the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) doi: 10.1111/tpj.12827

RESOURCE

Genome-guided investigation of plant natural product biosynthesis

Franziska Kellner[1,†], Jeongwoon Kim[2,†], Bernardo J. Clavijo[3], John P. Hamilton[2], Kevin L. Childs[2], Brieanne Vaillancourt[2], Jason Cepela[2], Marc Habermann[2], Burkhard Steuernagel[4], Leah Clissold[3], Kirsten McLay[3], Carol Robin Buell[2,*] and Sarah E. O'Connor[1,*]

[1]*Department of Biological Chemistry, The John Innes Centre, Norwich NR4 7UH, UK,*
[2]*Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA,*
[3]*The Genome Analysis Centre, Norwich, NR4 7UH, UK, and*
[4]*The Sainsbury Laboratory, Norwich, NR4 7UH, UK*

Received 16 December 2014; revised 27 February 2015; accepted 4 March 2015.
*For correspondence (e-mails sarah.o'connor@jic.ac.uk or buell@msu.edu).
†These authors contributed equally.

SUMMARY

The medicinal plant Madagascar periwinkle, *Catharanthus roseus* (L.) G. Don, produces hundreds of biologically active monoterpene-derived indole alkaloid (MIA) metabolites and is the sole source of the potent, expensive anti-cancer compounds vinblastine and vincristine. Access to a genome sequence would enable insights into the biochemistry, control, and evolution of genes responsible for MIA biosynthesis. However, generation of a near-complete, scaffolded genome is prohibitive to small research communities due to the expense, time, and expertise required. In this study, we generated a genome assembly for *C. roseus* that provides a near-comprehensive representation of the genic space that revealed the genomic context of key points within the MIA biosynthetic pathway including physically clustered genes, tandem gene duplication, expression sub-functionalization, and putative neo-functionalization. The genome sequence also facilitated high resolution co-expression analyses that revealed three distinct clusters of co-expression within the components of the MIA pathway. Coordinated biosynthesis of precursors and intermediates throughout the pathway appear to be a feature of vinblastine/vincristine biosynthesis. The *C. roseus* genome also revealed localization of enzyme-rich genic regions and transporters near known biosynthetic enzymes, highlighting how even a draft genome sequence can empower the study of high-value specialized metabolites.

Keywords: specialized metabolite, alkaloid, genome, *Catharanthus roseus*, vinblastine.

INTRODUCTION

Catharanthus roseus (L.) G. Don, a member of the euasterids I clade (Gentianales order, Apocynaceae family), produces monoterpene indole alkaloids (MIA), an exceptionally diverse class of specialized metabolites ('natural products') (O'Connor and Maresh, 2006). The monoterpene moiety of all MIAs is derived from an iridoid class of monoterpenes, secologanin, and the indole moiety is derived from the amino acid tryptophan. Thousands of MIAs derived from secologanin and tryptamine are produced in numerous plant families, with *C. roseus* producing a subset of approximately 100 MIAs. While *C. roseus* is best known for production of the bis-indole MIAs, vinblastine and vincristine, which are used in the clinic as anti-cancer agents, other bioactive MIAs such as raubasine, yohimbine, and alstonine are also produced by this plant (Figure 1) (Aslam *et al.*, 2010).

Ample transcriptomic and proteomic resources are now available for *C. roseus* (Murata *et al.*, 2008; Champagne *et al.*, 2012; Gongora-Castillo *et al.*, 2012; Van Moerkercke *et al.*, 2013; Verma *et al.*, 2014). While this information has dramatically accelerated the discovery of MIA biosynthetic genes, a whole-genome sequence will provide additional and important insights into the production, regulation, and evolution of these valuable metabolites. For example, numerous studies have shown that genes encoding

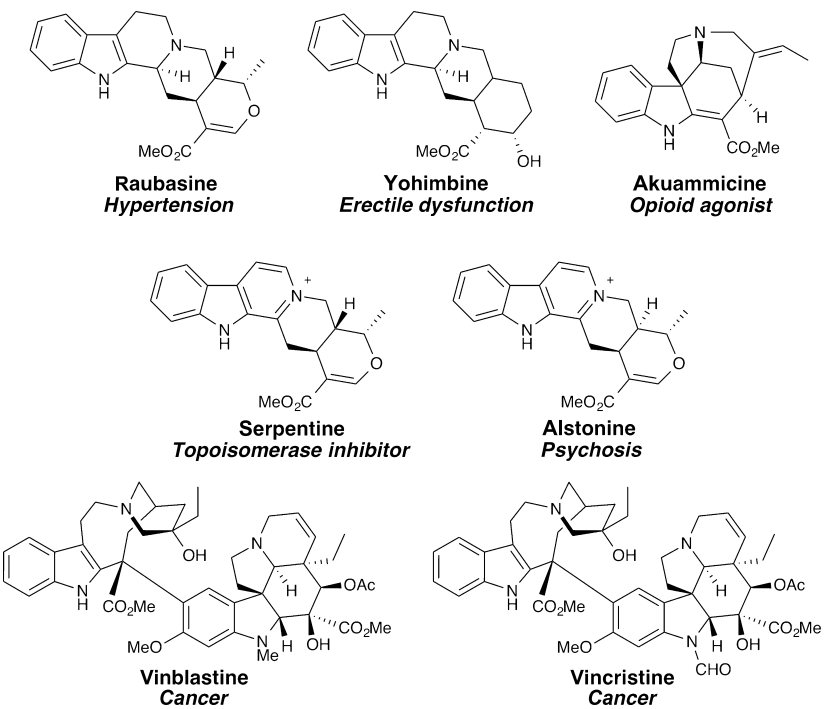

Figure 1. Representative monoterpene indole alkaloids (MIA) that are produced in *C. roseus.*

specialized metabolism in plants can be physically clustered in the genome (Frey *et al.*, 1997; Qi *et al.*, 2004; Amoutzias and Van de Peer, 2008; Field and Osbourn, 2008; Swaminathan *et al.*, 2009; Winzer *et al.*, 2012; Itkin *et al.*, 2013; Mugford *et al.*, 2013). While the reasons for clustering remain unresolved, one hypothesis is that evolutionary pressure for retention of a multigenic trait as a single locus facilitates the synthesis of the final product and prevents accumulation of toxic pathway intermediates (Takos and Rook, 2012; Nutzmann and Osbourn, 2014).

The extent to which gene clustering occurs in plant specialized metabolism is unclear, as the vast majority of plant species lacks the requisite genomic information required to explore this issue. Most gene clusters reported to date have been in crop species and *Arabidopsis thaliana*, as plant species that are valued for production of a single, specific, specialized metabolites have not been targets for genome sequencing due to limited fiscal and personnel resources. Despite the intense interest in *C. roseus* and the high economic value of its metabolites, genome sequence data for this plant was previously limited to the plastid (Ku *et al.*, 2013). *C. roseus*, a self-pollinating diploid ($2n$ = 2x = 16) with a moderate genome size (738 Mbp) (Guimaraes *et al.*, 2012), is an excellent candidate for genome sequencing with a next-generation sequencing approach.

Recent advances in genome sequencing technologies and assembly algorithms have resulted in generation of genome sequences for a wide range of plant species. While a near-complete genome sequence with scaffolds anchored into pseudomolecules to represent the chromosomes is limited to species with large research communities and/or those with major economic importance (Schnable *et al.*, 2009; The International Brachypodium Initiative, 2010; The Potato Genome Sequence Consortium, 2011; The Tomato Genome Consortium, 2012), high quality draft genome assemblies that represent genic regions of the genome can be generated by single investigators. These draft genomes provide not only insights into important biological processes, but also are a paradigm-changing resource for downstream analyses. Using sequencing-by-synthesis, we assembled a draft genome assembly for *C. roseus* that provided a near-comprehensive representation of the genic space. While this genome is not scaffolded into pseudomolecules representing the *C. roseus* chromosomes, it nevertheless provided substantial insights into specialized metabolism including physical clustering of secondary metabolism biosynthetic genes, evidence for sub- and neo-functionalizations following gene duplication, and transcriptional regulatory networks. Overall, this study highlights the importance that genomic data can play when investigating secondary metabolic pathways.

