Journal of Biotechnology 166 (2013) 122–134

Contents lists available at SciVerse ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest

Mei Xiao[a,1], Ye Zhang[a,1], Xue Chen[c,1], Eun-Jeong Lee[c,1], Carla J.S. Barber[b], Romit Chakrabarty[c], Isabel Desgagné-Penix[c], Tegan M. Haslam[c], Yeon-Bok Kim[b], Enwu Liu[b], Gillian MacNevin[c], Sayaka Masada-Atsumi[d], Darwin W. Reed[b], Jake M. Stout[b], Philipp Zerbe[e], Yansheng Zhang[f], Joerg Bohlmann[e], Patrick S. Covello[b], Vincenzo De Luca[d], Jonathan E. Page[b], Dae-Kyun Ro[c], Vincent J.J. Martin[g], Peter J. Facchini[c,*], Christoph W. Sensen[a]

[a] Department of Biochemistry and Molecular Biology, University of Calgary, 3330 Hospital Drive NW, Calgary, Alberta T2N 4N1, Canada
[b] National Research Council of Canada, 110 Gymnasium Place, Saskatoon, Saskatchewan S7N 0W9, Canada
[c] Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada
[d] Department of Biological Sciences and Centre for Biotechnology, Brock University, 500 Glenridge Avenue, St. Catharines, Ontario L2S 3A1, Canada
[e] Michael Smith Laboratories, University of British Columbia, 301-2185 East Mall, Vancouver, British Columbia V6T 1Z4, Canada
[f] CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, The Chinese Academy of Science, 430074 Wuhan, China
[g] Department of Biology, Concordia University, 7141 Rue Sherbrooke West, Montréal, Québec H4B 1R6, Canada

ARTICLE INFO

Article history:
Received 21 December 2012
Received in revised form 4 April 2013
Accepted 5 April 2013
Available online 16 April 2013

Keywords:
Roche-454 pyrosequencing
Bioinformatics
Illumina GA sequencing
Plant specialized metabolites
RNA-seq
Transcriptomics

ABSTRACT

Plants produce a vast array of specialized metabolites, many of which are used as pharmaceuticals, flavors, fragrances, and other high-value fine chemicals. However, most of these compounds occur in non-model plants for which genomic sequence information is not yet available. The production of a large amount of nucleotide sequence data using next-generation technologies is now relatively fast and cost-effective, especially when using the latest Roche-454 and Illumina sequencers with enhanced base-calling accuracy. To investigate specialized metabolite biosynthesis in non-model plants we have established a data-mining framework, employing next-generation sequencing and computational algorithms, to construct and analyze the transcriptomes of 75 non-model plants that produce compounds of interest for biotechnological applications. After sequence assembly an extensive annotation approach was applied to assign functional information to over 800,000 putative transcripts. The annotation is based on direct searches against public databases, including RefSeq and InterPro. Gene Ontology (GO), Enzyme Commission (EC) annotations and associated Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps are also collected. As a proof-of-concept, the selection of biosynthetic gene candidates associated with six specialized metabolic pathways is described. A web-based BLAST server has been established to allow public access to assembled transcriptome databases for all 75 plant species of the PhytoMetaSyn Project (www.phytometasyn.ca).

1. Introduction

Plant specialized metabolites have long been exploited through their use as flavors, pigments, medicines and industrial raw materials (Oksman-Caldentey and Saito, 2005; Fabricant and Farnsworth, 2001). Different from the role of primary metabolites in basic life functions, such as plant growth and development, specialized metabolites are mostly involved in mediating the interactions of plants with their environment, including the attraction of pollinators and defense against pathogens. These specialized compounds are characterized by an enormous diversity of chemical

* Corresponding author. Tel.: +1 403 220 7651; fax: +1 403 289 9311.
E-mail address: pfacchin@ucalgary.ca (P.J. Facchini).
[1] These authors contributed equally.

0168-1656
http://dx.doi.org/10.1016/j.jbiotec.2013.04.004

structures and can be categorized into several major groups based on their biosynthesis: polyketides, terpenes (isoprenoids), alkaloids, phenylpropanoids and flavonoids (Oksman-Caldentey and Inzé, 2004). Generally, each of these categories contains thousands of known compounds with many more awaiting discoveries.

Among the tens of thousands of plant specialized metabolites, many display potent biological activities and have been used extensively as pharmaceuticals (Rates, 2001). The *de novo* chemical synthesis of many of these metabolites has had limited success due to the typical occurrence of chiral centers; thus, naturally occurring and semi-synthetic compounds remain the main sources for commercial pharmaceutical applications. Nevertheless, the accumulation of many specialized metabolites in plants is low and depends on physiological, developmental and environmental factors (Oksman-Caldentey and Inzé, 2004). Access to such compounds is often inadequate and a reliance on the production of metabolites from naturally growing plants is not always sustainable. Metabolic engineering approaches have been used to increase specialized metabolite levels in plants (Dixon, 2005). However, it is often difficult to obtain desired compounds owing to the complexity of metabolic pathways and their regulation. Recently, plant biosynthetic pathways have been assembled in engineered microbial systems to produce targeted chemical compounds. For example, yeast has been engineered to produce a key precursor for the production of artemisinin (Ro et al., 2006) and a pathway intermediate leading to compounds such as morphine and codeine (Hawkins and Smolke, 2008). Although plant based metabolic engineering remains promising (De Luca et al., 2012), microbial production has several advantages over plant-based methods including (i) the relative ease of purifying target molecules using well established fermentation systems, (ii) the rapid growth rate of microorganisms compared with plants, and (iii) the improved optimization potential of microbial platforms using molecular, genetic and process engineering approaches.

Understanding the biosynthetic pathways is fundamental for the commercial production of specialized metabolites using these alternative approaches. Specialized plant metabolites often have long and complex biosynthetic pathways and it is generally challenging to identify all of the enzymes that catalyze the numerous metabolic transformations (Oksman-Caldentey and Inzé, 2004). Enzymes of plant specialized metabolic pathways are often encoded by large gene families and, generally, the specific functions of individual genes cannot be predicted strictly based on sequence analysis (Keeling and Bohlmann, 2006; Schuler and Werck-Reichhart, 2003; Nelson and Werck-Reichhart, 2011). The discovery of biosynthetic genes involved in plant specialized metabolism represents a unique challenge owing to the organization of many pathways as complex enzymatic networks producing several products, rather than simple linear schemes leading to a single compound (Hall et al., 2013). Moreover, most valuable specialized metabolites are derived from non-model plants, most of which have limited genomics resources (Fields and Johnston, 2005). A data-mining framework that integrates metabolomics, bioinformatics and functional genomics is essential to efficiently investigate specialized metabolite pathways in non-model plants.

Transcriptomics data mining is an efficient way to discover genes or gene families encoding enzymes involved in various metabolic pathways. High-throughput next-generation sequencing (NGS) technologies have revolutionized transcriptomics especially with the advent of RNA-sequencing (RNA-seq). This technology can be used to obtain RNA sequences on a massive scale with enormous sequencing depth. Despite these advantages, the sequence reads obtained from NGS platforms, such as Illumina, SOLiD and Roche-454, are often short (35–500 bp) compared with traditional Sanger sequencing (>700 bp) (Metzker, 2010). Correspondingly, a transcriptome must be reconstructed from raw reads using sequence assembly tools based on a reference genome, *de novo* assembly, or methods that combine both strategies (Martin and Wang, 2011).

RNA-seq has been applied to hundreds of non-model plants (Schliesky et al., 2012; Johnson et al., 2012). However, more comprehensive coverage of selected plant species is required to better understand the biosynthesis of specific specialized metabolites. The PhytoMetaSyn Project (www.phytometasyn.ca) has targeted 75 non-model plants that produce natural products belonging to three general categories: terpenoids, alkaloids and polyketides (Supplementary Table 1) (Facchini et al., 2012). Six subgroups (*i.e.* sesquiterpenes, diterpenes, triterpenes, monoterpenoid indole alkaloids, benzylisoquinoline alkaloids and polyketides) are the focus of efforts to identify novel biosynthetic genes responsible for the diversity of compounds produced in these 75 species. In this paper, the Roche-454 and Illumina GA NGS platforms (Suzuki et al., 2011) were used to sequence selected cDNA libraries. As a proof of concept and to compare the output transcriptome databases from two NGS technologies, we report the sequencing results of the first twenty plant species (Table 1). The bioinformatics pipeline developed to assemble and annotate the sequences is also described in detail.

2. Materials and methods

2.1. RNA extraction

The Trizol method was used to extract total RNA from plant organs and tissues (Chomczynski and Sacchi, 1987). When the polysaccharide and polyphenolic content was high, such as in roots or rhizomes, a modified CTAB method was used (Desgagne-Penix et al., 2010). The quality and quantity of isolated total RNA were evaluated on the basis of UV absorption ratios (*i.e.* 260/280 nm and 260/230 nm). All the samples showed a 260/280 nm ratio of between 1.9 and 2.1, and a 260/230 nm ratio in the range of 2.0–2.5. The 75 plant species used in this study are listed in Supplementary Table 2.

2.2. Poly(A)+ RNA purification, cDNA library preparation and next-generation sequencing

Poly(A)+ RNA purification, cDNA library preparation, emulsion-based PCR (emPCR) and sequencing was performed at the McGill University and Génome Québec Innovation Center (Montréal, Canada). The RNA content in all samples was quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific). RNA samples were further analyzed using an RNA 6000 Nano chip on a BioAnalyzer 2100 (Agilent Technologies) to validate RNA quality. Only samples with a BioAnalyzer RNA Integrity Number (RIN) of 7.5 or greater were used for sequencing. Poly(A)$^+$ RNA was purified from 20 to 40 μg of total RNA by two rounds of selection using oligo (dT) attached to magnetic beads and a Dynabeads mRNA Purification kit (Invitrogen). The cDNA libraries for Roche-454 pyrosequencing were constructed from 200 ng of mRNA using a cDNA Rapid Library kit (Roche) and subsequently amplified by emPCR as per the manufacturer's instructions. After amplification, the DNA carrying beads for each library were loaded onto one-half of a PicoTiterPlate and subjected to Roche-454 GS-FLX Titanium pyrosequencing. Image and signal processing of the raw output data was performed using GS Run Processor. Sequence reads with high-quality scores were written into Standard Flowgram Format (SFF) files.

