary Table 1). Vinca minor MSTRG.7518 was further ligated into donor DNA vector pPET104 under pTEF1 promoter (Kulagina et al. 2021).
Culture conditions
Strain small-scale feedings were performed in 200 μl of YPD liquid media (20 g 1−1 peptone, 10 g 1−1 yeast extract, 20 g 1−1 glucose) from the overnight 5 ml YPD cultures diluted 20 times (Kulagina et al. 2021), supplemented with 125 μM vincadifformine at 0h, and glucose to 20 g 1−1 at 24h, during 48h at 28°C and 200 rpm. The feedings were set up individually for each replicate and time point.

Sample analysis
The supernatants were collected by centrifugation (10 min, 7,000 g) and diluted 20 times in MeOH 100%, followed by vortexing and 15 min centrifugation at 20,000 g prior to injection. Compound identification was performed by ultra-performance liquid chromatography-mass spectrometry (UPLC-MS) as described previously (Parage et al. 2016) with an 8-min linear gradient from 15 to 50% acetonitrile (containing 0.1% formic acid). The selected mode of ion monitoring was employed for the following compounds: vincadifformine (m/z 339), RT = 5.68; 16-hydroxyvin- cadifformine (m/z 355), RT = 4.22.

Metabolomic analysis
Leaves, stems, flowers, and roots of V. minor were frozen in liquid nitrogen, then freeze-dried, ground into powder, and 2 replicates for each organ were sonicated with methanol and 0.1% of formic acid to extract metabolites. After centrifugations, extracts have been injected at 0.01 mg/ml on a UPLC system coupled to a quadrupole time-of-flight mass spectrometer, and the acquisition was made using full MS survey scans to acquire high-resolution MS1 spectra. LC-MS/MS scans to help identify compounds: vincadifformine (m/z 339), RT = 5.68; 16-hydroxyvin-cadifformine (m/z 355), RT = 4.22.

Results and discussion
Vinca minor genome sequencing, assembly, and analysis
The V. minor genome was sequenced usingONT long-read sequencing, resulting in 6,006,337 reads representing 45.47 GB of data, as well as with illumina short-read sequencing resulting in 177,102,894 reads representing 26.7 Gb of data (Supplementary Table 2). Illumina short-read k-mer counting estimated the V. minor genome length to be around 504,085 Mb, with the binomial distribution of the resulting k-mer graph typical of a diploid heterozygous genome with a heterozygosity rate of 1.54% (Table 2 and Supplementary Fig. 1). The genome was assembled from ONT long-reads using Flye (Kolmogorov et al. 2019). This Flye assembly consisted in 1,557.17 Mb distributed across 7,019 contigs with an N50 length of 0.5 Mb while its collapsing reduced length to 685.0 Mb, contig number to 296 and increased N50 length to 6.0 Mb. Finally, the collapsed assembly was polished twice using Illumina short-reads (Supplementary Table 2) with pilon (Walker et al. 2014), resulting in a final assembly length of 679 Mb (1.3 x the estimated assembly length), consisting of 296 scaffolds with an N50 scaffold length of 6.0 Mb (Table 2). 95.4% of this final assembly could be covered by illumina short reads, indicating a good level of continuity. The base level QV of 28.6878 (Table 2), corresponding to more than 99.8% accuracy, highlights the good quality of this final assembly. Genome completeness was determined by identifying conserved orthologs from the eudicot lineage dataset in the genome using BUSCO (Benchmarking Universal Single-Copy Orthologs). This analysis revealed that 96% of 2,326 eudicot lineage genes could be identified in the genome, indicating a very high level of completion (Fig. 2a).

RNA-seq-based gene model prediction using RNA-seq reads from the young leaves, old leaves, and adventitious roots of V. minor (Stander et al. 2020), were used to annotate a total of 29,624 protein-coding genes and 38,299 transcripts in the V. minor genome (Table 1), in comparison to the 34,363 and 22,617 transcripts reported in C. roseus and G. sempervirens (Franke et al. 2019). Depending on the RNAseq sample, 91.95–94.8% of reads uniquely mapped on this final assembly; 63.5–5.54% highlighted multiple matches and 1.19–2.51% did not match (Supplementary Table 3), indicating a good annotation quality. Among these 38,299 transcripts, 87% (33,357) presented a complete CDS based on TransDecoder ORF prediction. BUSCO analysis of the predicted transcripts revealed a complete BUSCO score of 88.6% and a low fragmentation score of 1.6%, indicating that a large number of genes with high-quality sequences have been annotated (Fig. 2a). These sequences were functionally annotated against Pfam, SwissProt, GO, and KEGG databases resulting in the annotation of 81.9% genes (24,272 predicted genes, Fig. 2b), and 82.4% transcripts (31,572 predicted transcripts, Supplementary Table 4, Fig. 2c). A quarter of the total genes were annotated by all 4 databases, whereas 53% of the genes were only annotated with GO and KEGG databases.

