Sequencing ESTs from *Phytophthora cinnamomi* infected avocado root cDNA libraries.

*Mahomed W*1,2, *van den Berg N.*1,2, *Myburg A.A.*1,2

1 Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa
2 Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa

Avocado (*Persea americana*) is a diploid angiosperm of the Lauraceae family and is an important commercial fruit crop in almost 50 countries around the world. One of the most serious threats to the avocado industry is *Phytophthora* root rot (PRR) and the impact of the disease is further magnified by the mounting tolerance to the use of phosphite, which is the only effective control strategy. Despite the importance of both the avocado crop and the effect *Phytophthora* has on its cultivation, there is a lack of molecular knowledge underpinning our understanding of defence strategies against the pathogen. The objective of this study was therefore to sequence cDNA libraries of avocado roots infected with PRR and to identify genes related to tolerance against PRR. To achieve this, 454 pyrosequencing of pre- and post-infected, tolerant avocado root cDNA libraries was conducted. A total of 2 Mb of data was produced using a single lane on the GS-FLX platform. Approximately 10,000 reads were generated from which contigs were constructed using the Newbler assembly program. Further characterisation of the sequence data was done using the CAS program and the NCBI database. Approximately 20 putative defence related genes were identified of which 10 were quantified with quantitative RT-PCR. Genes such as metallothionein, thaumatin and the pathogenesis related protein *PsemI* were found to be differentially regulated. This study is the first step in elucidating the avocado transcriptome as well as identifying host-specific defence responses of avocado and a root pathogen.

Secuenciación de librerías EST desde raíces de palto infectadas con *Phytophthora cinnamomi*

Palto (*Persea americana*) es una angiosperma diploide de la familia Lauraceae y es un importante cultivo frutal en casi 50 países del mundo. Uno de los más serios problemas en la industria del palto es *Phytophthora* root rot (PRR), y el impacto de la enfermedad es agravado por la falta de una efectiva estrategia de control. A pesar de la importancia de *Phytophthora* en el cultivo de palto, existe una falta de conocimiento que permita el entendimiento de la estrategia molecular de defensa contra el patógeno. El objetivo de este estudio fue por consiguiente el transcriptoma de las raíces de palto infectadas por *Phytophthora cinnamomi* y la identificación de genes relacionados al tolerancia contra PRR. Para lograr esto, se utilizó pirosecuencia 454 para construidas librerías de raíces de palto en variedades tolerantes, antes y después de la infección. Un total de 2 Mb de datos fue producido usando un solo carril de la plataforma GS-FLX. Cerca de 10 mil lecturas fueron generados, los contig fueron construidos usando el programa de ensamblaje Newbler. La caracterización posterior se realizó usando el programa CAS y la base de datos NCBI. Aproximadamente 20 genes putativos de defensa fueron identificados, de los cuales 10 fueron cuantificados con RT-PCR. Genes como metallothionein, thaumatin y los genes *PsemI*, relacionados con la patogénesis, fueron encontrados diferencialmente regulados. Este estudio es el primer paso en elucidar el transcriptoma del palto, así como de identificar respuestas hospedero-específicas en palto y un patógeno de raíces.

Introduction

Avocado (*Persea americana* Mill.) is an important agricultural crop in over 50 countries worldwide and is native to the western hemisphere. *P. americana* is a diploid angiosperm consisting of 24 chromosomes with approximately 8.83 x 10⁸ base pairs (bp). To date the avocado genome has not been published and a limited number of ESTs (8738) have been sequenced, annotated and released on the NCBI.

*Phytophthora* root rot (PRR), is caused by *Phytophthora cinnamomi* and considered to be the most destructive threat to the avocado industry (Coffey, 1987; Zentmyer, 1984; Zentmyer, 1955) with production relying heavily on the use of chemical control and tolerant rootstocks. Methods of control include the use of tolerant rootstocks such as Dusa® and phosphite trunk injections (Coffey, 1987; Giblin *et al.*, 2005), supported by planting in high organic matter soils and mulching to promote antagonistic microbial growth against *P. cinnamomi*. Phosphite trunk injections were developed in the...
early 1980s as a chemical means of controlling PRR and currently the use of phosphite fungicides has been shown to provide a measure of control (Coffey, 1987; Giblin et al., 2005; Kaiser et al., 1997), however it has recently been indicated that *P. cinnamomi* can develop resistance to phosphite after prolonged usage of the fungicide (Dobrowolski et al., 2008).