The cDNA libraries for Illumina GA sequencing were constructed from 10 μg of total RNA using the TruSeq Stranded mRNA Sample

Table 1
(a) Sequencing outputs using the 454 GS-FLX Titanium platform to establish reference transcriptomes for 20 non-model plant species. (b) Sequencing outputs using the Illumina GA platform to establish reference transcriptomes for 20 non-model plant species.

(a)

No.	Plant	Tissue	Roche GS-FLX Titanium							
			SRA accession	Number of raw reads	Number of cleaned reads	Average read length (bp)	Average transcript read depth (reads/bp)	Unigenes	Predicted full-length CDS	Intersects between 454 and Illumina predicted full length CDS
1	*Amsonia hubrichtii*	Leaf	SRS160815	715,432	627,544	330	8.4	30,348	13,857	10,955
2	*Arnica montana*	Leaf	SRS212543	739,124	660,971	430	8.2	36,819	19,850	15,751
3	*Centella asiatica*	Roots	SRS212540	712,683	626,382	418	10.3	26,212	15,520	13,078
4	*Chelidonium majus*	Stem	SRS150402	504,849	408,742	347	8.0	23,678	10,312	7586
5	*Catharanthus roseus*	Leaf	SRS160818	653,830	588,162	328	9.6	26,804	12,342	6432
6	*Dianthus superbus*	Leaf	SRS212539	752,802	659,677	409	9.4	32,649	17,568	13,251
7	*Eschscholzia californica*	Root	SRS160813	472,167	423,743	428	6.9	32,150	17,385	12,911
8	*Glaucium flavum*	Root	SRS212395	648,604	540,433	396	9.5	26,520	12,084	8199
9	*Hydrastis canadensis*	Rhizome	SRS212407	685,239	440,083	254	5.9	23,809	10,801	7617
10	*Hypericum perforatum*	Gland	SRS212625	732,402	528,871	400	7.5	36,542	24,715	18,257
11	*Lippia dulcis*	Leaf	SRS165538	1,261,768	1,071,907	451	8.2	81,081	31,606	11,817
12	*Matricaria recutita*	Flower	SRS212542	735,449	602,979	410	6.1	44,324	22,596	18,464
13	*Papaver bracteatum*	Stem	SRS160814	595,176	528,920	380	5.6	46,224	18,879	14,042
14	*Platanus occidentalis*	Leaf	SRS160811	599,373	434,169	343	9.9	25,457	11,552	7920
15	*Pseudolarix amabilis*	Root	SRS212537	733,434	621,506	380	7.8	35,710	15,033	10,362
16	*Sanguinaria canadensis*	Rhizome	SRS212403	653,689	571,822	417	10.1	25,652	11,787	9514
17	*Silene vulgaris*	Root	SRS160810	777,155	674,050	426	7.2	42,603	23,283	17,961
18	*Tabernaemontana elegans*	Leaf	SRS160816	708,517	622,051	379	8.8	28,744	14,646	12,082
19	*Valeriana officinalis*	Root	SRS165540	568,993	497,316	569	6.7	55,093	22,620	14,601
20	*Xanthium strumarium*	Leaf	SRS212544	593,315	525,997	413	8.4	28,589	14,751	8027
	Average			692,200	582,766	395	8.1	35,450	17,059	11,941

(b)

No.	Plant	Tissue	Illumina genome analyzer					
			SRA accession	Number of raw reads	Number of cleaned reads	Average transcript read depth (reads/bp)	Unigenes	Predicted full-length CDS
1	*Amsonia hubrichtii*	Leaf	SRS260004	75,544,560	65,984,652	32.1	48,557	48,557
2	*Arnica montana*	Leaf	SRS260895	67,378,080	51,890,730	36.8	54,401	54,401
3	*Centella asiatica*	Roots	SRS260019	77,853,840	70,505,918	70.1	37,852	37,852
4	*Chelidonium majus*	Stem	SRS259976	62,362,800	50,819,736	32.3	45,005	45,005
5	*Catharanthus roseus*	Leaf	SRS260010	77,661,840	65,442,826	34.3	38,260	38,260
6	*Dianthus superbus*	Leaf	SRS260015	68,008,560	61,358,556	58.4	37,887	37,887
7	*Eschscholzia californica*	Root	SRS259979	62,704,080	53,746,798	37.3	42,167	42,167
8	*Glaucium flavum*	Root	SRS259986	60,410,640	38,697,818	63.2	31,100	31,100
9	*Hydrastis canadensis*	Rhizome	SRS259993	71,077,680	61,254,386	87.8	33,335	33,335
10	*Hypericum perforatum*	Gland	SRS260013	64,844,040	55,545,066	27.7	47,054	47,054
11	*Lippia dulcis*	Leaf	SRS260896	59,685,424	53,617,274	33.0	41,220	41,220
12	*Matricaria recutita*	Flower	SRS260903	64,608,720	54,992,498	37.3	59,718	59,718
13	*Papaver bracteatum*	Stem	SRS259987	69,721,200	57,768,096	36.1	70,428	70,428
14	*Platanus occidentalis*	Leaf	SRS260016	74,883,840	65,746,902	23.3	66,228	66,228
15	*Pseudolarix amabilis*	Root	SRS260867	73,600,604	66,210,456	41.5	39,674	39,674
16	*Sanguinaria canadensis*	Rhizome	SRS259992	71,713,920	59,322,808	23.3	53,019	53,019
17	*Silene vulgaris*	Root	SRS260018	72,629,520	65,589,906	41.1	63,945	63,945
18	*Tabernaemontana elegans*	Leaf	SRS260005	75,544,560	64,407,924	43.4	52,822	52,822
19	*Valeriana officinalis*	Root	SRS260897	64,131,050	56,865,984	70.4	47,998	47,998
20	*Xanthium strumarium*	Leaf	SRS260904	71,174,880	54,773,006	38.9	35,182	35,182
	Average			69,276,992	58,727,067	43.4	47,293	47,293

Prep Kit (Illumina) according to the manufacturer's instructions. The quality and average length of cDNAs in each library were determined using a High Sensitivity DNA (Agilent Technologies) chip on a 2100 Bioanalyzer. For Illumina GA sequencing, 7 pmol of each library containing cDNA with lengths from 600 to 1200 base pairs (bp) were loaded into one lane of the flow cell to generate approximately 750,000 clusters per mm^2. The HCS 1.4 and CASAVA 1.6-1.8 software suites were used to obtain base calls and raw fastq reads.

2.3. Sequence quality control and cleaning

Quality scores and header information were extracted from SFF files generated from the 454 data. The pre-processing pipeline included several cleaning procedures, including clipping of adapter/primer sequences and window-based trimming of reads with Phred quality scores of less than 22. Low-complexity regions, including homopolymers were masked along with repeat regions identified based on similarity to the following: the RepBase14

database (Jurka et al., 2005), the Viridiplantae subset of NCBI reference sequences (Refseq), and the TIGR plant repeat database (Ouyang and Buell, 2004). Ribosomal RNA (rRNA) and ribosomal protein reference sets for related species were downloaded from NCBI and SILVA databases (Pruesse et al., 2007). Reads identified as rRNA and ribosomal protein sequences, and those shorter than 100 bp (not including masked regions) were removed from each 454 database. The Scylla component of the Paracel Filtering Package (PFP) (Paracel Inc, Pasadena, CA) was used to perform these steps.

Initial quality assessment for Illumina GA sequence data was based on FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) statistics, and Cutadapt (Martin and Wang, 2011) was used for adapter/primer trimming. Trimmed reads were further subjected to quality score conversion, trimming of reads with Phred quality scores of less than 25, and removal of read pairs with at least one member shorter than 35 bp using scripts written in-house. Reads were also trimmed at the 5′ end by 12–14 bp to account for bias associated with random priming. The rRNA and ribosomal protein-encoding sequence content was monitored by mapping the raw reads against the reference sets downloaded from NCBI and SILVA databases. No filtering step was applied if the composition was not substantial.

2.4. De novo transcriptome assembly

Assemblies of cleaned 454 sequence data were generated using MIRA (version 3.2) (Chevreux et al., 2004). The pre-processing functions in MIRA (Chevreux et al., 2004) were disabled and analyses were performed using the 'accurate' setting. Other *de novo* assemblers, Paracel Transcript Assembler (Paracel Inc), CLC assembly cell (CLC bio, Cambridge MA), and Newbler v2.3 (Margulies et al., 2005) were also evaluated by comparing assembly statistics. MIRA produced the largest number of contigs over 1000 bp using the highest proportion of 454 reads.

Short-read Illumina GA data were assembled using VelvetOases v0.1.16 (Zerbino and Birney, 2008). Optimal assembly of contigs in each database and representing a wide dynamic range of gene expression was acquired using a combined k-mer assembly approach. The pipeline generated eight parallel Oases runs for each paired-end read set using incremental k-mer settings increasing by units of 5 between 37 and 67. The generation of multiple copies of similar transcripts was expected among the different k-mer runs when merging the eight assemblies. The clustering tool CD-HIT-EST (Li and Godzik, 2006) was used to reduce redundancy by clustering nearly identical (greater than 99%) transcripts and extracting the longest representative within each cluster. The non-redundant pool of transcripts was further assembled using CAP3 (Huang and Madan, 1999) to combine contigs with significant overlaps (minimum 95% identity over at least 50 bp). Final assemblies were completed after contigs of less than 300 bp were discarded.

2.5. Functional annotation

Annotation of the 20 assembled transcriptome datasets was performed using the Magpie Automated Genomics Project Investigation Environment (MAGPIE) (Gaasterland and Sensen, 1996). MAGPIE automates sequence similarity searches against major public and internal target databases. TimeLogic Decypher Biocomputing systems (http://www.timelogic.com) were used to significantly accelerate similarity searches. Specifically, the TimeLogic Tera-BLAST algorithm was used to compare transcripts to the NCBI databases NR (non-redundant) and the viridiplantae subset of RefSeq (Pruitt et al., 2007). An expected *e*-value of $1e{-}3$ and a minimum alignment length of 30 bp were used. To obtain motif-level information, accelerated Hidden Markov Model (HMM) searches were performed against local instances of Interpro HMM libraries at an *e*-value of $1e{-}10$. The NCBI Conserved Domain Database (CDD) was also queried using RPS-BLAST for additional annotation information.