Next, we explored evolutionary WGD artifacts by identifying paralogous gene pairs across different plant species, and calculating the synonymous substitutions per synonymous site (Ks) for each gene pair (Fig. 2d). Sorting the paralogs in order of age (Ks) resulted in an initial high-density peak at low Ks representing the genes that are being duplicated at present. Over time, duplicated genes are lost which results in a L-shaped pattern or exponential decrease. Large-scale duplications, including WGD events, lead to substantial increases in the number of paralogs at a specific time point (Ks) which can be visualized as secondary peaks in the plot. The well-described and conserved whole-genome triplication event that is shared among eudicots (Jiao et al. 2012) was detected as such a secondary peak at around Ks = 2 in V. minor, C. roseus, S. lycopersicum, O. pumila, C. gigantea,
M. speciosa, and G. sempervirens (Fig. 2d and Supplementary Fig. 2). Vinca minor did not show any additional secondary peaks in the Ks plot, suggesting a lack of any additional recent WGDs. However, the previously reported post-y WGD (Kang et al. 2021; Rai et al. 2021) was evident in the paralog Ks plot of C. acuminata (Supplementary Fig. 2).

Lastly, we analyzed the composition in TEs of the V. minor genome. Repeat analysis showed that 48.03% of this genome consists of TEs, most of which being LTRs (34.74%) (Fig. 2e and Supplementary Table 5). Interestingly, the V. minor genome has the highest percentage of repetitive elements, compared to fellow Gentianales members, C. roseus (38.78%), C. gigantea (24.31%), G. sempervirens (19.38%), O. pumila (47.68%), and M. speciosa (46.76%, Fig. 2e). Vinca genome also present a higher TE proportion than C. acuminata (29.47%) and A. thaliana (10.26%) while S. lycopersicum (58.54%) and P. somniferum (72.58%) present a higher TE proportion than V. minor. Most LTRs were Copia elements in V. minor (18.5%), G. sempervirens (5.47%), C. roseus (11%), and C. acuminata (8.67%), whereas the 2 Rubiaceae species, O. pumila and M. speciosa, had a majority of Gypsy elements accounting for 21.09% and 19.05% of TEs in each genome, respectively (Fig. 2e and Supplementary Table 5). Such a distribution thus confirms that repeat elements are constantly expanding and contracting across different plant genomes, resulting in the disparity in genomic repeat profiles, even between closely related species (Lisch 2013).
Comparative genomic analysis

To attain additional insights into V. minor genome evolution, we compared the V. minor protein coding genes to the protein coding genes from 9 plant genomes, including 5 MIA-producing genomes (C. acuminate, C. roseus, G. sempervirens, M. speciosa, and O. pumila). OrthoFinder was used to generate orthologous (OG, gene families) across these broader plant lineages. In total, 92.2% of genes were assigned to 23,486 OGs across all species (Supplementary Table 6), with a mean OG size of 11.8 proteins. Comparing orthologs between MIA-producing and non-producing plants may provide interesting insights into which gene families are specifically conserved within the MIA-producing species. A total of 19,764 OGs were shared among the 6 MIA-producing plant proteomes (Supplementary Table 7). Of these, 4,592 OGs were specific to MIA-producers, and 11 OGs were common to all 6 MIA-producing species (Fig. 3a). Such conserved gene sets that are associated with MIA-production can shed light on the evolution and diversification of MIA-metabolism. For example, Gene ontology (GO) enrichment analysis revealed 387 biological processes, 84 cellular components and 154 molecular function GO terms enriched in V. minor genes present in the 19,764 OGs shared among the 6 MIA-producing plants. Among them, 20 are associated with transcription regulation (Supplementary Fig. 3), 75 to stimuli response including ABA, oomycetes, iron starvation, water deprivation, light, and 80 to plant development (including leaf development, vegetative-reproductive transition, cell growth, cell and nuclei division, PCD, secondary shoot development, and secondary cell wall development, autophagy). The V. minor genes present in the increased 4,592 OGs MIA-specific OGs (Fig. 3a) were found to be mainly implicated in biological processes associated with plant development (including endosperm development, endocytosis, petal formation, and seed germination:

