

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 British Columbia for a period of five months in 2010.
- My fellow students at the Bioinformatics and Computational Biology Unit for the hours of insightful 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 Integrative 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

<b>Acknowledgements</b> . . . . .	i
<b>List of Figures</b> . . . . .	vi
<b>List of Tables</b> . . . . .	ix
<b>List of Abbreviations</b> . . . . .	x
<b>Lexicographical conventions</b> . . . . .	xiii
<b>Chapter 1. An introduction to ultra-high-throughput DNA sequencing technologies and their application in genetics and functional genomics</b> . . . . .	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
<i>De novo</i> 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
	ii

<b>Chapter 2. A core bioinformatics workflow environment for ultra-high-throughput transcriptome data analysis</b>	41
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. <i>De novo</i> 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 assesment of Illumina short-reads	53
2.3.3. <i>De novo</i> 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
<b>Chapter 3. The assembly and annotation of a draft transcriptome sequence of a <i>Eucalyptus</i> hybrid tree</b>	81
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. <i>De novo</i> 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
	iii

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
<b>Chapter 4. Eucspresso: Towards the development of a <i>Eucalyptus</i> genome and transcriptome information resource . . . . .</b>	<b>122</b>
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
<b>Chapter 5. Concluding Discussion . . . . .</b>	<b>141</b>
<b>Summary . . . . .</b>	<b>147</b>

<b>Appendix A. Bioinformatics workflow</b> . . . . .	149
<b>Appendix B. Extendinator</b> . . . . .	150
<b>Appendix C. Transcriptome assembly</b> . . . . .	151
C.1. Evaluating contig contiguity of the assembled transcript sequences . . . . .	151
C.1.1. Full length <i>Eucalyptus</i> 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
<b>Appendix D. <i>De novo</i> assembled expressed gene catalog of a fast-growing <i>Eucalyptus</i> tree     produced by Illumina mRNA-Seq</b> . . . . .	157
<b>Bibliography</b> . . . . .	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 <b>Galaxy</b> framework with the "shuffleseq" tool. . . . .	51
2.2	The interface of the <b>FASTQ</b> shuffleseq tool described in the fastq_shuffleseq.xml file, as rendered by <b>Galaxy</b> . . . . .	52
2.3	The Illumina read quality assesment pipeline . . . . .	54
2.4	An example of <b>FASTQ</b> quality scores obtained from a 76 bp Illumina GAII paired-end run . . . . .	57
2.5	A <b>Galaxy</b> workflow which performs a <i>de novo</i> assembly with the Velvet assembler . . . . .	58
2.6	The assembly scoring function is a robust measure to select the kmer of the best <b>Velvet</b> assembly. . . . .	63
2.7	The effect of the expected coverage and the coverage cutoff parameters on a <b>Velvet</b> 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 <b>Galaxy</b> . . . . .	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 <i>Eucalyptus grandis</i> 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 <i>de novo</i> assembly of the <i>Eucalyptus</i> transcriptome. . . . .	91
3.3	Identifying the optimal expected coverage value to use for the <i>de novo</i> assembly of the <i>Eucalyptus</i> 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 <i>Eucalyptus</i> 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 <i>Eucspresso</i> . . . . .	127
4.2	Browsing and searching for contigs through the <i>Eucspresso</i> 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 <i>Eucspresso</i> <i>GBrowse</i> instance, indicating the position of contig_31 (IRX3) on the 8X <i>Eucalyptus</i> draft sequence. . . . .	137
4.10	The <i>Eucspresso</i> search interface . . . . .	138



## List of Tables

1.1	A selected list of short read sequence alignment tools currently available for academic use. . . . .	31
2.1	Third party applications that were added to the BCBU Galaxy server instance. . . . .	46
2.3	A list of tools newly developed to complement the existing tools available in the BCBU Galaxy server. . . . .	47
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. . . . .	55
2.6	Velvet assembly statistics for a single lane of paired 76 bp sequences from <i>Eucalyptus</i> xylem tissue trimmed to different lengths. . . . .	59
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 <i>Eucalyptus</i> xylem tissue trimmed to different lengths. . . . .	60
2.8	Velvet assembly statistics for a single lane of paired 76 bp sequences from <i>Eucalyptus</i> xylem tissue. . . . .	62
3.1	Comparing the assembled Velvet dataset before and after the coverage assisted extension. . . . .	96
3.2	Coding sequences predicted in the assembled dataset with different <i>ab initio</i> gene prediction software packages. . . . .	97
3.3	A summary of the representation of <i>Arabidopsis</i> , <i>Populus</i> and <i>Vitis</i> genes in the constructed public dataset ( <i>EucAll</i> ), and the assembled contig dataset at different e-value thresholds. . . . .	103
3.4	The top 30 genes identified in the xylogenic tissues, compared to photosynthetic tissues . . . . .	114
3.5	Top 30 photosynthetic genes identified as over-expressed in photosynthetic tissue compared to xylogenic tissue . . . . .	115
A.1	Velvet assembly statistics of contig longer than 1 000 bp for a single lane of paired 76 bp sequences from <i>Eucalyptus</i> xylem tissue trimmed to different lengths. . . . .	149

## 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 <i>de Bruijn</i> 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 <i>Staphylococcus aureus</i>
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
Scuff	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 interchangeably with 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.

## Chapter 1

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

### 1.1. Introduction

Eucalypt forest trees supply high quality raw material for the pulp, paper and wood industries, and have been identified as important role-players in the search for renewable energy resources. Eucalypts are hardy, fast growing and have a high dry matter production and resprouting potential, which makes them one of the most widely used tree species in industrial hardwood plantations (Forrest and Moore, 2008; Rengel *et al.*, 2009). In recent years, the global forestry industry has experienced a steady shift in location from the northern hemisphere to the tropics and subtropics, where it is actively competing with food crops for land space needed for expansion (Grattapaglia and Kirst, 2008). In South Africa, a recent report from the South African Department of Water Affairs and Forestry (DWAF) indicated that 1.25 million hectares (1.1%) of South Africa's total land area are covered by forestry plantations, of which 36% (450 000 hectares) are planted with *Eucalyptus* species (<http://www2.dwaf.gov.za/webapp/Documents/FSA-Abstracts2009.pdf>). The economic importance of plantation trees as renewable energy and biomass producing crops makes them excellent candidates for genetic improvement studies.

Eucalypts have a high fiber count of uniform nature, a sought after property that has created high demands in the pulp, paper and raw wood industries (Moore *et al.*, 2008). Large numbers of genes are affecting wood formation in forest trees, and have been actively investigated by various research groups focusing on key properties, such as wood density, pulp yield, cellulose content, fiber length and lignin content (for a review of the state of *Eucalyptus* breeding see Myburg *et al.*, 2005). Improvements to biomass yield and fiber quality with improved breeding programs and the direct application of biotechnology advances to crop development will play increasingly important roles in the future of the eucalypt forestry industry.

Woody biomass has been identified as important in the search for renewable energy resources. The United States Department of Energy (US-DOE) announced in 2007 their goal to reduce the usage of gasoline in the United States by 20% by the year 2017 (<http://genomicscience.energy.gov/biofuels/>). To achieve this, an expansion of the annual renewable fuel supply from a variety of plant materials, including grasses, woodchips and agricultural wastes needs to occur. The bioenergy initiative actively supported the research community in successfully determining the genomic sequence of the *Populus trichocarpa* genome (Tuskan *et al.*, 2006) and the *Eucalyptus grandis* genome (version 1.0 released in January 2011, <http://www.phytozome.net>) by the Joint Genome Institute (JGI). It is expected that fast growing, short-rotation woody crops such as *Eucalyptus* and *Populus* and their respective hybrids will contribute up to 30% of the biomass of the so-called "energy crops" (Hinchee *et al.*, 2009).

Advances in the fields of biotechnology, genetics and computer science have resulted in an unprecedented growth in the amount of biological data being generated on a daily basis by the scientific community. This aided the slow, but definitive paradigm shift from a hypothesis-driven scientific approach to a data-driven, explorative approach. Next-generation DNA sequencing technologies (NGS) have opened the floodgates in terms of biological sequence data generation. Since the first application of NGS by Margulies *et al.* (2005), various technological improvements have led to higher and higher base pair throughput from NGS platforms. As stated in the preamble of this document, the term ultra-high-throughput sequencing (uHTS) will be used in the rest of this manuscript to denote the different high throughput DNA

sequencing technologies (next generation sequencers, second generation sequencers and third generation sequencers, Werner, 2010).

High-throughput experiments now commonly investigate the range of gene expression products between different organisms, between tissues within organisms, or between tissues of the same organism in different disease states in order to investigate underlying molecular basis of a phenotype. Pyrosequencing technologies have effectively revolutionised the approach and turnover time needed to sequence and re-sequence genomes. Applications of uHTS technologies are evident in the advances made in the fields of mutation discovery, metagenomic characterisation, non-coding RNA and DNA-protein interaction discovery (Mardis, 2008). The data produced from these high-throughput experiments have resulted in a biological data glut, where gigabases of data are produced in a single experiment and biologists are now forced to design and follow efficient data management practices for experiments.

Sequencing large numbers of mRNAs from a sample forms the basis of the revolutionary expressed sequence tag method (EST) used for identifying genes during the human genome project (Adams *et al.*, 1991; Venter *et al.*, 2001). The costly nature, long experimental run time, low quality reads and general inability to detect transcripts expressed at a low level has hampered the technology from being widely used (Graveley, 2008). The parallel nature of next-generation sequencing makes it a ideal technology for transcriptome sequencing, generating hundreds of millions of short reads (35-350 base pairs (bp) long). Many research groups have employed a technology called mRNA-Seq (Section 1.3) to sequence at various levels of detail and complexity the transcriptomes of a diverse set of organisms (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mortazavi *et al.*, 2008; Novaes *et al.*, 2008; Nagalakshmi *et al.*, 2008). Transcriptome studies have revealed, among others, differences in transcript abundance, efficiency of the machinery active during intron removal and detection of alternatively spliced transcripts between different tissues and/or organisms of interest. Improvements in the technology in terms of read length, the ability to perform paired-end sequencing, strand-specific sequencing and improved algorithms to assemble short reads will provide even greater insight into the transcriptome landscape (Graveley, 2008).

The following sections will focus on the different ultra-high-throughput DNA sequencing platforms



available in the market with specific focus on the applications of these technologies to the fields of genetics and functional genomics. A brief discussion regarding the data management issues involved in working with and analysing data from these platforms is then followed by a section dedicated to defining the main problem statement of this study. The final section of the chapter includes an outline of the specific aims and requirements in order to achieve the goals of this study.

## 1.2. Ultra-high-throughput DNA sequencing platforms

Ultra-high-throughput sequencing (uHTS) technologies have been categorically assigned to one of the following groups: microelectrophoretic methods, sequencing by hybridisation, real-time observation of single molecules and cyclic-array sequencing (Shendure *et al.*, 2004). The current technological advances made with cyclic-array sequencing has proven this to be the most successful approach by far, as is evident in the implementation of this technology in various commercial products. These products, defined in the literature as Next Generation Sequencing (NGS) platforms, or more recently Second Generation Sequencing (SGS) platforms (Kislyuk *et al.*, 2005), include the 454 Genome Sequencer (Roche Applied Science, Margulies *et al.*, 2005), Solexa technology (Illumina Genome Analyser, Fedurco *et al.*, 2006; Turcatti *et al.*, 2008) and the SOLiD platform (Applied Biosystems, Shendure *et al.*, 2005). Very recently, the term of Third Generation Sequencers (TGS) emerged with the advent of single molecule sequencers (Schuster, 2008). Of these systems, the most prolific commercial offerings include the Heliscope Single Molecule Sequencer (Helicos, Braslavsky *et al.*, 2003) and the Single Molecule Real Time (SMRT) sequencing platform from Pacific Biosciences (Eid *et al.*, 2009), but the nanoball sequencing platform from Complete Genomics (Drmanac *et al.*, 2010) and the innovative Ion Torrent (unpublished) platforms are also available.

### 1.2.1. Cyclic array sequencing applications

The first practical implementations of uHTS technologies included the *de novo* sequencing and assembly of the *Mycoplasma genitalium* genome (Margulies *et al.*, 2005), and the re-sequencing of an evolved

*Escherichia coli* strain (Shendure *et al.*, 2005). Since these seminal papers were published, different applications have been developed in which high-throughput technologies were employed in various biological scenarios which will be discussed in Section 1.3. Although the different uHTS platforms use diverse DNA sequencing biochemistry and follow different methodologies in terms of array generation, a general workflow common to most technologies can be envisioned. Most cyclic-array technologies rely on the random fragmentation of a target DNA library, followed by the *in vitro* ligation of a specific set of adaptor sequences. In the case of paired-end sequencing, a so-called "jumping" library of mate-pair tags with a controllable distance between them is generated (Ng *et al.*, 2005; Shendure *et al.*, 2005). Following amplification of the target sequences on a custom array, the sequencing process is achieved by alternative cycles of flushing enzymes across a target array in order to drive a biochemical process. At every step during the sequencing process an image capture device is used to record the chemical reaction taking place at every position on the array. Various downstream computational approaches are then available to produce a string of characters with associated quality or confidence values representing the DNA sequence hybridised to the specific position on the array.

#### **454 GS FLX Pyrosequencing (Roche Applied Science)**

The 454 pyrosequencer relies on the principle of 'pyrosequencing' which employs the biochemical cleavage of a pyrophosphate molecule released during nucleotide incorporation by DNA polymerase in order to set off a chain of reactions, which will ultimately produce a burst of light from the cleavage of oxyluciferin by luciferase (Margulies *et al.*, 2005). Initially developed by 454 Life Sciences, the technology was the first widely adopted high-throughput sequencing technology and has a well-established user community. As per the general protocol, sequencing libraries are constructed that give rise to a mixture of short, adaptor-flanked fragments. These fragments are then clonally amplified with emulsion PCR inside picoliter reactors on a custom array, with amplicons captured to the surface of 28- $\mu\text{m}$  beads (Tawfik and Griffiths, 1998; Ghadessy *et al.*, 2001; Margulies *et al.*, 2005). A sequencing primer is hybridised to the universal adaptor at the appropriate position and orientation, and the pyrosequencing reaction initiated (Margulies *et al.*, 2005).

Several hundred cycles of pyrosequencing involves the inclusion of a single species of fluorescently-labeled nucleotides to the microtiter wells, and in wells where a base is incorporated, a pyrophosphate molecule is released. One reaction takes place for every base that is incorporated in the sequence, which leads to signal saturation when more than four or five bases are incorporated during homopolymer runs of the sequence (Margulies *et al.*, 2005). The nature of the technology results in asynchronous sequencing of the wells, in other words when the 'A'-base reaction takes place, multiple reactions might take place in some wells where more than one complimentary base is exposed. At the same time in wells where the template does not have a complimentary base no reaction will take place. The incorporation of bases is measured in sequence by a live capture of a charged coupled device (CCD, or camera) from the array.

At the time of writing, approximately 800 papers had been published making use of 454 pyrosequencing, including very diverse applications in metagenomics, novel and re-sequenced genomes and plasmids, population diversity determination, RNA discovery and function inferences, epigenetic studies, transcriptome studies and genome structural variant investigations (for a review on the use of high-throughput sequencing technologies in functional genomics, see Section 1.3). The GS FLX Titanium series produce between 400 and 600 million high quality bases per run with an average read length of 400 bases, which amounts to just over 100 million high quality reads per run. The long read lengths make this technology ideal for *de novo* genome sequencing projects of various organisms (<http://www.454.com>). The issue with the homopolymer run base calls is an inherent feature of the technology, and can only be overcome by employing a more sensitive light intensity detection system (Rothberg and Leamon, 2008).

### **Illumina Genome Analysis (Illumina)**

The development of the Illumina platform was derived from the initial work of Turcatti and colleagues on benzene-1,3,5-triacetic acid (BTA) and reversible deoxynucleotide terminators (Fedurco *et al.*, 2006; Turcatti *et al.*, 2008). The core methodology consists of adaptor-flanked DNA fragments of a couple of hundred base pairs that are amplified by a bridge PCR method. During this phase of the bridge PCR protocol, both forward and reverse primers are attached to a glass surface, in such a manner as to allow for the grouping of all amplified constructs from a single template in a cluster. During each

step of the bridge PCR, the reaction alternatively extends the template sequence with *Bst* polymerase and then denatures the double stranded sequence with formaldehyde (Turcatti *et al.*, 2008). After the amplification step, each cluster on the glass array should be represented by roughly 1 000 clonal amplicons, thus the initial concentration of the sequencing library needs to be known. The amplification process is highly parallelised, resulting in several million clusters amplified at distinct positions within each of the independent lanes on the array, or flow cell (Turcatti *et al.*, 2008). After cluster construction, the amplified constructs are denatured into single strands, and a sequencing primer is hybridised to the adaptor.

The sequencing process involves the single base-pair extension of the template sequence with a modified deoxynucleotide base. The deoxynucleotide base is modified in two ways; first, it is a reversible terminator, and secondly; it is fluorescently labeled to correspond to each of the four nucleotide bases. After incorporation of the modified deoxynucleotide base on the sequencing strand, chemical cleavage is needed to remove the 3' hydroxyl position, and the attached fluorescent molecule again starts a chain of reactions ending in the emission of a light signal. A CCD device captures the signal and the incorporated base is then computationally determined in downstream analysis of the images (with the Illumina analysis tools **Firecrest** and **Bustard**). The array is then prepared for the next cycle of base incorporation by enzymatically removing the blocking position of the incorporated base, and the next round of bases are flushed over the array. At every cycle of the sequencing process, only one base can be incorporated on the sequencing strand resulting in synchronous probe sequencing.

In contrast to the 454 sequencing, Illumina tends to focus on throughput rather than the lengths of the reads obtained from a sequencing run. At present, read lengths of up to 100 bp are possible, but there is a drop in quality of the reads as the read reaches the maximum read length. An example of the drop in quality of a 76 bp run of sequencing is presented in Figure 2.4, where a drop in base-quality can be observed from around base 68. The development of the paired-end protocol, where the both ends of the amplicons are sequenced, together with the extremely high-throughput (500 Gbp) on the HiSeq2000 platform, has made this technology ideal for genome re-sequencing and transcriptome studies where

the digital expression on a specific transcript can be measured (<http://www.illumina.com>). The factors limiting the technology to produce longer read lengths include the incomplete enzymatic cleavage of the fluorescent labels or terminal moieties, which leads to a decay in the detection signal and eventually leads to dephasing of the reaction (Shendure and Ji, 2008). Illumina technology suffers from a base substitution error, rather than an insertion or deletion as observed with the 454 platform. Average raw error rates have been reported to be in the order of 1-1.5%, but higher accuracy bases with error rates down to 0.1% can be achieved (Shendure and Ji, 2008).

### **SOLiD (Applied Biosystems)**

The original work of Shendure *et al.* (2005) and patents by McKernan *et al.* (2006) directly led to the development of the unique two-base encoding methodology behind Applied Biosystem's SOLiD system. As with the other systems discussed thus far, a fragmented DNA library of adaptor-flanked regions serve as the starting point for this technology. Cloning of the fragments is achieved with emulsion PCR, with the amplicons captured to the surface of 1 $\mu$ m beads (Dressman *et al.*, 2003). After breaking the emulsion, the amplicon-containing beads are immobilised to a solid planar substrate in order to generate a dense, disordered array of beads (Shendure and Ji, 2008). After the addition of a universal primer that ligates to the amplicons, the rather complex sequencing process can begin.

A notable difference between the SOLiD and the methods mentioned previously is that the sequencing reaction is driven by a DNA ligase rather than a polymerase, and is achieved by ligating a degenerate fluorescent octamer to the template (Shendure *et al.*, 2005). The octamer mixture is structured so that the identity of a specific base in the octamer corresponds to the fluorescent label of the octamer. After ligation and image capture with a CCD, the octamer is chemically cleaved between positions three and six, removing the fluorescent label. In effect progressive rounds of octamer ligation results in the sequencing of every fifth base (Shendure and Ji, 2008). After several cycles, the extended primer is denatured and the system is reset to its original state. The process is repeated, each time sequencing a different position in the octamer by either using an initial primer of a different length or by using a different position in the octamer as the fluorescent label. An additional complication to the system is

that an error correction method is in place. Effectively two adjacent bases correspond to the selected fluorescent label, and each base position is then queried twice, once as the first base and once as the second base, during a given cycle. A graphical representation of the sequencing cycle with the two base encoding system can be viewed on the company's website (<http://www.appliedbiosystems.com>).

The result from the two-base encoding system is that very accurate base qualities (>99.94 % accuracy) are achieved with the SOLiD system (<http://www.appliedbiosystems.com>). Read lengths were initially limited to 36 bp, but steadily increased to 75 bp. The high quality of the reads, as well as the very high-throughput of 300 Gb per run from the SOLiD 5500xl System puts it in the same application space as the Illumina platform. The confidence in the quality of the reads also provides a good platform for polymorphism studies. Since the output from the SOLiD system is in "color space" and not "base space", decoding of the reads into base space needs to occur before any analysis can be performed on the results. Most widely used sequence mapping and assembly tools have been adapted to cater for working in "color space", and a variety of converters exists which will convert "color space" reads to "base space".

### **Complete Genomics (Complete Genomics)**

Drmanac *et al.* (2010) described another DNA sequencing technology making use of self-assembling DNA nanoarrays and demonstrated it by re-sequencing three human genomes. The technology employs recursive restriction site cutting (type IIS restriction enzymes) and directional adaptor insertion methods to produce circled DNA replicated many times with a polymerase in order to create DNA nanoballs (Drmanac *et al.*, 2010). These nanoballs are attached to a photolithographic surface, and the sequence adjacent to the inserted directional adaptor sites sequenced using a high-accuracy combinatorial probe anchor ligation (cPAL) technology. cPAL uses degenerate anchors in order to read up to 10 bp adjacent to the inserted adaptor sites, with similar read accuracy across all the bases read. This method produced between 31-35 bp mate-paired reads.

Using nanoarray sequencing the average amount of sequence produced from three human genomes ranged from 124 Gb to 241 Gb, which corresponds to a coverage between 45X and 85X (Drmanac *et al.*, 2010). In terms of sequence quality and polymorphism calls, the authors achieved confident diploid calls

for up to 95% of the theoretical 98% of a Yoruban female genome (HapMap id: NA19240), with close to 94% of the SNP positions called (99.15% accuracy) in the HapMap phase I/II for the caucasian genome (NA07022).

Sequencing-by-synthesis, and sequencing-by-ligation-based technologies use chained reads, where the substrate for cycle  $N+1$  depends on the product of cycle  $N$ . The ligation-based approach described by Drmanac *et al.* (2010) uses an unchained approach, where complete probes are ligated to the target sequences, and the sequencing process does not depend on driving the reaction to completion with high concentrations of labeled nucleotides as used in other methods. Because of the lack of high concentrations of purified fluorescently labeled substrates, the average cost per sequenced genome was reduced to under US\$4 400. The short reads obtained from this technology and the late introduction of the commercial product to the market are some of the initial hurdles to overcome in order to ensure widespread adoption, but with the reduced cost this can be an attractive platform alternative to the Illumina and SOLiD platforms.

### 1.2.2. Single-molecule sequencing platforms

Single-molecule sequencers have been earmarked as the next big technological development aiming to achieve the target of sequencing a human genome for US\$1 000. At the time of writing, only the Helicos Biosciences system was available as a commercial application, but the commercial launch of the Pacific Biosciences Single Molecule Real Time (SMRT™) system was imminent according to the company. The Ion Torrent system was first announced at the 2010 Advances in Genome Biology and Technology (AGBT, <http://agbt.org>) meeting, and received much attention that warrants its inclusion in the following section. Oxford Nanopore's sequencing system is still in development, and little information is available on the technical aspects of the system, and is therefore not covered in this review.

#### SMRT™ sequencing (Pacific Biosciences)

The technology that led to the development of Pacific Biosciences' single-molecule sequencer was first described by Eid *et al.* (2009). The technology also relies on the incorporation of a fluorescently-labeled

nucleotide complementary to the target strand being sequenced. A notable difference with the nature of the fluorescently-labeled nucleotide, is that the nucleotide is labeled on the phosphate group. This labeling strategy has the effect that the fluorescent label is naturally cleaved from the nucleotide together with the phosphate group during nucleotide incorporation into the synthesized strand. Another unique feature of the Pacific Biosystems system is that rather than fixing the DNA template to an array and flushing enzymes across it, the DNA polymerase enzyme is fixed to the array, with fragmented DNA and labeled nucleotides flowing over the array. The technology involves binding a DNA polymerase ( $\Phi 29$ ) on a polyglycol-covered silica surface without direct interaction between the protein and the silica surface (Eid *et al.*, 2009). The seating of the polymerase protein occurs inside a zeptoliter ( $10^{-21}$  liter) well, which is small enough to allow a single fragmented DNA strand to enter, along with labeled nucleotides. Multiple wells are constructed in an aluminum cladding, known as the Zero-mode Waveguide (ZMW), in which the sequencing reaction occurs. Apart from functioning as a micro-reactor for the sequencing reaction, the ZMW reduces the background light noise which occurs in other wells on the ZMW, and allows for the detection of the light emitted from a single molecule of the fluorescently-labeled phosphate as nucleotides are incorporated by polymerase in real time (Single-molecule, real time (SMRT™) sequencing, Eid *et al.*, 2009). Since the whole process proceeds as fast as the DNA polymerase can incorporate bases into the template sequence, an average per base incorporation rate four orders of magnitude faster than second generation sequencers can be achieved. By simply manufacturing more wells on the ZMW, the reaction can occur in parallel, and comparable base pair throughput should be achievable in the future.

The use of SMRT™ sequencing has led to the development of a novel method of DNA circularisation, coined SMRTbell™, for consensus sequencing of the same molecule (Travers *et al.*, 2010). Using these circular templates which represents a linear DNA fragment, multiple passes of sequencing are performed, providing multiple copies of the same molecule. A demonstrative application of the technology was in re-sequencing a housekeeping gene (aroE132) with a single nucleotide difference between two strains of Multiple Resistance *Staphylococcus aureus* (MRSA, the FDA209 and Mu50 strains). By mixing the DNA fragments of the aroE132 gene from these two strains in different ratios, the robustness of the system to



detect the frequency of a single nucleotide difference within the samples was determined (Travers *et al.*, 2010).

Flusberg *et al.* (2010) showed that detection of DNA methylation without bisulfite treatment was possible with SMRT™ sequencing, avoiding some of the drawbacks of bisulfite sequencing, which includes the costly sample preparations used in methylation studies, the constraints in primer design of a treated genome, and the ambiguities in alignments of the generated sequences to the reference genome. By measuring the pulse duration from the phosphate cleavage by DNA polymerase of the labeled nucleotides, a difference in the polymer kinetics inside the ZMW well between methylated and non-methylated sites could be detected. The use of circular consensus sequencing aided in determining the parameters needed to measure methylated-adenosine sites, but methylated-cysteine and hydroxymethylcytosine detection needed additional kinetic sensitivity enhancements (Flusberg *et al.*, 2010).

Pacific Biosciences recently revealed read lengths up to 10 000 bp, and promises reads up to 50 000 bp in the near future. The high accuracy of the bases and confidence in detected variants of samples which are sequenced multiple times, are the major advantages of the technology, but the relatively low multiplexing capability of 3 000 ZMW wells in the commercial package is a drawback. However, the development system showcased at the 2010 AGBT meeting showed a massively parallel system, with over 80 000 ZMW wells capable of simultaneous sequencing in parallel. At the current sequencing speed of almost two nucleotides per second, this system has the potential to make real-time diagnostic sequencing a reality.

### **Heliscope Single Molecule Sequencer (Helicos Biosciences)**

The Helicos sequencer is a single molecule cyclic array sequencer. It was developed based on the research by Braslavsky *et al.* (2003). The key advantage of this technology over cyclic array sequencers is that there is no amplification step required during the sequencing process, which implies that the each signal detected on the array originates from a single molecule, and not a cluster of amplicons. A highly sensitive fluorescent detection system is used to directly interrogate single DNA molecules *via* sequencing-by-synthesis. Poly-A tailed fragmented DNA template molecules are captured by a

surface-tethered poly-T array, yielding an array of primed, single sequencing templates. Fluorescently labeled nucleotides and DNA polymerase are then systematically washed over the array, interspersed by chemical cleavage in order to detect the incorporated base *via* a CCD device.

Read lengths ranging from 35 bp to 70 bp have been reported with the system (Harris *et al.*, 2008; Pushkarev *et al.*, 2009), and read accuracy has been reported to be improved with a two-pass strategy in which the array of single molecules is sequenced, the original strand removed by denaturing, and the remaining strand re-sequenced (Harris *et al.*, 2008). This effectively yields a read in the opposite orientation from the template. This two-pass strategy can reduce the error rate from 2-7% to 0.2-1% (Shendure and Ji, 2008).

Due to the use of single molecules, a much higher density of unique fragments can fit on an array. Although the read length only ranges from 25 to 55 bases, the highly parallel nature of the technology allows it to achieve a throughput of between 21 and 35 Gb per run. The imaging system on the Helicos platform was designed for a theoretical throughput of 1Gb/hour, but this has not been achieved due to the practical constraints introduced by the chemical efficiency of the system. Functional genomic applications of the Helicos system have included the sequencing of a viral genome and BAC library (Harris *et al.*, 2008; Bowers *et al.*, 2009), digital gene expression of poly-A RNA transcripts generated by strand-specific reads (Lipson *et al.*, 2009; Ozsolak *et al.*, 2009) and ChIP-Seq applications (Goren *et al.*, 2010). The comparatively short average read length produced by the system, and the relatively late market introduction of the commercial application seem to be the major drawbacks in widespread adoption of the system.

## **Ion Torrent**

At the 2010 Advances in Genome Biology and Technology Meeting the founder of 454 Life Sciences, Johnathan Rothberg, revealed an innovative approach of sequencing DNA using a semiconductor system to detect the change in pH (due to the release of an hydrogen) when a base gets incorporated during sequencing (<http://www.agbt.org>, <http://www.iontorrent.com>). This technology, described as "Post light sequencing with semiconductor chips" lowers the capital investment needed to acquire a

sequencer to below US\$50 000, and the consumables for a run down to US\$500 per sequencing run (<http://www.iontorrent.com>). As of the beginning of 2011, no research articles have been produced applying the Ion Torrent system in a research environment, and commercial instances of the sequencer have not been sold. However, this technology promises affordable high-throughput sequencing available without a large capital investment.

### **1.3. High-throughput DNA sequencing applications in genetics and functional genomics**

The technological advances made with uHTS technologies have provided biologists with most of the required tools for a systematic approach to functional genomics. This has led to a gradual shift in focus from studying isolated parts of a system, to analysing DNA, RNA and proteins in context of the whole organism or cell. Genome re-sequencing efforts have led to better understanding and quantification of sequence and structural variation between individuals within species (Fullwood *et al.*, 2009; Pang *et al.*, 2010), and a more detailed blueprint of the genomic data organised in near complete chromosomes for most model organisms. Another consequential development was the understanding that the same physical blueprint, such as the genes embedded in a genome, exhibits massive variation in terms of functional post-transcriptional form and levels of transcript abundance (Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008; Pan *et al.*, 2008; Sultan *et al.*, 2008). The study of genotypic variation present in transcription products gains merit when there is an observable effect of a mutation on a phenotype. This, together with the observation that there are distinct differences in the structure and abundance of transcripts in a cell, necessitates the study of transcriptomes not only in an individuals, but in a specific tissue and in many individuals in order to observe transcriptional differences that can be associated with a condition. Both these approaches are relying on the use of uHTS technologies to provide the primary data for genome and transcriptome wide studies.

## ***De novo* genome sequencing**

Improvements in the chemistry used by sequencing platforms and the development of novel sequencing techniques such as paired-end sequencing have led to a gradual shift in sequencing applications from re-sequencing known genomes (Margulies *et al.*, 2005; Shendure *et al.*, 2005; Velicer *et al.*, 2006; Hofreuter *et al.*, 2006), to *de novo* sequencing and assembly of prokaryotic genomes (Tauch *et al.*, 2008; Reinhardt *et al.*, 2009), small eukaryotic genomes (DiGuistini *et al.*, 2010; Nowrousian *et al.*, 2010) and ultimately large eukaryotic genomes like that of the Giant Panda genome completely assembled from Illumina reads (Li *et al.*, 2010b). *De novo* genome sequencing with uHTS technologies has been thought an impossible task due the very short reads generated by these technologies, but mixing reads generated from different technologies which complement each other in terms of the read length, the quality of the bases in the reads, and the sequence throughput from these technologies have led to the development of cost-effective and *de novo* genome sequencing and assembly strategies (Aury *et al.*, 2008; DiGuistini *et al.*, 2010; Nowrousian *et al.*, 2010).

The most robust genome sequencing method is known as BAC-end sequencing. The fundamental approach to BAC-end sequencing is to perform a shotgun fragmentation of chromosomal DNA, and making use of Bacterial Artificial Clones (BAC) as vectors to sequence around 500 bp of each end of the vector insertion point (Venter and Smith, 1996). BAC-end sequencing has been very successfully applied in large genome sequencing projects, including the human genome project (Venter *et al.*, 2001), and was a key improvement over the generation of overlapping Yeast artificial chromosomes (YACs, Venter and Smith, 1996). Making use of uHTS technologies has enabled the sequencing of the large, complex and highly repetitive genome of barley from BACs (Wicker *et al.*, 2006; Steurnagel *et al.*, 2009). Another sequencing approach in contrast to BAC-end sequencing is the whole genome shotgun sequencing (WGS) of the organism in a single approach using NGS. Uncertainty over the feasibility of using only uHTS technologies to sequence a large genome was laid to rest with the publication of the Giant Panda genome (Li *et al.*, 2010b). There are certain tradeoffs between WGS and BAC sequencing, for example the increase in bioinformatics costs to assemble a genome produced from uHTS technologies.

For large complex genomes full of repeat elements such as the cereal genomes, alternative methods to BAC and WGS approaches exist. These methods aim to sequence very specific, pre-selected regions of the genome. Some of these methods include restriction analysis, where genomic DNA is treated with a restriction endonuclease, and then fragmented to remove abundant repeat fractions (Van Tassell *et al.*, 2008). Another approach can be isolating specific chromosomes for sequencing by means of chromosome sorting (Dolezel *et al.*, 2007; Simková *et al.*, 2008a,b).

The application of uHTS technologies to sequence plant genomes is fast gaining momentum. Since the initial sequencing of the first plant genome, *Arabidopsis* (AGI, The Arabidopsis Genome Initiative, 2000), large genome sequencing projects including rice (Goff *et al.*, 2002; Yu *et al.*, 2002), poplar (Tuskan *et al.*, 2006), maize (Schnable *et al.*, 2009) and soybean (Schmutz *et al.*, 2010) genomes have been completed by using Sanger sequencing. One of the first agriculturally important crops to make use of uHTS technology (454 sequencing) to complete a genome sequence was the consortium to sequence a heterozygous grape variety (Velasco *et al.*, 2007). More examples of completed genome projects making use of a mixture of traditional and high-throughput technologies include the cucumber genome (Huang *et al.*, 2009a), BAC sequences of the barley genome (Stearnagel *et al.*, 2009), and a genomic survey of the perennial grass *Miscanthus* (*Miscanthus x giganteus*, Swaminathan *et al.* 2010). A recent report on the applications of uHTS technologies in plant genomics revealed that the sequencing of the cacao (*Theobroma cacao*), apple (*Malus domestica*) and strawberry (*Fragaria vesca*) genomes currently underway make use of a mixture of Sanger and uHTS approaches (Imelfort and Edwards, 2009).

### **Genome re-sequencing and variant discovery**

Some of the first applications of uHTS technologies in a genomic context were the re-sequencing of the bacterial genomes of *Mycoplasma genitalium* (Margulies *et al.*, 2005), *Myxococcus xanthus* (Velicer *et al.*, 2006) and *Campylobacter jejuni* (Hofreuter *et al.*, 2006). In these projects, the microbes of interest were a lineage or strain that exhibits a biological phenotype different from the reference genome available for the species. These reference genomes served as template scaffolds onto which the generated sequences

were aligned, in order to detect single nucleotide polymorphism (SNP) and indel variations between the reference genome and the newly re-sequenced genome. The genomic differences were then related to the presence or absence of a biological phenotype, for instance antibiotic resistant genes or pathogenicity islands in the re-sequenced genomes.

Human cancer genomics has made great advances in terms of disease-specific re-sequencing efforts, revealing mutations in somatic tissues that are thought to contribute to tumor progression (Ley *et al.*, 2008; Mardis *et al.*, 2009; Pleasance *et al.*, 2010a). Exposure to detrimental environmental agents, such as tobacco smoke, has also led to genome re-sequencing of tissues under mutational pressure from these exposures, providing insight into the genome-wide carcinogenic effect of these agents (Pleasance *et al.*, 2010b). Data from these studies led directly to the design of genome-wide association studies (GWAS), which have the basic aims to identify genetic markers which can be used to predict an individual's risk to disease, and secondly to highlight the molecular processes involved in a disease, with the ultimate aim of identifying potential therapeutic targets. A natural feedback of information is present in determining genetic variation, where polymorphism information produced from genome re-sequencing efforts leads to the design of population-based marker arrays, which in turn prompts investigation in very specific, personal-whole genomes (Mir, 2009). Re-sequencing of genomes of agricultural importance tends to focus on adaptive evolutionary traits and the detection of novel genetic markers, especially where large differences in phenotypes are present in a species. The detection of a selective genomic sweep shared by broiler populations involving metabolic regulation and reproductive genes in modern chickens is an excellent example of identifying the effects of adaptive evolution and selection pressure in populations (Rubin *et al.*, 2010). Variant discovery and domestication studies have also been investigated in the silkworm (Xia *et al.*, 2009; Li *et al.*, 2010a), soybean (anchoring markers on the genome by Hyten *et al.*, 2010), and rice (Huang *et al.*, 2009b).

In human genetics, the search for disease phenotypes and population genetic markers led to the establishment of the 1 000 Genomes Project (<http://www.1000genomes.org>). The latest release of the data generated by the 1 000 genomes projects (released on 21 June 2010) included the data from three of

the completed subprojects. This release included the data from nearly 700 human genomes, and aims to produce an extensive catalog of human genetic variation, including SNP and structural variants. The final project will contain data described as "genomes of about 2000 unidentified people.....will be sequenced using next generation sequencing technologies" (<http://www.1000genomes.org>). This achievement somewhat overshadows the phenomenal achievement of the completion of the first draft human genome in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001), and builds on the example set by the re-sequencing efforts of the human genome by various other research groups (Bentley *et al.*, 2008; Wang *et al.*, 2008; Wheeler *et al.*, 2008; Ahn *et al.*, 2009; Kim *et al.*, 2009; McKernan *et al.*, 2009; Pushkarev *et al.*, 2009; Drmanac *et al.*, 2010; Schuster *et al.*, 2010), it also serves as an excellent showcase of the advances made possible by next generation sequencing during the last decade.

The development of high-throughput genotyping methods make the use of SNPs highly attractive in especially agricultural applications (De la Vega *et al.*, 2005). High-density SNP markers in a genome are ideally suited for the construction of high-resolution genetic maps, the investigation of evolutionary history within a population or species, and the discovery of marker-trait associations to aid marker assisted selection (MAS) in breeding programs. During the discovery of marker-trait associations, a dense set of markers are needed to cover the genome of interest to discover a casual mutation, or a SNP which is in linkage disequilibrium with a casual mutation for the trait of interest (Aranzana *et al.*, 2005). The construction of high-density genetic maps requires the genotyping of a large number of individuals, and platforms with the ability to genotype a large number of samples at a large number of polymorphic sites are desired. Successful applications of high-throughput genotyping experiments include the design of a barley SNP assay using the Illumina GoldenGate™ technology, providing the barley community with a platform to investigate diversity with over 3 000 markers (Close *et al.*, 2009). High-throughput genotyping assays have also been developed for the unsequenced genomes of white and black spruce (*Picea glauca* and *Picea mariana*, Pavy *et al.*, 2008), the complex genome of soybean which contains a high proportion of paralogous genes (Hyten *et al.*, 2008) and the allohexaploid genome of wheat (Akhunov *et al.*, 2009). A future application of uHTS technologies in genotyping, would be designing SNP arrays

for an organism for which a genome is not yet available, but for which gene information derived from technologies such as mRNA-Seq can be useful. A large number of EST sequences from different lines or individuals have already been used for marker identification in maize (Barbazuk *et al.*, 2007) and *Eucalyptus* (Novaes *et al.*, 2008). The authors of the *Eucalyptus* article reported close to 24 000 SNPs, and validated a proportion of the data with a success rate of close to 85%. Two more popular approaches to SNP detection in portions of the genome is to make use of specific fragments produced from selective amplification with restriction enzymes as demonstrated by van Orsouw *et al.* (2007) and the sequencing of restriction-site associated DNA (RAD) tags (Baird *et al.*, 2008).

Genome re-sequencing efforts also provide insight into other genome structural variations, such as indels, copy number variation, inversions and translocations occurring between different genomes. Re-sequencing of two naturally inbred *Arabidopsis* strains has led to the discovery of more than 800 000 SNPs and almost 80 000 indels ranging from 1 to 3 base pairs (Ossowski *et al.*, 2008). Finding longer indels between the genomes was reported as a problematic issue with the short reads (36 bp in length), but the use of paired-end reads as implemented by most current high-throughput technologies has resolved the problem (Ng *et al.*, 2006; Fullwood *et al.*, 2009). Structural variation detection has also been successfully employed in various human genome re-sequencing projects (McKernan *et al.*, 2009; Kim *et al.*, 2009; Pang *et al.*, 2010).

## **Transcriptome sequencing**

The transcriptome of an organism can be defined as the complete set of mRNA transcripts produced at any time in a cell. The transcriptome is by nature not in a steady state and across cell types, during different conditions in the cell's lifecycle, and in response to external and internal stimuli. The use of expressed sequence tags (ESTs) has become a standard in obtaining information regarding the coding, or expressed regions of an organism for which a sequenced genome is not yet available. Recently, the use of uHTS technologies has been applied to sequencing the RNA landscape of a cell, by making use of a



technology now commonly known as mRNA-Seq (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mortazavi *et al.*, 2008; Novaes *et al.*, 2008; Nagalakshmi *et al.*, 2008).

Various hybridisation-based methods have traditionally been used to study the transcriptome landscape, which have lately been complemented by sequence-based methods.. Traditionally, hybridisation-based methods involved labelling cDNA with a fluorescent dye, and then hybridising the cDNA to a set of probes on a microarray. Specialised array chips, such as exon-arrays have been designed specifically to identify spliced isoforms (Clark *et al.*, 2002; Frey *et al.*, 2005; Singer *et al.*, 2006; Kapur *et al.*, 2007), while genomic tiling arrays have been used to identify novel transcripts of already sequenced organisms (Bertone *et al.*, 2004; Cheng *et al.*, 2005; David *et al.*, 2006). The development of parallelised sequencing technologies have increased the use of sequence-based approaches to gene expression profiling and the genome-wide evaluation of chromatin immunoprecipitation (ChIP-seq) experiments. Some of the limitations of hybridisation-based methods include the dependency on knowledge of the sequence of the studied genome in order to manufacture the probes, the occurrence of inter-probe cross-hybridisation on the arrays, the presence of background noise and signal saturation, and some data-analysis issues in terms of normalisation of data between experiments (Eklund *et al.*, 2006; Okoniewski and Miller, 2006; Casneuf *et al.*, 2007; Royce *et al.*, 2007).

The development of tag-based sequencing methods which include cap analysis of gene expression (CAGE, Kodzius *et al.*, 2006), serial analysis of gene expression (SAGE, Velculescu *et al.*, 1995) and massively parallel signature sequencing (MPSS, Brenner *et al.*, 2000) allowed for the quantification of the amount of cDNA present in a biological sample. The advantages of these methods were that a unique hybridisation probe was not needed to detect each transcript and, in the case of SAGE analysis, multiple SAGE tags could be sequenced together providing several measurements simultaneously (Bertone *et al.*, 2005). The initial widespread adoption of these methods was hampered by the high cost of Sanger sequencing technology (Sanger *et al.*, 1977) used to determine the base pair composition of the sequence, and the technical problem that the very short tags (10-14 bp tags for SAGE analysis) generated by these technologies did not map uniquely to the reference genome (Bertone *et al.*, 2005; Wang *et al.*, 2009),

which made it very difficult to distinguish transcript isoforms from each other. An improvement in read length (21 bp Long-SAGE, Saha *et al.*, 2002) overcame some of these limitations, but the use of SAGE was prohibitively expensive until the power of HTS technologies was employed (Deep-SAGE, Nielsen *et al.*, 2006).

The development of a technology to sequence the transcriptome content of a biological sample has been achieved by the major high-throughput sequence technology companies (see Section 1.2 for an overview of the technologies). The premise of these technologies is the fragmentation of a population of RNA (total RNA, polyA-selected RNA), which is converted to a library of cDNA fragments with adapters attached to one or both ends. Each RNA molecule can then be sequenced in a high-throughput manner from one (single end sequencing) or both ends, resulting in reads that can vary from 35-450 bp in length depending on the technology used. Prior to sequencing, the RNA or cDNA molecules can be amplified, but sequencing of RNA without amplification has the added advantage of providing expression information in addition to the transcript sequences (Wilhelm and Landry, 2009). RNA-Seq technology is slowly reaching maturity, and it offers some key advantages over hybridisation-based technologies, with longer sequences than tag-based technologies, and a lower cost per base pair than traditional EST sequencing technologies. It has also been shown that RNA-Seq detects differential gene expression with greater sensitivity than expression (Li *et al.*, 2008a; Marioni *et al.*, 2008) and tiling microarrays (Hiller *et al.*, 2009).

Findings obtained with genome-wide analysis of transcribed sequences and potential transcriptional start sites indicated that the traditional genome-centric view of the protein coding regions of the genome needed to be replaced by a more complex transcript-centric view (Bertone *et al.*, 2004; Johnson *et al.*, 2005; Carninci *et al.*, 2006). These findings brought the idea that there is a defined set of isolated loci transcribed independently into doubt, and indicated that numerous overlapping coding and non-coding transcripts span the entire genome, and that those transcripts are of biological importance in the cell system, which in turn led to a renewed research interest in transcription and transcription-related products in a cell. Recently, with the use of RNA-Seq to determine the proportion of the genome which is

transcribed, evidence suggests that the initial estimation of transcription might have been excessively overestimated (van Bakel *et al.*, 2010). The earlier studies were based on tiling microarray data, and the recent studies indicated that the microarray platform is susceptible to a high rate of false positives (van Bakel *et al.*, 2010). In the recent study, most of the transcripts not mapping to exonic regions, mapped to introns, raising the possibility that these RNA-Seq fragments belong to pre-mRNAs (van Bakel *et al.*, 2010). This study indicated that most of the genome is not appreciably transcribed in levels associated with gene expression, but still leaves the question of what the function of low-level transcribed genomic regions are.

One of the initial applications of mRNA-seq derived data was the discovery of novel transcripts, with the simultaneous estimation of transcript abundance (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mortazavi *et al.*, 2008). Cloonan *et al.* (2008) sequenced poly-A captured RNA transcripts from two different mouse tissues, and demonstrated that alternative splice forms from transcriptionally active tissues were readily detectable with mRNA-seq. The sequencing approach they followed (not normalising the sequence libraries) led to the elucidation of transcript expression values, an approach initially proposed by Mortazavi *et al.* (2008) for mouse transcripts. Mortazavi *et al.* (2008) developed a measure of gene expression, measured in reads per kilobase of exon per million mapped sequence reads (RPKM), which is a normalised measure of exonic read density. The use of RPKM values was widely adopted, and various software packages utilise this measure to report gene expression. Furthermore, Cloonan *et al.* (2008) demonstrated that the *de novo* detection of gene models is possible with high levels of expression and alluded that allele specific expression detection is a near-certain possibility in transcript expression studies. In order to perform *de novo* prediction of gene models from a genome using mRNA-Seq, Denoeud *et al.* (2008) developed a software package **G-Mo.R-Se**, and applied it to the recently sequenced *Vitis vinifera* genome. The authors used mRNA-Seq (175 million Illumina reads) from four different tissues and identified new exons in known loci and alternative splice forms, as well as entirely new loci in the *Vitis* genome.

Data obtained from mRNA-Seq experiments led to investigations into the alternative splice complex-

ity of genes active in different tissues. Previous methods using microarray profiling and cDNA sequencing lacked the sensitivity or confidence due insufficient coverage needed to validate multiple splice events. In the human genome, alternatively spliced transcripts were estimated to occur in two thirds of the genes, but studies using mRNA-Seq estimated that 95% of multi-exon human genes in major human tissues showed evidence of alternative splicing (Pan *et al.*, 2008). Similar results were obtained in human embryonic kidney and B cell line tissues, where an average of 7.2 splice junctions per gene was identified, but employing a very lenient measure of one matched sequence to validate a synthetic splice junction (Sultan *et al.*, 2008). In *Arabidopsis*, the percentage of alternatively spliced genes was estimated at 42% for multi-exon genes (Filichkin *et al.*, 2010), which also surpasses the previous estimates of between 22% and 33% (Campbell *et al.*, 2006; Wang and Brendel, 2006; Chen *et al.*, 2007; Barbazuk *et al.*, 2008). Intron retention was the most prevalent form of alternative splicing in *Arabidopsis*, and was frequently associated with specific abiotic stresses of the plants, which led the authors to postulate the existence of a functional transcript regulation mechanism similar to the regulated unproductive splicing and translation (RUST) mechanism in animals (Lewis *et al.*, 2003; Filichkin *et al.*, 2010). These discussions regarding different splice forms being actively transcribed in a cell under certain conditions raised the question regarding in what quantities these splice forms are distributed across tissue types. In previous studies to quantify transcript expression from mRNA-Seq data, reads were not allocated to specific isoforms, but this feature was implemented in the **Cufflinks** software package (Trapnell *et al.*, 2010). The authors of **Cufflinks** detected 330 genes present in mouse myoblast tissue, which switched their dominant transcription start site or splice isoform during a time-series experiment. **Cufflinks** also no longer relies on any *a priori* information regarding the gene models of an organism, and is able to infer the gene models directly from the combination of mRNA-Seq data and a genome.

Antisense transcription has been shown to play an important regulatory role in the eukaryotic genome. A simple modification to the RNA-Seq method enabled the method to yield strand-specific transcripts (ssRNA-seq, Parkhomchuk *et al.*, 2009; Perkins *et al.*, 2009). The method incorporated a deoxyUTP during the second strand cDNA synthesis, followed by the destruction of the uridine-containing strand in

the sequencing library, thus allowing the polarity of the transcripts to be known. The method was applied to the yeast and mouse model organism datasets, yielding new information regarding promotor-associated and antisense transcription (Parkhomchuk *et al.*, 2009). Another genome-wide investigation of the transcriptional landscape using ssRNA-seq revealed the presence of subtle regulatory RNA and small RNA sequences in the genome of the bacterial pathogen *Salmonella enterica* serovar Typhi (Perkins *et al.*, 2009). The mapping of strand-specific reads to the *S. enterica* Typhi genome provided a single base pair resolution map of active transcriptional elements, resolving overlapping annotated transcripts previously made. The utilisation of ssRNA-seq data derived from large eukaryotic genomes will shed light on the content of the pervasively transcribed transcriptome in future studies.

The combination of high-density genome-wide genetic markers with expression profiling data to identify trait-associated gene expression patterns, or expression Quantitative Trait Loci (eQTL) in mapping populations is fast becoming a reality with the use of HTS technologies. Using data from 60 human Caucasian participants in the HapMap project, Montgomery *et al.* (2010) investigated the occurrence of detectable eQTLs from genome-wide collections of SNPs. The authors were also able to detect allele-specific expression from the same expression dataset, which would certainly form the basis for expression studies in hybrid mapping populations. According to the authors, a dataset of 10 million mappable fragments are required in order to quantify alternative and highly abundant transcripts (Montgomery *et al.*, 2010). A similar study of 69 lymphoblastoid cell lines derived from Nigerian HapMap participants identified over a thousand genes where genetic variation contributes to variation in expression and splicing (Pickrell *et al.*, 2010). Results from these studies confirm the observation that most eQTLs are located close to the gene's transcriptional start site, and that most eQTLs influence expression in a *cis* fashion (as oppose to *trans*-regulated expression). In addition to the ability to quantify the expression of different transcript isoforms, these studies also improved the annotation of the genome by detecting previously unannotated exons (Pickrell *et al.*, 2010).

mRNA-Seq has been shown to produce accurate measurements of the expression landscape of the genome with unprecedented accuracy. Data derived from mRNA-Seq experiments has been used to

detect the expression of known and previously unknown transcripts, to assemble transcriptomes from organisms with no genomic information, to detect allele specific expression patterns, and identify novel splice forms. Bioinformatics algorithms and data management approaches to handle these datasets are evolving at a rapid pace in order to handle mRNA-Seq data, and it is not uncommon for a software package to undergo several version updates in a short period of time as the nuances of these datasets are better understood. The computational needs of processing mRNA-Seq, or any uHTS dataset for that matter, varies according to the intended applications, from a large number of CPUs needed in loosely-coupled homology searches of tens of thousands of genes against public datasets in parallel, to the massive memory requirements of *de novo* assemblers, and must be considered when a high-throughput experiment is planned.

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

One of the strengths of ultra-high-throughput sequencing platforms is in the various practical applications it has in genetic and genomic studies. For each of these applications, there exists a core set of data analysis methods performed with the data in order to address the underlying biological questions. The core data analysis tools range from estimating the quality of the bases received from sequencing facilities, assembling of reads into larger contigs (transcriptomes or genomes), and mapping of reads to a target sequence in order to detect structural variation, evaluate transcript expression, and perform SNP mining or structural variation detection.

##### **Determining the quality of Illumina mRNA-Seq data**

Illumina results are generally presented to researchers in the FASTQ format, The preprocessing of the images is performed by the sequencing facility, since it uses the proprietary Illumina Pipeline to perform the base-calling from the image sources. The output from the Illumina Pipeline, or to be more specific, the BUSTARD tool, is a FASTQ formatted quality FASTA file (Figure 1.1). The FASTQ quality

values differ from the standardised Phred quality values prepared by Sanger-based sequencing machines and software pipelines, and also differs depending on the version of Illumina Pipeline that was used to perform the base calling. Phred-based quality scores are calculated by  $Q_{Phred} = -10\log_{10}\left(\frac{1}{\$error\_prob}\right)$ , where  $\$error\_prob$  is the probability of the base call being wrong (Ewing *et al.*, 1998; Ewing and Green, 1998). In order to present the score of a base in a single character, the  $Q_{Phred}$  score is converted to a corresponding American Standard Code for Information Interchange (ASCII) character. ASCII is an 8-bit character set defining alphanumeric characters widely used in the computer industry. Since ASCII 32 is the whitespace (spacebar) character, Phred scores use ASCII characters 32-126 to represent qualities from 0-93. The dynamic range of a Phred score ranges from 1.0 (a completely wrong base), through to  $10^{-9.3}$ , an extremely accurate base (Cock *et al.*, 2010). This is also known as the **fastq-sanger** format.

The Illumina FASTQ format encode base qualities in two different scoring systems. Illumina Pipeline (< version 1.3) defined a new scoring formula to determine the quality score:  $Q_{Solexa} = -\log_{10}\left(\frac{\$error\_prob}{1-\$error\_prob}\right)$ . The after-effect of this non-standard scoring formula resulted in a change of the ASCII-offset used to represent a base score. Since the  $Q_{Solexa}$  score's lower limit is -5, assuming a random read error probability of 0.75, a very low quality base will result in a whitespace character representing the quality score (this occurs because ASCII characters 0-32 are all whitespace characters). Due to the fact that whitespace characters can be interpreted differently by some computer operating systems, which should be avoided in setting a standard where the quality values are aimed to be represented in a single line of a text file (for example, the newline character is also a whitespace character), the ASCII offset of 64 was chosen. This resulted that ASCII 59-126 was used, providing the  $Q_{Solexa}$  score a dynamic range from -5 through to 62 inclusive (this format is generally known as the **fastq-solexa** format). After version 1.3 of the Illumina Pipeline, the scoring function changed to be compatible with the Phred standard, but the ASCII offset of +64 remained, and the format is now known as the **fastq-illumina** format (Illumina, 2008). For a review of the complete history of the FASTQ format, and also the introduction of the ABI Solid CFASTQ format (in color-space, not base or sequence space), please see Cock *et al.* (2010). The discussion above was required to introduce the concept of format conversions of raw Illumina

A)

```
@HWI-EAS121 0005 FC61APKAAXX:1:1:2358:16565#0/1
GTAGTAACTTGNCATTTGCTAGTGTGCTTGTGACATGTAGTTTTAGGTCATTTATTNATCTTTACTCTCAGGAATTCAG
+HWI-EAS121 0005 FC61APKAAXX:1:1:2358:16565#0/1
cddddeeeebKbbccccceeeeeeeeddeeeeeeeeeedda`b`bb`aa`daTdaedec`

@HWI-EAS121 0005 FC61APKAAXX:1:1:2358:4461#0/1
TTTTGATGTTGNCAGGATTACAAGAACAGCCATTTCTCTAGTGTGTTACTAGGGNGAGCAATACAGGAATTAATGGC
+HWI-EAS121 0005 FC61APKAAXX:1:1:2358:4461#0/1
YRa\\T\\a]]FVXZURVRVRZQZX]__bU_VUU]a]Va\\X[[Y[QZOZZ]RT[SVDZZR' Z' ZTGa]T\\KK'KaBBBB

@HWI-EAS121 0005 FC61APKAAXX:1:1:2358:19891#0/1
CAAGCGCAGGANGCCATGTGGACAATCAAGTCAACAACACGGGAAGTGTAGCCCCANTCATTGTCGTACCATGAGACCAG
+HWI-EAS121 0005 FC61APKAAXX:1:1:2358:19891#0/1
dddTdddadbKbbb]b`^`dddafffeeeeffffacdc^bbdadlb`_`GWIIYX[VXab[___aaa_a_^_Z^B

@HWI-EAS121 0005 FC61APKAAXX:1:1:2358:1852#0/1
GCAATACATGCNGTTACAAATACTTGATTGGAATGCATTCATTGTGCACGTGGGTANACTGCGGTGTGGGAATCAGCCT
+HWI-EAS121 0005 FC61APKAAXX:1:1:2358:1852#0/1
dddadeeeebKbbb]_ba`ffff^ffceffecfdaffffLdffffbdddY`XHW[VZYUYRa^Hab^a^^^acca\\

@HWI-EAS121 0005 FC61APKAAXX:1:1:2358:7138#0/1
CTGGTGTGCTTNCAATGCTCCTTTTCATGCTGAACCTGGATTGTGACCACTACATANATAACAGCAAGGCCGTCGCGAG
+HWI-EAS121 0005 FC61APKAAXX:1:1:2358:7138#0/1
dddadbbddblbbaccocffiffiffefefefeedffdbddffiffdbaGaaa``]]ad`_`bcbbbcbcdba^
```

B)

```
@HWI-EAS121 0005 FC61APKAAXX: Instrument name
1: Flowcell lane
1: Tile number within flowcell
2358: X-coordinate of cluster
7138: Y-coordinate of cluster
#0: Index number (if multiplexed)
/1: Member of paired-read (1 or 2)
```

Figure 1.1: An example of an Illumina FASTQ formatted mRNA-Seq file. The example presented above represents five 80 bp reads and the quality values associated with the reads (a). The sequence and quality header lines are denoted by the @ and + symbols, while the line following the header line represent the bases and the qualities associated with the specific base pair. Note that the whitespace lines in between the reads were inserted to improve readability of the format. The header file contains the following information separated by colons; the unique instrument name, the flowcell lane, the tile number within the flowcell, the 'x' coordinate and 'y' coordinate of the cluster within the tile, the index number for a multiplexed sample and if paired, the first or second member of a pair (b).



data. Some assembly tools perform the conversion between the Illumina formats (both `fastq-illumina` and `fastq-solexa`) to the traditional `fastq-sanger` formats if the input type is specified at run time. There are also standalone conversion tools available to translate between the different formats for use in analysis tools that do not provide the conversion ability.

### ***De Bruijn* graph-based genome and transcriptome assembly**

The short reads produced by uHTS technologies are not suited to be assembled by the same sequence assemblers as traditional Sanger sequencing reads. With longer Sanger reads, the assembly process relied on the overlapping of reads which fit together to generate a consensus sequence, or contig. Very short reads are not suited for the traditional overlap-layout-consensus based method of assembly (Zerbino and Birney, 2008). Because of the large numbers of reads that are produced, short reads have a much higher coverage over a specific region. An overlap-based method, where the actual reads are stored to generate a consensus sequence, has computational limitations when handling billions of reads where large numbers of reads have an overlap of all but one base pair. With overlap-based methods, each read forms a node of a graph, and the nodes are connected by an overlap metric between the nodes (Batzoglou, 2005).

A fundamental shift in the methodology behind aligning short reads was introduced in 2001, with the adaptation of *de Bruijn* graphs to represent and organise the relation between reads using an Eulerian path approach to assemble sequence reads (Pevzner *et al.*, 2001). In essence, *de Bruijn* graphs do not represent whole reads as nodes in a graph, but rather break the reads into words of a pre-defined length (length  $k$ , henceforth known as  $k$ mer(s)), and the reads are then organised in paths through the graph in a determined order. By using  $k$ mers rather than reads, the redundancy of the graph is inherently handled by the structure of the graph, without increasing the number of nodes in the graph. Every node in the graph thus represents a single  $k$ -mer (non-redundant), and have explicit links to the neighbors, or start and end positions of the  $k$ mer in a read (Pevzner *et al.*, 2001). Various research groups have since investigated the use of *de Bruijn* graphs in short read assembly software programs (Shah *et al.*, 2004; Bokhari and Sauer, 2005; Myers, 2005; Jiang *et al.*, 2007; Zerbino and Birney, 2008).

The Velvet program was one of the first *de novo* short read assemblers implementing the *de Bruijn*

graph assembly strategy. While transcriptome-specific assemblers were developed towards the end of this study, during the initial phases of this project **Velvet** was the only assembler found to produce cDNA contigs of reasonable length and quantity. Analysis with **Velvet** consists of two phases, first the indexing of the input reads with the desired kmer, and secondly the traversing and tracking of the kmers to construct the contigs. **Velvet** relies on coverage per kmer to eliminate erroneous nodes, resolve repeated kmers and find the path between the nodes which is most represented by coverage and constructs the output sequence (Zerbino and Birney, 2008). **Velvet** is an example of a memory hungry application, with massive memory requirements needed to store and traverse the kmer graphs. A recent experiment of a single lane of 76 bp paired sequence ( $\approx 40$  million reads), consumed close to 45 GB of RAM during assembly with a kmer of 41 bp. The developers of the **Velvet** package are continuously improving the memory footprint of the algorithms used.

Alternative assemblers which utilise the *de Bruijn* graph assembly approach include but is not limited to the **ABYSS** (Simpson *et al.*, 2009) and **OASES** (Zerbino *et al.*, unpublished) assemblers. **ABYSS** was used to successfully assemble the human transcriptome of a patient with follicular lymphoma (Birol *et al.*, 2009). Using **ABYSS**, the authors assembled  $\approx 65\,000$  contigs representing close to 30 Mb of the human transcriptome. The **OASES** assembler was developed as an extension to the **Velvet** assembler with the purpose of focusing on splice variant assembly of transcripts. The source code of the project was made public early in 2010, and at the time of writing no peer reviewed publications had been published using the application. These applications are viable alternatives for transcriptome assembly projects.

### **Mapping mRNA-Seq reads to a reference dataset**

The requirements of a short read mapper can be separated into a strategic requirement in terms of alignment accuracy, and a more practical requirement in terms of a time constraint (Trapnell and Salzberg, 2009). Firstly, the use of high-throughput sequence technologies for variant discovery in whole genomes requires the accurate, high confidence alignment of the short read to the target genome. In this application, the presence of repeat regions in the genome, as well as natural variation that occurs between the reference genome and the re-sequenced genome needs to be accounted for, and the short read

mapper needs to be robust enough to handle these issues confidently. Traditional alignment programs, such as BLAST (Altschul *et al.*, 1990) and BLAT (Kent, 2002) are also able to align short sequences to a target genome, but the algorithms used in these aligners are not optimised for very short reads (35-76 bp), and the time required by these aligners to perform billions of alignments hampers these programs from being serious contenders for high-throughput alignments.

RNA-derived reads can be mapped to a target sequence with different objectives; firstly, a fully sequenced, annotated genome where gene models are already predicted, and the mapped reads are used to calculate gene expression values; secondly an un-annotated or newly sequenced genome to detect gene models or infer new genes; or thirdly, a set of genes or coding regions from a unknown genome (typically the results from a *de novo* transcriptome assembly project). Several short read mapping software packages are available, some of the first mappers include ZOOM! (Lin *et al.*, 2008), MAQ (Li *et al.*, 2008b), Mosaik (Stromberg and Marth, 2008), SOAP (Li *et al.*, 2008d), SHRiMP (Rumble *et al.*, 2009) and Bowtie (Langmead *et al.*, 2009), with more recent updates to the algorithms implemented in SOAP2 (Li *et al.*, 2009b) and the successor to Bowtie, BWA (Li and Durbin, 2009, Table 1.1). These short read mappers typically works by selecting a defined wordsize usually from the beginning of the short read, and then requiring some number of these words to fit perfectly to the target to find a match, while mismatches are allowed to occur within the rest of the words (Li *et al.*, 2008d; Langmead *et al.*, 2009; Li *et al.*, 2009b; Li and Durbin, 2009). Another common approach is to create a subsequence, or a spaced seed, along the high quality 5' end of the short read sequence, and again with some mismatch threshold allowed, the seeds are aligned to the target (Lin *et al.*, 2008; Li *et al.*, 2009b; Rumble *et al.*, 2009). The next section describes in detail the difference in these two approaches, as implemented by the Bowtie and MAQ aligners.

### Mapping reads with the spaced seed approach

MAQ employs a spaced seed indexing strategy in order to align segments of a short read to a genome. A short read is effectively divided into four sets of words of equal length, called a spaced seed. By default, MAQ uses the first 28 bp of a short read for seed generation, and uses a word size of six to

Table 1.1: A selected list of short read sequence alignment tools currently available for academic use. These software tools perform essentially the same function in aligning reads generated from uHTS technologies to a target genome, but implementing different mathematical, statistical and programmatic approaches to achieve this goal.

Program name	Description	Reference
BFAST	BLAT-like Fast Accurate Search Tool for aligning re-sequence data to a genome. The program returns an accurate alignment for a candidate alignment location where the short read corresponds to the genome. It also includes support for two-base encoding sequences from the SOLiD platform.	Homer <i>et al.</i> (2009 $a,b$ )
Bowtie	A very efficient short read aligner implementing the Burrows-Wheeler transform in order to be memory efficient. Bowtie can align up to 25 million 35 bp reads per CPU hour.	Langmead <i>et al.</i> (2009)
BWA	An update of the MAQ package, based on a backward search with Burrows-Wheeler transform, effectively eliminating the alignment of repeated short reads.	Li and Durbin (2009)
ERANGE	Mapping mRNA-Seq data to genomes for quantification of transcript expression. Makes use of the Bowtie aligner.	Mortazavi <i>et al.</i> (2008)
Genome Mapper	Simultaneously aligning reads to multiple genomes by collapsing the corresponding regions of the genomes into a single graph structure. Used by the 1001 genomes project ( <a href="http://1001genomes.org">http://1001genomes.org</a> ) consortium.	Schneeberger <i>et al.</i> (2009)
RMAP	Used base quality scores in deciding the appropriate map position of a read on a reference sequence.	Smith <i>et al.</i> (2008)
Slider and SliderII	Specifically developed for the Illumina platform, and uses the probability files instead of the sequence files in order to perform the alignment to the reference sequence.	Malhis <i>et al.</i> (2009)
SOAP and SOAP2	Introduced gapped and ungapped alignments, and the use of a paired-end module. SOAP2 update of SOAP, implementing a Burrows-Wheeler transform algorithm.	Li <i>et al.</i> (2008 $d$ , 2009 $b$ )
TopHat	Uses BWA to perform multiple alignments to a genome with mRNA-Seq data in order to detect splice junctions.	Trapnell <i>et al.</i> (2009)
MAQ	One of the first short read aligners to implement mapping quality to the target genome. Not as computationally efficient as some of the other programs.	Li <i>et al.</i> (2008 $c$ )
Mosaik	Produces gapped alignments using the Smith-Waterman alignment algorithm, and forms part of a software suite which includes SNP calling.	Stromberg and Marth (2008)

generate the spaced seeds. If a perfect match between the read and the target sequence exists, then all of the spaced seeds will match the target. If, however, a mismatch is present in the target sequence, then one or possibly more of the spaced seeds will not match perfectly. When two mismatches are present between the short read and the target sequence, at most two of the spaced seeds will not have a perfect match (only one space seed will show a mismatch if the mismatches are close to each other, and do not span a space seed boundary). By aligning pairs of spaced seeds (there are six possible pairs for the 4 seeds) to the target, it is possible to identify the possible locations on the entire target sequence where the complete short read will match, allowing for at most two seed mismatches. The resulting list of candidate positions are then compared to the complete read extending from position 28 onwards without gaps to identify the correct mapping position. The sum of the qualities of the mismatched bases are then calculated and stored together with a random number and the hit positions in an index. When two short read sequences are mapped with the same mismatch quality scores, the one with the smallest random number is selected as the best possible alignment. MAQ can be configured to use up to 20 spaced seeds, and is then able to find all 28 bp seeds with up to 3 bp mismatches, although this means a mismatch ratio of more than 10% between the seed and the target sequence.

### **Mapping reads with the Burrows-Wheeler transform approach**

The Burrows-Wheeler transform (BWT) is a much more complicated method, but has the advantage of running substantially faster (up to 35x when compared to MAQ) than an index-based method, and with a smaller memory footprint (Langmead *et al.*, 2009). Originally developed for lossless file compression (Burrows and Wheeler, 1994), the transform involves building an extremely efficient transformation of the target sequence, and then mapping a short read one base at a time to the BWT target. This is achieved by combining the BWT with some opportunistic data structures and the building of a reverse index to minimize backtracking, to allow for an efficient search space (Ferragina and Manzini, 2000, 2001). Each new successively aligned character allows the algorithm to narrow down the possible location where a short read might match perfectly. It has been shown that the original implementation of MAQ and SOAP would take 35x and 300x longer than the corresponding *Bowtie* alignment (Langmead *et al.*, 2009).

Since the original publications of MAQ (development discontinued and replaced by BWA, Li and Durbin, 2009) and SOAP (updated as SOAP2, Li *et al.* 2009b), both of these these programs have been updated to utilise the BWT algorithm for building a transformed target sequence. The much smaller memory footprint (1.3 GB for the entire human genome), and the general 30x speedup of the BWT algorithm has made this approach currently the most widely used tool for mapping short reads to a target sequence.

### **Mapping high-throughput genomic reads to a genome**

High-throughput DNA sequencing is ideally suited for genome re-sequencing projects where variant discovery is the main focus (see section 1.3 for a review of re-sequencing projects). The fraction of short reads which map to the reference genome depends on several factors. If there is a minimal amount of variation between the reference and the re-sequenced genome, the alignment algorithms improved are capable to align from around 70-75% of single end reads to the reference genome, up to 85% with the BWA aligner, and up to 98% with paired-end reads (Langmead *et al.*, 2009; Li and Durbin, 2009). The quality of the sequencing library, the amount of repeat regions in the reference genome, the length of the reads and the insert size in the case of paired-end reads all influence the mappability of a short read. Paired-end reads improves the mappability of a sequenced fragment by having two reads with a known distance associated with the fragment. Paired-reads are specifically useful for improving fragment mappability in cases where one of the reads aligns to a repeat region in the genome sequence. It has been calculated that with 35 bp reads, the fraction of the human genome that is re-sequencable is 85%, and with paired-end reads with an insert of 170 bp, this fraction increases to 93% (Li *et al.*, 2008b). Any additional increase in short read mappability could only be obtained with an increase in read length and having datasets of varying insert sizes available.

