Journal of Biotechnology xxx (2013) xxx–xxx

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 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 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, 3280 Hospital Drive NW, Calgary, Alberta T2N 1N4, 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 xxx

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 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

* 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/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.jbiotec.2013.04.004

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 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

Please cite this article in press as: Xiao, M., et al., Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. J. Biotechnol. (2013), http://dx.doi.org/10.1016/j.jbiotec.2013.04.004