enriched in secondary cell wall development, autophagy). The

orthogroups between MIA-producing and non-producing plants may provide interesting insights into which gene families are specifically conserved within the MIA-producing species. A total of 19,764 OGs were shared among the 6 MIA-producing plant proteomes (Supplementary Table 7). Of these, 4,592 OGs were specific to MIA-producers, and 11 OGs were common to all 6 MIA-producing species (Fig. 3a). Such conserved gene sets that are associated with MIA-production can shed light on the evolution and diversification of MIA-metabolism. For example, Gene ontology (GO) enrichment analysis revealed 387 biological processes, 84 cellular components and 154 molecular function GO terms enriched in V. minor genes present in the 19,764 OGs shared among the 6 MIA-producing plants. Among them, 20 are associated with transcription regulation (Supplementary Fig. 3), 75 to stimuli response including ABA, oomycetes, iron starvation, water deprivation, light, and 80 to plant development (including leaf development, vegetative-reproductive transition, cell growth, cell and nuclei division, PCD, secondary shoot development, and secondary cell wall development, autophagy). The V. minor genes present in the increased 4,592 OGs MIA-specific OGs (Fig. 3a) were found to be mainly implicated in biological processes associated with plant development (including endosperm development, endocytosis, petal formation, and seed germination: (GO:0045943, GO:0040029, GO:0034243, GO:0001736, GO:0048451, GO:0090392, GO:2000014, GO:0090628, GO:0009554, GO:0048765, GO:0001736, GO:0000122, GO:0048765, GO:0090628, GO:0009554, GO:0035834, GO:0009709, Fig. 3e). Among the studied MIAs, 22 were identified at different confidence levels (Supplementary Table 10). As previously observed, vincamone was the most abundant MIA in the 4 organs with high amounts of its 9-methoxy derivative in leaves and stems (Supplementary Fig. 5) (Abouzeid et al. 2017). Interestingly, while vincadifformine was accumulated at low or trace levels in all organs, numerous derivatives were identified including minovincinine, 16-methoxyvincadifformine (ervinceine), 11-methoxyminovincinine (16-methoxyminovincinine), minovincine, and minovine for instance, especially in leaves and flowers. We also observed numerous 11/16-methoxy derivatives of vincadifformine including 16-methoxyvincadifformine (ervinceine), 11-methoxyminovincinine (16-methoxyminovincinine), and 11 methoxymnivin (16-methoxymnivin). By contrast, no acylated derivatives of vincadifformine such as echitovenine were detected and neither tetrahydroxalstonine, ajmalicine, tabersonine and derivatives (lochnerine, horhammerine) or catharanthine as expected. At the gene level, since C. roseus and Rauwolfia species have been widely used to elucidate MIA pathways, we used functionally validated MIA pathway genes of V. minor orthologs (Supplementary Table 9). Orthologs of all genes from the methylerythritol and monoterpene secoiridoid pathways, which provide secologanin, were identified. We thus observed that V. minor displays 4 potential copies of secologanin synthase (MSTRG.3111, MSTRG.22185, MSTRG.1442, and MSTRG.1443) as observed previously in C. roseus (Kellner et al. 2015b). In addition, orthologs of strictosidine synthase (STR) and strictosidine b-glucosidase (SGD) that ensure strictosidine formation and deglycosylation for downstream MIA synthesis were predicted (McKnight et al. 1990; Gerasimenko et al. 2002). Potential orthologs of all genes ensuring vincadifformine synthesis from geissoschizine synthase (GS) up to VS were also retrieved from the BLAST analysis (Tatsis et al. 2017; Caputi et al. 2018; Qu et al. 2019). Besides the 2 potential copies of VS as described in C. roseus, we also predicted 4 orthologs of tabersonine synthase (TS) (Caputi et al. 2018; Qu et al. 2019). Since tabersonine is not accumulated in V. minor, it is highly possible that these genes catalyze the synthesis of the other MIAs of interest. By contrast, we did not

Vinca minor metabolomics analysis and homology-based prediction of putative MIA gene orthologs

