Assembly, annotation and polymorphism analysis of a draft transcriptome sequence for a fast-growing *Eucalyptus* plantation tree

by

Charles Amadeus Hefer

Submitted in partial fulfillment of the requirements for the degree

*Philosophiae Doctor*

in the

Bioinformatics and Computational Biology Unit
Department of Biochemistry
Faculty of Natural and Agricultural Sciences
University of Pretoria

Pretoria

2011

© University of Pretoria
I, Charles Amadeus Hefer, declare that the thesis, which I hereby submit for the degree PhD(Bioinformatics) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: __________________      22 July 2011
Acknowledgements

- My supervisor, Prof F. Joubert from the Bioinformatics and Computational Biology Unit, and co-supervisor, Prof A.A. Myburg from the Department of Genetics for providing me with the required support to complete this study.
- Mr E. Mizrachi and Mr M. Ranik for collecting the biological material used in this study, and the hours of discussions we had to make sense of the results.
- The National Bioinformatics Network (NBN), the National Research Foundation (NRF) and the University of Pretoria for financial support.
- The South African Pulp and Paper Industry (Sappi) and Mondi group for financial support through Prof Myburg’s Forest Molecular Genetics group, awarded to me.
- DELL computers (SA), for graciously lending us the use of a computer with sufficient RAM to test various assembly algorithms with.
- Illumina technical support, for evaluating the development of GoldenGate and Infinium SNP arrays.
- Prof Jasper Rees for several hours of discussions during my first introduction to high throughput sequence data.
- Prof Shawn Mansfield for hosting me at the University of British Columbia for a period of five months in 2010.
- My fellow students at the Bioinformatics and Computational Biology Unit for the hours of insightfull discussions, especially Oliver, Nanette and Gordon.
- To my parents and brothers. Thank you for always supporting me.
Summary

Ultra-high throughput DNA sequencing technologies have rapidly changed the face of genomic research projects. Technologies such as mRNA-Seq have the potential to rapidly profile the expressed gene-catalog of non-model organisms, albeit with significant bioinformatics related costs and support required. This study developed automated data analysis workflows focused on the quality evaluation of mRNA-Seq reads, de novo transcriptome assembly, transcriptome annotation and digital gene expression profiling making use of data analysis tools available in the public domain and novel tools developed for this purpose. The developed workflows were made available in a private instance of the Galaxy workflow management system. The developed workflows were used to perform the de novo assembly of a gene-catalog of a Eucalyptus plantation tree. The fast growing and good wood properties of Eucalyptus tree species and their hybrids make them excellent renewable resources of fiber for pulp and paper, and woody biomass for bioenergy production. We produced an expressed gene-catalog of 18 894 de novo assembled contigs from Illumina deep mRNA-Seq of six sampled plant tissues. Using a novel coverage-assisted re-assembly approach, we were able to assemble near full-length biologically relevant transcripts. The assembly was evaluated in terms of contig quality and contiguity, and functional annotations were assigned. Digital expression profiling (FPKM values) of each contig across the tissues were calculated, which was used to identify of tissue-specific sets of expressed genes. Polymorphism analysis of 13 806 high-confidence contigs revealed a combined exon and untranslated region SNP density of 0.534 SNPs/100 bp, which provides a good opportunity for designing high-density SNP assays in the expressed regions of the Eucalyptus genome. The assembled and annotated gene catalog was made available for public use in a user-friendly, web-based interface as the Eucpresso database (http://eucpresso.bi.up.ac.za). The
developed database acts as a prelude to a more comprehensive mRNA-Seq whole-transcriptome repository, the *Eucalyptus* Genome Intergrative Explorer (EucGenIE), a resource that will focus on identifying transcriptional networks active during woody biomass development. Results from the study proved that current bioinformatics software tools and approaches can be used to successfully assemble and characterise a large proportion of the transcriptome of a complex eukaryotic organism. This approach can be used to characterise the gene catalog of a wide range of non-model organisms using only data derived from uHTS experiments.
Contents