To coordinate all search results for each contig, MAGPIE ranked individual hits into three tiers of confidence. For BLAST results, *e*-value cutoffs were set at $1e{-}35$, $1e{-}15$, and $1e{-}5$ for evidence levels 1, 2 and 3 respectively. For HMM results, *e*-value cutoffs were $1e{-}20$, $1e{-}14$ and $1e{-}10$ at percentage similarity requirements of 65, 45 and 25%, respectively. Putative functional descriptions were assigned to each contig by performing a weighted summary of search result annotations. Summaries were based on word frequency, lexical complexity and word length, in addition to the level and type of evidence, and the taxonomic distance between the subject and the query species. GO annotations and EC numbers were compiled from GIDs extracted from level-1 evidence and attached to individual contigs as additional functional information. Contigs were subsequently cross-mapped to one another based on common GO terms and level-1 evidence. As a step toward the incorporation of metabolomics data, putative transcript data was mapped to KEGG metabolic pathways based on EC numbers. A summary page holding all evidence and annotation was generated for each contig in MAGPIE.

2.6. Full-length coding sequence prediction

Annotated contigs were available for further analysis after the assembly and annotation of each sequence dataset. ESTScan (Lottaz et al., 2003) is a statistical hidden Markov model (HMM) program that can be used to discover patterns and was used for CDS detection. A refined HMM model was built using a set of full-length coding sequences for training. To determine the training set, the annotation evidence for each contig within an assembly was examined on all six open reading frames. The frame with the longest length of annotated sequence was scanned further. If the length of annotation within the selected frame was greater than 75% of the original contig length and contained putative start and stop codons, this frame was saved as a training set member. To ensure that the selected CDS has the maximum possible length, another scan was extended to flanking regions to search for possible start and stop codons.

After the double scanning was applied to every contig in an assembly, a set of putative full-length coding sequences was collected. This full-length coding sequence dataset was used to train ESTScan to build the HMM model (Iseli et al., 1999). After building the model, ESTScan was applied again to predict a putative CDS for every contig of the assembly. The CDS dataset from ESTScan could contain partial coding regions; thus, scanning of the original contig and annotation was repeated for every CDS generated. When both start and stop codons were found in the original contig and the annotation was longer than 75% of the original contig length, this putative CDS was retained as a full-length putative CDS. In contrast, partial putative coding regions were removed. The full-length putative CDS dataset was then combined with the full-length coding region dataset used to train the HMM model. Duplicated sequences were removed to generate the final predicted CDS dataset (Supplementary Fig. 1). To conservatively estimate the intersect between predicted full-length CDS sets generated by 454 and Illumina, Mega BLAST (Zhang et al., 2000) was used to compare respective sets at an *e*-value cutoff of 0.

2.7. Gene expression analysis

Gene expression levels were determined by quantifying the observed read abundance. Raw read counts were extracted from assembly files for each contig in the case of the 454 assemblies.

For Illumina GA data, counts were estimated by re-mapping raw short reads to the assembled contigs using Bowtie (Langmead et al., 2009). The RNA-Seq by Expectation-Maximization (RSEM) package (Li and Dewey, 2011) was used to resolve ambiguous mappings and to perform final quantifications. Only paired-end reads that mapped to a common contig were considered. Normalization was done by calculating FPKM values (Fragments Per Kilobase of exon model per Million mapped reads) for each contig.

2.8. Phylogenetic analysis

Amino acid alignments were performed using ClustalW and used to construct the phylogenetic trees by the neighbor-joining method.

3. Results

3.1. Plant species selection

The strategy for building NGS sequencing platforms for 75 non-model species is shown in Fig. 1. Plant species were selected on the basis of: (i) the availability of plant tissues, (ii) the commercial value of key metabolites, (iii) an inherent understanding of the biochemical diversity across related species, and (iv) the availability of biochemical resources responsible for the production of specific compounds. The selected species represent 31 plant families and produce compounds that include benzylisoquinoline alkaloids, monoterpenoid indole alkaloids, polyketides, and various terpenoids (Supplementary Table 1). Herein, we report detailed results from NGS sequencing analysis of 20 species. Assembled transcriptome databases for all 75 species can be found on the PhytoMetaSyn Project website (www.phytometasyn.ca) (Facchini et al., 2012). Of the 20 species selected here, nine plants were targeted as a source of alkaloids, 10 as producers of sesqui-, di- or tri-terpenoids, and one owing to the accumulation of polyketides. Targeted compound classes for each species are listed in Supplementary Table 2.

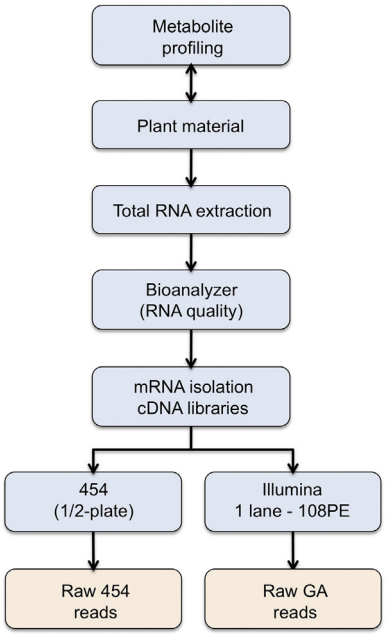

Fig. 1. Overall strategy for plant sample selection, preparation, and 454 and Illumina GA sequencing.

3.2. Assembly and read depth

The overall strategy of the bioinformatics pipeline used for (i) processing the 454 and Illumina GA raw reads, (ii) *de novo* assembly of raw reads into unigenes, and (iii) database construction is shown in Fig. 2. After the quality control procedures, the remaining raw 454 sequence data were used to assemble contigs. Across the 20 selected plant species, the 454 assemblies yielded datasets ranging from 23×10^3 to 55×10^3 unigenes and, on average, each assembly generated $35 \times 10^3 \pm 13 \times 10^3$ (standard deviation) independent contigs. The Illumina GA assembly output was between 31×10^3 to 70×10^3 unigenes and on average, the Velvet/Oases assembly pipeline yielded $47 \times 10^3 \pm 11 \times 10^3$ independent contigs. Details for each species are provided in Table 1. The number of sequencing reads obtained was primarily a function of the sequencing platform. Sequencing on the Illumina GA platform generally produced 100-fold more reads than the Roche-454 pyrosequencing and, consequently, Illumina GA-derived assemblies generated more contigs than 454-based assemblies (Luo et al., 2012). The average read depth was at least five-fold higher for Illumina GA transcript assemblies (Fig. 1a and b).

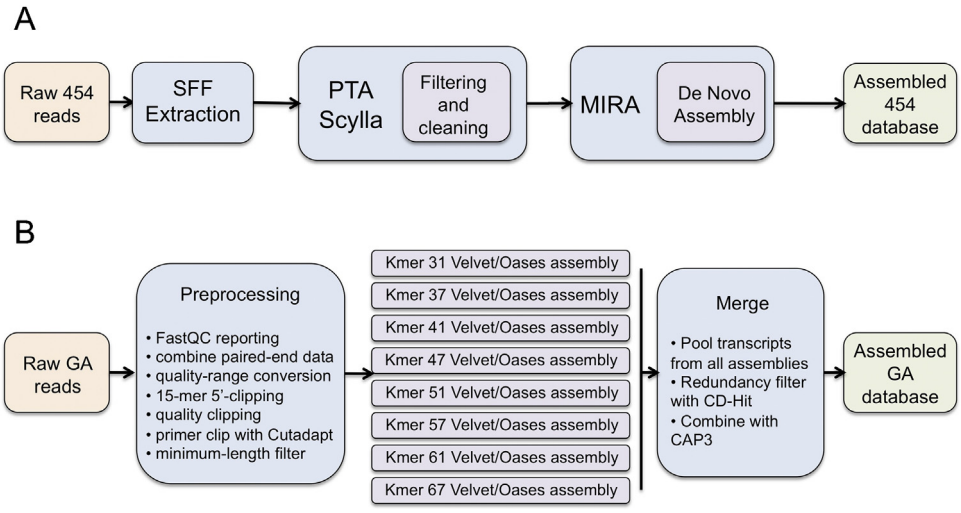

Fig. 2. Strategy of the bioinformatics pipeline used for the processing of raw reads, assembly of raw reads into unigenes, and construction of databases for (A) 454 and (B) Illumina GA (B) sequences.

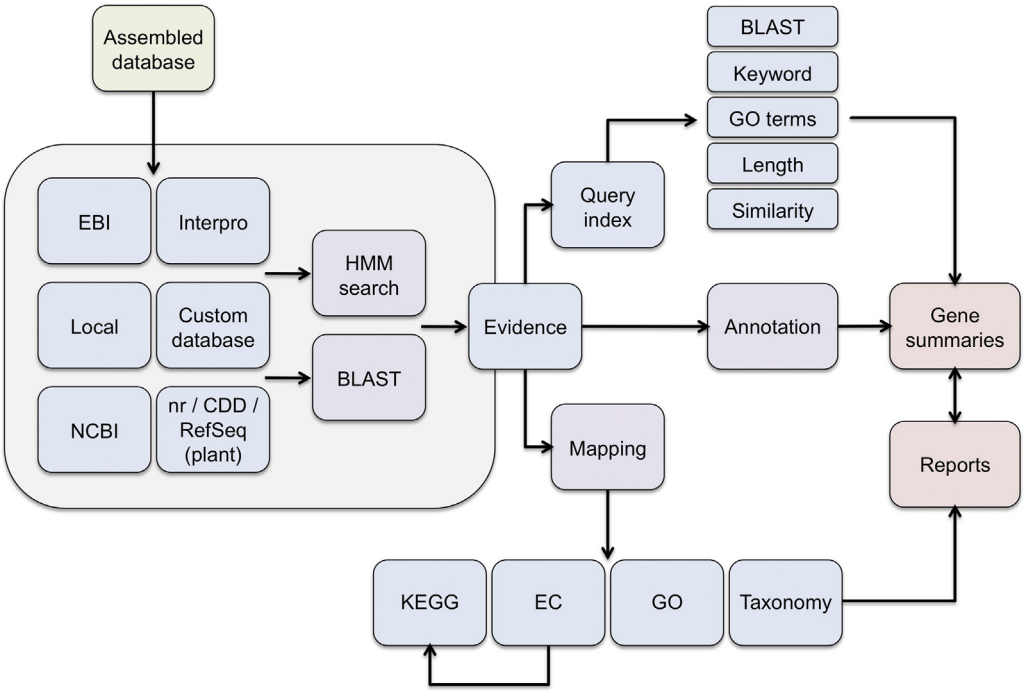

Fig. 3. Annotation and gene summary reports of unigenes assembled from 454 and Illumina GA sequence databases.