To guide future elucidations of MIA biogenesis in V. minor, we investigated the specific content of MIA biosynthetic genes of this plant in comparison to MIA content. A UPLC/HRMS analysis was thus performed on methanolic extracts of V. minor roots, stems, flowers, and leaves. Principal component analysis showed sample clustering per organ type thus highlighting the specificity of MIA accumulation in each organ with roots being the most different (Supplementary Fig. 4). Among the studied MIAs, 22 were identified at different confidence levels (Supplementary Table 10). As previously observed, vincamone was the most abundant MIA in the 4 organs with high amounts of its 9-methoxy derivative in leaves and stems (Supplementary Fig. 5) (Abouzeid et al. 2017). Interestingly, while vincadifformine was accumulated at low or trace levels in all organs, numerous derivatives were identified including minovincinine, 16-methoxyvincadifformine (ervinceine), 11-methoxyminovincinine (16-methoxyminovincinine), minovincine, and minovine for instance, especially in leaves and flowers. We also observed numerous 11/16-methoxy derivatives of vincadifformine including 16-methoxyvincadifformine (ervinceine), 11-methoxyminovincinine (16-methoxyminovincinine), and 11 methoxymnivin (16-methoxymnivin). By contrast, no acylated derivatives of vincadifformine such as echitovenine were detected and neither tetrahydroxalstonine, ajmalicine, tabersonine and derivatives (lochnerine, horhammerine) or catharanthine as expected. At the gene level, since C. roseus and Rauwolfia species have been widely used to elucidate MIA pathways, we used functionally validated MIA pathway genes of V. minor orthologs (Supplementary Table 9). Orthologs of all genes from the methylerythritol and monoterpene secoiridoid pathways, which provide secologanin, were identified. We thus observed that V. minor displays 4 potential copies of secologanin synthase (MSTRG.3111, MSTRG.22185, MSTRG.1442, and MSTRG.1443) as observed previously in C. roseus (Kellner et al. 2015b). In addition, orthologs of strictosidine synthase (STR) and strictosidine b-glucosidase (SGD) that ensure strictosidine formation and deglycosylation for downstream MIA synthesis were predicted (McKnight et al. 1990; Gerasimenko et al. 2002). Potential orthologs of all genes ensuring vincadifformine synthesis from geissoschizine synthase (GS) up to VS were also retrieved from the BLAST analysis (Tatsis et al. 2017; Caputi et al. 2018; Qu et al. 2019). Besides the 2 potential copies of VS as described in C. roseus, we also predicted 4 orthologs of tabersonine synthase (TS) (Caputi et al. 2018; Qu et al. 2019). Since tabersonine is not accumulated in V. minor, it is highly possible that these genes catalyze the synthesis of the other MIAs of interest. By contrast, we did not
find any potential true orthologues of catharanthine synthase (CS) and minovincinine 19-hydroxy-O-acetyltransferase (MAT) in agreement with the non-detection of catharanthine and echitovenine in V. minor. Finally, already known genes from V. minor including VmPiNMT, Vm16OMT, and VmTPT2 were also identified. In conclusion, all the results illustrated how a genome-based prediction of MIA biosynthetic genes can directly provide potential new gene candidates as exemplified by the TS orthologues. Furthermore, such a prediction could also be of high interest to guide the discovery of yet unknown MIA biosynthetic steps by mining the genomic environment of the predicted orthologues.

Identification of MIA gene clusters in Vinca minor led to the identification of a functional vincadifformine 16-hydroxylase

It has been well described that genes involved in similar specialized metabolisms can cluster on genomic regions in fungi and bacteria as well as in plant genomes (Nützmann et al. 2018). The current definition for plant BGCs states that BGCs contain at least 3 different types of nonhomologous enzymes and are typically co-expressed (Nützmann et al. 2016). However, additional genes sharing high-homology to cluster genes may exist within these regions, having arisen from tandem duplications. We, therefore, extended our V. minor genetic investigation by searching for co-
localization of metabolic genes of interest in the \textit{V. minor} genome. The predicted MIA orthologs (Supplementary Table 9) were thus anchored onto the \textit{V. minor} genome and used to search the surrounding genomic landscapes in order to identify associated biosynthetic genes or gene duplicates. On this basis, by using a custom built script, we captured the genomic regions that contained at least 2 genes of interest or gene duplicates, including alcohol dehydrogenases, dioxygenases, hydrolyases, methyltransferases, cytochrome P450s, transcription factors, acetyltransferases, and UDP-glycosyltransferases, which constitute the enzyme arsenal mostly involved in specialized metabolisms. A total of 23 putative co-localized gene regions of interest were identified comprising 387 genes and 16 putative MIA biosynthetic genes of interest (Supplementary Table 11). For instance, in gene cluster 1, a total of 2 cytochrome P450s, 3 acetyltransferases, 1 hydrolyase, 1 alcohol dehydrogenase, and 1 transcription factors clustered together in a region of 482 kb, some of them corresponding to putative orthologues of MIA biosynthetic genes. It is thus highly possible that this region corresponds to a potential reservoir of genes encoding yet unknown enzymes of the MIA pathway.