Acknowledgements .................................................. i

List of Figures ....................................................... vi

List of Tables ......................................................... ix

List of Abbreviations ................................................ x

Lexicographical conventions ........................................ xiii

Chapter 1. An introduction to ultra-high-throughput DNA sequencing technologies and their application in genetics and functional genomics ........................ 1

1.1. Introduction ....................................................... 1

1.2. Ultra-high-throughput DNA sequencing platforms ........................................ 4
   1.2.1. Cyclic array sequencing applications ........................................ 4
   1.2.2. Single-molecule sequencing platforms ....................................... 10

1.3. High-throughput DNA sequencing applications in genetics and functional genomics .... 14
   De novo genome sequencing ..................................................... 15
   Genome re-sequencing and variant discovery ................................... 16
   Transcriptome sequencing .......................................................... 19

1.4. Core analyses associated with ultra-high-throughput Illumina sequence mRNA-Seq data ... 25

1.5. High-throughput DNA sequencing data management ........................................ 34
   1.5.1. Widely-used bioinformatics workflow systems ................................ 35

1.6. Problem Statement ................................................... 39

1.7. Specific research questions and aims ............................................ 40
Chapter 2. A core bioinformatics workflow environment for ultra-high-throughput transcriptome data analysis

Chapter preface ............................................................. 41

2.1. Introduction ......................................................... 42

2.2. Materials and methods ............................................. 44
  2.2.1. BCBU Galaxy: Extending the public Galaxy framework .......... 44
  2.2.2. Illumina short-read base-quality evaluation workflow .......... 45
  2.2.3. De novo transcriptome assembly workflow ....................... 45
  2.2.4. Annotation of predicted protein sequences workflow .......... 48
  2.2.5. Expression profiling using Illumina mRNA-Seq short reads workflow .......... 48

2.3. Results and discussion ............................................ 49
  2.3.1. Extending the Galaxy framework .................................. 49
  2.3.2. Quality assessment of Illumina short-reads ...................... 53
  2.3.3. De novo transcriptome assembly using Illumina mRNA-Seq data .... 56
  2.3.4. Annotating assembled transcript sequences ..................... 65
  2.3.5. Using mRNA-Seq data to calculate transcript expressions values ...... 73

2.4. Conclusion .................................................................. 76

Chapter 3. The assembly and annotation of a draft transcriptome sequence of a Eucalyptus hybrid tree

Chapter Preface ............................................................. 81

3.1. Introduction ........................................................... 82

3.2. Materials and methods .............................................. 83
  3.2.1. Plant tissue collection, mRNA-Seq library preparation and sequence generation ........ 83
  3.2.2. De novo transcriptome assembly ................................... 84
  3.2.3. Prediction of coding sequences .................................... 86
  3.2.4. Inspecting contig contiguity ....................................... 87
  3.2.5. Homology searches ................................................. 88
  3.2.6. InterProScan ....................................................... 88
3.2.7. Calculating transcript coverage and expression ................................. 89
3.2.8. Single nucleotide polymorphism detection ............................. 90
3.3. Results ................................................................. 90
3.3.1. Assembly ............................................................. 90
3.3.2. Prediction of coding sequences ........................................ 95
3.3.3. Inspecting contig contiguity .............................................. 97
3.3.4. Homology searches .................................................. 102
3.3.5. InterProScan ......................................................... 102
3.3.6. Expression profiling .................................................. 104
3.3.7. Single nucleotide polymorphism (SNP) detection .................... 116
3.4. Discussion ............................................................... 116
3.5. Conclusion ............................................................... 121

Chapter 4. Eucspresso: Towards the development of a *Eucalyptus* genome and transcriptome

information resource ................................................................. 122

Preface ....................................................................................... 122
4.1. Introduction ........................................................................ 123
4.2. Materials and methods .......................................................... 124
4.2.1. MySQL database .......................................................... 124
4.2.2. TurboGears Web framework .......................................... 124
4.2.3. Custom Python controllers and R scripts............................... 125
4.3. Results and discussion ............................................................. 125
4.3.1. Eucspresso data model ..................................................... 125
4.3.2. Browsing and searching for a contig .................................... 126
4.3.3. Visualising a contig and associated annotation ................. 126
4.3.4. Search interface ............................................................ 136
4.4. Conclusion .......................................................................... 136