### **Mapping mRNA-Seq reads to a genome**

RNA-derived reads, such as those produced by mRNA-Seq, strand specific RNA-Seq and total-RNA-Seq protocols provided by Illumina require gapped alignments across gene splice junctions in order to map sequenced reads to eukaryotic genomes. The computational approach to map reads to exon-exon bound-

aries is different to genome derived short read mapping due to the possibility of a single read spanning across two exons that were joined during transcript processing. The first approach to solve this problem was to utilise the structure of known genes in determining the intron-exon boundaries of a gene, such as implemented in the ERANGE package (Mortazavi *et al.*, 2008). Another approach is to extract possible junction sequences from the aligned genomic sequence with some form of machine learning algorithm, for example a logistic regression classifier (Pan *et al.*, 2008) and a support vector machine-like approach (Schulze *et al.*, 2007; De Bona *et al.*, 2008). Unfortunately these methods only work for organisms for which gene models are available, as the gene models serve as a required input to delineate the intron-exon boundaries together with training data sets.

Because of the reliance on known gene models to map the RNA-Seq reads to fully sequenced genomes as mentioned before, these methods are limited in detecting novel splice junctions. Another approach to splice junction mapping was proposed and implemented by the two software packages TopHat (Trapnell *et al.*, 2009) and G-Mo.R-Se (Denoeud *et al.*, 2008). These packages utilise the power of a BWT mapping tool (initially only Bowtie, but Bowtie and BWA are now supported) to detect possible exons, and then by joining the exons which share transcripts, remap the data in order to detect possible splice junctions. Of the two packages, TopHat package is currently being actively maintained.

## 1.5. High-throughput DNA sequencing data management

Recent calculations from the Ontario Institute for Cancer Research indicated that since the advent of uHTS, the cost of sequencing a base has been dropping faster than the cost associated with storing a byte of data on a computational storage medium (Stein, 2010). The author investigated the historical trends in data storage prices *vs.* DNA sequencing costs, and found that the doubling time in sequenced base pair per dollar was less than six months, exceeding the drop in disk storage cost on a logarithmic scale. One of the fundamental problems in terms of sequence storage, is that a single base has multiple bytes associated with it. During a uHTS run where the bases incorporated during the sequencing process is captured by a CCD, the image needs to be converted from an image to a string representation, usually in

basespace, but colorspace is also gaining prevalence in order to prepare the data for input into a variety of analysis programs. A quality score is usually associated with the each base call, effectively doubling the storage space needed for a base. Format incompatibilities, such as the case of the FASTQ format (Section 1.4 on page 25) can require various duplicate versions of the same data to be stored as input files. Different analysis tools produce various output files, which can be thought of as different representations of a base, highlighting different features of the base, or the surrounding bases in terms of biological relevance. The problem in terms of storage cost and expansion capabilities is thus compounded by the already exponential growth of uHTS base throughput, and the non-linear relationship between a base of sequence and the space required to store the biological relevance of that base.

The nature of uHTS data requires a disciplined and structural approach to data management. The different file formats required by software packages require that the data be duplicated between analysis steps, increasing the data storage and computational cost associated with uHTS analysis. Tools developed for uHTS analysis are being made available to the community at a rapid pace, and an analysis environment where these tools can be distributed to various users for immediate use and implementation in data analysis workflows is essential.

### 1.5.1. Widely-used bioinformatics workflow systems

During the last decade, many bioinformatics research groups have dedicated resources to develop mature automated and semi-automated analysis environments. The implementations of these systems are as varied as the number of programming languages used to develop the system, and include executing complex analysis on local resources (Ergratis Orvis *et al.*, 2010; Kepler Ludäscher *et al.*, 2005; Galaxy Goecks *et al.*, 2010), on remote systems through web-services access (Taverna, Oinn *et al.*, 2004), or making use of distributed grid systems (Taverna, Galaxy). To evaluate different workflow systems, one needs to critically evaluate the the relative strengths and weaknesses of these cyberinfrastructure implementations.

Using dedicated, local resources for high-throughput data analysis has the the advantage of having



complete control over the number of CPU cycles dedicated to a project. The downside of local resources is firstly the cost of the resource, the cost of installing and maintaining a diverse set of analysis tools and systems on the servers, and the investment in human capacity to fully utilise and maintain the hardware components.

Web-services, grid and cloud computing offer attractive alternatives to overcome the initial capital investment in hardware (Stein, 2010). One of the fundamental requirements of utilising a remote resource for computing, is the access to fast and cheap network bandwidth to the remote server for data transfer, but this requirement often precludes the use of remote services from some institutions or research groups. Access to these remote computing sites is also limited to the availability of CPUs at the remote sites at any given time.

## **Taverna**

**Taverna** (Oinn *et al.*, 2004) was developed as part of the *myGrid* initiative for the composition and execution of workflows in the life sciences domain. **Taverna** relies on the Simplified conceptual workflow language (**Scuf1**) to represent each step of a workflow as a single task. A graphical user interface (GUI) was developed and packaged as part of **Taverna** which acts as a container in which **Scuf1**-based workflows can be constructed, without the need to learn the **Scuf1** language. The workflows in **Taverna** rely on the availability of programmatic access to bioinformatics repositories, such as **GenBank**, and analysis tools, such as the **EMBOSS** suite of tools at the European Bioinformatics Institute (EBI), **SOAPlab** (Senger *et al.*, 2003) and **BioMOBY** (Wilkinson and Links, 2002). Access to the tool or repository is granted through a web-service interface (Stein, 2002), which allows the consumer (the **Taverna** client) to query a database or start an analysis tool on the host server remotely. The advantages of this type of architecture is that data stored in large datacenters, such as the EBI, NCBI and DDBJ, are accessible to users across the world through a simple, standardised interface. Centers with access to large computational resources can also expose analysis web-services to the community, and therefore allow smaller research groups with limited resources to execute jobs with large computational requirements remotely. This service-oriented

design of **Taverna** also allows it to connect to services that can submit jobs on a grid-like environment for distributed computing.

**Taverna** has been successfully employed by many research groups, the biggest and most prominent is the integration of **Taverna** into the cancer Biomedical Informatics Grid (caBIG) project, where **Taverna** and the Web-service-Business Process Execution Language (WS-BPEL) are used in a service-oriented data analysis environment (Tan *et al.*, 2008, 2009; Missier *et al.*, 2009). As explained above, the service-oriented nature of **Taverna** relies on the ability to connect to a host server to interact with the data, but when the data is not mirrored on the host server, the data needs to be transferred to the compute elements. This requires that either a reliable, fast and inexpensive network connection is needed to connect to the remote services, or a duplication of the services needs to be present on a local network where the data is already present.

The nature of uHTS data in general does not lend it to be readily distributed to various computing locations. In most cases, the prohibitive factor is the cost and time needed to duplicate multi-GB datasets across many locations in order to perform analysis in parallel. Although the South African Research Network (SANREN, <http://meraka.org.za/sanren.htm>) has made great progress in terms of providing a fast and reliable cyberinfrastructure between South African research institutes and the rest of the world, the availability of reliable bandwidth at a high enough data throughput is still a major hurdle to overcome.

## **Kepler**

The **Kepler**-project (Ludäscher *et al.*, 2005, <https://kepler-project.org>) is an example of a data-driven, scientific data analysis and knowledge discovery pipeline. This **JAVA**-based application is very similar to the web-service-based implementation of **Taverna**, but relies on the **Ptolemy II** open-source software framework which support an actor-oriented pipeline design (Eker *et al.*, 2003). An actor can be seen as a step in the analysis pipeline, where multiple actors can be connected to each other *via* data channels. The **Ptolemy II** system was designed with heterogeneous data in mind, and has been very successfully implemented in automated pipelines by scientific groups (Lee and Zheng, 2005; Lee, 2009; Leung *et al.*, 2009).

## Ergatis

**Ergatis** is a workflow management system optimised for parallelised analysis of constructed pipelines making use of the **Sun Grid Engine** (SGE, Orvis *et al.*, 2010). It is a workflow management system targeted for working with genome sequence data, where analysis pipelines can be executed on a single server, or distributed across large computing clusters. **Ergatis** was developed making use of standard ontologies in bioinformatics, and supports input files in the Bioinformatics Sequence Markup Language format (<http://www.bsml.org>), the Sequence Ontology for sequence feature annotation (Eilbeck *et al.*, 2005), and the Gene Ontology format for functional annotations (Gene Ontology Consortium, 2001). The workflow system has the added capability of exporting results into a CHADO-based database (Mungall *et al.*, 2007), making it compatible with the GMOD set of tools (Stein *et al.*, 2002). The **Ergatis** system executes scripts or tools locally and does not require a web-service as interface, in contrast to TAVERNA and Kepler, and offers a flexible user interface to manage and control executing workflows.

## Galaxy

The **Galaxy** workflow system (Goecks *et al.*, 2010) has been used by several research groups for biological data analysis (Kosakovsky Pond *et al.*, 2009; Gaulton *et al.*, 2010; Peleg *et al.*, 2010). The goal of **Galaxy** is to serve as a layer of abstraction on top of a myriad of underlying tools, and serve them to regular users through an intuitive web interface. The inputs and results from various programs, as well as the parameters used for each of these programs are stored in a history of a project or analysis step, which can be shared with collaborators, used as a workflow for similar analysis steps, or archived for publications. Almost any scriptable piece of software, including custom Python, PERL and R scripts can be wrapped in the **Galaxy** interface allowing for the easy extension of the framework to include custom tools. **Galaxy** hides the underlying complexity of the programs imbedded in it allowing users to focus on scientific hypotheses, rather than technical issues associated with the software needed to perform the analysis used to address the biological questions.

## 1.6. Problem Statement

The hypothesis is formulated that by making use of data from Illumina mRNA-Seq deep sequencing data, the transcriptome of a complex eukaryotic organism like *Eucalyptus* can be successfully assembled and characterised to such an extent that biologically relevant and accurate information can be obtained regarding transcriptional control of growth and development.

In order to test the hypothesis, a structured approach is needed to first identify a suitable data management and data analysis framework to aid in the analysis of uHTS data. The data analysis framework will then be used to test the different parameters and settings of the software packages used to assemble and annotate the *Eucalyptus* transcriptome. The framework should be readily extendible with additional software tools that are not already implemented in the framework to aid in the analysis and construct automated workflows to perform the data analysis steps.

The workflows developed should then be used to perform a *de novo* assembly and homology-based annotation of the transcriptome of a *Eucalyptus grandis* x *Eucalyptus urophylla* plantation tree from deep sequenced mRNA-Seq data. The assembly should be validated as far as possible without the aid of the draft *Eucalyptus grandis* genome sequence, to validate that a *de novo* transcriptome assembly is indeed possible. The assembled gene catalog should be characterised and annotated with homologs from other angiosperm transcriptomes, and used to identify genes differentially expressed between xylogenic and photosynthetic tissues.

To allow access to the assembled gene catalog, a web-based system should be developed that stores the contigs and corresponding annotations, and allows users to browse and search for contigs based on the annotations assigned to the contigs. The gene expression (FPKM) of the contig in each of the sampled tissues used perform the assembly should additionally be made available in the user interface.

## 1.7. Specific research questions and aims

- With the current selection of open-source uHTS data management and analysis packages available, is it possible to develop automated software workflows that perform DNA sequence analysis? In each of the developed workflows, identify the key parameters that have an effect on the results from a workflow. Where software tools are not present in the selected data management system, these tools should either be developed or added to the system to successfully perform a *de novo* assembly and annotation of a transcriptome dataset.
- To what extent can a transcriptome of a complex organism like *Eucalyptus* be assembled and evaluated using only mRNA-Seq data? The workflows developed in the previous aim should be used to completely assemble and annotate a large eukaryotic transcriptome. The assembled transcriptome should be evaluated for contig contiguity and the presence of full-length contigs in the dataset, without the aid of the *Eucalyptus* genome sequence. Functional annotation of the transcripts should be made in an automated fashion, and the transcript dataset should be compared to other angiosperm datasets in terms of the number and diversity of the assembled contigs. Finally, the gene expression profiles (FPKM) values of the transcripts should be used to identify a set of differentially expressed genes in xylogenetic and phytosynthetic tissues.
- Development of an intuitive, web-based *Eucalyptus* specific transcriptome resource that enables users to query and browse the assembled transcriptome dataset based on annotations? The web-resource should serve as a central repository for the data generated in the previous aims, and should be considered as a development platform and extension point for future whole genome mRNA-Seq based transcriptome sequencing and expression studies in *Eucalyptus*.

## Chapter 2

# A core bioinformatics workflow environment for ultra-high-throughput transcriptome data analysis

## Chapter preface

This chapter describes the development of software tools in the form of `Galaxy` workflows to address very specific next-generation sequence analysis problems. The workflows address very specific bioinformatics steps during the analysis of uHTS transcriptome datasets. The developed workflows focus on evaluating the quality of data from an Illumina mRNA-seq run, introduce a *de novo* transcriptome assembly pipeline, describe an automated assemble pipeline, and also provides a framework for gene expression (FPKM) calculation of genes expressed from a genome where the gene models have not been defined yet.

A customised `Galaxy` server has been installed at the Bioinformatics and Computational Biology Unit (BCBU) research group, that contains a copy of tools available in the public `Galaxy` server, as well as new tools that are not available on the default server. These tools consists of either third party, open source applications in the public domain that were included in the BCBU `Galaxy` instance, or novel Python and R scripts that were developed specifically for the BCBU server.

The BCBU `Galaxy` server can be accessed at the following URL: <http://zoidberg.bi.up.ac.za:8882>

## 2.1. Introduction

The sheer volumes of data produced by high-throughput technologies are forcing the biological research community to adopt automated data analysis methodologies in order to investigate the underlying biological relevance of the data produced. These technologies have enabled relatively small research groups with moderate budgets to produce large amounts of DNA sequence data, which necessitated the bioinformatics community to develop user-friendly analysis environments geared towards data management and result sharing. The current lack of bioinformatics human capacity, technical support and computational hardware in most research institutions is generally considered the bottleneck in obtaining relevant biological answers to a hypothesis. Deploying flexible and user-friendly analysis systems which empower the laboratory scientist to assist in data analysis and interface with custom software solutions developed by the bioinformatics community will greatly relieve the demand for bioinformatics support in a research project, and will assist both the experimental biologist and bioinformaticist in interpreting experimental findings.

The field of bioinformatics is more often than not spoiled for choice when it comes to selecting the most appropriate software analysis tool to perform a specific analysis. New software tools are made available to the community on a weekly basis, and especially in a newly expanding field such as high-throughput sequencing applications, various analysis tools that perform essentially the same function, but following different methodological approaches are rapidly being developed. A good example is the wide range of short read alignment tools currently available to align results from mRNA-Seq data to a target genome (Table 1.1). Each of these software packages have been designed with specific criteria in mind, and selecting the most appropriate tool that fits an experimental design or computational environment is often a daunting task. Many research groups and consortia have developed software pipelines and automated systems which use specific tools to address the need for analysis automation (Mungall *et al.*, 2002; Durham *et al.*, 2005; Forment *et al.*, 2008). In general, these pipelines do not lend themselves to customisation in terms of the exchange of one analysis tool for another that is more suitable for an experiment, and often requires modifying various scripts in order to successfully replace a tool. The

need therefore exists for a bioinformatics workflow environment, where very complex analysis pipelines can be built *ad hoc* from a repository of tools, and these pipelines can then be executed with different datasets and parameters as input, and together with the results, shared with collaborators (Ludäscher *et al.*, 2005; Taylor *et al.*, 2007).

A successful bioinformatics analysis and workflow system needs to meet a diverse set of requirements. First, the initial development hurdle required to extend the system needs to be intuitive and relatively easy, it needs to be easily deployable and maintainable, scalable to various computational environments systems, as well as having a user-friendly interface for the users. The bioinformatics community currently employs a diverse range of tools and programming languages to develop analysis tools geared towards biological data mining. Traditionally, dynamic scripting languages, such as Python, PERL, PHP and RUBY have been used with great success in building complex analysis portals and resulted in large “Bio\*” community projects developing around these languages (Chapman and Chang, 2000; Stajich *et al.*, 2002; Goto *et al.*, 2003; Holland *et al.*, 2008 and <http://www.openbio.org>). The aim of these communities can be summarized as providing a standard set of tools, or modules to perform common bioinformatics tasks. These tasks generally involve parsing results from popular analysis tools, connecting to the application programming interface (API) of a widely-used analysis tool, or converting between different biologically relevant file formats. The extensive use of these scripting languages in the bioinformatics community can be attributed to the lower entry level knowledge required when compared to compiled languages such as C, C++ and JAVA when learning the language. This is evident in the popularity of these languages in many introductory courses to bioinformatics (Cohen, 2003; Boyle, 2004). Ideally, a bioinformatics analysis pipeline system should be ignorant in terms of the language a particular tool is written in, and should leverage the community expertise in term of skills and experience when new tools and features needs to be added to the workflow framework.

The modern biologist and laboratory scientist should ideally interact with an analysis workflow system in such a way that the underlying hardware requirements and nuances of running a specific tool should be obscured from the user interface, enabling the researchers to focus on interpreting the results obtained.



The **Galaxy** workflow system (Giardine *et al.*, 2005), introduced in Section 1.5.1 meets a large number of the criteria mentioned above for a successful bioinformatics pipeline system, and was therefore selected to serve as the basis of a system which caters for next generation data management and analysis. **Galaxy** has the ability to execute scripts or analysis programs concurrently on local computational resources, and do not require the use of remote resources to execute a specific job. Workflow systems such as **Taverna** (Oinn *et al.*, 2004) and **Kepler** (Ludäscher *et al.*, 2005) makes extensive use of remote servers and protocols to construct the workflows. With the limited bandwidth available in South Africa during the lifetime of this project, these workflow systems were not considered as viable contenders for a base workflow system to extend. The **Ergratis** (Orvis *et al.*, 2010) system was only published in 2010, which effectively excluded it from being used in this study.

The aim of the chapter is firstly to develop automated analysis pipelines which will perform analysis related to the quality evaluation of mRNA-Seq reads, the *de novo* assembly of a gene catalog, develop an automated functional annotation pipeline and perform expression profiling of gene transcripts using mRNA-Seq short reads. Secondly, for each of the workflows developed, some key parameters that have an effect on the output of the different tools will be investigated, and recommendations provided as to what ranges of these parameters should be considered when performing some of the analysis steps. In order to fully describe the parameters, different mRNA-Seq datasets were used as input to the workflows. The workflows developed in this chapter were used to perform a successful *de novo* assembly and annotation of a gene catalog described in Chapter 3.

## 2.2. Materials and methods

### 2.2.1. BCBU Galaxy: Extending the public Galaxy framework

The **Galaxy** framework (Giardine *et al.*, 2005) served as the base of extension for the development of the uHTS sequence analysis workflows. The public framework already contains a wide range of NGS analysis tools, and these tools were used wherever possible to construct the workflows. When a specific

analysis tool was missing from the public server, the tool was added to the BCBU Galaxy server. The tools added to the BCBU server either consisted of third party applications, such as the Velvet assembler that were developed by external authors, or custom Python and R scripts that were developed specifically for this project. The list of third party applications added to the BCBU server is provided in Table 2.1, and the newly developed tools added to the BCBU server in Table 2.3.

### 2.2.2. Illumina short-read base-quality evaluation workflow

The Illumina FASTQ quality evaluation was performed with scripts and tools already present in the Galaxy framework. The default installation of Galaxy already provides uHTS data analysis functionality focussed on mRNA-Seq quality evaluation. The workflow, "Illumina QC" evaluates the quality of the bases from the forward and reverse reads from an Illumina paired-end run. The output from the workflow includes a bar chart of the distribution of base quality values for every base in the sequenced mRNA-seq dataset. The workflow also produces a summary of the FASTQ statistics file, which reports the number of reads in the lane, the number of bases, and the number of unknown bases in the run. The quality control tools enable users to evaluate the quality values of especially the 3' end of bases in the input dataset, and make informed decisions for trimming bases from a dataset for use in downstream analysis.

### 2.2.3. De novo transcriptome assembly workflow

The *de novo* transcriptome assembly workflow made use of the *de Bruijn* graph-based assembler Velvet, and a FASTA statistics calculation script from the `cndsrc` package<sup>1</sup> to guide the user towards steps needed to perform a transcriptome assembly. Transcriptome assembly is not a straight-forward process, and during the workflow construction the effect of multiple parameters regarding the input dataset, such as sequenced read length and the effect of paired end reads, as well as the effect that different parameters provided to the assembler have on the final assembly were evaluated. A 76 bp *Eucalyptis grandis* Illumina-sequenced mRNA-seq dataset was used to illustrate the effect of these parameters. This dataset was trimmed to illustrate the effect various input data lengths (50 bp to 76 bp),

<sup>1</sup> <http://www.biostat.wisc.edu/~cdewey/software.html>, included in the Galaxy framework as the "faLen" tool

Table 2.1: Third party applications that were added to the BCBU Galaxy server instance. The category column indicates the location of the tool in the BCBU server, and the reference column describes the publication of the tool, or where applicable, the software package that the tool is part of.

Name	Category	Description	Reference
Exonerate alignment	Alignment	Alignment of EST or cDNA sequence to a target genome sequence	Slater and Birney (2005)
BLAST2GO pipeline	Annotation	Executes the b2gPipe command line interface of the BLAST2GO tool, requires a local installation of the BLAST2GO package and databases	Conesa <i>et al.</i> (2005)
BLASTXML2 BLAST2GO	Annotation	Re-formats BLAST results in XML format to a format required by the BLAST2GO application	Developed by lmanchon@univ-montp2.fr, open source
InterProScan	Annotation	Runs the InterProScan analysis tool, requires the installation of all the required InterPro datasets. Currently optimised to utilise 16 cores on a single server	Zdobnov and Apweiler (2001)
BLAST	BLAST	Performs a BLAST against one of the public databases available locally	Altschul <i>et al.</i> (1990)
BLAST two FASTA files	BLAST	Allows users to upload fasta files, creates the BLAST databases on demand, and performs a BLAST analysis	Altschul <i>et al.</i> (1990)
Circoletto BLAST visualisation	BLAST	Makes use of the Circoletto application to view BLAST results in text format	Darzentas (2010)
faLen stats	FASTA tools	Calculates the N50, min, max, 1st and 3rd Quartile, mean and median sequence lengths from a fasta file	<a href="http://www.biostat.wisc.edu/~cdewey/software.html">http://www.biostat.wisc.edu/~cdewey/software.html</a>
FASTQ shuffleseq	FASTQ tools	Shuffles two FASTQ files into one file, required by the Velvet assembler	Zerbino and Birney (2008)
GenScan	Gene Predictors	Calls the GenScan tool on a fasta file containing protein sequences	Burge and Karlin (1997)
Velvet assembly	NGS tools	Performs a Velvet assemble on a FASTQ file	Zerbino and Birney (2008)
Multiple Velvet assemblies	NGS tools	Allows a series of Velvet assemblies with a range of parameters	Zerbino and Birney (2008)
Oases assembly	Development	Performs an Oases assembly on a FASTQ file	Zerbino <i>et al.</i> , unpublished
DEGseq	Development	Calculates differential expression between lists of genes using FPKM as the measure of expression	Wang <i>et al.</i> (2010a)
Muscle alignment	Development	Uses Muscle to perform multiple sequence alignments	Edgar (2004)

Table 2.3: A list of tools newly developed to complement the existing tools available in the BCBU Galaxy server. The tools include R and Python scripts that perform specific analysis, or convert files between different formats that serve as input to the next tool in the analysis pipeline.

Name	Category	Description
Exonerate targetgff2gff3	Alignment	Converts the gff and text output from Exonerate to the GFF3 format
InterProScan RAW format converter	Annotation	Re-formats InterProScan RAW results to either a txt or XML based format. The XML format is required by the BLAST2GO application
InterProScan2 BLAST2GO	Annotation	Converts InterProScan XML results to a directory format required by the BLAST2GO application
Parse BLAST XML	BLAST	Provides the facility to extract custom fields from a BLAST XML file
Convert gff3 to gtf	Convert formats	Produces the compact GTF format from a GFF3 file
Convert qseq to fastq	Convert formats	Converts an Illumina qseq file to a fastq file
Extract FASTA region	FASTA tools	Extract regions from a FASTA file
Reverse fasta sequence direction	FASTA tools	Reverse all the sequences in the FASTA file
Retrieve longest transcripts	FASTA tools	Parses the OASES assembler assembly files, retrieves the longest assembled transcripts
Rename FASTA entries	FASTA tools	Rename the FASTA entries
Summary of FASTQ	NGS tools	Calculates the number of usable bases, the number of A, C, G and T bases and the theoretical base yield from a FASTQ summary statistics file
Summary statistics file		
SAM QC stats	NGS tools	Calculates the number of reads that map as pairs, as singles, and uniquely from a SAM file
TopHat QC stats	NGS tools	Calculates the same statistics from a TopHat generated SAM file
SNP filter	SNP tools	Filter a pileup file with more stringent constraints, such as the minimum distance between two SNPs
SNP summary	SNP tools	Generates a summary of a pileup file. Includes the average distances between SNPs

different sequencing approaches (paired *vs.* single end sequencing), and different assembly parameters (kmer, expected coverage, and coverage cutoff parameters) on the same dataset. The different assemblies obtained from running multiple iterations of the workflow were compared with each other by a robust scoring algorithm that takes the number of contigs and length distribution of the contigs into account to evaluate an assembly. The workflow is provided in the BCBU Galaxy server as the "Velvet assembly pipeline". The workflow also discusses ways to evaluate the contig contiguity of the assembled datasets against known transcript sequences using BLAST (Altschul *et al.*, 1990) and related tools.

#### **2.2.4. Annotation of predicted protein sequences workflow**

An annotation workflow that focus on the functional annotation of translated cDNA sequences by widely used tools such as BLAST2GO (Conesa *et al.*, 2005) and InterProScan (Zdobnov and Apweiler, 2001) was developed. The pipeline predicts protein sequences from the input cDNA sequence file, and assigns functional annotations such as Gene Ontology (Gene Ontology Consortium, 2001), KEGG (Ogata *et al.*, 1999) and PFAM (Finn *et al.*, 2010) to the predicted protein sequences. The workflow relies on finding homologous sequences in model organisms, on which the functional annotations is based. The workflow is made available as the "Annotation pipeline" workflow in the BCBU Galaxy server. The various components in the workflow were used to perform the functional annotation of a *de novo* assembled Eucalyptus transcriptome described in Chapter 3. The results from the annotation pipeline can easily be imported into a third party application database, such as the Eucspresso system (Chapter 4) for the visualisation of results.

#### **2.2.5. Expression profiling using Illumina mRNA-Seq short reads workflow**

One of the main uses of mRNA-Seq data is transcriptional profiling of expressed gene products across the genome. Steps involved in calculating transcript expression include mapping reads to a target genome, inferring read coverage, and calculating the number of short read fragments that map to a specific genomic position, albeit a known gene region or an unknown genomic region. The workflow makes use of the TopHat aligner (Trapnell *et al.*, 2009) to map short-reads to a target genome sequence,

and the CUFFLINKS (Trapnell *et al.*, 2010) program used to calculate the normalised expression value of the gene in fragments per kilobase of reads mapped per million mapped reads. The workflow describes the gene expression calculation of a genome sequence where the only resource to define the gene boundaries in the genome is a set of EST data. The EST dataset is aligned to the genome with the EST2GENOME mode of the EXONERATE (Slater and Birney, 2005) package. After the genomic positions of the putative gene models were identified, differentially expressed genes between two sets of tissues were identified with the R-package DEGseq (Wang *et al.*, 2010a).

## 2.3. Results and discussion

Several next-generation data analysis workflows were constructed and saved in the BCBU Galaxy server as re-usable workflows, specifically with the aim to evaluate the quality of initial Illumina mRNA-Seq input data, the parameters which influence the assembly of transcriptome datasets, annotation of predicted protein sequence datasets, and expression profiling of transcriptome making use of mRNA-Seq short-reads. The sections describing each of the workflows consist of an overview or aim of each workflow, a short discussion on the components of the workflow, and a description of the effect of the parameters that can serve as input to the workflow on the results from the analysis pipeline.

### 2.3.1. Extending the Galaxy framework

The Galaxy framework serves as a container to host data analysis tools. The framework has the ability to sequentially execute various analysis tools on specific input datasets, selected by the user. Each tool contained in the framework is represented by a XML file, which specifies the input parameters that are sent to the tool during programmatic execution. Jobs can be executed on a local server, or submitted to a job handler server, such as the Sun grid engine (SGE, <http://http://wikis.sun.com/display/GridEngine/Home>) that executes jobs on a cluster-based computing platform. The Galaxy server automatically keeps track of the status of the submitted jobs, and the results are displayed in the server (the "histories" pane) after the job has been completed. The server also enables the user to construct workflows, or sequential steps

that need to be performed given an input data set. The following section describes the steps required to add a very basic analysis tool to the **Galaxy** framework.

The results from paired-end sequencing on the Illumina platform, consist of two **FASTQ** quality files, one for reads sequenced in the 5' to 3' ( forward reads), and one for reads oriented in the 3' to 5' direction (reverse reads). The tool, named “**shuffleseq**”, joins two **FASTQ** formatted files from an Illumina file into one file, with the reads in the final file sorted in an alternate fashion of forward and reverse reads. This “shuffled” **FASTQ** file is a required format for the **Velvet** assembler, and the **shuffleseq** executable script forms part of the **Velvet** assembler distribution.

To extend the BCBU **Galaxy** server to contain the “**shuffleseq**” script, an XML file needs to be created that registers the tool in the server, and renders an interface to select the tool. The **shuffleseq** XML file is presented in Figure 2.1, and consists of the following sections. Lines 3-7 specify the command to be executed, and allows the definition of the names of the input parameters, as well as the required format of the input datasets (lines 9-15). The name and format of the output file to store in the database is defined in lines 16-19. **Galaxy** has a default interface to define automated software tests, and encourages test-driven development, which will not be discussed here. These automated tests can then be run during the development phase of the when adding a tool to ensure that pre-calculated results are obtained with a with pre-defined set of input parameters. In this example, the input parameters for the tests are defined in lines 21-24, and the expected output for a successful test in line 25. Documentation regarding the functionality of the tool is provided from lines 29 to 46 of the XML file. This XML file renders the interface shown in Figure 2.2.

The executable, in this case the **Python** script named “**fastq\_paired\_end\_shuffleseq.py**”, is presented in Figure 2.1. Lines 13-16 of the file handle error reporting, and lines 21-38 contain error handling code to ensure that the input and output files are readable and writeable. The execution of the **PERL** script occurs on line 45, surrounded again by some error handling code if the execution of the script fails. The crucial link between the XML file and the executable is defined in the `<command>` tag of the XML file, and the input parameters or options in the **Python** script. In effect, the **Galaxy** execution engine passes the



```

fastq_paired_end_shuffleseq.xml
1 <tool id="fastq_paired_end_shuffleseq" name="FASTQ_shuffleseq" version="0.1.a">
2 <description>on single end reads to make a Velvet input file</description>
3 <command interpreter="python">fastq_paired_end_shuffleseq.py
4 --input1="$input1"
5 --input2="$input2"
6 --output="$output_file"
7 </command>
8 <inputs>
9 <param name="input1" type="data"
10 format="fastq.fastqsanger.fastqcassandra.fastqillumina.fastqsolexa"
11 label="Left-hand Reads" />
12 <param name="input2" type="data"
13 format="fastq.fastqsanger.fastqcassandra.fastqillumina.fastqsolexa"
14 label="Right-hand Reads" />
15 </inputs>
16 <outputs>
17 <data name="output_file" format="fastq" />
18 </outputs>
19 <tests>
20 <test>
21 <param name="input1_file"
22 value="split_pair_reads_1.fastqsanger" ftype="fastq" />
23 <param name="input2_file"
24 value="split_pair_reads_2.fastqsanger" ftype="fastq" />
25 <output name="output_file" file="3.fastq" />
26 </test>
27 </tests>
28 <help>
29 What it does
30 This tool joins single end FASTQ reads from two separate files into a joined file,
31 formatted for Velvet input.
32 -----
33 Input formats
34 Left-hand Read:
35 @HWI-EAS91_1_30788AAXX:7:21:1542:1758/1
36 GTCAATTGACTGGTCAATAAATAAAGATAGGATC
37 +HWI-EAS91_1_30788AAXX:7:21:1542:1758/1
38 hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
39 Right-hand Read:
40 @HWI-EAS91_1_30788AAXX:7:21:1542:1758/2
41 GCTCCTAGCATCTGGAGTCTCTATCACCTGAGCCCA
42 +HWI-EAS91_1_30788AAXX:7:21:1542:1758/2
43 hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh hfhhVZSWehr
44 -----
45 Output
46 A multiple-fastq file, for example:
47 @HWI-EAS91_1_30788AAXX:7:21:1542:1758
48 GTCAATTGACTGGTCAATAAATAAAGATAGGATC
49 +HWI-EAS91_1_30788AAXX:7:21:1542:1758
50 hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
51 @HWI-EAS91_1_30788AAXX:7:21:1542:1758
52 GCTCCTAGCATCTGGAGTCTCTATCACCTGAGCCCA
53 +HWI-EAS91_1_30788AAXX:7:21:1542:1758
54 hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh hfhhVZSWehr
55 </help>
56 </tool>

fastq_paired_end_shuffleseq.py
1
2 Runs the velvet shuffleseq command on the two input files,
3 creating a correctly formatted fastq file for velvet
4
5 shuffleseq forms part of the Velvet package, developed by
6 D Zerbino. http://www.ebi.ac.uk/~zerbino/velvet/
7
8 @version: 0.0.1b
9 @author: charles.hefer@gmail.com
10 """
11 import sys, os, optparse
12
13 def stop_err(msg):
14     """Handles any errors"""
15     sys.stderr.write( "%s\n" % msg )
16     sys.exit()
17
18 def main():
19     Main program control
20     parser = optparse.OptionParser()
21     parser.add_option( "-a", "--input1", dest="input1",
22                       help="The first input fastq file" )
23     parser.add_option( "-b", "--input2", dest="input2",
24                       help="The second input fastq file" )
25     parser.add_option( "-o", "--output", dest="output",
26                       help="The output fastq file" )
27     (options, args) = parser.parse_args()
28
29     Open up the input files and write to the output file
30     try:
31         open(options.input1, "r").close()
32         open(options.input2, "r").close()
33         open(options.output, "w").close()
34     except TypeError, e:
35         raise stop_err("You need to define the input and output files:\n%s" % str(e))
36     except IOError, e:
37         raise stop_err("There was an error reading or writing to:\n%s" % str(e))
38
39     The executable path to the shuffleSequences.pl script
40     cmd = "/usr/local/velvet/shuffleSequences.pl"
41     options = "%s %s %s" % (options.input1, options.input2, options.output)
42
43     try:
44         os.system( "%s %s" % (cmd, options) )
45     except Exception, e:
46         raise stop_err(e)
47
48 if __name__ == "__main__":
49     main()
50
51

```

Figure 2.1: An example of code developed to extend the Galaxy framework with the "shuffleseq" tool. The .xml file (left) defines the interface to the tool, and specifies the input and output format requirements. The Python script (.py) on the right pass the input and output parameters from the xml file to the Perl script, located on the file system. This example illustrates the ease of extending the Galaxy framework. In just over 100 lines of code, additional functionality was added to the framework.



### FASTQ shuffleseq

**Left-hand Reads:**

**Right-hand Reads:**

---

**What it does**

This tool joins single end FASTQ reads from two separate files into a joined file, formatted for Velvet input.

---

**Input formats**

**Left-hand Read:**

```
@HWI-EA391.1.30788AAXX.7.21.1542.1758/1
GTC AATTGTACTGGGTCAATAC TAAAAAATAGGATC
+HWI-EA391.1.30788AAXX.7.21.1542.1758/1
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
```

**Right-hand Read:**

```
@HWI-EA391.1.30788AAXX.7.21.1542.1758/2
GTCCTAGCATCTGGAGTCTCTATCACCTGAGCCCA
+HWI-EA391.1.30788AAXX.7.21.1542.1758/2
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh*bfbbY2SWebR
```

---

**Output**

A multiple-fastq file, for example:

```
@HWI-EA391.1.30788AAXX.7.21.1542.1758
GTC AATTGTACTGGGTCAATAC TAAAAAATAGGATC
+HWI-EA391.1.30788AAXX.7.21.1542.1758
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EA391.1.30788AAXX.7.21.1542.1758
GTCCTAGCATCTGGAGTCTCTATCACCTGAGCCCA
+HWI-EA391.1.30788AAXX.7.21.1542.1758
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh*bfbbY2SWebR
```

Figure 2.2: The interface of the FASTQ shuffleseq tool described in the fastq\_shuffleseq.xml file, as rendered by Galaxy. The interface provides the user to select buttons to select the forward (left hand) and reverse (right-hand) reads that will be "shuffled" into a single file as output. A short description on the function of the tool, and an example of input formats is also provided.

following parameters to the Python script during execution: `python fastq_paired_end_shuffleseq.py --input1=path/to/input1/file --input2=/path/to/input2/file --output=/path/to/output/file`, and expects the result file to be present in the output file location. The PERL script could have been called directly by the XML file, but this example illustrates that any executable command can be wrapped in the Galaxy framework and executed.

### 2.3.2. Quality assesment of Illumina short-reads

The quality control of experimental data forms an integral part of any analysis pipeline. A workflow dedicated to calculating the average base quality, the number of usable bases and the total number of reads from an Illumina mRNA-Seq lane was developed (Figure 2.3, which is available as the *Illumina QC* workflow in the BCBU Galaxy server,). The typical yield in terms of bases from an Illumina GA IIx run is reported by the company to be between 37 Gbp and 45 Gbp (January 2011, <http://www.illumina.com>), and these ranges were observed in a recently produced dataset (Table 2.5).

The FASTQ file format stores the quality associated with every sequenced base of every read in the FASTQ file. Reads produced with the Illumina platform tend to show a drop in the quality of bases as the read length increases (Figure 2.4). In an attempt to filter erroneous sequences from dataset, it is often required to remove or trim a subset of bases from the 3' end of each read. In the case of paired-end sequencing, the reverse reads also tend to have lower quality values associated with the bases when compared to the forward reads (Table 2.5). Trimming the last few bases from the 3' end of the reads can improve the number of reads that aligns to a target sequences (read mapability), prevent the occurrence of false positives during SNP identification, and prevent misassembled contigs. The effect of read trimming will be further addressed in the sections regarding *de novo* assembly (Section 2.3.3) and read mapping to a reference genome (Section 2.3.5 on page 73). A good guideline for trimming the reads is to use an error rate of 1 in 100 bases during assembly and read mapping, which translates to a Phred quality score cutoff of 20. Several tools already exists in the public Galaxy server to trim the end of

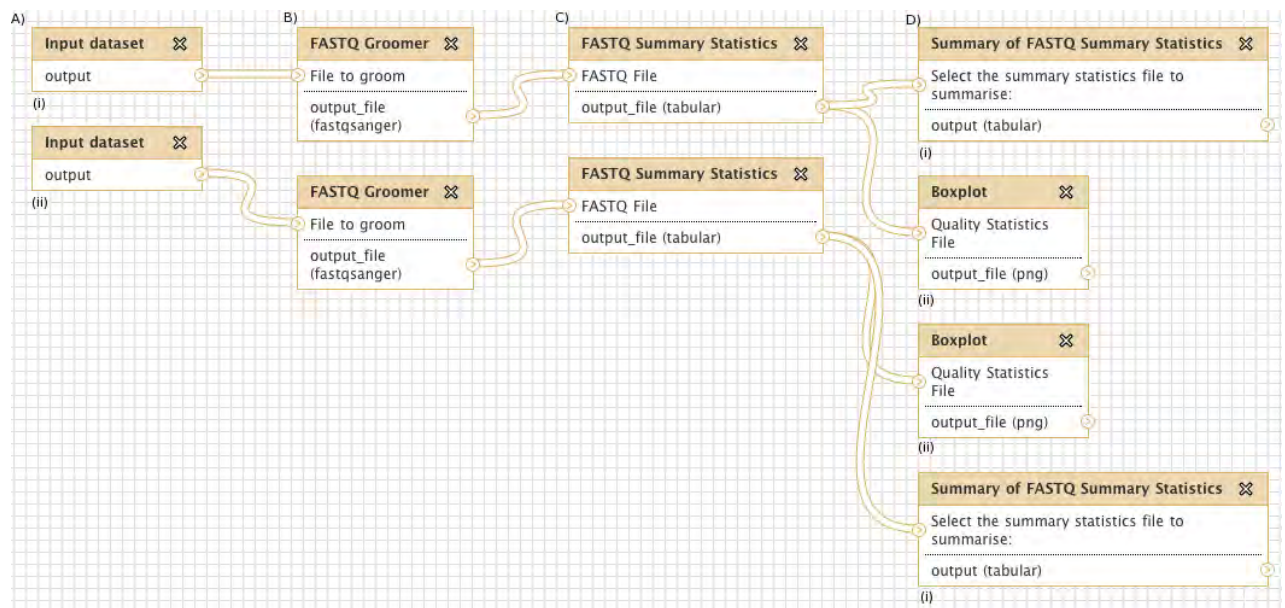


Figure 2.3: The Illumina read quality assesment pipeline. The first step after defining the input datasets (two FASTQ formated files (A), one that consists of the forward reads (A i) of a paired end run, and that consists of the reverse reads, A ii) is to convert the FASTQ values from the Illumina (1.3+) version to the FASTQSANGER format (FASTQ Groomer, B). Quality statistics per base are then calculated (FASTQ Summary Statistics, C), and a graphical summary of all the bases in the lanes produced (Boxplot, presented in Figure 2.4, D ii). From the FASTQ summary statistics, the number of reads and the number of bases present in each lane can be calculated (Summary of FASTQ summary statistics, Table 2.5, D i) .

Table 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. The theoretical yield was calculated as the total reads per lane times the read length. On average, the forward reads yielded 97.53% of the theoretical bases to pass the internal quality control performed by the sequencing center, while 96.83% of the reverse bases were useable. If seven usable lanes are considered per flowcell, an estimated 42 Gbp would have been produced from these lanes (please note that these lanes were not produced from the same flow cell).

Tissue	Read length	Total reads	Theoretical base yield	Useable base yield	Useable Gbp
Young leaf (a)	76 bp X 76 bp	38 675 726 (X 2)	2 939 355 176 (X 2)	5 714 978 949 (97.56% fwd, 96.87% rev)	5.71
Young leaf (b)	76 bp X 76 bp	40 644 094 (X 2)	3 088 951 144 (X 2)	6 005 687 472 (97.56% fwd, 96.86% rev)	6.01
Young leaf (c)	76 bp X 76 bp	40 603 294 (X 2)	3 085 850 344 (X 2)	5 999 955 671 (97.57% fwd, 96.86% rev)	6.00
Xylem (a)	76 bp X 76 bp	40 626 119 (X 2)	3 087 585 044 (X 2)	6 001 212 765 (97.54% fwd, 96.83% rev)	6.00
Xylem (b)	76 bp X 76 bp	41 212 187 (X 2)	3 132 126 212 (X 2)	6 084 735 293 (97.50% fwd, 96.76% rev)	6.00
Xylem (c)	76 bp X 76 bp	38 363 392 (X 2)	2 915 617 792 (X 2)	5 664 669 869 (97.48% fwd, 96.81% rev)	5.66

the reads based on either read length (the `FASTQ Trimmer` by column tool in `Galaxy`), or base quality (`FASTQ Quality Trimmer` by sliding window).

### 2.3.3. *De novo* transcriptome assembly using Illumina mRNA-Seq data

One of the main aims of this study was to perform a *de novo* assembly of a gene catalog from mRNA-Seq data generated from a range of primary and secondary *Eucalyptus* tissues (Chapter 3). A *de novo* assembly pipeline to achieve this goal typically consists of firstly formatting the input data to satisfy the requirements of the assembler, secondly perform the assembly, and finally evaluate the assembly (Figure 2.5, which is available as the “`Velvet assembly`” pipeline in the BCBU server). `Velvet` (Zerbino and Birney, 2008), the assembler used in this workflow, requires paired-end reads to be in a format where the first read of a fragment is directly followed by the second read of the fragment, as opposed to some other assemblers which require the reads from the same fragment to be in the same order, but in two different files. The “`shuffleseq`” tool, a script provided with the `Velvet` assembler and used to create the single file format, was wrapped in the BCBU `Galaxy` environment to allow for workflow integration (Section 2.3.1).

Input parameters of note that are specified for use during the graph-creation step of the `Velvet` assembly include the choice of kmer (Section 1.4), and the flag that specifies whether the input datasets are in paired-end format. During the graph traversal step, the expected coverage parameter and a coverage cutoff parameter is specified. The coverage cutoff parameter is used by the assembler to restrict highly connected nodes in the graph (repeat regions) from dominating the assembly. Changing each of these parameters results in differences in the properties of the final set of contigs produced from an assembly (see Section 1.4 for an overview of graph based *de novo* assemblers).

Currently no standardised protocol exists for steps needed to evaluate the success of a transcriptome assembly. Unlike the assembly of a genome sequence, where the aim is to assemble a single contig from all the reads provide, the aim of a transcriptome assembly can be viewed as the assembly of multiple, short fragments that represent mRNA molecules. The coverage of genome derived data is also distributed more

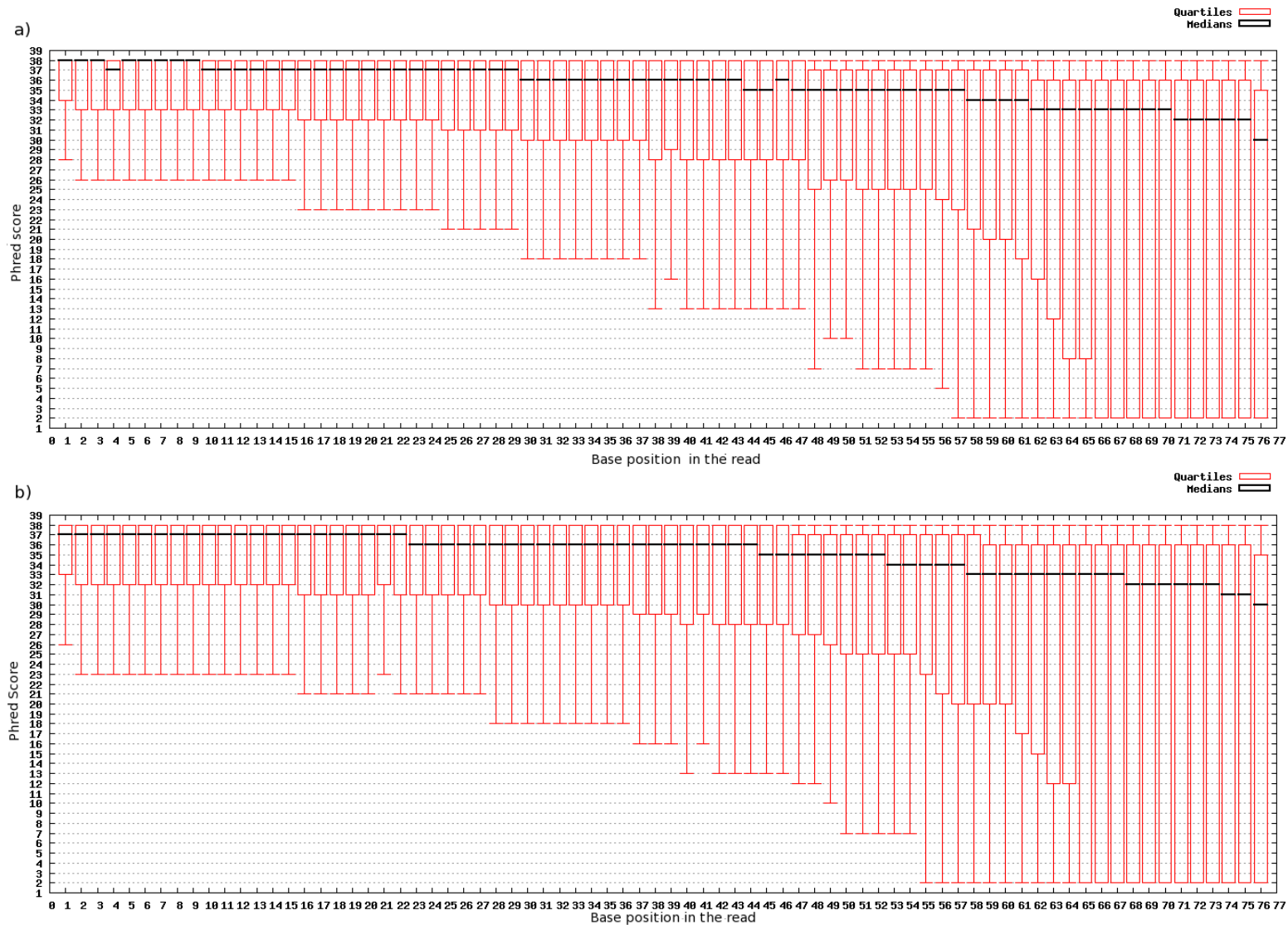


Figure 2.4: An example of FASTQ quality scores obtained from a 76 bp Illumina GAII paired-end run. The quality of each base is plotted on the y-axis, with the position of the base on the sequence on the x-axis. This lane contained around 38 million reads (2.8 billion bases) in the forward (a), and 38 million reads in the reverse (b) direction. The median (black line) and the standard error bars (red bars) for all the reads are shown in both directions. A quality drop is observable for bases closer to the 3' end (sharp increase in base-quality variation from base 56-58) and removing these bases with lower qualities might influence read mapping and assembling strategies.

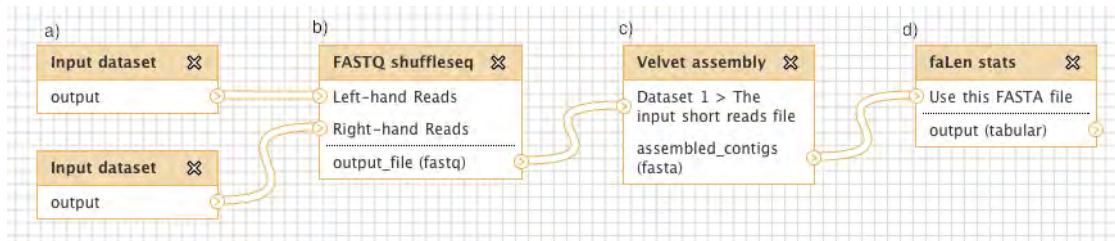


Figure 2.5: A Galaxy workflow which performs a *de novo* assembly with the Velvet assembler. The default input data (a) for this workflow is the forward and reverse FASTQ (fastqsanger) formatted mRNA-Seq reads. The reads are then reformatted with a "shuffleseq" script (b) to the correct input format for paired-end reads as required by Velvet, and the assembly is performed by Velvet (c). A script to calculate the N50, longest, mean and average sequence lengths is then run on the assembled fasta file.

evenly across the genome, with exceptions of the repeat regions, while the transcriptome data has varied coverage across a single transcript and between multiple transcripts. The variation in transcript coverage fluctuates due to the number of transcripts present in then sample mRNA pool, and the variation across a transcript has been postulated to be due to the folding patterns of the mRNA transcripts in the cell (Mortazavi *et al.*, 2008). There are several descriptive statistics available to assist in selecting the best possible assembly, namely the number of bases in the contigs (sum), the number of contigs (N), the contig length spread (minimum and maximum contig length, 1st and 3rd quartile length, mean and median length), and the N50 value. The N50 value is calculated as the contig length where 50% of the bases in the assembly are present in contigs of the reported length, or longer. A scoring function to empirically select the best assembly has been discussed on the Velvet users group mailing list<sup>2</sup>, and defined as:  $\frac{(N50_{all} * N_{long})}{Sum_{all} + \log(Sum_{long})}$ , where the long values are calculated for contigs longer than 1 000 bp. A higher score indicates a higher ratio between the bases located in the longer reads in the dataset and the bases assigned to short contigs. This scoring metric was also discussed on the community portal SeqAnswers<sup>3</sup>, and later implemented in an optimisation script for Velvet as a third party script, and although this scoring function has been defined for genome assemblies, it provides a good guideline when applied to transcriptome assemblies. In the sections discussed below, the score of the assemblies were calculated with the scoring function to give an indication of the function's performance on multiple assembled datasets.

<sup>2</sup> <http://listserver.ebi.ac.uk/mailman/listinfo/velvet-users>

<sup>3</sup> <http://seqanswers.com>



Table 2.6: Velvet assembly statistics for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue reads trimmed to different lengths (50 - 76 bp). The same assembly parameters (kmer 41) were used to illustrate the effect of sequence length on the assembly. Assemblies with the longest reads as input (65, 70 and 76 bp) generated the largest (N) assemblies, and the longest single contigs (max) were assembled with the 65 bp reads. The scoring function also indicates that the longer input reads generate better assemblies, except when the last 6 bp which were error prone are included. The 1 000 bp contig values (long contigs) used in the scoring function are presented in the Appendix I table A.1.

Read length (bp)	Number of contigs (N)	Sum of bases	Min (bp)	1st Quartile (bp)	Median (bp)	3rd Quartile (bp)	Max (bp)	Mean (bp)	N50 (bp)	Score
50	73 762	21 723 533	81	130	183	342	6 772	294.51	411	6.63
55	104 471	32 014 867	81	122	171	349	8 078	306.45	486	6.95
60	134 970	39 632 149	81	111	163	323	8 241	293.64	467	7.05
65	169 960	46 302 130	81	102	156	293	11 008	272.43	414	7.06
70	207 383	52 321 544	81	95	151	269	8 573	252.29	362	7.03
76	255 609	59 076 999	81	92	148	247	8 985	231.12	308	6.95

Low quality bases are generally present in the 3' end of Illumina reads (see Figure 2.4), and removing or trimming these reads tend to influence the subsequent assemblies. Assemblers using the *de Bruijn* graph approach, where kmers are used to find joins between reads and the high coverage paths between kmer nodes in the graph are used to assemble the contigs, have a higher tolerance towards low frequency erroneous bases in the input dataset (see Section 1.4). There also exists uncertainty about the optimal read length required to perform *de novo* transcriptome assemblies, and since longer reads require more reagents that influences the cost of sequencing this is an important consideration in project planning. Illumina mRNA-Seq paired-end reads from a deeply sequenced *Eucalyptus* xylem dataset were trimmed to a length ranging from 50 bp to 76 bp. The trimmed datasets were then assembled with the **Velvet** assembler (**Velvet assembly workflow**) with a defined kmer of 41 to determine the length of the input dataset reads that produced the best assembly. Table 2.6 indicates that longer reads produce longer individual contigs, but there is a decrease in overall assembly quality when the last 6 bp (low quality bases) of the 76 bp reads are not trimmed from the input dataset. The 55 bp assembly showed the largest N50 and the longest mean and median contigs, but if the additional  $\approx 7$  Mbp of sequence data gained



Table 2.7: Statistics for **Velvet** assembled contigs with a minimum length of 200 bp for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue reads trimmed to different lengths. The values in parentheses indicate the same statistics obtained with the same dataset, but where the datasets were treated as single and not paired-end reads. The 1 000 bp contig values (long contig) used in the scoring function for the single end assemblies are presented in the Appendix A Table A.1.

Read length (bp)	Number of contigs (N)	Sum of bases	Min (bp)	1st Quartile (bp)	Median (bp)	3rd Quartile (bp)	Max (bp)	Mean (bp)	N50 (bp)	Score
50	33 475 (31 519)	16 411 541 (14 581 289)	200	268 (245)	365 (328)	570 (527)	6 772 (5 571)	490.26 (462.62)	562 (535)	6.68 (6.63)
55	42 934 (43 283)	23 989 757 (22 004 217)	200	278 (253)	403 (248)	672 (577)	8 078 (8 078)	558.76 (508.38)	693 (615)	7.03 (6.95)
60	49 152 (50 771)	28 489 587 (26 957 786)	200	275.5 (258)	407 (359)	689 (603)	8 241 (8 241)	579.62 (530.97)	733 (653)	7.19 (7.08)
65	55 059 (56 990)	31 759 222 (30 633 000)	200	272 (260)	398 (366)	676 (610)	11 049 (11 049)	576.82 (537.52)	730 (660)	7.23 (7.14)
70	60 039 (61 683)	34 307 077 (33 463 851)	200	270 (262)	394 (371)	662 (615)	11 008 (10 757)	571.41 (542.51)	718 (664)	7.25 (7.18)
76	64 713 (65 989)	36 602 687 (36 070 026)	200	268 (264)	389 (375)	652 (621)	9 925 (10 873)	565.62 (546.61)	705 (669)	7.26 (7.22)

by the 60 bp, or the additional  $\approx 14$  Mbp of data when the 65 bp input dataset is considered, those assemblies can certainly be considered when evaluating an assembly. The scoring function calculated on these datasets provide a ranking system for the assemblies, but ultimately the choice of read length depends on the discretion of the researcher. Assembled contigs of a length between 81 bp and 200 bp most likely consist of small fragments of larger contigs, or very rare low coverage transcripts, and an additional constraint can be applied to the assembled dataset that contigs need to have a least a length of 200 bp to be considered for downstream analysis and annotation (Table 2.7). Because the **Velvet** assembler was developed for the *de novo* assembly of genomes, not transcriptomes, alternative spliceforms will be lost during assembly since the assembler returns the longest graph of the most coverage in the final assembly.

The assembly of the various trimmed datasets were repeated with the two lanes of the paired datasets provided separately to the assembler, effectively re-formatting the input data as two single-end datasets as oppose to a single paired-end dataset (Table 2.7, results in parentheses). Overall the single-end reads

did not perform worse than the paired-end assemblies, and even produced the same maximum length contigs in some cases. There is, however, a sampling bias in the data used for this single-end assembly, since the single-ends are not independently sampled fragments from the sequenced mRNA-Seq pool, but in fact represent sampled paired sequences. This simulated assembly of single end data thus does not represent the true effect of sequencing single-end *vs.* paired-end libraries, but rather reflects the difference in the assembler algorithm and the improvement achieved when enabling the paired-end flags. These values represent the practical best case scenario when single-end reads are used for assembly, and real independently sampled single-end assemblies will thus perform worse than reported here.

The graph traversing step of **Velvet** has multiple parameters that will ultimately affect the set of contigs assembled. One of the most notable parameters is the effect of kmer size (kmer of 41 - 63 bp) on the different assemblies, as presented in Table 2.8. The choice of kmer for assembly will vary with a change in length of the input reads, as well as the inherent sequence properties of the tissue or organism sampled. The scoring function defined above relates well to the a combination of the N50 value and the descriptive statistics of the assembly, and plotting the different assembly statistics as a fraction of the highest value of each parameter show that the scoring function can be successfully used as a guideline to select the best assembly for further analysis (Figure 2.6). The figure makes use of a normalised value for some descriptive statistics (N50, Sum and Score in Figure 2.6A) achieved during a specific kmer assembly according to the maximum value obtained (y-axis) across all kmers (x-axis), and can be used to graphically select the set of kmers that produce an assembly with a high score. The kmer of 51 (k51) produced an assembly containing 69 485 contigs, ranging from 200 bp to 8 451 bp in length. The scoring algorithm assigned a score of 7.13 to the k51 assembly, but the k49, k53 and k55 assemblies also achieved a high score. The best choice of a kmer to use in further assemblies depends on whether full length transcripts were assembled during any of these kmer assemblies, but the scoring algorithm does provide some measure of comparison between the assemblies.

The effect of two additional parameters during the graph traversal step, the expected coverage and coverage cutoff value, on the results from multiple assemblies is presented in Figure 2.7. The expected

Table 2.8: Velvet assembly statistics for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue. The same input parameters were used, except for the kmer-value to obtain these assemblies. Note a general trend that fewer contigs (N) and fewer total bases (Sum) are present in higher kmer assemblies, indicating that more contigs might be joined with longer kmers. The descriptive statistics in terms of median, mean and N50 values peak around the mid kmer (k49-k55) sizes. The assembly score was calculated to critically evaluate overall score of an assembly. All contigs longer than 200 bp were included in the analysis.

KmerNumber of contigs (N)	Sum of bases	Min (bp)	1st Quartile (bp)	Median (bp)	3rd Quartile (bp)	Max (bp)	Mean (bp)	N50 (bp)	Score	
k41	84 428	38 627 991	200	249	334	523	8 985	457.53	518	7.00
k43	81 527	38 434 796	200	250	339	538	8 862	471.44	543	7.05
k45	78 748	37 908 219	200	250	342	548	8 451	481.39	560	7.08
k47	75 732	37 110 906	200	250	345	557	8 451	490.03	576	7.10
k49	72 320	36 115 097	200	249	349	573	8 451	499.38	598	7.12
k51	69 485	35 124 810	200	250	351	581	8 451	505.50	613	7.13
k53	66 029	33 652 392	200	249	353	587	8 065	509.66	621	7.12
k55	62 391	31 953 361	200	248	351	593	8 582	512.15	632	7.12
k57	58 921	30 071 960	200	247	350	593	8 277	510.38	631	7.09
k59	54 966	27 877 831	200	246	349	591	9 622	507.18	626	7.04
k61	51 057	25 518 959	200	245	346	585	7 152	499.81	613	6.98
k63	46 684	22 658 706	200	244	338	563	6 360	485.36	585	6.89

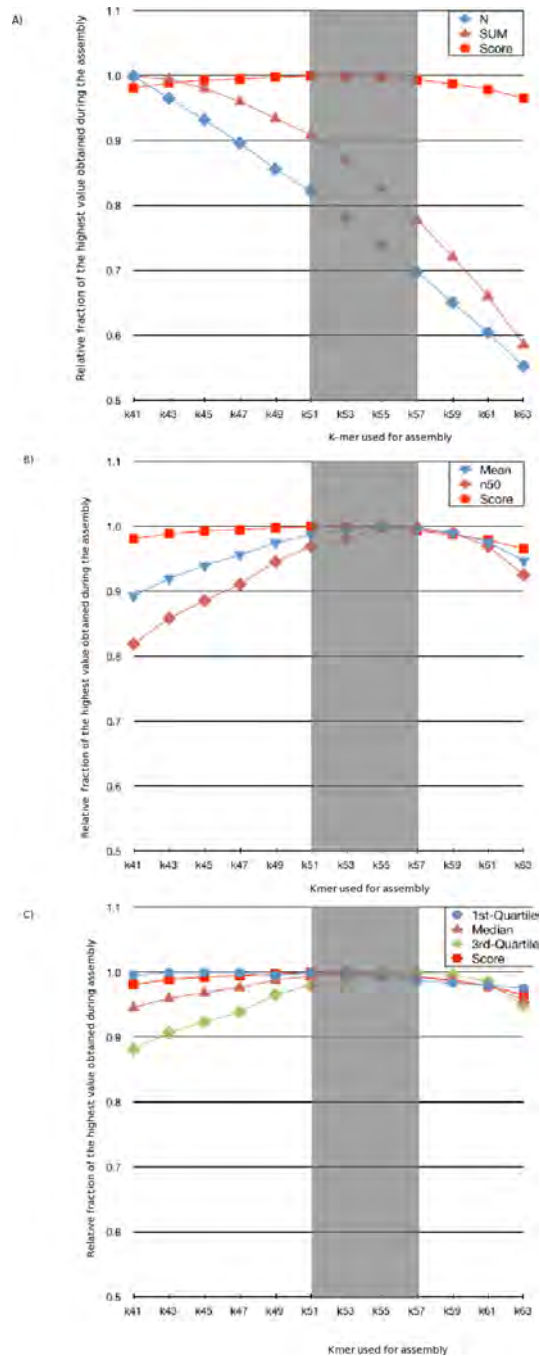


Figure 2.6: The assembly scoring function is a robust measure to select the kmer of the best Velvet assembly. The y-axis represents the value of a certain descriptive statistic obtained for a kmer as a fraction of the maximum value of that statistic (y-axis) across all kmers (x-axis). The scoring function is not sensitive to changes in total base count and number of contigs (a), and correlates well with the N50 and mean values (b) as well as the other descriptive statistics (c). The graphs were normalised so that the values correspond to a fraction of the maximum value achieved for each parameter across all kmer assemblies shown.

coverage parameter performs two key functions during the assembly. First, it is required to activate the paired-end read resolution function of *Velvet* (as stated in the *Velvet* manual), which programatically makes use of the insert size between pairs to join contigs; and secondly it assists in finding the optimal path through the nodes in the graph of kmers by searching for nodes in the graph that correspond to the expected coverage value. This assistance provides the assembler with a naive approach to filter the nodes in the graph based on the node coverage in order to determine optimal contigs (Zerbino and Birney, 2008). This approach is especially useful when a genome sequence is assembled, since the sequence coverage from a lane of genomic short-read data should have near uniform coverage, bar the repeat regions of the genome that should have higher coverage. The inherent properties of mRNA-Seq data, where coverage varies between transcripts based on the amount of transcript present in the sampled mRNA pool and across a single transcript based on the mRNA molecule's folding properties, the occurrence of alternative splicing, and the known 3' bias exhibited by mRNA-Seq technologies render this parameter less useful during transcriptome assemblies.

Figure 2.7A (left), indicates that for a transcriptome assembly, high expected coverage values produce the best possible assembly when evaluating the results based on the scoring function. The results were obtained by performing various assemblies with a constant set of parameters (insert length between paired reads = 150 bp, the coverage cutoff = 10X, and the kmer set to 51), but increasing the expected coverage value from 0 to 1 000 with each subsequent assembly. The graph shows that a higher expected coverage value can produce assemblies with longer mean length and N50 values (an expected coverage of 0 produced an assembly with an average N50 length of 1 018 bp, only 55% of the N50 value achieved by the assembly where the expected coverage was set to 1 000 (N50 = 1 854 bp)). These estimations of the expected coverage value are needed to assemble highly expressed transcripts to a complete length, and will remove lowly expressed transcripts from the assembly.

The coverage cutoff value effectively screens the contigs after graph generation, removing contigs that do not meet the minimum coverage cutoff value as specified. This parameter removes short, low coverage contigs from the assembly, and in general improves the assembly when set to a reasonable value

between 4 and 10 (Figure 2.7b). Setting the value too high will remove highly covered and good quality contigs, while a too low value will include short, low covered contigs which most likely originated from nucleotide errors in the sequence, or contain low covered introns that were captured when unprocessed mRNA molecules were selected before sequencing.

Varying the parameters used during an assembly has a measurable effect on the total number of contigs, the average contig length and the number of bases present in a transcript assembly. The quality of a transcriptome assembly is, however, not based on the global properties of the assembly, but on the presence of near complete or completely assembled cDNA transcripts in the assembly. By using known, well studied, full-length cDNA sets of genes the corresponding transcripts in the assembly can be evaluated. Figures 2.8, 2.9 and 2.10 presents six *Eucalyptus grandis* cellulose synthase (CesA) genes (Ranik and Myburg, 2006), and the results of performing a BLAST ( $e^{-100}$ ) of the CesA genes against assemblies from kmer 41 (Figure 2.8), kmer 51 (Figure 2.9) and kmer 61 (Figure 2.10) presented in Table 2.8. The CesA sequences (DQ014510.1, DQ014509.1, DQ014508.1, DQ014507.1, DQ014506.1 and DQ014505.1) are connected with colored banners of high similarity to regions present in contigs in the assembly dataset. Each CesA sequence can have similarity regions on multiple contigs present in the assembly. A perfect assembly will have a one-to-one ratio of CesA sequence to assembled contig with both sequences showing similarity along the entire length of the transcript. A subset of these CesA genes have been shown to have high expression in either primary or secondary cell formation tissues (Ranik and Myburg, 2006), and since these assemblies were performed with a single lane of xylem mRNA-Seq data, it can be expected that the lower abundant transcripts would not fully assemble. In order to select the best assembly parameters, a similar analysis should be repeated with different gene families that have a range of expression across multiple tissues.

#### **2.3.4. Annotating assembled transcript sequences**

Several good EST annotation pipelines exists in the public domain. These pipelines consists mainly of a set of scripts that calls a subset of tools sequentially to annotate a set of protein or DNA sequences. Few

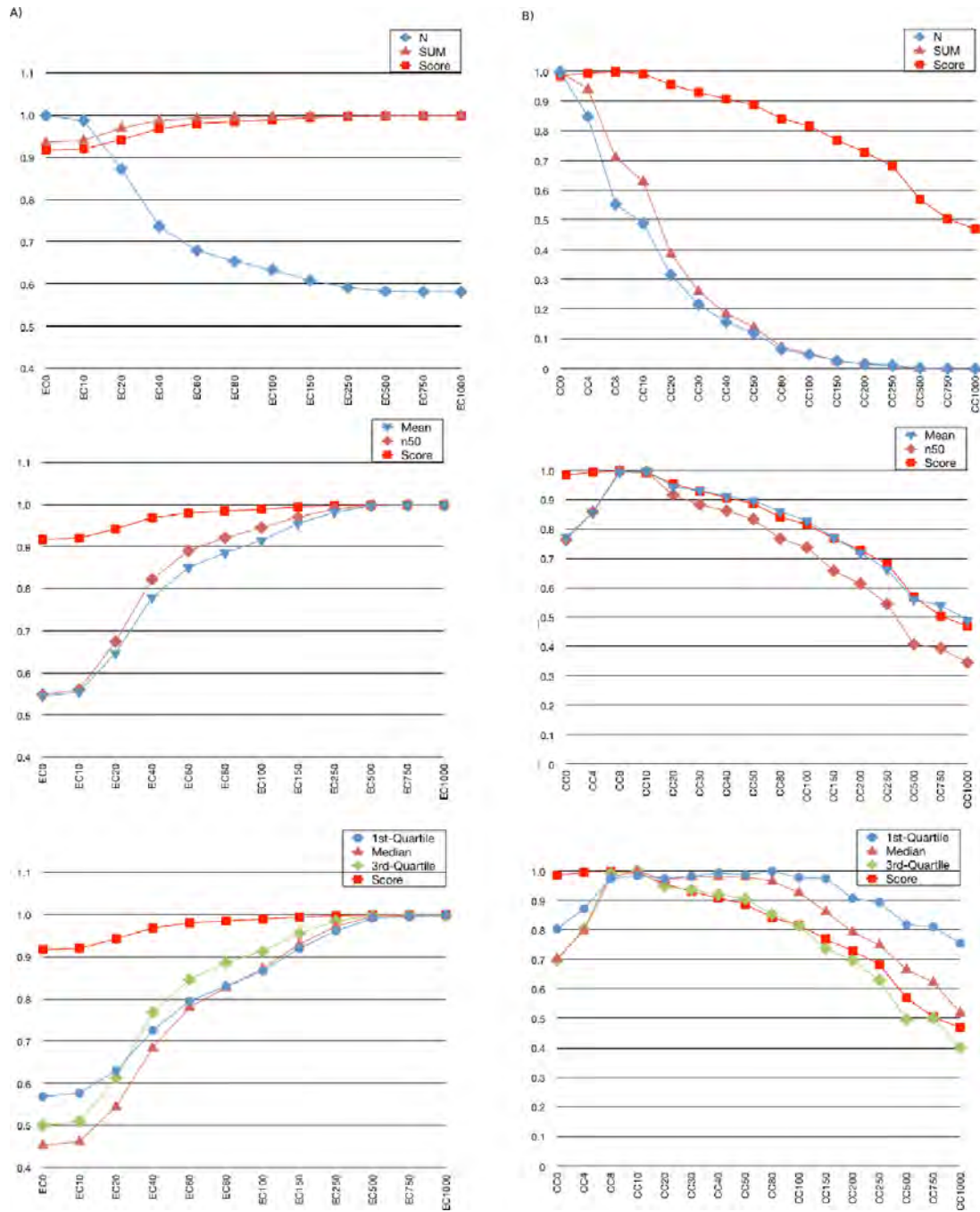


Figure 2.7: The effect of the expected coverage and the coverage cutoff parameters on a **Velvet** assembly. Due to the large dynamic range in transcript expression, high expected coverage values (A, left) produce the highest scoring assemblies. For the coverage cutoff parameter, it was found that the best **Velvet** assembly is achieved when the coverage cutoff parameter (B, right) ranges between 6 and 10. This will effectively remove low coverage contigs from the assembly while not removing the higher covered, longer contigs.