RESULTS

Genome sequence, assembly, and annotation

Using a whole-genome shotgun sequencing approach, we generated a draft genome sequence of the cultivar 'SunStorm™ Apricot'. Using 33 Gb of sequence from a single 400-bp fragment Illumina library, we generated an assem-

bly of 523 Mb with an N50 scaffold size of 26.2 kbp, only including scaffolds >200 bp (Table S1). If only scaffolds >1000 bp are included, the assembly represents 506 Mb with a larger N50 scaffold size of 27.3 kbp (Table S1). K-mer spectra representation analysis (http://www.tgac.ac.uk/tools-resoures/kat) shows a highly complete assembly with good k-mer copy-number distributions (Figure S1). The size cutoffs on scaffolds of 200 and 1000 bp were evaluated also in terms of their total and distinct k-mer counts, revealing that the smaller cutoff at 200 bp was the point at which most small repetitive sequences were discarded (Figures S1 and S2). Estimation of completeness of the assembly (scaffolds ≥1000 bp) with both Sanger-derived *C. roseus* Expressed Sequence Tags (ESTs) and the Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra *et al.*, 2007) pipeline revealed robust representation of genes in the assembly; 95.7% of the ESTs and 97.6% of the conserved CEGMA proteins were detected in the assembly (Table S1). 'SunStorm™ Apricot', like other *Catharanthus* cultivars, is self-pollinating, and to assess the degree of heterozygosity, sequencing errors, and/or potential misassembly, we aligned all of the genome reads to the assembly and assessed the rate of single nucleotide polymorphism (SNPs). Nearly 85% of the reads aligned to the assembly with a mapping quality score >20 or were multi-mapping suggesting a high degree of representation of the *C. roseus* genome in our assembly. Estimation of the sequencing and assembly error rate was <1 per 500 kb while the heterozygosity rate, as reflected by biallelic SNPs, was estimated as <1 per 1000 bp, consistent with an inbred cultivar and a high quality assembly.

Using the MAKER annotation package (Holt and Yandell, 2011; Campbell *et al.*, 2014), 33 829 genes were annotated (Table S2). To assess the quality of the assembly and annotation, we compared the predicted gene set with 32 known biosynthetic genes involved in MIA biosynthesis, including methylerythritol phosphate (MEP; upstream terpene biosynthesis) biosynthesis, iridoid biosynthesis, downstream alkaloid biosynthesis, and two transcription factors known to regulate MIA biosynthesis; all were present in the assembly further supporting that this assembly provides a robust representation of the genic regions of the *C. roseus* genome (Figure 2a and Table S3, orange entries). While several of the biosynthetic genes were partial due to their localization near the end of a scaffold, only one gene [SGD (alkaloid gene)], was substantially misassembled; manual examination revealed a collapsed genome assembly representing a minimum of two close SGD paralogs. SGD encodes 13 exons spanning more than 10 kb and was located on five separate scaffolds. Surprisingly, even with an N50 scaffold length of 27.3 kbp, SGD is the only examined biosynthetic genes for which we failed to readily identify the cognate gene sequence in our assembly. The vast majorities of the known MEP, iridoid, and alkaloid genes were located on scaffolds greater than the N50 size (Table S3), consistent with the high degree of representation of ESTs and CEGMA genes in the assembly and suggestive that the 523 Mb of assembly provides a near-comprehensive representation of the genic regions of the *C. roseus* genome. The remaining 215 Mb of the 738 Mb *C. roseus* genome that is not present in this genome assembly is likely primarily composed of repetitive sequences that are recalcitrant to the assembly process using short-read sequences and therefore are absent or substantially under-represented in all genome assemblies generated with short-read sequences (Hirsch and Buell, 2013).

Co-expression of monoterpene indole alkaloid pathway genes

A key feature of vinblastine/vincristine biosynthesis in *C. roseus* is induction of biosynthesis in seedlings following treatment with methyl jasmonate (MeJA). This increase in alkaloid content is correlated with increased expression of genes in the MIA biosynthetic pathway (Vazques-Flota and De Luca, 1998; Gongora-Castillo *et al.*, 2012). Analysis of gene expression profiles from available *C. roseus* transcriptomic data has already enabled successful identification of new biosynthetic genes (Geu-Flores *et al.*, 2012; Asada *et al.*, 2013; Besseau *et al.*, 2013). However, all previous analyses utilized *de novo* transcriptome assemblies, which do not provide full representation of the *C. roseus* transcriptome. Re-analysis of previously reported transcriptomic data (Gongora-Castillo *et al.*, 2012) using the newly sequenced draft genome has the potential to improve not only the resolution but also the accuracy of co-expression analyses, as the genome assembly provides a near-complete representation of the *C. roseus* gene repertoire and allows resolution of paralogs and splice isoforms that are collapsed in the transcriptome assembly. Using RNA-sequencing reads generated from a set of developmental tissues and sterile seedlings treated with MeJA (Gongora-Castillo *et al.*, 2012), we identified 956 genes whose expression was increased in 5 day or 12 day MeJA-treated seedlings compared with untreated sterile seedlings (Table S4 and Dataset S1). This set included the majority of the genes in the iridoid and alkaloid parts of the MIA pathway (Figure 2). To more broadly examine co-expression of MIA pathway genes, we performed hierarchical clustering with 15 681 genes using expression values from five developmental tissues and sterile seedlings treated with MeJA. Co-expression of pathway genes was readily apparent (Figure 3) with a large number of genes involved in MEP, iridoid, and alkaloid biosynthesis occurring in co-expression clusters with other genes in the pathway). Remarkably, co-expression clusters roughly correlating to distinct stages in the vinblastine/vincristine pathway were evident with a co-expression cluster of