Despite the importance of avocado and a 60 year attempt to unravel the host pathogen interaction PRR is still the most limiting factor in avocado production. Currently the information surrounding PRR and understanding resistance to *P. cinnamomi* involves analysis of root exudates (Botha & Kotze, 1989), chemical analysis of roots (Sánchez-Pérez et al., 2009), the application of chemicals to aid in suppression of the pathogen (Bekker et al., 2006) and biochemical studies (García-Pineda et al., 2010). There has also been histological studies performed on roots infected with *P. cinnamomi* (Phillips et al., 1987). The attraction of *P. cinnamomi* zoospores was investigated by Botha and Kotze in 1989 and it was found to be influenced by the composition of amino acids in root exudates.

Genetic information available for avocado is limited. To date molecular research has comprised of genetic relationship studies and the molecular characterization of the fruit and flowers. There has been some gene characterization of avocado fruit ripening genes (Cass et al., 1990; Chernys & Zeevaart, 2000; Christoffersen et al., 1984; Hammond-Kosack & Jones, 1997; Kupke et al., 2003). The greater part of molecular detail exists due to a continuous effort in marker development to assist in either elucidating genetic relationships amongst scions (Acheampong et al., 2008; Chanderbali et al., 2008; Chang et al., 1993; Clegg & Davis, 1989; Davis et al., 1998; Mhameed et al., 1997), or scion improvement (Chen et al., 2007; Clegg et al., 1992; Clegg et al., 1999; Lavi et al., 1991) and there is currently a preliminary genetic map available based on microsatellites, random amplified polymorphic DNA (RAPD) markers and DNA fingerprint (DFP) markers (Sharon et al., 1997).

The most recent molecular development in the fight against PRR was the identification of 70 microsatellite markers that were developed from over 8000 ESTs (Expressed sequence tags) in the hope of aiding in marker assisted breeding against PRR. ESTs were previously generated from fruiting and flowering organs of avocado with a lack of EST data available for avocado rootstocks, more specifically a lack of information on the avocado/*P. cinnamomi* interaction.

The aim of this study was to sequence ESTs of a tolerant avocado rootstock infected with *Phytophthora cinnamomi* using a high-throughput sequencing platform. The 454 platform was used to generate sequence data of a 6-12 hour and 24-72 hour plant response and to identify transcripts that were associated with the defence response. We identified 371 transcripts from avocado and quantified the gene expression of a selection of these ESTs, thereby providing the first molecular data for the avocado/*P. cinnamomi* interaction.

**Materials and methods**

**Plant material inoculation**

Nine month old tolerant R0.09 avocado (*Persea americana*) clonal plantlets were provided by Westfalia Technological services (Tzaneen, South Africa) and inoculated with *Phytophthora cinnamomi* mycelia. A concentration of 0.5 g/L mycelia was was then mixed into 112 kg of vermiculite in a mistbed. Plantlets were randomly grounded in vermiculite and constantly irrigated over a period of six weeks. Root material was harvested at 0 hour (uninfected), 3, 6, 12, 24, 48 and 72 hours post infection (hpi).

**RNA isolation and cDNA synthesis**

RNA isolations were done using the CTAB method (Chang et al., 1993). Three technical replicates per biological replicate were performed for RNA isolation.

Prior to mRNA isolation, different time points were combined into three libraries for pyrosequencing. The 0 h time point was regarded as the uninfected library, the 1-12 hour infection library composed of 6, and 12 h, while the 24-72 hour infection library comprised of 24, 48 and 72 h samples. Purification of mRNA was done according to manufacturer’s instructions using Oligotex (Oligotex™ mRNA kit, Qiagen, Valencia, California, USA).

DNA contamination was assessed by using intron flanking primers. F3H forward: (5'-TCTGATTTCGAGATCGACTCGC-3') and F3H reverse: (5'- TTAGACTTGGGCCACCCTGTTT-3')
primers were used to amplify a 300 bp fragment of the flavanone-3-hydroxylase gene from RNA as opposed to the 1200 bp fragment which is obtained from DNA.

cDNA libraries were synthesized with the Roche cDNA synthesis system (Roche Diagnostics, Mannheim, Germany) according to manufacturer’s instructions. cDNA libraries were purified using the Qiagen MinElute PCR Purification Kit (Qiagen, Valencia, California, USA) before being sequenced.