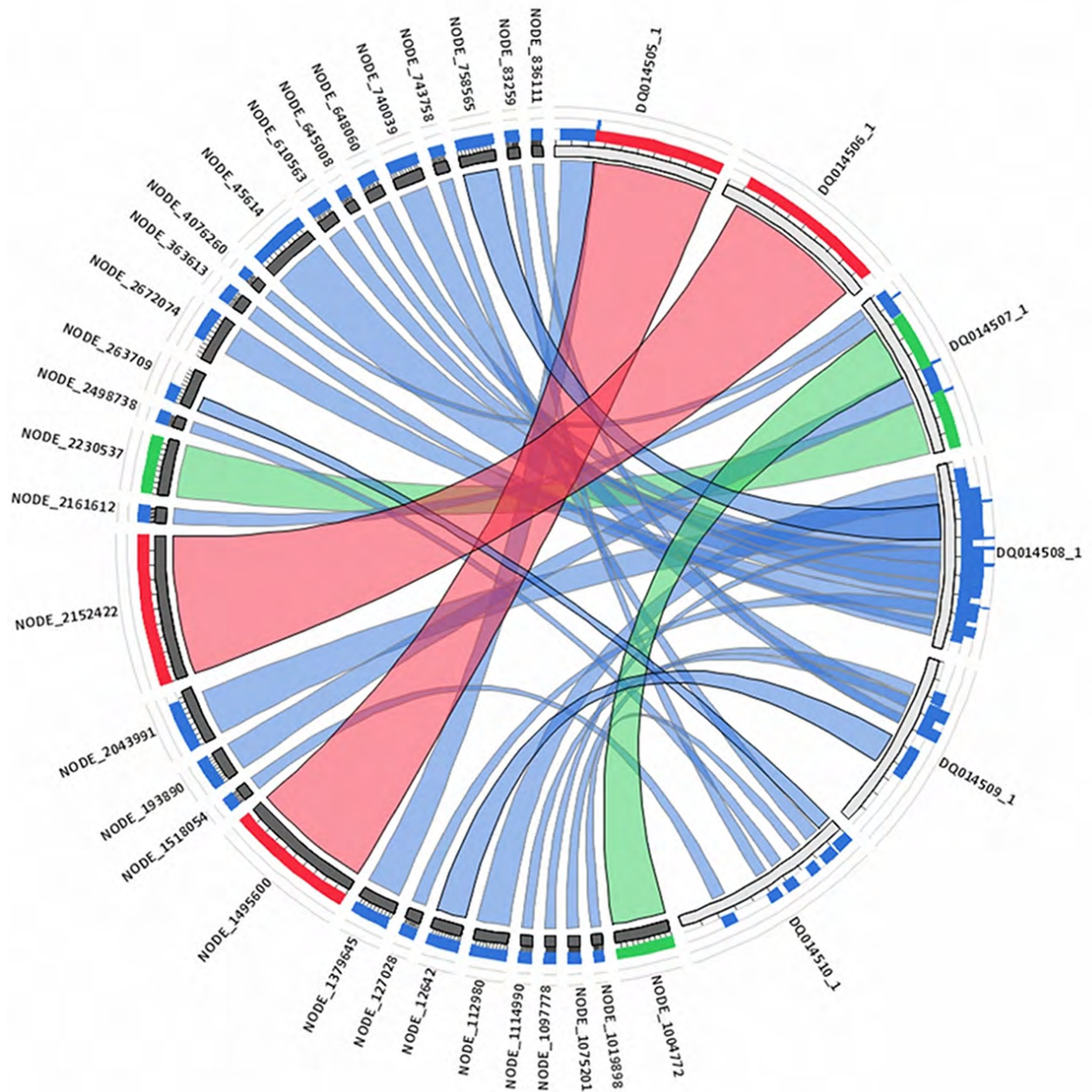


Figure 2.8: Alignment of the six full length CesA cDNA sequences against an assembly with a kmer size of 41 (k41). The CesA cDNA sequences (identifier starts with "DQ") show similarities to various contig sequences (identifiers "NODE") in the assembly. Blue ribbons indicate regions where the bit score of the alignment is < 25% of the maximum bit score in the dataset. Warmer colors (25% > green <= 50%, 50% > orange <= 75% and red > 75%) indicate higher bit scores. The two CesA cDNA sequences, DQ014506\_a and DQ014505\_1 are presented by near full length contigs NODE\_2152422 and NODE\_1495600. The cDNA sequence DQ014507\_1 is represented by two large contigs (NODE\_2230537 and NODE 1004772), while the remaining cDNA sequences are represented by various small contigs.



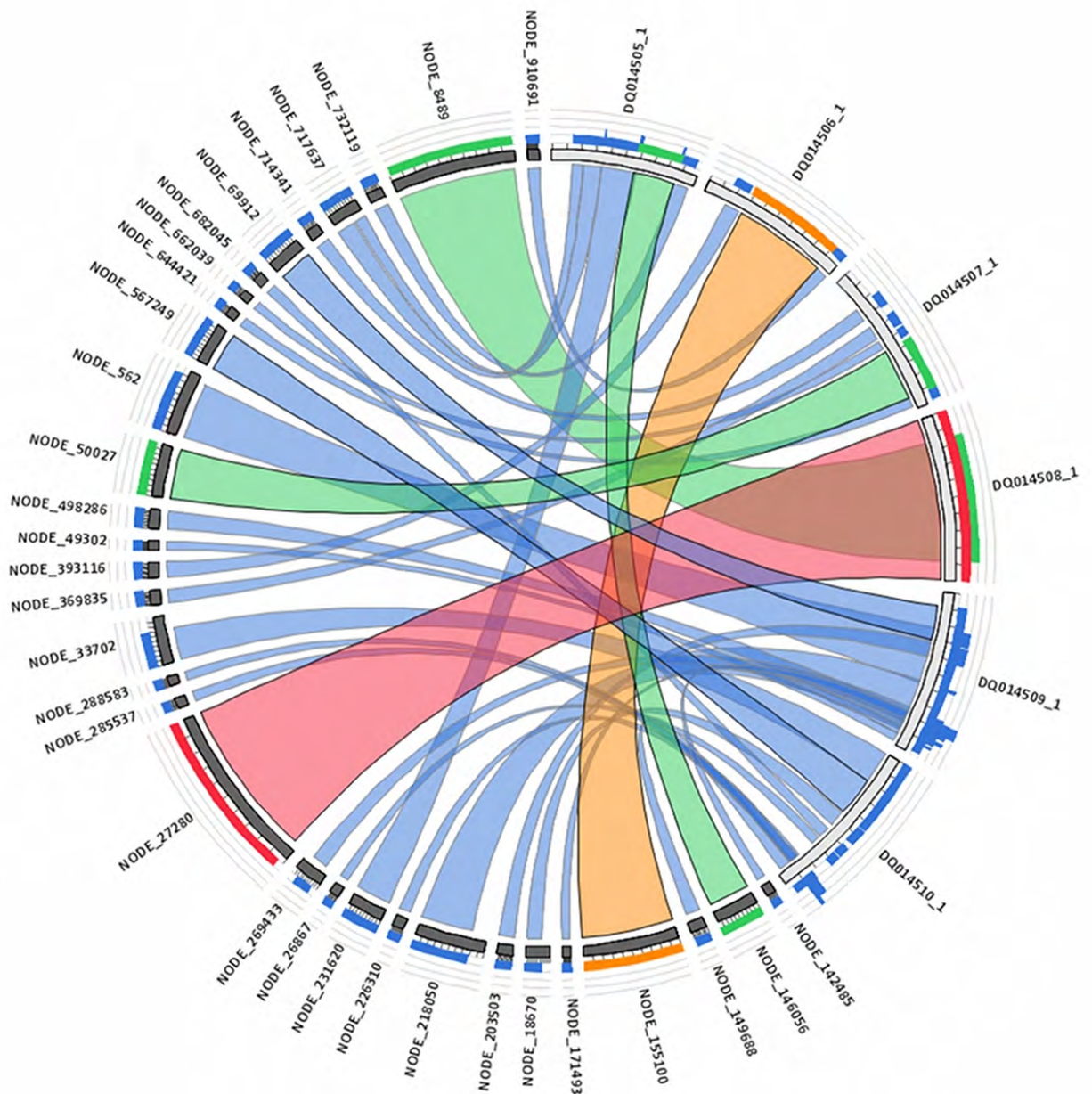


Figure 2.9: Alignment of the six full length Cesa cDNA sequences against an assembly with a kmer size of 51 (k51). The Cesa cDNA sequences (identifier starts with "DQ") show similarities (best BLAST hit) to various contig sequences (identifiers "NODE") in the assembly. Blue ribbons indicate regions where the bit score of the alignment is < 25% of the maximum bit score in the dataset. Warmer colors (25% > green <= 50%, 50% > orange <= 75% and red > 75%) indicate higher bit scores. The alignment indicate two copies of the cDNA sequence DQ014501\_1 in the assembly (NODE\_27280 and NODE8489). A partially assembled contig (NODE\_155100) that represent DQ014506\_1 can also be identified. The remaining Cesa's are represented by various shorter contigs in the dataset, indicating that there are still fragmented transcripts present in the assembly. The graph was generated with the Circoletto tool from the BLAST result file.

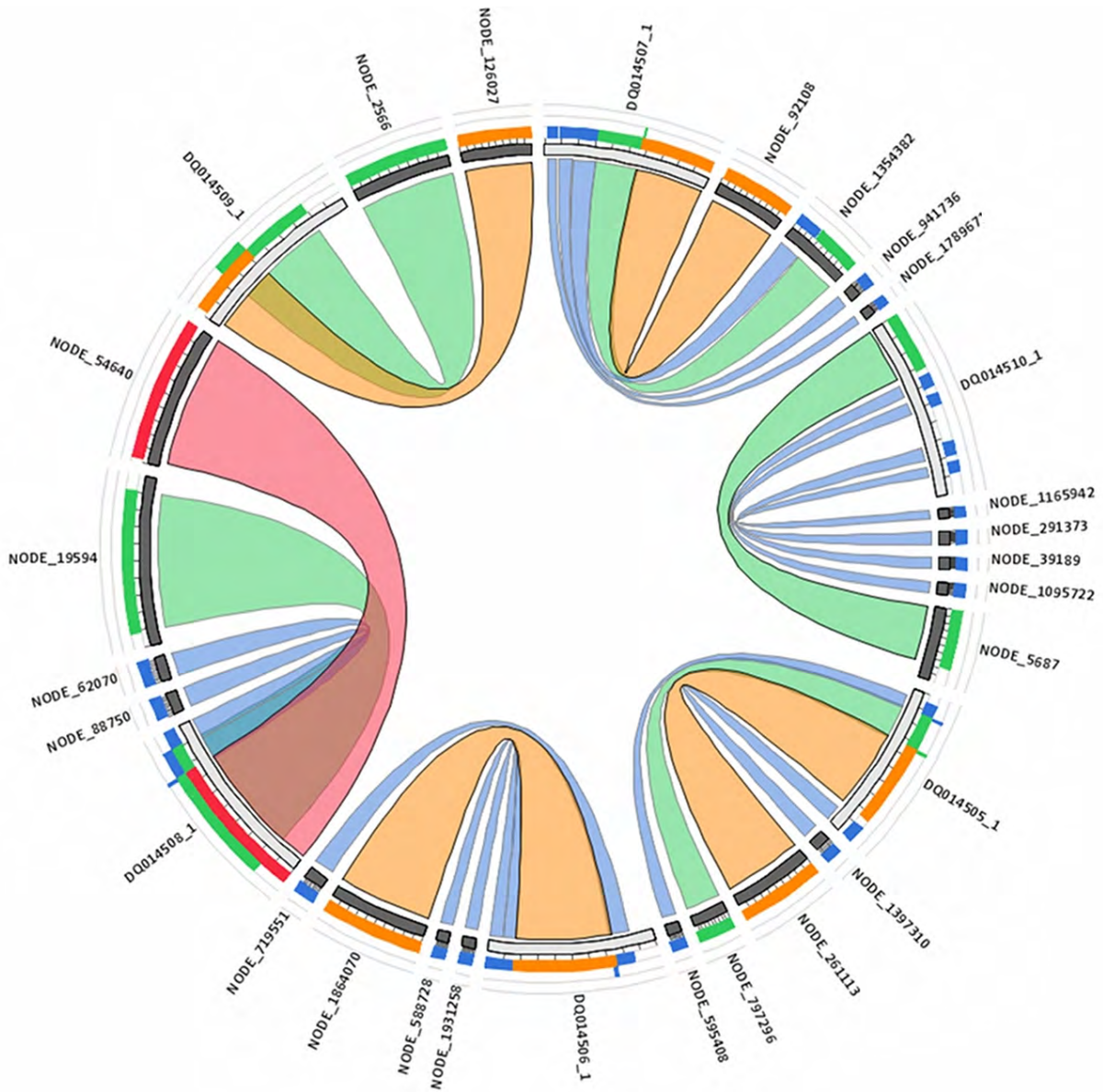


Figure 2.10: Alignment of the six full length CesA cDNA sequences against an assembly with a kmer size of 61 (k61). The CesA cDNA sequences (identifier starts with "DQ") show similarities (best BLAST hit) to various contig sequences (identifiers "NODE") in the assembly. Blue ribbons indicate regions where the bit score of the alignment is < 25% of the maximum bit score in the dataset. Warmer colors (25% > green <= 50%, 50% > orange <= 75% and red > 75%) indicate higher bit scores. The alignment represents the least fragmented assembly of the CesA cDNA sequences when compared to Figures 2.8 and 2.10. A duplicate assembled contig can be identified in the assembled dataset for sequence DQ014508\_1. Most of the remaining CesA cDNA sequences are represented by at least one or two large contigs in the assembly, although not all of them aligning across the whole length of the cDNA. The graph was generated with the Circoletto tool from the BLAST result file.



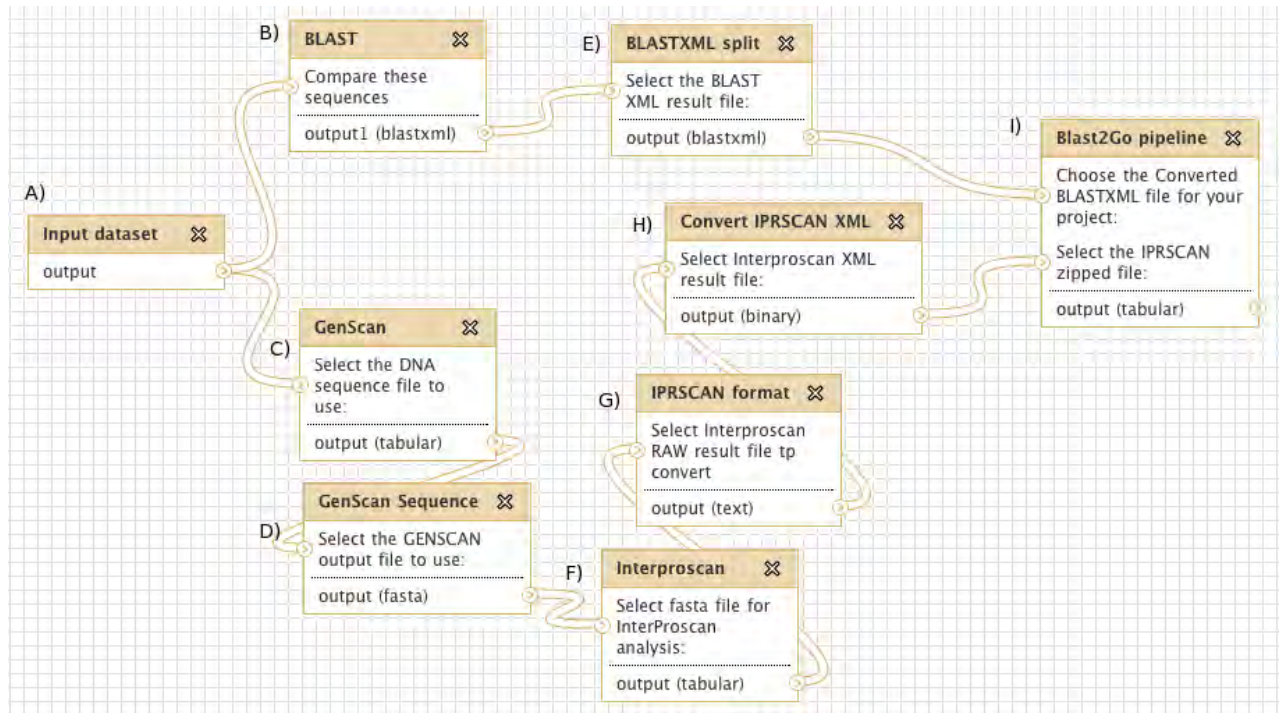


Figure 2.11: The automated annotation pipeline developed from tools available in Galaxy. The input for the pipeline (A) is a FASTA file containing cDNA sequence data. Protein sequence predictions are performed by GenScan (C) and the results converted to FASTA format, (D) and the resulting peptides submitted to the IPRSCAN pipeline. (F, G, and H) The input file is simultaneously submitted to BLAST (B and E) to perform homology searches (BLASTX), and the results of the IPRSCAN and BLAST searches used as input to the BLAST2GO pipeline (I) for further analysis.

pipelines allow the user to customise the different components used by the annotation pipeline specifically for the organism that is to be annotated. Two widely used tools, the InterProScan set of scripts and databases (Zdobnov and Apweiler, 2001), and the BLAST2GO annotation pipeline Conesa *et al.* (2005) were incorporated in the BCBU Galaxy server. The InterProScan annotation scripts and associated databases are often used to unknown protein sequences with protein feature, protein family and detected motifs present on the protein sequence. The BLAST2GO pipeline assigns functional annotations to the submitted cDNA or protein dataset, which consists of Gene Ontology, KEGG and InterPro accessions. An automated workflow (Figure 2.11, available as the "Annotation pipeline" workflow in the BCBU Galaxy server) was developed to use both these annotation pipelines to annotate a set of cDNA sequences from the transcriptome assembly pipeline described above.

The automated assembly workflow takes cDNA sequences as input (ESTs or contigs assembled from

mRNA-Seq data), performs a translation of the coding sequence into a putative protein and CDS sequence, and uses the predicted protein sequence to find protein family and protein feature annotations with IPRSCAN, the interface to the EBI's **InterProScan** tool. Results from IPRSCAN analysis are then converted to a format acceptable for the **BLAST2GO** annotation tool. The protein sequences are also used for a homology-based search against an external database (for instance, the NCBI's database of non redundant protein sequences), and the results parsed for use in the **BLAST2GO** annotation pipeline. **BLAST2GO** analysis is performed with the homology search (**BLAST**) and the IPRSCAN results as input, and an annotation (**.annot**) file is constructed. This **.annot** file can then be used as direct input into a **BLAST2GO** instance for the perusal of the annotations or imported into an external database.

The input sequences to the pipeline can consist of portions of genomic cDNA, full-length CDS or partial CDS sequences. The gene finder application, **GenScan** (Burge and Karlin, 1997) was used to predict a protein and CDS sequence from the input sequence. This is a very crude approach to cDNA translation and peptide sequence prediction, since partially assembled sequences will not have all the sequence signals present on the sequence required by **GenScan** to perform a reliable prediction of the exact intron and exon structure of the input sequence. This particular tool can, however, be replaced by any other gene prediction or cDNA translation tool in the workflow, as long as a protein sequence is the output from the alternative tool. The pipeline was used to perform a basic annotation of the 18 894 full-length, or partially assembled sequences of a *Eucalyptus grandis* x *Eucalyptus urophylla* transcriptome generated from mRNA-Seq data (Chapter 3).

The **InterProScan** analysis tool scans a given protein sequence against a range of protein signatures stored in the InterPro member databases. These signatures, present in the PROSITE, PRINTS, Pfam, ProDOM and SMART databases can then be used to provide functional annotations of the input protein sequence based on motifs present in the sequence. The **InterProScan** tool is a scalable and extensible system for protein feature annotation, and searches databases installed on a local server of the mentioned sources in order to find signature sequences. Results from the **InterProScan** analysis tool can be converted into XML, HTML or a TXT based file, which can be used to create a summary of the features

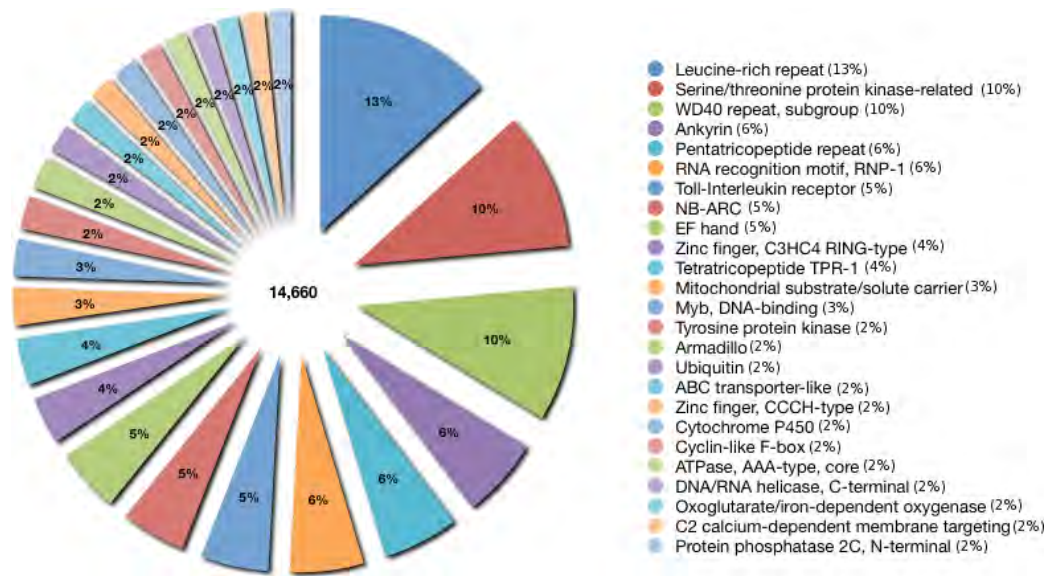


Figure 2.12: The 25 most prevalent protein family domains annotated in an assembled transcriptome dataset, expressed as a fraction of the total number of PFM annotations. The Leucine Rich Repeat (PF:PF00560) region was the annotation assigned in 13% of the annotations, and the dataset also represents annotations of kinases (PF00069 and PF07714) and the Myb transcription factor binding domains (PF00249). The figure was produced from the PFM annotations assigned to the 18 894 assembled contigs by the InterProScan tool.

found in the dataset on a global scale (Figure 2.12), or to view the signatures and features annotated on a specific sequence (Figure 2.13).

Various functional annotation projects use the Gene Ontology system to group sequences into related functional groups. The BLAST2GO annotation tool offers a wide range of statistical validations in assigning a functional classification to a protein sequence. The results from the annotation workflow produce an annotation file, generated by the command line interface (b2gPipe) of the BLAST2GO annotation tool. The pipeline expects BLAST XML results formatted in a specific manner, and a directory containing InterProScan XML results in order to complete the annotation. The BLASTXML2BLAST2GO and IPRSCANXML2BLAST2GO Galaxy extensions perform the simple conversions between the formats, and also execute the b2gPipe pipeline. BLAST2GO relies on a local installation of public Gene Ontology and Gene Ontology Accession databases to assign the Gene Ontology annotations to the sequences in the BLAST XML file. The annotation file produced can then be imported in a stand-alone version of the

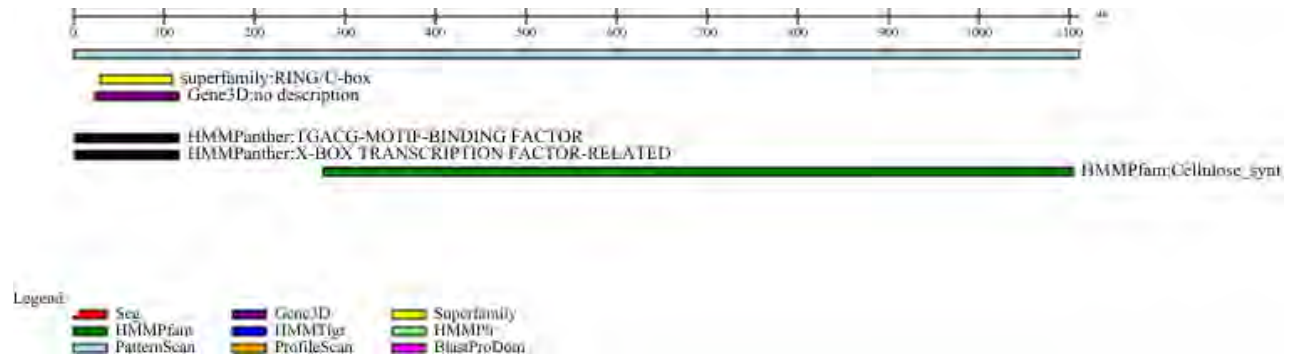


Figure 2.13: Protein features annotated by InterProScan present on the cellulose synthase 6 (CesA6) protein sequence assembled from reads derived from mRNA-Seq sequencing. The sequence represents the assembled contig with the highest homology to the Cesa6 (DQ014510.1) mRNA sequence, and was annotated by the InterProScan annotation pipeline. The annotation indicates the presence of a transcription factor binding motif (TGACC-motif, black box), a X-Box transcription factor-related motif (black box) on the 5' end of the sequence identified by HMMPanther. The same 5' region has also been identified as having a Ring/U-box superfamily signature (yellow box). The long green box represents the presence of the cellulose synthase protein family signature identified by HMMPfam. The image was generated from the RAW results by the InterProImageGenerator tool in Galaxy.

BLAST2GO tool, and can be used to summarise the overall ontology structure of the dataset, as well as inspect the annotations made to a single protein sequence.

### 2.3.5. Using mRNA-Seq data to calculate transcript expressions values

Many research groups have calculated gene transcript abundance levels with the aid of mRNA-Seq data (see Section 1.3 for a review of RPKM and FPKM calculations and other high-throughput sequencing applications in genetics and genomics). Mortazavi *et al.* (2008) showed that the differences in transcript abundance can span five orders of magnitude, and that the mRNA-seq methodology used was shown to be sensitive enough to detect even single copies of a transcript in a cell. A recent methods paper used mRNA-Seq data to detect novel transcripts and alternative spliceforms of transcripts, and was made available as the CUFFLINKS package (Trapnell *et al.*, 2010). CUFFLINKS performs a *de novo* prediction of splice junctions, and generates a set of detected gene models with their corresponding expression values (FPKM). The following section describes the workflow developed to detect transcript expression values for an organism where no annotated gene information is available (Figure 2.14, available as the "FPKM calculation" workflow in the BCBU server). The workflow starts of by mapping an input



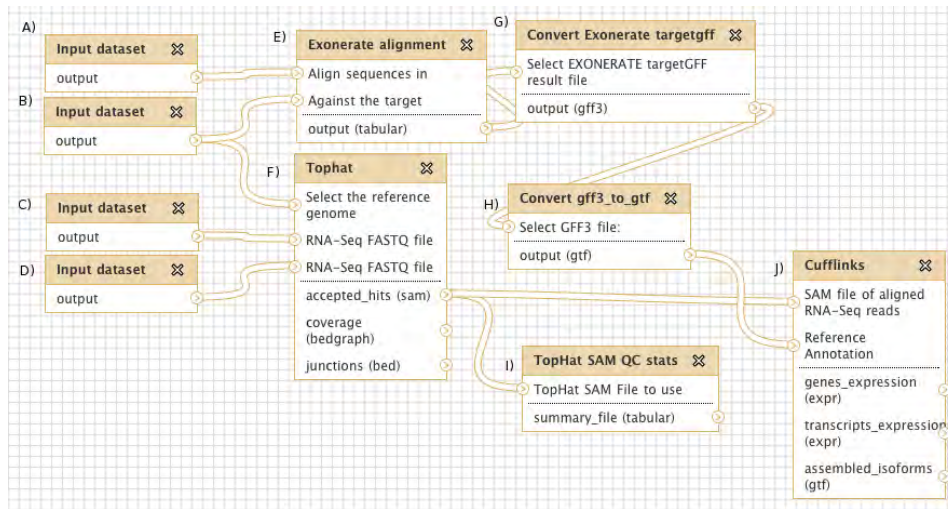


Figure 2.14: Calculating gene expression (FPKM) values for unigenes aligned regions from a genome with no gene models available. The input dataset for the workflow is a reference genome (B), the forward and reverse reads of an mRNA-Seq lane (C and D), and a FASTA file containing a set of ESTs (A). TopHat aligns the mRNA-Seq reads to the genome (F) and also against the splice junction regions using the Bowtie aligner. The alignment file (SAM format) is then used in calculating some short read mapping statistics (I), and as input for CUFFLINKS (J). The unigenes input dataset is aligned against the genome with EXONERATE (E), and the GFF output of EXONERATE is converted to the required GTF format (H) for CUFFLINKS. The GTF and SAM files are used to calculate the FPKM values (J).

set of mRNA-Seq reads to a target genome with TopHat (Trapnell *et al.*, 2009), as well as aligning a set of cDNA sequences to the genome with the EXONERATE aligner (Slater and Birney, 2005). The workflow further generates a gene model file from the cDNA alignment, and calculates the FPKM values for each of the transcripts present in the alignment.

The normalised transcript expression values (FPKM) are calculated by mapping reads to a target genome, constructing splice sites where reads span intron junctions, and then calculating the number of fragments that map per unit transcript. The TopHat mapping program (Trapnell *et al.*, 2009) was designed to determine the splice junction alignment when mapping to genome sequences. A single lane of 76 bp Illumina mRNA-Seq data was trimmed to shorter lengths and mapped to the *Eucalyptus grandis* draft genome sequence. Since longer reads require more reagents during sequencing, a key question to address is how a difference in read length influences the read mapability. Figure 2.15 indicates that there is an increase in the number of paired reads that map uniquely to a genome when the read length is increased from 40 bp to 50 bp, but beyond 50 bp there is not a marked difference in the number of paired

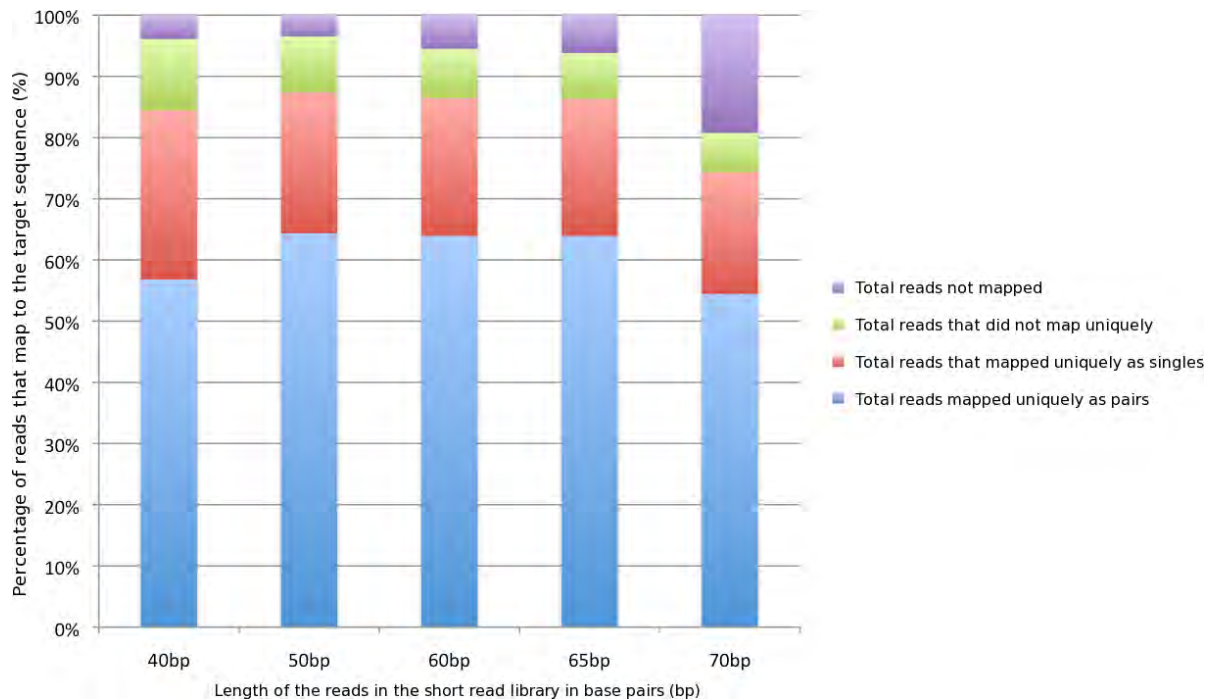


Figure 2.15: A breakdown of the number of reads which map uniquely, and non-uniquely as pairs or single reads to a target genome for different read lengths. No increase in read specificity can be detected when paired reads are longer than 50 bp in terms of unique paired mapping to the genome. Up to 97% (50-65 bp) of the reads were mappable to the genome, but this includes reads that map to regions outside gene models and within repeat regions. There is a significant increase in the number of reads that did not map to the genome when the read length was 70 bp.

reads that map to the genome. These results indicate that a paired read or fragment of 50 bp has a high enough specificity in the genome to map uniquely, and longer reads are not necessarily more specific. Reads longer than 70 bp shows a decrease in mappability, due to the stringency associated with the number of mismatches allowed when aligning a read to the target sequence. These mismatches have a higher probability to occur in longer reads, mostly due to the effect of sequencing errors in longer reads, but also due to SNPs present in a sequenced sample.

CUFFLINKS makes use of the genomic coordinates of genes or transcripts to calculate the FPKM expression value. The coordinates file needs to be supplied in the GTF (a condensed GFF3 file format) format to CUFFLINKS. The genome coordinates for a genome where no annotation, i.e. no GFF3 file exists, can be determined by performing a gapped alignment of cDNA sequences to the genome with EXONERATE. Output from EXONERATE needs to be reformatted to the GFF3 format and converted to the



GTF format before serving as input to CUFFLINKS. CUFFLINKS can calculate the FPKM values for the annotated genes present in the GTF file, or if no reference gene models are provided, it will identify new expressed transcripts.

Lists of genes and their expression values can serve as input to one of several statistical packages to determine groups of genes that are differentially expressed between experiments. The R package `DEGseq` (Wang *et al.*, 2010a) was used to determine a list of genes differentially expressed between immature xylem and young leaf tissue of a *Eucalyptus grandis* hybrid tree (Chapter 3). Figure 2.16 present the results from the `DEGseq` package used to determine differential expression. The figure presents the MA plot (where  $M = \log_2 \text{tissue}_1 - \log_2 \text{tissue}_2$ ,  $A = 1/2(\log_2 \text{tissue}_1 + \log_2 \text{tissue}_2)$ ) of differential expressed genes identified with a 2X fold change method to detect differential expression. The Venn diagrams below the MA plot shows the number of genes detected to be differentially expressed in immature xylem and in young leaf tissue, and the set of genes not being differentially expressed.

## 2.4. Conclusion

The management and data analysis of large DNA sequence datasets produced with high throughput biological experiments require sound data management principles, dedicated and sometimes specialised computational hardware, and a variety of software tools. The `Galaxy` framework was identified as one of many potential data management and automated data analysis workflow systems that can be used and adapted to analyse mRNA-Seq datasets. The framework can easily be extended to include new analysis tools, which can then be incorporated into complex workflows, which have the ability to make high throughput data analysis tools available to research groups. The framework effectively reduces the steep learning curve needed to master the command line interface of an analysis tool, by providing a web-based form to set the parameters used during the execution of the analysis program.

The quality evaluation of uHTS data is one of the first analysis steps when working with these datasets. The current Illumina pipeline (version 3.6) produced quality scores associated with each base of sequence in an format that differs from the standard Phred based format, which needs to be converted

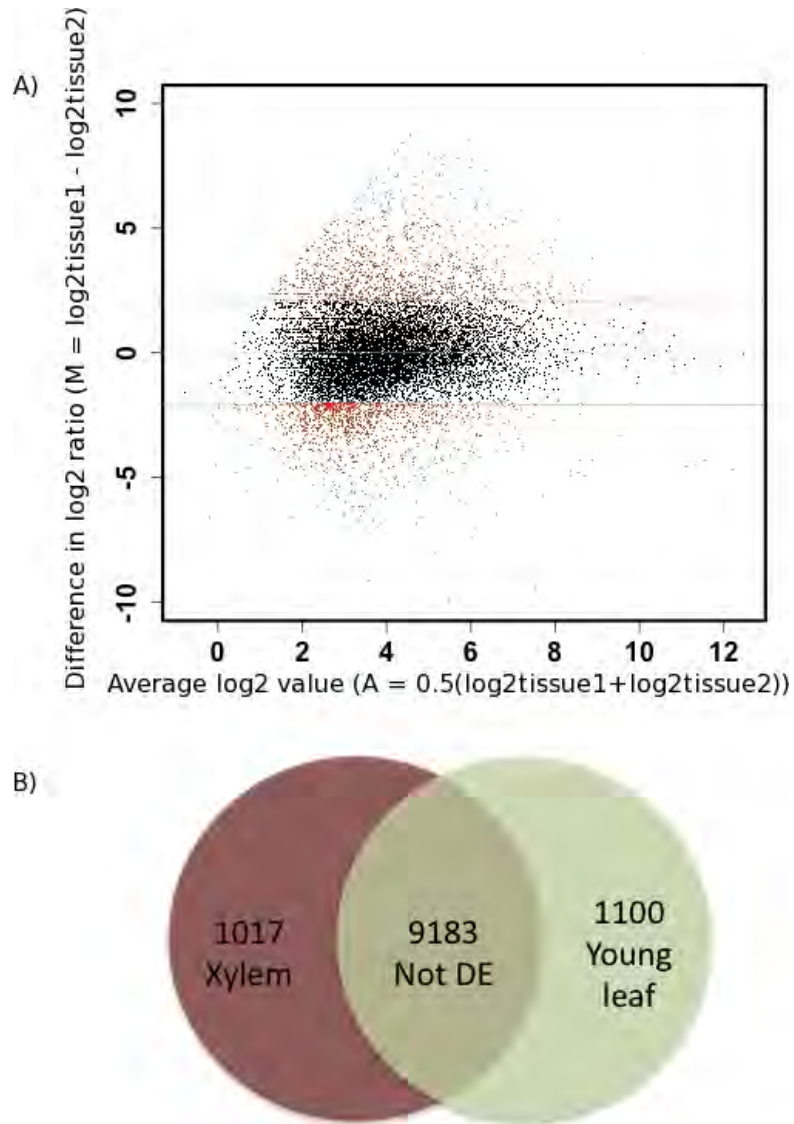


Figure 2.16: Genes identified as differentially expressed in immature xylem and young leaf tissues of a *Eucalyptus grandis* hybrid tree. The top figure (A) represent genes identified by the DEGseq tool as differentially expressed genes based on the MA (where where  $M = \log_2\text{tissue}_1 - \log_2\text{tissue}_2$ ,  $A = 1/2(\log_2\text{tissue}_1 + \log_2\text{tissue}_2)$ ) using a 2X fold change method. The Venn diagrams represent the same set of genes identified as being differentially expressed in immature xylem (brown) and young leaf tissue (green), and the genes that are not detected as being differentially expressed (Not DE, overlapping area).

to the standard **Phred** format. After conversion, a per base quality graph can be calculated for every base at every position of the read, and bases removed from the 3' ends of the reads. Depending on the amount of data available, it is recommended that a **Phred** quality value of 20 (base error rate of 1 in 100) is used as a guideline to trim the reads. Erroneous bases at the 3' ends of the reads have the ability to prohibit the alignment of a read to a target sequence as it increases the number of mismatches that will occur between a target sequence and the read, and also with graph-based assemblers it can create low coverage paths between the nodes of the graph. A default pipeline for the quality evaluation of short read Illumina data is available as the "Illumina QC" workflow in the BCBU **Galaxy** server installed at the University of Pretoria.

The assembly of a set of representative cDNA sequences from a pool of mRNA reads is still a challenging endeavor. A workflow which makes use of the **Velvet** assembler to assemble contigs was developed to assist in performing multiple assemblies and keep track of the results. The workflow re-formats the input datasets to the format required for Velvet, performs the assembly of the input datasets, and produces a basic statistics file summarising the assembly. *De Bruijn* assemblers have a very high memory footprint, and hardware with the required RAM is required to successfully complete the assembly. A dataset containing 35 million short reads of (35-50 bp in length) typically requires up to 120 GB of RAM, depending on the size of kmer used during assembly. A recent thread on the SeqAnswers forums<sup>4</sup> stated that the following formula can be used to calculate the amount of RAM needed for a genome assembly:  $RAM = -109635 + 18977 * ReadLength + 86326 * GenomeSize + 233353 * NumberOfReads - 51092 * kmer$ . No such formula exists to calculate the amount of RAM needed for a transcriptome assembly, mainly due to the uncertainties of transcriptome size, and number of alternative isoforms that can be present in a sample. For a typical Illumina dataset consisting of reads 76 bp long, a kmer value between 51 and 55 were found to produce the best assembly using a scoring function that takes into account the number of bases as well as the number and length of contigs present in an assembly. The choice of kmer, expected coverage and coverage cutoff depends greatly on the size and characteristics of the biological sample, as well as the amount and quality of sequence data used for

---

<sup>4</sup> <http://seqanswers.com/forums/showthread.php?t=2101>

the assembly, and therefore no conclusion can be reached in terms of the best parameters to use. One important aspect when evaluating the contiguity of the assembled transcripts is the comparison against known, full-length cDNA sequences in order to identify missassembled contigs and critically evaluate an assembly.

The availability of transcriptome specific assembly software, such as **trans-ABYSS** (Robertson *et al.*, 2010), **OASES** (Zerbino *et al.*, unpublished) and the recently released **Trinity** (Grabherr *et al.*, 2011) software packages will in future make *de novo* assemblies of full-length transcripts a standard bioinformatic operation. The **Velvet**-based assembler approach described here does not deal with the assembly of alternative splice forms, and may assemble some partial transcripts, but the analysis described did result in the assembly of near full-length, contiguous biological molecules, as described in Chapter 3.

Functional annotation of a set of assembled transcripts occurs mainly through homology-based searches to identify sequences similar to a newly sequenced organism. Both the **InterProScan** and **BLAST2GO** pipelines makes use of homology-based searches and functional protein domain signatures to assign functional annotation to a contig. These annotation pipelines have been used with great success to functionally annotate a vast range of EST and cDNA datasets (Vizoso *et al.*, 2009; Coetzer *et al.*, 2010; Arnaiz *et al.*, 2010; Blanca *et al.*, 2011; Mondego *et al.*, 2011), The **InterProScan** pipeline assigns **PROSITE**, **PRINTS**, **Pfam**, **ProDOM** and **SMART** annotations to each contig in the cDNA file, with the **BLAST2GO** pipeline makes use of these protein features to assign Gene Ontology, **KEGG** and **InterPro** categories to the contigs. The results from the pipeline is presented in a format that can be viewed by the **BLAST2GO** application, or parsed to a delimited text file that can be imported to a database system.

Gene expression calculated with mRNA-Seq data is reported to be more robust than microarray data (Li *et al.*, 2008a; Marioni *et al.*, 2008; Hiller *et al.*, 2009). Estimating gene expression values from known and novel genome models and transcripts aids in identifying pathways and functional gene classes that are over-expressed between different tissues or conditions. Functional expression analysis of different tissues and/or different stages of development can be viewed as the first steps to a complete functional characterisation of a species of interest. The first step in estimating gene expression is to re-align or

map the Illumina short-read data to the target genome and a set of splice junctions. Results show that for the *Eucalyptus grandis* genome, paired end reads longer than 50 bp do not increase the mapability of the fragments, when reads were aligned with the TopHat program (Trapnell *et al.*, 2009). This value will differ between different organisms, but can be used as a guideline to determine gene expression for organisms of similar genome complexity as eucalypts. Several statistical approaches have been developed to model the distribution of RNA-Seq data across a transcriptome (Langmead *et al.*, 2010; Srivastava and Chen, 2010; Trapnell *et al.*, 2010; Wang *et al.*, 2010a) and correct for transcript length (Oshlack and Wakefield, 2009), positional (Bohnert and Ratsch, 2010) and content bias of the technology (Hansen *et al.*, 2010). Improvements to the CUFFLINKS package to incorporate various normalisation methods for the detection of differential expression makes it a valuable benchmark to use for expression analysis (Trapnell *et al.*, 2010; Roberts *et al.*, 2011). The DEGseq package makes use of three different published methods (Marioni *et al.*, 2008; Bloom *et al.*, 2009; Tang *et al.*, 2009) and two novel methods to identify differential expression using mRNA-Seq data, and also serves as a good alternative starting point for differential expression analysis. Both CUFFLINKS and DEGseq are available as tools in the BCBU Galaxy server. Investigations of transcriptome wide gene expression data assist in the selection of target genes of interest for genetic modification and the elucidation of complex traits when combined with population genetic data.

The workflows described here serve as a starting point to a whole range of uHTS DNA sequence analyses. The Galaxy environment facilitates easy incorporation of new tools, results storage and tracking, and a common interface to store and share analysis pipelines and results. Key parameters that can influence the output of the individual analysis tools that make up the workflows have been discussed and guidelines provided regarding the effect of these parameters on a dataset. The guidelines provided should, however, be used with caution, as they are only applicable to the datasets and organism evaluated. The workflows described here have been used to perform the *de novo* assembly of a gene catalog from mRNA-Seq, the subsequent annotation of the assembled gene catalog as well as the expression profiling of the assembled transcripts as described in Chapter 3.

## Chapter 3

# The assembly and annotation of a draft transcriptome sequence of a *Eucalyptus* hybrid tree

## Chapter Preface

The following publication resulted from the worked described in this chapter:

- Mizrachi, E., Hefer, C.A., Ranik, M., Joubert, F. and Myburg, A.A., 2010. *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. **BMC Genomics**, Volume 11, 681.

Several of the figures used in this chapter were also used in the above mentioned publication. The manuscript is attached as Appendix D.

### *Author contributions:*

C.A. Hefer performed the *de novo* assembly and automated annotation, participated in data analysis, and drafted the chapter. E. Mizrachi helped sample the biological material, prepared the libraries, participated in the *de novo* assembly and data analysis, M. Ranik prepared the libraries, helped sample the biological material and participated in data analysis. F. Joubert participated in data analysis. A.A. Myburg conceived of the study, and participated in its design and coordination and participated in data analysis.

### 3.1. Introduction

In South Africa, 36% (450 000ha) of the total land area used for commercial forestry comprises of eucalypt species (DWAF report, <http://www2.dwaf.gov.za/webapp/Documents/FSA=Abstracts2009.pdf>). The *Eucalyptus* genome released early in 2011 (<http://www.phytozome.net>) is only the second forest hardwood tree for which a genome sequence is available. Together with the genome sequence of *Populus trichocarpa* (Tuskan *et al.*, 2006), the *Eucalyptus* genome sequence provides researchers with interests in woody biomass production unique opportunities to elucidate the underlying biochemical and genetic components of wood properties and cellulose production. Eucalypt and poplar trees have been earmarked as potential bioenergy crops (Hinchee *et al.*, 2009), which adds to the existing value of these plantation crops in the pulp, paper and timber industries (Moore *et al.*, 2010).

Accurately identifying gene models in a newly sequenced genome relies heavily on the presence of evidence of expression of potential gene models in order to reduce the number of false positives identified using computational gene finders. Despite the availability of uHTS technology, by the the end of 2009 precious few eucalypt unigene and EST datasets had been made available to the scientific community, mostly due to the commercial interests in the species (Hibino, 2009). The EST datasets that were available consisted mostly of Sanger sequenced datasets (Rasmussen-Poblete *et al.*, 2008; Rengel *et al.*, 2009) and 454 (Roche Life Sciences) generated EST datasets (Novaes *et al.*, 2008). A collection of EST resources in the public domain is now accessible from the *EucalyptusDB* resource (<http://eucalyptusdb.bi.up.ac.za>), and consists of ESTs and unigenes derived from seedlings and different leaf and xylem tissues from various eucalypt species.

Sequencing gene specific tags of the mRNA content of a cell was first demonstrated during the human genome project (Adams *et al.*, 1991), and has in the past two decades been used to profile the transcriptomes of many organisms (Boguski *et al.*, 1993, 1994; Sterky *et al.*, 1998; Seki *et al.*, 2002; Dias Neto *et al.*, 2000; Rasmussen-Poblete *et al.*, 2008). The advent of ultra-high-throughput sequencing technologies, especially the use of mRNA-Seq has enabled the genome wide identification of novel expressed transcripts in various tissues and organisms (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mor-

tazavi *et al.*, 2008), the identification of alternative splicing events (Pan *et al.*, 2008; Sultan *et al.*, 2008; Filichkin *et al.*, 2010) and quantification of transcript abundance (Mortazavi *et al.*, 2008; Trapnell *et al.*, 2010). Transcriptome profiling has mostly been performed for model organisms, although early access to genome sequences has been used to profile gene expression in non-model organisms, with reference-based transcriptome assemblies performed for the *Pachycladon* (Collins *et al.*, 2008), *Melitaea* (Vera *et al.*, 2008) and *Cucumis* (Wu *et al.*, 2010) genomes.

The following sections describe the *de novo* assembly, annotation and transcriptome profiling of a *Eucalyptus* hybrid tree. By performing deep mRNA sequencing of six different tissues with Illumina technology, reads ranging from 35-55 bp long were assembled into 18 894 contigs longer than 200 bp. The assembled contigs were evaluated for contig contiguity and assembly quality, and transcript composition compared to the homologous transcripts available for the *Populus trichocarpa*, *Vitis vinifera* and *Arabidopsis thaliana* angiosperms. Annotation of the assembled contigs was performed based on homology search results against the above mentioned angiosperm transcriptome datasets, as well as additional annotation including protein family and protein feature annotations, gene ontology classification and functional pathway classifications. The transcript abundance of the assembled contigs was calculated in each of the sampled tissues, and a set of transcripts over-expressed in woody when compared to non-woody tissues were identified. The deep sequencing of the tissues also allowed for the identification of possible polymorphism sites in the assembled gene catalog, providing insight in the heterozygosity present in the hybrid transcriptome.

## 3.2. Materials and methods

### 3.2.1. Plant tissue collection, mRNA-Seq library preparation and sequence generation

Six different tissues from a six year old ramet of a commercially grown *E. grandis* x *E. urophylla* hybrid clone (GUASPI, Sappi forest Research) sampled consisted of xylem, immature xylem, phloem, shoot tip, and young and mature leaf tissue. After total RNA extraction and polyA enrichment, paired-end libraries



with an approximate average insert length of 200 bp were synthesized. The libraries were sequenced on an Illumina Genome Analyser (version I, II and IIx) equipped with a paired-end module. Further details regarding the sampling and laboratory methods are described in the Materials and Methods section of Mizrahi *et al.*, 2010 (Appendix D).

### 3.2.2. *De novo* transcriptome assembly

A single paired-end file was created containing the reads from the various tissues and sequence lanes. After removing reads containing regions of consecutive low quality bases (4 consecutive "N"s), a total dataset of 3.9 Gb of sequence was used for the assembly. The *de novo* transcriptome assemblies were performed with the *de Bruijn* graph-based assembler **Velvet** (Zerbino and Birney, 2008). Various assemblies were performed to firstly identify the optimal kmer length, and then the expected coverage cutoff that resulted in the assembly of the final set of transcripts. A stringent average coverage cutoff of 8X was used to remove entire contigs with low coverage.

#### Extending the assembly

The short read assembler, **Velvet** (Zerbino and Birney, 2008) showed superior performance over other short read assemblers, and although the assembler was developed for genome assembly, it managed to assemble sufficiently long contigs of representative mRNA-transcripts. The assembler requires an estimation of the coverage across a transcript in order to correctly join nodes in the *de Bruijn* graph representing each contig. If large discrepancies in coverage happen to occur across a contig, the genome assembler tends to break the contig into two or more shorter sequences. Due to the variable nature of transcript expression, a coverage assisted re-assembly of the assembled contigs was performed. The re-assembly process involved mapping the dataset of short reads to the assembled contigs, and calculating the average sequence depth of each transcript. The matching read and associated mate pair reads that mapped to any given transcript were then extracted from the total dataset and together with the calculated average coverage and the original contig used in a reference based approach to re-assemble

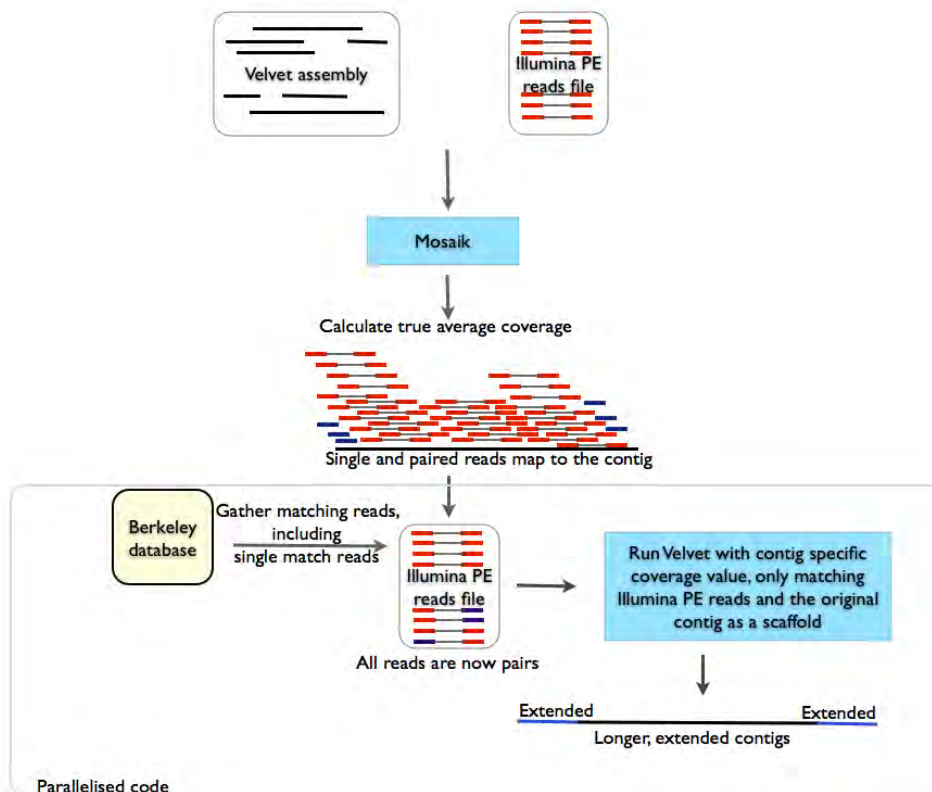


Figure 3.1: A schematic flow diagram of the coverage-assisted re-assembly process. First, a mapping process (using Mosaik, Stromberg and Marth, 2008) is followed where all the Illumina reads are mapped to a contig from the initial *de novo* Velvet assembly, then the average coverage of the contig is calculated. The short-reads will map as pairs (red) or single reads (blue) to the assembled contig. All mate-pairs (of the red and blue reads) that mapped to the contig are then extracted from the Berkeley database and stored in a separate file. These pairs are then, together with the appropriate coverage setting and the contig as a backbone, submitted to Velvet for re-assembly.

the contig (see Figure 3.1 for a graphical representation of the process, and Appendix B for the Python code).

A Berkeley database (BDB, Oracle, 2009) was constructed to facilitate the storage of the mate-pair information for the 35 million paired-ends reads in an efficient manner. The high performance and scalability of the BDB storage system made BDB more suited for the task at hand than relational database systems or flat-file storage (Oracle, 2009). The BDB system is designed to be embeddable in a programmatic fashion, and have the ability to handle multiple concurrent queries. The mate pair information was stored as tuples in the database, with the name of the entry as the lookup or key value. The

key-value storage allowed for the fast querying of the data, but with the initial computational overhead of initializing the database. The `Python` API provided access to the database via a dictionary-like object, and streamlined the extraction of mate-pair information from the database.

The average coverage per contig for the assembled contigs were calculated using the `Mosaik` assembler (Stromberg and Marth, 2008). The mapping parameters used a hash size of 12, and limited the number of hash positions on a contig to 100, as per user documentation. Variations of the input parameters did not yield significantly different results. In addition to the coverage value for each contig, the assembler returns a list of short reads which aligned to each contig. These reads and their respective pairs were then extracted from the BDB using custom `Python` scripts.

The contig coverage and short reads that mapped to a specific contig as determined by `Mosaik` were used in a reference based re-assembly. The expected coverage parameter was customized to represent the calculated coverage, and the short reads were submitted as paired reads to the assembler, with the original contig as the reference template. The reference based assembly had a relatively small memory footprint, since only the reads that mapped to the contig were used during the assembly, and a pipeline was developed to run the re-assembly process in parallel on the 24-core server used for assembly.

The release of a *de novo* transcriptome assembler, `OASES` (Zerbino *et al.*, unpublished and in beta release) prompted the re-assembly of the contigs using the same parameters that was used for the original `Velvet`-based assembly. The `OASES` assembler does not accept any parameters regarding the expected coverage values since it attempts to estimate the coverage during assembly. The `OASES`-assembled contigs were then compared with the `Velvet` assembled transcripts and a set of full-length cDNA sequences from GenBank.

### 3.2.3. Prediction of coding sequences

In order to provide supporting information for the contiguity of the assembled contigs, multiple *ab initio* coding sequence (CDS) predictions were performed on the assembled and extended contigs (Table 3.2). `GENSCAN` (Burge and Karlin, 1997), `GeneMark` (Borodovsky and McIninch, 1993), `AUGUSTUS`

(Stanke and Waack, 2003), GLIMMER (Salzberg *et al.*, 1999) and GeneID (Guigo *et al.*, 1992) are all Markov model-based prediction tools for the prediction of coding sequences from genomic DNA. Markov-based prediction tools are trained on a predefined known dataset of known features associated with a coding sequence such as a transcriptional start site (TSS), 5' and 3' untranslated regions (UTRs), start codons, splice donors and acceptors, *etc.* These training sets are then used to perform *ab initio* coding sequence predictions. For all of the previously-mentioned predictors, the *Arabidopsis* training data set was used to predict the coding regions of the contigs. The prediction of coding sequences played an important role in validating the assembled contigs in terms of possible misassemblies occurring in the dataset.

#### 3.2.4. Inspecting contig contiguity

The nature of the assembled contigs in terms of assembling a complete full length contig, and in terms of identifying possible misassemblies, were inspected by selecting 33 full-length *Eucalyptus* cDNA sequences representing various different gene families, and using these as reference templates for the assembled contigs (Section C.1.1 in Appendix B). The homology search tool, BLAST (Altschul *et al.*, 1990), was used with a stringent e-value cutoff ( $1e^{-100}$ ) to find the corresponding assembled contig that matched each of the Genbank cDNA sequences. A global alignment (Needle, Rice *et al.*, 2000) was then performed between the cDNA sequence and the assembled contig, and the cDNA sequence and the predicted coding sequence from the GENSCAN analyses were considered in order to evaluate the contig contiguity.

The coverage per base pair (CBP) was calculated separately for all of the cDNA sequences, the assembled contigs and the predicted CDS with BWA aligner (Li and Durbin, 2009, see section 3.2.7 for a description of the parameters used). The coverage values and the alignment information were then used to construct a graph which represents the coverage across the alignment between the three sequences. This representation of the sequences allowed for the fast identification of misassembled contigs from Illumina data in comparison to the full cDNA sequences obtained from Sanger sequencing.

### 3.2.5. Homology searches

Homology-based analyses were used to evaluate the size categories and completeness of the assembled contig dataset. The complete peptide datasets of *Arabidopsis thaliana* (TAIR9, Huala *et al.*, 2001), *Populus trichocarpa* (Version 2, Tuskan *et al.*, 2006) and *Vitis vinifera* (Jaillon *et al.*, 2007) were compared to two *Eucalyptus* datasets, the assembled transcriptome, and a dataset of all publicly available *Eucalyptus* sequences at that time (August 2009). The publicly available dataset (henceforth known as the EucAll dataset) consisted of 45 442 entries from GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/index.html> downloaded on 27 July 2009), 13 930 entries from a *Eucalyptus* Wood (*EucaWood*) unigene and ESTs resource (Rengel *et al.*, 2009), leaf tissue ESTs (120 661 entries from JGI-produced sequences), and 190 106 unigenes and singlets from 454 data (Novaes *et al.*, 2008). The aim was to identify the sequence homologs of the *Arabidopsis*, *Vitis* and *Populus* protein datasets present in the *Eucalyptus* datasets with homology-based searches. BLAST searches were performed against the *Eucalyptus* datasets with e-value thresholds of  $1e^{-5}$ ,  $1e^{-10}$  and  $1e^{-20}$ , and a High Scoring Pair (HSP) minimum alignment length of 100 bp (33 amino acids). The set of results were further separated based on the size of the hit (*Eucalyptus*) sequence. The proportion of genes shared among four angiosperm species (*Eucalyptus*, *Arabidopsis*, *Poplar* and *Vitis*) were also determined with BLAST ( $1e^{-10}$ , min HSP alignment length of 100 bp) analysis, identifying genes common to all four species, and genes shared between the assembled *Eucalyptus* contigs and each of the other three angiosperm species.

### 3.2.6. InterProScan

The InterProScan tool was used to detect protein predictive models or signatures in the assembled dataset. InterProScan relies on integrative data stored in the InterPro database (Hunter *et al.*, 2009) which aggregates diverse information from multiple databases, including Gene3D, PANTHER, Pfam, PIR, PRINTS, ProDom, ProSITE, SMART, SUPERFAMILY and TIGRFAM data. In the 2009 release of InterPro close to 58 000 different signatures were present in the database, and together with the over 16 000 UniProtKB entries formed a valuable tool for protein functional annotation.

### 3.2.7. Calculating transcript coverage and expression

Average coverage per contig was calculated by mapping the short reads to the assembled contigs with the BWA aligner (Li and Durbin, 2009), and averaging the coverage per base pair (CBP) for every base in the assembled transcript. The alignment allowed for a 0.04 fraction of missing alignments given the predicted 2% uniform error rate of Illumina reads, also allowing for one gap in the sequence alignment. During alignment, deletions were disallowed within 16 bp of the 3' end of the sequences, and within 5 bp of the 5' end. A gap opening penalty of eleven and gap extension penalty of four were used for the scoring matrices, and the mean insert size for a paired read to be considered as being mapped properly was set to 200 bp.

The Fragments per Kilobase of exon per Million mapped (FPKM, initially developed by Mortazavi *et al.* (2008) as Reads per Kilobase of exon per Million mapped, RPKM, but redefined as FPKM by Trapnell *et al.*, 2010) were derived from mapping the short reads to the assembled contigs with the BOWTIE short read aligner (Langmead *et al.*, 2009). The resulting alignment files (SAM format, Li *et al.*, 2009a) were then used as input for the CUFFLINKS software program (Trapnell *et al.*, 2010) in order to calculate the FPKM values. The parameters for the BOWTIE alignment allowed for three mismatches in the seed (first 28 bp from the 5' end of the sequence), but no gaps in the alignment. A mean insert size of 200 bp was used for the correct alignment of paired-end reads.

Differential transcripts in the xylogenetic (woody tissues which include the xylem and immature xylem datasets) *vs.* non-xylogenetic (non-woody tissues which include the shoot tip, mature and immature leaf samples) were detected by filtering the transcripts to only contain transcripts with an expression value  $>2X$  in either of the two groups of tissues. KEGG and gene ontology analysis of the set of differentially expressed transcripts were performed with the BiNGO Cytoscape plugin (Maere *et al.*, 2005) and the Paintomics (García-Alcalde *et al.*, 2010) web server.

### 3.2.8. Single nucleotide polymorphism detection

Results from short read mapping performed with the *Bowtie* short read alignment tool (Langmead *et al.*, 2009) were used to detect single nucleotide polymorphisms in the dataset. Possible polymorphisms were detected using *SAMTOOLS* (Li *et al.*, 2009a). *SAMTOOLS* applies a default filtering for SNPs using the following rules; (a) discard SNPs within the 3 bp flanking region around a potential indel; (2) discard SNPs covered by three or fewer reads; (3) discard SNPs covered by no read with a mapping quality higher than 60; (4) in any 10 bp window, if there are three or more SNPs, discard them all; and (5) discard SNPs with a consensus quality lower than 10 (Li *et al.*, 2008b). Potential SNPs were then filtered to contain only SNPs with coverage of at least 8X, where the minor allele occurs at least 4X. Only SNPs with a higher PHRED based quality score than 20 were included in the final results.

## 3.3. Results

### 3.3.1. Assembly

Multiple assemblies were performed with a defined set of input parameters using different values to evaluate which parameters resulted in the longest transcript sequences in the most contigs (Figure 3.3). The final assembly was performed with the following input parameters: kmer=31, expected coverage value=1000 and coverage cutoff value=8. The optimal kmer and expected coverage values were selected by performing a range of assemblies varying the kmer values from (kmer=19, 21, 23, 25, 27, 29, 31, 33) and expected coverage (EC=10, 25, 50, 75, 100, 250, 500, 750 and 1000) input parameters (Figure 3.2 and Figure 3.3). Each assembly was scored with the following scoring algorithm:  $Score = \frac{(N_{50_{all}} * N_{long})}{Sum_{all} + log(Sum_{long})}$ , where contigs longer than 1 000 bp were considered as long contigs (Section 2.3.3). The conservative coverage cutoff value (8X average coverage of a contig) was chosen to prevent low covered contigs from entering the assembly. A summary of the final assembly is presented in Table 3.1.

After assembly, a coverage-assisted re-assembly was performed on the assembled contigs. The resulting assembly contained 23.27 Mbp of sequence in 38 597 contigs *vs.* the 22.88 Mbp sequence in 38

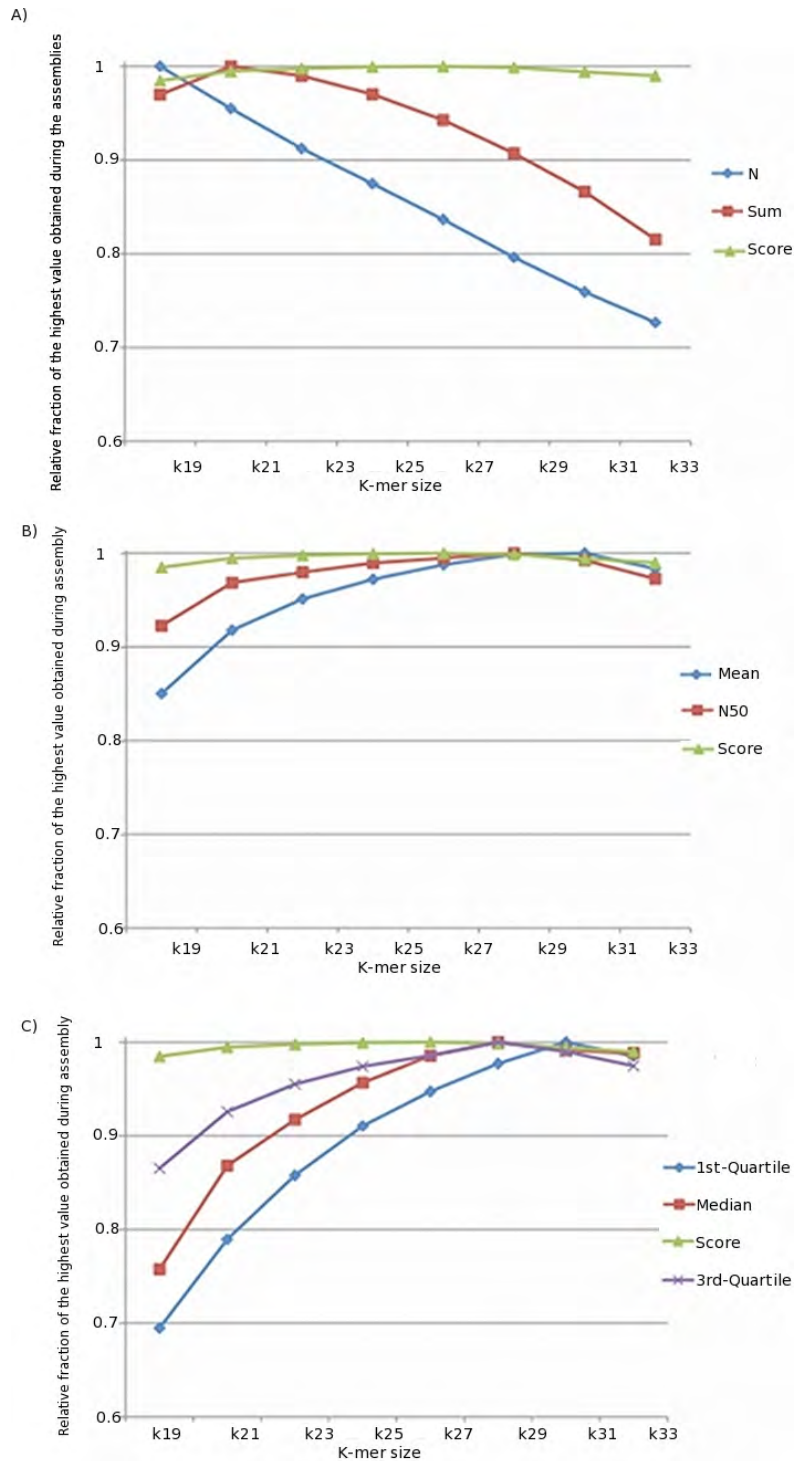


Figure 3.2: Identifying the optimal kmer used for the *de novo* assembly of the *Eucalyptus* transcriptome. The y-axis represent the relative fraction of the highest value obtained for each parameter during assemblies. The scoring function for each assembly is plotted together with assembly parameters such as number of contigs (N), the total sum of bases in the assembly (Sum) in (A), the mean contig size (Mean) and N50 value in (B) and the spread of contig sizes (1st-quartile length, median and 3rd-quartile length) in (C) for each assembly where the kmer value varied from 19 to 33. The final assembly using a kmer of 31 was further used to detect the optimal expected coverage value (Figure 3.3)



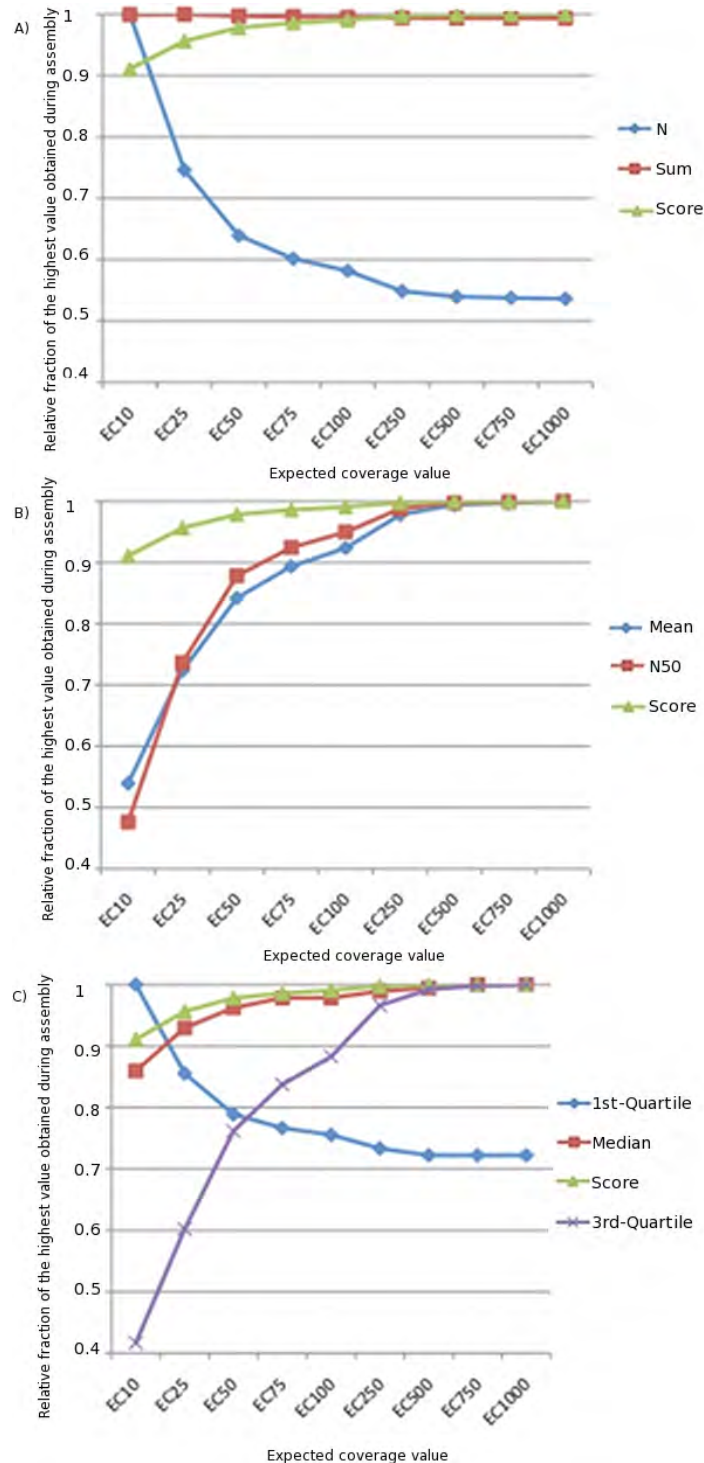


Figure 3.3: Identifying the optimal expected coverage value to use for the *de novo* assembly of the *Eucalyptus* transcriptome. The y-axis represent the relative fraction of the highest value obtained for each parameter during the assemblies. The scoring function for each assembly is plotted together with assembly parameters such as number of contigs (N), the total sum of bases in the assembly (Sum) in (A), the mean contig size (Mean) and N50 value in (B) and the spread of contig sizes (1st-quartile length, median and 3rd-quartile length) in (C) for each assembly where the expected coverage parameter varied from 10 to 1 000. The final assembly was performed with an expected coverage value of 1 000.