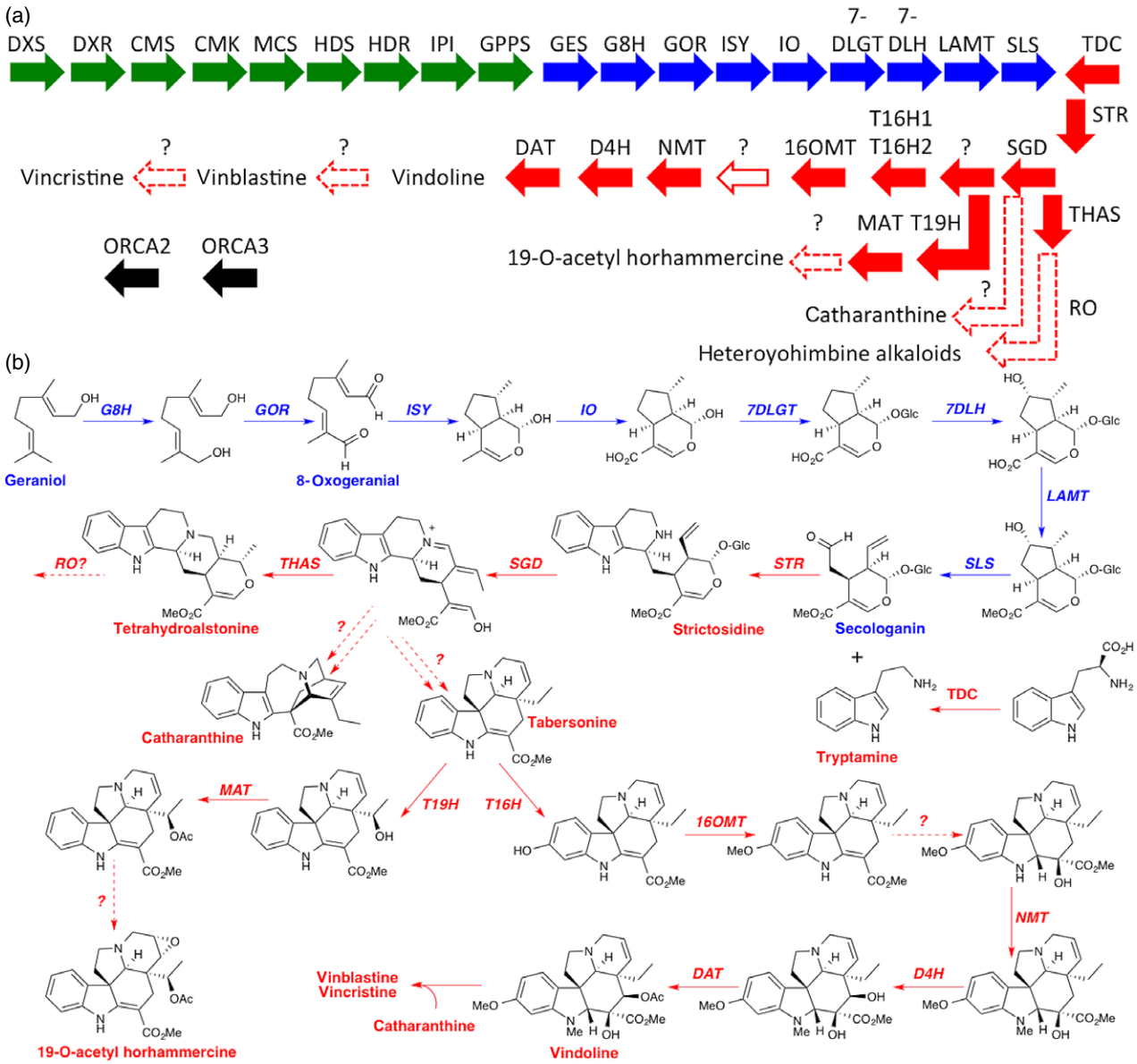

Figure 2. Branches of the monoterpene indole alkaloid (MIA) pathway in *C. roseus* for which biosynthetic genes have been identified.
Missing genes in the shown pathway branches are represented by dashed arrows.
(a) Green arrows represent methylerythritol phosphate (MEP)/upstream terpene biosynthesis genes; blue arrows represent iridoid biosynthesis genes; red arrows represent downstream alkaloid biosynthesis genes; black arrows represent transcription factors known to regulate MIA biosynthesis.
(b) Chemical representation of the iridoid (blue) and downstream alkaloid (red) biosynthetic pathway.
DXS, 1-deoxy-D-xylulose 5-phosphate synthase 2; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; CMS, 4-diphosphocytidyl-methylerythritol 2-phosphate synthase; CMK, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MCS, 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS, GCPE protein; HDR, 1-hydroxy-2-methyl-butenyl 4-diphosphate reductase; IPI, plastid isopentenyl pyrophosphate, dimethylallyl pyrophosphate isomerase; GPPS, geranyl pyrophosphate synthase; GES, plastid geraniol synthase; G8H, geraniol 8-hydroxylase; GOR, 8-hydroxygeraniol oxidoreductase; ISY, iridoid synthase; IO, iridoid oxidase (CYP76A26); 7DLGT, UDP-glucose iridoid glucosyltransferase; 7DLH, 7-deoxyloganic acid 7-hydroxylase; LAMT, loganic acid methyltransferase; SLS, secologanin synthase; TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGD, strictosidine β-glucosidase; T16H1, tabersonine 16-hydroxylase 1 (CYP71D12); T16H2, tabersonine 16-hydroxylase 2 (CYP71D351); 16OMT, 16-hydroxytabersonine *O*-methyltransferase; NMT, 16-hydroxy-2,3-dihydro-3-hydroxytabersonine *N*-methyltransferase; D4H, desacetoxyvindoline-4-hydroxylase; DAT, deacetylvindoline 4-*O*-acetyltransferase; THAS, tetrahydroalstonine synthase; RO, reticuline oxidase-like protein; T19H, tabersonine/lochnericine 19-hydroxylase (CYP71BJ1); MAT, minovincinine 19-hydroxy-*O*-acetyltransferase.

genes from the upstream MEP pathway (Figure 3a), a co-expression cluster of genes from the iridoid pathway (Figure 3b), and a co-expression cluster of genes at the terminus of the iridoid pathway and initiation of the alkaloid portion of the pathway (Figure 3c). This level of tight co-regulation of genes within the three major components

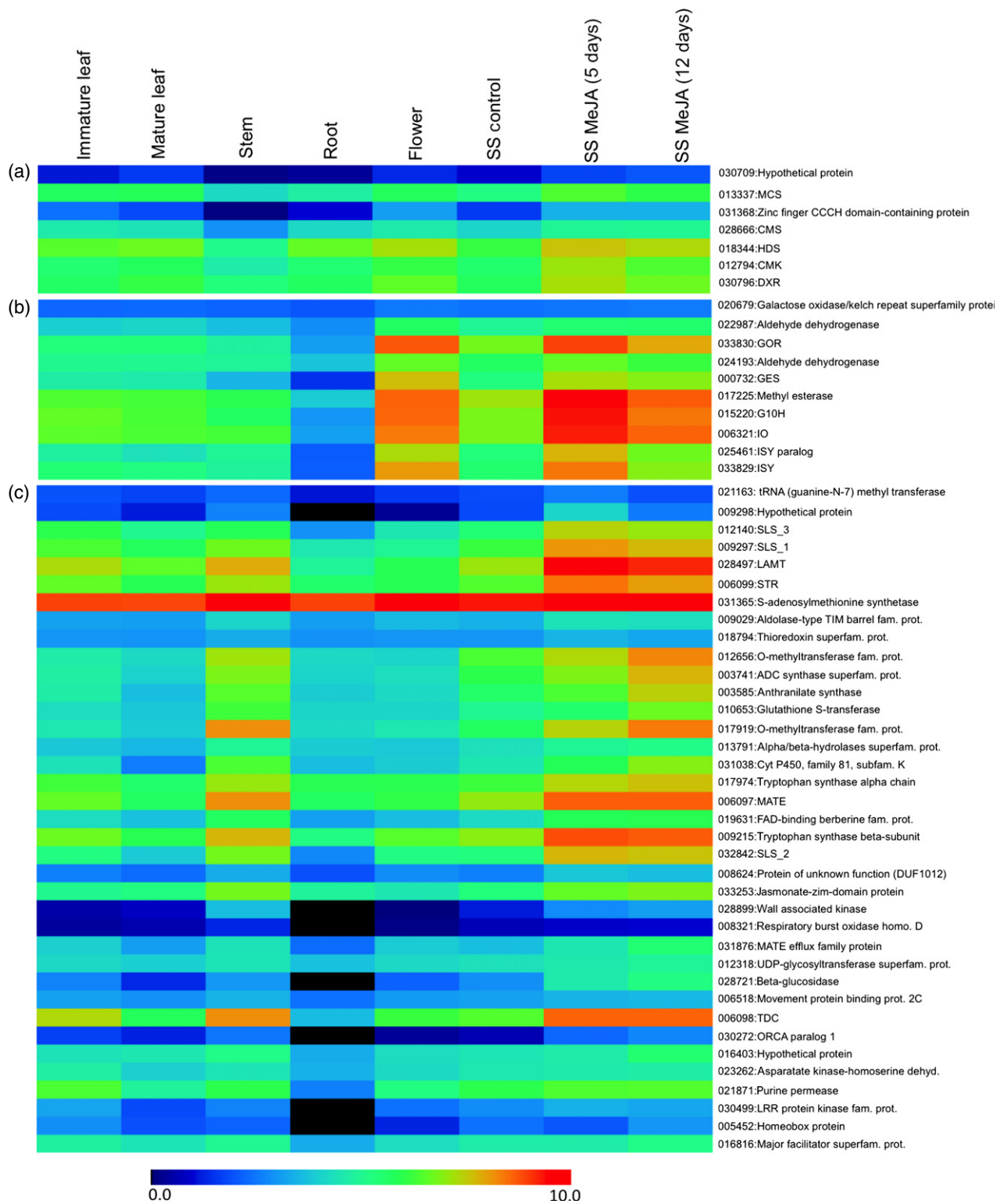

Figure 3. Hierarchical clustering (log_2 of FPKM) of expression of 15 681 *C. roseus* genes in primary stems, flowers, mature leaves, immature leaves, roots, sterile seedlings, and sterile seedlings after 5 days' and 12 days' treatments with methyl jasmonate.
(a) Co-expression cluster containing genes from the upstream MEP pathway (MCS, CMS, HDS, CMK, DXR).
(b) Co-expression cluster containing genes from the iridoid pathway (GOR, GES, G10H, IO, ISY paralog, ISY).
(c) Co-expression cluster containing genes at the terminus of the iridoid pathway and initiation of the alkaloid portion of the pathway (SLS1, SLS2, SLS3, LAMT, STR, TDC), two MATEs, and an ORCA paralog.

of this pathway suggests coordinated regulation of biosynthesis of precursors and intermediates throughout the pathway.

Gene clusters in monoterpene indole alkaloid biosynthesis

While physical clustering of specialized metabolism pathway genes is common in prokaryotes and fungi, the phenomenon of non-homologous gene clustering on plant chromosomes was first noted for 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) biosynthesis in maize in 1997 (Frey *et al.*, 1997). Subsequently, other biosynthetic gene clusters encoding for plant specialized metabolism have been reported, such as the thalinol (thale cress) (Field and Osbourn, 2008), avenacin (oat) (Qi *et al.*, 2004), noscapine (opium poppy) (Winzer *et al.*, 2012), monoterpenes (tomato) (Matsuba *et al.*, 2013), and steroidal glycoalkaloid (potato and tomato) (Itkin *et al.*, 2013) clusters. We used the *C. roseus* assembly to assess whether any known MEP, iridoid, or alkaloid biosynthetic genes are physically clustered in the *C. roseus* genome (Figure 2a). Most strikingly, TDC and STR were both located on a 30 kbp scaffold (Figure 4a). TDC is a pyridoxal dependent aromatic acid decarboxylase that generates tryptamine from tryptophan. In contrast, STR is a 'Pictet-Spenglerase' that condenses tryptamine with the iridoid secologanin to form strictosidine, the central biosynthetic intermediate for all known MIAs (Figure 2b). While the scaffold containing T16H1 and its paralog, T16H2, did not contain any other genes, sequence from a bacterial artificial chromosome (BAC) physically linked two scaffolds that not only contain the P450-encoding genes T16H1 and T16H2, but also 16OMT, the *O*-methyltransferase that methylates the hydroxyl group installed by T16H1 or T16H2 (Figure 4c).

