Contents lists available at SciVerse ScienceDirect

Gene

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

# Untangling the transcriptome from fungus-infected plant tissues

Sheng Zhu [a,1], Yong-Mei Dai [b], Xin-Ye Zhang [a,c], Jian-Ren Ye [b], Ming-Xiu Wang [a], Min-Ren Huang [a,*]

[a] *Jiangsu Key Laboratory for Poplar Germplasm Enhancement and Variety Improvement, Nanjing Forestry University, Nanjing 210037, China*
[b] *Institute of Forest Protection, Nanjing Forestry University, Nanjing 210037, China*
[c] *Hubei Province Forestry Science Academy, Wuhan 430079, China*

ARTICLE INFO

*Article history:*
Accepted 13 February 2013
Available online 1 March 2013

*Keywords:*
RNA-Seq
Assembly
Discrimination
Origin

ABSTRACT

The development of sequencing technology allows low-cost generation of sequence data. The huge amount of raw sequence data now available has introduced many challenges associated with analysis of these large-scale data banks. For example, it is very important to distinguish materials of plant and fungal origin in fungus-infected plant tissue. The origin of transcripts that were sequenced from Library 895-M6 (poplar tissue infected by *Marssonina brunnea*) on Illumina/Solexa GA IIx was determined by combining three methods: (1) based on the taxonomic information of homologous sequences; (2) based on the reference genome sequence; (3) based on the transcriptome sequence of the host and its pathogen obtained from Library 895 (poplar) and Library M6 (*M. brunnea*) as well as Library 895-M6 (mixture of poplar and *M. brunnea*). We idenified accurately the origin of 80,978 (99.5%) contigs in the mixed poplar and *M. brunnea* sample (Library 895-M6) by integrating the results from the three methods. The results of this study demonstrate that a combination of these three approaches described here is an effective strategy for determining the origin of sequences in a mixed pool, and provides a basis for further transcriptome analysis of the mixed sample.

## 1. Introduction

Fungal diseases are responsible for reduced yields of trees and crops. For example, Marssonina leaf spot is a common foliage disease in Poplar tree species caused by the pathogenic fungus *Marssonina brunnea* (Han et al., 2000). In the process of developing an effective strategy for controlling fungal diseases of plants, the first step is to understand plant diseases at the molecular level (Martin et al., 2003; Schwessinger and Ronald, 2012; Varshney et al., 2009; Zimaro et al., 2011). To analyze and compare the transcriptome of both fungus-infected and healthy (or normal) plant tissues is possibly one of the most straightforward and effective ways to understand the molecular mechanism of fungal plant diseases (Bonfante and Genre, 2010; Guimil et al., 2005; Venu et al., 2011).

DNA microchip(Stoughton, 2005) and expressed sequence tag (EST) (Nagaraj et al., 2007) technology are two kinds of traditional and useful technology for studying the transcriptome. Nevertheless, the ability to study the transcriptome via these two technologies is limited by their own shortcomings. DNA microchip technology relies on the presence of a suitable reference sequence and associated annotation, and results in a high false-positive rate (Shendure, 2008); further, EST technology cannot provide information on transcripts of relatively low-level expression (Wang et al., 2009). With the advance of DNA sequencing technology, the emergence of RNA-Seq technology depending upon fast, low-cost, single-base resolution and a high-throughput next-generation sequencing (NGS) platform has all its merits and compensates for almost all their shortcomings (Metzker, 2010; Wang et al., 2009). Hence, RNA-Seq technology has begun to be applied in the transcriptome of a wide range of eukaryotic organisms, including animals (e.g. *Homo sapiens* Wang et al., 2008), plants (e.g. *Arabidopsis thaliana* Lister et al., 2008) and fungi (e.g. *Saccharomyces cerevisiae* Nagalakshmi, 2008).