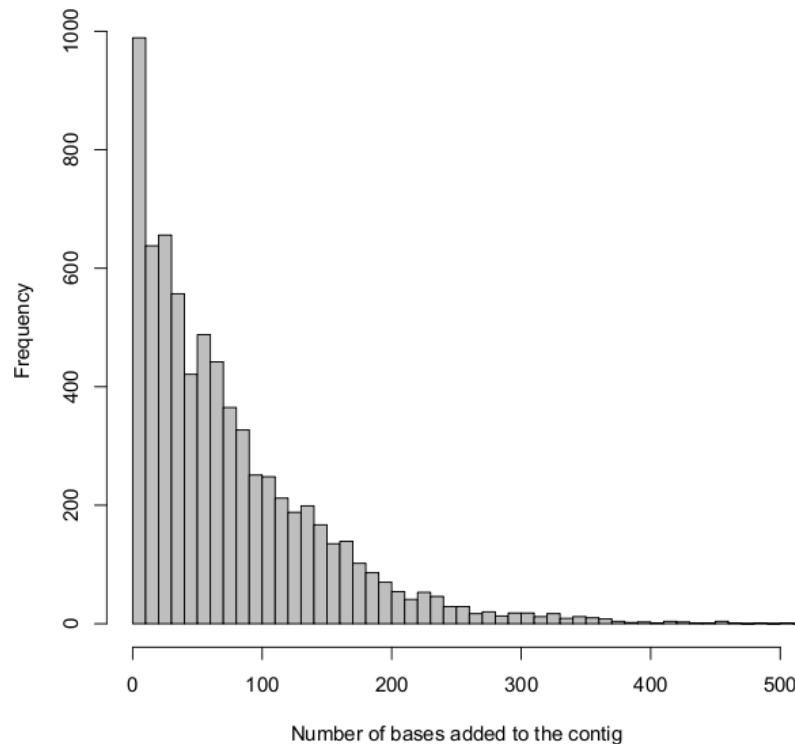


Figure 3.4: The number of bases per contig added during the extension of the assembly. The theoretical upper limit of the number of bases added by the extension step is 400 bp (200 bp for each end of the contig, which corresponds to the sequenced library size) + the standard deviation of the sequenced DNA fragment sizes. 99% of the added bases per contig were shorter than 336 bp.

597 contigs before extension. Although the maximum contig length did not improve, the average length of the shorter contigs did improve overall in the re-assembled dataset (Table 3.1). The mean contig length improved from 592.88 bp to 728.49 bp (22%), and number of unknown bases (N) in the assembly increased from 396 029 to 405 429 (2.3%). Figure 3.4 indicates that 99% of the additional bases added to the assembly per contig were shorter than 336 bp. The theoretical limit with which a single contig can be expected to be extended, was estimated as the insert size of the sequenced DNA library (200 bp) x 2 (one for each end of a contig) and adding a standard deviation for the library insert size (Figure 3.5). This would cater for the cases where one sequence from the mate pair library occurred at the beginning or end of the assembled contig, and the other fragment of the paired sequence were added during the extension step.

A closer look at the top 10 contigs where more than 500 bp was added to the assembly during

```

*****
NODE_10522_before -----TGGCAAAA 8
NODE_10522_after CCCAATTTATCGCAGTTTCAAGCTCAACTTATAAACAGGCCCGTGAATATATTTGAAATTTCCAAAAAGCTCGCAAAA 80

*****
NODE_10522_before ----- 88
NODE_10522_after CTCCTATCAACCAGCAACCCAGATCTTAGAAATCCAAACATCTATAGCCAAAGATCTACCAAAACAAGTTAAGAAA 160

*****
NODE_10522_before ----- 168
NODE_10522_after CCTTGAAGCCGACAGGATCCCGAAGGATGCTATCACCTCAGCAGTTTTTTGTTGAATTCAAAAGTTCCCTTCAGGTTTTCC 240

*****
NODE_10522_before ----- 248
NODE_10522_after TTCCCTGTGATTCCTCCTCAGCCACCGGTTTCTCTTGTGTCTTCTGCCCACTTCCCCAGCTGCCCTGGTGACCTGCTA 320

*****
NODE_10522_before ----- 328
NODE_10522_after TATGCACCCAGCAGCCATGCGGCCCGGTCAACACATAACGGTTACTCATGATGCGCGGACCCCTGAGTGCTTGTCTGT 400

*****
NODE_10522_before ----- 408
NODE_10522_after CTCGGCAGCTGCTATTGCAGATTTTGTCTTCCGAAACCTGGAACCTCTGGTCCACTTCTCTCATTTTCTCATTCACCTA 480

*****
NODE_10522_before ----- 488
NODE_10522_after CACTTGTCCAGCACTGAATTTTTCACITAGACCGATCTTCTGGTCTATAGAAGCAACTGTGCTGTAGCAAGTTGCAGTG 560

*****
NODE_10522_before ----- 568
NODE_10522_after ACAAGGTGTTTCTATCCAATGTTTTGCTGTGTAAAGGCATCTTGCCTAAGACAAAATCCCTTAGCTAGCATGTGCT 640

*****
NODE_10522_before ----- 648
NODE_10522_after TACCACTTCCCTGCTTCCGGACAGGAGATGCAGCAGAGCTCCCTGTTCTCAGTTTCAGTTGAAAATGAGAAAGCAT 720

*****
NODE_10522_before ----- 728
NODE_10522_after CATCTGGTCTTTTGAATCTGGTGCTAGAGCTATAGTGACTGCCTGATGACAAATGGTGCACCCGAAAGCATAGCTGCA 800

*****
NODE_10522_before ----- 808
NODE_10522_after GTCCTGCTCCCTCAGGATCTTGAAGTAAACATAAGCAATCTGGCACCGTTCAATCACTATGCAATTCACACGATC 880

*****
NODE_10522_before ----- 888
NODE_10522_after GATTTACCGGAAAAGAGAAAGAACTCCCTTATGTTTTGCTCAGATGCTGTCAAGGAAAGGTTATTGACTTTCACCGTCC 960

****
NODE_10522_before TTAT----- 892
NODE_10522_after TTATGCGCATGACGGGCTTTGTGTTGGGTGTTCAATTECGAGGAGAG 1009

```

Figure 3.5: The effect of performing a coverage-assisted re-assembly on a single contig (contig\_10522). An additional 124 bp were added to the contig during assembly. 72 bp were added to the beginning and 45 bp to the end of the contig due to a better estimate of the expected coverage of the contig.

extension revealed that the extensions still yielded biologically relevant molecules, as shown with the alignment of the sequences against known protein coding sequences and against the sequences present in the pre-extended dataset. For example, after the initial assembly contig\_68291 (Figure 3.6) had a region of low quality or coverage bases (the result from the stringent 8X coverage cutoff parameter) at positions 65 and position 1832, spanning 40 and 54 bases respectively. During the re-assembly step, when the contig sequence acted as a template sequence for the extended assembly and the 8X coverage cutoff value was not enforced anymore, these regions of unknown bases were extended and repeated, resulting in a total extension of 1 485 bp of low quality bases. By replacing these regions of low quality bases with a stretch of four consecutive Ns (NNNN), and aligning the contig before and after extension, the alignment indicates that the contig after extension actually had bases removed from the beginning of the sequence due to the presence of the polyA region which could not be overcome by the assembler (position 1-171 of the before-extension contig, see Figure 3.6). The alignment also indicates that a region initially consisting of low quality bases at position 1838 of the pre-extension contig was resolved during the extension step.

After applying a further restriction to the assembly to only include contigs equal to or longer than 200 bp, the final assembly contained 18 894 contigs representing 22 108 288 bp of sequence data (Table 3.1). The mean contig length was 908 bp, with the longest contig consisting of 12 053 bp. The N50 value of the final assembly was 1640 bp. These sequences were then used for further contig validation, coding sequence prediction and annotation.

### **3.3.2. Prediction of coding sequences**

In order to determine whether or not the assembled transcripts were full-length, contiguous biological molecules, coding sequence (CDS) predictions were performed on the assembled contigs to identify CDS, open reading frames (ORFs) and transcriptional start and stop sites. The success rate of various CDS detection software tools ranged from identifying 10 400 (7 776 single-exon and 2 624 multi-exon) contigs containing valid coding sequences to 18 894 (16 568 single and 2 326 multi-exon) CDS containing contigs



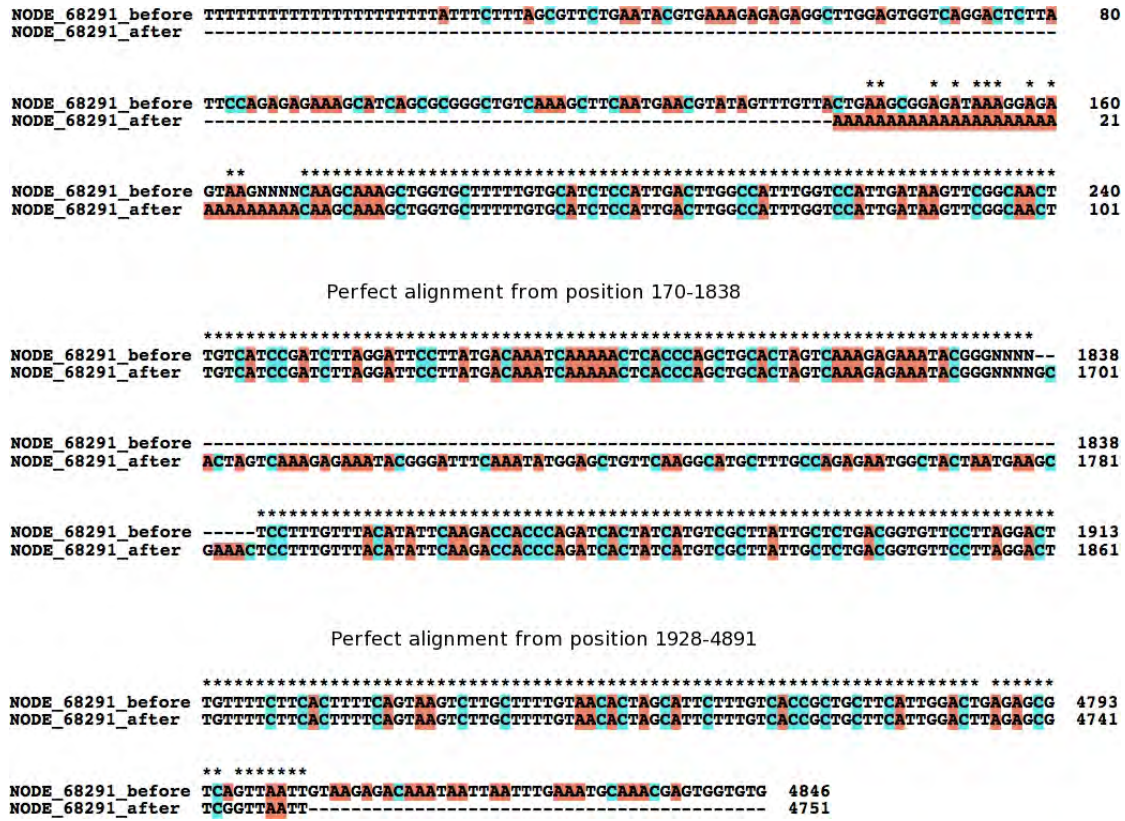


Figure 3.6: The alignment of contig\_68291 before and after extension. The alignment shows that although 1 485 bases was reportedly added to the contig during extension, these bases mostly consisted of the extension of a low quality region containing Ns. The extension did however resolve a 88 bp region of these low quality bases. The contig after extension also showed removed regions at the start and end of the original contig, due to the presence of a polyA region at the beginning of the sequence. The full alignment of the two sequences is available in Appendix C.1.3.

Table 3.1: Comparing the assembled Velvet dataset before and after the coverage assisted extension. The most notable effect is observed in the increased length of the shorter reads (the Q1, median and Q3 values).

	Velvet assembly	After assisted re-assembly	Final assembly (>=200 bp)
Number of contigs	38 597	38 597	18 894
Amount of bases in assembly	22 883 310	23 272 382	22 108 288
Shortest contig length (bp)	61	61	200
First quartile length (Q1) (bp)	64	89	470
Median contig length (bp)	137	358	908
Third quartile length (Q3) (bp)	856	1 078	1573
Maximum contig length (bp)	12 053	12 053	12 053
Mean contig length (bp)	592.88	728.49	1170.12
N50 length (bp)	1 550	1 570	1640
Number of Ns in assembly (bp)	396 029 (1.73 %)	405 439 (1.74 %)	405 238 (1.83 %)

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

Number of predicted exons	GENSCAN	GeneMark	AUGUSTUS	GLIMMER	GeneID
Single exon	10 887	8 320	11 134	7 776	16 568
Multiple exons	4 827	10 365	4 770	2 624	2 326
Total CDS predicted sequences	15 714	18 685	15 904	10 400	18 894

by the GLIMMER and GeneID software packages respectively (Table 3.2). GeneID assigned single-exon status to each of the input contigs, a clear over-estimation of the number of contigs present in the assembly, and the results were disregarded in further analysis. The prediction of single exon coding sequences ranged from 38.70% of the complete dataset with the GeneMark prediction tool, to around 70% (69.28%, 70.00% and 74.76% with the GENSCAN, AUGUSTUS and GLIMMER tools respectively), with a maximum of 87.69% by GeneID. When comparing the GENSCAN, GLIMMER and AUGUSTUS results, a total of 15 275 (94.85%) out of the maximum of 15 904 CDS-containing sequences were predicted by at least two of the CDS prediction tools. GENSCAN predicted more than 98% of the total coding sequences predicted by this subset of predictors, and the results from GENSCAN were subsequently used in downstream analysis.

Further analysis showed that 6 294 (39.57%) of the 15 904 predicted CDS had both start (ATG) and stop (TAA, TGA or TAG) codons present as the first and last codons of the sequence, while 13 660 (81.91%) had one of the features present. Predicted partial CDS sequences with neither the start nor stop codons present as the first and last positions of the contig comprised 14.19% (2 258 contigs) of the total dataset.

### 3.3.3. Inspecting contig contiguity

In order to gain confidence in the quality of the assembled contigs, several sequence alignment approaches were followed to ensure that the assembled contigs were representative of biologically relevant contiguous sequences and not assembly artifacts. Full length *Eucalyptus* cDNA sequences were retrieved from GenBank, and aligned with the corresponding assembled contig and predicted CDS results from GenScan (Figure 3.7 and the alignment of the predicted amino acid sequence on contig\_5550 and the GenBank sequence AF197329.1 in Figure 3.8). The alignment showed mismatches in the 5' and 3' UTR

regions between the GenBank and assembled contig sequences, but a high proportion of similarity in the CDS alignments. For each of the 33 cDNA sequences (see Appendix C.1.1) a global alignment between the cDNA, the contig and predicted CDS sequence was performed to evaluate the contig contiguity. The short read library was then mapped to the cDNA, predicted CDS and assembled contig, and the depth of coverage plotted across the length of each of the sequences. The multiple sequence alignment and the coverage plots of the sequences were then used to construct a coverage-alignment plot for each of the cDNA sequences (Figure 3.9 and Appendix C.1.2). Gaps in the alignment between the three sequences as presented as gaps in the coverage across the region, and where regions of dissimilar sequence occur, the coverage across the region will aid in detecting possible misassemblies.

Using the full-length cDNA sequences as template, 23 of the 33 (69%) comparisons revealed the presence of indels in either the cDNA sequence, the assembled contig, or the predicted CDS sequence. For the purpose of this analysis, indels were defined as any insertion or deletion in the alignment between the sequences longer than five base pairs. Of the 23 sequences where indels were detected, 17 (74%) had indels within the predicted coding sequence, with the remaining indels present in the predicted UTR regions. Twenty eight (85%) of the 33 sequences inspected contained both the 5' and the 3' UTRs, while the remaining sequences contained at least one UTR sequence.

Inspection of the zinc transporter cDNA sequence AF197329 and its corresponding assembled contig\_5550 showed some initial indels present in the 5' region of the assembled sequence when compared to the cDNA sequence (Figure 3.9A). Various single base pair mismatches occur within the predicted coding sequence (e.g. position 92 on the assembled contig), with a six base pair indel present at position 686 of the assembled contig. The coverage was calculated across the indel as 40X, indicating that the indel is likely present in the mRNA-Seq sequenced sample, and is not an artifact of a missassembly. The alignment of the three sequences is presented on the x-axis of Figure 3.9A, with the coverage across each sequence plotted on the y-axis. The 6 bp gap in the alignment where the indel is present, is indicated by a gap in the coverage (red line) of the graph. More indels were detected in the 3' region of the assembled transcript.

```

contig_5550      CGGAAAGTTGATGGAGACGAGGGAACCTAGTGTAGCGCC---CGCCGCCGTCCCT-----CGGGGATCGCGTCCGATCTCTCCGTCTCCGGAGC-----AGCAAGATGAGCCGATGATTTCTGAACATGGACAGATCATT 128
contig_5550_cds -----ATGAGCACGACATGATTTCTGAACATGGACAGATCATT----- 36
AF197329.1     -----GGAGACGAGGGAACCTAGTGTGGAGCGCAACCGCCTCCGCCCCGAAAAGCACCAGCGCCGCTCGATCTCTCCATCTCCGGAGCTAGTCCGGAAGCGGCAAGATGAGCACCGATGATTTCTGAACATGGGACATCATT 140

*****
contig_5550      GAAGTGTGTGGAGATGTCGAAGCCATGGAAAACAAGCCAGGTGGTGTAAAGGTATCGCCAGAGGCCCTTTGTGTTTTTCAGACCTTAGAAAATAGTTCCGAAAGATGCAAGAGAGAGGTCAAACTCTATGAAGAACTTTTGTTCGGGTG 278
contig_5550_cds GAAGTGTGTGGAGATGTCGAAGCCATGGAAAACAAGCCAGGTGGTGTAAAGGTATCGCCAGAGGCCCTTTGTGTTTTTCAGACCTTAGAAAATAGTTCCGAAAGATGCAAGAGAGAGGTCAAACTCTATGAAGAACTTTTGTTCGGGTG 186
AF197329.1     GAAGTGTGTCAAAGTGTCCAGCCATGGAAAACAGCCAGGTGGTGTAAAGGTATCGCCAGAGGCCCTTTGTGTTTTTCAGACCTTAGAAAATAGTTCCGAAAGATGCAAGAGAGAGGTCAAACTCTAGAAAACCTTTTAAATTCGGGTG 290

*****
contig_5550      GTGCTTTGTATTATCTTCATGAGCATAGAAAGTAGTTGGTGGTATTGAAAGCCAAAGTCTTCCATTCTCACAGATGCAGGCCATTGTGCTGCGGATGTTGCCCTCATTTGCAATACTCTATTCTATTGGGCATCAGGTTGGGAGGCA 428
contig_5550_cds GTGCTTTGTATTATCTTCATGAGCATAGAAAGTAGTTGGTGGTATTGAAAGCCAAAGTCTTCCATTCTCACAGATGCAGGCCATTGTGCTGCGGATGTTGCCCTCATTTGCAATACTCTATTCTATTGGGCATCAGGTTGGGAGGCA 336
AF197329.1     GTGCTTTGTATTATCTTCATGAGCATAGAAAGTAGTTGGTGGTATTGAAAGCCAAAGTCTTCCATTCTCACAGATGCAGGCCATTGTGCTGCGGATGTTGCCCTCATTTGCAATACTCTATTCTATTGGGCATCAGGTTGGGAGGCA 440

*****
contig_5550      ACTCCACGCCAGTCTTATGGTTCTTCCGAATGAAAATTTGGTGCACCTGTCTCCATCCAGATTATATGGCTACTTGGTGGGATTTGTTGTATGAAAGCAATAGAAAAGACTAATCAATGGTCAACAAGAAAGTTCATGGCTTCTCATG 578
contig_5550_cds ACTCCACGCCAGTCTTATGGTTCTTCCGAATGAAAATTTGGTGCACCTGTCTCCATCCAGATTATATGGCTACTTGGTGGGATTTGTTGTATGAAAGCAATAGAAAAGACTAATCAATGGTCAACAAGAAAGTTCATGGCTTCTCATG 486
AF197329.1     ACTCCACGCCAGTCTTATGGTTCTTCCGAATGAAAATTTGGTGCACCTGTCTCCATCCAGATTATATGGCTACTTGGTGGGATTTGTTGTATGAAAGCAATAGAAAAGACTTATCAATGGTCAACAAGAAAGTTCATGGCTTCTCATG 590

*****
contig_5550      TTCATTACTGCTCCCTTTGGTCTTCTTGTGAATATTGCCATGGCAGTGTACTGGGTACCGATCACAGCCATGGTCTAGTGCATGCAATGGTCAATGGCATAGTGCACATGAACTAGTGCATCAAGTCT 728
contig_5550_cds TTCATTACTGCTCCCTTTGGTCTTCTTGTGAATATTGCCATGGCAGTGTACTGGGTACCGATCACAGCCATGGTCTAGTGCATGCAATGGTCAATGGCATAGTGCACATGAACTAGTGCATCAAGTCT 636
AF197329.1     TTCATTATGCTCCCTTTGGTCTTCTTGTGAATATTGCCATGGCAGTGTACTGGGTACCGATCACAGCCATGGCAATGGTCAATGGCATAGTGCACATGAACTAGTGCATCAAGTCT 734

*****
contig_5550      CATAGCCATGAGGATCACGGTGTATGTCATCTATGGATTAAACCGTCAAGAAACATGACCATCATCATCAATGAGAAAGATTTCAAAGACCATGCTGATCAACATCATGACCATGAAGCGGCTTAACTGAGCCGGTTTTGTCAGACTTGC 878
contig_5550_cds CATAGCCATGAGGATCACGGTGTATGTCATCTATGGATTAAACCGTCAAGAAACATGACCATCATCATCAATGAGAAAGATTTCAAAGACCATGCTGATCAACATCATGACCATGAAGCGGCTTAACTGAGCCGGTTTTGTCAGACTTGC 786
AF197329.1     CATAGCCATGAGGATCACGGTGTATGTCATCTATGGATTAAACCGTCAAGAAACATGACCATCATCATCAATGAGAAAGATTTCAAAGACCATGCTGATCAACATCATGACCATGAAGCGCTTAACTGAGCCCTTTTGTGAGACTTGC 884

*****
contig_5550      TGGGAAGCTGAAAGTAACTCAAACCTGGAAACAAAACAAGACCAACCACTAATAATTAACATACAAGGGTCTTACTTCTATGTTGTTGGGATTCATTCAAAGTGTGGGTGTGATGATTTGGGGTGCAAATATATGGATTAAAGCCCGAG 1028
contig_5550_cds TGGGAAGCTGAAAGTAACTCAAACCTGGAAACAAAACAAGACCAACCACTAATAATTAACATACAAGGGTCTTACTTCTATGTTGTTGGGATTCATTCAAAGTGTGGGTGTGATGATTTGGGGTGCAAATATATGGATTAAAGCCCGAG 936
AF197329.1     TGGGAAGCTGAAAGTAACTCAAACCTGGAGCCAAAACAAGACCAACCACTAATAATTAACATGCAAGGGTCTTACTTCTATGTTGTTGGGATTCATTCAAAGTGTGGGTGTGATGATTTGGGGTGCAAATATATGGATTAAAGCCCGAG 1034

*****
contig_5550      TGGAAAATTGTGCACTTGTATTGACACATGATATTCACAGTAATGTGCTTTGGCAACAATAACCGTCTGCTACCTAACATTTTGGAGGTTTTGTGGAAGTACCCCTAGAGAGATTGATGCCACTAGGCTCGAGAGTGGACTTCGGGAG 1178
contig_5550_cds TGGAAAATTGTGCACTTGTATTGACACATGATATTCACAGTAATGTGCTTTGGCAACAATAACCGTCTGCTACCTAACATTTTGGAGGTTTTGTGGAAGTACCCCTAGAGAGATTGATGCCACTAGGCTCGAGAGTGGACTTCGGGAG 1086
AF197329.1     TGGACGATTGTGCACTTGTATTGACACCTGATATTCACAGTAATGTGCTTTGGCAACAATAACCGTCTGCTACCTAACATTTTGGAGGTTTTGTGGAAGTACCCCTAGAGAGATTGATGCCACTAGGCTCGAGAGTGGACTTCGGGAG 1184

*****
contig_5550      ATGGATGAAGTGAATGCAAGTCCATGAATTGCACATTTGGGCTATAAACGGTTGGAAAAGGTGCTATTAGCCTGCCATGTCAAAAATAAGGGGTGACGCCAATGCGGATAATGCTTGGCAAGGTCGTGGAATACATCAAGAGAGATCAAG 1328
contig_5550_cds ATGGATGAAGTGAATGCAAGTCCATGAATTGCACATTTGGGCTATAAACGGTTGGAAAAGGTGCTATTAGCCTGCCATGTCAAAAATAAGGGGTGACGCCAATGCGGATAATGCTTGGCAAGGTCGTGGAATACATCAAGAGAGATCAAG 1236
AF197329.1     ATGGATGAAGTGAATGCAAGTCCATGAATTGCACATCTGGGCTATAAACGGTTGGAAAAGGTGCTATTAGCCTGCCATGTCAAAAATAAGGGGTGACGCCAATGCGGATAATGCTTGGCAAGGTCGTGGAATACATCAAGAGAGATCAAG 1334

*****
contig_5550      ATAAAGTCAACCTACCAATCAAATAGAAAACAAGTAGATTTCCGGAAGGTGAATGATTTAGTTATGGCATTTGTATAATGGAATGGCAGGCTTGGGGTCAAAATTTGGCTTTAAAGTGTGTTAGATATTTGCAATTTGGAGCTTTTTTCTCGTAGCT 1478
contig_5550_cds ATAAAGTCAACCTACCAATCAAATAGAAAACAAGTAGATTTCCGGAAGGTGAATGATTTAGTTATGGCATTTGTATAATGGAATGGCAGGCTTGGGGTCAAAATTTGGCTTTAAAGTGTGTTAGATATTTGCAATTTGGAGCTTTTTTCTCGTAGCT 1272
AF197329.1     ATAAAGTCAACCTACCAATCAAAGTAGAAAACAAGTAGATTTGGAAGGTGAATGATTTAGTTATGGCATTTGTATAATGGAATGGCAGGCTTGGGGTCAAAATTTAGGTTCTAAATTTGTTGGATGTTGCGCTTTTCTGGTAGCT 1484

*****
contig_5550      GGGCCTTTGAGGCCCTTCAGGAGTATGATGTAATGTTCCGTTCTCCTTTGTTGGAACTTT-----ATGTTTTAAG----- 1547
contig_5550_cds ----- 1272
AF197329.1     GGGCCTTTGAGGCCCTTCAGGAGTATGATGTAATGTTCCGTTCTCCTTTGTTGGAACTTT-----ATGTTTTAAG----- 1628

```

Figure 3.7: Alignment of the full length cDNA sequence AF197329.1, the assembled contig\_5550, and the predicted coding sequence. Note that some gaps appear in the predicted contig upstream (5' UTR) of the ATG site and in the 3' UTR region downstream of the translation stop (TAG) site. There is a six-base-pair insertion present at position 686 of the cDNA sequence and various single nucleotide mismatches are visible in the alignment. The protein sequence alignment between contig\_5550 and AF197329.1 is presented in Figure 3.8.





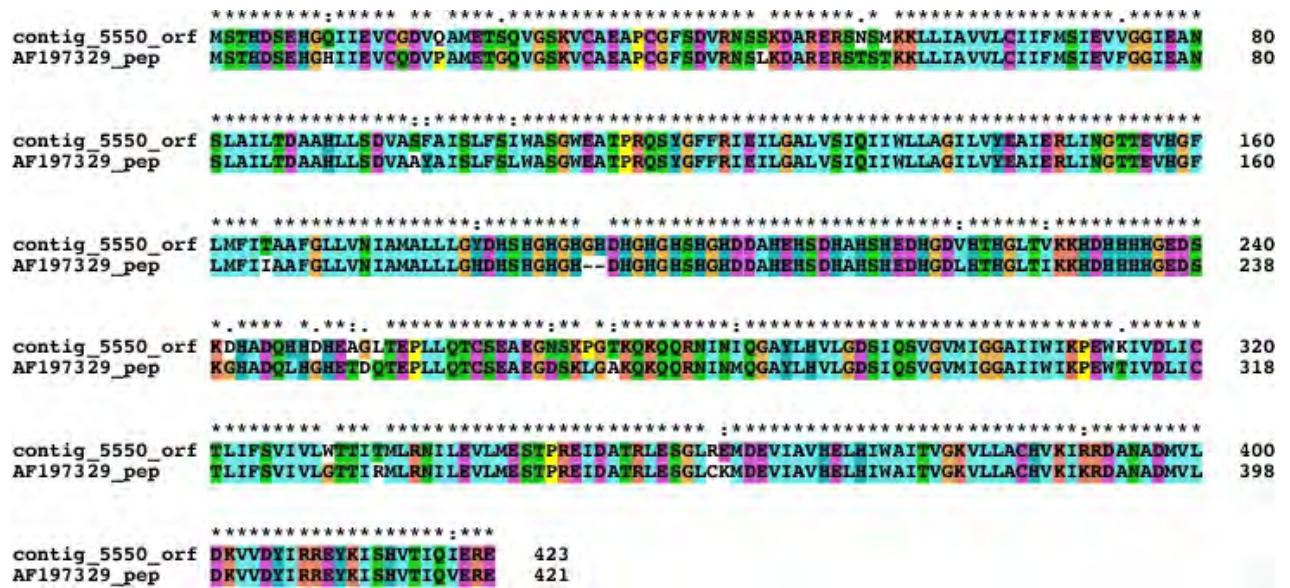


Figure 3.8: Alignment of the protein coding sequence of contig\_5550 and the full length cDNA sequence AF197329.1. The six basepair insert in the assembled contig (contig\_5550) coded for the amino acids glycine and histidine (at position 191 and 192) of the amino acid sequence. Alignment differences between the two sequences can be attributed to the species differences and natural variation between the two organisms represented by the amino acid sequences.

The *de novo* transcriptome assembler OASES (Zerbino *et al.*, unpublished which is based on the Velvet assembler) was used to assemble a transcriptome using the same kmer parameters as was used during the Velvet assembly. The OASES assembler corrects for the difference in expected coverage across transcripts in a dataset, and is able to assemble alternative isoforms of a transcript. By comparing the assembled Velvet contig (contig\_5550) to the assembled OASES transcripts, six shorter transcripts were identified in the OASES assembled dataset, with one of the transcripts (locus\_19278) suggesting that alternative isoforms of the transcript are present in the sequenced biological sample (Figure 3.9B). The OASES assembler holds the promise to be able to detect alternative isoforms of a transcript, but at the time of this study, it was found that it performs this function at the expense of assembling full-length transcripts.

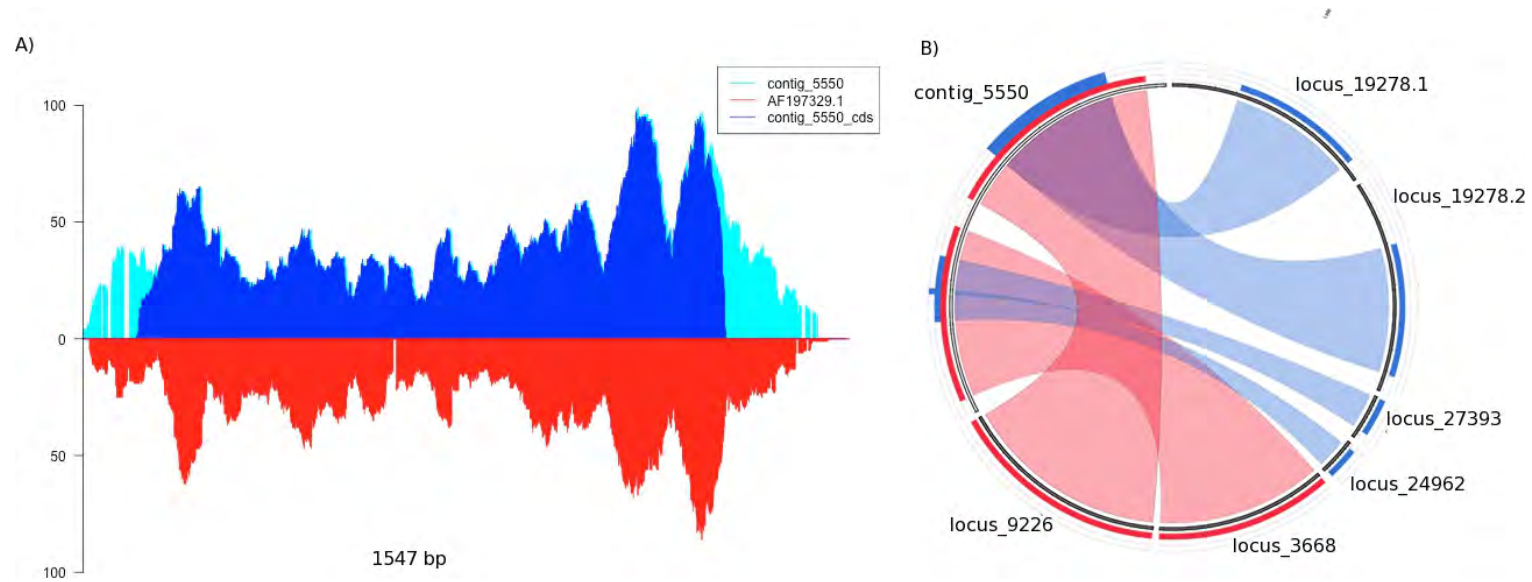


Figure 3.9: Alignment coverage figure of the full length cDNA sequence AF197329.1, the assembled homologous contig (contig\_5550) and the predicted CDS (A) and the OASES assembled transcripts (B). In figure A, the coverage per base are plotted on the y-axis, with the sequence on the x-axis (1 574 bp long). The cyan and blue bars represent the calculated coverage of the assembled contig (cyan) and predicted CDS sequence (blue). The red bars represent the coverage of the genbank sequence (AF197329.1). The six bp indel present in the GenBank sequence is indicated as a gap in the red coverage plot. In figure B, the assembled contig is presented as a light grey box at the top left of the figure. The red bands indicates regions of high similarity between the assembled contig and the loci assembled with OASES, while blue bands indicate lower similarity scores. The figure illustrates that multiple loci are being assembled by OASES at the cost of assembling a single contiguous sequence when compared to the Velvet assembly.

#### 3.3.4. Homology searches

The assembled contig diversity was inspected with the aid of homology-based comparisons of the contigs against the different angiosperm transcriptomes. The *EucAll* (see Section 3.2.5) and assembled gene catalog were binned into six different size categories, and the transcriptome sequences of *Arabidopsis*, *Populus* and *Vitis* compared against the binned sequences (Table 3.3). The results indicate that the assembled contig represented the same sequence diversity present in the *EucAll* dataset, but that a larger number of contigs (1 865, 4 543 and 2 887 vs 6 185, 15 286 and 9 010 for *Arabidopsis*, *Vitis* and *Populus* respectively) are present in the larger size categories (>2 000 bp) of the assembled contigs.

When comparing the assembled contig dataset against the selected angiosperm datasets, a large percentage of the contigs (82% or 15 505 contigs) matched at least one other angiosperm gene sequences (BLAST e-value cutoff at  $1e^{-10}$  and min HSP length of 100 bp, Figure 3.10). Between the *Populus* and the assembled *Eucalyptus* datasets, 14 769 sequences were common, while *Eucalyptus* and *Vitis* shared 14 883 sequences. Between *Eucalyptus* and *Arabidopsis* there were a common set of 14 231 sequences, while 3 552 sequences in *Eucalyptus* did not show similarity to any of the selected angiosperm transcriptomes at an e-value cutoff of  $1e^{-10}$ .

#### 3.3.5. InterProScan

The InterProScan pipeline annotated protein features and/or signatures on 10 557 (56%) of the 15 904 assembled contigs. During annotation, 2 504 distinct protein families (PFAM) were detected, assigning family information to 85% (9 028 contigs) of the 10 557 annotated contigs. PANTHER analysis provided 4 274 distinct functional annotations, with 7 589 (40.16%) sequences annotated and 7 056 sequences (37.43%) were classified in 724 distinct superfamilies, while 1 076 profiles were detected in 5 438 sequences. Conserved domains identified with TIGR HMM models contributed 869 (4.6%) of the total annotations utilising 492 models, and 364 (1.9%) sequences were annotated with 241 Protein Information Resource (PIR) domain identifiers (Figure 3.11).



Table 3.3: A summary of the representation of *Arabidopsis*, *Populus* and *Vitis* genes (number of sequences in brackets) in the constructed public dataset (*EucAll*), and the assembled contig dataset at different e-value thresholds. The assembled contigs contained the same number of homologous contigs as the EucAll dataset (27 939 and 26 848 sequences in *Arabidopsis*), but contained more longer contigs than the publicly available *Eucalypus* datasets (> 2 000 bp).

Angiosperme	e-value	<i>EucAll</i> dataset						Assembled contigs					
		>200bp	>300bp	>500bp	>1000bp	>2000bp	>3000bp	>200bp	>300bp	>500bp	>1000bp	>2000bp	>3000bp
<i>Arabidopsis</i> (33 410)	$1e^{-5}$	27 939	27 394	25 593	17 245	2 002	199	26 845	26 020	24 512	18 516	6 862	2 177
	$1e^{-10}$	26 587	26 202	24 662	16 903	1 940	199	25 538	24 757	23 390	17 744	6 602	2 114
	$1e^{-20}$	24 302	24 129	23 093	16 279	1 865	191	23 242	22 545	21 485	16 569	6 185	1 978
<i>Vitis</i> (75 983)	$1e^{-5}$	63 777	62 197	56 085	36 655	4 862	1 118	59 231	57 312	53 600	40 913	17 716	7 791
	$1e^{-10}$	61 167	59 932	54 585	35 975	4 750	1 088	56 462	54 632	51 231	39 301	16 897	7 374
	$1e^{-20}$	55 264	54 713	50 806	34 412	4 543	989	50 953	49 274	46 526	36 064	15 286	6 522
<i>Populus</i> (45 779)	$1e^{-5}$	38 723	37 835	34 827	23 340	3 107	483	36 922	35 737	33 487	25 348	10 197	3 673
	$1e^{-10}$	36 981	36 308	33 730	22 891	3 038	466	35 131	34 011	31 987	24 395	9 813	3 521
	$1e^{-20}$	33 082	32 789	31 034	21 736	2 887	401	31 546	30 560	28 936	22 451	9 010	3 171

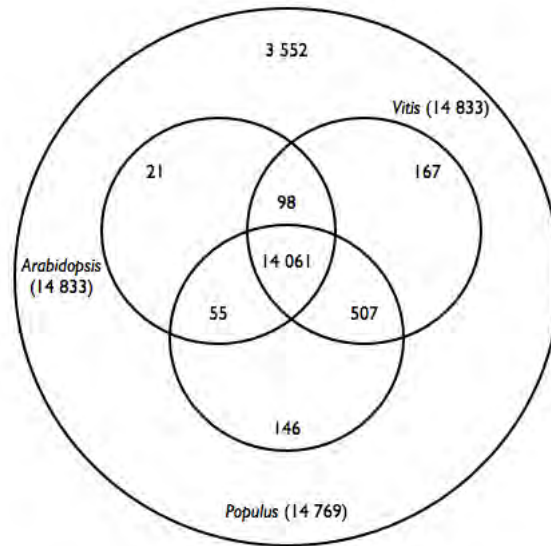


Figure 3.10: Similarity search results of the assembled *Eucalyptus* transcripts against three angiosperm species. In total, 15 505 contigs had homologous sequences in either *Populus* (14 769), *Vitis* (14 833) or *Arabidopsis* (14 883). The results were filtered to contain only high similarity results (e-value  $1e^{-10}$  and a minimum HSP length of 100 bp or 33 amino acids). There were 3 552 *Eucalyptus* sequences that were assembled but did not have homologous counterparts in the selected angiosperm datasets with these filter parameters.

### 3.3.6. Expression profiling

Relative gene expression in terms of Fragments of reads mapped Per Kilobase of exon per Million mapped reads (FPKM, Trapnell *et al.*, 2010) was calculated by mapping the six different mRNA-Seq samples back to the assembled transcriptome, and calculating the transcript abundance with the TopHat (Trapnell and Salzberg, 2009) and Cufflinks (Trapnell *et al.*, 2010) software packages. The expression ratio of the xylogenic tissues (average expression in xylem and immature xylem) *vs.* the non-xylogenic tissues (average expression in shoot tips, mature and young leaf) were used to identify genes which are differentially expressed between the woody and non-woody tissues. A 2X threshold was set, and 3 602 (19.06%) genes were identified with higher expression in the xylogenic tissues, while 879 (4.65%) genes were expressed 2X higher in the non-xylogenic tissues (Figure 3.13A). The expression profile of the selected genes indicate that the genes selected show patterns of co-expression across different tissues (Figure 3.13B).

Gene ontology (GO) category analysis of the over expressed genes in the xylogenic tissues (Figure



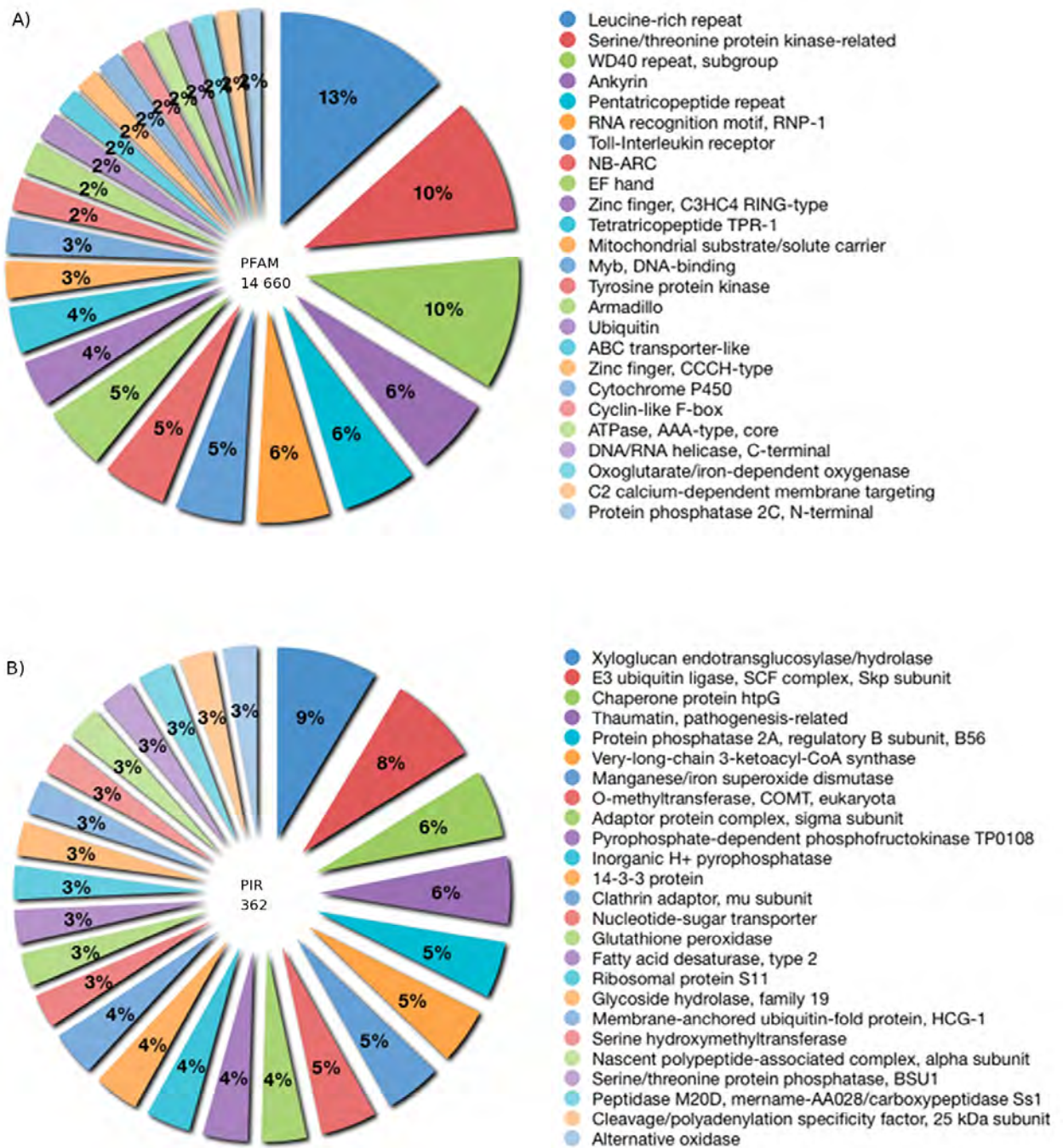


Figure 3.11: The 20 most prevalent protein family (PFAM) and protein information resource (PIR) annotations from InterProScan analysis. The pie charts represent the frequency of the top 20 annotations based on PFAM (a), and PIR (b) annotations. The number of annotations in each annotation category is presented in the center of each pie chart. Leucine repeats and protein kinase-related family members were the most prevalent protein families, and hydrolases, ligases and chaperone protein domains the most frequently annotated PIR features.

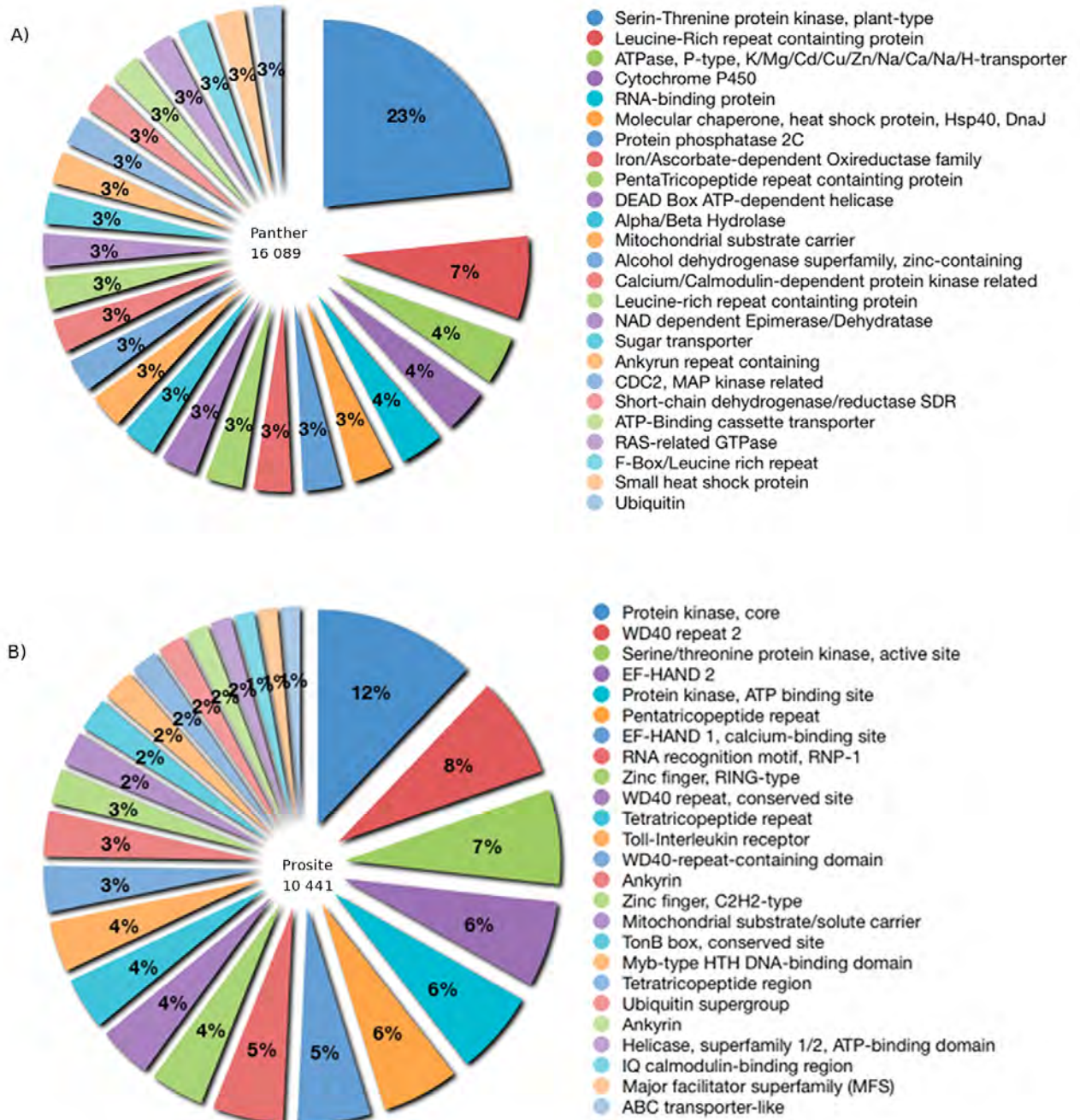


Figure 3.12: The 20 most prevalent Panther (a) and Prosite (b) annotations from InterProScan analysis. Protein kinase signatures were the most prevalent in both annotation sets, as well as the WD40 and leucine-rich repeats.



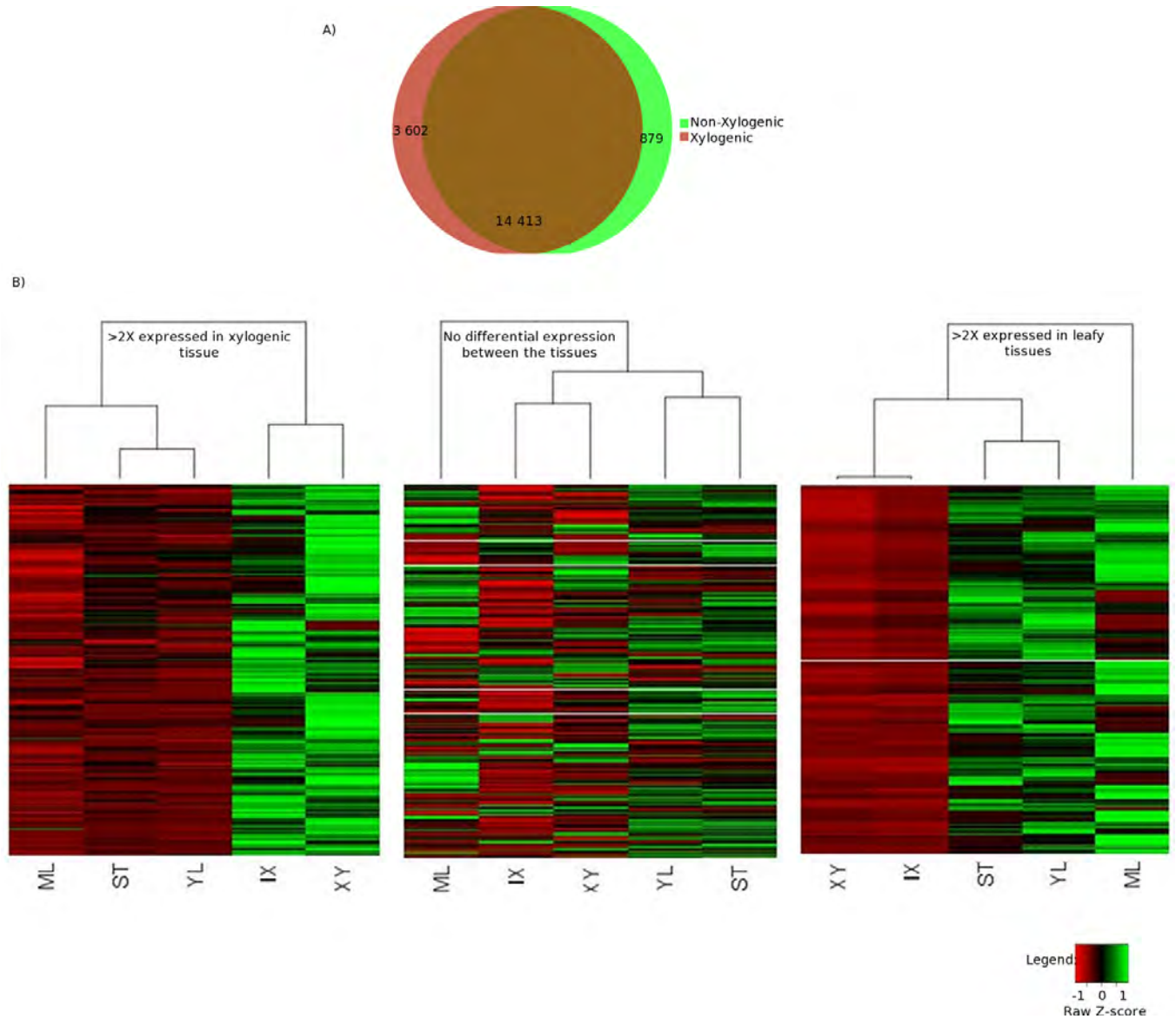


Figure 3.13: Identifying over-expressed xylogenic and non-xylogenic genes (non-xylogenic tissues include mature leaf (ML), shoot tip (ST) and young leaf (YL) tissues, while xylogenic tissues include the immature xylem (IX) and xylem (XY) tissues). Of the 18 894 genes, genes which are expressed 2X higher in xylogenic than non-xylogenic tissues and *vice versa* were identified (A). The expression profiles of the three sets of genes indicate a trend towards co-expression of the genes in the different tissues (B).



3.14, Figure 3.15 and Figure 3.16) and the non-xylogenetic tissues (Figure 3.14B) indicated an abundance in transporter associated, catalytic activity and membrane associated proteins in the xylogenetic tissues. Additional structural components over represented include the vacuole and the plasma membrane, both indicative of transport activity in these tissues. Photosynthetic biological processes and plastid associated genes were most prevalent in the non-xylogenetic tissues, as expected for these photosynthetic tissues.

By mapping the *Arabidopsis* homologs of the 3 602 genes identified as being over-expressed in xylogenetic tissues to the starch and sucrose metabolism pathway (KEGG map00500) in KEGG, xylem over-represented enzymes in the KEGG pathway were identified. The enzymes fructokinase (EC:2.7.1.4), dehydrogluconokinase (EC:2.7.1.13), UDP glucose pyrophosphorylase (EC:2.7.7.9) and alpha-galacturonosyl-transferase (EC:2.4.1.43) showed the largest differentiation in expression in the xylogenetic tissue (Figure 3.17). In the photosynthesis pathway (KEGG map00195) the photosystem II enzymes psbR, psbS and psbP were the most abundant, while the psaD, psaL and psaM photosynthesis I enzymes were the most differentially expressed (Figure 3.18). The annotations of the top 30 genes identified as differentially expressed are presented for xylogenetic (Table 3.4) and photosynthetic tissues (Table 3.5).

From Table 3.4 several known secondary cell wall proteins were identified as being over-expressed in xylogenetic tissues, which validates the approach of performing a de novo assembly with mRNA-Seq data, and making use of the short-read data to infer transcript expression. This included genes involved in growth and shoot development (AT3G53980, Che *et al.*, 2006, AT3G23090, Yuen *et al.*, 2003, AT1G15080, Katagiri *et al.*, 2005), heat shock, disease and stress response pathways (AT5G12030, Wehmeyer and Vierling, 2000, AT5G59720 and AT4G10250, Nishizawa *et al.*, 2006, AT3G53260, Wanner *et al.*, 1995, AT2G35980, Zheng *et al.*, 2004, AT3G51780, Doukhanina *et al.*, 2006, AT2G39530, Cartieaux *et al.*, 2003). Two proteins of unknown function (AT1G0961, Brown *et al.*, 2005) and AT3G0998 that contains the domain of unknown function (DUF662) have also been identified among others as being over-expressed in xylogenetic tissue. More importantly, xylem development genes, such as those identified as being active in the xylem development transcriptional network (AT4G28380, Ko *et al.*, 2006), those involved in secondary cell wall construction (AT5G60490 and AT5G03170, Andersson-Gunnerås *et al.*,

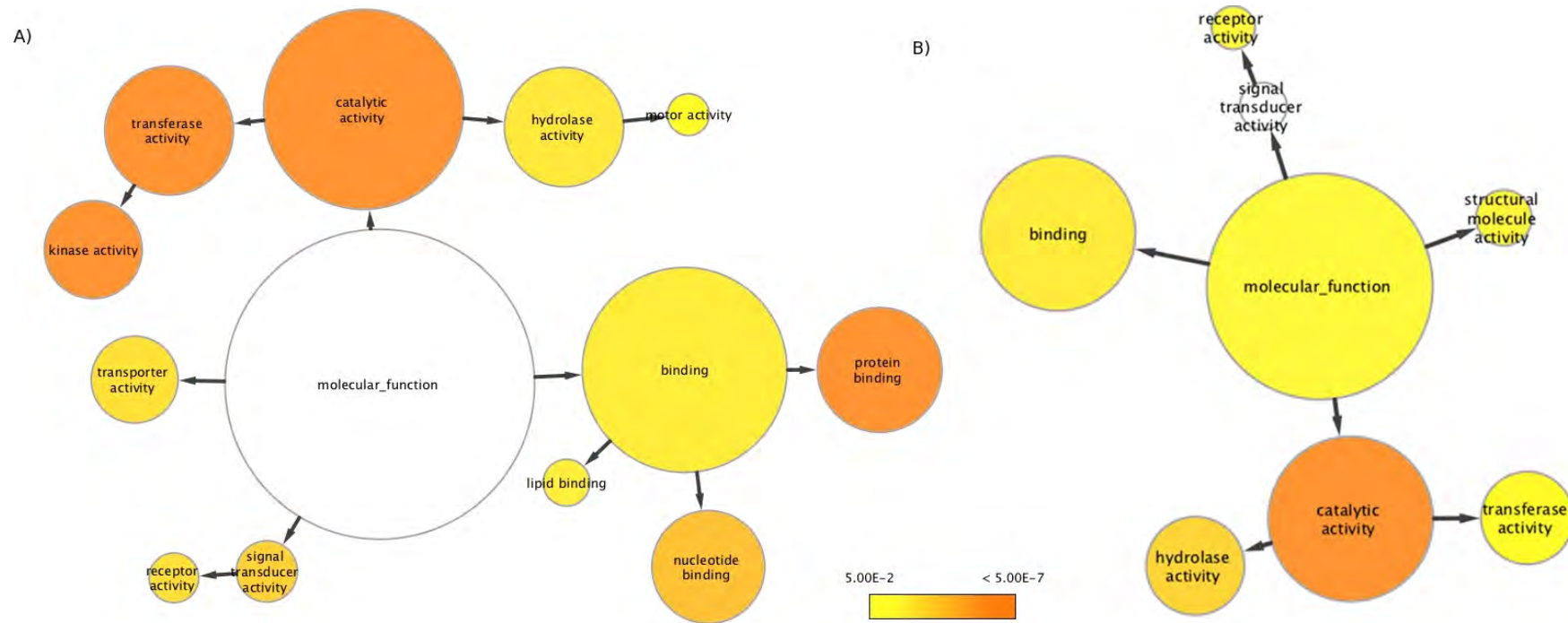


Figure 3.14: Over-represented molecular function gene ontology terms of genes over-expressed in xylogenic and photosynthetic tissues. AMIGO results of over-represented molecular function gene ontology terms in xylogenic (A) and photosynthetic (B) tissues. Xylogenic tissues contained an overrepresented set of terms associated with protein binding and genes with a catalytic activity, especially kinase and transferase activities.

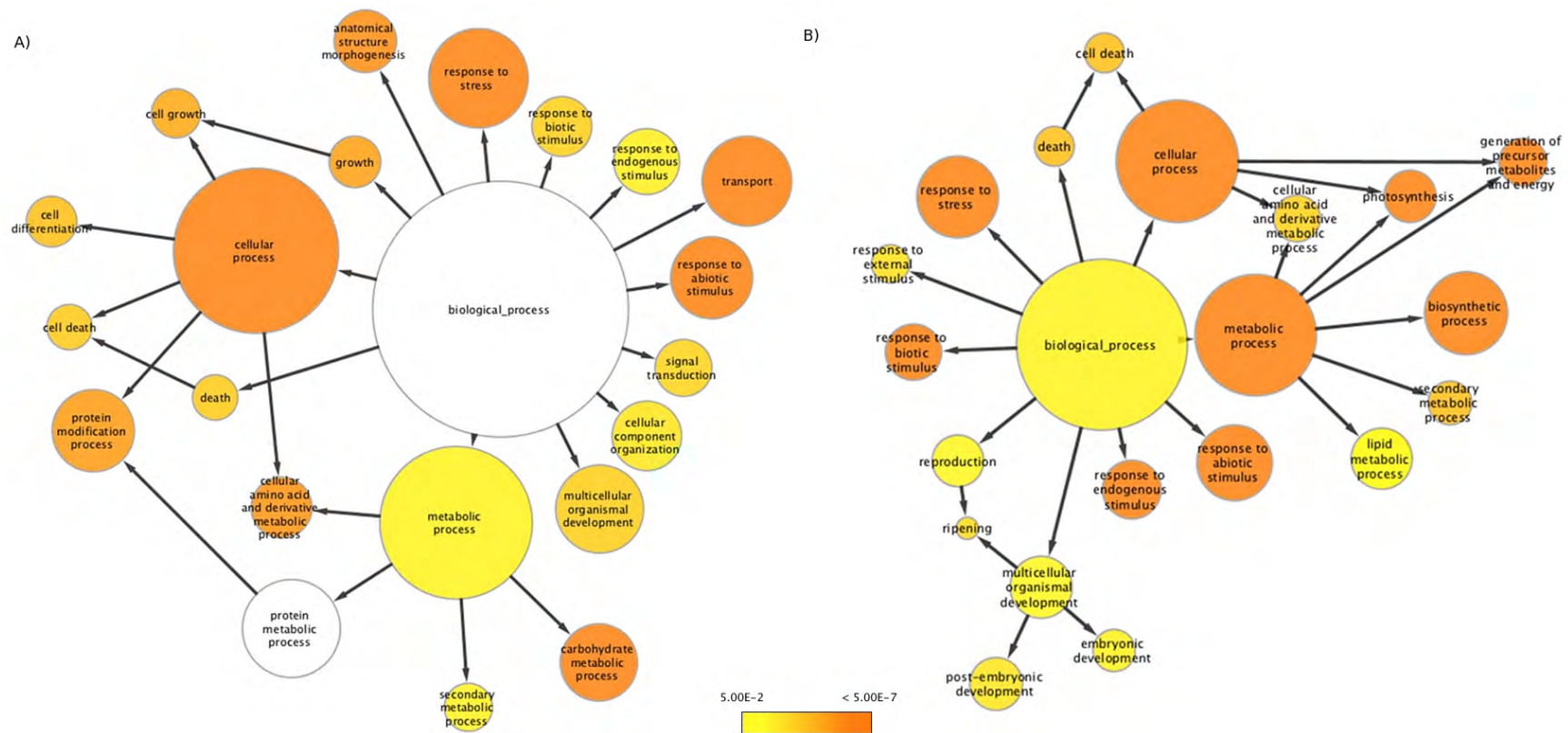


Figure 3.15: Over-represented gene ontology terms of genes over-expressed in xylogenic and photosynthetic tissues. AMIGO results of over-represented genes in xylogenic (A) and photosynthetic (B) tissues. Growth and protein modification processes dominated the xylogenic tissues, while processes associated with biosynthesis and photosynthesis processes were abundant in the photosynthetic tissue dataset.

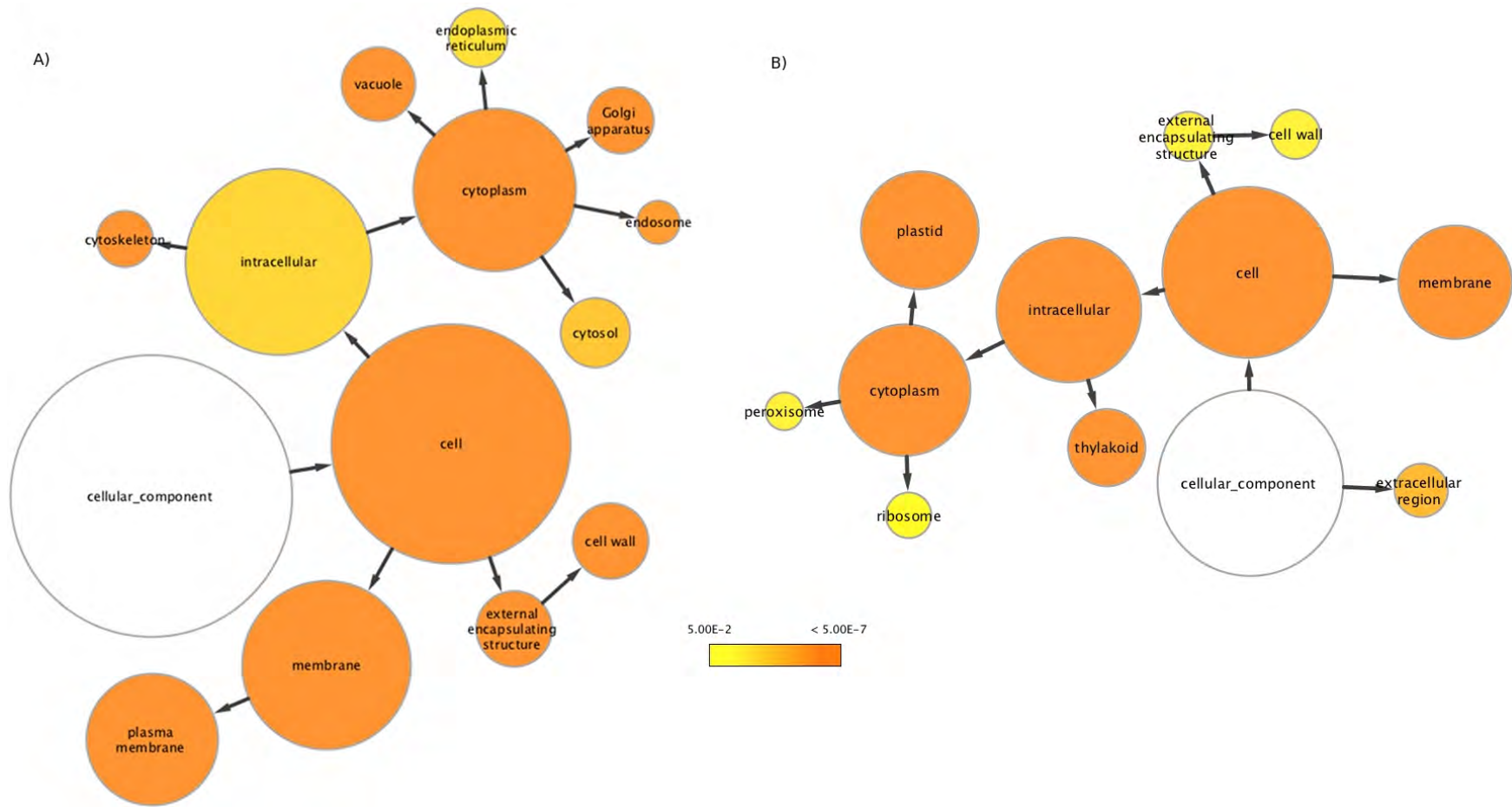


Figure 3.16: Over-represented cellular component gene ontology terms of genes over-expressed in xylogenic and photosynthetic tissues. AMIGO results of over-represented genes in xylogenic (A) and photosynthetic (B) tissues. Cell wall and plasma membrane components were identified as over-represented term in the xylogenic tissues, while terms associated as part of the plastid were over-represented in the photosynthetic set of genes.



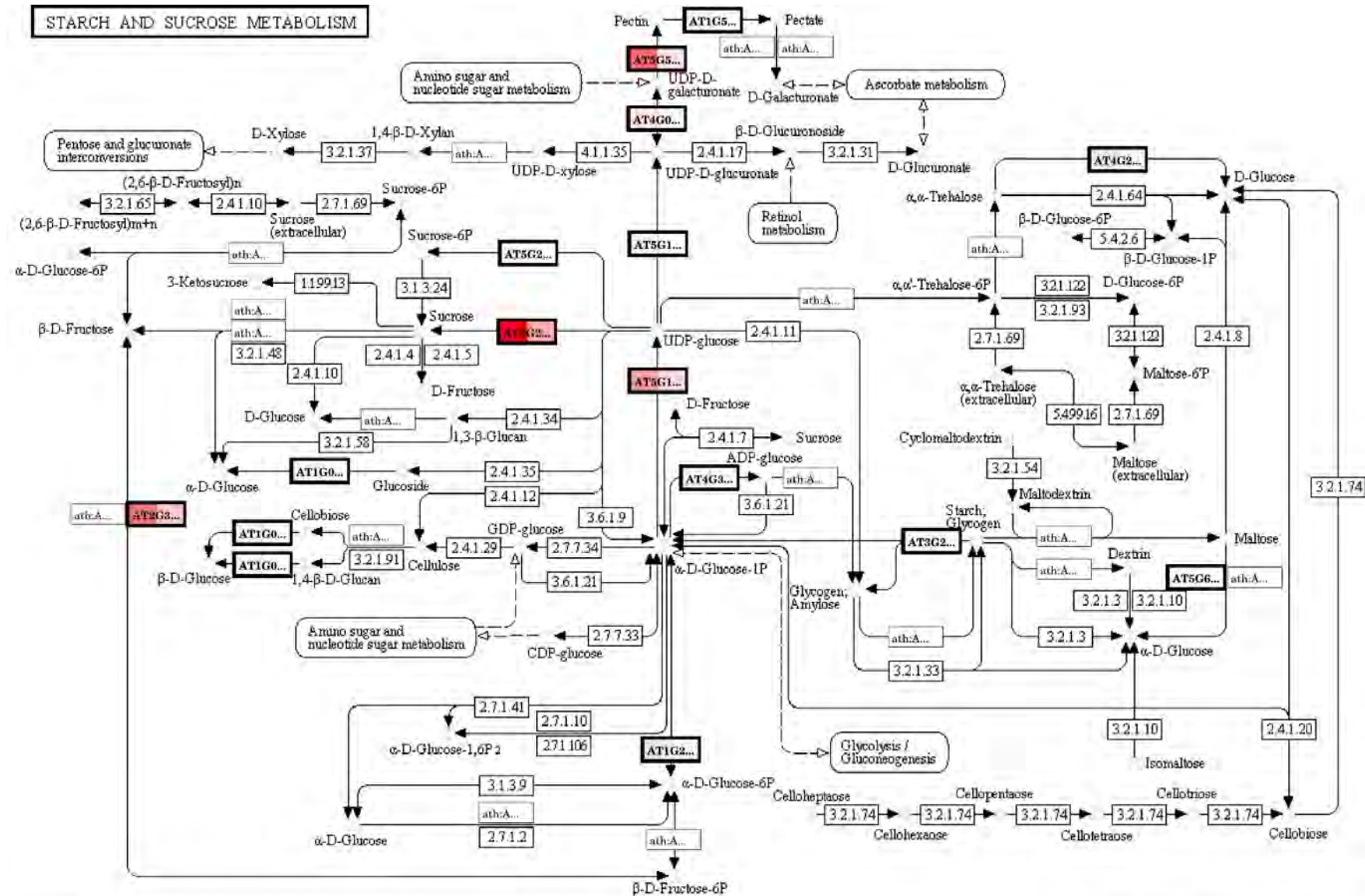


Figure 3.17: Differential gene expression between the xylogenetic and photosynthetic genes represented on the starch and sugar metabolism KEGG pathway. The starch and sugar metabolism pathway were used to identify enzymes higher expressed in xylogenetic than photosynthetic tissues. The enzymes are highlighted relative to their expression in both xylogenetic (left) and photosynthetic (right) tissues, where a dark red indicates a higher expression of the enzyme in the pathway. Results were generated by the Paintomics web-server.