Prospects for biosynthetic gene discovery

Despite decades of effort, the complete MIA pathway has not been fully elucidated in any plant species. After observing physical clustering of the known MIA genes TDC/STR and T16H1/T16H2/16OMT, we systematically mined the genome sequence flanking each known MEP, iridoid, and alkaloid biosynthetic gene to identify genes that may encode undiscovered MIA biosynthetic genes. Flanking the validated TDC and STR cluster was a gene encoding a multi-antimicrobial extrusion protein (MATE, 006097; Figure 4a and Table S3), a class of transporter implicated in transport of natural product biosynthetic intermediates

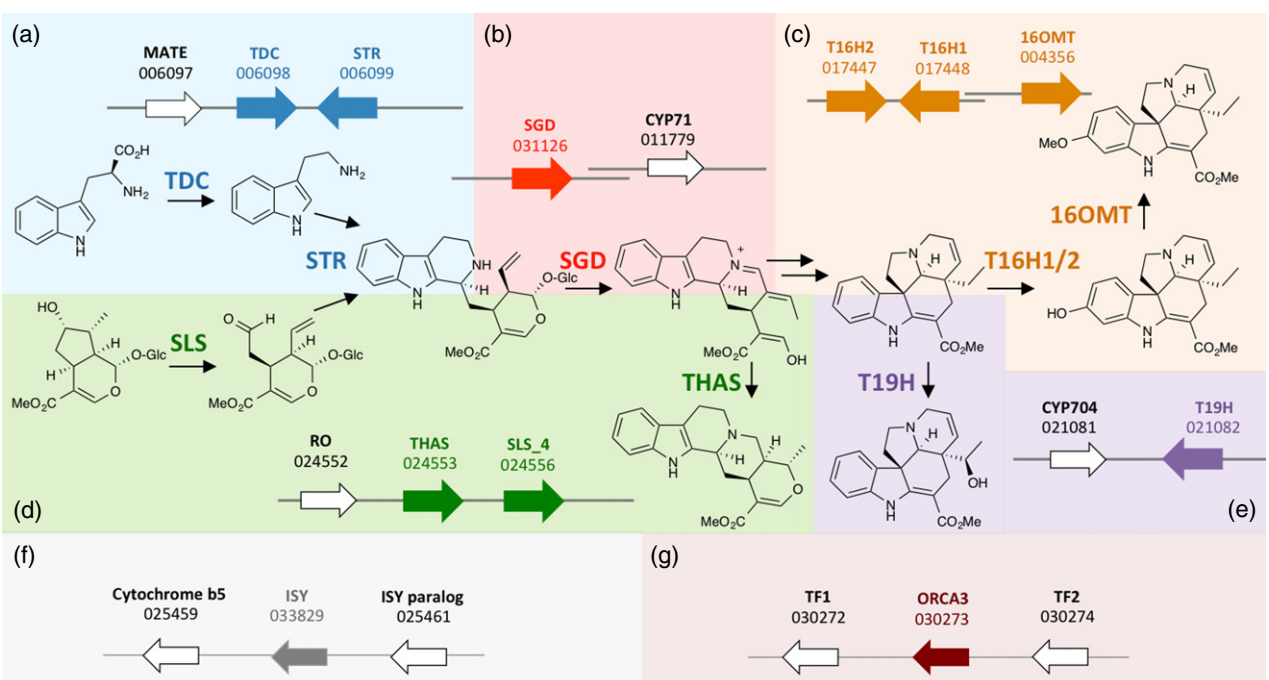

Figure 4. Physical clustering of known MIA pathway genes.
Potential uncharacterized genes in MIA biosynthesis (shown as white arrows) that cluster with known MIA pathway and regulatory genes (shown as filled arrows). Examples of gene duplication are also shown.
(a) WGS scaffold 3045674 (29 490 bp) containing TDC, STR and an uncharacterized MATE (blue box).
(b) BAC_SGD contigs c2 and c11 containing SGD and an uncharacterized P450 (red box).
(c) BAC_T16H contigs c3 and c1 containing T16H1, T16H2, and 16OMT (orange box).
(d) WGS scaffold 3063455 (76 903 bp) containing SLS paralog (SLS_4), THAS and uncharacterized RO (green box).
(e) WGS scaffold 2964965 (34 685 bp) containing T19H and an uncharacterized P450 (purple box).
(f) WGS scaffold 3047130 (46 594 bp) containing ISY, ISY paralog and uncharacterized cytochrome b5 (grey box).
(g) WGS scaffold 3061731 (40 593 bp) containing transcription factor ORCA3 plus two similar uncharacterized transcription factors (maroon box).

(Debeaujon *et al.*, 2001; Morita *et al.*, 2009). While many bacterial and fungal biosynthetic gene clusters contain transporters, the presence of a transporter gene encoded in a plant biosynthetic cluster is rare (Krattinger *et al.*, 2009). This transporter is also tightly co-expressed with STR and TDC (Figure 3c). In total, 61 predicted proteins in the *C. roseus* genome encode a MATE domain (Pfam ID PF01554). Notably, a substantial fraction of these, 15 of 61 MATE domain containing proteins, are up-regulated in response to MeJA (Table S4), suggesting that this transporter class is important in *C. roseus* specialized metabolism. We hypothesize that this MATE is involved in transport of an alkaloid intermediate or product. However, extensive functional characterization of this transporter will be required to support this hypothesis.

For the iridoid pathway gene SLS, four very close paralogs were annotated in the genome (Table S3). Physically clustered next to one of those paralogs (SLS_4) is an alcohol dehydrogenase (tetrahydroalstonine synthase, THAS, 024553) and a reticuline oxidase-like protein (RO, 024552) (Figure 4d and Table S3). Recent work has confirmed that the alcohol dehydrogenase THAS is responsible for the conversion of strictosidine aglycone (Figure 2) to the MIA tetrahydroalstonine, a monomeric MIA of the heteroyohimbine structural class (Stavrinides *et al.*, 2015). This is a crucial example of a gene that acts at the critical branch point of SGD, where the chemical diversity of the different classes of MIAs emerges. The best-characterized homolog of RO is the berberine bridge enzyme (*Eschscholtzia californica* (Dittrich and Kutchan, 1991), which catalyzes carbon-carbon bond formation in an alkaloid pathway unrelated to MIA biosynthesis. We speculate, based on the chemical similarity of the respective biochemical reactions, that RO could be involved in oxidizing tetrahydroalstonine to form alstonine, or oxidizing raubasine to serpentine (Figures 1 and S3). We conclude that at least one gene encoding heteroyohimbine biosynthesis is physically clustered with one of the distinct paralogs of SLS, which controls secologanin biosynthesis, a biosynthetic step two steps upstream (Figures 2 and 4d), clearly linking the iridoid and alkaloid portions of the pathway.

Additional uncharacterized genes that may encode enzymes involved in specialized metabolism were observed. For example, a P450-domain encoding gene (021081) is proximal to T19H and a cytochrome b5 (025459) flanks ISY (Figure 4e,f). This cytochrome b5 does not appear to have an obvious pathway role, although cytochrome b5 enzymes are known to enhance the activity of P450 enzymes (Im and Waskell, 2011). While nothing could be inferred about the genomic context of SGD from the whole-genome assembly, a BAC containing SGD also contained a cytochrome P450 (011779) (Figure 4b). Improvement of the *C. roseus* genome with additional scaffolding information, which would extend the length of contiguous sequences, may reveal additional gene clusters and provide leads to the discovery of hitherto elusive pathway enzymes that catalyze new biochemical reactions.

Gene duplication and sub-/neo-functionalization of monoterpene indole alkaloid biosynthetic paralogous genes

Nearly all plant genomes sequenced to date have undergone whole-genome, segmental, and/or tandem duplications resulting in paralogous gene families in which genes can undergo sub-functionalization, neo-functionalization, or pseudogenization as evidenced by lineage-specific evolution of plant specialized metabolic genes (Rensing *et al.*, 2008; Chae *et al.*, 2014). To examine the evolution of MIA biosynthetic pathway genes in *C. roseus*, we explored the genomic context, gene structure, and phylogenetic relationships of select MIA biosynthetic pathway genes (Table S3). We first noted the presence of multiple paralogs of several MIA genes with a subset likely derived from tandem duplication events, consistent with the observation in several angiosperms that specialized metabolic genes were more significantly enriched in local (tandem) duplication events as compared with whole-genome duplication events (Chae *et al.*, 2014). Two T16H paralogs [88% nucleotide (nt) identity] were tandemly duplicated (5500 bp) on a single 32 kbp scaffold. Recent work has demonstrated that these T16H paralogs are functionally distinct. While T16H2 is predominantly responsible for alkaloid biosynthesis in leaves (Besseau *et al.*, 2013), the differential expression profiles of T16H1 and T16H2 in various tissues suggests that this duplication represents expression sub-functionalization as the biochemical activity of the duplicated genes remains identical (Besseau *et al.*, 2013). The gene *7DLGT* (Asada *et al.*, 2013) is found within a family of similar glucosyltransferases (Table S3), although there is no evidence to suggest that these paralogs are involved in alkaloid biosynthesis. ISY (Geu-Flores *et al.*, 2012) is located in close proximity to a close paralog (025461), representing another gene duplication (Figure 4f). While the ISY paralog exhibits identical biochemical activity to ISY *in vitro*, *in planta* silencing by VIGS suggests it may not be physiologically involved in iridoid biosynthesis (Munkert *et al.*, 2014).