3.3. Annotation of transcripts and functional assignment

The pipeline to generate annotation and summary reports for unigenes derived from 454 and Illumina GA sequencing is shown in Fig. 3. For visualization purposes, similarity search results are assigned one of three confidence levels based on the statistical values for each contig. The total number of unigenes for each of the 20 species that (i) annotated with any level of confidence, (ii) annotated with high-confidence (level 1), (iii) mapped to GO terms, and (iv) associated with EC numbers are summarized in Table 2. For assemblies generated from 454 sequences, the average numbers of unique contigs associated with any annotation evidence, and with high-level annotation evidence, were 28×10^3 and 21×10^3 representing 82 and 60%, respectively, of the mean total number of unigenes. Although the average number of 454-generated unique contigs associated with GO terms was 28×10^3, the average number of unigenes assigned an EC number was only 3854 representing 80 and 10%, respectively, of the mean total number of unigenes. For assemblies generated from Illumina GA sequences, the average numbers of unique contigs associated with any annotation evidence, and with high-level annotation evidence, were 42×10^3 and 32×10^3 representing 89 and 68%, respectively, of the mean total number of unigenes. Although the average number of Illumina GA-generated unique contigs associated with GO terms was 41×10^3, the average number of unigenes assigned an EC number only 5835, which represent 88 and 12%, respectively, of the mean total number of unigenes. Annotation details are available through MAGPIE, which allows the query of databases according to different criteria including keyword, GO term, sequence similarity and unigene identification (Supplementary Figs. 2–4).

3.4. Full-length coding sequence prediction

Owing to the high proportion of non-coding sequences (*e.g.* introns and intergeneric regions) in plant genomes, the accurate prediction of a CDS directly from genome sequence assemblies is often difficult (Furuno et al., 2003). Transcriptome analysis provides an advantage in this regard because the sequence data represents expressed genes. The assembled and annotated databases were searched to retrieve full-length coding sequences based on the evidence collected in the gene summaries (Supplementary Fig. 5). The candidates were used to train the Markov model-based program for ESTScan reference which was then used to scan the assembled unigenes for full-length CDS candidates. On average, more full-length unigenes were retrieved from assemblies generated using Illumina GA sequences (63.2%) compared with 454 databases (48.1%) (Tables 1 and 2). The occurrence of common full-length CDS in the 454 and Illumina GA datasets, predicted using strict Mega BLAST comparisons, was approxiamtely 74.5 and 26.9%, respectively (Table 1).

3.5. Biosynthetic gene representation in six specialized metabolic pathways

Deep transcriptome databases allow the establishment of a transcriptional profile for selected non-model plants and facilitate the discovery of novel biosynthetic genes responsible for the production of unique compounds of various species. Examples of the application of NGS to the elucidation of various plant specialized metabolic pathways are provided below.

Sequiterpenes – Sesquiterpenoids are 15-carbon (C15) compounds, many of which are pharmacologically and biologically active. Sesquiterpene synthases (sesqui-TPSs; Chen et al., 2011) catalyze the conversion of farnesyl diphosphate (FPP) to several structurally diverse C15 hydrocarbons that can undergo further modification (*e.g.*, oxidation, reduction and methylation) yielding a vast array of specialized metabolites. The biosynthesis of specific sesquiterpenes in non-model plants is limited owing partially to the low abundance of many compounds. The combined application of NGS, metabolically engineered microbes and traditional enzyme purification methods can expedite the isolation and characterization of sesqui-TPSs. For example, the valerian plant (*Valeriana officinalis*) the Aztec sweet herb (*Lippia dulcis*) and *Thapsia garganica* synthesize the mild-sedative valerenic acid, the natural sweetener hernandulcin and the anti-cancer drug thapsigargin, respectively (Fig. 4). These pathways, which have not been

Table 2
Annotation summaries for the transcriptomes of the 20 non-model plant species generated using 454 GS-FLX Titanium and Illumina GA sequencing.

No.	Plant	Roche GS-FLX Titanium					Illumina genome analyzer				
		Unigenes	Overall annotated	High level annotated	GO annotated	EC number allocated	Unigenes	Overall annotated	High level annotated	GO annotated	EC number allocated
1	*Amsonia hubrichtii*	30,348	25765	18526	25591	3052	48,557	45627	40661	45441	7017
2	*Arnica montana*	36,819	32263	26140	32029	5090	54,401	52721	39832	51355	7729
3	*Centella asiatica*	26,212	24395	20078	24287	3469	37,852	32981	28280	32712	5189
4	*Chelidonium majus*	23,678	19635	13977	19460	2368	45,005	42057	33449	41956	6092
5	*Catharanthus roseus*	26,804	21840	15430	21631	2571	38,260	38092	22543	37204	3454
6	*Dianthus superbus*	32,649	27381	20744	27128	3994	37,887	30911	23585	30525	4196
7	*Eschscholzia californica*	32,150	28430	21403	28194	4221	42,167	38332	32677	38063	6545
8	*Glaucium flavum*	26,520	20945	15645	20725	2719	31,100	31100	19669	31100	3231
9	*Hydrastis canadensis*	23,809	20443	15491	20230	2511	33,335	33335	20898	33335	3637
10	*Hypericum perforatum*	36,542	33494	26240	33245	5414	47,054	43278	35022	42943	6627
11	*Lippia dulcis*	81,081	56129	33787	55024	5067	41,220	33675	26131	33300	4537
12	*Matricaria recutita*	44,324	37842	30478	37525	6537	59,718	59718	44395	59718	8977
13	*Papaver bracteatum*	46,224	33168	24381	32767	4988	70,428	56463	37334	53039	6793
14	*Platanus occidentalis*	25,457	21454	15722	21287	3113	66,228	61037	51227	60602	10163
15	*Pseudolarix amabilis*	35,710	27101	19659	25813	3429	39,674	34844	27998	33375	4863
16	*Sanguinaria canadensis*	25,652	20493	15938	20301	2621	53,019	47247	40122	46890	7715
17	*Silene vulgaris*	42,603	35096	26181	34611	4688	63,945	49566	36970	48762	6343
18	*Taberneamontana elegans*	28,744	24918	18843	24758	3024	52,822	52822	40825	52822	7090
19	*Valeriana officinalis*	55,093	42647	28842	42175	4866	47,998	33241	24224	32842	3557
20	*Xanthium strumarium*	28,589	24810	19383	24574	3343	35,182	28134	19693	27776	2946
	Average	35,450	28,912	21,344	28,568	3854	47,293	42,259	32,277	41,688	5835

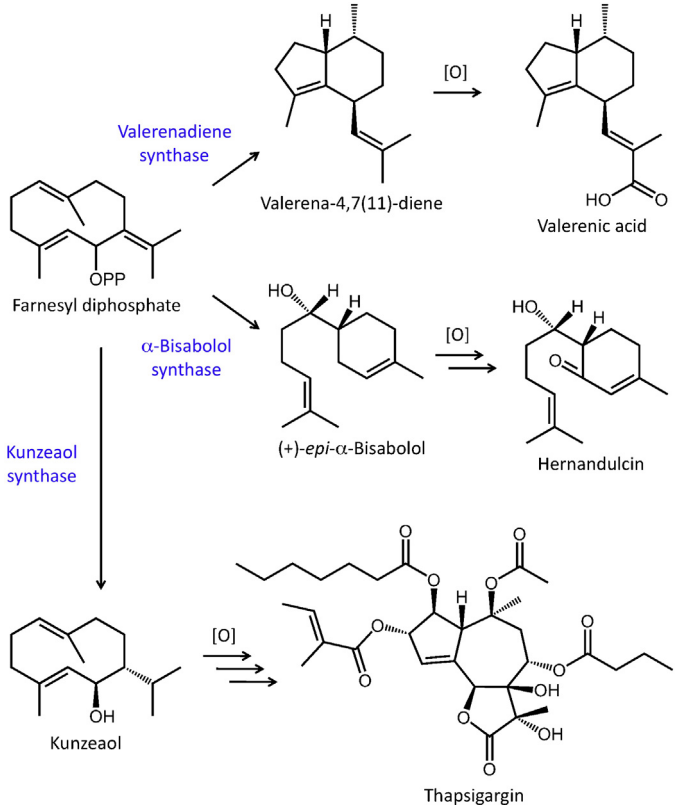

Fig. 4. Biosynthesis of the sesquiterpenes valerenic acid, thapsigargin and hernandulcin from farnesyl diphosphate (FPP). Multiple arrows indicate more than one enzymatic step. Several oxidative reactions [O] have not been characterized.

extensively studied, are now being successfully investigated using a gene isolation and characterization pipeline based on the NGS resources described herein (Pyle et al., 2012; Attia et al., 2012; Pickel et al., 2012). Nine full-length sesqui-TPS cDNAs were isolated from the assembled 454 and Illumina GA databases of these three plants. Subsequent expression of the full-length cDNAs in the FPP over-producing yeast, EPY300 (Ro et al., 2006, 2008) facilitated the *in vivo* characterization of new activities without costly and time-consuming *in vitro* analysis. Following purification of milligram quantities of each reaction product, the chemical identities of the novel sesquiterpenes were determined using various analytical techniques, including NMR, GC-MS and LC-MS. Similar approaches can be applied to expedite functional identification of other novel sesqui-TPSs and downstream enzymes that modify the sesquiterpene backbones.