As a result of the continuing advances in sequencing technology, a large amount of genome and transcriptome sequences is generated every day from different sequencing platforms for studying various biological problems (Ozsolak and Milos, 2011). RNA-Seq technology can be applied to nearly all aspects of gene expression and regulation, such as the identification of novel transcripts (or gene), splice variants and transcript quantification (Robertson et al., 2010; Wang et al., 2009). Therefore, it has become a routine, but complex, task to analyze massive amounts of RNA-Seq data. Analysis of the whole transcriptome of both infected and healthy tissues is possibly one of the most effective ways to gain a full understanding of host–pathogen interaction at the molecular level (Dhiman et al., 2009; Soanes and Talbot, 2005; Wu et al., 2010). Because of the disadvantage of microarray and expressed sequence tag (EST) technology in terms of studying the transcriptome, the molecular mechanism of interaction between plant and pathogen is explored using RNA-Seq technology instead of microarray and EST technology. However, there are few published studies about the discrimination

*Abbreviation:* bp, Base pairs; EST, Expressed sequence tag; Gb, Gigabases; hpi, Hour post inoculation; IGV, Integrative Genomics Viewer; Mb, Megabases; NGS, Next-generation sequencing; ORF, Open reading frame; PE1, paired-end library 1; PE2, paired-end library 2; PSA, Potato sugar agar; Q1, First quartile score; Q3, 3rd quartile score; QC, Quality control; RH, Relative humidity; SRA, Short Reads Archive.

[*] Corresponding author. Tel./fax: +86 025 85427412.
*E-mail address:* njmrhuang@njfu.edu.cn (M.-R. Huang).

[1] These authors contributed equally to this work.

0378-1119/$ – see front matter. Crown Copyright © 2013 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.02.023

between plant and fungus origin of RNA-Seq data from plant tissues infected with plant pathogen or phytopathogen. It was vitally important to distinguish between pathogenic and plant sequences in mixed EST or RNA-Seq collections from the fungus-infected plant tissues for further analysis of the transcriptome, such as identifying novel transcripts and gene expression analysis (Hsiang and Goodwin, 2003). So, discrimination between the host and the host pathogen in the mixed pool is valuable for dissecting the transcriptome of the infected tissues, for understanding the mechanism of host–pathogen interactions and for developing strategies of fungal disease control (Hsiang and Goodwin, 2003; Jantasuriyarat et al., 2005).

To study the transcriptome from fungus-infected poplar leaf tissue, we used an Illumina/Solexa Genome Analyzer IIx to generate RNA-Seq data, and obtained 3.0 gigabase (Gb) paired-end reads from Library 895 (uninfected poplar leaf tissues), 2.5 Gb paired-end reads from Library M6 (the conidia of *M. brunnea*) and 2.8 Gb paired-end reads from Library 895-M6 (poplar leaf tissues infected with *M. brunnea*). We present the transcriptome analysis of the mixed library, including RNA-Seq data quality assessment, de novo transcript assembly, assembly assessment, determination of the transcript origin in the fungus-infected tissues and transcriptome annotation. This study provides a resource for defining the origin of transcripts in the mixed samples and for facilitating the expression level of genes during host–pathogen interaction.

## 2. Materials and methods

### 2.1. Sample preparation and sequencing

Three different samples were used in this study. Two libraries were constructed from leaf tissues of poplar clone NL895 (*Populus* × *euramericana* CL NL895 (a clone cultivated in our laboratory); labeled as Library 895) without infection by pathogens, and from the conidia of the poplar leaf pathogens *M. brunnea* f. sp. *multigermtubi* (labeled as Library M6) grown on potato–sugar–agar (PSA). The other library (labeled as Library 895-M6) was constructed from poplar leaf tissues infected with *M. brunnea*. The healthy poplar leaf tissues were inoculated with spore suspensions (80,000 spores/ml), incubated at 22 °C, 100% relative humidity (RH) with a cycle of 12 h light/12 h darkness, and harvested at 96 h post inoculation (hpi) and samples were frozen quickly in liquid nitrogen. Total RNA was isolated with TRIzol® Reagent (Invitrogen) according to the manufacturer's instructions, following the removal of genomic DNA with DNaseI (TaKaRa). Total RNA was converted to cDNA using a QuantiTect Reverse Transcription Kit (Qiagen) following the standard protocol. Library 895 was obtained from five uninfected poplar leaves and 895-M6 was obtained from 6 infected poplar leaves. Sequencing of the transcriptome for the three libraries was done with an Illumina/Solexa Genome Analyzer IIx using paired-end technology.