Four near-identical SLS genes (94–98% nt sequence identity; Table S3) were identified on four separate scaffolds. All four paralogs are expressed in leaf tissue (Dataset S1). Despite several existing NCBI (National Center for Biotechnology Information) entries that report identical or near-identical sequences to those four SLS paralogs, only SLS_2 (L10081.1 99.7% nt identity with SLS_2) has been demonstrated as a functional secologanin synthase (Irmler *et al.*, 2000). These isoforms cannot be distinguished using short-read derived transcriptome data alone, which highlights how a draft reference genome can greatly improve existing expression profile datasets.

Strictosidine synthase (STR) has been previously shown to be involved in MIA biosynthesis (McKnight *et al.*, 1990) by catalyzing a Pictet-Spengler condensation between tryptamine and secologanin (Figure 2b) (O'Connor and Maresh, 2006). Twelve STR-domain (PF03088) containing genes were identified in the predicted *C. roseus* proteome. Phylogenetic analyses show that six STR family genes clustered with the previously characterized STR from the MIA pathway (006099; Figure S4). Similar to a number of MIA genes (Gongora-Castillo *et al.*, 2012), five of the STR-domain containing genes that clustered with 006099 were up-regulated in response to MeJA treatment (Figure S4 and Table S4), with only a single STR-domain containing gene (007860) in this cluster not up-regulated by MeJA treatment. Interestingly, 007860 is not expressed in any developmental tissue or MeJA treatment examined suggesting that this gene does not function in MIA biosynthesis. Five additional STR-domain containing genes more diverged from 006099 were not altered in gene expression following MeJA treatment. These paralogs of STR are not likely to catalyze the condensation of tryptamine and secologanin, since they lack the crucial catalytic residues found in the characterized STR enzyme. STR-like genes are found in many plants that do not produce alkaloids; for example, *A. thaliana* has more than 10 proteins with amino acid identity of >27% to *C. roseus* STR. While the functions of these STR-like proteins remain unclear, at least one STR homolog (*Vitis vinifera*) has been shown to exhibit hydrolase activity (Hicks *et al.*, 2011).

Transcription factors and transcriptional regulatory networks

Previous studies reported ORCA2 (Menke *et al.*, 1999) and ORCA3 (van der Fits and Memelink, 2000) are transcription factors (TFs) whose expression is induced by MeJA treatment and may regulate MIA biosynthesis. These two TFs were annotated in the genome assembly and, consistent with previous reports (Menke *et al.*, 1999; van der Fits and Memelink, 2000), both were up-regulated in response to MeJA treatment (Tables S3 and S4). Further analyses identified two additional putative TFs (030272 and 030274), adjacent to ORCA3, suggestive of tandem duplication events (Figures 4g and S5), both of which encode AP-2 domain DNA-binding proteins (PF00847). Similar to ORCA2 and ORCA3, expression of these two TFs was induced by MeJA treatment (Table S4). In particular, putative AP2 TF1 (030272) was co-expressed with TDC, STR, LAMT, and SLS1/2/3 (Figure 3c), suggesting that this TF may be involved in the production of MIA in *C. roseus.* Notably, these putative TFs (030272 and 030274) and ORCA3 (030273) are physically clustered on the same scaffold, yet 030272 and 030273 (ORCA3) have correlated expression profiles highly similar to MIA biosynthetic genes, while 030274 differs in expression patterns, suggesting expression-based neo-functionalization (Figure S5).

With access to our genome-based expression profiles and genome localization data, we constructed a transcriptional regulatory network for ORCA2 and then data-mined this for new, putative genes that may be involved in the MIA pathway using functional annotations and physical clustering within the genome. Using mutual rank analysis, which has been extremely productive in identifying co-expressed genes in the MIA pathway (Giddings *et al.*, 2011; Geu-Flores *et al.*, 2012), we generated a transcriptional regulatory network for ORCA2 (Figure 5 and Table S5). Within the ORCA2 network, we observed significant enrichment for genes encoding known MIA biosynthetic pathway components (six genes, *P*-value = 1.26e-07), as well as P450s (six genes, *P*-value = 0.0041), ABC transporters (three genes, *P*-value = 0.0365), MATE transporters (five genes, *P*-value = 4.96e-05), which are excellent candidates for undiscovered biosynthetic genes. We also noted that genes annotated as oxygenases, dehydrogenases/reductases, acyl transferases and glycosyltransferases, which also encode protein families that are excellent candidates for undiscovered secondary metabolic enzymes, were present in the network. Notably, two of the potential candidate genes, a MATE and an UDP-glucosyl transferase, were physically co-localized on the same scaffold as a validated MIA gene (*P*-value = 0.0426; Figure 5 and Table S5).

Within the ORCA2 network we found four TFs and we generated transcriptional networks for these as well (Figure 5). As with ORCA2, we observed substantial enrichment of genes encoding for MIA biosynthesis, P450s, ABC transporters, MATE transporters, TFs, as well as genes physically co-localized on the same scaffold as known MIA genes (Figure 5b). Overall, seven MIA pathway genes, 27 P450s, 14 MATEs, 15 ABC transporters, and 25 transcription factors were tightly co-expressed with ORCA2 or the four TFs (13335, 18593, 22277, or 28120), suggesting a robust level of secondary metabolite and transcriptional regulation is controlled by ORCA2. Notably, we identified 10 genes from these networks that have not been previously associated with MIA biosynthesis that are located on the same scaffold as a gene involved in MIA biosynthesis (Table S3). Functional characterization of these genes will provide a framework to rigorously test the link between co-regulation and genomic position.

DISCUSSION

The selective pressures that shape the organization of specialized metabolic biosynthetic genes remain elusive. It has been proposed that clustering of biosynthetic genes within the genome may have evolved to prevent the buildup of toxic intermediates (Field and Osbourn, 2008; Takos and Rook, 2012). However, while the genes encoding STR and TDC are physically clustered in the genome, the substrates of these enzymes – tryptophan, tryptamine, and secologanin – are not particularly toxic or unstable (McCoy and

