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

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 Brittish Columnbia 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 `Eucspresso` database (http://eucspresso.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

List of Figures

List of Tables

List of Abbreviations

A	Adenine nucleotide base
AGBT	Advances in Genome Biology and Technology meeting
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
BAC	Bacterial Artificial Clone
BDB	Berkeley Database
BTA	Benzene-1,3,5-Triacetic Acid
BWT	Burrows-Wheeler Transform
bp	base pairs
C	Cytosine nucleotide base
caBIG	cancer Biomedical Informatics Grid
CBP	Coverage per Base Pair
CCD	Charged Coupled Device
CDS	Coding DNA Sequence
contig	A multiple alignment of reads, which is converted into contiguous genomic sequence
cPAL	combinatorial Probe Anchor Ligation
DNA	Deoxyribonucleic Acid
DOE	Department of Energy
DWAF	Department of Water Affairs and Forestry
EST	Expressed sequence tag(s)

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 referes 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 `Eucspresso` datasource systems are available in a subversion repository upon request.