Diterpenes – Diterpenes are among the largest classes of plant specialized metabolites with several thousand known structures, many of which are biologically active. The core structures of diterpenes are produced from geranylgeranyl diphosphate (GGPP) by diterpene synthases (di-TPSs) (Chen et al., 2011). In the PhytoMetaSyn Project, the gymnosperm golden larch (*Pseudolarix amabilis*) and several other species were selected based on the occurrence of specialized diterpenes. Integrated assemblies derived from root-specific cDNA libraries subjected to 454 and Illumina GA sequencing proved particularly advantageous, allowing the efficient identification of seven putative di-TPS candidates. The discovery of full-length sequences was ~40% higher when combining both the 454 and Illumina GA datasets, compared with the individual assemblies. Phylogenetic analysis supported the functional annotation of the retrieved di-TPS candidates (Fig. 5), all seven of which are members of the conifer-specific TPS-d3 family (Martin et al., 2004; Keeling et al., 2011; Chen et al., 2011). One

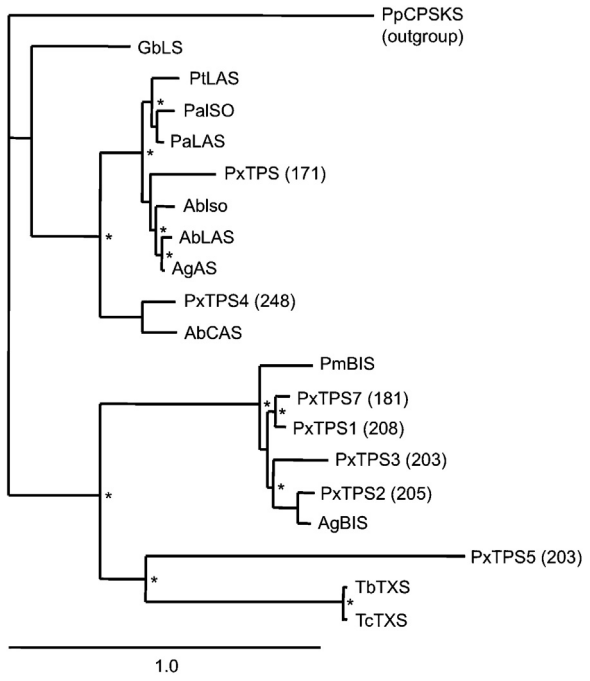

Fig. 5. Phylogenetic analysis of *Pseudolarix amabilis* diterpene synthase candidates. Maximum likelihood tree of di-TPS candidates of *P. amablis* within the TPS-d3 family rooted with *Physcomitrella patens ent*-copalyl diphosphate/*ent*-kaurene synthase (PpCPS/KS). Asterisks highlight bootstrap support of 85% and higher. Numbers in parentheses represent read counts in the 454 databases. Abbreviations and accession numbers: PxTPS, *Pseudolarix amablis* terpene synthase; PpCPS/KS, *Physcomitrella patens* copalyl diphosphate synthase/kaurene synthase (BAF61135); AbCAS, *Abies balsamea cis*-abienol synthase (JN254808); AbLAS, *A. balsamea* levopimaradiene/abietadiene synthase (JN254805); AbIso, *A. balsamea* isopimaradiene synthase (JN254805); PaIso, *Picea abies* isopimaradiene synthase (AY473620); PaLAS, *P. abies* levopimaradiene/abietadiene synthase (AAS47691); GbLS, *Gingko biloba* levopimaradiene synthase (AF331704); PtLAS, *Pinus taeda* levopimaradiene/abietadiene synthase (AY779541); AgAS, *Abies grandis* abietadiene synthase (U50768); PmBIS, *Pseudotsuga menziesii* y-bisabolene synthase (AY906868); AgBIS, *A. grandis* a-bisabolene synthase (AF006195); TbTXS, *Taxus brevifolia* taxadiene synthase (U48796); TcTXS, *T. cuspidata* taxadiene synthase (DQ305407).

of the putative *P. amabilis* enzymes (PxTPS6) clustered with known bifunctional class I/II di-TPSs of spruce (*Picea* spp.), fir (*Abies* spp.) and pine (*Pinus* spp.) diterpene resin acid biosynthesis (Keeling et al., 2011). One candidate di-TPS (PxTPS4) most closely grouped with *cis*-abienol synthase of balsam fir (*Abies balsamea*) (Zerbe et al., 2012). Four sequences (PxTPS1, PxTPS2, PxTPS3 and PxTPS7) were closely related to α-bisabolene synthases, which are sesqui-TPS with structural features similar to those of the three-domain di-TPSs. One candidate (*Px*TPS5) showed relatively high read counts in the 454 library indicating abundant transcript levels. *Px*TPS5 was also phylogenetically more distant from other candidates, but proximal to *Taxus* spp. taxadiene synthases (Fig. 5) suggesting that the enzyme possesses a distinct and potentially new function not previously identified in a di-TPS member of the pine family.

Triterpenes – Triterpenes display a wide range of interesting human health-related properties and many plants used in Asian herbal medicine feature bioactive triterpenes. The biosynthesis of the bioactive ursane triterpenes, asiatic acid and medicassic acid, in *Centella asiatica* are purportedly formed from 2,3-oxidosqualene *via* α-amyrin with subsequent oxidative functional group additions by cytochromes P450 at positions 2, 6, 23 and 28. In the PhytoMetaSyn Project, selection of appropriate cytochrome P450 candidates was based on assembled 454 and Illumina GS sequence databases of *C. asiatica* root cultures in which asiatic acid accumulation was induced more than 3-fold by methyl jasmonate treatment. An assembled Illumina GA sequence database was also prepared for control *C. asiatica* root cultures. Examples of cytochrome P450 candidates showing the highest induced expression levels, estimated by read counts and FPKM values, in response to jasmonate treatment are provided in Supplementary Table 3. The enzymatic function of the corresponding candidate enzymes is currently under investigation by heterologous expression of the corresponding cDNAs in yeast.

Monoterpenoid indole alkaloids (*MIAs*) – MIAs are a class of over 2000 known compounds, many of which possess important pharmacological activities. The 454 and Illumina GA databases of three species (*i.e. Catharanthus roseus*, *Tabernaemontana elegans* and *Amsonia hubrichtii*) were searched for genes putatively involved in MIA biosynthesis. Transcripts encoding geraniol 10-hydroxylase (G10H), 10-hydroxygeraniol oxidoreductase (10HGO), loganic acid *O*-methyltransferase (LAMT), secologanin synthase (SLS), tryptophan decarboxylase (TDC) and strictosidine synthase (STR) were identified in all three databases, whereas strictosidine β-D-glucosidase (SGD) with at least 70% amino acid sequence identity to the characterized enzyme form *C. roseus* could only be described in *C. roseus* and *T. elegans* (Supplementary Table 4). The *A. hubrichtii* databases contained a putative SGD with 57% amino acid sequence identity to the characterized *C. roseus* enzyme. The available datasets also suggested that transcripts encoding LAMT were more highly represented in young leaves of some species than in others. In contrast genes encoding 16-methoxy-2,3-dihydro-3-hydroxytabersonine *N*-methyltransferase (NMT), desacetoxyvindoline 4-hydroxylase (D4H) and deacetylvindoline acetyltransferase (DAT), which are uniquely involved in vindoline biosynthesis, were only represented within *Catharanthus roseus*, a result consistent with the unique occurrence of vindoline in the genus *Catharanthus.* The validity of the results obtained for known MIA biosynthetic genes (Supplementary Table 4) suggests that the same approach will identify candidate genes encoding novel enzymes involved in MIA metabolism.

Benzylisoquinoline alkaloids (*BIAs*) – more than 2500 BIAs, many of which possess potent pharmacological properties, have been identified in plants belonging mostly to the families Papaveraceae, Ranunculaceae, Berberidaceae and Menispermaceae (Ziegler and Facchini, 2008). Many of the enzymes involved in BIA biosynthesis have been identified from a limited number of plants including opium poppy (*Papaver somniferum*) and Japanese goldthread (*Coptis japonica*), yet the majority of catalysts responsible for the immense structural diversity of BIAs in other plants have not been characterized. Tapping into the vast biosynthetic potential of plants requires access to genes from a variety of species. The transcriptome databases from 20 BIA-producing species represent such a repository of unique biosynthetic genes responsible for the diverse alkaloid content of these plants. Based on the categorization of known BIA biosynthetic enzymes into discrete protein families (*e.g.* cytochromes P450, *O*- and *N*-methyltransferases, various NADPH-dependent reductases, FAD-linked oxidoreductases, acyl-CoA-dependent acetyltransferases and 2-oxoglutarate-dependent dioxygenases) numerous orthologous and paralogous candidate genes can be selected from these databases for functional characterization.

The utility of NGS-based transcriptome databases from related plant species for the identification of novel biosynthetic enzymes is shown by focusing on *N*-methylation as a common functional group modification in BIA metabolism. Three alkaloid type-specific *N*-methyltransferases (NMTs) have been already been characterized: coclaurine *N*-methyltransferase (CNMT), tetrahydroprotoberberine *N*-methyltransferase (TNMT) and pavine *N*-methyltransferase (PavNMT) (Fig. 6) (Liscombe et al., 2009). (+)-Magnoflorine is an antimicrobial alkaloid produced in several plant species *via* the *N*-methylation of (*S*)-corytuberine (Minami et al., 2008). Although the enzyme responsible for the formation of the quaternary

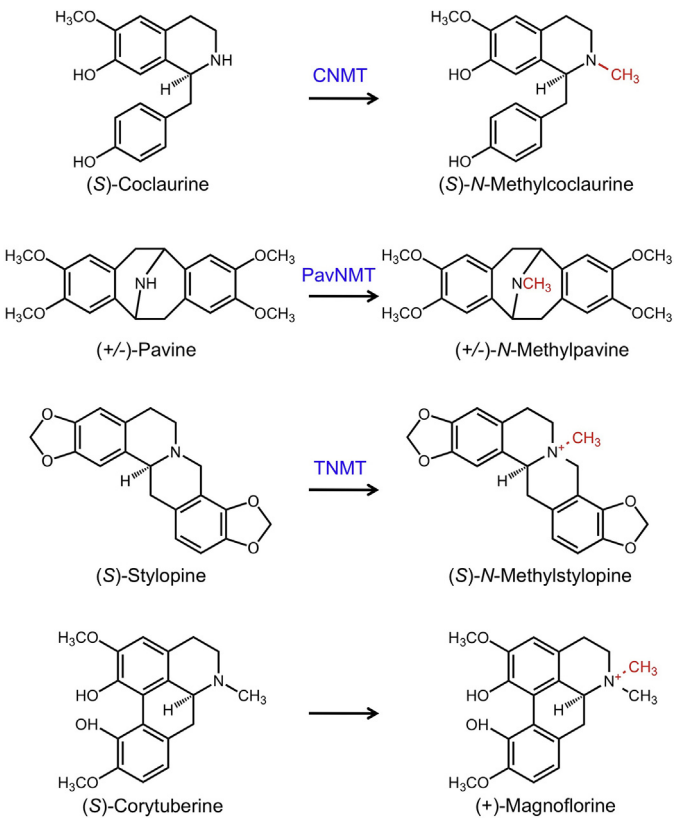

Fig. 6. *N*-Methylation reactions catalyzed by CNMT, PavNMT and TNMT, and the proposed NMT-catalyzed conversion of (*S*)-corytuberine to (+)-magnoflorine. CNMT catalyze the formation of (*S*)-*N*-methylcoclaurine from the 1-benzylisoquinoline (*S*)-coclaurine, PavNMT converts (*S*)-pavine to (*S*)-*N*-methylpavine, and TNMT *N*-methylates several different protoberberine alkaloids such as (*S*)-stylopine.