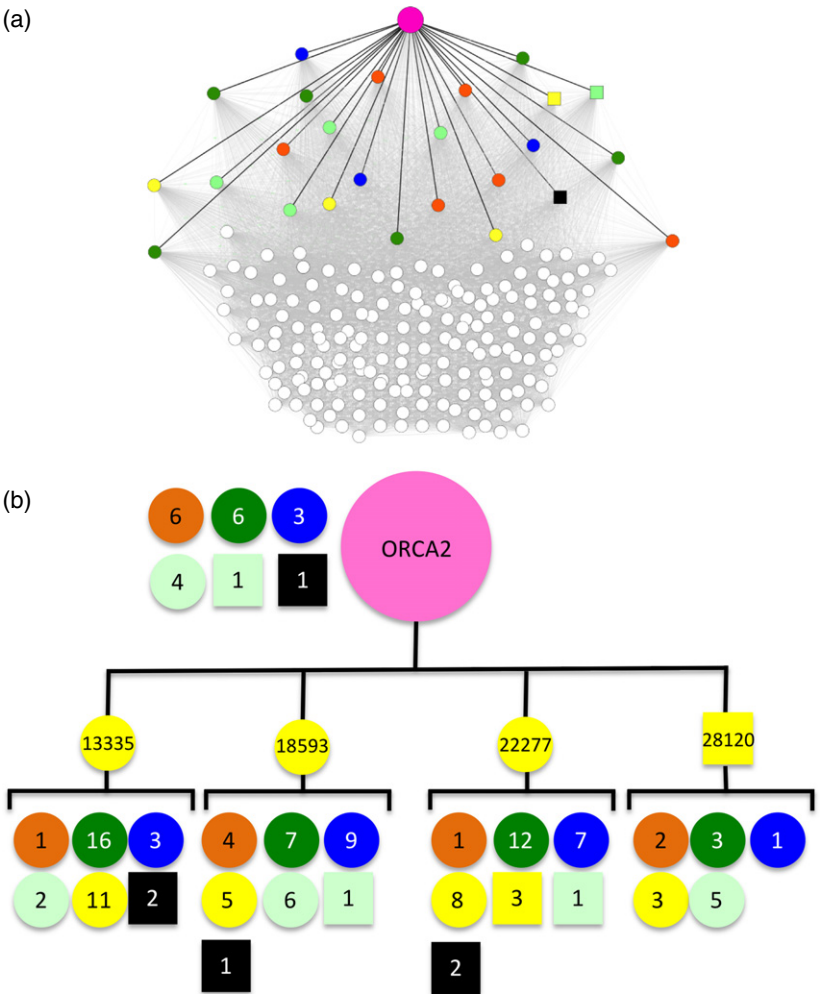

Figure 5. Transcriptional regulatory networks governed by ORCA2.
(a) ORCA2 co-expression network. A co-expression network for ORCA2 is shown with the top 200 genes co-expressed with ORCA2. Genes of interest are represented as circles colour-coded for each category: transcription factor (yellow), MIA-associated gene listed in Figure 2 (orange), P450 gene (dark green), MATE transporter (light green) and ABC transporter (blue). If a gene within one of these functional categories is also physically co-located with a validated MIA-associated gene (Table S3), it is depicted as a square. The black square represents an UDP-glucosyl transferase that is physically co-located with a validated MIA gene listed in Table S3.
(b) Additional transcriptional networks governed by ORCA2. Genes co-expressed by OCRA2 are shown at the top of the cascade using the same color- and shape-coding as in Figure 5(a) to represent function and physical co-localization, respectively. Four transcription factors tightly co-expressed with ORCA2 are shown as yellow circles (13335, 18593, 22277) and a yellow square (28120) with the genes tightly co-expressed with these transcription factors shown at the bottom using colour- and shape-coding for function and physical location.

O'Connor, 2006; Galan *et al.*, 2007). Similarly, the intermediates associated with T16H1/T16H2/16OMT are also stable. Physical clustering is also cited as a mechanism to ensure co-regulation, which is observed with TDC/STR/MATE cluster. However, not every cluster exhibits co-regulation. For example, the RO-like protein exhibits a different expression profile compared to its pathway neighbors THAS and SLS (Figure 2 and Dataset S1). It is possible that clusters of genes represent different stages of pathway evolution. For example, both TDC and STR are required for the MIA pathway to transition from synthesis of iridoids to synthesis of alkaloids (Figure 2) and the clustering of TDC and STR may reflect a distinct stage of evolution in which these alkaloids emerged from the iridoid monoterpenes. Furthermore, gene duplication either through whole-genome, segmental, or tandem (local) duplication provides a mechanism for sub- or neo-functionalization that may involve differential expression of the paralogs. Our survey of genes within the MIA biosynthetic pathway revealed a genome abundant with gene duplication events that have led to expression sub-functionalization and neo-functionalization of paralogous genes involved in MIA biosynthesis. Access to additional genomes from species within the Apocynaceae and/or related families should shed light on the evolution of MIA biosynthetic pathway genes including physical clustering.

Genome sequencing is not the only approach by which clustered pathways can be discovered. The recent elegant elucidation of the noscapine pathway used an F_2 mapping population to show that the biosynthetic pathway enzymes are tightly linked (Winzer *et al.*, 2012). While this is a powerful approach, mapping requires at least two cultivars with qualitatively different levels of specialized metabolites, which are not available for every plant species, or for every specialized metabolite of interest. For example, most *C. roseus* cultivars produce a relatively similar profile of MIAs (Magnotta *et al.*, 2006).

This study highlights how an inexpensive draft genome – at a cost accessible to an individual research group – that provides a comprehensive representation of the genic regions of a genome can be used to investigate a complex specialized metabolic pathway. While additional scaffolding

will likely provide additional insights into *C. roseus* metabolism, this draft genome sequence nevertheless can facilitate identification of additional biosynthetic steps in the MIA pathway, provide an improved understanding of the cellular and subcellular localization of MIA compounds, and enable further dissection of the transcriptional regulatory mechanisms of the MIA pathway. The capacity to rapidly and inexpensively generate quality genome sequence data provides an important addition to the growing set of approaches that can be used to unravel plant specialized metabolism.

The genome mining strategies that have revolutionized the field of microbial natural products may never be completely applicable to plant metabolism, given that plant pathways are both incompletely and unpredictably clustered. Nevertheless, the prospect of obtaining draft quality plant genomes at a reasonable cost will further accelerate the speed at which we can unravel pathways for complex molecules such as vincristine and vinblastine.

EXPERIMENTAL PROCEDURES

Plant material, genome sequencing, assembly, and assessment

Genomic DNA was isolated from purified nuclei of young leaves of 3-month-old plants of *C. roseus* 'SunStorm™ Apricot' grown in a growth chamber at 25°C with 12 h light/12 h dark as previously described (Brenchley *et al.*, 2012). A single TruSeq genomic DNA library was constructed (398-bp fragment size) and sequenced on an Illumina HiSeq at The Genome Analysis Centre (Norwich, UK, http://www.tgac.ac.uk/main-icons/platforms/sequencing-platforms/) generating 374 771 760 101 nucleotide paired-end reads. Reads were assembled using ABYSS (Simpson *et al.*, 2009) using a k-mer size of 71. In total, the assembly consists of 79 302 scaffolds >200 bp representing 522 653 749 bp with an N50 scaffold size of 26 249 bp. The quality of the assembly was assessed by alignment of 20 181 *C. roseus* ESTs using GMAP (2014-05-30 v2) and by the identification of 248 conserved eukaryotic genes using CEGMA v2.4 (Parra *et al.*, 2007).

To assess the quality and representation of the *C. roseus* genome in the assembly, individual Illumina reads were adapter trimmed and quality filtered with CUTADAPT (v.1.2.1, https://code.google.com/p/cutadapt/) (Martin, 2011) using a minimum quality value of 10 and a minimum trimmed read length of 30 bp. The cleaned reads were aligned to the assembly (min 200 bp scaffolds) using BWA-MEM (http://arxiv.org/abs/1303.3997v2; v.0.7.8) in single-end mode using the –M option. Duplicate reads were marked in the BAM output using PICARD MARKDUPLICATES (http://broadinstitute.github.io/picard/; v.1.106). Alignments surrounding putative indels were refined using the IndelRealigner tool from the GATK package (v3.3.0) (DePristo *et al.*, 2011). Single nucleotide polymorphisms (SNPs) were called using the GATK HAPLOTYPECALLER using a confidence threshold of 30 or calling and emitting variants (-stand_call_conf 30, -stand_emit_conf 30) and a minimum read mapping score of 20 (-mmq 20). Insertion/deletion calls were filtered from the VCF file using VCFTOOLS (v0.1.12b) (Danecek *et al.*, 2011). The remaining SNPs were hard filtered for quality and maximum and minimum depth using VCFTOOLS vcf-annotate (D = 100/Q = 30/q = 10/d = 5/r) and homozygous SNPs representing sequencing and assembly errors were identified using a custom Perl script.

To determine the heterozygosity rate, the *C. roseus* Illumina genomic DNA reads were aligned to the genome assembly using BWA-MEM (v.0.7.8) and duplicate read alignments were marked using PICARD MARKDUPLICATES (v.1.86). Alignments around insertions/deletions were refined using the GATK INDELREALIGNER (v.2.8.1) and SNPs were called using SAMTOOLS MPILEUP (Li *et al.*, 2009) and converted into VCF format using VCFTOOLS (v.0.1.19). SNPs were filtered using VCFTOOLS VCF-ANNOTATE (v.0.1.11) using a hard filter for quality and maximum and minimum depth (D = 100/Q = 20/q = 10/d = 5/r).

Genome annotation

Genome annotation was performed using the MAKER annotation pipeline (release 1103) (Holt and Yandell, 2011; Campbell *et al.*, 2014). The *C. roseus* genome was masked by REPEATMASKER using the RepBase repeat library (www.repeatmasker.org) (Jurka *et al.*, 2005). An initial HMM for the *ab initio* gene prediction program, SNAP (Korf, 2004), was trained using MAKER-generated alignments of *C. roseus* Sanger-derived ESTs downloaded from GenBank and *C. roseus* transcript assemblies (Gongora-Castillo *et al.*, 2012) to the masked *C. roseus* genome sequence. MAKER gene predictions were generated using the initial SNAP HMM with UniProt SwissProt plant protein alignments and *C. roseus* EST and transcript alignments as evidence. A subset of the first high-confidence SNAP gene predictions was used to train a second SNAP HMM. MAKER was then run with the second SNAP HMM using the same protein and transcript alignments as evidence. A subset of high-confidence predictions from the second SNAP HMM was used to train Augustus (Stanke and Waack, 2003). A final MAKER run was performed using *C. roseus* EST and transcript assembly alignments, UniProt SwissProt plant protein alignments, and *A. thaliana* TAIR10 protein alignments as evidence. SNAP and Augustus were used for gene predictions using the *C. roseus* trained HMMs, and FGENESH was also used for gene predictions using the tomato gene matrix (Salamov and Solovyev, 2000). For the final MAKER run, the single exon EST alignments >250 nucleotides were allowed as evidence. Additionally, MAKER was set to limit the use of single exon ESTs when generating final gene models, and MAKER was allowed to output unsupported (*ab initio*) gene models. The hmmscan tool within the HMMER3 package (Eddy, 2011) was used to identify MAKER-predicted genes containing Pfam protein domains (Finn *et al.*, 2014). Using only scaffolds >1000 bp, a final set of high-confidence gene predictions were identified from those genes that were supported by transcript or protein evidence and/or containing a Pfam domain. Functional annotation was determined in a hierarchical manner using evidence from alignment to the TAIR10 *A. thaliana* proteome, followed by Pfam domain composition, and finally alignment to annotated SwissProt plant proteins. Functional descriptions for all MIA pathway genes (SI, Table S3) were transitively annotated using cloned GenBank entries from *C. roseus.*