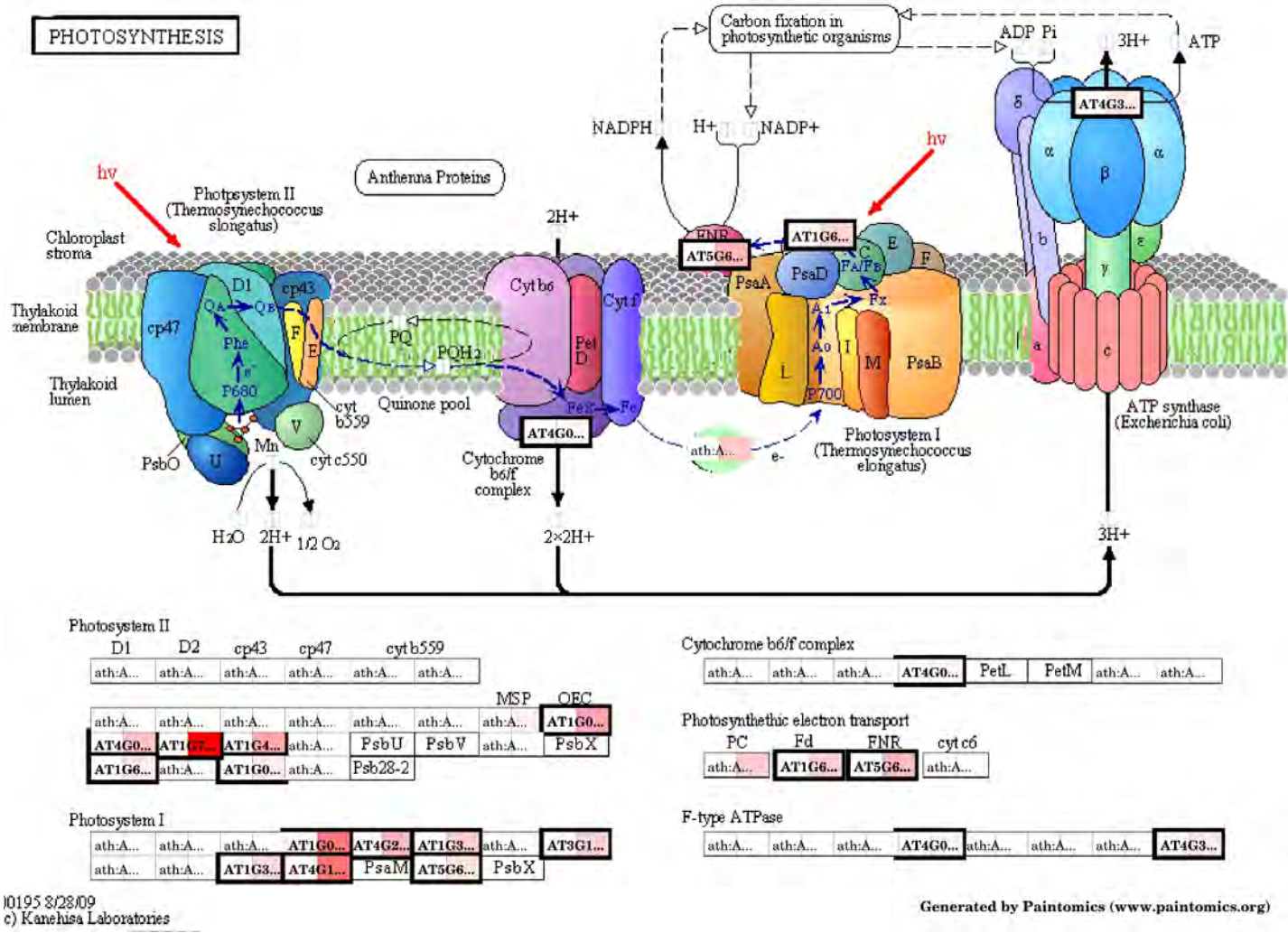


Figure 3.18: Differential gene expression between the xylogenic and photosynthetic genes represented on the photosynthesis metabolism KEGG pathway. The pathway indicates several enzymes higher expressed in photosynthetic tissues compared to xylogenic tissues. The enzymes are highlighted relative to their expression in both xylogenic (left) and photosynthetic (right) tissues, where a dark red indicates a higher expression of the enzyme in the pathway. Results were generated by the Paintomics web-server.

Table 3.4: The top 30 genes identified in the xylogenic tissues, compared to photosynthetic tissues. The ratio between xylogenic and photosynthetic expression were used to select the genes with the biggest differential expression. Only genes with a match (e-value  $< e^{-10}$ ) to an *Arabidopsis* homolog were included in the list.

Contig Name	Arabidopsis homolog	Description	Ratio
contig_139	AT3G53980.2	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	437.20
contig_4304	AT5G12030.1	A. thaliana heat shock protein 17.6A; Unfolded protein binding	388.05
contig_2918	AT5G59720.1	Heat shock protein 18.2	382.59
contig_368	AT1G09610.1	unknown protein	362.02
contig_14996	AT3G09980.1	unknown protein	351.51
contig_16352	AT3G53260.1	Phenylalanine ammonia-lyase	332.46
contig_954	AT2G35980.1	Yellow leaf specific gene 9	235.42
contig_29940	AT4G28380.1	Leucine-rich repeat family protein	221.68
contig_319	AT5G60490.1	FLA12	186.99
contig_35092	AT4G10250.1	Heat shock protein 22.0	185.77
contig_6941	AT5G01300.1	Phosphatidylethanolamine-binding family protein	176.87
contig_17263	AT1G24530.1	Transducin family protein / WD-40 repeat family protein	172.50
contig_13899	AT3G16920.1	Chitinase	170.75
contig_31438	AT3G51780.1	A. thaliana BCL-2-associated Athnogene 4; protein binding	165.51
contig_2525	AT3G23090.1	unknown protein	164.42
contig_4068	AT3G16920.1	Chitinase	161.67
contig_24841	AT1G15080.1	Acid phosphatase / phosphatidate phosphatase	157.65
contig_21284	AT2G39530.1	Integral membrane protein	158.06
contig_1039	AT4G12980.1	Auxin-responsive protein	148.23
contig_63769	AT4G33430.1	BRI1-Associated receptor kinase; kinase/ protein binding / protein heterodimerization	145.16
contig_41003	AT1G50590.1	Pirin	134.02
contig_59694	AT2G30490.1	Ttrans-cinnamate 4-monooxygenase	131.12
contig_3127	AT5G60020.1	Laccase 17	129.04
contig_3811	AT1G27440.1	Catalytic/ glucuronoxylan glucuronosyltransferase	126.59
contig_1532	AT3G16920.1	Chitinase	125.30
contig_17037	AT1G73140.1	unknown protein	124.19
contig_2707	AT5G03170.1	FLA11	122.99
contig_27288	AT2G30395.1	Thalianaovate family protein 17	120.44
contig_65667	AT1G72510.2	unknown protein	116.05
contig_69508	AT3G16920.1	Chitinase	114.86

Table 3.5: Top 30 photosynthetic genes identified as over-expressed in photosynthetic tissue compared to xylogenic tissue. Only genes with a *Arabidopsis* homolog (e-value <  $e^{-10}$ ) were considered for selection.

Contig Name	Arabidopsis homolog	Description	Ratio
contig_17098	AT5G38430.1	Ribulose biphosphate carboxylase small chain 1B / RuBisCO small subunit 1B (RBCS-1B)	393.58
contig_14337	AT2G47400.1	CP12-1, CP12	221.94
contig_22811	AT3G15353.1	Methallothionein 3; copper ion binding	191.79
contig_93397	AT4G27450.1	Unknown protein	171.45
contig_21245	AT5G47230.1	Ethylene responsive element binding factor 5; DNA binding transcription factor	166.67
contig_40682	AT3G01500.3	Carbonic anhydrase 1; carbonate dehydratase/ zinc ion binding	141.56
contig_86098	AT3G19390.1	Cysteine proteinase, putative / thiol protease	141.38
contig_31364	AT1G44575.1	Nonphotochemical quencing (NPQ4); chlorophyll binding / xanthophyll binding	123.27
contig_76583	AT5G22430.1	Unknown protein	91.95
contig_3750	AT5G04660.1	Electron carrier/ heme binding / iron ion binding / monooxygenase/ oxygen binding	91.66
contig_93320	AT4G37360.1	Electron carrier/ heme binding / iron ion binding / monooxygenase/ oxygen binding	91.57
contig_65926	AT1G76080.1	Chloroplastic drought-induced stress protein of 32KD (CDSP32)	75.40
contig_51400	AT4G29270.1	Acid phosphatase class B family protein	72.53
contig_37387	AT5G59320.1	Lipid transfer protein 3 (LTP3)	64.85
contig_46787	AT2G34430.1	Chlorophyll binding ( LHB1B1, LHCB1.4, LHB1B1)	46.80
contig_74523	AT5G48480.1	Unknown protein	40.42
contig_84512	AT4G37300.1	Maternal effect embryo arrest 59 (MEE59)	36.75
contig_32402	AT4G00430.1	Plasma membrane intrinsic protein (TMP-C, PIP1;4, PIP1E); water channel	33.88
contig_93894	AT4G24000.1	Cellulose synthase / transferase, transferring glycosyl groups (ATCSLG2, CSLG2)	33.51
contig_49907	AT3G10450.1	Serine carboxypeptidase like 7; serine-type carboxypeptidase	31.02
contig_61965	AT3G54420.1	Chitinase	26.92
contig_54188	AT1G12090.1	Extensin-like protein; lipid binding	25.54
contig_25739	AT1G79040.1	Photosystem II subunit R (PSBR)	24.49
contig_92707	AT1G68570.1	Proton-dependent oligopeptide transport (POT) family protein	23.61
contig_95912	AT4G25000.1	Alpha-amylase-like (ATAMY1, AMY1)	23.06
contig_37372	AT5G09640.1	Serine-type carboxypeptidase/ sinapoyltransferase (SNG2)	23.01
contig_38811	AT4G03280.1	Photosynthetic electron transfer C (PETC, PGR1)	22.80
contig_83181	AT1G73270.1	Serine carboxypeptidase-like 6 (SCPL6)	22.03
contig_95420	AT5G41120.1	Esterase / lipase / thioesterase family protein	21.92
contig_89772	AT3G03980.1	Short-chain dehydrogenase/reductase (SDR) family protein	21.91



2006, AT3G16920, Brown *et al.*, 2005, AT2G30490, Bayer *et al.*, 2006, AT1G27440, Bosca *et al.*, 2006, AT1G73140 which contains DUF231, Bischoff *et al.*, 2010) and lignin biosynthesis (AT5G60020, Sibout *et al.*, 2005 ) have been identified as up-regulated in the xylogenic tissues.

### 3.3.7. Single nucleotide polymorphism (SNP) detection

SNP diversity was investigated in a subset of the assembled contigs which were deemed to consist of high quality, full length coding genes. The high quality contigs were selected based on the decision tree shown in Figure 3.19. The total contig dataset was separated in CDS and non-CDS-containing reads, and further classified according to homology matches (BLAST e-value of  $e^{-10}$  and a minimum HSP length of 100 bp) of the contigs against various datasets. The 13 806 contigs which contained a predicted CDS and showed high levels of homology against angiosperm protein datasets (*Arabidopsis thaliana*, *Vitis vinifera* and *Populus trichocarpa*) were selected for polymorphism analysis.

A total of 106 658 possible SNPs were observed in these 13 806 contigs. The average SNP density in a predicted coding sequence was 0.21 SNP/100 bp (16 969 SNPs), while the SNP density in the predicted UTR regions was seven fold higher (1.43 SNP/100 bp, 89 689 SNPs). The overall SNP density (CDS and UTR regions) was 0.53 SNP/100 bp, with an average of 7.72 SNPs detected per contig.

## 3.4. Discussion

Deep Illumina mRNA-Seq data analysis of six different tissues of an actively growing six year old *Eucalyptus grandis* x *Eucalyptus urophylla* tree was used to assemble and annotate 18 894 expressed gene transcripts (Table 3.1), producing a well-annotated gene catalog of expressed eucalypt transcripts. The assembly process consisted of performing multiple assemblies of the data with the Velvet assembler in order to identify the set of input parameters that produces the longest contigs with the most bases, corresponding to near full length gene models (Figure 3.3). The assemblies were evaluated with a scoring function that accounts for the number of bases, the number of contigs and the length of contigs to evaluate an assembly (Section 2.3.3). The final assembly (assembly parameters: kmer=31, expected

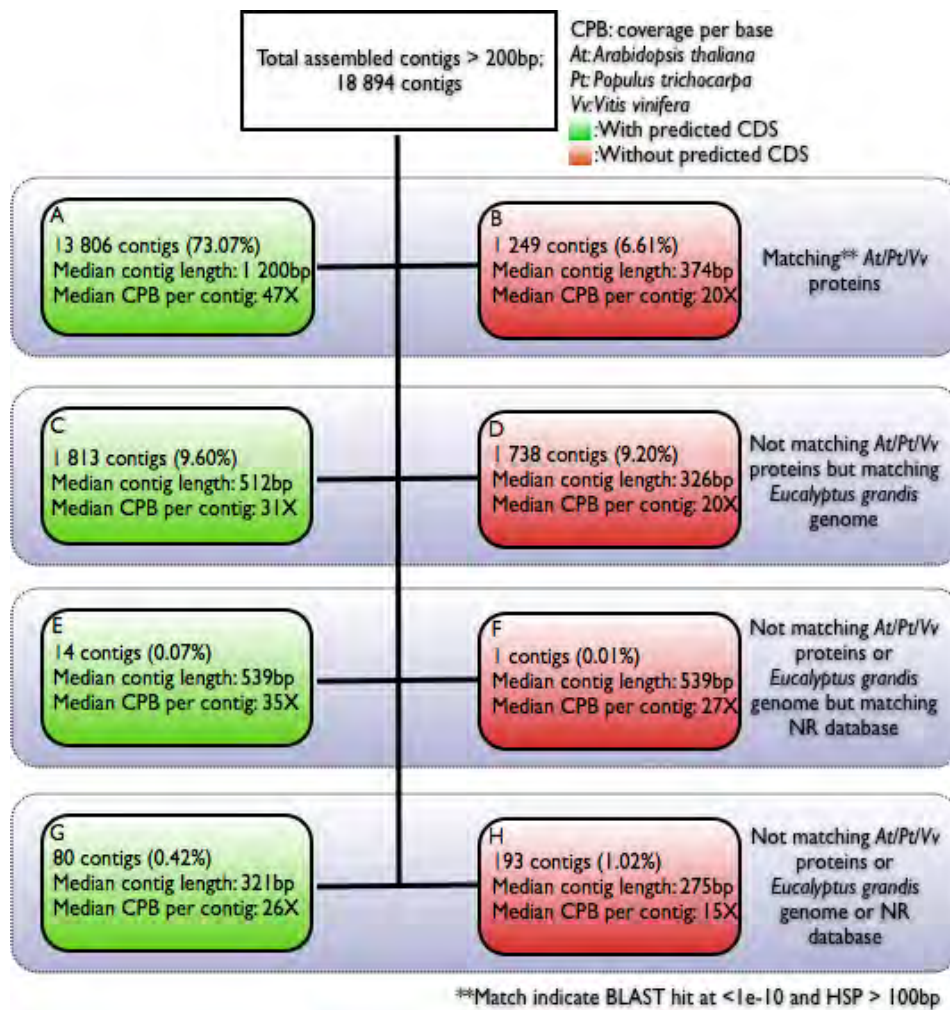


Figure 3.19: Selection of high quality, high confidence contigs for polymorphism detection. The total dataset were queried for contigs that matched against *A. thaliana*, *P. trichocarpa* or *V. Vinifera* proteins, and seperated based on the presence of a predicted CDS (A and B). The remaining contigs were then used to identify matches against the *E. grandis* genome sequence (B and C), and the NCBI non-redundant (NR) protein database (E and F, and G and H).

coverage = 1 000 and a coverage cutoff value of 8X) consisted of 22.8 million bases in approximately 40 000 contigs (Table 3.1). A novel assembly strategy, where the expected coverage value of each individual contig was calculated and the contig together with all the reads that matched to the contig were used for a coverage assisted re-assembly, yielded an additional 400 000 bases to the assembly, with most of the additional bases added to the shorter contigs (Section 3.2.2, Table 3.1 and Figure 3.4). The final assembly, containing only contigs longer than 200 bp, comprised of 22.1 Mbp transcript catalog in 18 894 contigs with an N50 length of 1 640 bp. Further inspection of the extended contigs indicated that most of the additional reads added during the coverage-assisted re-assembly were added to the start and end of the *de novo* assembled contigs (Figure 3.5 and Appendix B), with the exceptions of some low quality regions (Ns in the assembled contig), that became artificially elongated during the re-assembly process. This dataset represents the most complete gene catalog for a Eucalyptus forest tree produced making use of uHTS technology data (Novaes *et al.*, 2008).

Due to the nature of the assembler used, the assembled dataset would not contain full length alternative transcripts of the gene models assembled. *De Bruijn* graph assemblers returns the longest contigs with the most coverage as a consensus contig, and alternative fragments would be lost. *De novo* transcriptome specific assemblers, such as **trans-ABYSS** (Birol *et al.*, 2009), **OASES** (Zerbino *et al.*, unpublished) and **Trinity** (Grabherr *et al.*, 2011), became available at the end of the lifetime project and were not considered as alternative assemblers. The chosen assembler did however manage to assembly long contiguous transcripts that could be used for transcriptome profiling.

*Ab initio* coding sequence prediction tools were used to identify transcriptional start and stop sites in the assembled dataset. These prediction tools were developed to make use of genomic sequence, where it relies on finding sequence features in a predefined order, for example, in a 5' to 3' direction. These methods take into account the presence of promotor regions, the TSS, 5' UTR, start codon, exons, splice donors, introns, splice acceptors, stop codons, 3' UTR and polyA tail. In the case of partially assembled contigs producing coding sequences, when a feature needed for that stage of the HMM prediction state is not present, the predictor would be unable to exit the current state and fail to continue searching for

features in the rest of the sequence, resulting in a negative or incomplete CDS prediction. This can be classified as a false negative prediction, since the gene product is present in the assembly, but the gene model is incomplete. Much of the variation in the prediction of coding sequences can be attributed to the low sensitivity (70%) and specificity (60%) generally observed by *ab initio* gene prediction software (Blanco and Guigó, 2005), and together with the incomplete nature of the assembled contigs, most of the variation in coding sequence prediction results are explained (Table 3.2). The GENSCAN tool predicted 98% of the total coding sequences predicted by a combination of GLIMMER, GENSCAN and AUGUSTUS. The predicted CDS from GENSCAN were subsequently used to evaluate the contiguity of the assembled contigs.

A total of 33 full-length cDNA sequences representing a range of gene families were used to inspect the contiguity of the assembled contigs and predicted CDS sequences. Short indels were present in most of the UTR regions of the assembled sequences when compared to the full-length cDNA sequences, with a very low frequency of indels present in the CDS sequences. No gross misassemblies were observed in the tested dataset (Appendix C.1.2). Results from the *de novo* transcriptome assembler OASES were also compared to the assembled dataset, but the extended Velvet assembly produced longer, higher quality contigs in general. To further assess the quality of the assembled dataset, an in depth comparison between the predicted gene models based on the genome sequence will be performed when the gene models become available, but the current analysis provided great confidence in the quality and contiguity of the *de novo* assembled gene catalog.

The diversity of the assembled contigs was firstly evaluated by performing various homology-based searches against other angiosperm datasets. The assembled dataset represented longer, more diverse sequences than the previously available public dataset (*EucAll*), and over 14 000 contigs showed high similarity with other angiosperm species. A subset of the sequences did not show any homology to known angiosperm proteins, and these will be further investigated when the full set of gene models are available from the *Eucalyptus* genome annotation effort. InterProScan analysis provided the second measure of diversity for the assembled dataset. Over 10 000 protein sequences were annotated with a functional domain, allocating sequences to over 2 500 distinct protein families. These annotation together with the

Gene Ontology annotations made to the assembled dataset assigned valuable functional annotations to the sequences, which became especially useful during the expression profiling of the sequences.

By assigning relative expression values, in the form of FPKM values to each of the genes for each of the tissues sampled and sequenced, genes highly expressed in wood forming (xylogenic) and photosynthetic tissues were identified. The results indicate, as expected, that the xylogenic tissues have an over-abundance of transporter-associated, catalytic- and membrane-associated genes expressed, as well as an over-expressed set of structural proteins. Photosynthetic pathways and processes were the most abundant in the leafy and phloem tissues. A similar approach was followed in Mizrahi *et al.* (2010), where genes for which a high correlation in terms of expression patterns with some of the primary cell wall genes was observed. The database of expression patterns developed will serve as an starting point for more in depth analysis of expression correlation and tissue specific expression of various genes and pathways in future studies.

In the 13 806 contigs that were considered for putative SNP detection, 16 696 SNPs were identified in coding regions (0.206 SNPs/100 bp, 89 962 SNPs were identified in UTRs), resulting in an overall SNP density for coding and non-coding regions of 0.534 SNPs/100 bp (compared to genomic SNP density of one SNP per 17 bp, Külheim *et al.*, 2009). Furthermore, the theoretical designability of Illumina GoldenGate and Infinium HD Genotyping assays (<http://www.illumina.com>) was determined. This analysis ignored the presence of introns in the sequence, and is thus an over-estimation of the number of possible SNPs that can be used in the assays. Of the 106 658 putative SNPs, 73% (77 631) passed the initial 50 bp flanking window filter where no other polymorphisms should be present in order for the probes to bind, of which 16% (12 285 SNPs or 0.17 SNPs/100 bp) occurred within predicted coding regions. For the 60 bp window, a total of 12 070 coding regions SNPs (0.168 SNPs/100 bp) and 64 225 UTR SNPs (1.207 SNPs/100 bp) were detected. Assay designability performed by the Illumina support team (<http://www.illumina.com/support>) revealed that 68 606 (90%) of the SNPs had an Infinium HD Assay designability score higher than 0.8, and 68 579 (90%) had GoldenGate Genotype designability scores of 1.0. These results indicate that by designing the SNP assays based on the coding regions of the

genomic sequence, these two Illumina platforms could be useful for SNP genotyping and genetic mapping of thousands of expressed genes in a interspecific hybrid pedigree.

### 3.5. Conclusion

In this study we successfully assembled a draft gene catalog of an *Eucalyptus grandis* x *Eucalyptus urophylla* hybrid clone using deep mRNA-Seq from six different sampled tissues. The assembled transcriptome was evaluated in terms of contig contiguity and homology to other angiosperm transcriptomes. The assembled dataset does not contain only full length transcripts, but through investigation into the structure and nature of the assembled contigs, it can confidently be described as the most complete gene catalog hitherto of a single *Eucalyptus* tree. The level of completeness of the transcripts can only be fully evaluated when a complete, annotated genome sequence becomes available.

Functional annotations were assigned to the assembled transcriptome dataset, providing insight to the active transcriptional landscape of the organism. The expression profile of each assembled contig in the six sampled tissues were calculated and used to identify over-expressed genes in xylogenic and photosynthetic tissues. Several genes known to be active in secondary cell-wall formation (such as FLA11 and FLA12) and lignin biosynthesis (such as LAC17) were identified in the list of top 30 genes over-expressed in xylogenic tissues.

The dataset produced can be considered as a first step towards identifying transcriptional control networks active in a fast-growing wood-forming organism. Transcriptional profiles of individual trees with different genetic background (mapping populations), disease and physiological states will soon become available, which will soon shed more information on the level of gene co-expression and underlying active transcriptional modules involved in wood formation.

## Chapter 4

# Eucspresso: Towards the development of a *Eucalyptus* genome and transcriptome information resource

## Preface

This chapter describes the development of a public data resource that contains sequences and annotations for the 18 894 *de novo* assembled transcripts of a *Eucalyptus grandis* x *Eucalyptus urophylla* hybrid tree (Chapter 3). The resource was developed to provide users with access to the annotation and sequence data described in Chapter 3, and was published as part of the research manuscript describing the *de novo* assembly of the *Eucalyptus* hybrid transcriptome (joined first author publication):

- Mizrachi, E., Hefer, C.A., Ranik, M., Joubert, F. and Myburg, A.A., 2010. *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. **BMC Genomics**, Volume 11, 681.

Author contributions: E. Mizrachi, M. Ranik and A.A Myburg assisted in the general design of the features in the database, F. Joubert assisted with some technical challenges during development, and C.A. Hefer developed and designed the database and web interface.

The database resource, **Eucspresso** is available at the following URL:

<http://eucspresso.bi.up.ac.za>. Public access is granted to all the entries in the database.



## 4.1. Introduction

The release of the *Eucalyptus grandis* genome sequence and gene model annotation (Version 1.0, <http://www.phytozome.net>) in January 2011 provided forest tree geneticists with an opportunity to investigate gene targets for the genetic manipulation of the most abundant plantation tree in the Southern hemisphere. Traditionally, after the completion or release of a newly sequenced genome sequence, the immediate focus of research programmes shifts towards defining the characteristics of each functional unit in the genome. This translates to, among others, the identification and annotation of genes, the identification of gene expression regulation mechanisms, regions on the genome associated with certain traits and finally genomic targets for the genetic manipulation of the organism of interest. It is imperative that access to the different datasets and annotations associated with a sequenced genome is made available in a user friendly and easily accessible form to support research on the organism.

Several widely used plant genomics databases already exists for a variety of plant species (*Arabidopsis* Garcia-Hernandez *et al.*, 2002, *Zea mays* Lawrence *et al.*, 2004, *Populus* Sjödin *et al.*, 2009, *Brachypodium* and *Oryza* Zhao *et al.*, 2004), with some resources available for a range of plant species (<http://phytozome.net>, PlantGDB (Duvick *et al.*, 2008)). The focus of these resources range from performing comparative genomics and transcriptomics between plants, to hosting gene expression datasets. To facilitate research on the newly sequenced *Eucalyptus grandis* genome sequence, we envisioned the development of a *Eucalyptus*-focussed mRNA-seq gene expression database. As a first step to the development of such an mRNA-seq repository, we focussed on the development of *Eucspresso*, a module of the resource that focusses on the expression of genes in a eucalypt hybrid plantation tree.

The availability of a *de novo* assembled gene catalog of an *Eucalyptus grandis* x *Eucalyptus urophylla* F1 hybrid tree and its associated annotations, tissue specific gene expression information and close angiosperm homologs (Chapter 3 and Mizrachi *et al.*, 2010) necessitated the need to develop a central database to store the annotations for each of the 18 894 contigs in the dataset. The aim of the database is to provide access to the basic annotations performed on the dataset via a user-friendly, web-based interface. The interface has to cater for different search scenarios, where the user can search for contig

names, homolog IDs and sequences (BLAST), annotations and lists of terms or IDs. The interface also has to link to a genome browser instance of the 8X *Eucalyptus grandis* genome assembly to identify the genomic locations of the assembled transcripts.

## 4.2. Materials and methods

### 4.2.1. MySQL database

The database backend consisted of a **MySQL** database that stores the assembled transcript sequences and associated annotations. The **Eucspresso** data model was based on the open source **BioSQL** sequence data model (<http://www.biosql.org>), where each entry in the database inherits from a single **BioEntry** table. This design allowed for the effective storage of metadata, such as entry names, text-based descriptions and accessions in a single, indexable table that enhances the search capabilities of the database. Programmatic access to the entries in the database was provided through the **Python** based object relational mapper (ORM) **SQLAlchemy** (<http://www.sqlalchemy.org>), which also handles the field or property inheritance between the objects stored in the database.

### 4.2.2. TurboGears Web framework

The **TurboGears** (version 1.09b, <http://www.turbogears.org>) web framework was used to develop the **http** interface to the database. **TurboGears** enforces a model-view-controller design paradigm, with a software layer that provides access to the database backend or the model, logic code in a **Python** environment as the controllers, and a templating system to generate the viewable **HTML** code. As mentioned, the framework uses an ORM to construct custom **Python** objects that can be passed to and from the different layers. The **Genshi** templating engine (<http://genshi.edgewall.org>) provides a XML-based templating framework that is converted to the viewable **HTML** pages. **Eucspresso** is served by the default **CherryPy** web-server (<http://www.cherrypy.org>) at the current URL (<http://eucspresso.bi.up.ac.za>).

### 4.2.3. Custom Python controllers and R scripts

Python and R scripts were developed to provide the logic that interacts with the data model and perform on-demand analysis that enhances the interface. The Python simple object access protocol (SOAP) was used to access the remote KEGG server (<http://soap.genome.jp/KEGG.wsdl>) to render KEGG pathways with the annotated enzyme highlighted on the pathway. The GO graphs are downloaded upon request from the AMIGO web server (<http://amigo.geneontology.org>), and stored on the local server. After the KEGG maps and GO images are retrieved from the remote servers, the images are stored locally which are then used if the image is requested again. R-scripts are used to display the FPKM expression values of the selected gene as a bar chart.

## 4.3. Results and discussion

### 4.3.1. Eucspresso data model

The central entity of the Eucspresso data model is the BioEntry table (Figure 4.1). All data types stored in the database inherit properties from the BioEntry table. Search indices have been created for the BioEntry.Id, BioEntry.Accession, BioEntry.Identifier, BioEntry.Description and BioEntry.Name columns. The BioEntry.Datatype field stores the value of the child table that inherits the properties from the BioEntry table. By creating a single point of inheritance (the BioEntry table), a search can be performed across all datatypes at the same time, which increases the efficiency of searching. The BioEntry table stores a primary identifier of each of the entries in the Eucspresso database and contains over 1.5 million records.

The BioSequence table stores the sequence information related to each of the 18 894 contigs in the database. Each annotation associated with a contig has a foreign key (foreign keys are not shown in Figure 4.1) that relates the annotation to the contig. This allows the user to search for a contig and display the annotation, as well as search for a keyword term in the annotation field, and display all the contigs that share the annotation.

SQLAlchemy was used to construct the queries to the database, and provide custom objects that represent entries in the database. These custom data mappers makes use of the foreign key constraints between the Python data objects to build custom objects that are send to the Genshi template system to render the HTML pages in a browser.

#### 4.3.2. Browsing and searching for a contig

The primary entry point to the database is the contig browsing table (Figure 4.2). The table consists of a `ToscaWidgets` (<http://www.toscawidgets.org>) grid interface that uses JavaScript object notation (JSON) to populate the display table with a subset of entries (by default 25 sequences, but the user can customise it). The table is sortable on the contig name and length columns. The table contains the best homology based search (BLAST) result, and the first description of each of the GO, EC and InterProScan annotation assigned to the contig. Searching is possible based on Arabidopsis (AT) accession and description, GO, EC and InterPro annotation description, as well as the contig name. The results from searching is displayed in the same table, after a JSON request was submitted to the server and the results of the query returned back to the browser (Figure 4.2B).

#### 4.3.3. Visualising a contig and associated annotation

A summary of the annotations of a contig is presented as a "Summary" tab when the user clicks on the "View" link in the contig browsing table. The summary tab contains detail regarding the contig such as the length and GC content, the length of the GenScan predicted ORF, the closest homolog of the sequence found in either of the *Arabidopsis*, *Populus* or *Vitis* protein sequences, and an overview of the GO and KEGG annotations for the contig (Figure 4.3A). More detail is presented in each of the tabs at the top of the page. The "Sequence Detail" tab presents the cDNA and predicted protein sequence of the contig, as well as links to download the sequences (Figure 4.3B).

The top 20 BLAST results against the *Arabidopsis*, *Populus* and *Vitis* transcriptome datasets are presented in the "Homology search results" tab, with links to the TAIR (*Arabidopsis*) and Phytozome (*Populus* and *Vitis*) entry for each of the homologous sequences (Figure 4.4A). The "Gene Ontology" tab

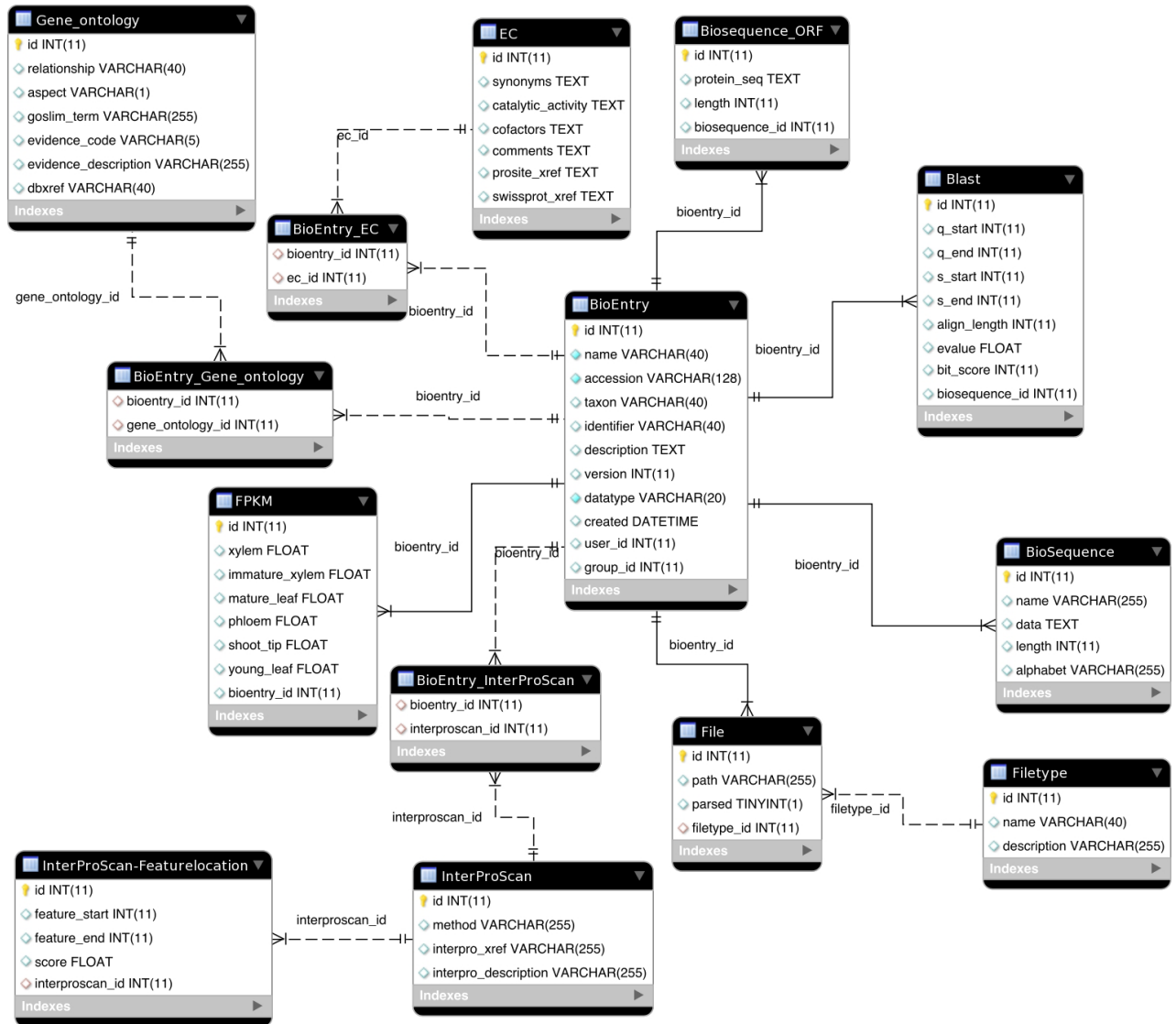


Figure 4.1: Entity relationship diagram of the main datatypes in *Eucspresso*. All the datatypes inherit attributes from the BioEntry table. The description and accession attributes of the BioEntry table are used for searching, and any link between different results occur through the BioEntry table.

**Eucspresso**  
The *Eucalyptus* gene expression database

Welcome Browse Advanced Search Contact FAQs

**Assembled Sequences**

The 18,894 sequences assembled during the experiment. To view more detail of a particular sequence, click on the view icon. The sequences are sortable by Name and Length, just click on the header pane of the table. You can also perform a simple search through the table based on the sequence name by clicking on the search icon at the bottom left of the table. To search for annotations, i.e retrieve all the sequences that was annotated with a specific GO, EC or InterPro accession, follow the search link.

**All Sequences**

View	Name	Length (bp)	Best At ortholog ID	Best At ortholog description	GO description	EC description	InterPro description
	contig_21861	1313	AT5G16220.1	octicosapeptide/Phox/Bem1p (PB1) domain			
	contig_10000	1061	AT1G33490.1	unknown protein	membrane		
	contig_10001	2390	AT5G57360.1	ZTL (ZEITLUPE); protein binding / ubiquitin	scavenger receptor activity		Kelch repeat type 2
	contig_10010	3107	AT4G24680.1	FUNCTIONS IN: molecular_function unknown			
	contig_10011	2132	AT5G10360.1	EMB3010 (embryo defective 3010); structure	ribosome	Protein-synthesizing GTPase.	Ribosomal protein S6e
	contig_10017	2664	AT3G46220.1	unknown protein			Protein of unknown function DUF2042
	contig_10018	1264	AT4G17510.1	UCH3 (UBIQUITIN C-TERMINAL HYDROLASE)	intracellular	Ubiquitin thioesterase.	Peptidase C12, ubiquitin carboxyl-terminal
	contig_10019	2034	AT4G03420.1	unknown protein			Protein of unknown function DUF789
	contig_10020	949	AT3G09110.1	unknown protein			Protein of unknown function DUF674
	contig_10021	700	AT5G47680.1	FUNCTIONS IN: molecular_function unknown	tRNA (guanine-N1-)-methyltransferase activity	tRNA (guanine-N(1)-)-methyltransferase.	tRNA (guanine-N1-)-methyltransferase, eu

Quick Search  Search AT Accession Search Clear Download

10 Page 1 of 1890 Displaying 1 to 10 of 18894 items

**Eucspresso**  
The *Eucalyptus* gene expression database

Welcome Browse Advanced Search Contact FAQs

**Assembled Sequences**

The 18,894 sequences assembled during the experiment. To view more detail of a particular sequence, click on the view icon. The sequences are sortable by Name and Length, just click on the header pane of the table. You can also perform a simple search through the table based on the sequence name by clicking on the search icon at the bottom left of the table. To search for annotations, i.e retrieve all the sequences that was annotated with a specific GO, EC or InterPro accession, follow the search link.

**All Sequences**

View	Name	Length (bp)	Best At ortholog ID	Best At ortholog description	GO description	EC description	InterPro description
	contig_31	3376	AT5G17420.1	IRX3 (IRREGULAR XYLEM 3); cellulose synthase	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
	contig_2805	3442	AT5G44030.1	CESA4 (CELLULOSE SYNTHASE A4); cellulose synthase	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
	contig_27025	599	AT3G07330.1	ATCSLC6 (CELLULOSE-SYNTHASE LIKE)	transferase activity		
	contig_268	3308	AT4G18780.1	IRX1 (IRREGULAR XYLEM 1); cellulose synthase	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
	contig_22590	3797	AT5G05170.1	CEV1 (CONSTITUTIVE EXPRESSION OF CELLULOSE SYNTHASE)	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
	contig_21138	2517	AT4G07960.1	ATCSLC12 (CELLULOSE-SYNTHASE LIKE)	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Glycosyl transferase, family 2
	contig_19509	4145	AT4G39350.1	CESA2 (CELLULOSE SYNTHASE A2); cellulose synthase	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
	contig_18438	2179	AT2G21770.1	CESA9 (CELLULOSE SYNTHASE A9); cellulose synthase	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
	contig_18095	3780	AT3G03050.1	CSLD3 (CELLULOSE SYNTHASE-LIKE D)	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
	contig_1406	2679	AT5G03760.1	ATCSLA09; mannan synthase/transferase	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Glycosyl transferase, family 2

Quick Search cellulose Search AT Description Search Clear Download

10 Page 1 of 2 Displaying 1 to 10 of 12 items

Figure 4.2: Browsing and searching for contigs through the Eucspresso web interface. The table consist of a ToscaWidget table, that sends queries to the database through a JSON controller. The entries can be sorted by contig name and length (A) and dynamic searches can be performed on the entries in the table. Searching for the "cellulose" keyword that occurs in the "AT description" column, returns 12 items to the table (B). A link to the detailed description of the contig in the table is provided by clicking on the "View" column in the table.

A)

**Eucspresso**  
The *Eucalyptus* gene expression database

Welcome Browse Advanced Search Contact FAQs

Summary Sequence detail Homology search results Gene Ontology Enzyme Commission InterProScan results Tissue-specific expression GBrowse

**Summary of: contig\_31**

Below is a short summary of the most important features and annotations of contig\_31. For a more detailed view of the annotations and features of this record, select one of the tabs above.

**Sequence detail**

Length	3376 base pairs
GC content	45.59 %

**Predicted Open Reading Frame**

Predicted ORF length	1010 Amino Acids
----------------------	------------------

**Homology results**

Best Arabidopsis ID	AT5G17420.1
Best Arabidopsis Description	IRX3 (IRREGULAR XYLEM 3); cellulose synthase
Organism	<i>Arabidopsis thaliana</i>

**Ontology results**

Gene ontology terms	GO:0016021; GO:0016760; GO:0008270; GO:0005515; GO:0030244
Gene ontology classes	Cellulose biosynthetic process; Protein binding; Cellulose synthase (udp-forming) activity; Integral to membrane; Zinc ion binding

**Enzyme commission results**

Enzyme Commission terms	2.4.1.12
Enzyme Commission description	Cellulose synthase (UDP-forming).

B)

**Eucspresso**  
The *Eucalyptus* gene expression database

Welcome Browse Advanced Search Contact FAQs

Summary Sequence detail Homology search results Gene Ontology Enzyme Commission InterProScan results Tissue-specific expression GBrowse

**Sequence: contig\_31**

Below is the assembled sequence, as well as the longest predicted Open Reading Frame of the sequence from GenScan.

DNA Sequence Protein sequence (largest ORF)


Sequence name	contig_31
Download	FASTA file 
Length (AA)	1010 Amino Acids
Sequence	<pre> NLDGQVCEICGDEVGLTVDGDLFVACNECGFPVCRPCYEYERREGSQLCPQCKTRYKRLKGSPRVGGDDDEEDIDDLHEHEFNIEDEQNKHKYMAEAMLHGKMSYGRGPED DDNAQFPFSVIAGGRSRPVSGEPFISSYGHGEMPSSLHKRVHPYPISEPGSERWDEKKGWERMDDWKLQGGNLGPEPDDINDPDMAMIDEARQLSRKVP IASSKINP YRMVIVARLAILAFFLRYRILNPHDAFGLWLTSLICEIWFAPSWLDQFPKWFPI DRETYLDRLSLRYEREGEPNMLSPVDVVFVSTVDPKPEPLVTGNTVLSILAMDY PVDKISCVSDDGASMLTFESLSEAEFARKWVFPCKKFSIEPRAPEMYFTLKDIDYLDKDVQPTFKERRAMKREYEEFKVRINALVAKAAKVPPEGWIMQDGTWPQGN TKDHPGMIQVFLGHSGGLDADGNELPRLVYVSREKRPFGHKKAGAMNALVRVSGVLTNAPFMLNLDCHYINNSKAVREAMCFLMDPQIGRKCVCYVQFPQRFIDGIDAN DRYANRNTVFFDINMKGLDGIQGPVYVGTGCFVRRQALYGYEPPKPKRPMVSCDCCPCFGRKKLPKYSKHSANGDAADLQGMDDDKELLMSEMNEKFKGQSAIFVT STLMEQGGVPPSSPAALLKEAIIHVISCYEDKTEWGTGELGWYIGSITFEDILTGFKMHCRGWRISYCMKPRAPKGSAPINLSRLNQLRWALGSVEIFFSHHSPVWYG YKGGKWLWERFAYVNTTIYPFTSLPLLYCTLPAICLLTDFKIMPAISTFASLFFIALFMSIFATGILELRWGSVSEEWWRNQFVWIGVSAHLFAVVQGLLKVLAG IDTNFTVTSKASDDEDFGELYAFKWTLLIPPTLIIINLVGVVAGISDAINNGYQAWGPLFGKLFPAFVWILHLYPFLKGLMGRQNRTPPTIVVIVSVLLASIFSLWLWR IDPFVLKTKGPDTRKCGINC </pre>

Figure 4.3: Contig summary and sequence detail tab for contig\_31, the assembled cellulose synthase IRX3 gene (A). Download links for the cDNA and predicted protein sequence in FASTA format are provided (B).



(Figure 4.4B) contains a description of the GO category that the sequence was annotated with, and links to a graph based representation of the ontology term, as rendered by the AmiGo server (Figure 4.5). The gene ontology page (Figure 4.5A) contains a link to download all the contigs in that GO category as a FASTA file (Figure 4.5A) and a graphical representation of the GO term (Figure 4.5B).

If a KEGG annotation is available for a contig, a highlighted KEGG map is drawn by the KEGG server by sending a SOAP request to the server, and the image shown in the "Enzyme commission" tab. Each map has an enzyme highlighted in yellow, which corresponds to the enzymes associated with the contig (Figure 4.6). For every enzyme annotation (EC number) associated with the assembled contig, a pathway image is generated. The hyperlink to the EC commission table links to a short description of the enzyme in the pathway, and a FASTA file containing all the contigs annotated with the EC number (screenshot not shown). The InterProScan results tab (Figure 4.7) displays a line diagram of the predicted protein sequence, indicating the annotated protein features on the sequence. The tab also contains a table summary of the features found on the protein sequence, and links to the InterPro entry of the feature in the InterPro (<http://www.ebi.ac.uk/interpro/>) database.

Transcript expression for the contigs was calculated by the **Cufflinks** (Trapnell *et al.*, 2010) program (see Chapter 3 Section 3.2.7), and the expression values for each of the six sequenced tissues displayed in a table and as a bar graph (Figure 4.8). The bar graph is created by an R-script (**Rpy2 Python** package) that extracts the values from the database, and the created image displayed by the browser. The IRX3 *Eucalyptus* gene (contig\_31), is highly expressed in woody tissues (xylem and immature xylem), compared to green leaf tissues (shoot tips and young and mature leaf).

The 8X coverage version of the *Eucalyptus grandis* genome became publicly available (during August of 2010) and the assembled contigs were aligned to the first draft genome sequence in order to inspect contig contiguity and to view the *de novo* assembled contig together with public EST data on the draft genome sequence. The generic genome browser, **GBrowse** (version 2.26) was used to visualize the results from aligning the assembled contigs, as well as the Illumina short-reads to the genome sequence. The "GBrowse" tab available in **Eucspresso** renders the genomic position of the assembled contig on the

A)

**Eucspresso**  
The *Eucalyptus* gene expression database

Welcome Browse Advanced Search Contact FAQs

Summary Sequence detail Homology search results Gene Ontology Enzyme Commission InterProScan results Tissue-specific expression GBrowse

**Homology search results**  
Homology based annotation of the *Eucalyptus* transcripts was performed by performing a BLAST against the *Arabidopsis thaliana* (TAIR 9), *Vitis vinifera* (Jailon *et al.*, 2007) and *Populus trichocarpa* datasets.

The Hit Accession column links to the entry in GenBank.

Arabidopsis thaliana Vitis vinifera Populus trichocarpa

TAIR locus	Hit description	Alignment length	Query start	Query end	Subject start	Subject end	e-value	Bit score
TAIR:AT5G17420.1	IRX3 (IRREGULAR XYLEM 3); cellulose synthase   Symbols: IRX3, CESA7, ATCESA7, MUR10	1026	3	3077	16	1026	0.0	4778
TAIR:AT5G05170.1	CEV1 (CONSTITUTIVE EXPRESSION OF VSP 1); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA3, IXR1, ATCESA3, ATH-B, CEV1	1069	21	3077	5	1065	0.0	3738
TAIR:AT4G39350.1	CESA2 (CELLULOSE SYNTHASE A2); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA2, ATH-A, ATCESA2	1088	3	3077	16	1082	0.0	3611
TAIR:AT4G32410.1	CESA1 (CELLULOSE SYNTHASE 1); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA1, RSW1	1073	3	3032	16	1065	0.0	3586
TAIR:AT5G64740.1	CESA6 (CELLULOSE SYNTHASE 6); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA6, IXR2, E112, PRC1	1078	3	3077	16	1083	0.0	3556
TAIR:AT5G44030.1	CESA4 (CELLULOSE SYNTHASE A4); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA4, IRX5, NWS2	1065	48	3077	17	1049	0.0	3555
TAIR:AT2G21770.1	CESA9 (CELLULOSE SYNTHASE A9); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA9, CESA09	1086	3	3077	16	1086	0.0	3525
TAIR:AT5G09870.1	CESA5 (CELLULOSE SYNTHASE 5); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA5	1078	3	3077	16	1068	0.0	3523
TAIR:AT2G25540.1	CESA10 (CELLULOSE SYNTHASE 10); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA10	1048	36	3041	23	1055	0.0	3489
TAIR:AT4G18780.1	IRX1 (IRREGULAR XYLEM 1); cellulose synthase/ transferase, transferring glycosyl groups   Symbols: CESA8, IRX1, ATCESA8, LEW2	1021	63	3077	8	980	0.0	3348
TAIR:AT4G24000.1	ATCSLG2; cellulose synthase/ transferase/ transferase, transferring glycosyl groups   Symbols: ATCSLG2, CSLG2	384	705	1829	22	379	0.0	637

B)

**Eucspresso**  
The *Eucalyptus* gene expression database

Welcome Browse Advanced Search Contact FAQs

Summary Sequence detail Homology search results Gene Ontology Enzyme Commission InterProScan results Tissue-specific expression GBrowse

**Ontology**  
Gene ontology terms associated with contig\_31. These ontology terms are associated with the contig\_31 through the homology based annotation performed on the record. Blast2GO was used to transfer the annotations from the blast results to the contig.

The Gene Ontology Accession column links to the respective Gene Ontology entry, and will provide you with a list of all the contigs that was annotated in this ontology class.

Gene ontology Accession	Description	Ontology class
GO:0016760	cellulose synthase (UDP-forming) activity	Molecular Function
GO:0005515	protein binding	Molecular Function
GO:0030244	cellulose biosynthetic process	Biological Process
GO:0016021	integral to membrane	Cellular Component
GO:0008270	zinc ion binding	Molecular Function

Figure 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. The angiosperm sequence identifier links to entries in the TAIR and Phytosome databases (A). The molecular function ontology classes "cellulose synthase", "protein binding" and "zinc ion binding", the cellular component "integral to membrane" and the biological process "cellulose biosynthetic process" are associated with the contig (B).

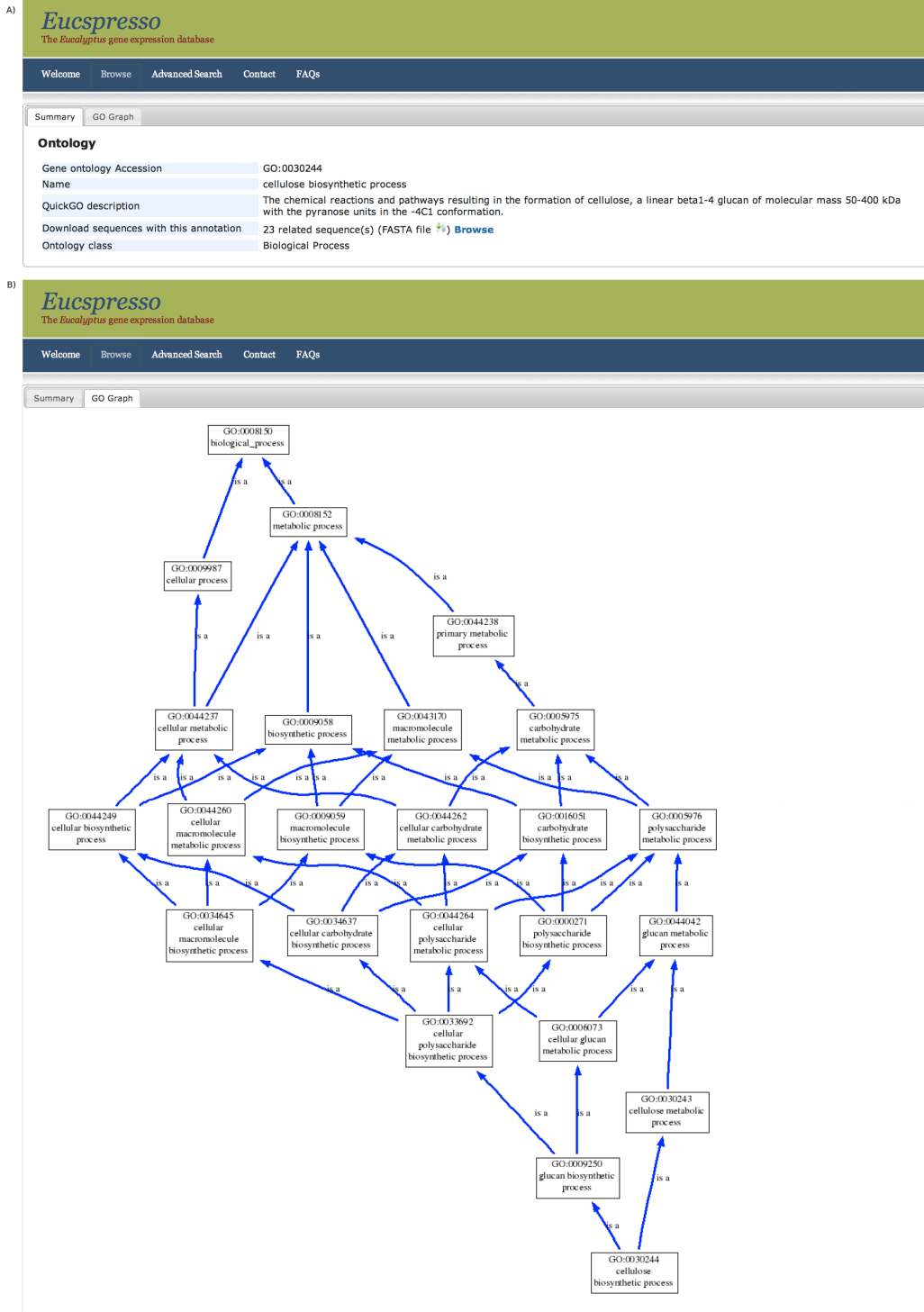


Figure 4.5: Gene ontology annotations for contig\_31, the assembled cellulose synthase IRX3 gene. A summary of the GO biological process category "cellulose biosynthetic process". A FASTA file containing the 23 FASTA sequences also annotated with the GO term (GO:0030244) is available as download (A). The GO graph of the GO term as rendered by the AmiGO web server is available in the "GO Graph" tab (B).

**Enzyme Commission**

Enzyme commission terms associated with the sequence.

Term	Description	Synonyms	Catalytic Activity	Cofactors	Comments
2.4.1.12	Cellulose synthase (UDP-forming).	UDP-glucose--beta-D-glucan glucosyltransferase. UDP-glucose-beta-D-glucan glucosyltransferase. UDP-glucose-cellulose glucosyltransferase.	UDP-glucose + (1,4-beta-D-glucosyl)(n) = UDP + (1,4-beta-D-glucosyl)(n+1).		Involved in the synthesis of cellulose.; A similar enzyme utilizes GDP-glucose (cf. EC 2.4.1.29).

**KEGG maps of the EC terms**

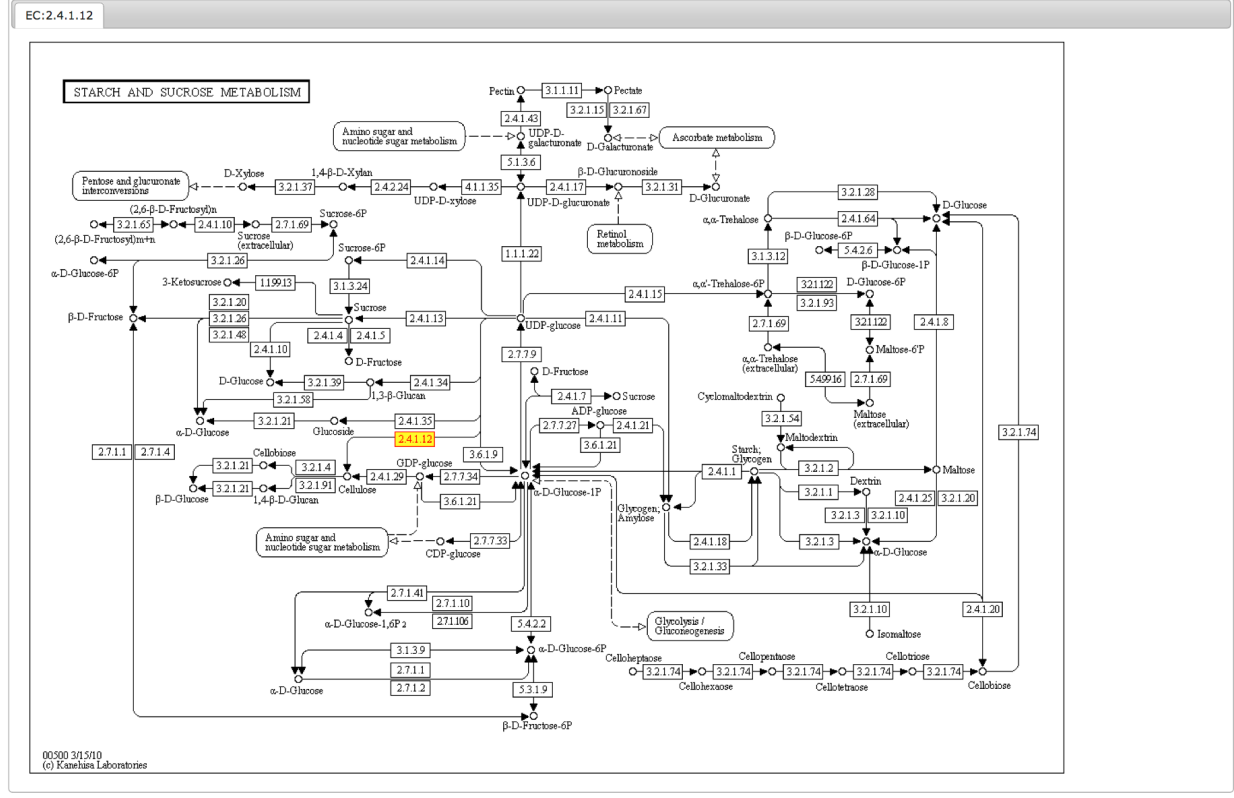


Figure 4.6: The cellulose synthase enzyme (EC:2.4.1.12) is highlighted on the starch and sucrose metabolism KEGG map.

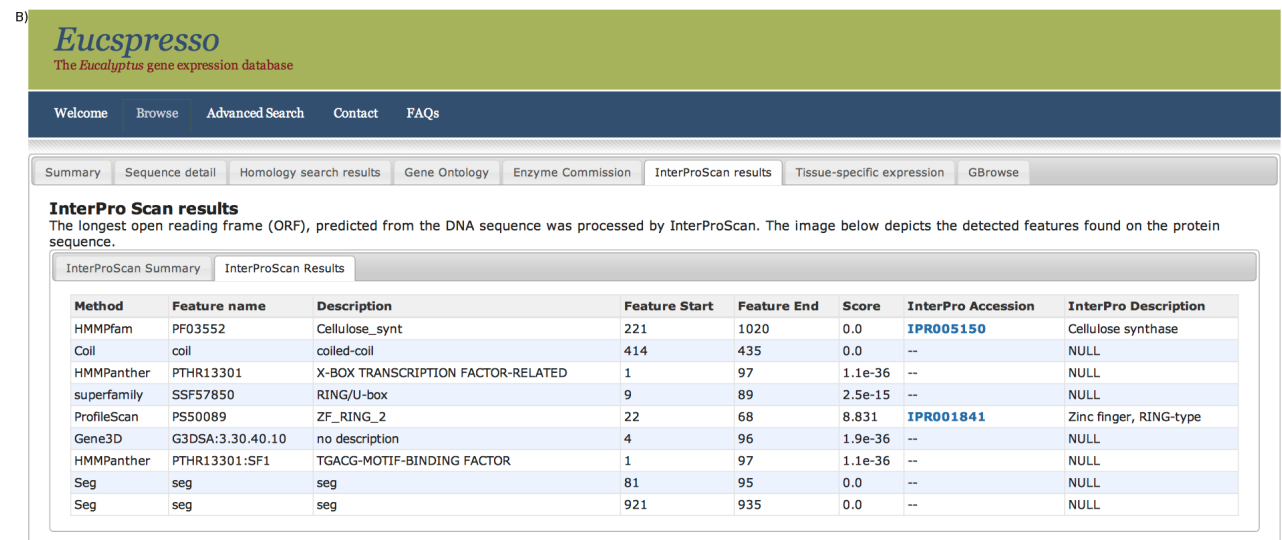
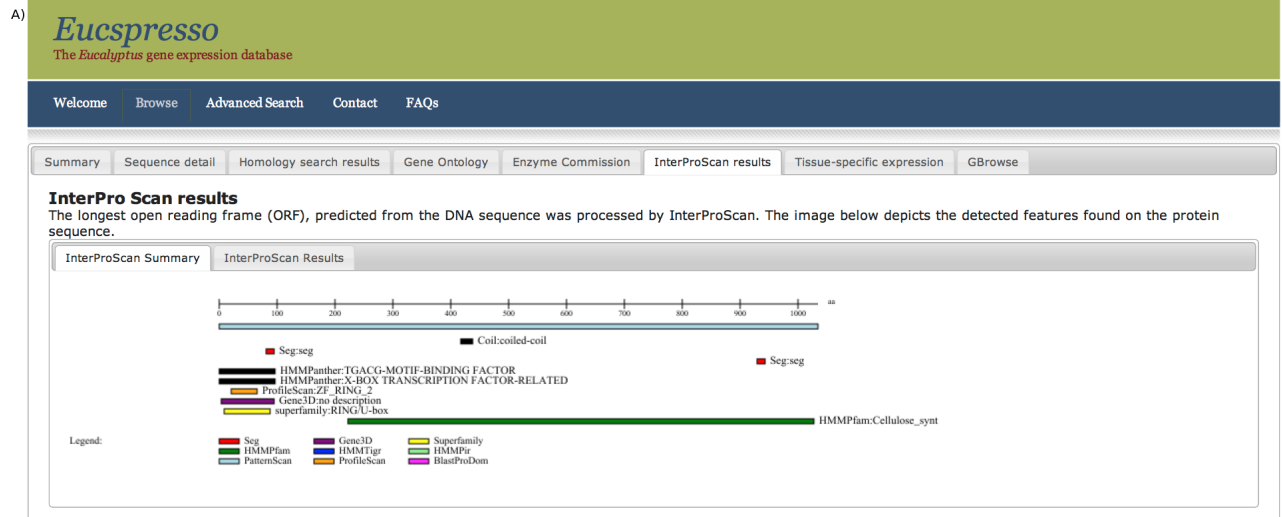


Figure 4.7: The InterProScan results tab describing protein features found on the predicted protein sequence from contig\_31. The contig contains the protein family domain for cellulose synthase (PF03552) and a zinc finger domain (PS50089) identified by the HMMPfam and ProfileScan tools (A and B). Some additional binding motifs were found close to the 5' of the sequence (A). Links to the InterPro entries of the cellulose synthase protein family and zinc finger domains are provided as blue text.

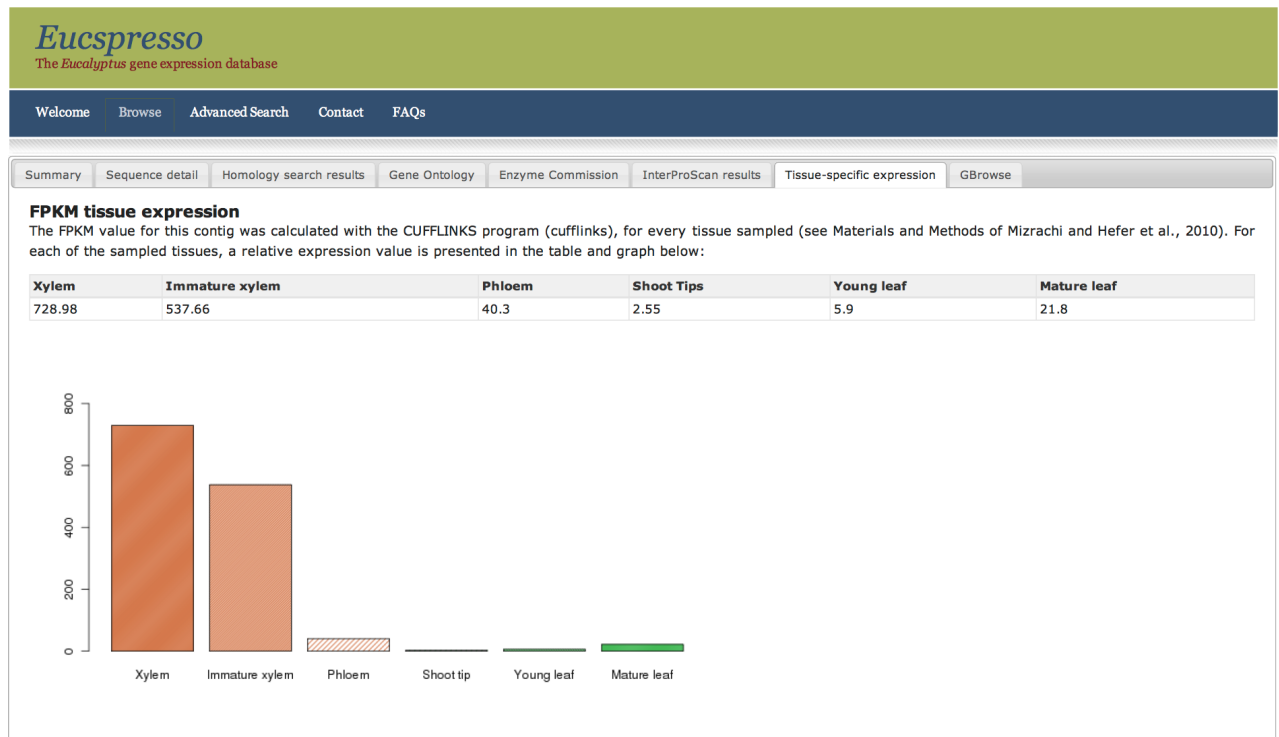


Figure 4.8: The FPKM expression values of contig\_31, a secondary cell wall synthesis gene (cellulose synthase, IRX3). The gene is highly expressed in woody tissues (FPKM value of 728.98 in xylem and 537.66 in immature xylem), and has a low expression value in leafy tissues (FPKM of 2.55 in shoot tips, 5.9 in young leaf, and 21.8 in mature leaf).

genome sequence. The user needs to manually request the **GBrowse** rendering option, since the rendering of the short-read track is time consuming. The short-reads can be visualised as a coverage plot, or individual reads aligned to the genome sequence.

#### 4.3.4. Search interface

In addition to the search interface available in the "Browse contig" interface (Figure 4.2), two additional search modules are available in **Eucspresso**. Under the "Advance Search tab", a keyword or accession number search can be used to filter the entries in the database. The "Keyword Search" tab offers the user the ability to construct complex queries using boolean search operators on a combination of datatypes and descriptors (Figure 4.10A). The search query interface is constructed as a set of predefined fields, or widgets (using **ToscaWidgets**), that dynamically constructs the SQL query with **SQLAlchemy**. The results of the search query are displayed in the same format as the "Browse and search" table discussed in Section 4.3.2.

The "Accession Search" tab allows for the opportunity to upload a combination of accessions, from the same datatype (GO accessions) or a mix of datatypes (GO, KEGG and InterPro accessions) and retrieve the contigs that were annotated with the terms (Figure 4.10B). A non-redundant set of sequences is returned to the user, and the results are again displayed in the "Browse and search" table format for further perusal of specific contigs.

## 4.4. Conclusion

The aim of the **Eucspresso** database (<http://euspresso.bi.up.ac.za>) was to serve as a central repository for the *de novo* assembled gene catalog described in Chapter 3. Although the resource currently contains data related to the specific *Eucalyptus* hybrid tree sequenced, it forms part of a bigger vision to build a genomic resource for *Eucalyptus* mRNA-Seq based expression data. Access to the **Eucspresso** data repository is provided through the web protocol as a easy to use interface to browse the contigs and annotations. The interface also provided several search interfaces to filter the data in such a matter



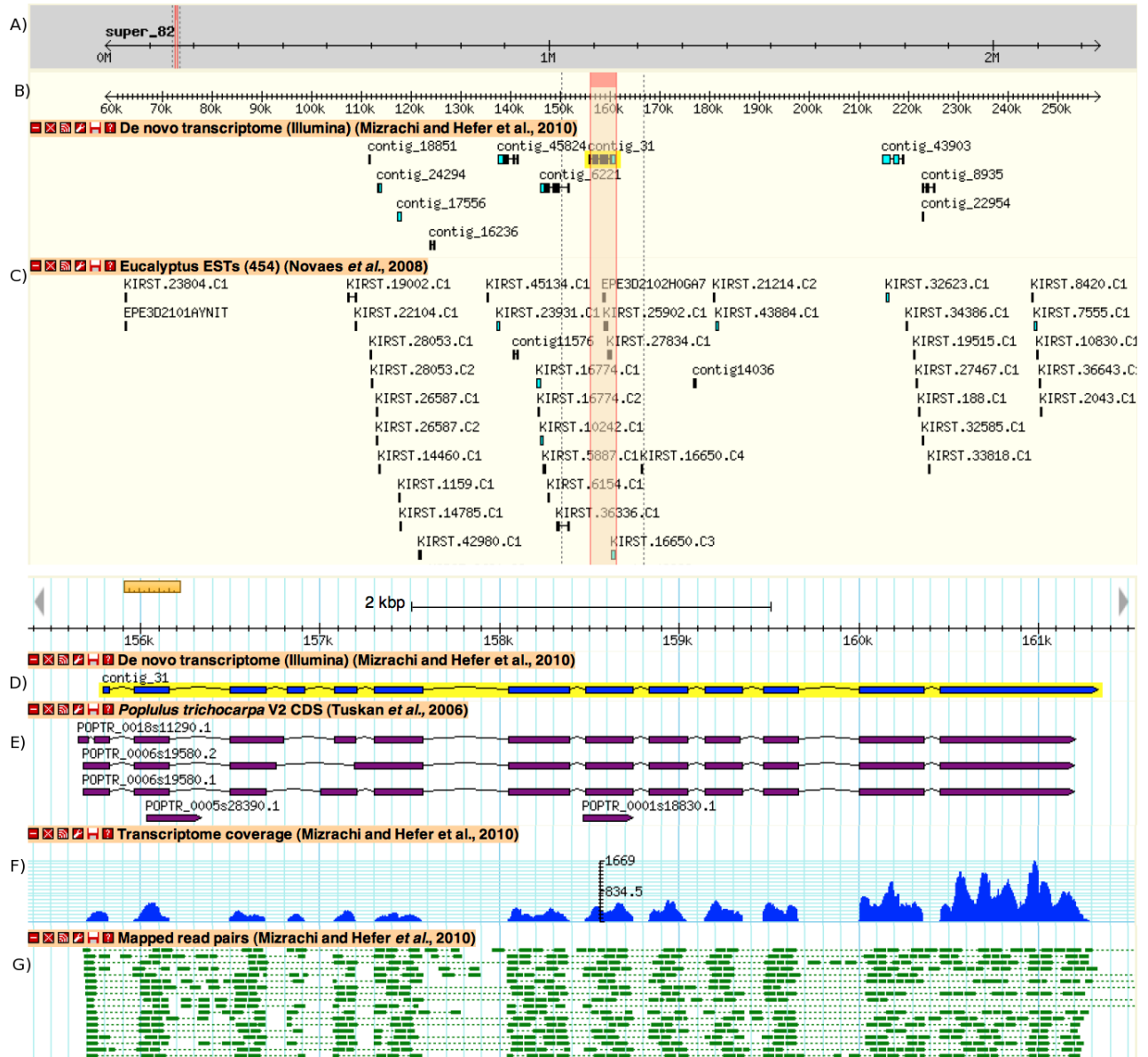
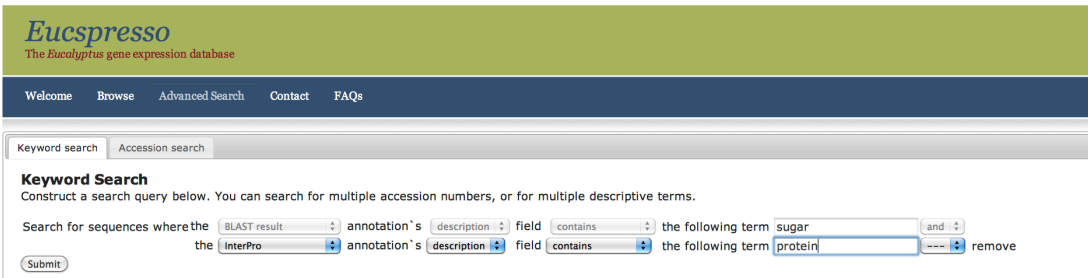


Figure 4.9: The Eucspresso GBrowse instance, indicating the position of contig\_31 (IRX3) on the 8X *Eucalyptus* draft sequence (scaffold 82, A). The assembled contig is shown in relation to other assembled contigs (B) and some 454 EST data (C) from Novaes *et al.* (2008). When focussing on the highlighted area, the complete transcript is shown (D) together with the closest *Populus* homolog that aligned to the same position in the genome (E). The coverage plot (F) represents the Illumina mRNA-Seq data aligned to the genome sequence, that was used to assemble the contig. The short-reads can be viewed when the user zooms in on the contig (G).