**Pyrosequencing and Bioinformatics**

Libraries were sequenced by Inqaba Biotec (Sunnyside, South Africa) on the GS-FLX platform. Approximately 3 µg of cDNA was supplied for each library. A third of each library was tagged with a different ten nucleotide tag (Uninfected tag- 5’CGTGTCTCTA’3, 1-12 hour infection tag- 5’CTCGCGTGTC’3, 24-72 hour infection tag- 5’TAGTATCAGC’3) and sequenced on a single lane.

Contigs were assembled using the Newbler assembler version 1.1.02.15 (Roche). Reads were trimmed before contig assembly. Low quality reads were not included and the assembly was analyzed and annotated using dCAS (Desktop cDNA Annotation System) Version 1.4.1 Build 3791 and CLC Free Workbench software (CLC bio, Cambridge, MA). The BLASTX tool was used (using the PAM 30 matrix) in order to produce short and nearly exact matches.

**Validation of sequencing data with quantitative RT-PCR**

Defence related genes that showed differential regulation between the cDNA libraries were selected for quantification using qRT-PCR at the 0, 3, 6, 12, 24 and 48 hour time points. Selected defence genes were: thaumatin, thaumatin-like product, metallothionein-like product, leucine rich repeat resistance protein-like product, a PR10 product- PsemI, putative universal stress product, profilin-like product, seven transmembrane protein MLO and cytochrome P450-like TBP (TATA box binding protein). Quantitative PCR was carried out using the Bio-Rad CFX96 real-time PCR detection system (Bio-Rad Laboratories, Hercules, CA). Statistical analysis was performed using the JMP® program version 9.0.0 (SAS Institute, Inc., Cary, NC). A student t test was carried out to determine significant differences between expression levels.

**Results**

**454 pyrosequencing and EST generation**

Three cDNA (uninfected, early and late) libraries were sequenced on a single lane on the GS FLX platform and generated a total of 2 Mb of data (after trimming and quality control) consisting of 9953 reads resulting in the assembly of 371 contigs. These contigs comprised of 1407 reads from the uninfected, 3584 reads from the 1-12 hour infection and 4962 reads from the 24-72 hour infection. The average read lengths for the libraries were 216.4 bp for uninfected, 217.5 bp for the 1-12 hour library and 215.9 bp for the 24-72 hour library (Figure 1).

**EST identification and classifications**

After analysis using the dCAS software, 367 novel ESTs were identified for *Persea americana*. The program used BLASTX amino acid comparisons to screen for homology of the contigs against the NCBI non-redundant (NR) database. A large proportion of the sequence generated could be associated with a hypothetical function based on protein sequence, 45 of the 371 contigs had no similarity to previously annotated sequence (Table 1). Of the 371 contigs identified, only two sequences showed homology to previously identified fructose- bisphosphate aldolase and metallothionein type-II protein from avocado with the remaining 369 having not been previously identified. Manual BLAST annotation did not influence transcript identification.

Contigs were grouped into functional classes according to the GO (Gene Ontology) and KOG (Eukaryotic Orthologous Groups) databases. Nine percent of contigs were grouped into the unknown functional class in the KOG database while 44.5 % of contigs from the GO classification were represented by unknown functions (Figure 2). The categories of post-translational modification; translation, ribosomal structure and biogenesis; signal transduction mechanisms and general function prediction contained a combined total of 34.8 % of all contigs. Over 20 putative defence related genes have been identified although this is not reflected in the GO or KOG classifications. Defence-related genes ranging from general defence-related genes (metallothioneins, thaumatin and universal stress...
proteins) to more specific oomycete defence-related genes (pathogenesis related protein PR10, oxysterol binding protein) have been identified.