Bacterial artificial chromosome library construction, screening, and sequencing

A BAC library was constructed from 'SunStorm™ Apricot' using the pINDIGOBAC-5 vector in *E. coli* strain DH10B by Bio S&T (www.bios&t.com). The library was pooled in a 96-well plate with each well containing approximately 500 independent primary clones. The library was confirmed to have 10× coverage and an average insert size of 155 kb. For screening, genomic sequence

specific primers for SGD and T16H2 were designed (Table S6) and BAC pools were screened by Bio S&T. Positive BACs were fingerprinted and BAC DNA was used to construct a single TruSeq DNA library that was sequenced on a MiSeq (150-nt or 250-nt paired-end reads). BAC assemblies were performed using MIRA, version 4.0rc4 (http://sourceforge.net/projects/mira-assembler). The first 50 000 read pairs of each dataset were used for further processing. Each read was aligned to the vector sequence (EU140754.1) and the genome sequence of *E. coli* strand K12, substrand DH10B (CP000948.1). Only reads with a BLASTN hit (Zhang *et al.*, 2000) with an e-value less than 1E-10 to the vector within 1 Kb distance to the restriction site and reads without any hit to *E. coli* or vector were used in the assembly. Remaining vector parts were clipped from the assembled contigs.

Expression abundances and differentially expressed genes. Expression abundance in developmental tissues and MeJA-treated seedlings was determined using RNA-seq data from a previous study (Gongora-Castillo *et al.*, 2012). Reads (National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) accessions SRR122239, SRR122243, SRR122244, SRR122245, SRR122251, SRR122252, SRR122253, and SRR122254) were assessed for quality using FASTQC (v 0.10.0) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and cleaned of adaptors and low quality sequences using CUTADAPT version (v.1.4.1) (Martin, 2011) allowing a minimum quality score of 20 and a minimum read length of 30 nt. Reads were aligned to the *C. roseus* genome using BOWTIE (v.0.12.7) (Langmead *et al.*, 2009) and TOPHAT version (v.1.4.1) (Trapnell *et al.*, 2009) with a minimum and maximum intron size of 5 and 19 000 bp, respectively. The segment length was set to 15 bp and all other defaults were maintained. Expression quantification in fragments per kilobase exon model per million mapped reads (FPKM) was performed using CUFFLINKS (v.1.3.0) (Trapnell *et al.*, 2010). Hierarchical clustering was performed on 15 681 genes which had an FPKM value >0 in 5 day MeJA-treated seedlings and an FPKM value >1 in mature leaves using the MEV: MULTIEXPERIMENT VIEWER software (http://www.tm4.org/).

Differentially expressed genes following treatment of sterile seedlings (SRR122243) with MeJA (SRR122244, SRR122245) were determined using TOPHAT output BAM files with EDGER (Robinson *et al.*, 2010). MeJA differentially expressed genes were defined based on two criteria: (1) $\log_2$ fold change compared with control sterile seedling >1; and (2) *P*-value <0.05 for either 5 day or 12 day MeJA treatment using a coefficient variance of 0.4.

Identification of known MIA genes in the *Catharanthus roseus* genome

Transcript and peptide sequences for 30 previously published MIA pathway genes and two TFs were collected from NCBI. *C. roseus* annotations corresponding to each MIA pathway or TF gene were identified by sequence alignment using GMAP (release 2014-05-30) (Wu and Watanabe, 2005) and BLAST (Altschul *et al.*, 1990) analyses, followed by manual inspection (Table S3).

Identification of orthologs and paralogs

For phylogenetic analyses of known MIA genes, ORTHOMCL (v.1.; Li *et al.* 2003), analysis was performed with default parameters using predicted *C. roseus* proteins along with the predicted proteomes of just *Amborella trichopoda* (v.1, http://www.amborella.org), *Arabidopsis thaliana* (v.10, https://www.arabidopsis.org), grapevine (v.1, http://www.phytozome.net/grape.php), and tomato (v.2.4, http://www.phytozome.net/tomato.php). Phylogenetic trees (Figures S4 and S5) were generated using MEGA6 (Tamura *et al.*, 2013) with the default parameters using a MUSCLE alignment with UPGMB methods, neighbor-joining methods with Poisson model and pairwise deletion, and bootstrap $n = 1000$.

Analysis of transcriptional regulatory networks

Mutual rank analysis was performed to identify genes co-expressed with ORCA2 using the filtered FPKM matrix that only includes genes that have FPKM ≥5 at least under at least one tissue type or treatment. The top 200 ranked genes were selected for data-mining. Six categories of genes including MIA pathway genes, P450s, TFs, ABC-, MATE transporters, and genes physically co-localized with previously characterized MIA genes were defined based on functional annotation and physical location of each gene. Genes containing PFAM domains PF01554 were classified as MATE transporters and genes containing PF00005 or PF00664 as ABC transporters, respectively. With the top 200 ranked genes identified from ORCA2 analysis, the co-expression network was visualized using CYTOSCAPE (v.3.2.0) (Shannon *et al.*, 2003).

Accession numbers and data access

Raw genome reads are available in the NCBI SRA under BioProject number PRJNA252611. The assembled genome (*C. roseus* SunStorm™ Apricot v.1.0) has been deposited in the NCBI Whole-Genome Shotgun Sequence database under the accession JQHZ00000000. The version described in this paper is version JQHZ01000000. BACs containing the T16H1, T16H2, and 16OMT cluster and SGD have been deposited in NCBI WGS under accession numbers ERP006960 and PRJEB7256, respectively. Accession numbers for cloned genes described in this study are listed in Table S3. FASTA files of the annotated genes, transcripts, and peptides, scaffolds (≥200 and ≥1000 bp), and the annotation in GFF3 format are available for download from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.hs593). A searchable genome browser, functional annotation tool, and BLAST database for the *C. roseus* SunStorm™ Apricot v1.0 annotated genome is available at The Medicinal Plant Genomics Resource (http://medicinalplantgenomics.msu.edu/).

ACKNOWLEDGEMENTS

Funds to support this work were made available by the European Research Council (ERC R20359) to S.E.O., by the Michigan State University Foundation to C.R.B., and by the National Science Foundation (grant no. IOS–1126998) to K.L.C. F.K. is supported by a studentship from the University of East Anglia. We gratefully acknowledge the following members of TGAC, Sophie Janacek (Project Manager), Lawrence Percival-Alwyn (Library Construction), Rachel Piddock (DNA sequencing), Richard Leggett and Darren Waite (Sequencing Informatics). We thank Anna Stavrinides and Fernando Geu-Flores for helpful discussions regarding the function of RO. The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

F.K., J.K., J.P.H., B.V., K.L.C., J.C., B.S. and C.R.B. performed data analyses. B.C. performed the assembly. L.C. and K.M. performed library construction and sequencing. F.K. prepared plant material, DNA samples and BACs. F.K., J.K., K.L.C., C.R.B., S.E.O. wrote the paper. All authors assisted in revisions.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Kmer spectra divided by copy number on the assemblies.

Figure S2. Total and distinct 31-mers for different cutoffs on the assembled sequences.

Figure S3. Proposed reaction of RO is oxidation of a tetrahydro-β-carboline to a β-carboline.

Figure S4. Expansion of the strictosidine synthase domain encoding gene family in *Catharanthus roseus.*

Figure S5. Duplication of AP2 class transcription factors in *Catharanthus roseus.*

Table S1. Assembly and annotation metrics for the *Catharanthus roseus* genome.

Table S2. Gene model annotation metrics.

Table S3. *Catharanthus roseus* genes involved in the biosynthesis and regulation of monoterpene indole alkaloids and their paralogs described in this study.

Table S4. Genes differentially expressed in *Catharanthus roseus* sterile seedlings after 5 or 12 days treatment with methyl jasmonate.

Table S5. Mutual rank analysis of the ORCA2 transcriptional cascade.

Table S6. Primer sequences used for screening the *Catharanthus roseus* BAC library.

Dataset S1. Scaffold location, gene annotation, and expression abundances of annotated *Catharanthus roseus* genes in a developmental tissue series and in seedlings treated with methyl jasmonate.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Amoutzias, G. and Van de Peer, Y. (2008) Together we stand: genes cluster to coordinate regulation. *Dev. Cell*, **14**, 640–642.