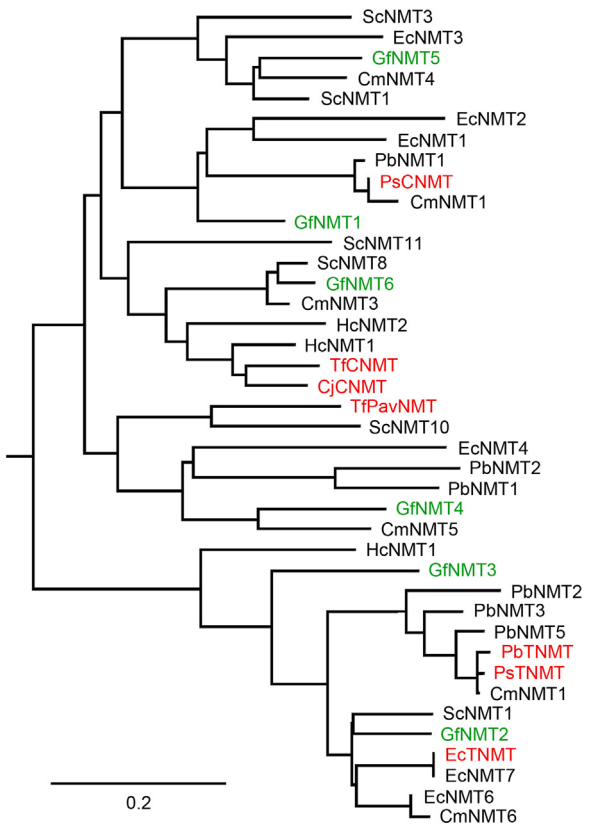

Fig. 7. Phylogenetic analysis of NMT candidates in the 454 and Illumina GA databases of six BIA-producing plant species listed in Tables 1 and 2. A total of 33 full-length cDNA sequences NMTs were found. Several functionally characterized NMTs are indicated in red. Sequences from *Glaucium flavum* are shown in green. Abbreviations and accession numbers: PsCNMT, *Papaver somniferum* coclaurine *N*-methyltransferase (AAP45316); TfCNMT, *Thalictrum flavum* coclaurine *N*-methyltransferase (AY610508); CjCNMT *Coptis japonica* coclaurine *N*-methyltransferase (BAB71802); TfPavNMT, *Thalictrum flavum* pavine *N*-methyltransferase (EU883010); EcTNMT, *Eschscholzia californica* tetrahydroprotoberberine *N*-methyltransferase (EU882977); PbTNMT, *Papaver bracteatum* tetrahydroprotoberberine *N*-methyltransferase (EU882994); PsTNMT, *Papaver somniferum* tetrahydroprotoberberine *N*-methyltransferase (DQ028579); Cm, *Chelidonium majus*; Gf, *Glaucium flavum*; Hc, *Hydrastis canadensis*; Sc, *Sanguinaria canadensis*. (For interpretation of the references to color in figure legend, the reader is referred to the web version of the article.)

ammonium in (+)-magnoflorine has not been identified (Fig. 6), CNMT from *Coptis japonica* was reported to *N*-methylate a broad range of substrates including corytuberine (Minami et al., 2008). Among the six BIA-producing plant species as listed in Table 1, *Glaucium flavum* is known to accumulate substantial quantities of (+)-magnoflorine (Novák and Slavík, 1974) suggesting that an efficient corytuberine *N*-methyltransferase is represented among the NMT homologs in the transcriptome database for this plant. A phylogenetic tree was constructed using predicted amino acid sequences of the NMT homologs from all six BIA-producing species (Table 1) and several functionally characterized NMTs from related plants (Fig. 7). Six full-length paralogs distributed in three different NMT subclades were identified from the *G. flavum* 454 and Illumina GA sequence databases. Based on the extensive sequence similarity, most of the candidate genes are expected to encode NMTs involved in BIA metabolism. However, empirical enzyme characterization is required to confirm precise catalytic function. For example, GfNMT1 is the most likely CNMT functional ortholog in *G. flavum*, GfNMT2 and GfNMT3 are expected to exhibit TNMT activities with unique or overlapping substrate preferences, and GfNMT4 could function as a PavNMT (Fig. 7). In contrast, the predicted amino acid sequences of GfNMT5 and GfNMT6 are sufficiently distinct to suggest unique substrate specificities. Considering that (*S*)-corytuberine exhibits structurally similarity to the CNMT substrate (*S*)-coclaurine, GfNMT5 or GfNMT6 are candidates for a predicted corytuberine *N*-methyltransferase in *G. flavum* (Fig. 6). Gene triage is essential for selecting priority candidates from large gene families.

Polyketides – Polyketides are another group of structurally diverse and biologically active metabolites. The use of St. John's Wort (*Hypericum perforatum*) for the treatment of depression has been attributed to compounds that include the polyketides hypericin and hyperforin (Nahrstedt and Butterweck, 2010). The biosynthesis of the prenylated acylphloroglucinol hypericin, which accumulates leaf and flower glands (Soelberg et al., 2007; Karppinen et al., 2007), has been partially elucidated (Karppinen et al., 2008; Karppinen and Hohtola, 2008). The acylphloroglucinol moiety of hyperforin is produced by a polyketide synthase (PKS) that condenses isobutyryl-CoA with three molecules of malonyl-CoA to form phlorisobutyrophenone (PIBP). PIBP is then prenylated three times using dimethylallyl diphosphate (DMAPP) as the donor, and once using geranyl diphosphate (GPP), to yield hyperforin. Genes encoding PIBP synthase and the aromatic prenyltransferases (PTs) are not known (Fig. 8).

To identify aromatic PTs in *H. perforatum*, the search term 'prenyltransferase' was used to query the Illumina GA dataset, which returned 67 candidates annotated as UbiA prenyltransferase-like proteins. This family includes the prokaryotic 4-hydroxybenzoate octaprenyltransferase, 1-4-dihydroxy-2-naphthoate octaprenyltransferases involved in menaquinone biosynthesis, and farnesyltransferase/geranylgeranyltransferase-type alpha subunits. A tBLASTn query was performed using the

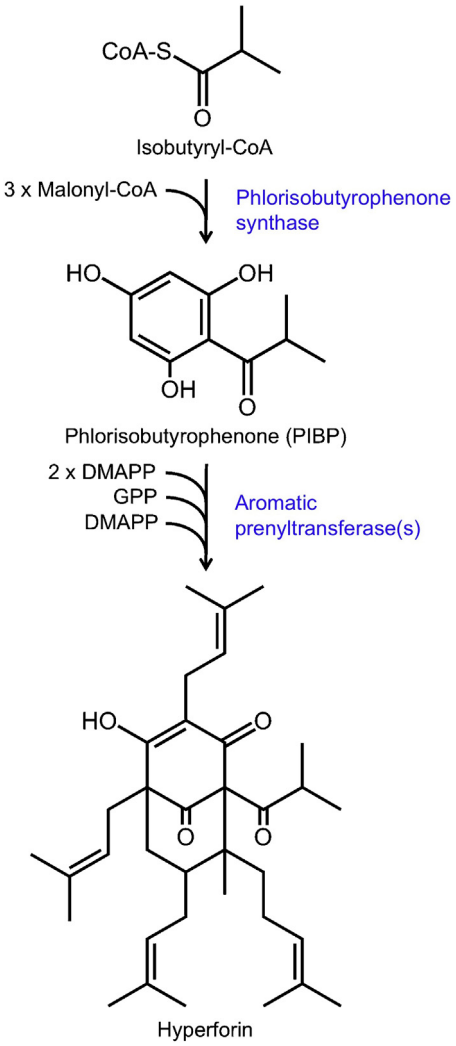

Fig. 8. Putative biosynthetic pathway for hyperforin. The order of prenylation reactions involving dimethylallyl diphosphate (DMAPP) and geranyl diphosphate (GPP) as donors is unknown.

nucleotide sequence of the aromatic PT involved in the bitter acid pathway in hop (*Humulus lupulus*) (Tsurumaru et al., 2012) and 29 sequences were retrieved encoding the UbiA-like proteins identified from the text search. Eleven candidates homologous to the hop PT are expressed in *H. perforatum* glands (Supplementary Table 5). Partial CDS unigenes obtained from the Illumina GA dataset were used as BLASTn queries against the 454 database, which extended one of the contigs. Since the transcriptome databases were produced from isolated glands, PT candidates with the highest read counts are likely involved in hyperforin biosynthesis.

3.6. BLAST portal for transcriptome databases

Assembled 454 and Illumina GA databases for the 75 plant species in the PhytoMetaSyn Project are available at http://www.phytometasyn.ca/index.php?option+com_php&Itemid+103 (Supplementary Fig. 6). Users can access a BLAST search page, submit their sequences in FASTA format, choose the species they wish to query, specify the search type (*i.e.* BLASTn, tBLASTx, tBLASTn) and set the *e*-value threshold before submission. Results can either be displayed in the browser or sent to the user by email.