### 2.2. Data availability

The RNA-Seq data generated in this study have been submitted to the Short Reads Archive (SRA) under the accession numbers SRR504342 (Library 895), SRR504340 (Library M6) and SRR504343 (Library 895-M6). The ESTs used in the study are available in GenBank under the accession numbers CX167465–CX187487, including CX167465–CX177330 for leaves of poplar I69 (*Populus deltoids* CL Lux (I-69/55) inoculated by *M. brunnea* (72 hpi) and CX177331–CX187487 for leaves of poplar I45 (*P. euramericana* CL I-45/51) inoculated with *M. brunnea* (72 hpi). The genome sequence of *M. brunnea* f. sp. *multigermtubi* is composed of 90 scaffolds totaling nearly 52 × $10^6$ bp and is available from GenBank (accession number AFXC00000000). The genome sequence of *Populus trichocarpa* (poplar) is composed of 2518 scaffolds containing 423 × $10^6$ bp and is available from Phytozome (http://www.phytozome.net/poplar, Phytozome version 7.0 in JGI).

### 2.3. De novo assembly

FastQC (version 0.10.0; http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) in combination with fastx_quality_stats in the package FASTX Toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/commandline.html) were used to calculate the quality values across all bases at each position in FASTQ files coming from an Illumina/Solexa Genome Analyzer IIx. The low-quality end regions of reads from the three libraries were trimmed with fastx_trimmer in the package FASTX Toolkit.

Transcriptome short reads or RNA-Seqs were de novo assembled using the RNA-Seq assembler Trinity (version 2012-01-25) (Grabherr et al., 2011). RNA-Seq reads were mapped to the assembled contigs using the short-read mapper Bowtie (version 0.12.7) (Langmead et al., 2009). The alignment of RNA-Seq reads to the corresponding genomic sequences was performed with the splice junction aligner Tophat (version 1.3.3)(Trapnell et al., 2009). The resulting contigs were located on the reference genome using BLAT (version 34)(Kent, 2002) with >90% identity and 80% coverage. RNA-Seq assembly results were visualized using Integrative Genomics Viewer (IGV; version 2.0.x) (Robinson et al., 2011) and Tablet (version 1.12.02.06) (Milne et al., 2010). The read coverage across all locations in all contigs was calculated with SAMtools (version 0.1.18) (Li et al., 2009) and VarScan (version 2.2.8) (Koboldt et al., 2009).

### 2.4. Discrimination between plant and fungus origin in the mixed pool

Here, we used three methods to distinguish between fungus and plant ETS-like sequences in poplar leaf tissue infected with *M. brunnea*. The first method is based on homology search (Hsiang and Goodwin, 2003; Sharma et al., 2012). The homology search of the assembled contigs against the NR protein database was done using BLASTX (version 2.2.25) with an *E*-value cutoff of $1E^{-10}$; following the taxonomic information of aligned sequences in the NR database was used as the evidence of identification of the origin of assembled contigs in the mixed pool. The results of the origin of contigs in Library 895 and Library M6 identified by the first method were used to calculate the expected value of the false-positive rate for the method.

The second method depended on the available reference genome sequences of both poplar (Phytozome version 7.0 in JGI) and *M. brunnea* f. sp. *multigermtubi* (GenBank accession number AFXC00000000) (Jantasuriyarat et al., 2005). The assembled contigs from the mixed library were aligned with the genome sequence of poplar and *M. brunnea* using the program BLAT (version 34) (Kent, 2002) and alignments with >90% identity and 80% coverage were retained. The contigs fall into two classes, poplar and *M. brunnea* origin, based on the information of the reference genome to which the contig sequences can be aligned. The results of the origin of contigs in Library 895 and Library M6 identified by the second method were used to calculate the expected value of the false-positive rate for the method.

The final method utilized two organism transcriptome sequences (Jantasuriyarat et al., 2005), including *M. brunnea* (Library M6) and poplar clone NL895 (Library 895). By the alignment of the contigs in the mixed library to the cDNA sequences from both Library 895 and Library M6 using BLAST (version 2.2.25), the origin of the aligned cDNA sequences of both poplar and *M. brunnea* was used to determine the origin of the corresponding contigs in the poplar leaf tissues after infection with *M. brunnea*. The results of the origin of contigs in Library 895 and Library M6 identified by the third method were used to calculate the expected value of the false-positive rate. When self-aligning (Library M6 against Library M6 or Library 895 against Library 895), the best hit excluding itself was selected as evidence for the identification of its origin.

The origin of a contig was identified manually by combination with the putative ORF, reference genome and NR database