A)



B)

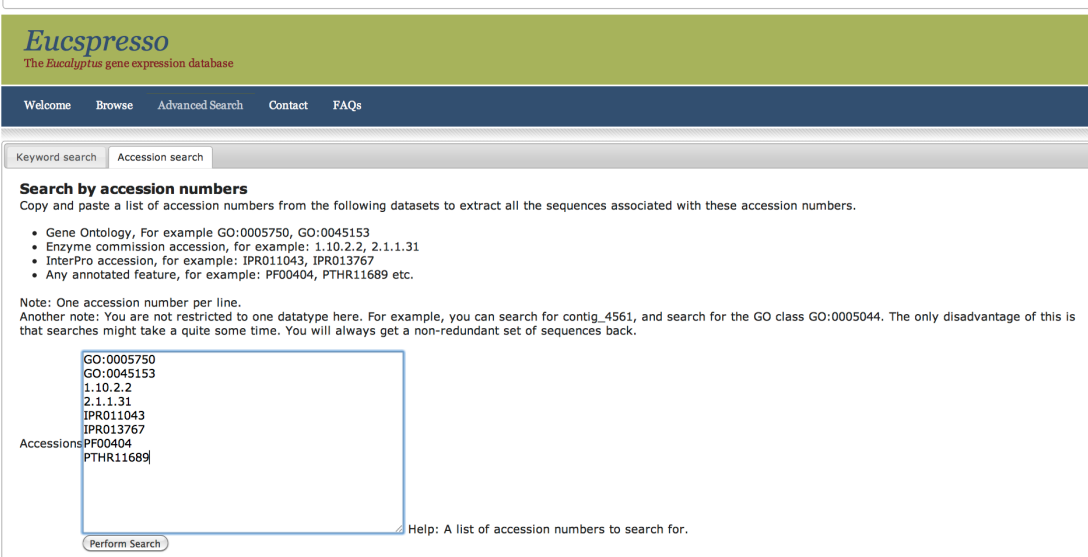


Figure 4.10: The Eucspresso search interface. Users can construct boolean searches based on accession IDs or keywords present in the EC, InterPro, GO and homology based annotations (A), as well as combine accession numbers from various datasets to retrieve non-redundant lists of contigs from the database (B).

as to focus on very specific subsets of the data. At any level of browsing, the specific contig or set of contigs of interest can be downloaded in FASTA format for further analysis in 3rd party applications.

Searches by common identifier, such as a specific GO category or KEGG identifier, can be used to explore very specific functional classes or metabolic pathways in terms of the sequences present in such a category. The genome browser interface provides additional confidence to the quality of the assembly process followed in Chapter 3, especially where EST data from Sanger sequence data or longer 454 reads are available to support the *de novo* assembled expressed transcripts.

The first version of annotation for the *Eucalyptus grandis* version 1.0 genome sequence was released early in 2011 (<http://www.phytozome.net>). The mRNA-Seq data used to assemble the transcriptome in this project is also available as an additional track in the **Phytozome** *Eucalyptus* genome browser (<http://www.phytozome.net>), and can be used to aid the identification of gene and exon boundaries of predicted gene models. The genome resource and predicted gene models available in **Phytosome** will be used to recalculate the FPKM values available in **Eucspresso**, and together with additional mRNA-seq experiments, including deep sequencing mRNA-Seq data of additional tissues, mRNA-Seq from disease challenged plants, and population based eQTL and mQTL data, a new resource is in the process of being developed. This new *Eucalyptus* resource (the *Eucalyptus* Genome Integrative Explorer, or **EucGenIE**), will focus primarily on the data from a multitude of mRNA-Seq experiments, and will complement genetic and genomic resources already available for woody plants.

Whole-transcriptome based expression experiments are fast becoming the standard to interrogate the transcriptional landscape of an organism. With more of these experiments being performed, a central repository can be envisioned where a multitude of experiments can be stored and combined to identify transcriptional networks. Similar resources are already publically available for microarray experiments (Manfield *et al.*, 2006; Obayashi *et al.*, 2007; Mutwil *et al.*, 2011), where data from several experiments can be combined to identify clusters of co-expressed genes. With the greater sensitivity of mRNA-Seq data to detect lowly expressed transcripts (Marioni *et al.*, 2008), algorithms and techniques developed for

microarray expression analysis can aid the elucidation of the transcriptional networks of the *Eucalyptus* forest tree.

## Chapter 5

# Concluding Discussion

Ultra-high-throughput DNA sequencing technologies have revolutionised the field of genomics. The advances made have led to the successful *de novo* sequencing of genomes (Tauch *et al.*, 2008; Reinhardt *et al.*, 2009; DiGuistini *et al.*, 2010; Nowrousian *et al.*, 2010; Li *et al.*, 2010*b*), large scale genome re-sequencing (Margulies *et al.*, 2005; Shendure *et al.*, 2005; Hofreuter *et al.*, 2006; McKernan *et al.*, 2009; Drmanac *et al.*, 2010; Pleasance *et al.*, 2010*a,b*), transcriptome profiling (Cloonan *et al.*, 2008; Denoeud *et al.*, 2008; Mortazavi *et al.*, 2008; Wilhelm and Landry, 2009; Wang *et al.*, 2010*b*), genome-wide DNA methylation mapping (Lister *et al.*, 2008; Hashimoto *et al.*, 2009; Flusberg *et al.*, 2010; Sun *et al.*, 2011) and protein-DNA interaction studies (Valouev *et al.*, 2008; Kuznetsov, 2009; Goren *et al.*, 2010). These studies lead us to formulate the hypothesis that a large proportion of the transcriptome of complex eukaryotes can be successfully *de novo* assembled, annotated and characterised using only mRNA-Seq data. The first objective of the study was to identify a suitable uHTS framework to store large sequence datasets, perform data analysis, and keep track of the results produced inside a web-based framework. Secondly, automated analysis workflows had to be developed to perform a set of pre-defined analysis on uHTS datasets, and, where needed, novel tools developed to complete the workflows. The *de novo* assembly of the transcriptome of a *Eucalyptus* hybrid tree was identified as a key validation of the developed hypothesis and tools, and the transcriptome was annotated and characterised without the aid of a genome sequence. The assembled transcriptome and annotations were then used to develop and populate a stand-alone transcriptome expression profiling database that forms part of a larger *Eucalyptus* genome

information resource (The Eucalyptus Genome Information Resource, **EucGenIE**), in anticipation of the release of annotated gene models from the *Eucalyptus* genome sequencing project (US Department of Energy and the Joint Genome Initiative, <http://www.phytozome.net>).

The **Galaxy** web framework (Goecks *et al.*, 2010) was identified as a suitable framework to store and manage large next-generation sequencing datasets, and also host the myriad of analysis tools available to perform analysis on uHTS data. The **Galaxy** framework provided the ability to connect input and output datasets of different analysis tools to create automated workflows. These workflows can then be shared between research groups and individuals. Widely-used ultra-high-throughput data analysis tools were incorporated into automated workflows, addressing tasks such as the quality evaluation of next-generation sequence data, *de novo* assembly of a transcriptome, mapping of short reads to a target genome and subsequent relative gene expression (FPKM) calculation, and the annotation of a set of assembled cDNA sequences. The design of these workflows led to the development of additional analysis tools and the extension of the **Galaxy** framework to include novel tools to perform the above-mentioned functions. All newly developed tools and wrappers have been incorporated in the local BCBU **Galaxy** server instance.

Critical evaluation of the developed workflow components identified several key parameters that influence the results from uHTS analysis tools. The **Velvet** (Zerbino and Birney, 2008) assembler was shown to be a reliable transcript assembler, assembling reliable, long, contiguous contigs. One critical shortfall of the assembler is that the assembly of alternative transcripts is not possible using **Velvet**, a problem that is being addressed by the development of the transcriptome specific assemblers **OASES** (Zerbino *et al.*, unpublished), **trans-ABYSS** (Birol *et al.*, 2009) and **Trinity** (Grabherr *et al.*, 2011). One of the key parameters to consider during the assembly, the expected coverage parameter, provided the most robust assembly when set high enough (a value of 1 000 was used in the final assembly) to allow for highly expressed transcripts. Another key parameter with great influence on the results obtained from the assembler, the kmer-value, needs to be independently verified for each transcriptome dataset, since it will vary with the complexity of the transcriptome and the length of the short reads sequenced. It

was also observed that paired-end reads from an Illumina sequenced cDNA library of larger than 50 bp did not significantly improve unique read mappability to a reference genome sequence as complex as the *Eucalyptus grandis* genome. The InterProScan (Zdobnov and Apweiler, 2001) and BLAST2GO (Conesa *et al.*, 2005) annotation pipelines were successfully incorporated in the BCBU Galaxy server, making high throughput annotation pipelines available in an easy to use web framework. For differential gene expression, the CUFFLINKS (Trapnell *et al.*, 2010) set of software tools, as well as the DEGseq R-package (Wang *et al.*, 2010a) provided various statistical approaches to model mRNA-Seq transcript sampling and identify differentially expressed genes in a sample dataset.

The workflows developed were used to perform a *de novo* assembly and annotation of the transcriptome of a *Eucalyptus grandis* x *Eucalyptus urophylla* hybrid tree from Illumina mRNA-Seq data. Six different tissues were sampled and a gene catalog consisting of 18 894 near full length transcripts were assembled. The assembled gene catalog was evaluated based on contig contiguity, contig diversity and similarity (BLAST) to other angiosperm transcriptome datasets. A novel transcriptome assembly approach was developed, where an assembled contig was used in a coverage-directed re-assembly approach in an attempt to extend the contig sequences. Although the assembly approach followed did not allow for the assembly of alternative transcripts, the set of transcripts assembled were shown to contain contiguous, near full-length biologically relevant molecules. The assembled transcriptome was annotated with Gene Ontology, KEGG and various InterProScan-related terms, identifying a range of assembled transcripts present in the assembly. The Illumina short-read data was then used to identify a set of transcripts over-expressed in xylogenic *vs.* leafy tissues (and *vice versa*). The study showed that current bioinformatics software tools and approaches can be used to assemble and characterise a large proportion of the transcriptome of a complex eukaryotic organism. This approach can be used to successfully characterise the gene catalog of a wide range of organisms using only data derived from uHTS experiments.

A Python based web framework (TurboGears) was used to develop a user-friendly, intuitive web interface to browse and interact with the assembled and annotated *Eucalyptus* hybrid gene catalog.



A MySQL database stored the relations between the assembled contigs and the functional annotations associated with each of the transcripts. The SQLAlchemy object relational mapper was implemented to perform queries on the relational database, and also provided the ability to construct *ad hoc* queries via the advanced search interface. The resource, Eucspresso, was developed with the aim to serve as a transcriptome expression module for a larger framework, EucGenIE, that will cater for the storage and analysis of data of a wide range of mRNA-Seq based whole-transcriptome experiments. The availability of such a range of whole transcriptome expression datasets will in future aid the discovery of transcriptional regulation networks, gene co-expression clusters and regulatory elements and will complement existing databases for forest research (PopGenIE, Sjödin *et al.*, 2009).

In conclusion, it was shown that by making use of deep Illumina mRNA-seq data, it is possible to assemble and characterise a gene catalog of a complex eukaryote without the use of any genomic information. Analysis tools and workflows were developed to address different steps in the assembly and annotation process, and these workflows implemented in a web-based framework. The study produced the most complete *de novo* assembled gene catalog to date for a forest tree from uHTS data (longer, more complete contigs than what was possible by a similar study using 454 data by Novaes *et al.*, 2008). The study was one of the first to make use of Illumina mRNA-Seq data to characterise the transcriptome of a large eukaryote, and a similar approach was followed with the characterisation of the Chickpea transcriptome (Garg *et al.*, 2011). Velvet and OASES, as well as trans-ABYSS were evaluated during the Chickpea transcriptome assembly, and it was found that OASES performed slightly better than Velvet when evaluating assemblies based on the N50 and mean transcript lengths. The findings from the Chickpea study supports the decision to make use of a *de Bruijn* graph assembler such as Velvet for *de novo* transcriptome assemblies, but also illustrates the rapid improvement of assembly algorithms with the finding that OASES performed better on the Chickpea dataset. When considering future *de novo* transcriptome assembly projects, the advances made in the algorithms for assembly needs to be carefully considered and several assemblers evaluated before selecting the best assembly. Improvements to the read

length of Illumina mRNA-Seq data and the algorithms used for *de novo* transcriptome will soon result in transcriptome profiling of species with very little or no genomic resources becoming commonplace.

The study also resulted in a bioinformatics workflow environment in which uHTS data can be used for transcriptome assembly, transcript annotation and transcript expression profiling. The developed **Eucspresso** transcriptome resource provided early access to the transcriptome landscape of *Eucalyptus*, and provided users with the gene expression profiles of six different sequenced tissues in a *Eucalyptus grandis* x *Eucalyptus urophylla* hybrid tree. The Illumina short-read data was made available to the EUCAGEN (<http://eucagen.org>) consortium to aid the annotation of the recently sequenced *Eucalyptus grandis* genome, and the short-reads are available as a separate track on the current (Version 7.0) release of Phytozome. Future work that directly follows from the findings in this study includes the development of a *Eucalyptus* genome integrative explorer (**EucGenIE**), that will serve as a primary repository for several re-sequenced genome sequences, as well as transcriptome datasets from several individuals used in a *Eucalyptus* genome mapping population, and several disease specific transcriptome datasets.

With the availability of the complete set of gene models predicted from the *Eucalyptus grandis* genome sequence, the use of *Eucalyptus* mRNA-Seq experimental data will move towards identifying alternative transcript spliceforms, alternative transcriptional start sites, and identify differential gene expression within tissues and under different environmental conditions. Whole-genome transcriptional profiles, when used in conjunction with population wide quantitative trait (Quantitative Trait Loci, QTL) association data, can lead to the identification of clusters of co-expressed genes associated with specific traits (Brem and Kruglyak, 2005). The availability of these genome wide, and population wide datasets will allow for future studies that test directly for the effect of allele specific expression in heterozygotes. For example, where heterozygous loci are present in a population, and the two copies of the transcript are present at different levels between individuals, the effect can possibly be ascribed to the effect of cis-acting regulatory elements that affect gene expression (Wittkopp *et al.*, 2008; Gilad *et al.*, 2009). The combination of genome-wide genomic and transcriptomic datasets and population genetic information

provides researchers with a powerful approach to identify the system-wide phenotypic effect of small molecular changes on the genome, a new field of study that can be considered genetical genomics.

## Appendix A

### Bioinformatics workflow

Table A.1: *Velvet* assembly statistics of contigs longer than 1 000 bp for a single lane of paired 76 bp sequences from *Eucalyptus* xylem tissue reads trimmed to different lengths. The assemblies were all performed with a kmer setting of 41. These statistics were used to calculate the assembly score, as discussed in Section 2.3.3 on page 56 and presented in Table 2.6.

Read length	N	Sum	Min	1st Quartile	Median	3rd Quartile	Max	Mean	N50
50	2 644	3 853 938	1 000	1 118	1 300	1 611.5	6 772	1 457.61	1 424
55	5 045	7 722 735	1 000	1 138	1 342	1 709	8 078	1 530.77	1 512
60	6 458	10 216 572	1 000	1 149	1 371	1 770	8 241	1 582.00	1 574
65	7 165	11 547 759	1 000	1 160.5	1 393	1 804	11 049	1 611.69	1 609
70	7 548	12 288 379	1 000	1 162	1 395	1 823	11 008	1 628.03	1 627
76	7 857	12 917 451	1 000	1 164	1 415	1 848	9 925	1 644.06	1 643

## Appendix B

# Extindinator

The `Python` script used for a coverage-assisted re-assembly of contigs, also known as "extindinator" is provided on the following pages. A graphical representation of the process is provided in Figure 3.1. The program selects an entry from the assembled contigs file, and performs an alignment of the short reads to the selected contig and calculated the true coverage of the contig. After alignment, the program extracts all the short reads together with their respective mate-pairs from a `Berkeley` database, and sends the contig as well as the sampled short reads to `Velvet` with the calculated coverage parameter to perform a directed contig assembly.



```
"""
Extindinator:
    An interative approach to try and improve contig sizes.

1) Map all the short reads to a contig, get the reads that mapped.
2) Extract the pairs
    2a) Connect to a database, get all the reads that match
    2ab) Convert to fasta
3) Assemble with Velvet
    3a) Parameter range cc_9 ec [9,50,100,200,1000]
    3b)Join the longest assemblies in one file (best_assembly.fa)

@requires: Biopython
@requires: bsddb3
@author: charles.hefer@gmail.com
"""

import sys
import getopt
from datetime import datetime
from Bio import SeqIO
import os
import subprocess
import time
from multiprocessing import Process
import bsddb3

global usage
usage = """
Extindinator: An iterative approach to extext Velvet contigs

Usage: python start_extindinator.py [options] short_reads.fa contigs.fa

++Bowtie options++
\t-f\t--short_reads_type\Either fa for fasta, or fq for fastq, default is fa
\t-b\t--bowtie_mismatch\tNumber of mismatches allowed during the bowtie matching of the short reads to the contig
\t-m\t--max_bowtie_processes\tMax number of bowtie processes
\t-t\t--threads\tNumber of threads for Bowtie, this times the #processes = number of CPUs

++Global options++
\t-h\t\t--help\tThis help message
"""

global cwd
cwd = os.getcwd()
global bowtie_build_cmd
bowtie_build_cmd = "/usr/local/bowtie/bowtie-build"
global bowtie_cmd
bowtie_cmd = "/usr/local/bowtie/bowtie"
global bdb
bdb = "./pairs.db"

class UsageEx(Exception):
    """The standard exception"""
    def __init__(self, msg):
        """
        Sets the exception message
        @var msg: The exception message thrown
        """
        self.msg = msg

def now():
    """
    Converts the current time to a string format

    @requires: datetime.datetime
    @return: A string reprepresentation of datetime.now()
    """
    curr_time = datetime.now()
    return curr_time.strftime("%c")

def get_number_of_processes(process):
    """Returns the number of processes returned by grep
    ps -eaf | grep processname
    Subtract the grep itself, and the extra newline that comes through.
    @var process: The process to grep for
    @type process: String

    @return: The number of process as an int
    """
    num_procs = subprocess.Popen("ps -eaf | grep '%s'" % process, shell=True, stdout=subprocess.PIPE)
    output = num_procs.stdout.readlines()
    i = len(output) - 2
    return i
```



```
def multiprocess_start(cmd):
    """
    Executes the command as a multiprocess
    """
    process = subprocess.call(cmd, shell=True, stdout=subprocess.PIPE)
    return process

def prepare_bowtie_build(dir, filename, max_bowtie_processes):
    """
    Sets the command to run bowtie build on the contig
    """
    #the resulting build has a _ewbt extension
    #and is in the ./bowtie dir
    cmd = "%s %s %s_ewbt" % (bowtie_build_cmd, dir+filename, "./bowtie/"+filename)

    while get_number_of_processes("bowtie_build") >= max_bowtie_processes:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()

def prepare_bowtie_align(short_reads_filename, ewbt_filename, bowtie_mismatch, max_bowtie_processes, threads, short_reads_filetype):
    """
    Aligns the short reads to the file
    """
    cmd = "%s -%s -n %s --alfa=%s.match -p %s %s %s %s.out" % (bowtie_cmd,
                                                             short_reads_filetype,
                                                             bowtie_mismatch,
                                                             "bowtie/"+ewbt_filename,
                                                             threads,
                                                             "bowtie/"+ewbt_filename,
                                                             short_reads_filename,
                                                             "bowtie/"+ewbt_filename)

    while get_number_of_processes("bowtie") >= max_bowtie_processes:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()

def save_biopython_entry(dir, entry, format):
    """
    Saves the biopython object in the correct format
    """
    try:
        handle = open(dir+"/"+ entry.name + "/" + entry.name+".fa", "w")
    except IOError, e:
        print(e)
        sys.exit()
    SeqIO.write([entry], handle, format)
    handle.close()
    return dir+"/"+entry.name + "/" + entry.name + ".fa"

def bowtie_watcher(contig, max_bowtie_processes, bowtie_mismatch, short_reads_filename):
    """
    Somehow manages the number of bowtie executables that can be started
    """
    #Get the current number of bowties running
    current = get_number_of_processes("bowtie")
    while current > max_bowtie_processes:
        time.sleep(10)
    else:
        bowtie_dir = prepare_bowtie_dir(contig.name)
        contig_file_name = save_biopython_entry(bowtie_dir, contig, "fasta")
        bowtie_builder(contig_file_name)
        bowtie_aligner(contig_file_name, bowtie_mismatch, short_reads_filename)

def split_fasta_file(handle, dir):
    """
    Takes every entry, create an output file for that entry in the dir
    """
    entries = SeqIO.parse(handle, "fasta")
    for entry in entries:
        out = open(dir+entry.name.replace(" ", "").replace("\\", "").replace("|", "_").replace("/", "_").replace("(", "_").replace(")", "_"), "w")
        SeqIO.write([entry], out, "fasta")
        out.close()

def create_mates_file(base_name, database_name):
    """ Iterates over ./bowtie/base_name.match, and returns all the mated
    that is found in the berkeley database
    Creates a file basenome.fa in ./mates
    """

    try:
        handle = open("./bowtie/%s.fa_ewbt.match" % (base_name), "r")
        out_handle = open("./mates/%s.fa" % (base_name), "w")
    except IOError, e:
        #No alignments found... can do nothing about that
        #should this be reported?
```





```
        return None
    entries = SeqIO.parse(handle, "fasta")

    mate_pairs = []
    pairs = bsddb3.hashopen(bdb, "r")
    for entry in entries:
        out_handle.write(">%s\n" % entry.name)
        out_handle.write("%s\n" % pairs[entry.name].split(",")[0])
        out_handle.write(">%s\n" % entry.name)
        out_handle.write("%s\n" % pairs[entry.name].split(",")[1])
    out_handle.close()

def faLen_stats(file):
    """
    Returns the result from running faLen on the file
    #TODO: Rewrite use subprocess
    """
    import popen2

    output = []

    cmd = "faLen < %s | stats" % (file)
    process = popen2.Popen3(cmd)
    process.wait()
    result = process.fromchild.readlines()
    for line in result:
        line = line.replace(" ", "")
        output.append(line.split("=")[1].rstrip())
    output.append("\n")
    return output

def velveth_runner(filename, kmer):
    """
    Runs velveth on the file, hashing for the kmer
    """
    velvet_exe = "/usr/local/velvet/velveth"
    cmd = "%s ./velvet/%s/assembly %s -fasta -shortPaired ./mates/%s -long ./fasta/%s" % \
        (velvet_exe, filename, kmer, filename, filename)

    while get_number_of_processes("velveth") >= 20:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()
    time.sleep(2)

def get_coverage(filename):
    """
    Returns the coverage value stored in ./mates/cov_stats.csv
    """
    file = open("./mates/cov_stats.csv", "r")
    for line in file:
        if filename in line:
            cols = line.split(",")
            contig_length = int(cols[1])
            bases = int(cols[2].rstrip())
    return bases/float(contig_length)

def velvetg_runner(filename):
    """
    Runs velvetg in the file, hashing for the kmer
    """
    velvet_exe = "/usr/local/velvet/velvetg"
    coverage = get_coverage(filename)

    cmd = "%s ./velvet/%s/assembly -ins_length 200 -ins_length_sd 80 -exp_cov %s -cov_cutoff 8" % \
        (velvet_exe, filename, coverage)

    print cmd

    while get_number_of_processes("velveth") >= 20:
        time.sleep(5)
    process = Process(target=multiprocess_start, args=(cmd,))
    process.start()

def save_longest_entry(entry_name, contigs_file, location):
    """
    Finds the longest entry in the contigs_file, rename it to the
    entry name [minus the extension], and saves it in the locatoion
    """
    try:
        contigs_handle = open(contigs_file, "r")
        location_handle = open(location+"/%s" % entry_name, "w")
    except IOError, e:
        print(e)

    longest_entry = None
    longest_length = 0
```



```
entries = SeqIO.parse(contigs_handle, "fasta")
for entry in entries:
    if len(entry.seq) > longest_entry:
        longest_entry = entry
        longest_length = len(entry.seq)

#Rename the longest_entry
longest_entry.id = entry_name.replace(".fa","")
longest_entry.name = ""
longest_entry.description = ""
#write to the location
SeqIO.write([longest_entry], location_handle, "fasta")
location_handle.close()
contigs_handle.close()

#update the report
#remove the entries that did not grow for fasta
#repeat = True

def main(argv = None):
    """
    The main program flow
    """
    print("%s Extindinator started" % now())

    #Get all the arguments
    if argv is None:
        argv = sys.argv
    try:
        try:
            opts, args = getopt.getopt(argv[1:], "b:h:m:t:f:",
                                       ["bowtie_mismatch=",
                                        "max_bowtie_processes=",
                                        "threads=",
                                        "short_reads_type"
                                        "help"])

            bowtie_mismatch = 2
            max_bowtie_processes = 1
            threads = "2"
            short_reads_filetype = "f"

            for opt, value in opts:
                if opt in ("b", "--bowtie_mismatch"):
                    bowtie_mismatch = value
                if opt in ("m", "--max_bowtie_processes"):
                    max_bowtie_processes = int(value)
                if opt in ("t", "--threads"):
                    threads = value
                if opt in ("f", "--short_reads_type"):
                    if value == "fq":
                        short_reads_filetype = "q"
                if opt in ("h", "--help"):
                    print(usage)
                    raise sys.exit()

            except getopt.error, e:
                print(e)
                raise UsageEx(e)

            #test the presence of the contigs and short read files
            try:
                print("%s Validating the short reads file: %s" % (now(), args[0]))
                short_reads_filename = cwd+"/"+args[0]
                short_reads_handle = open(args[0], "r")
                print("%s Validating the contigs file: %s" % (now(), args[1]))
                contigs_handle = open(args[1], "r")
            except IOError,e:
                print(e)
                raise UsageEx(e)

            #Prepare the directory structure
            #this can be made more intelligent
            try:
                os.system("rm -rf bowtie")
                os.system("rm -rf fasta")
                os.system("rm -rf mates")
                os.system("rm -rf velvet")
            except OSError:
                pass
            try:
                os.mkdir("bowtie")
                os.mkdir("fasta")
                os.mkdir("mates")
                os.mkdir("velvet")
            except OSError, e:
                os.system("rm -rf bowtie/*")
```



```
os.system("rm -rf fasta/*")
os.system("rm -rf mates/*")
os.system("rm -rf velvet/*")

try:
    report_handle = open("report.csv", "w")
except IOError, e:
    print(e)
    sys.exit()

#The step is to parse the contigs file
print("%s Parsing the contigs file into ./fasta" % now())
split_fasta_file(contigs_handle, "./fasta/")
contigs_handle.close()

fasta_entries = os.listdir("./fasta")

#generate an file with the initial lengths
print("%s Generate the initial report template" % now())
report_handle.write("Sequence_entry,init_length\n")
for fasta_file in fasta_entries:
    #get the sequence length
    entry_length = int(falLen_stats("./fasta/%s" % fasta_file)[1])
    report_handle.write("%s,%s\n" % (fasta_file, entry_length))
report_handle.close()

while 1:
    fasta_entries = os.listdir("./fasta")

    if len(fasta_entries) == 0:
        break

    print("%s Building the Bowtie indices" % now())
    for fasta_entry in fasta_entries:
        prepare_bowtie_build("./fasta/", fasta_entry, max_bowtie_processes)
    time.sleep(2)
    #Need to wait for all the processes to finish
    while get_number_of_processes("bowtie_build") > 0:
        time.sleep(5)

    print("%s Running Bowtie aligner with %s mismatches" % (now(), bowtie_mismatch))
    print("Stdout from Bowtie to follow...this can be ignored")
    for fasta_entry in fasta_entries:
        prepare_bowtie_align(short_reads_filename, fasta_entry+"_ewbt", bowtie_mismatch, max_bowtie_processes, threads, short_
        #give the os time to register
        time.sleep(2)
    #Need to wait for all the processes to finish
    time.sleep(5)
    while get_number_of_processes("bowtie") > 0:
        time.sleep(5)
    print("%s Done with the Bowtie aligner" % (now()))

    print("%s Preparing to find the mates" % (now()))
    for fasta_entry in fasta_entries:
        #change the name
        fasta_entry = ".".join(fasta_entry.split(".")[::-1])
        create_mates_file(fasta_entry, bdb)
    time.sleep(5)
    print("%s Mates now in ./mates" % now())

    print("%s Calculating the coverage statistics" % now())
    mate_entries = os.listdir("mates")
    try:
        mate_entries.remove("cov_stats.csv")
    except:
        pass
    cov_stats_handle = open("mates/cov_stats.csv", "w")
    cov_stats_handle.write("Contig_name,Lenght,Bases_in_mates")
    cov_stats_handle.write("\n")
    for mate_entry in mate_entries:
        contig_length = int(falLen_stats("./fasta/%s" % mate_entry)[1])
        pairs_bases = int(falLen_stats("./mates/%s" % mate_entry)[1])
        cov_stats_handle.write("%s,%s,%s" % (mate_entry, contig_length, pairs_bases))
        cov_stats_handle.write("\n")
    cov_stats_handle.close()
    time.sleep(5)
    print("%s Finished with the coverage statistics, in ./mates/cov_stats.csv" % now())

    print("%s Preparing for the velvet hashing " % now())
    for entry in mate_entries:
        try:
            os.mkdir("./velvet/%s" % entry)
        except OSError, e:
            pass
        velveth_runner(entry, "31")
    time.sleep(5)
    while get_number_of_processes("velveth") > 0:
        time.sleep(5)
```



```
print("%s Done with the velvet hashing" % now())

print("%s Preparing for the velvet assembly " % now())
for entry in mate_entries:
    velvetg_runner(entry)
time.sleep(5)
while get_number_of_processes("velveth") > 0:
    time.sleep(5)
print("%s Done with the velvet assembly" % now())

print("%s Getting the longest entry for every assembly" % now())
for entry in mate_entries:
    #the contigs resides in velvet/entry/assembly/contigs.fa
    save_longest_entry(entry, "velvet/%s/assembly/contigs.fa" % entry, "fasta/")
print("%s All the longest entries now back in ./fasta" % now())

print("%s Adding the newest data to the report.csv file" % now())
#Append to the reports file
os.system("mv ./report.csv ./report.csv.prev")
reports_handle = open("report.csv.prev", "r")
report_out_handle = open("report.csv", "w")
report_out_handle.write("Sequence_entry,init_length\n")
for line in reports_handle:
    if line.startswith("Sequence_entry"):
        continue
    line = line.rstrip()
    cols = line.split(",")
    #the name of the entry is the first col
    try:
        entry_length = int(faLen_stats("./fasta/%s" % cols[0])[1])
        cols.append("%i" % entry_length)
    except IndexError, e:
        pass
    outline = ",".join(cols)
    report_out_handle.write(outline + "\n")
report_out_handle.close()
reports_handle.close()
print("%s Updated the report.csv file" % now())

#Now, check the report file, if the last entry is smaller or equal to
#the second last entry, then call the entry finished
#remove from ./fasta/
#and append to finished_contigs.fa
print("%s remove the contigs that does not want to grow any more" % now())
report_handle = open("report.csv", "r")
for line in report_handle:
    if line.startswith("Sequence_entry"):
        continue
    print line
    line = line.rstrip()
    cols = line.split(",")
    if int(cols[-1]) <= int(cols[-2]):
        print cols[0]
        os.system("less ./fasta/%s >> finished_contigs.fa" % cols[0])
        os.system("rm ./fasta/%s" % cols[0])
        os.system("rm ./mates/%s" % cols[0])
        print os.listdir("fasta")
        print os.listdir("mates")

    print("%s And start over again?" % now())
print("%s Done" % now())

except UsageEx, err:
    print(usage)

if __name__ == "__main__":
    if len(sys.argv) < 3:
        print(usage)
        sys.exit()
    else:
        sys.exit(main())
```

## Appendix C

# Transcriptome assembly

## C.1. Evaluating contig contiguity of the assembled transcript sequences

### C.1.1. Full length *Eucalyptus* cDNA sequences

The following table contains the 34 full length CDS sequences used to validate the assembly. The functional role of the 33 sequences ranges from transcription factors, transporter genes, structural and developmental proteins, indicating that the assembled transcriptome successfully assembled near full length genes, including the 5' and 3' UTR regions for a wide variate of mRNA sequences.

Accession	Contig_id	Description	length	FPKM
AB465730.1	contig_87094	Eucalyptus grandis AGL mRNA for agamous-like protein, complete cds.	1184	17.98
AB479542.1	contig_10798	Eucalyptus grandis mRNA for transcription factor Myb, complete cds.	666	14.02
AB479543.1	contig_45922	Eucalyptus grandis mRNA for transcription factor GRAS family protein, complete cds.	1485	13.00
AB479544.1	contig_94920	Eucalyptus grandis mRNA for 1-aminoacyclopropane-1-carboxylate oxidase, complete cds.	1288	81.75

AB479545.1	contig_56935	Eucalyptus grandis mRNA for transcription factor squamosa promoter binding protein like, complete cds	1940	43.35
AF029976.1	contig_93436	Eucalyptus grandis MADS box protein (EGM2) mRNA, complete cds.	920	13.01
AF197329.1	contig_5550	Eucalyptus grandis zinc transporter (EgZnT1) mRNA, complete cds.	1635	17.08
AF197330.1	contig_2649	Eucalyptus grandis calcineurin-like protein (EgCBL1) mRNA, complete cds.	951	27.21
AY150283.1	contig_11286	Eucalyptus grandis fertilization independent endosperm development protein mRNA, complete cds	1626	18.87
AY263807.1	contig_68957	Eucalyptus grandis SOC1-like floral activator MADS3 mRNA, complete cds.	1112	21.66
AY263808.1	contig_52396	Eucalyptus grandis SOC1-like floral activator MADS4 mRNA, complete cds.	980	8.80
AY263809.1	contig_6043	Eucalyptus grandis SVP-like floral repressor mRNA, complete cds.	855	20.09
DQ014506.1	contig_2805	Eucalyptus grandis cellulose synthase 2 (CesA2) mRNA, complete cds.	3471	226.37
DQ014507.1	contig_31	Eucalyptus grandis cellulose synthase 3 (CesA3) mRNA, complete cds.	3452	220.59
DQ014509.1	contig_4202	Eucalyptus grandis cellulose synthase 5 (CesA5) mRNA, complete cds.	3712	137.25
DQ014510.1	contig_19509	Eucalyptus grandis cellulose synthase 6 (CesA6) mRNA, complete cds.	3782	97.32
DQ227992.1	contig_6857	Eucalyptus grandis thioredoxin h mRNA, complete cds.	354	133.93

DQ227993.1	contig_69050	Eucalyptus grandis sucrose synthase (SuSy1) mRNA, complete cds.	2498	250.38
DQ227994.1	contig_40644	Eucalyptus grandis sucrose synthase (SuSy3) mRNA, complete cds.	2508	220.28
EF179384.1	contig_24067	Eucalyptus grandis UDP-glucose dehydrogenase (UGDH) mRNA, complete cds.	1443	812.03
EF534216.1	contig_319	Eucalyptus grandis fasciclin-like arabinogalactan protein (FLA1) mRNA, complete cds.	1179	666.30
EF534217.1	contig_4434	Eucalyptus grandis fasciclin-like arabinogalactan protein (FLA2) mRNA, complete cds.	1125	180.66
EF534218.1	contig_2707	Eucalyptus grandis fasciclin-like arabinogalactan protein (FLA3) mRNA, complete cds.	1033	224.10
EF534219.1	contig_2477	Eucalyptus grandis beta-tubulin (TUB1) mRNA, complete cds.	1583	285.33
EF534220.1	contig_64905	Eucalyptus grandis beta-tubulin (TUB2) mRNA, complete cds.	1654	55.93
EF534223.1	contig_4441	Eucalyptus grandis beta-tubulin (TUB5) mRNA, complete cds.	1607	307.08
EF534224.1	contig_100	Eucalyptus grandis alpha-tubulin (TUA1) mRNA, complete cds.	1657	674.32
EU737107.1	contig_2692	Eucalyptus grandis UTP-glucose 1 phosphate uridylyltransferase (UGP) mRNA, complete cds.	1431	153.30
EU737108.1	contig_33128	Eucalyptus grandis UDP-D-glucuronate carboxy-lyase (UXS1) mRNA, complete cds.	1041	158.60
EU770570.1	contig_2246	Eucalyptus grandis iron-sulfer cluster scaffold protein ISU1 (ISU1) mRNA, complete cds.	756	78.07

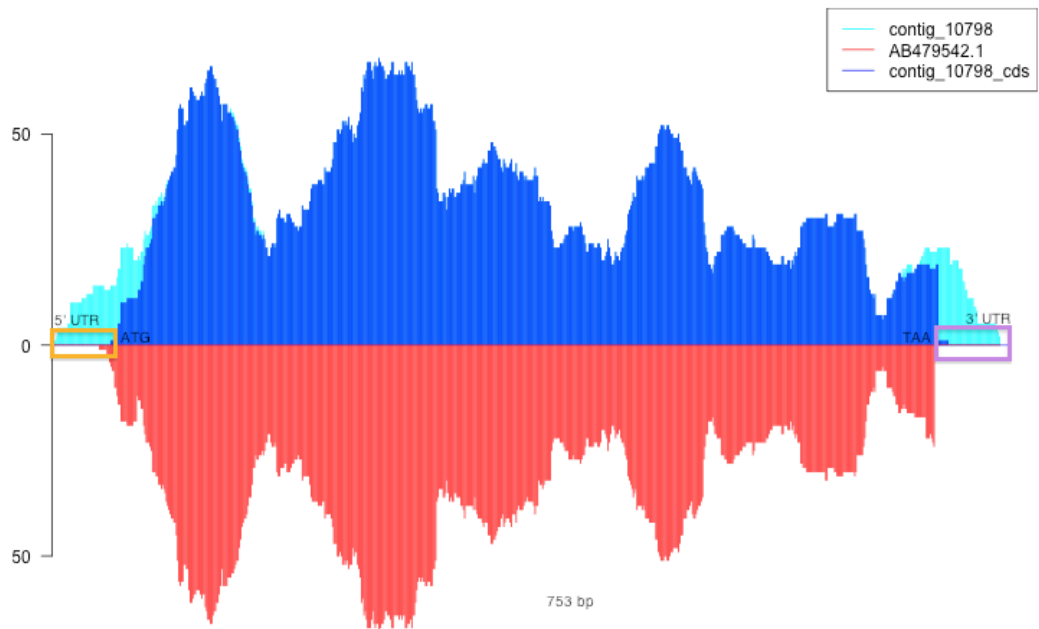
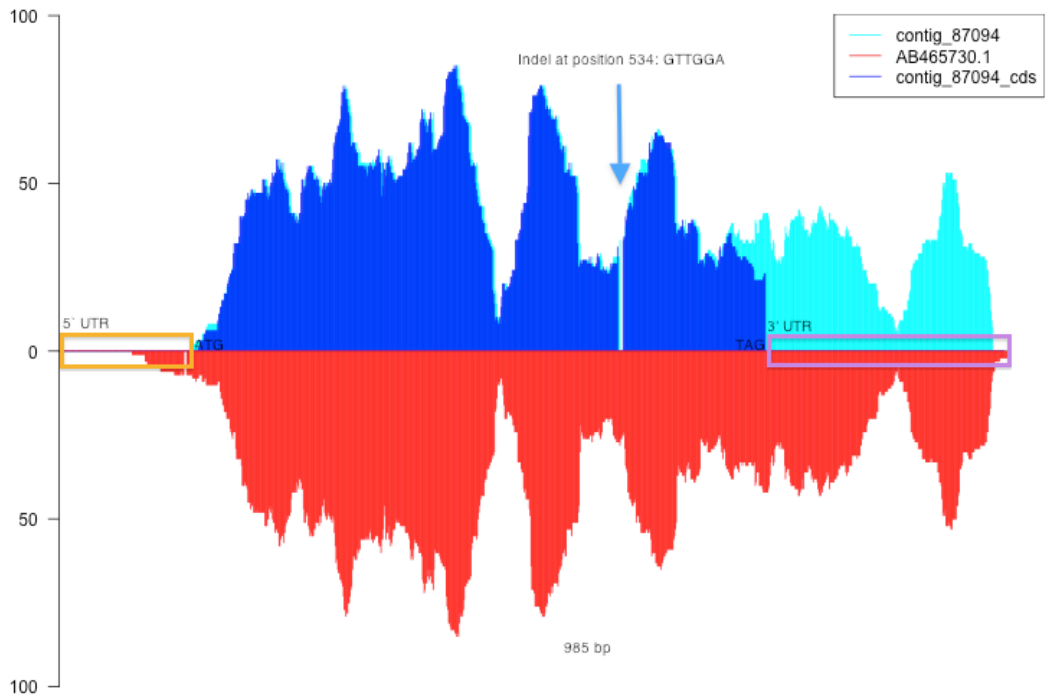


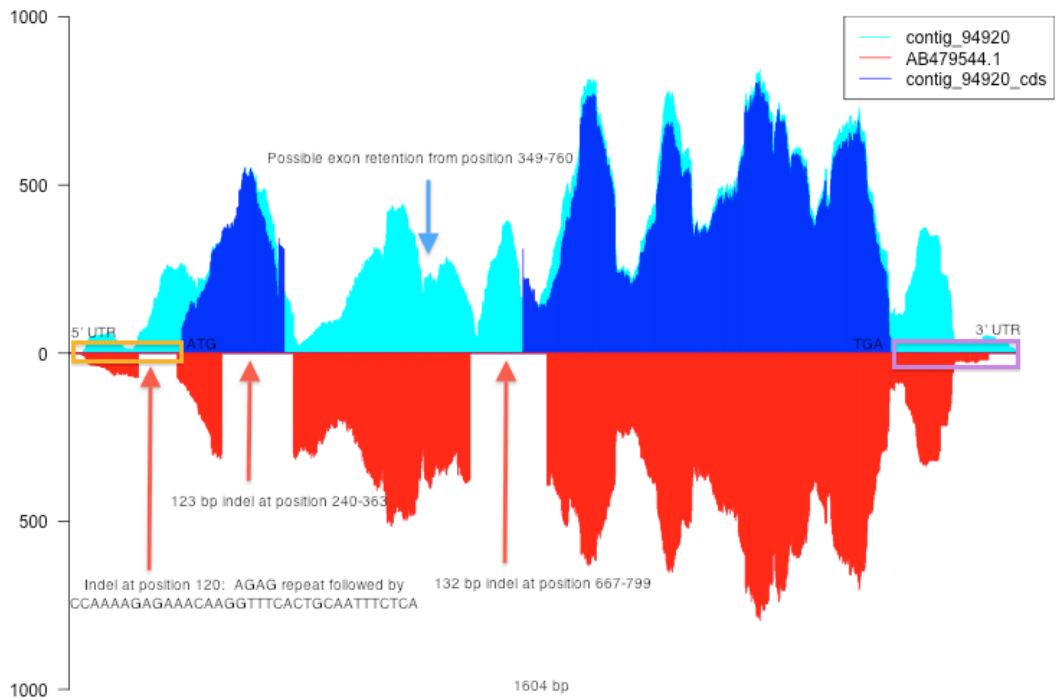
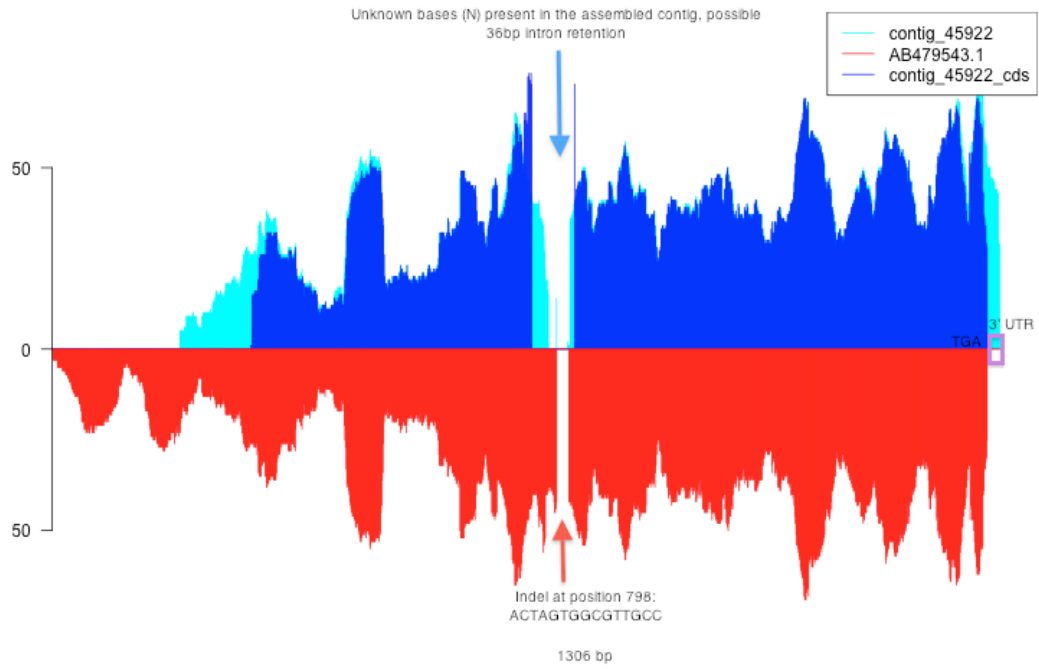
EU770571.1	contig_31483	Eucalyptus grandis iron-sulfer cluster scaffold protein NFU4 (NFU4) mRNA, partial cds.	869	13.30
EU770572.1	contig_15010	Eucalyptus grandis iron-sulfer cluster scaffold protein ISA1 (ISA1) mRNA, partial cds.	822	25.81
EU770573.1	contig_25291	Eucalyptus grandis iron-sulfer cluster scaffold protein NFS1 (NFS1) mRNA, partial cds.	871	16.29

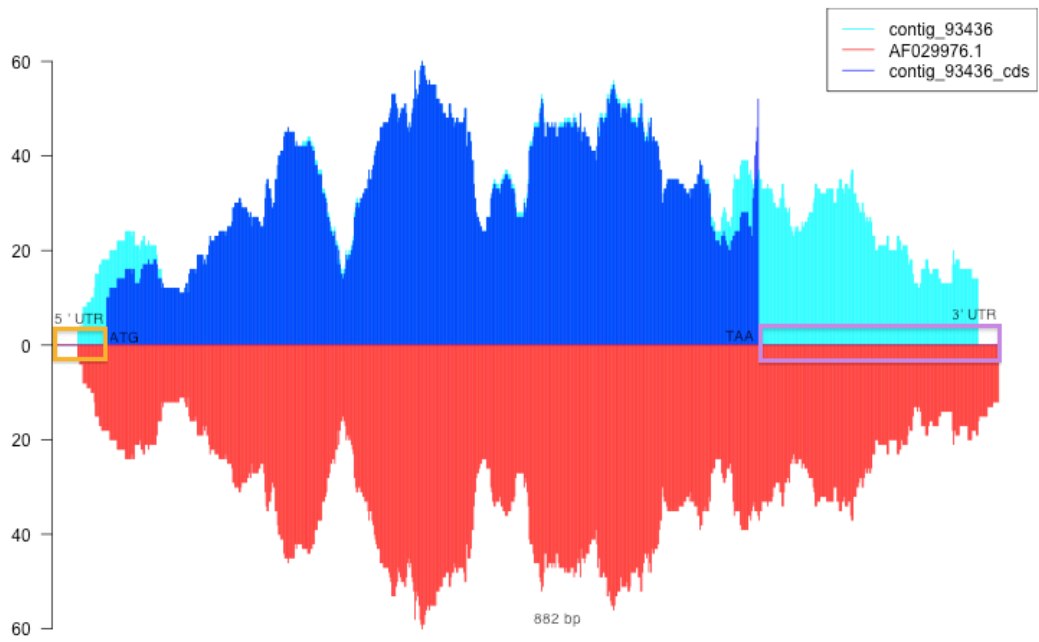
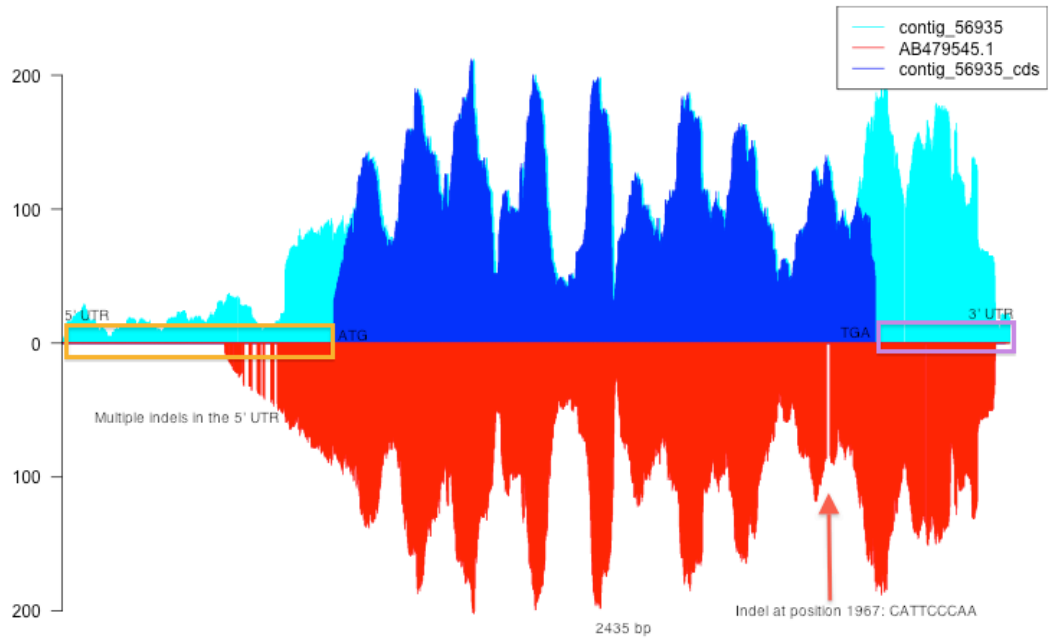
---

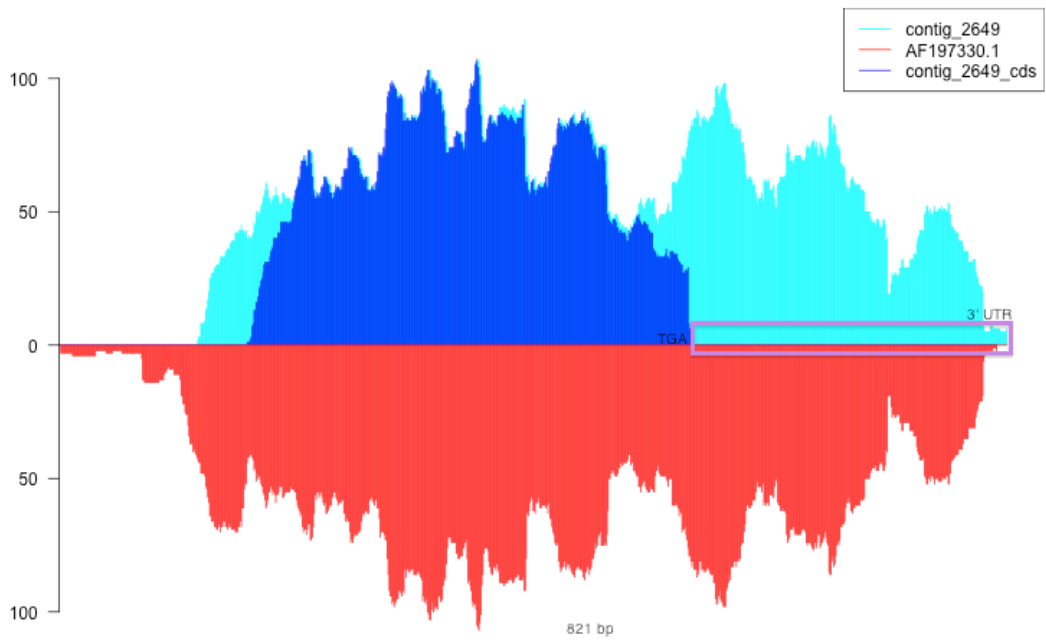
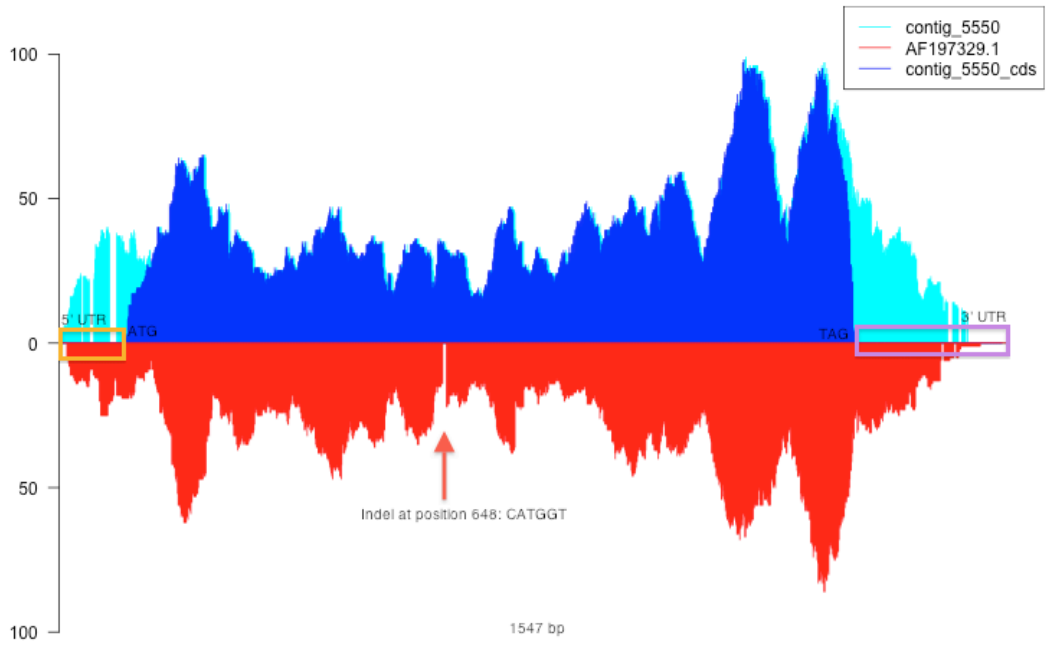
### C.1.1.2. Alignment coverage graphs of the 33 full length cDNA sequences and assembled contigs

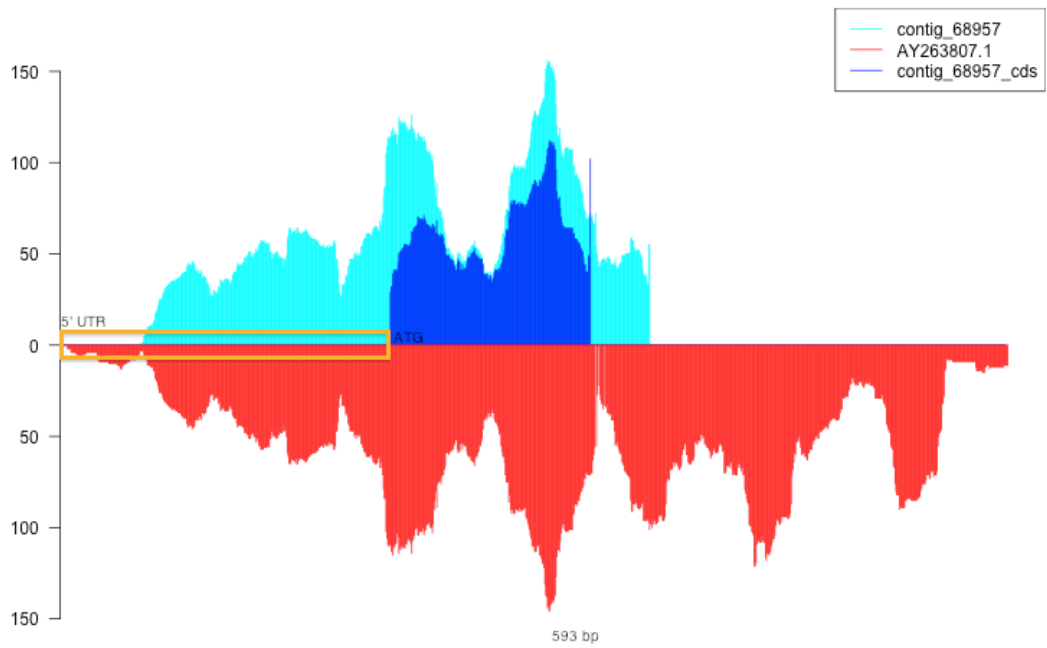
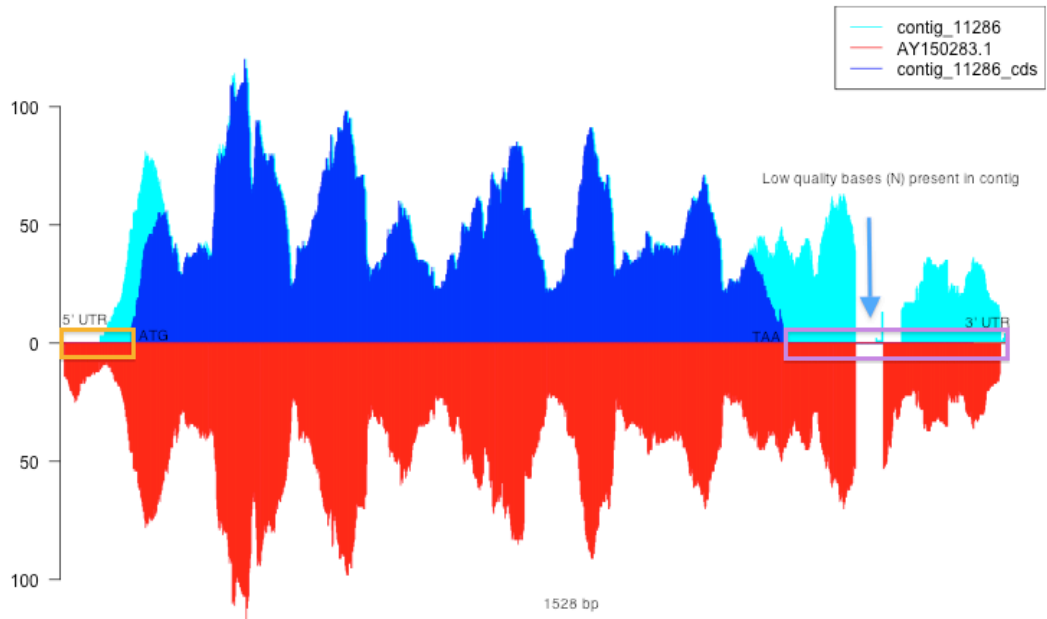
Comparison of 33 *de novo*-assembled contigs of the *Eucalyptus grandis* x *Eucalyptus urophylla* clone compared to the reference contigs obtained from Sanger sequencing. Peak heights indicates the actual coverage per base (CPB) across the contig. The CPB of the assembled contig is shown in cyan, the CPB of the predicted CDS in dark blue, and the CPB of the reference sequence in red. Where present, the 5' UTR (orange box) and the 3' UTR (purple box) is indicated. Large gaps in the global alignment between the sequences are indicated by gaps in the graph, and possible reasons for the gap annotated on each graph. The graphs are also available as supplementary material for the article by Mizrachi *et al.* (2010).