In addition, physically co-localized genes may also be conserved between chromosomes within or between species which is known as synteny. A synteny analysis was thus performed between \textit{V. minor} and the well-characterized \textit{C. roseus}. The currently available \textit{C. roseus} assembly (Franke et al. 2019) is still highly fragmented and contigs can be misoriented. The dot plot comparing \textit{V. minor} and \textit{C. roseus} clearly indicated an apparent collinearity between the 2 genomes albeit it can be partially masked due to improper contig orientation and fragmentation in \textit{C. roseus} (Supplementary Fig. 6a). However, we were able to find a conserved organization with 237 of the 296 \textit{V. minor} contigs having at least one hit with a \textit{C. roseus} contig and 438 of the 2,090 \textit{C. roseus} contigs having at least one hit with a \textit{V. minor} contig (Supplementary Fig. 6, a and b and Table 12). Among the other contigs, 197 \textit{V. minor} contigs and 394 \textit{C. roseus} contigs display several matches (Supplementary Fig. 6, a and b) as exemplified with \textit{C. roseus} contig 60 and \textit{V. minor} contig 4699 (Supplementary Fig. 6c). Indeed, 1 sterol transport-associated gene (MSTRG.18143), 2 SCARECROW-like transcription factor (MSTRG.18145, MSTRG18147), 1 LACCASE (MSTRG.18149), and 1 metalloproteinase (MSTRG.18150) can be found in the 70 kb-long fragment of \textit{V. minor} contig 4699. A similar gene content and organization can be observed in the 110 kb-long syntenic fragment of \textit{C. roseus} contig 60, composed of 1 metalloproteinase (CRO_133159), 1 LACCASE (CRO_133160), 4 SCARECROW-like (CRO_133163, CRO_133164, CRO_133165, CRO_133166), and 1 sterol transport-associated gene (CRO_133167). On the other hand, it is well known that synteny between different species may unveil related genes that are involved in specific metabolite biosynthesis and allow the transfer of gene discoveries made in one species to phylogenetically related species. For example, in \textit{C. roseus}, 2 genes coding T16H isoforms (T16H1 and T16H2) with the associated 16OMT were found to be physically co-located on the \textit{C. roseus} genome (Franke et al. 2019; Kellner et al. 2015b). By microsynteny analysis, we identified a similar methoxylation gene cluster (cluster 22) in \textit{V. minor} on contig_2420 (Fig. 5a and Supplementary Table 11). This cluster contained 2 putative cytochrome P450s (MSTRG.7518 and MSTRG.7523) with high homology to \textit{C. roseus} T16Hs (78.4% and 77.4 nucleotide sequence identity, Supplementary Tables 9 and 11), as well as 5 methyltransferases (MSTRG.7519, MSTRG.7522, MSTRG.7526, MSTRG.7528, and MSTRG.7529; 63.5–75.6% nucleotide sequence identity to \textit{C. roseus} 16OMT, Supplementary Tables 9 and 11), including the functionally validated Vm16OMT (MSTRG.7522) that was found to methylate 16-hydroxyvincadifformine (Stander et al. 2020). Interestingly, this gene organization also corresponds to the methoxylation cluster identified in \textit{C. roseus} composed of rankinidine and humantenine-11-hydroxylation (RH11) and rankinidine and humantine-11-O-methyltransferase (RH11OMT) which catalyze the hydroxylation and methylation of the MIA oxindole scaffold in \textit{C. roseus} (Franke et al. 2019). As expected, the synteny comparison between \textit{V. minor} scaffold contig_2420, \textit{C. roseus} scaffold cro_v2_scaffold_16 and \textit{C. roseus} scaffold_505 thus revealed considerable similarities between the 3 methoxylation clusters. Furthermore, we compared the relative proportion of genes and TEs in cluster 22 to their proportion on scaffold 2420 (Fig. 5b).