4. Discussion

The ultimate goal of the PhytoMetaSyn Project is to establish rational design and engineering principles to reconstruct biosynthetic pathways leading to high-value plant specialized metabolites in microorganisms. A critical step is the establishment of deep transcriptome databases for a wide range of plant species producing a variety of bioactive compounds. As shown herein, dynamic transcript mapping coupled with NGS technologies provides a comprehensive catalog of novel biosynthetic genes involved in specialized metabolic pathways. 454 and Illumina GA have been two of the most widely used NGS platforms applied to more than 50 plant species. Although both technologies are reliable for generating transcriptome databases owing to their wide and deep sequence coverage, each is associated with common problems including sequence artifacts, poor quality reads and primer/adaptor contamination (Patel and Jain, 2012). The 454 platform is affected by a high error rate in homopolymeric tracts and sequences resulting from artificial amplification (Margulies et al., 2005; Quince et al., 2009; Gomez-Alvarez et al., 2009). Illumina GA technology circumvents these issues, but inherent base-calling errors generally cause systemic bias in the sequence data (Erlich et al., 2008; Nakamura et al., 2011). Illumina sequencing has been shown to yield longer and more accurate contigs compared with 454 pyrosequencing, despite the shorter read-lengths (Luo et al., 2012), owing to substantially more extensive sequence coverage. In our study, the number of reads and percentage of unigenes matching reference sequences were greater for the Illumina GA compared with the 454 platform. A notable consequence was the increased likelihood of identifying full-length open reading frames. However, each NGS platform has inherent advantages and disadvantages (Egan et al., 2012). Plant genomes and to some extent transcriptomes contain highly repetitive sequences that are often difficult to assemble from short-read sequence data, providing a utility for complementary 454 pyrosequencing, which generates longer reads capable of bridging gaps in the Illumina GA data. Whereas methods are available to effectively merge genomic 454 and Illumina GA assemblies (DiGuistini et al., 2009), the use of Velvet/Oases or MIRA for hybrid assembly for RNASeq data did not yield improved results as neither tool was optimal for both data types. As such, we have not yet adapted a method to merge 454 and Illumina GA assemblies, or to generate combined assemblies. However, manual alignment of overlapping contigs from each platform is an effective approach to further reduce errors, fill gaps and extend the coverage of partial transcripts.

Since reference genomes for the 20 plant species used in this work, and most of the 75 species targeted by the PhytoMetaSyn Project, are not yet available, the *de novo* assembly of transcripts was required. Even when a reference genome is available, *de novo* assembly is necessary to recover transcripts corresponding to missing regions of an assembled genome. Major advantages of *de novo* transcriptome assembly include independence from the correct alignment of sequence reads to known splice sites and the prediction of novel splicing sites, which are required for reference-based assembly. Major challenges for the *de novo* assembly of higher eukaryotic transcriptomes include difficulties associated with handling increasingly large datasets, the wide range in the coverage depth of transcripts, and the identification of alternative splice variants. In this study, the MIRA and Velvet/Oases platforms were used for assembly of 454 and Illumina GA data, respectively. We demonstrated the ability to isolate full-length gene candidates encoding targeted biosynthetic enzymes involved in a variety of specialized metabolic pathways from assembled databases generated by both sequencing technologies (Fig. 3 and Table 2). The availability of the assembled and annotated databases on the user-friendly MAGPIE interface (Supplementary Figs. 2–5) facilitates exploration of the gene summaries for the triage of candidate

genes. The relative gene expression levels represented by FPKM (Fragments Per Kilobase of exon model per Million mapped reads) associated with our sequence datasets is indispensable to determine whether or not the abundance of transcripts corresponding to selected contigs is within the expected range compared with other known biosynthetic genes in a specific specialized metabolic pathway.

Over the course of the present study, NGS sequencing technologies have advanced rapidly. Second-generation instruments, such as 454 GS FLX and Illumina GA, have been followed over the past two years by several benchtop sequencing platforms (so-called 'third-generation' technologies) including 454 GS Junior (Roche), MiSeq (Illumina), Ion Proton (Life Technologies), Single-Molecular Real-Time Sequencer (Pacific Biosciences), and the Heliscope Single-Molecule Sequencer (Helicos Biosciences) (Thompson and Milos, 2011; Thudi et al., 2012). Remarkably, these instruments provide longer sequence reads in a shorter time and at lower cost, compared with 454 and Illumina GA. 'Fourth-generation' technologies (Egan et al., 2012) will potentially elevate sequencing to unprecedented levels. The widespread adoption of the most novel NGS approaches however is limited by their increased trade off of sequencing accuracy for volume and the need for extensive creation of dedicated bioinformatics tools to support downstream data manipulation and analysis. Nevertheless, currently and widely available sequencing and bioinformatics platforms provide sufficiently accurate and annotated databases to perform state-of-the-art biochemical genomics. In addition to the PhytoMetaSyn Project, other deep plant transcriptome surveys using similar technology platforms include the 1KP Project (http://www.onekp.com), which is targeting transcriptomes from 1000 plant species (Schliesky et al., 2012; Johnson et al., 2012), and the Medicinal Plant Genomics Resource (http://www.medicinalplantgenomics.msu.edu), which provides transcriptome and metabolome resources for 14 medicinal plants (Góngora-Castillo et al., 2012). Three species investigated by MPGR overlap with those reported herein. A brief examination of Illumina assembly statistics between MPGR and our own transcript data show comparable mean unigene lengths (Supplementary Table 6). Our assemblies generate about 50% of the final number of contigs compared with MPGR datasets. Both are within a reasonable and expected range, and differences likely result from variations in assembly methods and input data. Overall, our data are complementary to those produced by other projects.

The acquisition of novel biosynthetic genes with diverse catalytic functions will provide a road map for the specialized metabolic engineering of plants and microorganisms leading to the commercial production of high-value bioproducts. Our transcriptome databases include a vast number of full-length gene sequences for 75 plant species targeted by the PhytoMetaSyn Project, 20 of which are described here in detail. Combined with targeted metabolomics, deep transcriptome databases are an essential resource for the identification and characterization of novel biosynthetic enzymes. In addition to the examples discussed here, which were drawn from six large classes of specialized metabolites, the catalog of biosynthetic genes and enzymes involved in plant specialized metabolism continues to expand. Among the targeted compounds types of the PhytoMetaSyn Project, more than 40,000 terpenes, 12,000 alkaloids and thousands of polyketides are known to occur in nature and many more await discovery (Facchini et al., 2012). Mining of transcriptome databases for the selected plant species will generate hundreds of biosynthetic genes encoding enzymes with novel catalytic activities and variants with similar functions, but different biochemical features. Genes encoding variants might display improved expression characteristics in plants and microorganisms, providing metabolic engineering options for the optimization of synthetic biosystems designed to produce high-value plant metabolites. The use of emerging 'plug-and-play' technology, which employs various combinations and permutations of biosynthetic genes to engineer multi-step biosynthetic pathways in microorganisms (Facchini et al., 2012), will accelerate the discovery of novel enzymes, and the reconstruction and optimization of natural and unnatural product pathways based on combinatorial biochemistry.

Several studies have partially reconstituted various plant natural product pathways in *Escherichia coli* or yeast (*Saccharomyces cerevisiae*) leading to the formation of taxadiene, a key isoprenoid intermediate in taxol biosynthesis (Engels et al., 2008; Ajikumar et al., 2010), amorphadiene, the sesquiterpene olefin precursor to artemisinin (Martin et al., 2003), artemisinic acid, the immediate precursor to artemisinin (Ro et al., 2008), the diterpene fragrance precursors *cis*-abienol (Zerbe et al., 2012) and sclareol (Caniard et al., 2012), and reticuline, a key intermediate in the biosynthesis of codeine and morphine (Minami et al., 2008; Hawkins and Smolke, 2008). Nevertheless, the deployment of most plant metabolic pathways in microbial hosts still requires the isolation and functional characterization of many unknown biosynthetic genes. Even when all biosynthetic genes required for the formation of a specific compound have been isolated from one plant and reconstituted in a microorganism, the specific catalytic characteristics of each enzyme can be inappropriate for the efficient operation of the metabolic pathway in a heterologous system. In such cases, the overall metabolic flux will be limited by the enzyme step with the lowest catalytic efficiency (Ajikumar et al., 2010). The availability of enzyme variants from a wide variety of plant species, as described in this work, provides a possible empirical solution to such metabolic engineering bottlenecks.

Acknowledgements

Sequencing was performed at the McGill University and Génome Québec Innovation Center. This work was supported by Genome Canada, Genome Alberta, Genome Quebec, Genome Prairie, Genome British Columbia, the Canada Foundation for Innovation, the Government of Alberta, the Ontario Ministry of Research and Innovation, the Government of Saskatchewan, the National Research Council of Canada and private sector partners. P.J.F., V.J.J.M., D.K.R. and V.D.L. were also funded by the Canada Research Chair program. J.B. was also supported by the Distinguished University Scholar program at the University of British Columbia.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbiotec.2013.04.004.

References

Attia, M., Kim, S.U., Ro, D.K., 2012. Molecular characterization of (+)-*epi*-α-bisabolol synthase from *Lippia dulcis* – the entry point enzyme for the natural sweetener, hernandulcin, biosynthesis. Archives of Biochemistry and Biophysics 527, 37–44.

Ajikumar, P.K., Xiao, W.H., Tyo, K.E., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T.H., Pfeifer, B., Stephanopoulos, G., 2010. Isoprenoid pathway optimization for taxol precursor overproduction in *Escherichia coli*. Science 330, 70–74.

Caniard, A., Zerbe, P., Legrand, S., Cohade, A., Valot, N., Magnard, J.-L., Bohlmann, J., Legendre, L., 2012. Discovery and functional characterization of two diterpene synthases for sclareol biosynthesis in *salvia sclarea* (L.) and their relevance for perfume manufacture. BMC Plant Biology 12, 119.

Chen, F., Tholl, D., Bohlmann, J., Pichersky, E., 2011. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant Journal 66, 212–229.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E., Wetter, T., Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Resource 14, 1147–1159.