**Species similarity between avocado and other plants**

We have observed significant sequence homology between *Vitis vinifera* (grape) and avocado when the species origin of the sequence similarity was investigated. A general trend was seen over all three libraries that a large proportion of the sequence showed homology to *V. vinifera* sequences on the NCBI. When further investigated, it was found that the grape vine featured in the top ten hits of every contig that was annotated (Figure 3). As can be seen majority of the species similarities relate to a large variety of plants that have been collectively categorised as other. Other defined species such as *Arabidopsis*, *Oryza sativa* etc were included due to the amount of sequence data available in the NCBI database of these organisms. The lack of similarities to avocado highlights the lack of genetic data present for the crop. The interpretation of the sequence similarity between these two organisms will have to be studied further to understand if it holds any significance.

**Validation of sequencing data with quantitative RT-PCR**

The sequencing data was validated by eight of the nine selected sequences. The thaumatin, thaumatin-like, metallothionein-like protein, leucine-rich repeat resistance protein-like product, seven transmembrane protein MLO and cytochrome P450-like TBP (TATA box binding protein) genes all showed the highest up-regulation from a time point belonging to the cDNA library of its identification. The PR10 product- PsemI was up-regulated in both infected libraries while the putative universal stress product and profilin-like product showed the highest expression at a time point different to that of which it was identified.

**Table 1 Contig classification for cDNA libraries of *P. cinnamomi* infected avocado roots.** Contigs were classed as unidentified, identified or hypothetically identified if the sequence homology search revealed that there was no similarity, significant similarity or inferred structural function respectively.

<table>
<thead>
<tr>
<th>NR data BLAST</th>
<th>Uninfected</th>
<th>1-12 hour</th>
<th>24-72 hour</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidentified</td>
<td>5</td>
<td>16</td>
<td>24</td>
<td>45</td>
</tr>
<tr>
<td>Hypothetical protein</td>
<td>23</td>
<td>54</td>
<td>75</td>
<td>152</td>
</tr>
<tr>
<td>Genes identified</td>
<td>15</td>
<td>69</td>
<td>89</td>
<td>173</td>
</tr>
<tr>
<td>Total contigs</td>
<td>43</td>
<td>139</td>
<td>189</td>
<td></td>
</tr>
</tbody>
</table>

**Discussion**

We sequenced the first set of transcriptomic data for the *avocado*/*Phytophthora cinnamomi* interaction. A single lane of pyrosequencing on the GS FLX platform generated 2.0 Mb (of a potential 2.5 Mb) of data, consisting of 9953 reads that assembled into 371 contigs. In addition to identifying ESTs we were also able to identify ten defence-related genes. An important objective of this study was to generate EST data of an avocado rootstock infected with *P. cinnamomi*. This data identified the genes involved in cellular processes and defence mechanisms thereby providing the first platform for studying the important avocado root/*P. cinnamomi* interaction on a molecular level.
The 371 contigs were grouped into 38 and 21 functional classes based on the KOG and GO databases respectively. The majority of sequences however had unknown functions. Due to the depth of sequencing achievable, transcriptome studies identify many transcripts that are not yet characterised and many that have unknown functions even when annotated using a database such as Gene ontology (Coram et al., 2008). The top 10 mostly populated functional groupings according GO classification revealed that 44.5% of assembled contigs were represented by unknown functions followed by functional groups of other, cellular component, biological process, stress response, ribosome structure, cell wall related, protein binding, mitochondrion and ATP binding. According to the KOG database much of sequence data matched categories of general function prediction only, meaning that transcripts were not unequivocally assigned to a certain group. The KOG database revealed that the top ten classes that the contigs grouped into started with general function prediction followed by signal transduction, unknown function, translation and ribosomal structure, chaperones, carbohydrate metabolism, intracellular trafficking, transcription, cytoskeleton and inorganic ion transport and metabolism. Furthermore, the presence of unidentified reads in this study is not isolated and other studies have also produced sequence that did not align to any sequence present in NCBI datasets (Klionsky et al., 1984; Weber et al., 2007).

Pyrosequencing generated 9953 reads with an average read length of just under the platform threshold of optimal 250 bp read lengths. The long read lengths generated along with their ease of assembly and library preparation make pyrosequencing the technology of choice for the de novo transcriptome assembly of non-model organisms (Collins et al., 2008; Meyer et al., 2009). This relatively large dataset was easily generated due to the efficiency of the high-throughput sequencing to produce large amounts of ESTs (Vega-Arreguin et al., 2009). The ESTs generated for avocado in this study have already proved to be useful by providing transcript data to address other important questions. These ESTs are valuable for use in gene discovery for an organism that has very no genome data available.

