Genome-guided investigation of plant natural product biosynthesis

Franziska Kellner1,1, Jeongwoon Kim2,†, Bernardo J. Clavijo3, John P. Hamilton2, Kevin L. Childs2, Brieanne Vaillancourt2, Jason Cepela2, Marc Habermann2, Burkhard Steuernagel2, Leah Clissold2, Kirsten McLay3, Carol Robin Buell2,* and Sarah E. O'Connor1,*

1Department of Biological Chemistry, The John Innes Centre, Norwich NR4 7UH, UK, 2Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA, 3The Genome Analysis Centre, Norwich, NR4 7UH, UK, and 4The 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
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) (Guimarães 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 generic 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 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-
ably of 523 Mb with an N50 scaffold size of 26.2 kb, 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 kb (Table S1). K-mer spectra representation analysis (http://www.tgac.ac.uk/tools-resources/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 kb, SGD is the only examined biosynthetic gene 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
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 (log2 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., 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 secologamn 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.
information, which would extend the length of contiguous whole-genome assembly, a BAC containing SGD also be inferred about the genomic context of SGD from the P450 enzymes (Im and Waskell, 2011). While nothing could appear to have an obvious pathway role, although cytochrome b5 enzymes are known to enhance the activity of alkaloid portions of the pathway.

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 (tetrhydroalstonine 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 tetrhydroalstonine, 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 tetrhydroalstonine 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 secoliganin 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 secoliganin synthase (Immler 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.
with previous reports (Menke et al., 1999; van der Fits and Memelink, 2000), both 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 000609 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
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 F2 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 \textit{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 \textit{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 (388-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 \textit{abys} (Simpson et al., 2009) using a k-mer size of 71. In total, the assembly consists of 79 302 scaffolds \textgreater 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 \textit{C. roseus} ESTs using \\textit{snap} (2014-05-30 v2) and by the identification of 248 conserved eukaryotic genes using \textit{cegma} v2.4 (Parra et al., 2007).

To assess the quality and representation of the \textit{C. roseus} genome in the assembly, individual Illumina reads were adapter trimmed and quality filtered with \textit{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 \textit{bwa-mem} (http://archiv.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 \textit{picard markduplicates} (http://broadinstitute.github.io/picard/; v.1.106). Alignments surrounding putative indels were refined using the IndelRealigner tool from the \textit{gatk} package (v3.3.0) (DePristo et al., 2011). Single nucleotide polymorphisms (SNPs) were called using the \textit{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 \textit{vcftools} (v0.1.12b) (Danecsek et al., 2011). The remaining SNPs were hard filtered for quality and maximum and minimum depth using \textit{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 \textit{C. roseus} Illumina genomic DNA reads were aligned to the genome assembly using \textit{bwa-mem} (v.0.7.8) and duplicate read alignments were marked using \textit{picard markduplicates} (v.1.86). Alignments around insertions/deletions were refined using the \textit{gatk indelrealigner} (v.2.8.1) and SNPs were called using \textit{samtools mpileup} (Li et al., 2009) and converted into VCF format using \textit{vcftools} (v.0.1.19). SNPs were filtered using \textit{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 \textit{maker} annotation pipeline (release 1103) (Holt and Yandell, 2011; Campbell et al., 2014). The \textit{C. roseus} genome was masked by \textit{repeatmasker} using the RepBase repeat library (www.repeatmasker.org) (Jurka et al., 2005). An initial HMM for the \textit{ab initio} gene prediction program, \textit{snap} (Korf, 2004), was trained using \textit{maker}-generated alignments of \textit{C. roseus} Sanger-derived ESTs downloaded from GenBank and \textit{C. roseus} transcript assemblies (Gongora-Castillo et al., 2012) to the masked \textit{C. roseus} genome sequence. \textit{Maker} gene predictions were generated using the initial SNAP HMM with UniProt SwissProt plant protein alignments and \textit{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. \textit{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 \textit{maker} run was performed using \textit{C. roseus} EST and transcript assembly alignments, UniProt SwissProt plant protein alignments, and \textit{A. thaliana} TAIR10 protein alignments as evidence. SNAP and Augustus were used for gene predictions using the \textit{C. roseus} trained HMMs, and FGENESH was also used for gene predictions using the tomato gene matrix (Salamov and Solovyov, 2000). For the final \textit{maker} run, the single exon EST alignments \textgreater 250 nucleotides were allowed as evidence. Additionally, \textit{Maker} was set to limit the use of single exon ESTs when generating final gene models, and \textit{Maker} was allowed to output unsupported \textit{(ab initio)} gene models. The hmmscan tool within the \textit{hmmers} package (Eddy, 2011) was used to identify \textit{maker}-predicted genes containing Pfam protein domains (Finn et al., 2014). Using only scaffolds \textgreater 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 \textit{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 \textit{C. roseus}.

**Bacterial artificial chromosome library construction, screening, and sequencing**

A BAC library was constructed from ‘SunStorm™ Apricot’ using the pINDIGOBAC-5 vector in \textit{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

© 2015 The Authors

specific primers for SGD and T16H2 were designed (Table S6) and BAC pools were screened by Bio S&T. Positive BACs were finger-printed 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 strain K12, sub-strand DH10B (CP000948.1). Only reads with a BLAST 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 adapters. Each read was aligned to the vector sequence (EU140754.1) and the genome sequence of E. coli strain K12, sub-strand DH10B (CP000948.1). Only reads with a BLAST 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 adapters. Each read was aligned to the vector sequence (EU140754.1) and the genome sequence of E. coli strain K12, sub-strand DH10B (CP000948.1). Only reads with a BLAST 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.

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 BLAST (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 MEAGB (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 ABC transporters and genes containing PF00005 or PF00664 as MATE 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 PRJNA2526811. The assembled genome (C. roseus SunStorm™ Apricot v1.0) has been deposited in the NCBI Whole-Genome Shotgun Sequence database under the accession JH1Z00000000. The version described in this paper is version JH1Z01000000. BACS containing the T16H1, T16H2, and 160MT cluster and SGD have been deposited in NCBI WGS under accession numbers ERP006960 and PRJEB27256, 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-1128898) 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


SUPPORTING INFORMATION

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


Genome biology of specialized metabolism