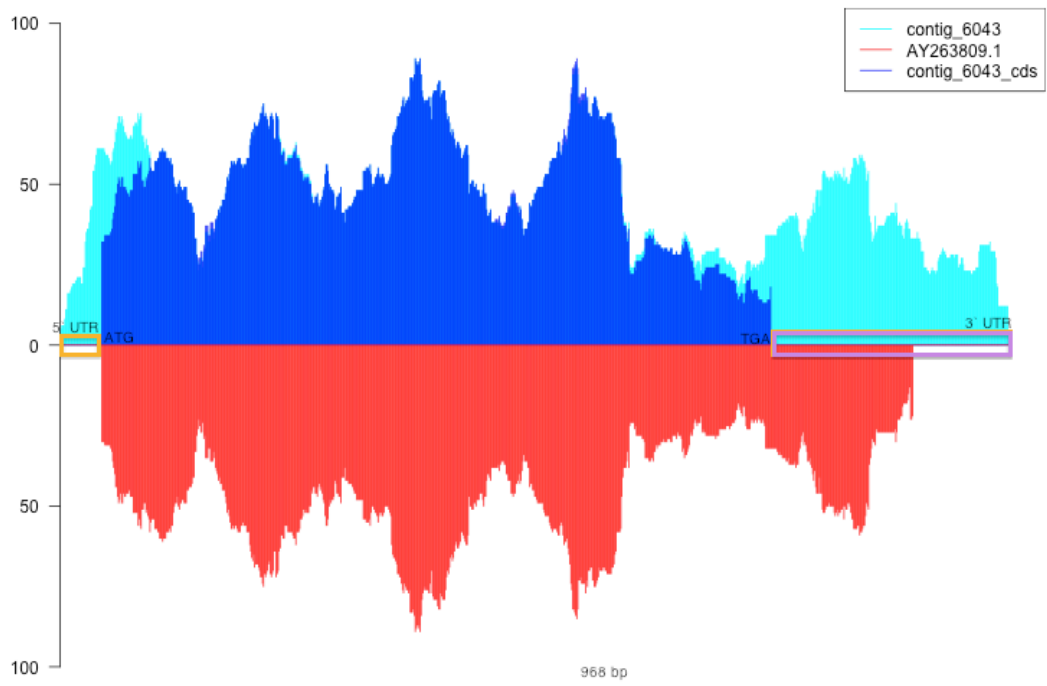
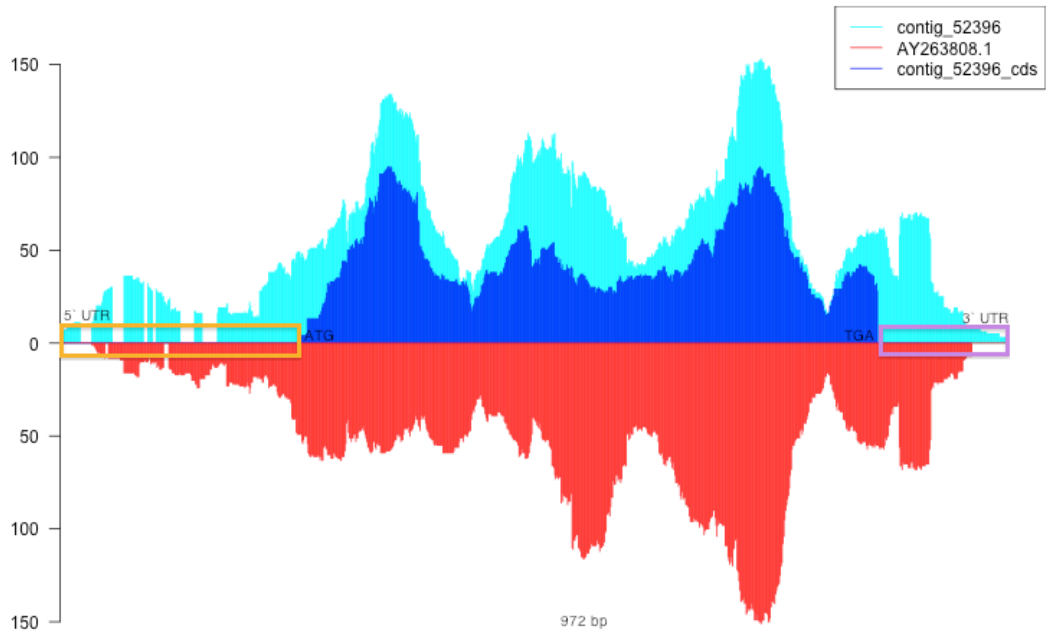


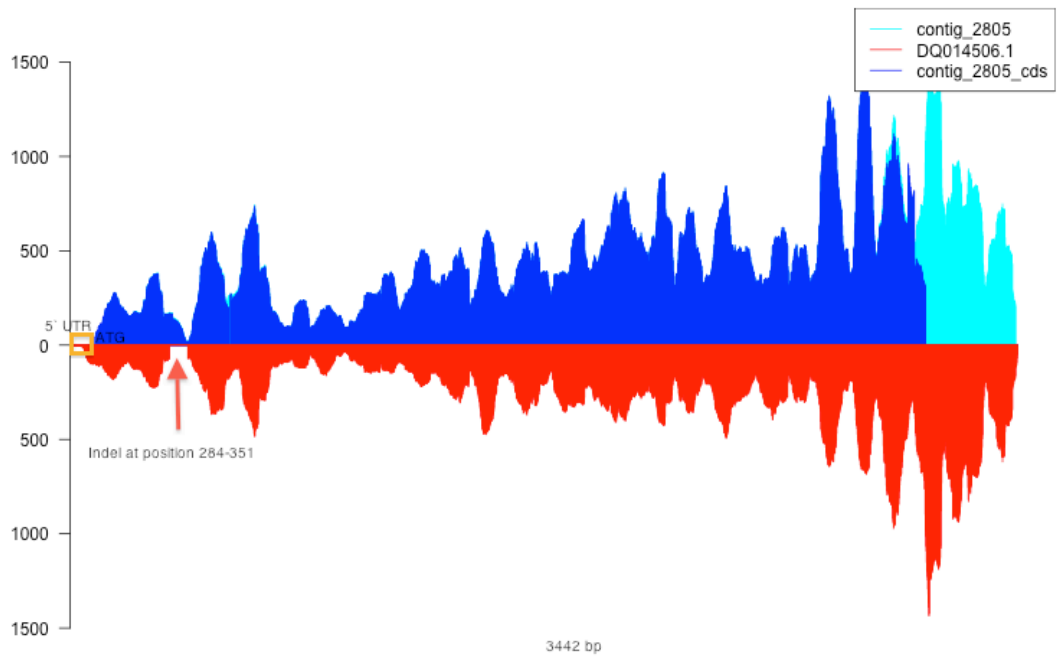
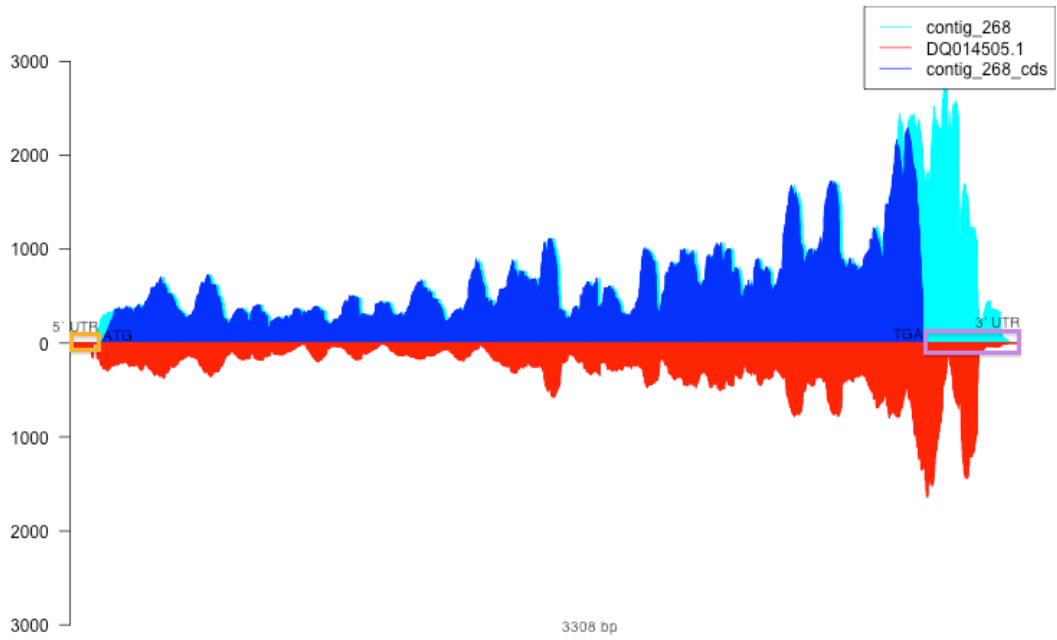


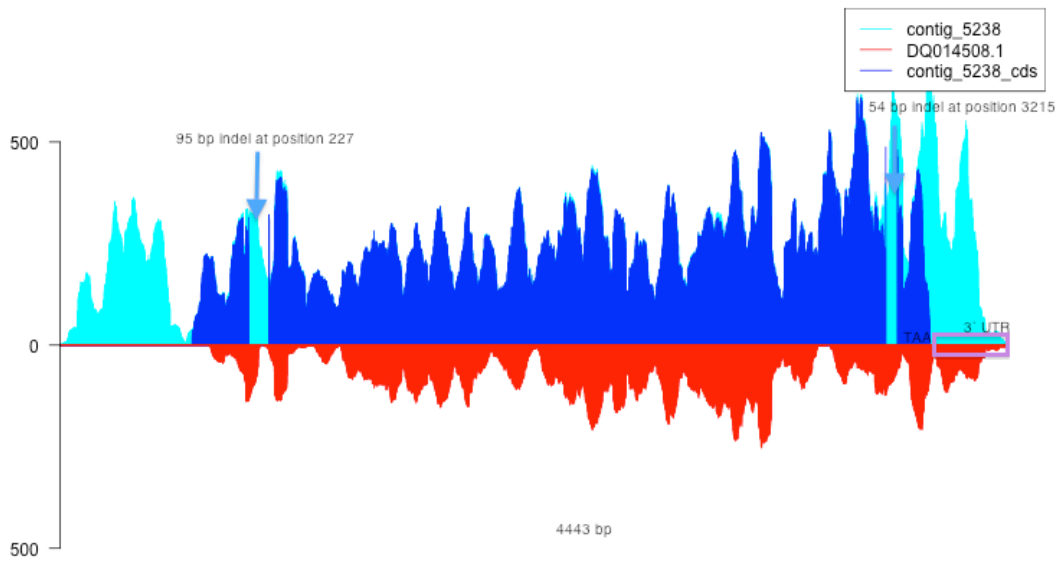
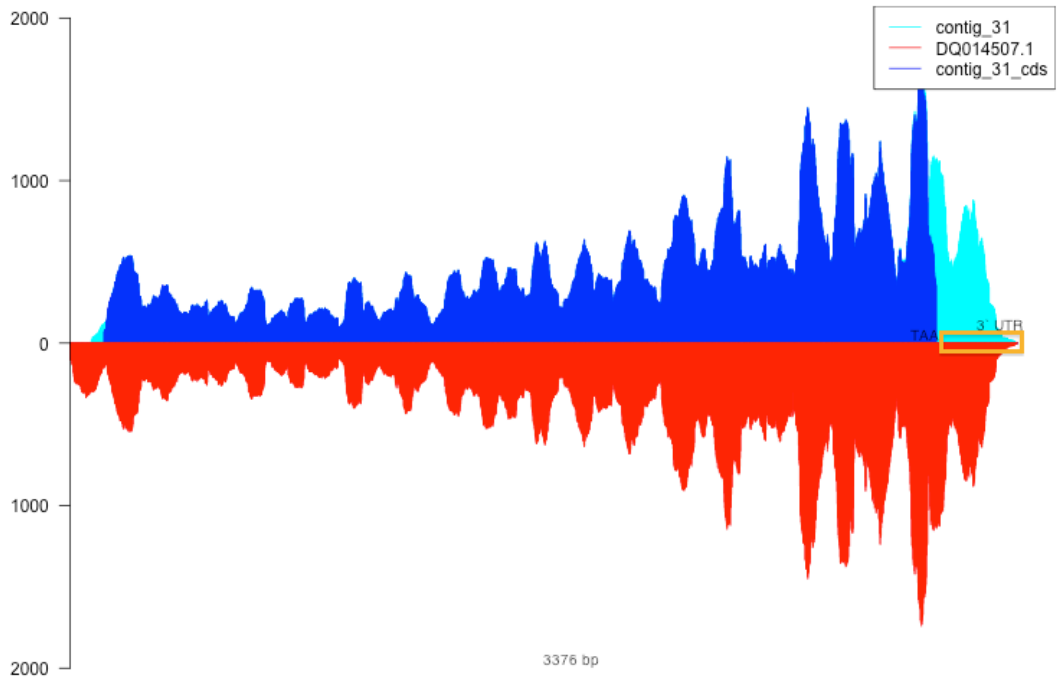


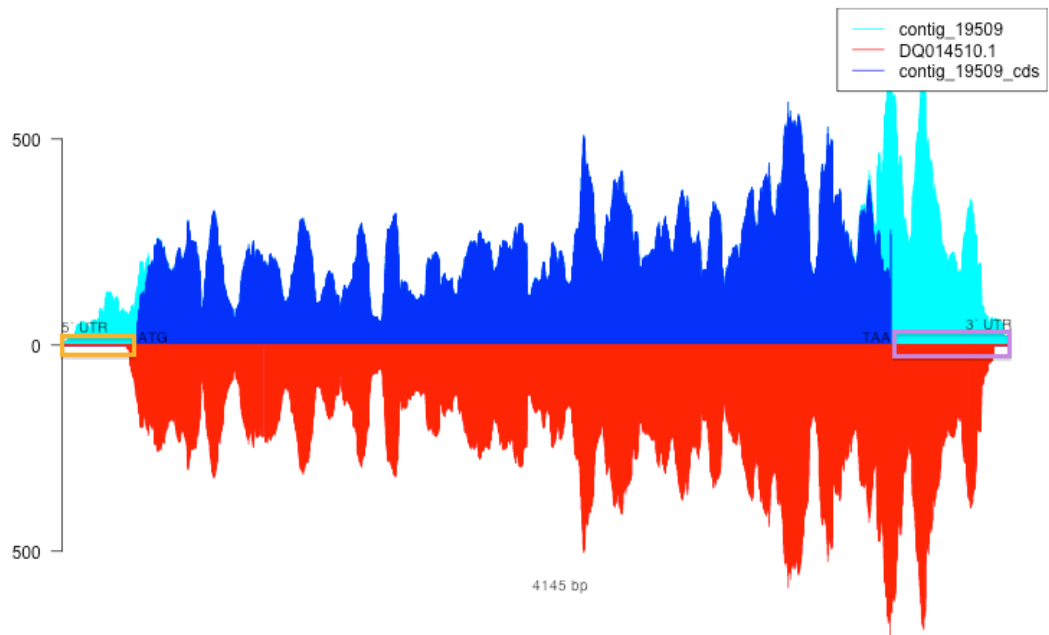
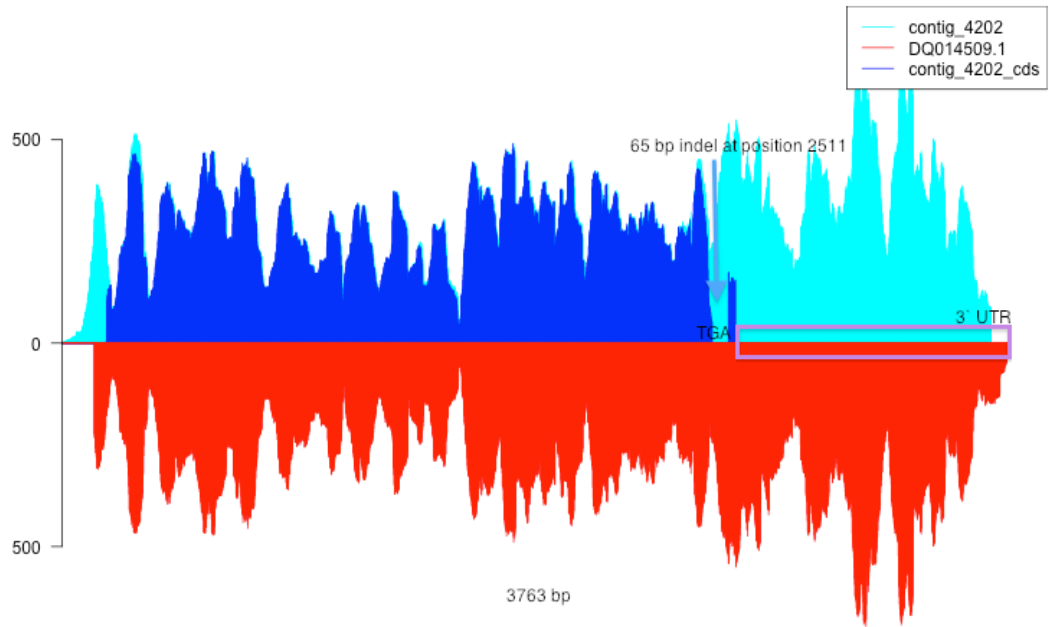


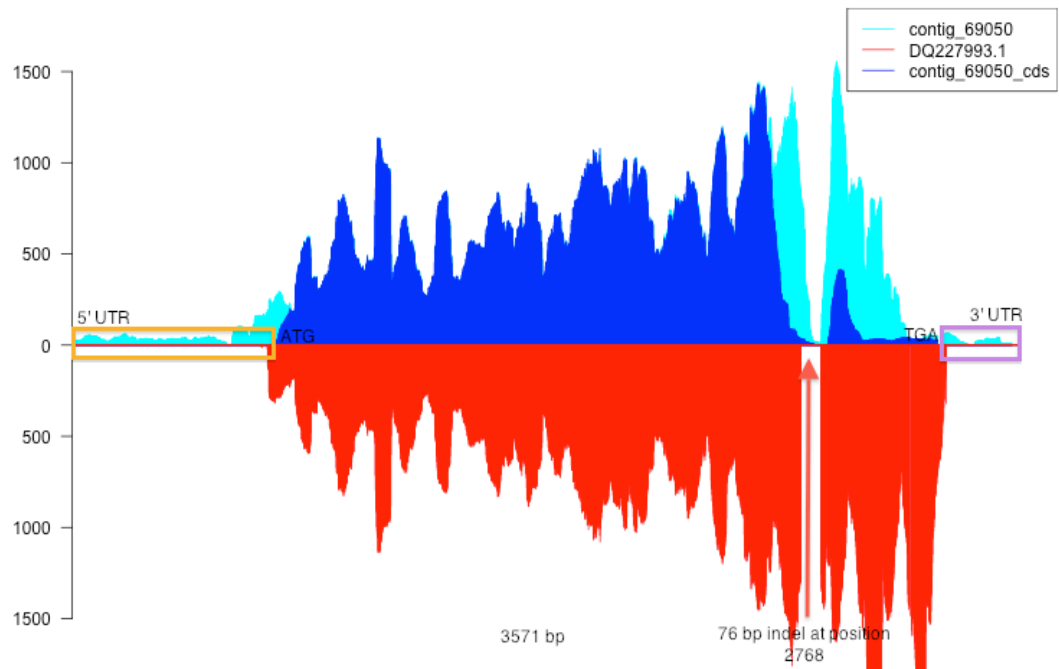
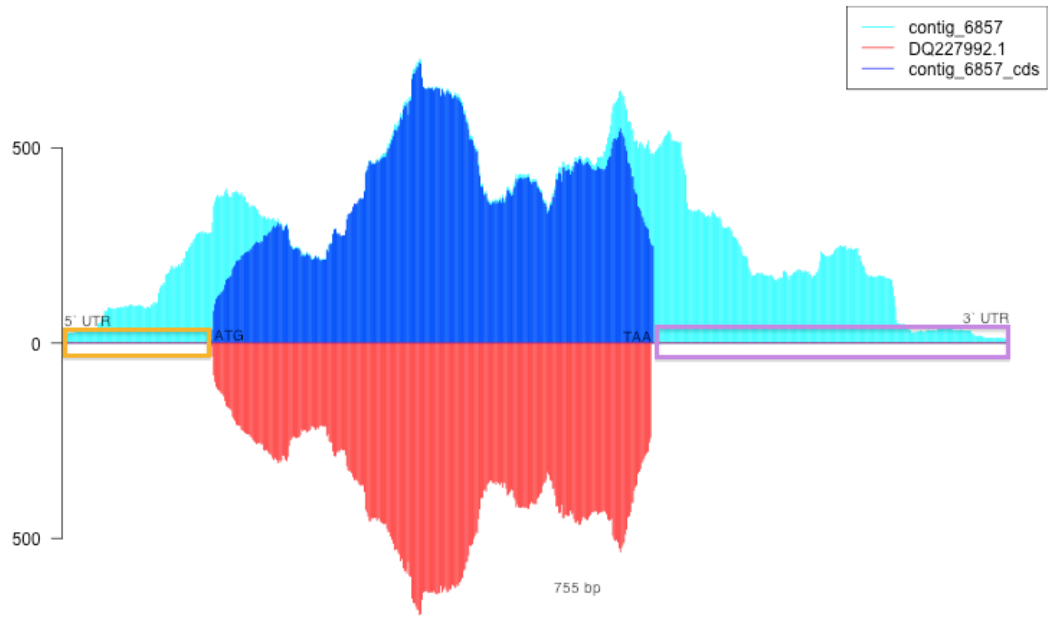


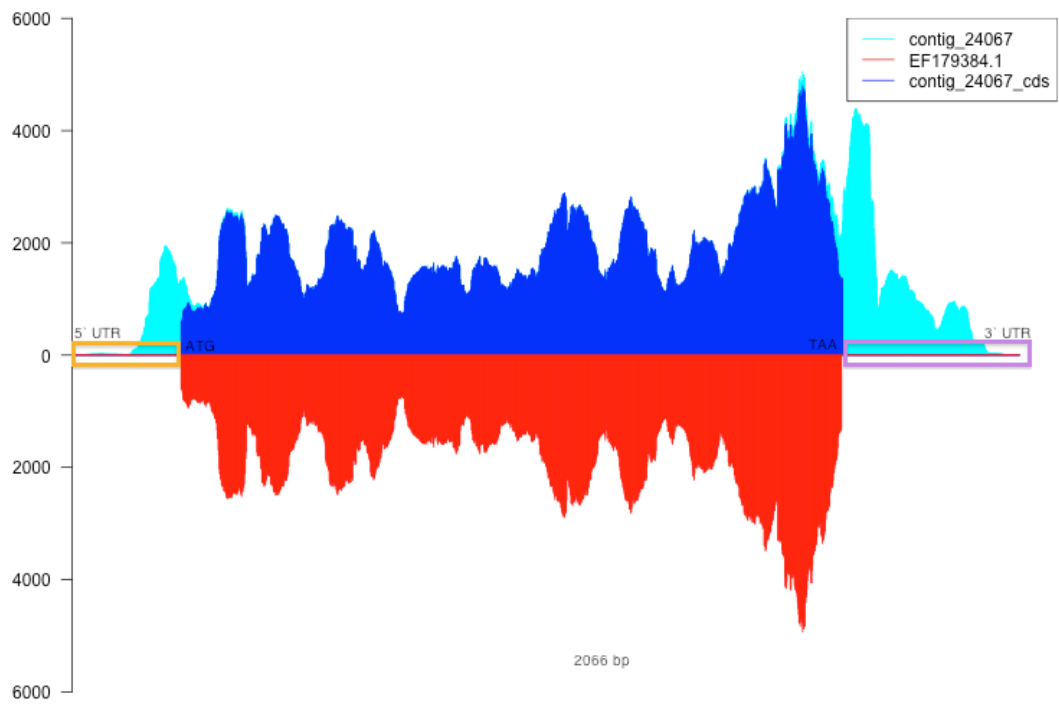
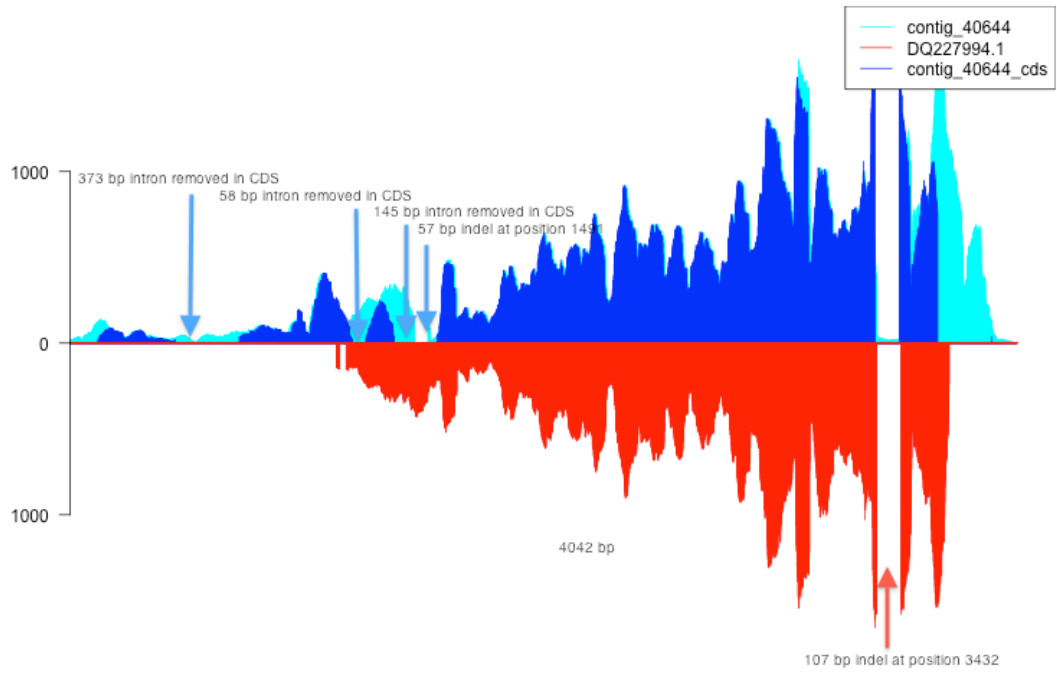


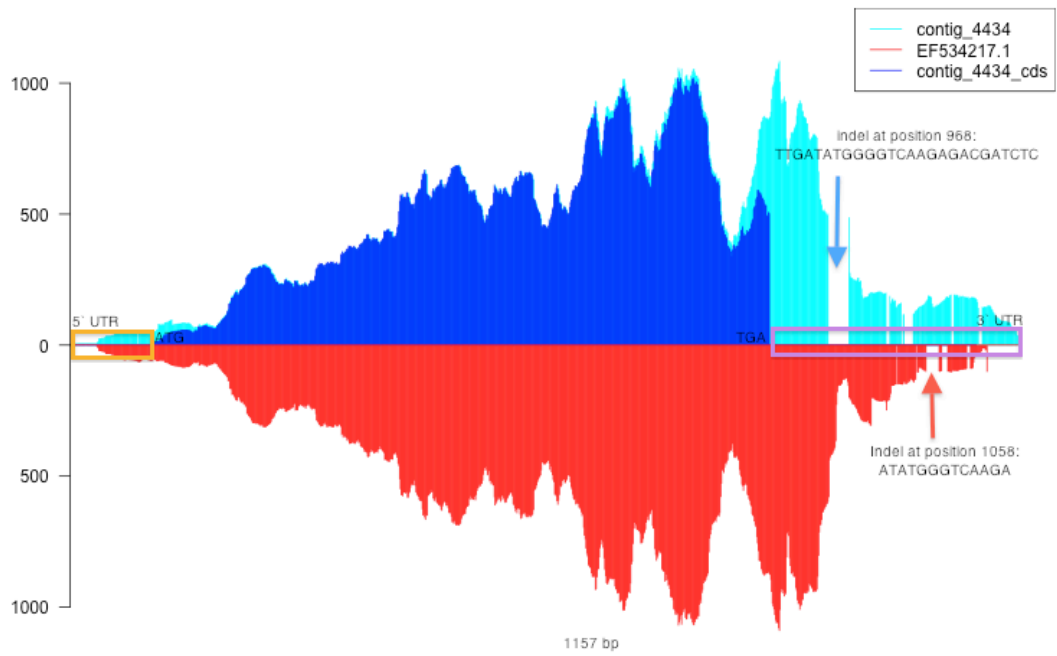
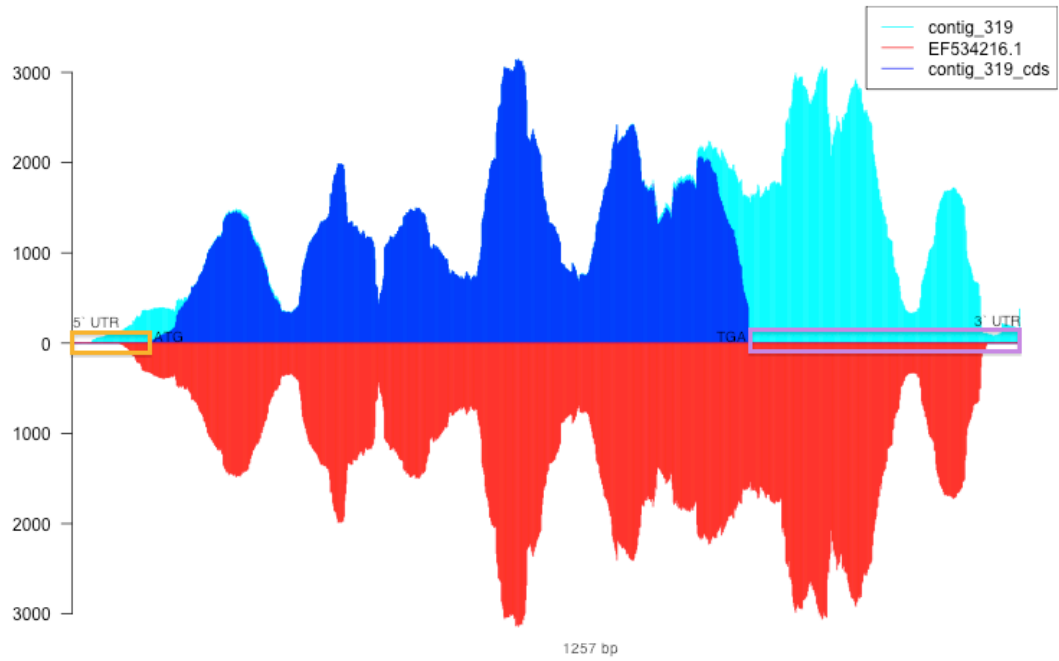




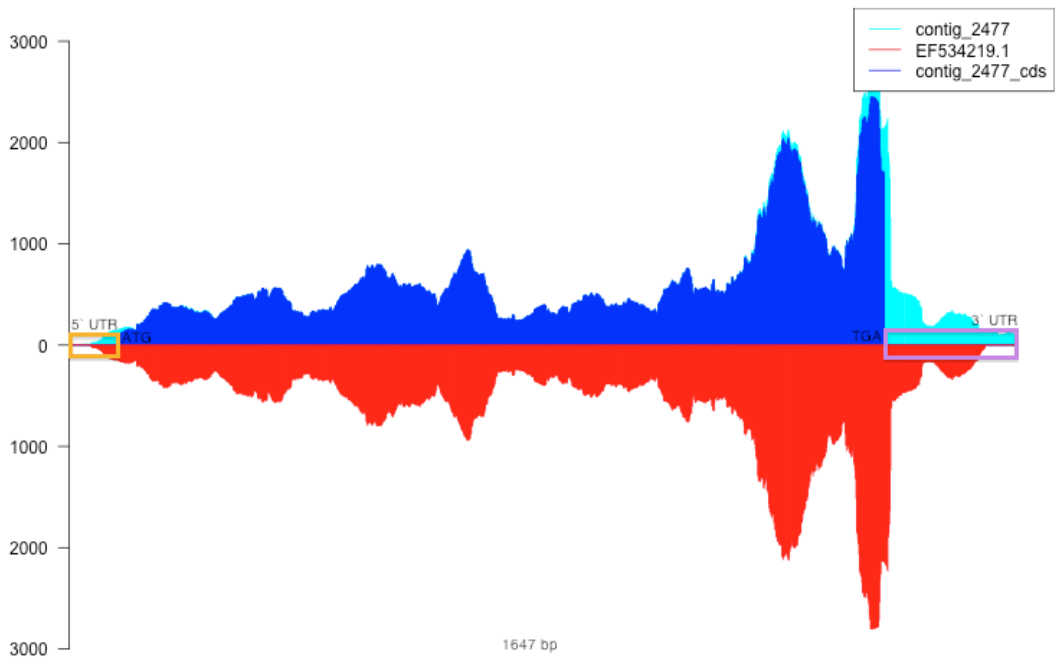
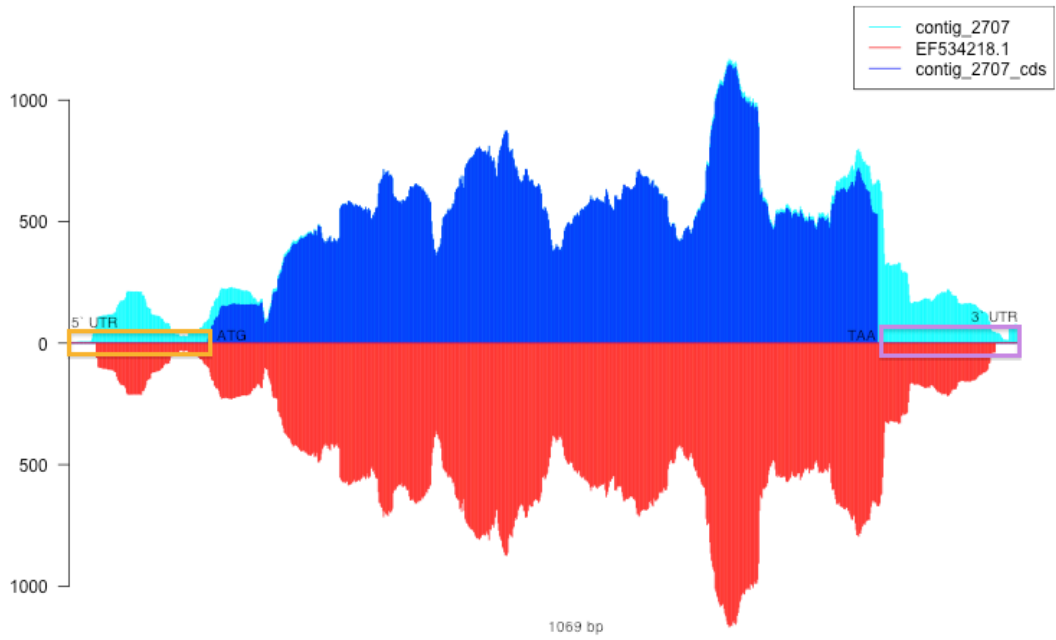


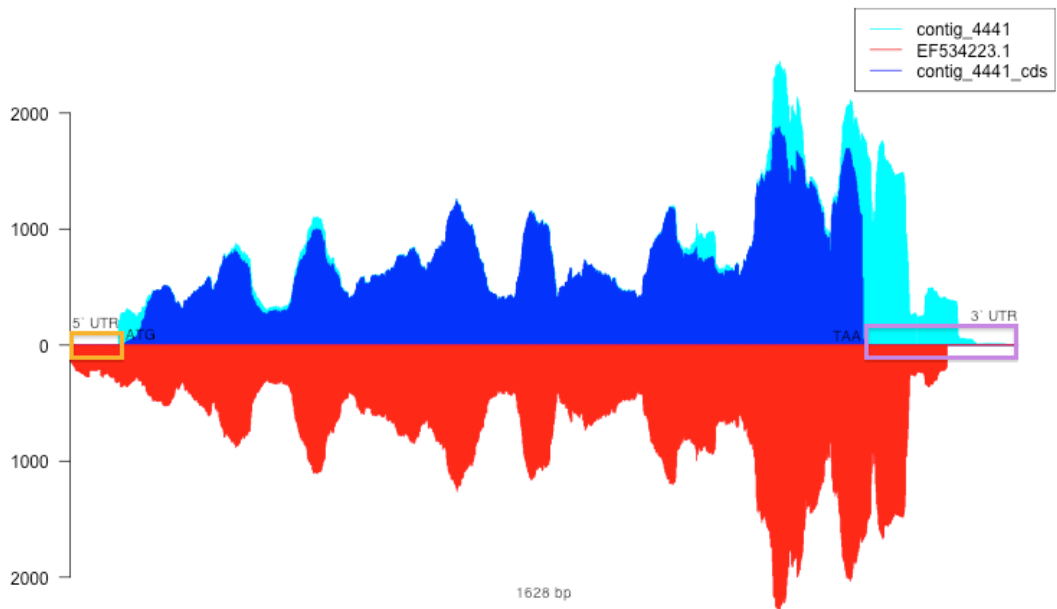
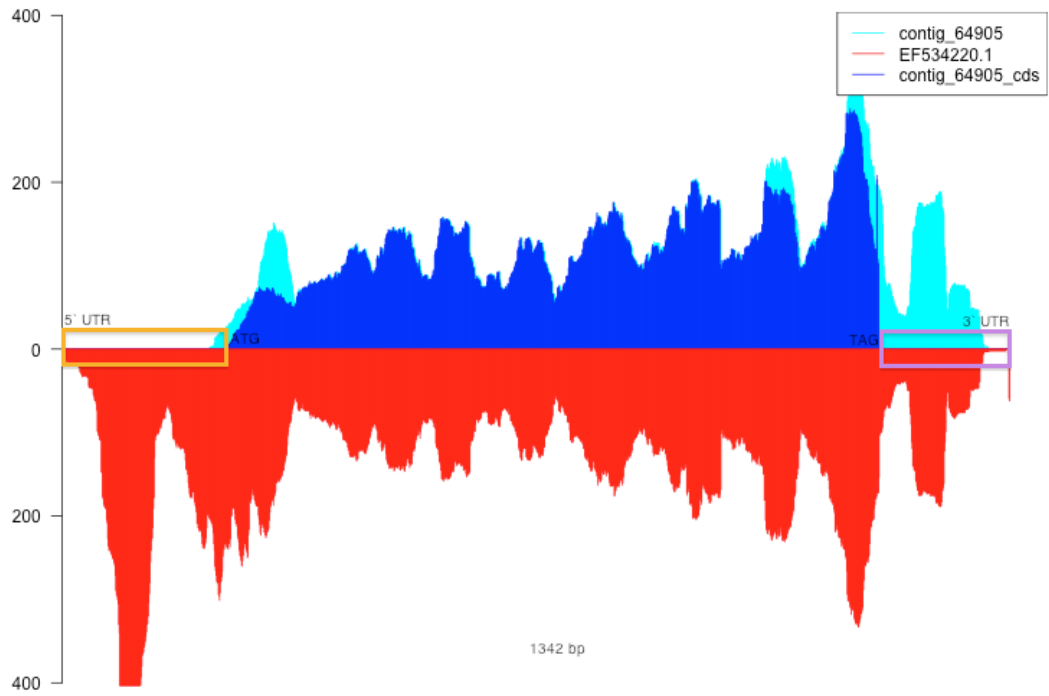


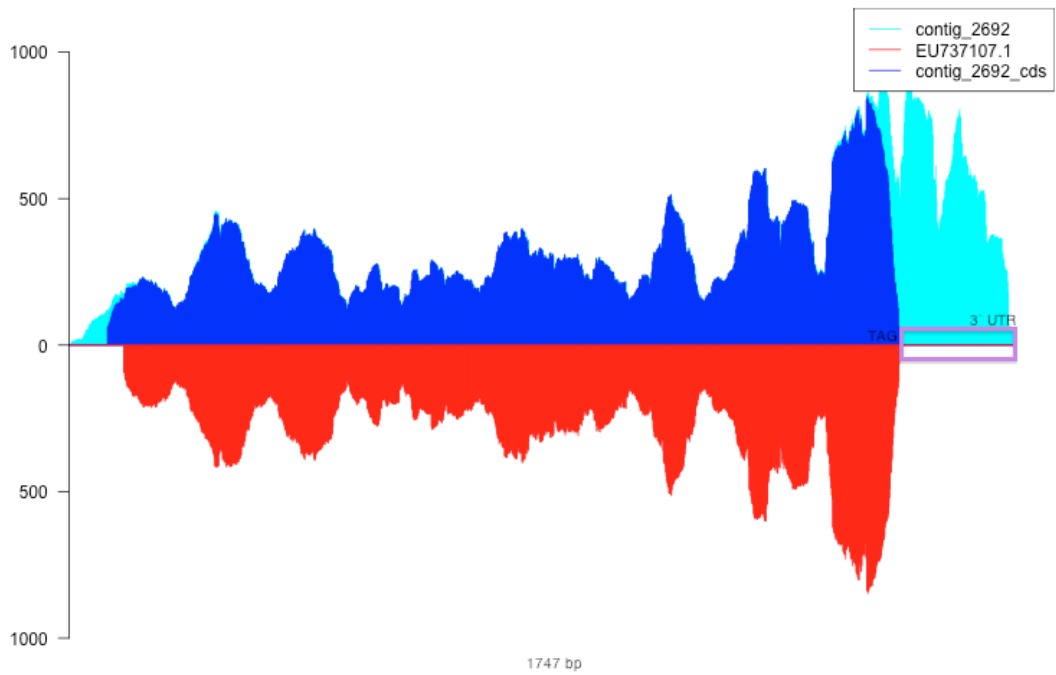
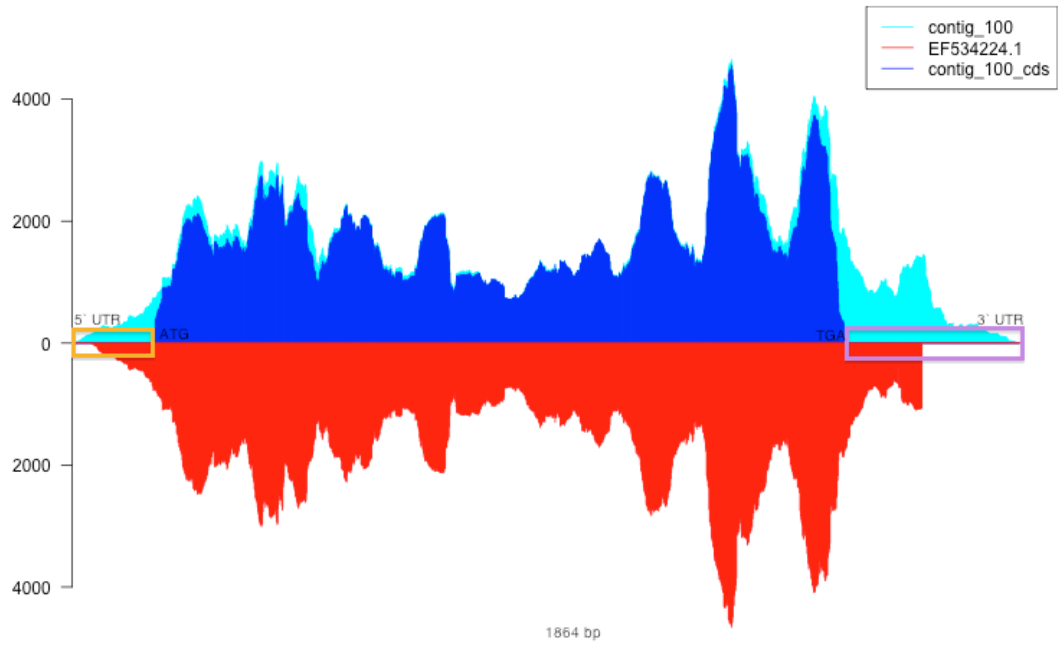


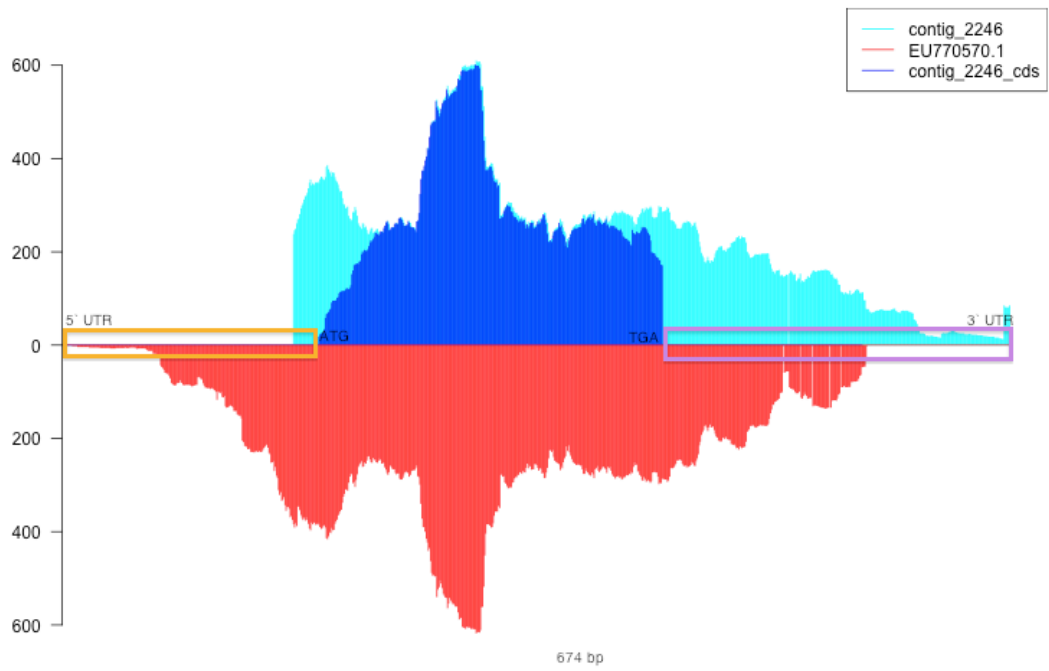
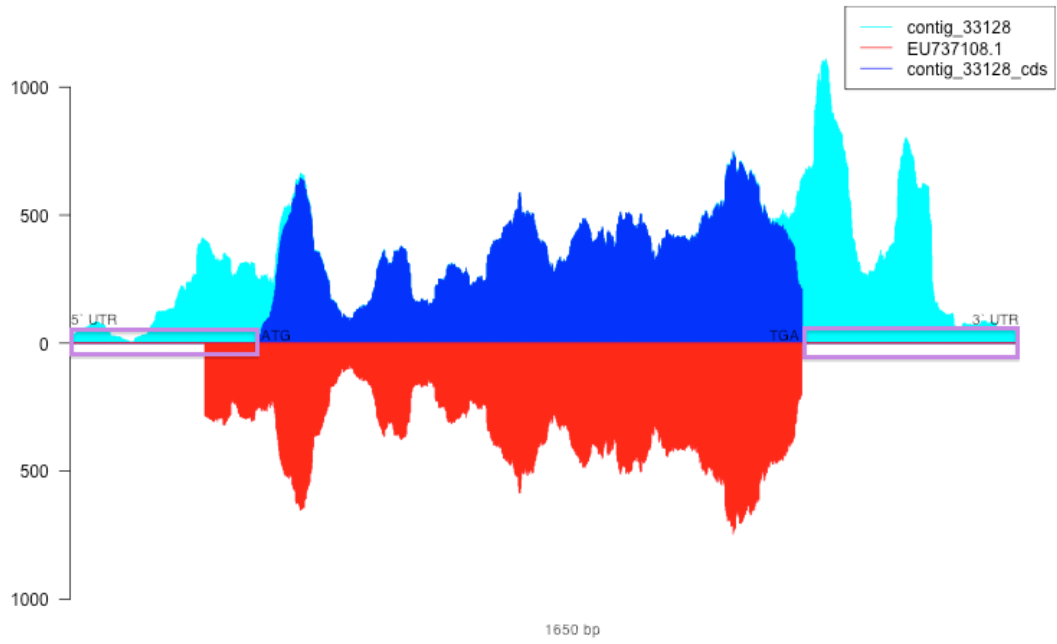


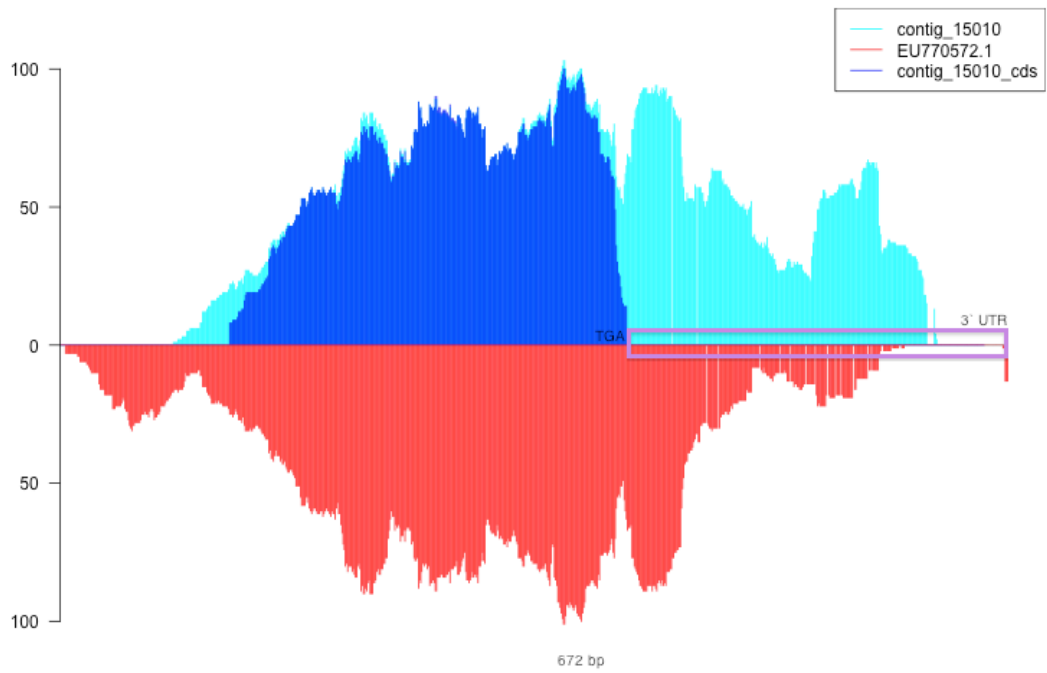
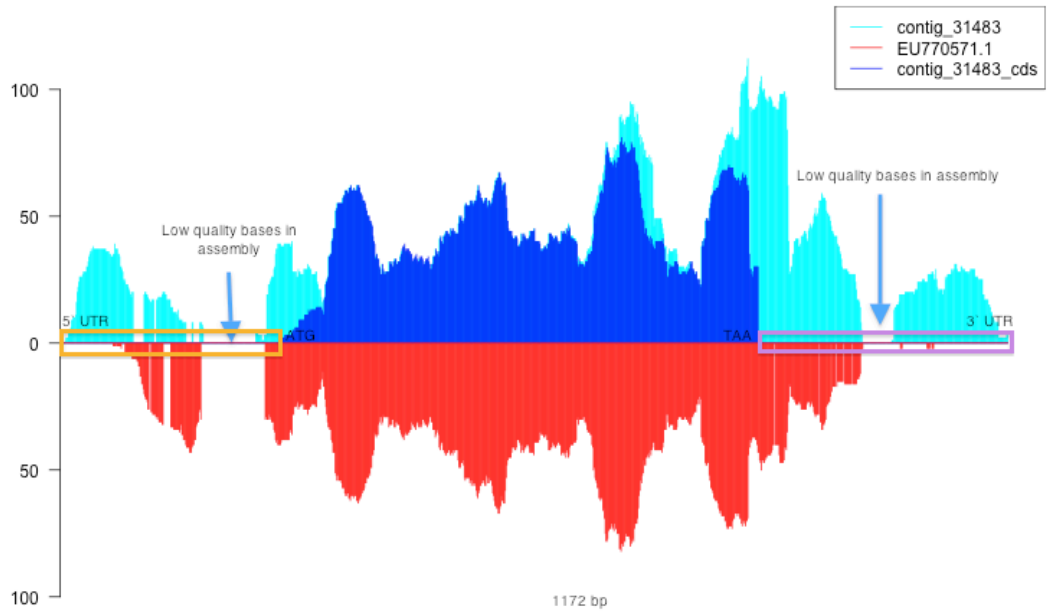


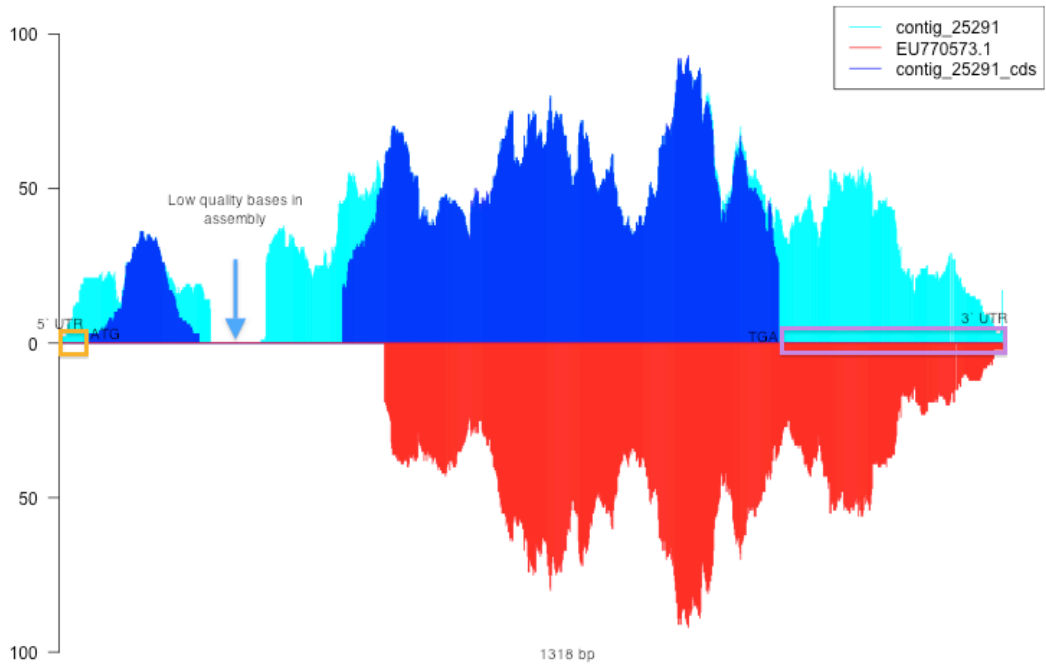












### **C.1.3. Alignment of contig 68291 before and after extension**

The complete alignment of contig or node 68291 before and after the coverage-assisted re-assembly of the dataset. The alignment was performed with the the `ClustalW` program, and no editing of the alignment was performed. The alignment shows that although 1 485 bases was reportedly added to the contig during extension, these bases mostly consisted of the extension of a low quality region containing Ns. The extension did however resolve a 88 bp region of these low quality bases. The contig after extension also showed removed regions at the start and end of the original contig, due to the presence of a polyA region at the beginning of the sequence. An extract from the alignment is presented in Figure 3.6.



NODE_68291_before	TTTTTTTTTTTTTTTTTTTTTTT	AGAGAGGCTTGGAGTGGTCAGGACTCTTA	80
NODE_68291_after	-----	-----	
NODE_68291_before	TTCCAGAGAGAAAGCATCAGCGCGGGCTGTCAAAGCTTCAATGAACGTATAATTGTTACTGAAGCGGAGATAAAGGAGA	** * * * *	160
NODE_68291_after	-----	-----AAAAAAAAAAAAAAAAAAAAA	21
NODE_68291_before	GTAAAGNNNNCAAAGCAAAGCTGGTGC TTTTGTGCATCTCCATTGACTTGGCCAATTGGTCCATTGATAAGTTCGGCAACT	**	240
NODE_68291_after	AAAAAAAAAAACAAAGCAAAGCTGGTGC TTTTGTGCATCTCCATTGACTTGGCCAATTGGTCCATTGATAAGTTCGGCAACT		101
NODE_68291_before	CAGGCGATTCTGGCTTTGCGTCGGTGAAGTTCCTCGTACTGCGCAATGGCGTCCGCTCTAGCCGGCGATGATTGGCTA		320
NODE_68291_after	CAGGCGATTCTGGCTTTGCGTCGGTGAAGTTCCTCGTACTGCGCAATGGCGTCCGCTCTAGCCGGCGATGATTGGCTA		181
NODE_68291_before	GATCCACGAGCAGCCGCCGAGCTGGGCTCCGGGAGCCACCGGAGCTGGGCTCCACGAGCTTCCGGGAGGCGTGGAAC		400
NODE_68291_after	GATCCACGAGCAGCCGCCGAGCTGGGCTCCGGGAGCCACCGGAGCTGGGCTCCACGAGCTTCCGGGAGGCGTGGAAC		261
NODE_68291_before	GGCCCGCCAGATGTGTTCCGCGGGAGCGGGAGGCAGGACGACGAGGAGGAGCTCCGGTGGCCGCCATCGAACGGCTGCC		480
NODE_68291_after	GGCCCGCCAGATGTGTTCCGCGGGAGCGGGAGGCAGGACGACGAGGAGGAGCTCCGGTGGCCGCCATCGAACGGCTGCC		341
NODE_68291_before	AACGTAATGACCGCTCCGAAAAGGCATGCTGAAGCAAGTACTTGATACTGGGAGGGTGGTCCAGCAAGAAGTGGACGTGA		560
NODE_68291_after	AACGTAATGACCGCTCCGAAAAGGCATGCTGAAGCAAGTACTTGATACTGGGAGGGTGGTCCAGCAAGAAGTGGACGTGA		421
NODE_68291_before	CCAACCTCGGAATGACAGGACAAAGAGCAGTTGATGGAGAGCATCCCTTAAGGTTCCGGAAGAAGCAATGAGAGGTTCTTG		640
NODE_68291_after	CCAACCTCGGAATGACAGGACAAAGAGCAGTTGATGGAGAGCATCCCTTAAGGTTCCGGAAGAAGCAATGAGAGGTTCTTG		501
NODE_68291_before	AGGAGATTGAGAGACAGGACTGATAGGGTCCGGATCGAAATTCGAAAGATCGAAGTCCGGTGTGAGCATTATCTGTAGA		720
NODE_68291_after	AGGAGATTGAGAGACAGGACTGATAGGGTCCGGATCGAAATTCGAAAGATCGAAGTCCGGTGTGAGCATTATCTGTAGA		581
NODE_68291_before	AGGAGACGTGTACGTTGGAAGCAGAGCTCTCCCTACCTTCTCAATGCCACTATGAACCGGATAGAGAGTGTCTTGGAC		800
NODE_68291_after	AGGAGACGTGTACGTTGGAAGCAGAGCTCTCCCTACCTTCTCAATGCCACTATGAACCGGATAGAGAGTGTCTTGGAC		661
NODE_68291_before	TTATTCCGCTAGCCCCATCGAAGAAGAGAAAAATTCAGATACTTAAGGACGTGAACGGATTAGTACGGCCTTCGAGGATG		880
NODE_68291_after	TTATTCCGCTAGCCCCATCGAAGAAGAGAAAAATTCAGATACTTAAGGACGTGAACGGATTAGTACGGCCTTCGAGGATG		741
NODE_68291_before	ACCCTACTTTTGGGTCCACC GGGAGCTGGGAAGACAACA TTGTTGCTGGCACTTGCTGGGAAACTAGACAGCGATCTGAG		960
NODE_68291_after	ACCCTACTTTTGGGTCCACC GGGAGCTGGGAAGACAACA TTGTTGCTGGCACTTGCTGGGAAACTAGACAGCGATCTGAG		821
NODE_68291_before	GGTAACGGGAAAAATCACCCTACTGTGGTCCAGAGCTAAACGAATTTGTTCTCAAAGGACTTGCCTTATATCAGCCAAC		1040
NODE_68291_after	GGTAACGGGAAAAATCACCCTACTGTGGTCCAGAGCTAAACGAATTTGTTCTCAAAGGACTTGCCTTATATCAGCCAAC		901
NODE_68291_before	ATGATCTTCACTATGGGGAATGACAGTTAGAGAGACATTGGACTTCTCGGGTCCGCTGTTGGGTGTAGGGACAAGGTAT		1120
NODE_68291_after	ATGATCTTCACTATGGGGAATGACAGTTAGAGAGACATTGGACTTCTCGGGTCCGCTGTTGGGTGTAGGGACAAGGTAT		981
NODE_68291_before	GAGATGCTTGCAAGACTCTCCAGGCGAGAGGGGAAGCCGAATCAAACCTGATCCCGAAATTGACGCTTTTATGAAGGC		1200
NODE_68291_after	GAGATGCTTGCAAGACTCTCCAGGCGAGAGGGGAAGCCGAATCAAACCTGATCCCGAAATTGACGCTTTTATGAAGGC		1061
NODE_68291_before	CACAGCTCTGTCCGGTCAAAGAGACAAGCTTGGTCACTGATTAATACTCAAGATTCTTGGATTGGATATCTGTGCAGACA		1280
NODE_68291_after	CACAGCTCTGTCCGGTCAAAGAGACAAGCTTGGTCACTGATTAATACTCAAGATTCTTGGATTGGATATCTGTGCAGACA		1141
NODE_68291_before	TTATGGTCCGAGATGAGATGCGAAGGGCATTTCAGGTGGACAAAAAAGCGCTTACAAACCGGAGAGATGTTAGTAGGA		1360
NODE_68291_after	TTATGGTCCGAGATGAGATGCGAAGGGCATTTCAGGTGGACAAAAAAGCGCTTACAAACCGGAGAGATGTTAGTAGGA		1221
NODE_68291_before	CCAGCAAAGGCTCTTTTTATGGATGAAATATCCACAGGGTGGACAGTTCACCTACTTTTCAAATTTGCAAATTCATGAG		1440
NODE_68291_after	CCAGCAAAGGCTCTTTTTATGGATGAAATATCCACAGGGTGGACAGTTCACCTACTTTTCAAATTTGCAAATTCATGAG		1301
NODE_68291_before	GCAGATGGTTCAATTTATGATGTCAACATGATCATCTCATTTGCTTCAGCCGGCTCCTGAGACTTATGATCTTTCGATG		1520
NODE_68291_after	GCAGATGGTTCAATTTATGATGTCAACATGATCATCTCATTTGCTTCAGCCGGCTCCTGAGACTTATGATCTTTCGATG		1381





NODE_68291_before	***** ACATTATCCTTCTCTCGGAGGGT	***** ACGTCCTCGAGTTTTTCGAGCACATGGGA	1600
NODE_68291_after	ACATTATCCTTCTCTCGGAGGGTCAAGTCGTCTACCAAGGTCACGAGAGAAACGTCCTCGAGTTTTTCGAGCACATGGGA		1461
NODE_68291_before	***** TTCAAGTGCCTTCAAAGGAAAGGAGTTGCCGACTTCTTGCAAGAAAGTGACATCTAAGAAAGATCAAGAACAGTATTGGTT	***** TTCAAGTGCCTTCAAAGGAAAGGAGTTGCCGACTTCTTGCAAGAAAGTGACATCTAAGAAAGATCAAGAACAGTATTGGTT	1680
NODE_68291_after	TTCAAGTGCCTTCAAAGGAAAGGAGTTGCCGACTTCTTGCAAGAAAGTGACATCTAAGAAAGATCAAGAACAGTATTGGTT		1541
NODE_68291_before	***** CAAGAAACCAACCTTTCCAATACGTTTCTGTAGATGATTTCTGTCATGGATTCAAATCTTTTCACATTGGCCAAACATC	***** CAAGAAACCAACCTTTCCAATACGTTTCTGTAGATGATTTCTGTCATGGATTCAAATCTTTTCACATTGGCCAAACATC	1760
NODE_68291_after	CAAGAAACCAACCTTTCCAATACGTTTCTGTAGATGATTTCTGTCATGGATTCAAATCTTTTCACATTGGCCAAACATC		1621
NODE_68291_before	***** TGTCATCCGATCTTAGGATTCTTATGACAAATCAAAAATCACCAGCTGCACCTAGTCAAAGAGAAATACGGNNNN--	***** TGTCATCCGATCTTAGGATTCTTATGACAAATCAAAAATCACCAGCTGCACCTAGTCAAAGAGAAATACGGNNNN--	1838
NODE_68291_after	TGTCATCCGATCTTAGGATTCTTATGACAAATCAAAAATCACCAGCTGCACCTAGTCAAAGAGAAATACGGNNNNGC		1701
NODE_68291_before	-----	-----	1838
NODE_68291_after	ACTAGTCAAAGAGAAATACGGGATTTCAAATATGGAGCTGTTCAAGCATGCTTTGCCAGAGAAATGGCTACTAATGAAGC		1781
NODE_68291_before	***** ----TCCTTTGTTTACATATTCAAGAACCACCCAGATCACTATCATGTCGCTTATTGCTCTGACGGTGTCTCTTAGGACT	***** ----TCCTTTGTTTACATATTCAAGAACCACCCAGATCACTATCATGTCGCTTATTGCTCTGACGGTGTCTCTTAGGACT	1913
NODE_68291_after	GAAACTCCTTTGTTTACATATTCAAGAACCACCCAGATCACTATCATGTCGCTTATTGCTCTGACGGTGTCTCTTAGGACT		1861
NODE_68291_before	***** GAAATGCCAGTAGGGTCAGTGCAAGATGGAGGGAAAGTTTTTGGAGCACTTTTCTTTCAGCTTGATCAATGTATGTTCAA	***** GAAATGCCAGTAGGGTCAGTGCAAGATGGAGGGAAAGTTTTTGGAGCACTTTTCTTTCAGCTTGATCAATGTATGTTCAA	1993
NODE_68291_after	GAAATGCCAGTAGGGTCAGTGCAAGATGGAGGGAAAGTTTTTGGAGCACTTTTCTTTCAGCTTGATCAATGTATGTTCAA		1941
NODE_68291_before	***** TGGAATGGCGGAACTTGCAATGACCGTTTTCCAGGCTTCCGTGTTCTATAAGCAGAGAGATTTCTTGTTTTACCCCGCTT	***** TGGAATGGCGGAACTTGCAATGACCGTTTTCCAGGCTTCCGTGTTCTATAAGCAGAGAGATTTCTTGTTTTACCCCGCTT	2073
NODE_68291_after	TGGAATGGCGGAACTTGCAATGACCGTTTTCCAGGCTTCCGTGTTCTATAAGCAGAGAGATTTCTTGTTTTACCCCGCTT		2021
NODE_68291_before	***** GGGCTTTGGCTTGCCTATTGGGTCTCCGAAATCCGTTGTCTTTCATGGAATCAGGGATATGGATCATCTTAACATAC	***** GGGCTTTGGCTTGCCTATTGGGTCTCCGAAATCCGTTGTCTTTCATGGAATCAGGGATATGGATCATCTTAACATAC	2153
NODE_68291_after	GGGCTTTGGCTTGCCTATTGGGTCTCCGAAATCCGTTGTCTTTCATGGAATCAGGGATATGGATCATCTTAACATAC		2101
NODE_68291_before	***** TACACCATTTGGCTTCGCTCCAGCGGCCAGCAGGTTCTTCAAGCAATCTTGGCATTCTTTGGCATCCATCAGATGGCACT	***** TACACCATTTGGCTTCGCTCCAGCGGCCAGCAGGTTCTTCAAGCAATCTTGGCATTCTTTGGCATCCATCAGATGGCACT	2233
NODE_68291_after	TACACCATTTGGCTTCGCTCCAGCGGCCAGCAGGTTCTTCAAGCAATCTTGGCATTCTTTGGCATCCATCAGATGGCACT		2181
NODE_68291_before	***** GTCCCTCTTTGGTTCATTGCTGAGTTGGGAGAACTCAGGTTGTGCGCAAAACCCCTGGGAACCTTCACTTTGCTAATGG	***** GTCCCTCTTTGGTTCATTGCTGAGTTGGGAGAACTCAGGTTGTGCGCAAAACCCCTGGGAACCTTCACTTTGCTAATGG	2313
NODE_68291_after	GTCCCTCTTTGGTTCATTGCTGAGTTGGGAGAACTCAGGTTGTGCGCAAAACCCCTGGGAACCTTCACTTTGCTAATGG		2261
NODE_68291_before	***** TTTTCGTTCTTGGAGGATTTATTGTTTCAAAAAACGACATCGAGCCATGGATGATATGGGGATATTACGTATCTCCTATG	***** TTTTCGTTCTTGGAGGATTTATTGTTTCAAAAAACGACATCGAGCCATGGATGATATGGGGATATTACGTATCTCCTATG	2393
NODE_68291_after	TTTTCGTTCTTGGAGGATTTATTGTTTCAAAAAACGACATCGAGCCATGGATGATATGGGGATATTACGTATCTCCTATG		2341
NODE_68291_before	***** ATGTATGGGC AAAATGCTATAGTGATGAATGAATTCCTCGACAAAAGATGGAGCACGCGTAACGAGGATAC TAGAATTA	***** ATGTATGGGC AAAATGCTATAGTGATGAATGAATTCCTCGACAAAAGATGGAGCACGCGTAACGAGGATAC TAGAATTA	2473
NODE_68291_after	ATGTATGGGC AAAATGCTATAGTGATGAATGAATTCCTCGACAAAAGATGGAGCACGCGTAACGAGGATAC TAGAATTA		2421
NODE_68291_before	***** TGAGCCACAGTTGGAAAAAGTCTTTTGAAGTCTCGAGGTTTCTTCTGTAACAAGATATTGGTATTGGATCTGCATTGGAG	***** TGAGCCACAGTTGGAAAAAGTCTTTTGAAGTCTCGAGGTTTCTTCTGTAACAAGATATTGGTATTGGATCTGCATTGGAG	2553
NODE_68291_after	TGAGCCACAGTTGGAAAAAGTCTTTTGAAGTCTCGAGGTTTCTTCTGTAACAAGATATTGGTATTGGATCTGCATTGGAG		2501
NODE_68291_before	***** CACTGTTGGGTTTTCACTCCTCTTCAACATCTTGTTTGTGTCAGCATTGACTTGGTTAAATCCTTTGGGAGATGCAAAA	***** CACTGTTGGGTTTTCACTCCTCTTCAACATCTTGTTTGTGTCAGCATTGACTTGGTTAAATCCTTTGGGAGATGCAAAA	2633
NODE_68291_after	CACTGTTGGGTTTTCACTCCTCTTCAACATCTTGTTTGTGTCAGCATTGACTTGGTTAAATCCTTTGGGAGATGCAAAA		2581
NODE_68291_before	***** GCAGTTGCTCGGATGAAGAGGCGGATAAAGAAGAAAAACAATCAATTGCTCTTGC AACCTGCGAAAAGGAAATCGACAT	***** GCAGTTGCTCGGATGAAGAGGCGGATAAAGAAGAAAAACAATCAATTGCTCTTGC AACCTGCGAAAAGGAAATCGACAT	2713
NODE_68291_after	GCAGTTGCTCGGATGAAGAGGCGGATAAAGAAGAAAAACAATCAATTGCTCTTGC AACCTGCGAAAAGGAAATCGACAT		2661
NODE_68291_before	***** GCAAGTGAGAAAGTTCTTCTGAAATCGTTAGCATTTCAGAGAAATATACAGAGAAAGGGATGGTTCTGCCATTCCAACCC	***** GCAAGTGAGAAAGTTCTTCTGAAATCGTTAGCATTTCAGAGAAATATACAGAGAAAGGGATGGTTCTGCCATTCCAACCC	2793
NODE_68291_after	GCAAGTGAGAAAGTTCTTCTGAAATCGTTAGCATTTCAGAGAAATATACAGAGAAAGGGATGGTTCTGCCATTCCAACCC		2741
NODE_68291_before	***** TTTCTCTGCGTTCAACCATGTGAACACTACTCGTGGAATGCGCTGCAGAAAAGGAGTCAAGGAGTTGAGGAAAGCCGT	***** TTTCTCTGCGTTCAACCATGTGAACACTACTCGTGGAATGCGCTGCAGAAAAGGAGTCAAGGAGTTGAGGAAAGCCGT	2873
NODE_68291_after	TTTCTCTGCGTTCAACCATGTGAACACTACTCGTGGAATGCGCTGCAGAAAAGGAGTCAAGGAGTTGAGGAAAGCCGT		2821
NODE_68291_before	***** CTCCAACTGTTGAGAGATGTCAGTGGCGCTTTTCAGACCAAGGGTACTCACAGCATTGGTGGGGTTAGTGGTGCTGGAAA	***** CTCCAACTGTTGAGAGATGTCAGTGGCGCTTTTCAGACCAAGGGTACTCACAGCATTGGTGGGGTTAGTGGTGCTGGAAA	2953
NODE_68291_after	CTCCAACTGTTGAGAGATGTCAGTGGCGCTTTTCAGACCAAGGGTACTCACAGCATTGGTGGGGTTAGTGGTGCTGGAAA		2901



NODE_68291_before	***** GACAACCCCTCATGGATGTGCTAG	***** AGGAAGTATTAGCATCTCCGGATACCCTA	3033
NODE_68291_after	GACAACCCCTCATGGATGTGCTAGCAGGAAGGAAGACAGGTGGTTACATAGAAAGGAAGTATTAGCATCTCCGGATACCCTA		2981
NODE_68291_before	***** AAAACCAATCAACGTTTGTCTCGGGTCAGTGGTACTGTGAACAGAACGACATTCACCTCGCCTAACGTCACCTGTCTACGAA	***** AAAACCAATCAACGTTTGTCTCGGGTCAGTGGTACTGTGAACAGAACGACATTCACCTCGCCTAACGTCACCTGTCTACGAA	3113
NODE_68291_after	AAAACCAATCAACGTTTGTCTCGGGTCAGTGGTACTGTGAACAGAACGACATTCACCTCGCCTAACGTCACCTGTCTACGAA		3061
NODE_68291_before	***** TCCCTCCTATACTCAGCCTGGCTTCGTCTTTCCTCCGACATTAAGACTCAAACCTCGCAAGATGTTTGTGGAAGAAAGTTAT	***** TCCCTCCTATACTCAGCCTGGCTTCGTCTTTCCTCCGACATTAAGACTCAAACCTCGCAAGATGTTTGTGGAAGAAAGTTAT	3193
NODE_68291_after	TCCCTCCTATACTCAGCCTGGCTTCGTCTTTCCTCCGACATTAAGACTCAAACCTCGCAAGATGTTTGTGGAAGAAAGTTAT		3141
NODE_68291_before	***** GGAGTTGGTTGAGCTCAACCCATCAGAAACGCGCTTGTCTGGGCTTCTGGTGTGATGGCCTTTTCGACTGAGCAAAGAA	***** GGAGTTGGTTGAGCTCAACCCATCAGAAACGCGCTTGTCTGGGCTTCTGGTGTGATGGCCTTTTCGACTGAGCAAAGAA	3273
NODE_68291_after	GGAGTTGGTTGAGCTCAACCCATCAGAAACGCGCTTGTCTGGGCTTCTGGTGTGATGGCCTTTTCGACTGAGCAAAGAA		3221
NODE_68291_before	***** AGCGGCTGACAATAGCTGTAGAGTTGGTGGCTAATCCATCTATTATCTTTATGGACGAACCAACCTCCGGCCTTGATGCT	***** AGCGGCTGACAATAGCTGTAGAGTTGGTGGCTAATCCATCTATTATCTTTATGGACGAACCAACCTCCGGCCTTGATGCT	3353
NODE_68291_after	AGCGGCTGACAATAGCTGTAGAGTTGGTGGCTAATCCATCTATTATCTTTATGGACGAACCAACCTCCGGCCTTGATGCT		3301
NODE_68291_before	***** AGAGCAGCCGCCATCGTGTGCGTACGGTGAGGAACACGGTGGATACAGGGAGGACTGTGTTTGCACGATTCACCAGCC	***** AGAGCAGCCGCCATCGTGTGCGTACGGTGAGGAACACGGTGGATACAGGGAGGACTGTGTTTGCACGATTCACCAGCC	3433
NODE_68291_after	AGAGCAGCCGCCATCGTGTGCGTACGGTGAGGAACACGGTGGATACAGGGAGGACTGTGTTTGCACGATTCACCAGCC		3381
NODE_68291_before	***** GAGCATTGACATTTTTGAAGCTTTTGTAGAGTTGCTATTAATGAAAAGAGGCGGGCGGGTCATTTATGCTGGCCCTCTTG	***** GAGCATTGACATTTTTGAAGCTTTTGTAGAGTTGCTATTAATGAAAAGAGGCGGGCGGGTCATTTATGCTGGCCCTCTTG	3513
NODE_68291_after	GAGCATTGACATTTTTGAAGCTTTTGTAGAGTTGCTATTAATGAAAAGAGGCGGGCGGGTCATTTATGCTGGCCCTCTTG		3461
NODE_68291_before	***** GTCGCCATTCCCAAGCTCGTAGAATATTTTGAAGGCTGTCCCAGGGGTTCCGAAAGATCAGGGATGGTCACAAATCCAGCC	***** GTCGCCATTCCCAAGCTCGTAGAATATTTTGAAGGCTGTCCCAGGGGTTCCGAAAGATCAGGGATGGTCACAAATCCAGCC	3593
NODE_68291_after	GTCGCCATTCCCAAGCTCGTAGAATATTTTGAAGGCTGTCCCAGGGGTTCCGAAAGATCAGGGATGGTCACAAATCCAGCC		3541
NODE_68291_before	***** ACATGGATGCTTGAAGTGAGTCTCCGGCAGTTGAGGCTCAGCTCAGGTCGACTTCGCAGATATTACCCAACTCTGA	***** ACATGGATGCTTGAAGTGAGTCTCCGGCAGTTGAGGCTCAGCTCAGGTCGACTTCGCAGATATTACCCAACTCTGA	3673
NODE_68291_after	ACATGGATGCTTGAAGTGAGTCTCCGGCAGTTGAGGCTCAGCTCAGGTCGACTTCGCAGATATTACCCAACTCTGA		3621
NODE_68291_before	***** CCTTTTAAAGCGGAACCAAGACCTGATCAAAGAGCTTAGTACCCAGCCAGGCTGCAAAGATCTCCACTTCCCTACCG	***** CCTTTTAAAGCGGAACCAAGACCTGATCAAAGAGCTTAGTACCCAGCCAGGCTGCAAAGATCTCCACTTCCCTACCG	3753
NODE_68291_after	CCTTTTAAAGCGGAACCAAGACCTGATCAAAGAGCTTAGTACCCAGCCAGGCTGCAAAGATCTCCACTTCCCTACCG		3701
NODE_68291_before	***** AGTACTCAACCTTTCTCTACTCAGTGCAAGGCTTGTCTTGGAACAGCAGCTGGCTTACTGGAGAAATCCTCAGTAC	***** AGTACTCAACCTTTCTCTACTCAGTGCAAGGCTTGTCTTGGAACAGCAGCTGGCTTACTGGAGAAATCCTCAGTAC	3833
NODE_68291_after	AGTACTCAACCTTTCTCTACTCAGTGCAAGGCTTGTCTTGGAACAGCAGCTGGCTTACTGGAGAAATCCTCAGTAC		3781
NODE_68291_before	***** AACGCCATCCGGTTCTTTATGACCATAGTCATCCGATTTGTTTGGTTAATAATCTGGGATAAAGGACAGCAGACGAC	***** AACGCCATCCGGTTCTTTATGACCATAGTCATCCGATTTGTTTGGTTAATAATCTGGGATAAAGGACAGCAGACGAC	3913
NODE_68291_after	AACGCCATCCGGTTCTTTATGACCATAGTCATCCGATTTGTTTGGTTAATAATCTGGGATAAAGGACAGCAGACGAC		3861
NODE_68291_before	***** CAAGCAACAAGACCTGATGAATCTTTTGGGAGCCATGTACGCAGCTGTGCTTTTCTTGGGGCCACAAATGCTTCTGCTG	***** CAAGCAACAAGACCTGATGAATCTTTTGGGAGCCATGTACGCAGCTGTGCTTTTCTTGGGGCCACAAATGCTTCTGCTG	3993
NODE_68291_after	CAAGCAACAAGACCTGATGAATCTTTTGGGAGCCATGTACGCAGCTGTGCTTTTCTTGGGGCCACAAATGCTTCTGCTG		3941
NODE_68291_before	***** TGCAGTCTATAGTCGCCATTGAGAGGACAGTCTTCTACCGTGAACGAGCAGCTGGAATGTAATCTCCGCTGCCATACGCA	***** TGCAGTCTATAGTCGCCATTGAGAGGACAGTCTTCTACCGTGAACGAGCAGCTGGAATGTAATCTCCGCTGCCATACGCA	4073
NODE_68291_after	TGCAGTCTATAGTCGCCATTGAGAGGACAGTCTTCTACCGTGAACGAGCAGCTGGAATGTAATCTCCGCTGCCATACGCA		4021
NODE_68291_before	***** TTTGCTCAGGTGGCTATTGAGACAATTTATGTAGCGATTGAGACATGGTCTACAGTCTTCTCCTTACTCGATGATTGG	***** TTTGCTCAGGTGGCTATTGAGACAATTTATGTAGCGATTGAGACATGGTCTACAGTCTTCTCCTTACTCGATGATTGG	4153
NODE_68291_after	TTTGCTCAGGTGGCTATTGAGACAATTTATGTAGCGATTGAGACATGGTCTACAGTCTTCTCCTTACTCGATGATTGG		4101
NODE_68291_before	***** GTTCCA GTGGAAGGCGGGGAGGTTCTTGTGGTTCTACTACTACATACTGATGTGCTTCACTACTTCACGATGATGGAA	***** GTTCCA GTGGAAGGCGGGGAGGTTCTTGTGGTTCTACTACTACATACTGATGTGCTTCACTACTTCACGATGATGGAA	4233
NODE_68291_after	GTTCCA GTGGAAGGCGGGGAGGTTCTTGTGGTTCTACTACTACATACTGATGTGCTTCACTACTTCACGATGATGGAA		4181
NODE_68291_before	***** TGATGGTTGTAGCATTGACACCAGGCCACCAGATAGCTGCCATTGTGATGTCTTCTTCTTCCGAGCTTCTGGAACCTGTTC	***** TGATGGTTGTAGCATTGACACCAGGCCACCAGATAGCTGCCATTGTGATGTCTTCTTCTTCCGAGCTTCTGGAACCTGTTC	4313
NODE_68291_after	TGATGGTTGTAGCATTGACACCAGGCCACCAGATAGCTGCCATTGTGATGTCTTCTTCTTCCGAGCTTCTGGAACCTGTTC		4261
NODE_68291_before	***** TCTGGCTTCCTTATCCCTAGCCCGCAAATTCCTGTATGGTGGAGGTGGTATTACTGGGCTTCCAGGTGGCATGGACGCT	***** TCTGGCTTCCTTATCCCTAGCCCGCAAATTCCTGTATGGTGGAGGTGGTATTACTGGGCTTCCAGGTGGCATGGACGCT	4393
NODE_68291_after	TCTGGCTTCCTTATCCCTAGCCCGCAAATTCCTGTATGGTGGAGGTGGTATTACTGGGCTTCCAGGTGGCATGGACGCT		4341
NODE_68291_before	***** GTACGGTCTTGTCACTCTCAAGTGGCGGCAAGAAATGGCAATCTCGAAATACAGGAGCCGGCAACATGCCGTTGAAGC	***** GTACGGTCTTGTCACTCTCAAGTGGCGGCAAGAAATGGCAATCTCGAAATACAGGAGCCGGCAACATGCCGTTGAAGC	4473
NODE_68291_after	GTACGGTCTTGTCACTCTCAAGTGGCGGCAAGAAATGGCAATCTCGAAATACAGGAGCCGGCAACATGCCGTTGAAGC		4421



```

*****
NODE_68291_before AGTTCC TGAAGGTAGAAC TGGGT-----GGTTGCTCACATCGGCTGGGTCTTCTC 4553
NODE_68291_after  AGTTCC TGAAGGTAGAACTGGTTTTGACTACAGCTTCC TCCCGCTGTCGCGTTGCTCACATCGGCTGGGTCTTCTC 4501

*****
NODE_68291_before TTTTCTTTGCTTCGCTTACGGCATCAAGTTCTCAATTTCCAGAGGAGATAAAACCGATGGCAAACAGTTCTCATT 4633
NODE_68291_after  TTTTCTTTGCTTCGCTTACGGCATCAAGTTCTCAATTTCCAGAGGAGATAAAACCGATGGCAAACAGTTCTCATT 4581

*****
NODE_68291_before CTGGCTAGATTTTGAAACGTTAAACGTAGGCCATCATGTAAATTAAGGATGATAGGCCACTAAAGAGTCTCCCTCCTCC 4713
NODE_68291_after  CTGGCTAGATTTTGAAACGTTAAACGTAGGCCATCATGTAAATTAAGGATGATAGGCCACTAAAGAGTCTCCCTCCTCC 4661

*****
NODE_68291_before TGTTTTCTTCACTTTTCACTAAGTCTTGCTTTTGTAACTAGCATTCCTTTGTCACCGCTGCTTCATTGGACTGAGAGCG 4793
NODE_68291_after  TGTTTTCTTCACTTTTCACTAAGTCTTGCTTTTGTAACTAGCATTCCTTTGTCACCGCTGCTTCATTGGACTGAGAGCG 4741

** *****
NODE_68291_before TCAGTTAATTGTAAAGAGCAAATAATTAAATTTGAAATGCAAACGAGTGGTGTG 4846
NODE_68291_after  TCGGTTAATT----- 4751

```

## Appendix D

*De novo* assembled expressed gene catalog of a  
fast-growing *Eucalyptus* tree produced by Illumina  
mRNA-Seq



RESEARCH ARTICLE

Open Access

# *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq

Eshchar Mizrahi<sup>1†</sup>, Charles A Hefer<sup>2†</sup>, Martin Ranik<sup>1</sup>, Fourie Joubert<sup>2</sup>, Alexander A Myburg<sup>1\*</sup>

## Abstract

**Background:** *De novo* assembly of transcript sequences produced by short-read DNA sequencing technologies offers a rapid approach to obtain expressed gene catalogs for non-model organisms. A draft genome sequence will be produced in 2010 for a *Eucalyptus* tree species (*E. grandis*) representing the most important hardwood fibre crop in the world. Genome annotation of this valuable woody plant and genetic dissection of its superior growth and productivity will be greatly facilitated by the availability of a comprehensive collection of expressed gene sequences from multiple tissues and organs.