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

Aslam, J., Khan, S.H., Siddiqui, Z.H. *et al.* (2010) Catharanthus roseus (L.) G. Don. An important drug: it's applications and production. *Pharmacie Globale (IJCP)*, **4**, 1–16.

Besseau, S., Kellner, F., Lanoue, A. *et al.* (2013) A pair of tabersonine 16-hydroxylases initiates the synthesis of vindoline in an organ-dependent manner in *Catharanthus roseus*. *Plant Physiol.* **163**, 1792–1803.

Brenchley, R., Spannagl, M., Pfeifer, M. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Campbell, M.S., Law, M., Holt, C. *et al.* (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524.

Chae, L., Kim, T., Nilo-Poyanco, R. and Rhee, S.Y. (2014) Genomic signatures of specialized metabolism in plants. *Science*, **344**, 510–513.

Champagne, A., Rischer, H., Oksman-Caldentey, K.-M. and Boutr, M. (2012) In-depth proteome mining of cultured *Catharanthus roseus* cells identifies candidate proteins involved in the synthesis and transport of secondary metabolites. *Proteomics*, **12**, 2536–2547.

Danecek, P., Auton, A., Abecasis, G. *et al.* (2011) The variant call format and VCFTools. *Bioinformatics*, **27**, 2156–2158.

Debeaujon, I., Peeters, A.J.M., Léon-Kloosterziel, K.M. and Koornneef, M. (2001) The TRANSPARENT TESTA12 gene of Arabidopsis encodes a multidrug secondary transporter-like protein required for flavonoid sequestration in vacuoles of the seed coat endothelium. *Plant Cell*, **13**, 853–871.

DePristo, M.A., Banks, E., Poplin, R. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498.

Dittrich, H. and Kutchan, T.M. (1991) Molecular cloning, expression, and induction of berberine bridge enzyme, an enzyme essential to the formation of benzophenanthridine alkaloids in the response of plants to pathogenic attack. *Proc. Natl Acad. Sci. USA*, **88**, 9969–9973.

Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195.

Field, B. and Osbourn, A.E. (2008) Metabolic diversification- independent assembly of operon-like gene clusters in different plants. *Science*, **320**, 543–547.

Finn, R.D., Bateman, A., Clements, J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.* **42**, 222–230.

van der Fits, L. and Memelink, J. (2000) ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science*, **289**, 295–297.

Frey, M., Chomet, P., Glawischnig, E. *et al.* (1997) Analysis of a chemical plant defense mechanism in grasses. *Science*, **277**, 696–699.

Galan, M.C., McCoy, E. and O'Connor, S.E. (2007) Chemoselective derivatization of alkaloids in periwinkle. *Chem. Commun.* **31**, 3249–3251.

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

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

Gongora-Castillo, E., Childs, K.L., Fedewa, G. *et al.* (2012) Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PLoS One*, **7**, e52506.

Guimaraes, G., Cardoso, L., Oliveira, H., Santos, C., Duarte, P. and Sottomayor, M. (2012) Cytogenetic characterization and genome size of the medicinal plant *Catharanthus roseus* (L.) G. Don. *AoB Plants*, **2012**, pls002.

Hicks, M.A., Barber, A.E., Giddings, L.A., Caldwell, J., O'Connor, S.E. and Babbitt, P.C. (2011) The evolution of function in strictosidine synthase-like proteins. *Proteins*, **79**, 3082–3098.

Hirsch, C.N. and Buell, C.R. (2013) Tapping the promise of genomics in species with complex, nonmodel genomes. *Annu. Rev. Plant Biol.* **64**, 89–110.

Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.

Im, S.C. and Waskell, L. (2011) The interaction of microsomal cytochrome P450 2B4 with its redox partners, cytochrome P450 reductase and cytochrome b(5). *Arch. Biochem. Biophys.* **507**, 144–153.

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

Itkin, M., Heinig, U., Tzfadia, O. *et al.* (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, **341**, 175–179.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) RepBase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **14**, 59.

Krattinger, S.G., Laguda, E.S., Spielmeyer, W., Singh, R.P., Huerta-Espino, J., McFadden, H., Bossolini, E., Selter, L.L. and Keller, B. (2009) A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science*, **323**, 1360–1363.

Ku, C., Chung, W.-C., Chen, L.-L. and Kuo, C.-H. (2013) The complete plastid genome sequence of Madagascar periwinkle *Catharanthus roseus* (l.) g. don: plastid genome evolution, molecular marker identification, and phylogenetic implications in asterids. *PLoS One*, **8**, e68518.

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

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.

Li, H., Handsaker, B., Wysoker, A. *et al.* (2009) The Sequence Alignment/Map format and SAMTools. *Bioinformatics*, **25**, 2078–2079.

Magnotta, M., Murata, J., Chen, J. and De Luca, V. (2006) Identification of a low vindoline accumulating cultivar of *Catharanthus roseus* (L.) G. Don by alkaloid and enzymatic profiling. *Phytochemistry*, **67**, 1758–1764.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12.

Matsuba, Y., Nguyen, T.T., Wiegert, K. *et al.* (2013) Evolution of a complex locus for terpene biosynthesis in Solanum. *Plant Cell*, **25**, 2022–2036.

McCoy, E. and O'Connor, S.E. (2006) Directed biosynthesis of alkaloid analogs in the medicinal plant *Catharanthus roseus*. *J. Am. Chem. Soc.* **128**, 14276–14277.

McKnight, T.D., Roessner, C.A., Devagupta, R., Scott, A.I. and Nessler, C.L. (1990) Nucleotide sequence of a cDNA encoding the vacuolar protein strictosidine synthase from *Catharanthus roseus*. *Nucleic Acids Res.* **18**, 4939.

Menke, F.L., Champion, A., Kijne, J.W. and Memelink, J. (1999) A novel jasmonate- and elicitor-responsive element in the periwinkle secondary metabolite biosynthetic gene Str interacts with a jasmonate- and elicitor-inducible AP2-domain transcription factor, ORCA2. *EMBO J.* **18**, 4455–4463.

Morita, M., Shitan, N., Sawada, K., Van Montagu, M.C.E., Inze, D., Rischer, H., Goossens, A., Oksman-Caldentey, K.M., Moriyama, Y. and Yazaki, K. (2009) Vacuolar transport of nicotine is mediated by a multidrug and toxic compound extrusion (MATE) transporter in *Nicotiana tabacum*. *Proc. Natl Acad. Sci. USA*, **106**, 2447–2452.

Mugford, S.T., Louveau, T., Melton, R. *et al.* (2013) Modularity of plant metabolic gene clusters: a trio of linked genes that are collectively required for acylation of triterpenes in oat. *Plant Cell*, **25**, 1078–1092.

Munkert, J., Pollier, J., Miettinen, K. *et al.* (2014) Iridoid synthase activity is common among the plant progesterone 5β-reductase family. *Mol. Plant*, **8**, 136–152.

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.

Nutzmann, H.W. and Osbourn, A. (2014) Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **26**, 91–99.

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

Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R. and Osbourn, A. (2004) A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl Acad. Sci. USA*, **101**, 8233–8238.

Rensing, S.A., Lang, D., Zimmer, A.D. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

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

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**, 516–522.

Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, 215–225.

Stavrinides, A., Tatsis, E., Foureau, E., Caputi, L., Kellner, F., Courdavault, V. and O'Connor, S.E. (2015) Unlocking the diversity of alkaloids in *Catharanthus roseus:* nuclear localization suggests metabolic channeling in secondary metabolism. *Chem. Biol.* **22**, 336–341.

Swaminathan, S., Morrone, D., Wang, Q., Fulton, D.B. and Peters, R.J. (2009) CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell*, **21**, 3315–3325.

Takos, A.M. and Rook, F. (2012) Why biosynthetic genes for chemical defense compounds cluster. *Trends Plant Sci.* **17**, 383–388.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729.

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

The Potato Genome Sequence Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

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., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.

Van Moerkercke, A., Fabris, M., Pollier, J., Baart, G.J., Rombauts, S., Hasnain, G., Rischer, H., Memelink, J., Oksman-Caldentey, K.M. and Goossens, A. (2013) CathaCyc, a metabolic pathway database built from *Catharanthus roseus* RNA-Seq data. *Plant Cell Physiol.* **54**, 673–685.

Vazques-Flota, F.A. and De Luca, V. (1998) Jasmonate modulates development and light regulated alkaloid biosynthesis in *Catharanthus roseus*. *Phytochemistry*, **49**, 395–402.

Verma, M., Ghangal, R., Sharma, R., Sinha, A.K. and Jain, M. (2014) Transcriptome analysis of *Catharanthus roseus* for gene discovery and expression profiling. *PLoS One*, **9**, e103583.

Winzer, T., Gazda, V., He, Z. *et al.* (2012) A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, **336**, 1704–1708.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214.