We investigated the sequence homology between the avocado sequence data generated and other plant species. 20-30% of the contigs generated show similarities to the grapevine (Vitis vinifera) when the BLAST sequence homologies were analysed. The large amount of sequence homology observed for these two species raises questions regarding species divergence and evolution as well as that of functional gene conservation across species. The sequence similarities were distributed equally over all contig libraries and were not restricted to any specific group or functional class of transcript, thus precluding the sole homology of highly conserved genes.

Figure 4 Read length distributions of uninfected, 1-12 hour and 24-72 hour infection cDNA libraries from P. cinnamomi infected avocado roots. (a) Average read length for all three libraries was 245 bp.
The largest number of classifications belongs to the classes of general function prediction and unknown functions; this is followed by signal transduction mechanisms and translational mechanisms.
It has previously been indicated that a correlation exists amongst transcript abundance in 454 data and qRT-PCR analysis (Klionsky et al., 1984) and the quantification of genes using qRT-PCR serves as a validation method for pyrosequencing data (Coram et al., 2008). The defence genes were chosen to represent a broad range of defence responses, either Phytophthora specific (thaumatin, PR-10, LRR resistance protein-like protein), general defence-related transcripts (universal stress related protein, metallothionein, thaumatin-like protein) and general function (cytochrome P450, profilin-like protein, MLO protein). The expression data validates the high-throughput sequence data. Eight of the nine selected sequences showed congruency between the two methods indicated by an up-regulation at time points specific to the library in which the transcript showed the highest sequence representation. This indicates the use of the platform for expression profiling.

The chosen method performed well albeit a few pitfalls regarding the experimental design. Although average concentrations of RNA were above 1000 ng/µl but after mRNA isolation this value was reduced significantly, this factor needs to be considered when planning the number of RNA extractions and could also be partially due to the mRNA isolation step being repeated. The amount of data that can be produced using high-throughput sequencing minimises the need for other methods to enrich for low copy number transcripts (Simon et al., 2009) but normalization would yield higher unique transcript identification (Weber et al., 2007) and a combination of different methods of sequencing can be employed (time and cost dependent) to generate more depth and better assembly of the data. Since our aim was to investigate general transcript expression and generate ESTs from the transcriptome of an avocado rootstock in response to Phytophthora cinnamomi and not identify only differentially expressed genes, it was chosen to sequence without normalization or SSH (selective subtractive hybridization), because SSH enriches for a specific set of genes (Diatanchenko et al., 1996). The use of SSH to enrich for differentially expressed genes may in future be used less frequently due to the large amounts of sequence data generated by high-throughput platforms or be used in conjunction with the high-throughput sequencing platforms to generate more complete differential libraries for organisms (Sánchez-Pérez et al., 2009).

In conclusion we have characterised/identified 371 ESTs from a tolerant avocado rootstock that will form a basis for further transcriptomic work and gene expression studies to be done. It will help in the elucidation of the transcriptome, identification of markers for improved rootstock breeding and screening and aid in making microarray analysis for avocado root expression a reality.

The gene discovery initiated by this study will allow biochemical studies such as those by García-Pineda et al (2010) to be carried out and is also important for the development of the rootstock breeding, which has been indicated as a key aspect of improving and sustaining the avocado industry (Chang et al., 1993). The defence genes identified in this study can be studied further in other rootstocks and is the beginning of identifying the genes that play a role in the host-pathogen interaction.

The data generated from de novo pyrosequencing of avocado root could also be used in marker development as demonstrated in Meyer et al. (2009). One of the transcripts (pathogenesis-related protein Psem1) already shows potential as a marker according to observations in other studies (Ekramoddoulah et al., 2000). The thaumatin transcript is another plausible marker judging from its expression profile with the thaumatin II gene increasing partial resistance to P. infestans in transgenic potato (Filippov et al., 2005). This research is critical in understanding the infection and response processes in the interaction because it cannot be conclusively answered by studies on non-hosts (Rookes et al., 2008).

Acknowledgments

We are grateful for the support by Westfalia Technological Services for providing the plantlets. This work was supported by funding from the NRF and Hans Merensky Foundation.

References

(http://www.indexmundi.com)