Chapter 5. Concluding Discussion ...................................................... 141

Summary ..................................................................................... 147
Appendix A. Bioinformatics workflow ........................................... 149

Appendix B. Extendinator .......................................................... 150

Appendix C. Transcriptome assembly ........................................... 151

C.1. Evaluating contig contiguity of the assembled transcript sequences . . . . . . . . . . . . . . . . . . . 151
  C.1.1. Full length Eucalyptus cDNA sequences ........................................ 151
  C.1.2. Alignment coverage graphs of the 33 full length cDNA sequences and assembled contigs . 155
  C.1.3. Alignment of contig 68291 before and after extension ..................... 156

Appendix D. De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree
  produced by Illumina mRNA-Seq ................................................ 157

Bibliography ................................................................. 158
List of Figures

1.1 An example of an Illumina FASTQ formatted mRNA-Seq file ............... 27

2.1 An example of code developed to extend the Galaxy framework with the "shuffleseq" tool. .... 51
2.2 The interface of the FASTQ shuffleseq tool described in the fastq_shuffleseq.xml file, as rendered by Galaxy. .......................................................... 52
2.3 The Illumina read quality assessment pipeline ........................................ 54
2.4 An example of FASTQ quality scores obtained from a 76 bp Illumina GAII paired-end run ...... 57
2.5 A Galaxy workflow which performs a de novo assembly with the Velvet assembler .......... 58
2.6 The assembly scoring function is a robust measure to select the kmer of the best Velvet assembly. 63
2.7 The effect of the expected coverage and the coverage cutoff parameters on a Velvet assembly .... 66
2.8 Alignment of the six full length CesA cDNA sequences against an assembly with a kmer size of 41
   (k41). .................................................................................. 67
2.9 Alignment of the six full length CesA cDNA sequences against an assembly with a kmer size of 51
   (k51). .................................................................................. 68
2.10 Alignment of the six full length CesA cDNA sequences against an assembly with a kmer size of 61
    (k61). .................................................................................. 69
2.11 The automated annotation pipeline developed from tools available in Galaxy .................... 70
2.12 The 25 most prevalent protein family domains annotated in the assembled transcriptome dataset,
    expressed as a fraction of the total number of Pfam annotations ....................... 72
2.13 Protein features annotated by InterProScan present on the cellulose synthase 6 (CesA6) protein
    sequence assembled from reads derived from mRNA-Seq sequencing .................. 73
2.14 Calculating gene expression (FPKM) values for unigene aligned regions from a genome with no gene
    models available ................................................................... 74
2.15 A breakdown of the number of reads which map uniquely, and non-uniquely as pairs or single reads to a target genome for difference read lengths. .............................................. 75

2.16 Genes identified as differentially expressed in immature xylem and young leaf tissues of a *Eucalyptus grandis* hybrid tree. ................................................................. 77

3.1 A schematic flow diagram of the coverage-assisted re-assembly process. ........................................ 85
3.2 Identifying the optimal kmer used for the *de novo* assembly of the *Eucalyptus* transcriptome. .... 91
3.3 Identifying the optimal expected coverage value to use for the *de novo* assembly of the *Eucalyptus* transcriptome. ................................................................. 92
3.4 The number of bases per contig added during the extension of the assembly ................................. 93
3.5 The effect of performing a coverage assisted re-assembly on a single contig. ............................... 94
3.6 The alignment of contig_68291 before and after extension ..................................................... 96
3.7 Alignment of the full length cDNA sequence AF197329.1, the assembled contig_5550, and the predicted coding sequence. ...................................................... 99
3.8 Alignment of the protein coding sequence of contig_5550 and the full length cDNA sequence AF197329.1 .................................................................................. 100
3.9 Alignment coverage figure of the full length cDNA sequence AF197329.1, the assembled homologous contig, the predicted CDS and the OASES assembled transcripts. ................. 101
3.10 Similarity search results of the assembled *Eucalyptus* transcripts against three angiosperm species. 104
3.11 The 20 most prevalent protein family (PFAM) and protein information resource (PIR) annotations from InterProScan analysis. ........................................ 105
3.12 The 20 most prevalent Panther and Prosite annotations from InterProScan analysis. ............... 106
3.13 Identifying over-expressed xylogenic and non-xylogenic genes ............................................. 107
3.14 Over-represented molecular function gene ontology terms of genes over-expressed in xylogenic and photosynthetic tissues ................................................. 109
3.15 Over-represented biological process gene ontology terms of genes over-expressed in xylogenic and photosynthetic tissues ................................................. 110
3.16 Over-represented cellular component gene ontology terms of genes over-expressed in xylogenic and photosynthetic tissues ................................................. 111
3.17 Differential gene expression between the xylogenic and photosynthetic genes represented on the starch and sugar metabolism KEGG pathway ......................................................... 112