Eighteen genes and 176 TE have been annotated on cluster 2 whilst 510 genes and 6,780 TE have been annotated on scaffold 2420. An exact Poisson test revealed a significant gene enrichment (F-value = 0.004751285) in cluster 2 as compared to scaffold 2420. This result is in agreement with the already described TEs enrichment in numerous BGCs (Field et al. 2011; Field and Osbourn 2012; Winzer et al. 2012; Krokida et al. 2013; Li et al. 2021; Shen et al. 2021), supporting TEs involvement in the formation of metabolic gene clusters by providing homologous sequences for genomic recombinations and rearrangements (Huang et al. 2012; Bennetzen and Wang 2014).

To further describe the putative \textit{V. minor} methoxylation cluster, we next calculated the relative expression level of each gene as TPM using data from the \textit{V. minor} transcriptome (Stander et al. 2020; Fig. 5c). The highest expression levels were obtained for the O-methyltransferases Vm16OMT, OMT3, and OMT2 in young leaves where MIA metabolism is high. While P450_2 was only expressed at trace levels in the studied organs, P450_1 was substantially expressed in young leaves (>100 TPM) which prompted us to perform its functional validation. To conduct this assay, P450_1 (VmMSTRG.7518) was expressed in yeast following a CRISPR/Cas9-mediated integration. The resulting yeast strain, as well as a strain transformed by an empty vector, were then cultured and fed with vincadifformine before analysis of the resulting products with ultra-performance liquid chromatography-mass spectrometry (UPLC–MS; Fig. 5d).
comparison of selected ions (m/z 339 for vincadifformine, m/z 355 for 16-hydroxyvincadifformine) to retention times of standards allowed to establish that P450_1 (VmMSTRG.7518) hydroxylated vincadifformine into 16-hydroxyvincadifformine as compared to yeast transformed with empty vector. This result thus confirms that VmMSTRG.7518 and Vm16OMT form a vincadifformine methoxylation cluster responsible for the formation of 16-methoxyvincadifformine (ervinceine) in leaves of V. minor. This throws light on the conservation of a methoxylation gene cluster organization in V. minor, C. roseus, and G. sempervirens.

**Conclusion**

Here, we describe the genome of the lesser periwinkle, V. minor, which is the first Apocynaceae native to central and southern Europe sequenced to date (Fig. 1c). While V. minor shares numerous genomic and evolutionary traits with the other sequenced MIA producing plants, the careful examination of gene organization also sheds light on the existence of several putative gene clusters whose future functional characterization will probably lead to the discovery of new MIA biosynthetic enzymes. Similarly, gene synteny analysis combined with gene functional validation allowed the identification of V. minor vincadifformine...
methoxylation cluster as already described in C. roseus and G. semperivirens for tabersonine and rankinidine, respectively. This reinforces the potential interest of this genomic data set for gene discovery and contributes to the construction of an MIA-oriented pan-genome. Therefore, beyond widening our knowledge on MIA biosynthesis, the future elucidations of MIA biosynthetic routes resulting from this study will also allow the development of new MIA supply approaches based on the gene transfer in heterologous hosts such as yeast or bacterium (Courdavault et al. 2021). These metabolic engineering strategies allowed creating efficient cell factories producing natural products on demand to face unstable and limited production in plants, as reported for strictosidine, vindoline, and vinblastine (Brown et al. 2015; Kulagina et al. 2021; Liu et al. 2021; Zhang et al. 2022).

**Data availability**

Raw DNA-seq and the genome assembly have been deposited in the NCBI database under the BioProject accession number: PRJNA873287 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA873287). The genome annotation, transcript expression abundances, and script used for searching for MIA gene clusters are available on the figshare: https://doi.org/10.6084/m9.figshare.20749096.v1.

Supplemental material available at G3 online.

**Acknowledgments**

We thank access and support to the CCSC computing resources (Cascimodot Federation, CNRS, Orléans). We are thankful to the chemical ecology platform of the Institut de Recherche sur la Biologie de l’Insecte (IRBI; Tours, France) for the access to the UPLC/quadrupole time of flight mass spectrometer.

**Funding**

We acknowledge funding from the EU Horizon 2020 research and innovation program (MIAMI project-Grant agreement N°814645), ARD-CVL Biopharmaceutical program of the Région Centre Val de Loire (ETOPOCentre project), and ANR (project MIACYC—ANR-20-CE43-0010).

**Conflict of interest**

RFD and HJ are CEO and CTO of Future Genomics Technologies, respectively.

**Literature cited**


Communicating editor: J. Ma