Chomczynski, P., Sacchi, N., 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. Analytical Biochemistry 162, 156–159.
De Luca, V., Salim, V., Masada Atsumi, S., Yu, F., 2012. Mining the biodiversity of plants: a revolution in the making. Science 336, 1658–1661.
Desgagne-Penix, I., Khan, M.F., Schriemer, D.C., Cram, D., Nowak, J., Facchini, P.J., 2010. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biology 10, 252.
DiGuistini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R., Hirst, M., Mardis, E., Marra, M.A., Hamelin, R.C., Bohlmann, J., Breuil, C., Jones, S.J.M., 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biology 10, R94.
Dixon, R.A., 2005. Engineering of plant natural product pathways. Current Opinion in Plant Biology 8, 329–336.
Egan, A.N., Schlueter, J., Spooner, D.M., 2012. Applications of next-generation sequencing in plant biology. American Journal of Botany 99, 175–185.
Engels, B., Dahm, P., Jennewein, S., 2008. Metabolic engineering of taxadiene biosynthesis in yeast as a first step towards Taxol (Paclitaxel) production. Metabolic Engineering 10, 201–206.
Erlich, Y., Mitra, P.P., delaBastide, M., McCombie, W.R., Hannon, G.J., 2008. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nature Methods 5, 679–682.
Fabricant, D.S., Farnsworth, N.R., 2001. The value of plants used in traditional medicine for drug discovery. Environmental Health Perspectives 1, 69–75.
Facchini, P.J., Bohlmann, J., Covello, P.S., De Luca, V., Mahadevan, R., Page, J.E., Ro, D.K., Sensen, C.W., Storms, R., Martin, V.J.J., 2012. Synthetic biosystems for the production of high-value plant metabolites. Trends in Biotechnology 30, 127–131.
Fields, S., Johnston, M., 2005. Whither model organism research? Science 307, 1885–1886.
Furuno, M., Kasukawa, T., Saito, R., Adachi, J., Suzuki, H., Baldarelli, R., Hayashizaki, Y., Okazaki, Y., 2003. CDS annotation in full-length cDNA sequence. Genome Research 13, 1478–1487.
Gaasterland, T., Sensen, C.W., 1996. MAGPIE: automated genome interpretation. Trends in Genetics 12, 76–78.
Gomez-Alvarez, V., Teal, T.K., Schmit, T.M., 2009. Systematic artifacts in metagenomes from complex microbial communities. ISME Journal 3, 1314–1317.
Góngora-Castillo, E., Childs, K.L., Fedewa, G., Hamilton, J.P., Liscombe, D.K., Magallanes-Lundback, M., Mandadi, K.K., Nims, E., Runguphan, W., Vaillancourt, B., Varbanova-Herde, M., Dellapenna, D., McKnight, T.D., O'Connor, S., Buell, C.R., 2012. Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. PLoS ONE 7, e52506.
Hall, D.E., Zerbe, P., Jancsik, S., Quesada, A.L., Dullat, H., Madilao, L.L., Yuen, M., Bohlmann, J., 2013. Evolution of conifer diterpene synthases. Plant Physiology 161, 600–616.
Hawkins, K.M., Smolke, C.D., 2008. Production of benzylisoquinoline alkaloids in *Saccharomyces cerevisiae*. Nature Chemical Biology 4, 564–573.
Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Research 9, 868–877.
Iseli, C., Jongeneel, C.V., Bucher, P., 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. In: Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, 7, pp. 138–214.
Johnson, M.T.J., Carpenter, E.J., Tian, Z., Bruskiewich, R., Burrls, J.N., Carrigan, C.T., Chase, M.W., Clarke, N.D., Covshoff, S., dePamphills, C.W., Edger, P.P., Goh, F., Graham, S., Greiner, S., Hbberd, J.M., Jordon-Thaden, I., Kutchan, T.M., Leebens-Mack, J., Melkonian, M., Miles, N., Myburg, H., Patterson, J., Pires, J.C., Ralph, P., Rolf, M., Sage, R.F., Soltis, D., Pamela, S., Stevenson, D., Steward Jr., C.N., Surek, B., Thomsen, C.J.M., Villarreal, J.C., Wu, X., Zhang, Y., Deyholos, M.K., Wong, G.K-S., 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS ONE 7, e50226.
Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110, 462–467.
Karppinen, K., Hohtola, A., 2008. Molecular cloning and tissue-specific expression of two cDNAs encoding polyketide synthases from *Hypericum perforatum*. Journal of Plant Physiology 165, 1079–1086.
Karppinen, K., Hokkanen, J., Mattila, S., Neubauer, P., Hohtola, A., 2008. Octaketide-producing type III polyketide synthase from *Hypericum perforatum* is expressed in dark glands accumulating hypericins. FEBS Journal 275, 4329–4432.
Karppinen, K., Hokkanen, J., Tolonen, A., Mattila, S., Hohtola, A., 2007. Biosynthesis of hyperforin and adhyperforin from amino acid precursors in shoot cultures of *Hypericum perforatum*. Phytochemistry 68, 1038–1045.
Keeling, C.I., Bohlmann, J., 2006. Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. New Phytologist 170, 657–675.
Keeling, C.I., Weisshaar, S., Ralph, S.G., Jancsik, S., Hamberger, B., Dullat, H.K., Bohlmann, J., 2011. Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (*Picea* spp.). BMC Plant Biology 11, 43.
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25.
Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323.
Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
Liscombe, D.K., Ziegler, J., Schmidt, J., Ammer, C., Facchini, P.J., 2009. Targeted metabolite and transcript profiling for elucidating enzyme function: isolation of novel *N*-methyltransferases from three benzylisoquinoline alkaloid-producing species. Plant Journal 60, 729–743.
Lottaz, C., Iseli, C., Jongeneel, C.V., Bucher, P., 2003. Modeling sequencing errors by combining hidden Markov models. Bioinformatics 19, ii103–ii112.
Luo, C., Tsementzi, D., Kyrpides, N., Read, T., Konstantinidis, K.T., 2012. Direct comparisons of Illumina versus Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS ONE 7, e30087.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.
Martin, D.M., Fäldt, J., Bohlmann, J., 2004. Functional characterization of nine Norway Spruce TPS genes and evolution of gymnosperm terpene synthases of the TPS-d subfamily. Plant Physiology 135, 1908–1927.
Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nature Reviews Genetics 12, 671–682.
Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., Keasling, J.D., 2003. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. Nature Biotechnology 21, 796–802.
Metzker, M.L., 2010. Sequencing technologies – the next generation. Nature Reviews Genetics 11, 31–46.
Minami, H., Kim, J.S., Ikezawa, N., Takemura, T., Katayama, T., Kumagai, H., Sato, F., 2008. Microbial production of plant benzylisoquinoline alkaloids. Proceedings of the National Academy of Sciences of the United States of America 105, 7393–7398.
Nahrstedt, A., Butterweck, V., 2010. Lessons learned from herbal medicinal products: the example of St. John's Wort (perpendicular). Journal of Natural Products 73, 1015–1021.
Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-U.I-Amin Md Ogasawara, N., Kanaya, S., 2011. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research 39, e90.
Nelson, D., Werck-Reichhart, D., 2011. A P450-centric view of plant evolution. Plant Journal 66, 194–211.
Novák, V., Slavík, J., 1974. Further alkaloids from *Glaucium flavum* CR. Collection of Czechoslovak Chemical Communications 39, 3352–3356.
Oksman-Caldentey, K.M., Inzé, D., 2004. Plant cell factories in the post-genomic era: new ways to produce designer specialized metabolites. Trends in Plant Sciences 9, 433–440.
Oksman-Caldentey, K.M., Saito, K., 2005. Integrating genomics and metabolomics for engineering plant metabolic pathways. Current Opinion in Biotechnology 16, 174–179.
Ouyang, S., Buell, C.R., 2004. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Research 32, D360–D363.
Patel, R.K., Jain, M., 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE 7, e30619.
Pickel, B., Drew, D.P., Manczak, T., Weitzel, C., Simonsen, H.T., Ro, D.K., 2012. Identification and characterization of a kunzeaol synthase from *Thapsia garganica*: implications for the biosynthesis of the pharmaceutical thapsigargin. Biochemical Journal 448, 261–271.
Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glöckner, F.O., 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35, 7188–7196.
Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): acurated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 35, D61–D65.
Pyle, B.W., Tran, H.T., Pickel, B., Haslam, T.M., Gao, Z., Macnevin, G., Vederas, J.C., Kim, S.U., Ro, D.K., 2012. Enzymatic synthesis of valerena-4,7(11)-diene by a unique sesquiterpene synthase from the valerian plant (*Valeriana officinalis*). FEBS Journal 279, 3136–3146.
Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F., Sloan, W.T., 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Methods 6, 639–641.
Rates, S.M.K., 2001. Plants and source of drugs. Toxicon 39, 603–613.
Ro, D.K., Ouellet, M., Paradise, E.M., Burd, H., Eng, D., Paddon, C.J., Newman, J.D., Keasling, J.D., 2008. Induction of multiple pleiotropic drug resistance genes in yeast engineered to produce an increased level of antimalarial drug precursor, artemisinic acid. BMC Biotechnology 8, 83.

Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., Chang, M.C., Withers, S.T., Shiba, Y., Sarpong, R., Keasling, J.D., 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440, 940–943.

Schliesky, S., Gowik, U., Weber, A.P., Bräutigam, A., 2012. RNA-Seq assembly – are we there yet? Frontiers in Plant Science 3, 220–230.

Schuler, M.A., Werck-Reichhart, D., 2003. Functional genomics of P450s. Annual Review of Plant Biology 54, 629–667.

Soelberg, J., Jörgensen, L.B., Jäger, A.K., 2007. Hyperforin accumulates in the translucent glands of Hypericum perforatum. Annals of Botany 99, 1097–1100 (Erratum, Annals of Botany 100, 679).

Suzuki, S., Ono, N., Furusawa, C., Ying, B.W., Yomo, T., 2011. Comparison of sequence reads obtained from three next-generation sequencing platforms. PLoS ONE 6, e19534.

Thompson, J.F., Milos, P.M., 2011. The properties and applications of single-molecule DNA sequencing. Genome Biology 12, 217.

Thudi, M., Li, Y., Jackson, S., May, G.D., Varshney, R.K., 2012. Current state-of-art of sequencing technologies for plant genomics research. Briefings in Functional Genomics 11, 3–11.

Tsurumaru, Y., Sasaki, K., Miyawaki, T., Uto, Y., Momma, T., Umemoto, N., Momose, M., Yazaki, K., 2012. HlPT-1, a membrane-bound prenyltransferase responsible for the biosynthesis of bitter acids in hops. Biochemical and Biophysical Research Communications 417, 393–398.

Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for aligning DNA sequences. Journal of Computational Biology 7, 203–214.

Zerbe, P., Chiang, A., Yuen, M., Hamberger, B., Hamberger, B., Draper, J.A., Britton, R., Bohlmann, J., 2012. Bifunctional *cis*-abienol synthase from *Abies balsamea* discovered by transcriptome sequencing and its implications for diterpenoid fragrance production. Journal of Biological Chemistry 287, 12121–12131.

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

Ziegler, J., Facchini, P.J., 2008. Alkaloid biosynthesis: metabolism and trafficking. Annual Review of Plant Biology 59, 735–769.