3.18 Differential gene expression between the xylogenic and photosynthetic genes represented on the photosynthesis KEGG pathway ................................................................. 113

3.19 Selection of high quality, high confidence contigs for polymorphism detection .................................. 117

4.1 Entity relationship diagram of the main datatypes in *Eucspresso* .................................................. 127

4.2 Browsing and searching for contigs through the *Eucspresso* web interface. .............................. 128

4.3 Contig summary and sequence detail tab for contig_31, the assembled cellulose synthase IRX3 gene. 129

4.4 The homology search results of the contig against a set of selected angiosperm transcriptomes, and a summary of the GO category that the sequence is associated with ........................................ 131

4.5 Gene ontology annotations for contig_31, the assembled cellulose synthase IRX3 gene. .............. 132

4.6 The cellulose synthase enzyme (EC:2.4.1.12) is highlighted on the starch and sucrose metabolism KEGG map ................................................................. 133

4.7 The InterProScan results tab describing protein features found on the predicted protein sequence (contig_31). .................................................................................. 134

4.8 The FPKM expression values of contig_31, a secondary cell wall synthesis gene (cellulose synthase, IRX3). .................................................................................. 135

4.9 The *Eucspresso* GBrowse instance, indicating the position of contig_31 (IRX3) on the 8X *Eucalyptus* draft sequence. ......................................................... 137

4.10 The Eucspresso search interface .............................................................................................. 138
## List of Tables

1.1 A selected list of short read sequence alignment tools currently available for academic use.  

2.1 Third party applications that were added to the BCBU Galaxy server instance.  

2.3 A list of tools newly developed to complement the existing tools available in the BCBU Galaxy server.  

2.5 The theoretical and usable base (bases identified as A, G, C and T) yield for six Illumina GA IIx 76 bp paired-end lanes.  

2.6 Velvet assembly statistics for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue trimmed to different lengths.  

2.7 Statistics for Velvet assembled contigs with a minimum contig length of 200 bp for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue trimmed to different lengths.  

2.8 Velvet assembly statistics for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue.  

3.1 Comparing the assembled Velvet dataset before and after the coverage assisted extension.  

3.2 Coding sequences predicted in the assembled dataset with different *ab initio* gene prediction software packages.  

3.3 A summary of the representation of *Arabidopsis*, *Populus* and *Vitis* genes in the constructed public dataset (*EucAll*), and the assembled contig dataset at different e-value thresholds.  

3.4 The top 30 genes identified in the xylogenic tissues, compared to photosynthetic tissues.  

3.5 Top 30 photosynthetic genes identified as over-expressed in photosynthetic tissue compared to xylogenic tissue.  

A.1 Velvet assembly statistics of contig longer than 1,000 bp for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue trimmed to different lengths.  

## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine nucleotide base</td>
</tr>
<tr>
<td>AGBT</td>
<td>Advances in Genome Biology and Technology meeting</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>ASCII</td>
<td>American Standard Code for Information Interchange</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial Artificial Clone</td>
</tr>
<tr>
<td>BDB</td>
<td>Berkeley Database</td>
</tr>
<tr>
<td>BTA</td>
<td>Benzene-1,3,5-Triacetic Acid</td>
</tr>
<tr>
<td>BWT</td>
<td>Burrows-Wheeler Transform</td>
</tr>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine nucleotide base</td>
</tr>
<tr>
<td>caBIG</td>
<td>cancer Biomedical Informatics Grid</td>
</tr>
<tr>
<td>CBP</td>
<td>Coverage per Base Pair</td>
</tr>
<tr>
<td>CCD</td>
<td>Charged Coupled Device</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding DNA Sequence</td>
</tr>
<tr>
<td>contig</td>
<td>A multiple alignment of reads, which is converted into contiguous genomic sequence</td>
</tr>
<tr>
<td>cPAL</td>
<td>combinatorial Probe Anchor Ligation</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DOE</td>
<td>Department of Energy</td>
</tr>
<tr>
<td>DWAF</td>
<td>Department of Water Affairs and Forestry</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed sequence tag(s)</td>
</tr>
</tbody>
</table>
G Guanine nucleotide base
GB Gigabyte(s), or 1 073 741 842 bytes
Gbp Gigabase(s) pair, or 1 000 000 000 nucleotide bases
GUI Graphical User Interface
GWAS Genome-Wide Association Studies
ha Hectares
HMM Hidden Markov Model
Indel Insertion/deletion of a base in a sequence
JGI Joint Genome Institute
kmer A word size, of length k. Used by de Bruijn graph assemblers
MAS Marker Assisted Selection
MB Megabyte(s) or 1 048 576 bytes
Mbp Megabasepair(s) or 1 000 000 nucleotide bases
miRNA micro RNA
MRSA Multiple Resistance Staphylococcus aureus
mRNA messenger Ribonucleic Acid
N Used to represent the total number of sequences or contigs in an assembly
NGS Next-generation sequence(ing) technologies, includes the 454 Sequencer from Roche, Illumina's GA sequencers and ABI's SOLiD system
N50 The length where 50% of the bases in an assembly occurs in contigs longer than this number
PCR Polymerase Chain Reaction
PIR Protein Information Resource
PPT Pentatricopeptide
read(s) Refer to a DNA string of base pairs
RNA Ribonucleic Acid
RDBMS Relational Database Management System
RPKM  Reads Per Kilobase of exon Per Million mapped sequenced reads
RUST  Regulated Unproductive Splicing and Translation
Scufl  Simplified Conceptual Workflow Language
SGS   Second Generation Sequencers, see NGS
SMRT™ Single Molecule Real Time
SMRTbell™ A circular DNA template for SMRT™ sequencing
SNP   Single Nucleotide Polymorphism
snRNA small nuclear RNA
ssRNA strand-specific RNA
T     Thymine nucleotide base
TAIR  The Arabidopsis Information Resource
TGS   Third Generation Sequencers, refers to single molecule sequencers
TIGR  The Institute for Genomic Research
TSS   Transcriptional start site
uHTS  Ultra-High-Throughput DNA Sequencing, includes NGS, SGS and TGS
UTR   Untranslated region(s)
US-DOE United States Department of Energy
WGS   Whole Genome Sequencing
ZMW   Zero-mode waveguide used in SMRT™ sequencing
Lexicographical conventions

- *Short-reads* refers to reads from the Illumina GAII analyser, *pairs* refer to the forward and reverse sequences from the Illumina Paired End protocol.

- The names of software packages are indicated by the **TYPEWRITER** font, and are all in capital letters unless general naming convention dictates the use of **CamelCase** or lower case letters.

- Wherever there is a reference to a technology-sequence type, for instance Sanger sequence or Illumina sequence, or 454 sequence, it refers to a sequence generated from that specified technology. This also holds true for reference to a technology, i.e. there will be references to 454, which refers to the technology behind the Roche 454 sequencing platform.

- The SMRT™ and SMRTbell™ trademarks are registered by Pacific Biosciences.

- In this document, the term "ultra-high-throughput sequencing technologies" (uHTS) is used interchangeable with the the collective term for the so called Next-Generation (NGS) or Second-Generation (SGS) DNA sequencing platforms, and includes the Third-Generation (TGS) DNA sequencing single molecule platforms.

- The complete codebase of both the Galaxy instance, and the Eucpresso datasource systems are available in a subversion repository upon request.