**Results:** We present an extensive expressed gene catalog for a commercially grown *E. grandis* × *E. urophylla* hybrid clone constructed using only Illumina mRNA-Seq technology and *de novo* assembly. A total of 18,894 transcript-derived contigs, a large proportion of which represent full-length protein coding genes were assembled and annotated. Analysis of assembly quality, length and diversity show that this dataset represent the most comprehensive expressed gene catalog for any *Eucalyptus* tree. mRNA-Seq analysis furthermore allowed digital expression profiling of all of the assembled transcripts across diverse xylogenic and non-xylogenic tissues, which is invaluable for ascribing putative gene functions.

**Conclusions:** *De novo* assembly of Illumina mRNA-Seq reads is an efficient approach for transcriptome sequencing and profiling in *Eucalyptus* and other non-model organisms. The transcriptome resource (Eucspresso, <http://eucspresso.bi.up.ac.za/>) generated by this study will be of value for genomic analysis of woody biomass production in *Eucalyptus* and for comparative genomic analysis of growth and development in woody and herbaceous plants.

## Background

Ultra-high-throughput second-generation DNA sequencing technologies from companies such as Roche (454 pyrosequencing), Illumina (sequencing by synthesis, Solexa GA) and Applied Biosystems (sequencing by ligation, SOLiD), are increasingly being used for novel exploratory genomics in small to medium-sized laboratories. "Short-read" (36 - 72 nt) technologies such as those of Illumina and Applied Biosystems have proven to be exceptionally successful in a wide variety of whole-transcriptome investigations [1-5], but most of these studies have relied on prior sequence knowledge

such as an annotated genome for qualitative and quantitative transcriptome analyses.

Genome assembly of short sequences without any auxiliary knowledge has primarily utilized 454 sequencing data, due to the longer individual read lengths of 150-400 base pairs (bp). However, short-read sequencing (Illumina GA and SOLiD) has been successfully used for *de novo* assembly of small bacterial genomes (2-5 Mbp), where 36 bp reads have been assembled [6-8] and hybrid approaches, where genomes are *de novo* assembled using a combination of reads from multiple sequencing platforms to overcome the inherent limitations of each technology, have been used to successfully assemble genomes of up to 40 Mbp [9,10]. More recently, the sequencing of the giant panda genome was demonstrated [11] using *de novo* assembly of sequence derived from a single platform (Illumina), but utilizing a

\* Correspondence: [zander.myburg@fabi.up.ac.za](mailto:zander.myburg@fabi.up.ac.za)

† Contributed equally

<sup>1</sup>Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa

Full list of author information is available at the end of the article



combination of different insert sizes, allowing assembly of an estimated 94% of the genome (2.25 Gbp). *De novo* assembly of large, highly repetitive and highly heterozygous eukaryotic genomes from short-read data remains a challenge.

In transcriptome studies, 454 pyrosequencing has proven very useful for generating ESTs representing the majority of expressed genes. This has enabled gene discovery in a variety of previously uncharacterized eukaryotic organisms with no or little *a priori* DNA sequence information [12-16]. However, relatively few published studies have attempted *de novo* assembly of whole-transcriptome sequences from short-read data such as that generated by Illumina GA or SOLiD technologies. Assembly of short (36-72 bp) read data into accurate, contiguous transcript sequences has only recently been reported [17-19] demonstrating that assembly of long, potentially full-length, transcript assemblies is indeed possible.

*Eucalyptus* tree species and hybrids presently constitute the most widely planted ( $\approx 20$  Mha) and commercially important hardwood fibre crop in the world. They are mainly utilized for timber, pulp and paper production [20]. Their fast growth rates and wide adaptability may in future allow sustainable and cost efficient production of woody biomass for bioenergy generation [21,22]. *Eucalyptus* will soon be only the second forest plantation genus (after *Populus*) for which a reference genome sequence will be completed by end 2010 [23]. To support the genome annotation effort, there is much value in having a dataset of genes with strong transcriptional evidence across a range of tissues and developmental stages. Until recently, limited amounts of *Eucalyptus* EST/unigene data were available in public databases, mainly due to the fact that commercial interests have necessitated private EST collections [24]. As of March 2010, aside from a mixed-species collection of  $\approx 56,000$  nucleotide sequences on NCBI ( $\approx 37,000$  of which are Sanger EST sequences) and which contain extensive redundancy, the largest effort to date to generate a comprehensive catalogue of expressed genes in a single *Eucalyptus* species was based on 454 sequencing of cDNA fragments from *E. grandis* trees [15]. While this study provided an excellent representation of expressed genes and gene ontology classes in *E. grandis*, the relatively short lengths of the assembled contigs (mean length of 389 bp for all contigs longer than 200 bp) meant that very few complete gene models were represented. There remains therefore a fundamental need for a high-quality expressed gene catalog for *Eucalyptus*, to support genome annotation efforts and discern authentically expressed genes from predicted gene models, as well as for future genomics research, which will include transcriptome, proteome and metabolome profiling.

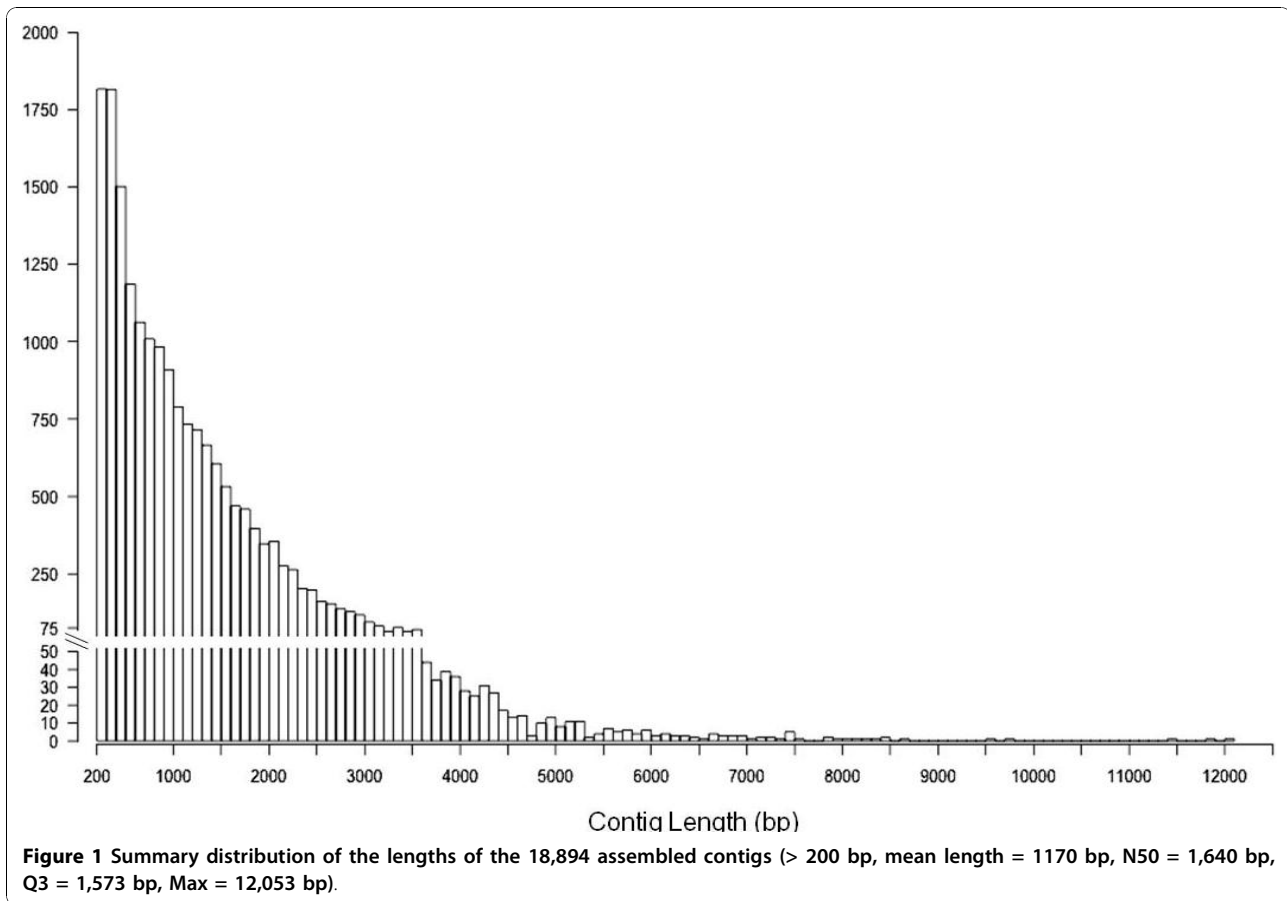
In the process of producing such a high-quality expressed gene catalog for *Eucalyptus*, we addressed three main questions: First, is it feasible to *de novo* assemble Illumina mRNA-Seq data into contiguous, near full-length gene model sequences for *Eucalyptus*? Second, what genes make up the expressed gene catalog for a fast-growing *Eucalyptus* plantation tree? Finally, can we re-use the mRNA-Seq data to create a tissue and organ-specific digital expression profile for each assembled contig? We addressed these questions by generating a comprehensive set of expressed gene sequences from a commercially grown *Eucalyptus* hybrid (*E. grandis*  $\times$  *E. urophylla*) clone using Illumina mRNA-Seq technology and *de novo* short-read assembly. We report herein the complete annotation of the expressed gene catalog based on comparative analysis with the published *Arabidopsis thaliana* [25], *Populus trichocarpa* [26] and *Vitis vinifera* [27] protein-coding datasets. We describe an interactive database of annotated transcript sequences, coding sequences (CDSs) and derived protein sequences (Eucspresso, <http://eucspresso.bi.up.ac.za/>, CA Hefer, E Mizrachi, AA Myburg, F Joubert, unpublished), which will be continuously updated and curated in association with the *Eucalyptus* Genome Network (EUCAGEN, <http://www.eucagen.org>) as part of an effort to initiate a publicly accessible database for *Eucalyptus* transcriptomics research similar to that produced for *Populus* [28].

## Results

### *De novo* assembly, validation and annotation of contigs

In total, 62 million paired-end reads of raw mRNA-Seq data (6.90 Gbp) representing poly(A)-selected RNA from six *Eucalyptus* tissues and varying in lengths from 36 bp to 60 bp, were generated in 14 lanes on Illumina GA and GAIi instruments. Following a sequence filtering process to exclude low quality and ribosomal RNA-derived reads, we assembled 36 million paired-end reads (3.93 Gbp, Additional file 1 - Table S1 and Figure S1, NCBI Sequence Read Archive accession SRA012408) of non-normalized mRNA sequence, using the Velvet short-read assembler (version 0.7.30, [29]). In total, 18,894 RNA-derived contigs were assembled (comprising 22.1 Mbp of transcriptome sequence) that were greater than 200 bp in length (mean = 1170 bp, Figure 1 and Additional file 2), with a median coverage per base (CPB) per contig of 37 $\times$ , ranging from 8 $\times$  (minimum coverage cut-off for assembly) to 5,262 $\times$  (Additional file 1 Figure S2).

We performed *ab initio* CDS prediction using GENSCAN [30] and found that 15,713 contigs (83.2%) contained a predicted CDS (Additional file 1 Table S3). Analysis of the predicted coding sequences using Anacoda [31] identified 6,208 contigs that contained

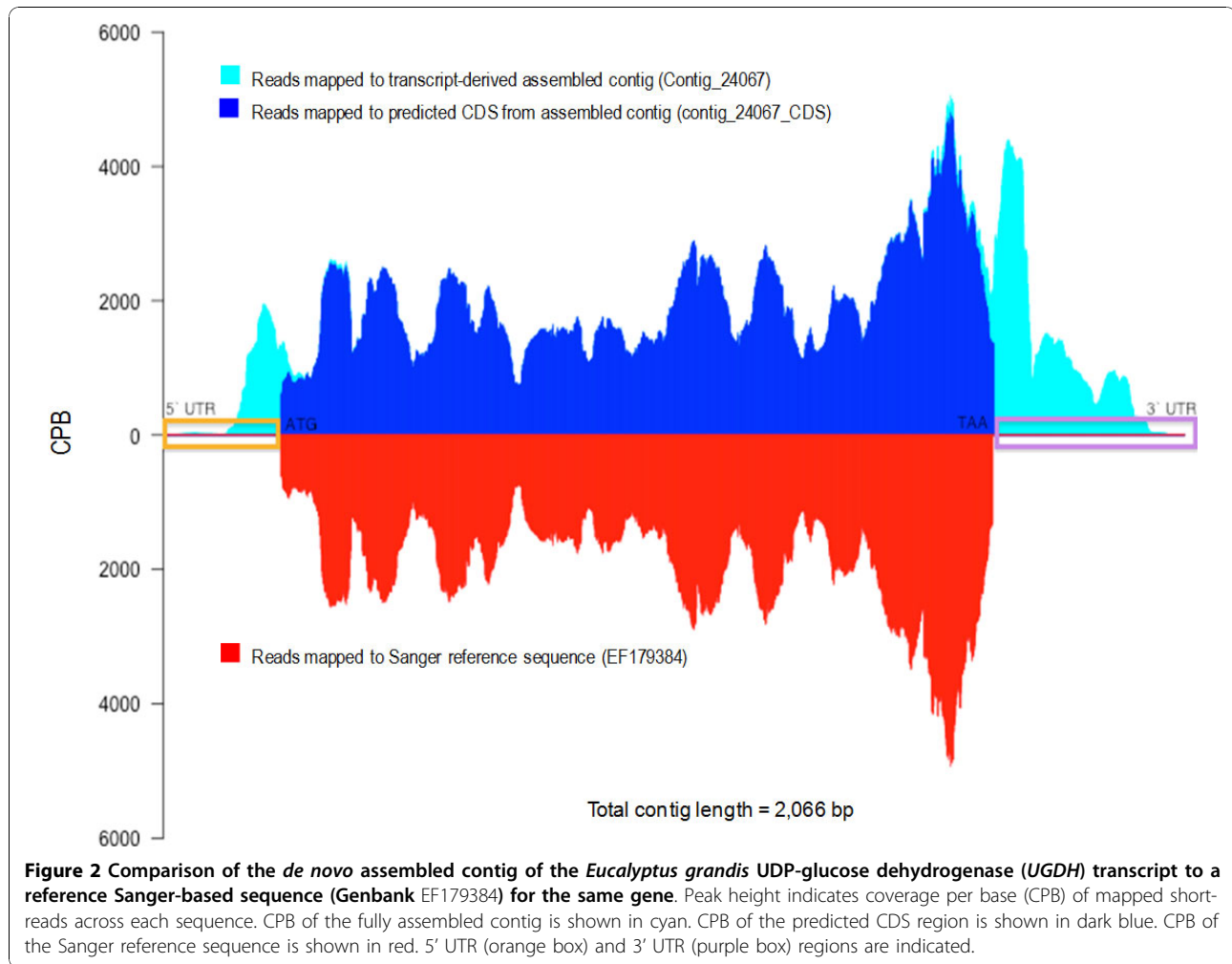


putatively full-length CDSs (i.e. containing start and stop codons), 4,610 predicted to contain a start but no stop codon, 4,874 predicted to contain a stop but no start codon, and only 21 with neither. To ascertain the quality of Velvet assembly of short reads into long contiguous coding sequences, we compared a subset of 35 of our transcript-derived contigs to corresponding Sanger-sequenced, full-length, cloned *Eucalyptus grandis* mRNA sequences in NCBI (Figure 2 and Additional file 3). Paired reads were independently mapped to each Sanger reference sequence, the *de novo* assembled Velvet contig and its corresponding predicted CDS. A Needleman-Wunsch alignment of these three sequences was used for contiguity validation of the assembled contigs. Independently, each sequence had 100% coverage validation across the contig, except in cases of low quality assembly ('N's inserted by Velvet), which occurred in regions of coverage lower than 8x per base. Of the 35 transcript-derived contigs evaluated, 25 (71%) assembled completely with a 5' UTR, 3' UTR, as well as a contiguous coding sequence matching that of the reference mRNA sequence. We found several cases where, despite high coverage, our transcript-derived contigs differed from the Sanger reference sequence due to indels, but

these were generally in the UTR regions and likely represent allelic differences between the F1 hybrid individual and the reference sequences (Additional file 3).

Of the 18,894 assembled contigs, 18,606 (98.48%) exhibited significant similarity (BLASTN, -10, [31]) to the preliminary draft 8X DOE-JGI *E. grandis* genome assembly (<http://eucalyptusdb.bi.up.ac.za/>) consistent with the origin of the mRNA contigs (an F1 hybrid of *E. grandis* and *E. urophylla*). We further characterized the assembled contigs by high stringency BLASTX analysis (-10 confidence, minimum 100 bp high scoring pair (HSP) match length) to protein datasets from three reference sequenced angiosperm genera (*Arabidopsis*, *Populus* and *Vitis*). Cumulatively, 15,055 contigs (79.68%) exhibited high similarity to *Arabidopsis* (14,235 contigs), *Populus* (14,769 contigs) or *Vitis* proteins (14,833 contigs, Additional file 1 Figure S3). Of the 15,055 contigs with high similarity to *Arabidopsis*, *Populus* or *Vitis* proteins, 13,806 (91.70%) also contained predicted coding sequences (Figure 3A), while 1,249 (8.30%) did not (Figure 3B), possibly due to low expression of these transcripts which would have resulted in lower coverage and shorter contigs that represented only a fraction of the open reading frame (or mostly





**Figure 2** Comparison of the *de novo* assembled contig of the *Eucalyptus grandis* UDP-glucose dehydrogenase (UGDH) transcript to a reference Sanger-based sequence (Genbank EF179384) for the same gene. Peak height indicates coverage per base (CPB) of mapped short-reads across each sequence. CPB of the fully assembled contig is shown in cyan. CPB of the predicted CDS region is shown in dark blue. CPB of the Sanger reference sequence is shown in red. 5' UTR (orange box) and 3' UTR (purple box) regions are indicated.

UTR sequence). Predicted codon usage and amino acid frequencies in the proteome represented by the *Eucalyptus* expressed gene catalog were very similar to those of expressed gene catalogs from *Arabidopsis* and *Populus* (Additional file 1 Figure S4 and Figure S5).

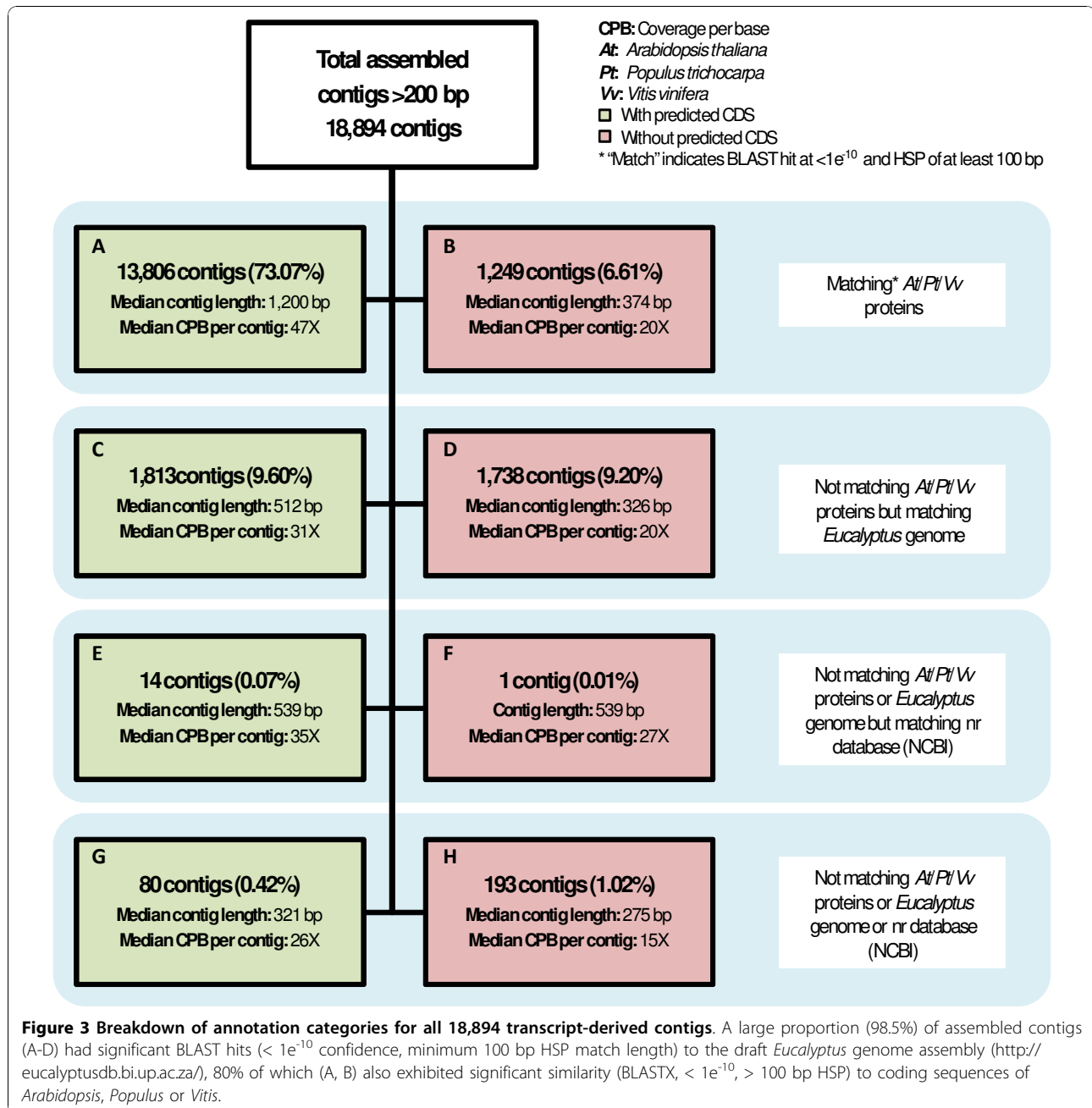
To compare the completeness of our expressed gene catalogue to that of all publicly available gene sequence data for *Eucalyptus*, we generated a separate dataset, termed EucALL, containing all publicly available *Eucalyptus* gene sequence data to date (March 2010). This included all NCBI unigenes and ESTs, assembled 454 EST data from *E. grandis* leaf tissue (DOE-JGI, <http://eucalyptusdb.bi.up.ac.za/>), assembled 454 EST data produced by Novaes and colleagues [15], and the EucWood contig dataset [33]. We compared the representation of *Arabidopsis* genes in the EucALL dataset and in our assembled *E. grandis* × *E. urophylla* (EGU) transcript dataset by BLASTX at significance levels of  $< 1e^{-05}$ ,  $< 1e^{-10}$  and  $< 1e^{-20}$  (Additional file 1 Table S2). While the overall numbers of hits were

higher in the EucALL dataset, these were mostly in the lower size ranges. For our *de novo* assembled contigs, a much higher number of significant hits in contigs larger than 2000 bp in size (6,602 compared to 1,940 at significance  $< 1e^{-10}$ ) suggested that a greater proportion of our contigs represent full-length gene models than the publicly available *Eucalyptus* gene sequence set (EucALL).

#### Functional annotation of the expressed gene catalog

The transcript-derived contig sequences were annotated according to several functional annotation conventions, including Gene Ontology (GO - <http://www.geneontology.org/>), KEGG (<http://www.genome.jp/kegg/>) and InterProScan (<http://www.ebi.ac.uk/Tools/InterProScan/>). The numbers and assortment of allocated GO categories provides a good indication of the large diversity of expressed genes sampled from the *Eucalyptus* transcriptome (Figure 4). This was also reflected in the diversity of InterProScan categories identified (Additional file 1 Figure S6 and Figure S7), as well as the



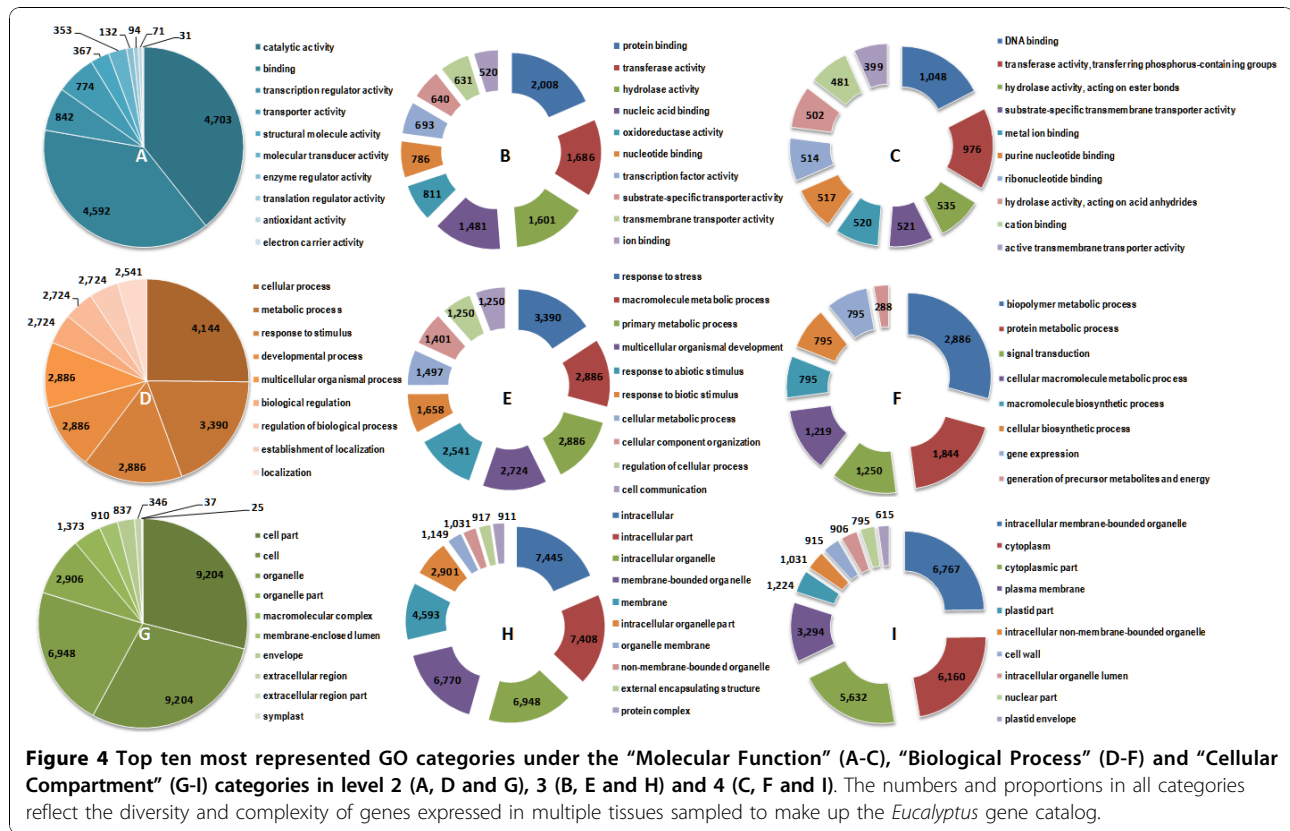


comprehensive coverage of biochemical processes by KEGG annotation, which was similar to that of the entire *Arabidopsis* gene catalog (Additional file 1 Figure S8).

#### Digital expression profiling

An accepted method of identifying large scale differences in gene expression is to use EST abundance as an indicator of transcript abundance. This method has been implemented and validated in numerous studies using Sanger-derived ESTs [34,35], as well as 454-

pyrosequencing methods [13,36-39]. Quantitative transcriptome analysis using ultra-high-throughput sequencing technologies such as Illumina and SOLiD has been shown to be accurate and highly correlated with other quantitative methods such as RT-qPCR and microarray analysis [1,5]. To quantify tissue-specific transcript abundance reflected in our short-read dataset, we combined data (multiple lanes in most cases) generated from the same tissues and mapped six tissue-specific datasets (Additional file 1 Table S1) to the assembled gene catalog using Bowtie [40]. Following this, we used



the Cufflinks [41] program (<http://cufflinks.cbcb.umd.edu>), which provides relative abundance values by calculating Fragments Per Kilobase of exon per Million fragments mapped (FPKM) as validated previously [2]. This enabled the allocation of a tentative digital expression profile for each transcript-derived contig (Additional file 4).

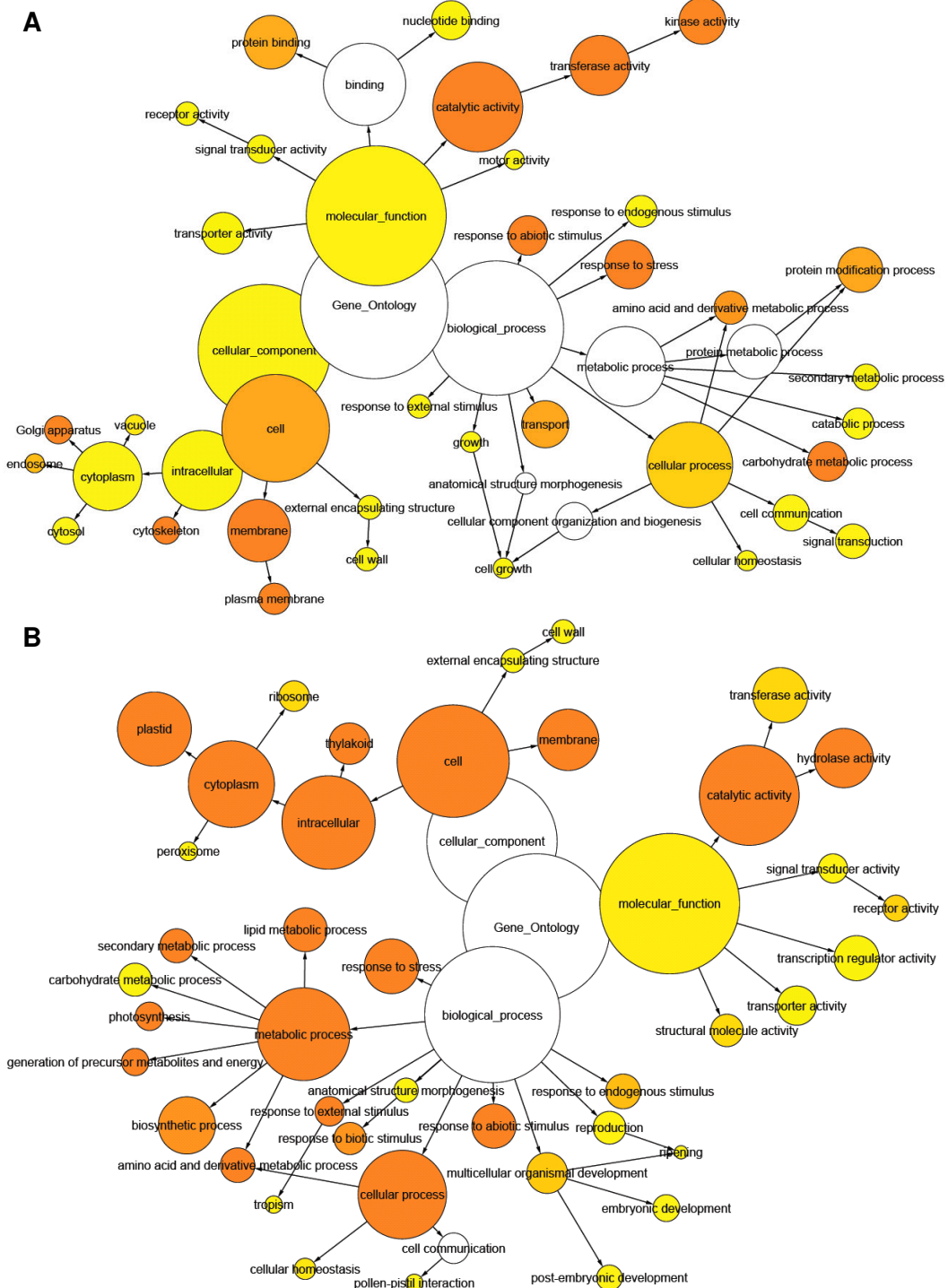
To compare between two general tissue types that are of interest for woody biomass production, we evaluated groups of genes whose FPKM values were greater than two-fold higher in woody (xylogenic) tissues (average FPKM of immature xylem and xylem: 1,897 annotated contigs) or leaf (non-xylogenic) tissues (average FPKM of shoot tips, young leaves and mature leaves: 1,531 annotated contigs). GO categories over-represented in the xylem-upregulated set compared to the leaf set (Figure 5A) was representative of developing woody tissues, with significant enrichment ( $p < 0.05$ ) in signalling (“kinase activity”), carbohydrate metabolism, and genes associated with the Golgi, cytoskeleton and the plasma membrane - consistent with an emphasis on delivery of biopolymers to the cell wall. In contrast, gene categories significantly enriched ( $p < 0.05$ ) in leaf tissue compared to woody tissue (Figure 5B) were associated with photosynthesis (“plastid”, “thylakoid”, “photosynthesis”),

growth and energy production (precursor metabolites, “lipid biosynthesis”, “amino acid metabolism”).

We also interrogated our transcriptome data using the “core xylem gene set” identified in *Arabidopsis* by Ko and colleagues [42]. Of the 52 genes identified by the authors as markers of secondary xylem formation in *Arabidopsis*, 33 had putative homologues in the *Eucalyptus* transcriptome (BLASTX,  $< 1e^{-10}$ ) and in total 43 contigs were identified. Of these, 40 (93%) showed greater than two-fold “Xylem” to “Leaf” digital expression profile ratios and six were only detected in xylem tissues (Additional file 1 Table S4). Most of the expression profiles were also highly correlated with that of secondary cell wall-specific *Eucalyptus* cellulose synthase genes, similar to the patterns previously observed in *Arabidopsis*. These results are comparable to the 80% (51 out of 63 genes) reported recently for the same set of *Arabidopsis* homologs in *Populus* [43], which provided further support for the biological validity of the short-read-based digital expression profiles associated with the *Eucalyptus* expressed gene catalog.

#### Public data resource

We constructed a public data resource, Eucpresso (<http://eucpresso.bi.up.ac.za>), which provides a



**Figure 5 Over-represented GO categories in xylem (A - 1,897 annotated contigs) and leaf (B - 1,531 annotated contigs) tissues.** All genes with a FPKM value more than two-fold higher in one tissue type versus the other were considered for the analysis. Data were analyzed using BiNGO (Maere et al. 2005). Node size is proportional to the number of genes in each category and colors shaded according to significance level (white - no significant difference, yellow - FDR = 0.05, Orange - FDR < 0.05).



searchable interface to the assembled contigs. The database can be queried based on closest homologous entry in the *Arabidopsis thaliana* (TAIR9), *Populus trichocarpa* (Version 2.0) and *Vitis vinifera* (Sept 2009 build) sequence data sets. Simple and compound keyword searches can be performed based on all of the functional annotation terms and the predicted coding and protein sequences can be obtained for all contigs. Finally, the tissue-specific (FPKM) digital expression profile and the location of each contig in the draft 8X *E. grandis* genome assembly (<http://eucalyptusdb.bi.up.ac.za/>) can be viewed from within Eucspresso.

## Discussion

We have assembled nearly 19,000 expressed gene sequences from xylogenic and non-xylogenic tissues of an actively growing *Eucalyptus* plantation tree using only Illumina mRNA-Seq technology and *de novo* short-read assembly. Quality control comparisons to full-length, cloned, Sanger-derived transcript sequences from *Eucalyptus*, as well as multiple lines of evidence such as CDS prediction and Pfam prediction showed that the transcript assemblies are robust and that thousands of full-length coding sequences and their respective 5' and/or 3' UTR regions were successfully assembled. Comparison of assembled gene models to gene catalogs of other angiosperm species by BLAST analysis and functional annotation (GO, InterProScan and KEGG category numbers and proportions, Figure 4 and Additional file 1 - Figure S6, Figure S7 and Figure S8) indicate that we have sampled an expansive and diverse expressed gene catalog representing a large proportion of the genes expressed in mature *Eucalyptus* trees across a variety of woody and non-woody tissues. Comparison to all publicly available *Eucalyptus* DNA sequence suggests that we have sampled a more comprehensive set of genes, which is also more complete in length (Additional file 1 - Table S2) from a single eucalypt tree genotype than has been available to date for the entire genus. Additionally, using a validated approach to quantify mRNA-Seq data we have produced an informative database of transcript abundance across six *Eucalyptus* tree tissues, which, due to the depth of sequencing, results in higher sensitivity and wider dynamic range than Sanger or 454-derived EST counts usually associated with this type of analysis.

A concern associated with *de novo* assembly of transcript sequences, be it Sanger derived [33] or 454 sequence derived [15] assemblies, is the contiguity of assembled sequences. This concern intuitively increases as the read length decreases, and may be one of the main reasons why most transcriptome *de novo* assembly approaches have utilized technologies with longer read lengths to date. We provide several lines of evidence

which jointly support the contiguity of transcript sequences assembled in our study using short-read data. First, a high proportion of the contigs exhibited high-confidence BLASTX similarity to protein sequences from annotated gene catalogs of three angiosperm species *Arabidopsis*, *Populus* and *Vitis* (Figure 3). Second, a large proportion of the contigs contained long, near full-length, predicted CDSs (Figure 3). Third, InterProScan analysis predicted 45,687 protein domains, which is indicative of contiguous, in-frame predicted protein sequences (Additional File 1). Finally, a random subset of the contigs, which represented a variety of length and read coverage, were validated by direct alignment to previously published, Sanger sequenced, full-length *Eucalyptus* genes that were directly cloned from cDNA (Additional File 3).

Assigning biological significance to *de novo* assembled contigs should be approached with caution. In our study, 13,806 assembled gene models (73.07% of the total assembled contigs, Figure 3A) were considered high confidence annotations due to the presence of a significant high stringency BLAST hit in other angiosperm species, as well as a predicted CDS. These contigs had relatively high coverage per base (CPB) values (median 47X) as compared to contigs lacking a predicted CDS (median CPB of 20X or lower, Figure 3B and 3D and Supplemental Table S3). Thus, a lack of CDS prediction was generally associated with low gene expression level and low CPB, which resulted in 'N's inserted by Velvet in the contig sequences (Figure 3B and 3D and Supplemental Table S3). The assembly quality and annotation of these sequences could be improved in future by even deeper sequencing and the addition of data from new tissue types. Another possible source of error is the spurious prediction of CDSs in long, non-coding RNAs, which has been previously shown to occur [44,45]. It is notable that of the 1,813 *Eucalyptus*-derived contigs with no significant BLAST hit to other angiosperms, but containing a predicted CDS (Figure 3C), only 81 contigs had predicted InterProScan domains. Additionally, the median CDS to contig length ratio was 0.33, as compared to 0.62 in the 13,806 high confidence contigs in Figure 3A, which suggests that many of these CDS predictions may be false positives. *De novo* assembled transcriptome datasets lack the ability to distinguish and classify the lower confidence annotations, an exercise that is beyond the scope of this study, albeit one that can be resolved once a genome-based predicted set of gene models is available.

Validation of the digital expression (FPKM) profiles using the "core xylem gene set" identified in *Arabidopsis* [42] has precedence in similar investigations in conifers [46], cotton [47] and poplar [43]. This analysis, combined with the results shown in Figure 5A and Figure





5B, lend support to the biological significance of digital expression profiles derived from short-read sequencing technology, which will assist in the discovery and annotation of novel *Eucalyptus* genes - and using the genome sequence, promoters - playing key roles in growth and development, and particularly in woody biomass production. The Eucpresso online resource produced from this study, as well as future comparative analysis with other woody species such as *Vitis* and *Populus*, will be valuable for studying the unique biology of woody perennials.

## Conclusions

Taking into consideration the number, length, coverage and quality of assembled gene models, as well as their digital expression profiles, this dataset surpasses several previous *de novo* transcriptome assemblies using Illumina [17,18] or 454 technology [13-16]. This can primarily be attributed to the amount of data generated (3.93 Gbp of non-rRNA derived reads), the diversity of tissues sampled and strategy of paired-end sequencing, as well as read-length (mostly 50-60 bp, compared to only 36 bp in earlier studies). Our dataset was generated using several generations of Illumina GA technology, but considering the current throughput of Illumina sequencing (up to 100 Gbp per flowcell), a gene catalog of this scale can now be produced using a single lane of Illumina mRNA-Seq. Finally, non-normalized short-read data will be extremely useful for downstream applications such as digital gene expression profiling and detection of alternative transcript structure, once reference models are available from the genome.

## Methods

### Plant tissue collection

Tissues from a six-year-old ramet of a commercially grown *E. grandis* × *E. urophylla* hybrid clone (GUSAP1, Sappi Forestry, Kwambonambi, South Africa) were collected in a clonal field trial and immediately frozen in liquid nitrogen, as previously described by Ranik and Myburg [48]. The following tissues were sampled from approximately breast height (1.35 m) on the main stem following bark removal: immature xylem (outer glutinous 1-2 mm layer comprising early developing xylem tissue) and xylem (after removal of the immature xylem layer, 2-mm-deep planing including xylem cells in advanced stages of maturity). Early developing phloem tissue including small amounts of cambial cells was collected by scraping the first 1-2 mm layer from the inner surface of the bark. Additionally, we sampled shoot tips (soft green termini of young crown tip branches containing shoot primordia and apical meristems), young leaves (rapidly-growing leaves in the process of unfolding) and mature leaves (older, fully expanded leaves of the current growth season).

### Paired-end mRNA-Seq library preparation and sequence generation

Total RNA was extracted from the six tissues using the protocol described previously [49]. Total RNA quality and concentration were determined using the Agilent RNA 6000 Pico kit (Agilent, Santa Clara, CA) on a 2100 Bioanalyzer (Agilent). Enrichment of polyA<sup>+</sup> RNA was performed using the Oligotex midi kit (Qiagen, Valencia, CA). Two hundred nanograms of polyA<sup>+</sup> RNA were fragmented in 1× RNA fragmentation solution (Ambion, Austin, TX) at 70°C for 5 minutes. The fragmented RNA was precipitated with three volumes of ethanol and re-dissolved in water. Double-stranded cDNA was synthesized using the cDNA Synthesis System (Roche, Indianapolis, IN) according to manufacturer's instructions using random hexamers (Invitrogen, Carlsbad, CA) to prime the first strand cDNA synthesis. Paired-end libraries with approximate average insert lengths of 200 base pairs were synthesized using the Genomic Sample Prep kit (Illumina, San Diego, CA) according to manufacturer's instructions. Prior to cluster generation, library concentration and size were assayed using the Agilent DNA1000 kit (Agilent) on a 2100 Bioanalyzer (Agilent). Libraries were sequenced on a Genome Analyzer equipped with a paired-end module (versions I, II and Iix, Illumina).

### *De novo* assembly of mRNA-Seq data

After removing sequences containing low quality bases ('N's) or single base repeats and ribosomal RNA sequences, the 3.93 Gbp dataset was used for assembly and subsequent coverage per base (CPB) estimation for each assembled contig. We assembled the filtered Illumina paired-end (PE) reads using Velvet version 0.7.30 [29]. Previous studies [1-3,50] have demonstrated that mRNA-Seq technology produces uneven coverage over a transcript, which prompted us to follow a coverage-assisted reference assembly strategy. Using Mosaik (<http://bioinformatics.bc.edu/marthlab/Mosaik>) to align the filtered Illumina PE sequences to the assembled contigs, the average coverage per contig was calculated. A custom script was then developed to extract the pairs of sequences that mapped to each contig, and using that contig as a template, each contig was re-assembled using Velvet with the associated expected coverage parameter set to the Mosaik average coverage value for that contig.

### Contig validation

The degree to which the assembled contigs represented long, contiguous RNA transcript sequences, was evaluated by aligning 35 Velvet contigs and their respective predicted CDSs to full-length, cloned, Sanger-derived *Eucalyptus* reference sequences present in NCBI. CPB was calculated for the sequences using BWA [51] and a



global pairwise alignment of the sequences was performed using the Needle package from EMBOSS [52]. Plots were constructed from the alignments with the CPB on the y-axis of the plot. Zero coverage values were assigned to gaps in the alignments. This revealed where gaps and/or potentially misassembled regions were present in the assembled contigs, and to what depth these contigs were sequenced.

#### Coding sequence prediction

Coding sequence predictions were performed using GENSCAN [30] and AUGUSTUS [53], predicting 15,713 and 15,904 proteins respectively. The difference in coding sequences predicted could be attributed to the different training data sets used and inherent difficulty of predicting coding sequences from incomplete genomic sequences. The GENSCAN results (15,713 predicted proteins) were used in downstream analyses.

#### Annotation of assembled contigs

Homology searches were performed against public sequence databases. The newest versions as of February 2010 of the protein sequences of *Arabidopsis* (TAIR 9), *Vitis* (Sept 2009 build) and *Populus* (version 2.0, Phytozome) were used to construct the individual BLAST datasets. The *Eucalyptus* public dataset (EucAll) consisted of 45,442 entries in Genbank (downloaded March 2010), 13,930 entries from the *Eucalyptus* Wood unigenes and ESTs [33], *E. grandis* leaf tissue ESTs (120,661 entries from DOE-JGI-produced 454 sequences, <http://eucalyptusdb.bi.up.ac.za/>) and 190,106 Unigenes and singlets from *E. grandis* 454 data [15]. The BLAST e-value threshold was set at  $1e^{-10}$ , with a minimum alignment length of 100 nucleotides (33 amino acids). Functional annotation (GO and KEGG) was performed using BLAST2GO [54], using the default annotation parameters (BLAST e-value threshold of  $1e^{-06}$ , Gene Ontology annotation threshold of 55). InterPro annotations were performed using InterProScan (<http://www.ebi.ac.uk/Tools/InterProScan/>).

#### Coverage and FPKM determination

Sequence depth and base coverage were calculated using BWA (Lin et al. 2009) and the FPKM values estimated by aligning the Illumina reads to the assembled transcriptome using Bowtie [40] and estimating the expression level of each predicted transcript (FPKM value) using Cufflinks (<http://cufflinks.cbc.umd.edu>) [41].

#### Additional material

Additional file 1: Supplemental Tables S1-S3 and Supplemental Figures S1-S8 referred to in text.

Additional file 2: FASTA formatted sequences of all 18,894 assembled contigs.

Additional file 3: Contig validation, Needleman-Wunsch alignment figures.

Additional file 4: Table containing all 18,894 contig names and calculated FPKM values for six tissues (immature xylem, xylem, phloem, shoot-tips, young leaves and mature leaves). Eucpresso (<http://eucpresso.bi.up.ac.za/>) - Online database with mRNA contig sequences and their Blast, GO, KEGG, Pfam annotations. The short-read sequence data have been submitted to the NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) under accession SRA012408.

#### Acknowledgements

The authors would like to acknowledge J. Rees and J.-M. Celton of the University of the Western Cape (Cape Town, South Africa) for assistance with Illumina GA sequencing. Plant materials were kindly provided by Sappi Forestry (Kwambonambi, South Africa). This work was supported through a strategic research grant from the South African Department of Science and Technology (DST) and by research funding from Sappi and Mondi, through the Wood and Fibre Molecular Genetics (WFMG) Programme, the Technology and Human Resources for Industry Programme (THRIP) and the National Research Foundation (NRF) of South Africa.

#### Author details

<sup>1</sup>Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa. <sup>2</sup>Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Pretoria, 0002, South Africa.

#### Authors' contributions

EM drafted the manuscript, helped sample the material, prepared the libraries, participated in the *de novo* assembly and data analysis, and helped design Eucpresso. CAH performed the *de novo* assembly and automated annotation, participated in data analysis, designed the database Eucpresso, and helped draft the manuscript. MR prepared the libraries, helped sample the material and participated in data analysis. FJ participated in data analysis and the design of Eucpresso. AAM conceived of the study, and participated in its design and coordination and helped to draft the manuscript and participated in data analysis, and helped design Eucpresso. It is the authors' opinion that EM and CAH contributed equally as first authors to this manuscript. All authors have read and approved the final version of the manuscript.

Received: 30 May 2010 Accepted: 1 December 2010

Published: 1 December 2010

#### References

1. Cloonan N, Forrest ARR, Kollé G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5**(7):613-619.
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621-628.
3. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing.** *Science* 2008, **320**(5881):1344-1349.
4. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al: **mRNA-Seq whole-transcriptome analysis of a single cell.** *Nat Methods* 2009, **6**(5):377-382.
5. Wilhelm BT, Landry JR: **RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing.** *Methods* 2009, **48**(3):249-257.
6. Farrer RA, Kemen E, Jones JDG, Studholme DJ: **De novo assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads: RESEARCH LETTER.** *FEMS Microbiol Lett* 2009, **291**(1):103-111.



7. Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J: **De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer.** *Genome Res* 2008, **18**(5):802-809.
8. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ: **Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.** *Nat Methods* 2009, **6**(4):291-295.
9. DiGiustini S, Liao N, Platt D, Robertson G, Seidel M, Chan S, Docking TR, Birol I, Holt R, Hirst M: **De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data.** *Genome Biology* 2009, **10**(9).
10. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC: **De novo Assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis.** *PLoS Genet* 2010, **6**: e1000891.
11. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2010, **463**(7279):311-317.
12. Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM: **Shedding light on an extremophile lifestyle through transcriptomics.** *New Phytol* 2009, **183**(3):764-775.
13. Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL: **Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*.** *BMC Genomics* 2009, **10**(234).
14. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **10**(219).
15. Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome.** *BMC Genomics* 2008, **9**(312).
16. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17**(7):1636-1647.
17. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, et al: **De novo transcriptome assembly with ABySS.** *Bioinformatics* 2009, **25**(21):2872-2877.
18. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A: **Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics.** *Mol Biol Evol* 2009, **26**(12):2731-2744.
19. Wu T, Qin Z, Zhou X, Feng Z, Du Y: **Transcriptome profile analysis of floral sex determination in cucumber.** *J Plant Physiol* 2010, **167**(11):905-913.
20. Eldridge K, Davidson J, Harwood C, van Wyk G: *Eucalypt domestication and breeding* Oxford: Clarendon Press; 1993.
21. FAO: **Forests and Energy.** *FAO Forestry Paper No* 2008, **154**, (Rome):(ISBN 978-992-975-105985-105982).
22. Hincee M, Rottmann W, Mullinax L, Zhang C, Chang S, Cunningham M, Pearson L, Nehra N: **Short-rotation woody crops for bioenergy and biofuels applications.** *In Vitro Cell Dev Biol - Plant* 2009, **45**(6):619-629.
23. Myburg AA, Grattapaglia D, Tuskan GA, Schmutz J, Barry K, Bristow J, The Eucalyptus Genome Network: **Sequencing the *Eucalyptus* genome: Genomic resources for renewable energy and fiber production.** *Plant & Animal Genome XVI Conference: January 12-16, 2008; San Diego, CA* 2008.
24. Hibino T: **"Post-genomics" research in *Eucalyptus* in the near future.** *Plant Biotechnol* 2009, **26**(1):109-113.
25. Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin X, et al: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**(6814):796-815.
26. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam M, Ralph S, Rombauts S, Salamov A, et al: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313**(5793):1596-1604.
27. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449**(7161):463-467.
28. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S: **The *Populus* Genome Integrative Explorer (PopGenIE): A new resource for exploring the *Populus* genome.** *New Phytol* 2009, **182**(4):1013-1025.
29. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821-829.
30. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**(1):78-94.
31. Pinheiro M, Afreixo V, Moura G, Freitas A, Santos MAS, Oliveira JL: **Statistical, computational and visualization methodologies to unveil gene primary structure features.** *Methods Inf Med* 2006, **45**(2):163-168.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
33. Rengel D, Clemente HS, Servant F, Ladouce N, Paux E, Wincker P, Couloux A, Sivadon P, Grima-Pettenati J: **A new genomic resource dedicated to wood formation in *Eucalyptus*.** *BMC Plant Biol* 2009, **9**(36).
34. Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneras S, et al: **Poplar carbohydrate-active enzymes. Gene identification and expression analyses.** *Plant Physiol* 2006, **140**(3):946-962.
35. Pavy N, Laroche J, Bousquet J, Mackay J: **Large-scale statistical analysis of secondary xylem ESTs in pine.** *Plant Mol Biol* 2005, **57**(2):203-224.
36. Hale MC, McCormick CR, Jackson JR, DeWoody JA: **Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): The relative merits of normalization and rarefaction in gene discovery.** *BMC Genomics* 2009, **10**(203).
37. Kristiansson E, Asker N, Förlin L, Joakim DGJ: **Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing.** *BMC Genomics* 2009, **10**(345).
38. Schwarz D, Robertson HM, Feder JL, Varala K, Hudson ME, Ragland GJ, Hahn DA, Berlocher SH: **Sympatric ecological speciation meets pyrosequencing: Sampling the transcriptome of the apple maggot *Rhagoletis pomonella*.** *BMC Genomics* 2009, **10**(633).
39. Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB: **Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing.** *Plant Physiol* 2007, **144**(1):32-42.
40. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009, **10**(3).
41. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**(5):511-515.
42. Ko JH, Beers EP, Han KH: **Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*.** *Mol Genet Genomics* 2006, **276**(6):517-531.
43. Dharmawardhana P, Brunner AM, Strauss SH: **Genome-wide transcriptome analysis of the transition from primary to secondary stem development in *Populus trichocarpa*.** *BMC Genomics* 2010, **11**(1):150.
44. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES: **Distinguishing protein-coding and noncoding genes in the human genome.** *Proc Natl Acad Sci USA* 2007, **104**(49):19428-19433.
45. Dinger ME, Pang KC, Mercer TR, Mattick JS: **Differentiating protein-coding and noncoding RNA: Challenges and ambiguities.** *PLoS Comput Biol* 2008, **4**(11):1-5.
46. Pavy N, Boyle B, Nelson C, Paule C, Ciguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H, et al: **Identification of conserved core xylem gene sets: Conifer cDNA microarray development, transcript profiling and computational analyses.** *New Phytol* 2008, **180**(4):766-786.
47. Betancur L, Singh B, Rapp RA, Wendel JF, Marks MD, Roberts AW, Haigler CH: **Phylogenetically distinct cellulose synthase genes support secondary wall thickening in *Arabidopsis* shoot trichomes and cotton fiber.** *J Integr Plant Biol* 2010, **52**(2):205-220.
48. Ranik M, Myburg AA: **Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis.** *Tree Physiol* 2006, **26**(5):545-556.
49. Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees.** *Plant Mol Biol Report* 1993, **11**(2):113-116.
50. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*.** *Cell* 2008, **133**(3):523-536.
51. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754.
52. Rice P, Longden I, Bleasby A: **EMBOSS: the European molecular biology open software suite.** *Trends Genet* 2000, **16**(6):276-277.
53. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19**(SUPPL 2):ii215-ii225.



54. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO:** A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, **21**(18):3674-3676.

doi:10.1186/1471-2164-11-681

**Cite this article as:** Mizrachi *et al.*: *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* 2010 **11**:681.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)





## Bibliography

- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B. and Moreno, R. F. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project *Science* **252**, 5013, 1651–6. 3, 82
- AGI, The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana* *Nature* **408**, 6814, 796–815. 16
- Ahn, S.-M., Kim, T.-H., Lee, S., Kim, D., Ghang, H., Kim, D.-S., Kim, B.-C., Kim, S.-Y., Kim, W.-Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S., Kim, C. G., Cha, J.-Y., Kim, K.-H., Lee, B., Bhak, J. and Kim, S.-J. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**, 9, 1622–1629 ISSN 1549-5469 (Electronic); 1088-9051 (Linking). 18
- Akhunov, E., Nicolet, C. and Dvorak, J. (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay *Theor Appl Genet* **119**, 3, 507–17. 18
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 3, 403–410. 30, 46, 48, 87
- Andersson-Gunnerås, S., Mellerowicz, E. J., Love, J., Segerman, B., Ohmiya, Y., Coutinho, P. M., Nilsson, P., Henrissat, B., Moritz, T. and Sundberg, B. (2006) Biosynthesis of cellulose-enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis *Plant J* **45**, 2, 144–65. 108
- Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P. and Nordborg,

- M. (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes *PLoS Genet* **1**, 5, e60. 18
- Arnaiz, O., Goût, J.-F., Bétermier, M., Bouhouche, K., Cohen, J., Duret, L., Kapusta, A., Meyer, E. and Sperling, L. (2010) Gene expression in a paleopolyploid: a transcriptome resource for the ciliate *Paramecium tetraurelia* *BMC Genomics* **11**, 547. 79
- Aury, J.-M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F. and Wincker, P. (2008) High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies *BMC Genomics* **9**, 603. 15
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. and Johnson, E. A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, 10, e3376. 19
- Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L. and Schnable, P. S. (2007) SNP discovery via 454 transcriptome sequencing *Plant J* **51**, 5, 910–8. 19
- Barbazuk, W. B., Fu, Y. and McGinnis, K. M. (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges *Genome Res* **18**, 9, 1381–92. 23
- Batzoglou, S. (2005) *Encyclopedia of genomics, proteomics and bioinformatics* chapter Algorithmic Challenges in Mammalian Genome Sequence Assembly John Wiley and Sons. 28
- Bayer, E. M., Bottrill, A. R., Walshaw, J., Vigouroux, M., Naldrett, M. J., Thomas, C. L. and Maule, A. J. (2006) Arabidopsis cell wall proteome defined using multidimensional protein identification technology *Proteomics* **6**, 1, 301–11. 116
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance,

P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoshler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovskiy, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 7218, 53–59 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18

Bertone, P., Gerstein, M. and Snyder, M. (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res* **13**, 3, 259–274 ISSN 0967-3849

(Print); 0967-3849 (Linking). 20

- Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M. and Snyder, M. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 5705, 2242–2246 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 20, 21
- Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., Horsman, D. E., Connors, J. M., Gascoyne, R. D., Marra, M. A. and Jones, S. J. M. (2009) De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 21, 2872–2877 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 29, 118, 142
- Bischoff, V., Nita, S., Neumetzler, L., Schindelasch, D., Urbain, A., Eshed, R., Persson, S., Delmer, D. and Scheible, W.-R. (2010) TRICHOME BIREFRINGENCE and its homolog AT5G01360 encode plant-specific DUF231 proteins required for cellulose biosynthesis in Arabidopsis *Plant Physiol* **153**, 2, 590–602. 116
- Blanca, J., Cañizares, J., Roig, C., Ziarsolo, P., Nuez, F. and Picó, B. (2011) Transcriptome characterization and high throughput SSRs and SNPs discovery in Cucurbita pepo (Cucurbitaceae) *BMC Genomics* **12**, 104. 79
- Blanco, E. and Guigó, R. (2005) *Bioinformatics: A practical guide to the analysis of genes and proteins* chapter Predictive methods using DNA sequences, 116–142 3 John Wiley and Sons. 119
- Bloom, J. S., Khan, Z., Kruglyak, L., Singh, M. and Caudy, A. A. (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays *BMC Genomics* **10**, 221. 80
- Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993) dbEST–database for "expressed sequence tags" *Nat Genet* **4**, 4, 332–3. 82
- Boguski, M. S., Tolstoshev, C. M. and Bassett, D. E., Jr (1994) Gene discovery in dbEST *Science* **265**, 5181, 1993–4. 82
- Bohnert, R. and Rättsch, G. (2010) rQuant.web: a tool for RNA-Seq-based transcript quantitation *Nucleic*

- Acids Res* **38**, Web Server issue, W348–51. 80
- Bokhari, S. H. and Sauer, J. R. (2005) A parallel graph decomposition algorithm for DNA sequencing with nanopores *Bioinformatics* **21**, 7, 889–96. 28
- Borodovsky, M. and McIninch, J. (1993) GENMARK: parallel gene recognition for both DNA strands *Comput Chem* **17**, 2, 123–133. 86
- Bosca, S., Barton, C. J., Taylor, N. G., Ryden, P., Neumetzler, L., Pauly, M., Roberts, K. and Seifert, G. J. (2006) Interactions between MUR10/CesA7-dependent secondary cellulose biosynthesis and primary cell wall structure *Plant Physiol* **142**, 4, 1353–63. 116
- Bowers, J., Mitchell, J., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G. M., Marappan, S., McInerney, P., Platt, A., Roy, A., Siddiqi, S. M., Steinmann, K. and Thompson, J. F. (2009) Virtual terminator nucleotides for next-generation DNA sequencing. *Nat Methods* **6**, 8, 593–595 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 13
- Boyle, J. (2004) Bioinformatics in undergraduate education: Practical examples *Biochemistry and Molecular Biology Education* **32**, 4, 236–238. 43
- Braslavsky, I., Hebert, B., Kartalov, E. and Quake, S. R. (2003) Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* **100**, 7, 3960–3964 ISSN 0027-8424 (Print); 0027-8424 (Linking). 4, 12
- Brem, R. B. and Kruglyak, L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast *Proc Natl Acad Sci U S A* **102**, 5, 1572–7. 145
- Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* **18**, 6, 630–634 ISSN 1087-0156 (Print); 1087-0156 (Linking). 20
- Brown, D. M., Zeef, L. A. H., Ellis, J., Goodacre, R. and Turner, S. R. (2005) Identification of novel

- genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics *Plant Cell* **17**, 8, 2281–95. 108, 116
- Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 1, 78–94 ISSN 0022-2836 (Print); 0022-2836 (Linking). 46, 71, 86
- Burrows, M. and Wheeler, D. (1994) A block-sorting lossless data compression algorithm Technical Report 124 Digital Equipment Corporation. 32
- Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M. and Buell, C. R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis *BMC Genomics* **7**, 327. 23
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustinich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A. and Hayashizaki, Y. (2006) Genome-wide analysis of mammalian promoter architecture and evolution *Nat Genet* **38**, 6, 626–35. 21
- Carteaux, F., Thibaud, M.-C., Zimmerli, L., Lessard, P., Sarrobert, C., David, P., Gerbaud, A., Robaglia, C., Somerville, S. and Nussaume, L. (2003) Transcriptome analysis of Arabidopsis colonized by a plant-growth promoting rhizobacterium reveals a general effect on disease resistance *Plant J* **36**, 2, 177–88. 108
- Casneuf, T., Van de Peer, Y. and Huber, W. (2007) In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics* **8**, 461 ISSN 1471-2105 (Electronic); 1471-2105 (Linking). 20
- Chapman, B. and Chang, J. (2000) Biopython: Python tools for computational biology *SIGBIO NewsL* **20**, 2, 15–19 ISSN 0163-5697. 43
- Che, P., Lall, S., Nettleton, D. and Howell, S. H. (2006) Gene expression programs during shoot, root,

- and callus development in Arabidopsis tissue culture *Plant Physiol* **141**, 2, 620–37. 108
- Chen, F.-C., Wang, S.-S., Chaw, S.-M., Huang, Y.-T. and Chuang, T.-J. (2007) Plant Gene and Alternatively Spliced Variant Annotator. A plant genome annotation pipeline for rice gene and alternatively spliced variant identification with cross-species expressed sequence tag conservation from seven plant species *Plant Physiol* **143**, 3, 1086–95. 23
- Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S. and Gingeras, T. R. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 5725, 1149–1154 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 20
- Clark, T., Sugnet, C. and Ares, M. (2002) Genomewide analysis of mRNA processing in Yeast using splicing-specific microarrays *Science* **296**, 5569, 907–910. 20
- Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J. and Grimmond, S. M. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**, 7, 613–619 ISSN 1548-7105 (Electronic). 3, 20, 22, 82, 141
- Close, T. J., Bhat, P. R., Lonardi, S., Wu, Y., Rostoks, N., Ramsay, L., Druka, A., Stein, N., Svensson, J. T., Wanamaker, S., Bozdog, S., Roose, M. L., Moscou, M. J., Chao, S., Varshney, R. K., Szucs, P., Sato, K., Hayes, P. M., Matthews, D. E., Kleinhofs, A., Muehlbauer, G. J., DeYoung, J., Marshall, D. F., Madishetty, K., Fenton, R. D., Condamine, P., Graner, A. and Waugh, R. (2009) Development and implementation of high-throughput SNP genotyping in barley *BMC Genomics* **10**, 582. 18
- Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 6, 1767–1771 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 26
- Coetzer, N., Gazendam, I., Oelofse, D. and Berger, D. K. (2010) SSHscreen and SSHdb, generic software



- for microarray based gene discovery: application to the stress response in cowpea *Plant Methods* **6**, 10. 79
- Cohen, J. (2003) Guidelines for Establishing Undergraduate Bioinformatics Courses *Journal of Science Education and Technology* **12**, 4, 449–456. 43
- Collins, L. J., Biggs, P. J., Voelckel, C. and Joly, S. (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences *Genome Inform* **21**, 3–14. 83
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research *Bioinformatics* **21**, 18, 3674–6. 46, 48, 70, 143
- Darzentas, N. (2010) Circoletto: visualizing sequence similarity with Circos *Bioinformatics*. 46
- David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. and Steinmetz, L. M. (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**, 14, 5320–5325 ISSN 0027-8424 (Print); 0027-8424 (Linking). 20
- De Bona, F., Ossowski, S., Schneeberger, K. and Ratsch, G. (2008) Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**, 16, i174–80 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 34
- De la Vega, F. M., Lazaruk, K. D., Rhodes, M. D. and Wenz, M. H. (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPLex Genotyping System *Mutat Res* **573**, 1-2, 111–35. 18
- Denoëud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O. and Artiguenave, F. (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biol* **9**, 12, R175 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 3, 20, 22, 34, 82, 141
- Dias Neto, E., Correa, R. G., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva, W., Jr, Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., Carvalho, A. F., Matsukuma, A., Baia, G. S., Simpson, D. H., Brunstein, A., de Oliveira, P. S., Bucher, P., Jongeneel, C. V., O'Hare, M. J., Soares, F., Brentani, R. R., Reis, L. F., de Souza, S. J. and Simpson, A. J. (2000) Shotgun sequencing of the

- human transcriptome with ORF expressed sequence tags *Proc Natl Acad Sci U S A* **97**, 7, 3491–6. 82
- DiGuistini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan, S., Docking, T., Birol, I., Holt, R., Hirst, M., Mardis, E., Marra, M. A., Hameling, R. C., Bohlmann, J., Breuil, C. and Jones, S. J. M. (2010) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data *Genome Biology* **9**, R94. 15, 141
- Dolezel, J., Kubaláková, M., Paux, E., Bartos, J. and Feuillet, C. (2007) Chromosome-based genomics in the cereals *Chromosome Res* **15**, 1, 51–66. 16
- Doukhanina, E. V., Chen, S., van der Zalm, E., Godzik, A., Reed, J. and Dickman, M. B. (2006) Identification and functional characterization of the BAG protein family in *Arabidopsis thaliana* *J Biol Chem* **281**, 27, 18793–801. 108
- Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. and Vogelstein, B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations *Proc Natl Acad Sci U S A* **100**, 15, 8817–22. 8
- Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcharding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M. and Reid, C. A. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 5961, 78–81  
ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 4, 9, 10, 18, 141
- Durham, A. M., Kashiwabara, A. Y., Matsunaga, F. T. G., Ahagon, P. H., Rainone, F., Varuzza, L. and Gruber, A. (2005) EGene: a configurable pipeline generation system for automated sequence analysis

*Bioinformatics* **21**, 12, 2812–3. 42

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., Lushbough, C. and Brendel, V. (2008) PlantGDB: a resource for comparative plant genomics *Nucleic Acids Res* **36**, Database issue, D959–65. 123

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res* **32**, 5, 1792–7. 46

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 5910, 133–138 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 4, 10, 11

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005) The Sequence Ontology: a tool for the unification of genome annotations *Genome Biol* **6**, 5, R44. 38

Eker, J., Janneck, J., Lee, E. A., Liu, J., Liu, X., Ludvig, J., Sachs, S. and Xiong, Y. (2003) Taming heterogeneity - the Ptolemy approach *Proceedings of the IEEE* **91**, 1, 127–144. 37

Eklund, A. C., Turner, L. R., Chen, P., Jensen, R. V., deFeo, G., Kopf-Sill, A. R. and Szallasi, Z. (2006) Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat Biotechnol* **24**, 9, 1071–1073 ISSN 1087-0156 (Print); 1087-0156 (Linking). 20

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 3, 186–194 ISSN 1088-9051 (Print); 1088-9051 (Linking). 26

Ewing, B., Hillier, L., Wendl, M. C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 3, 175–185 ISSN 1088-9051 (Print); 1088-9051 (Linking). 26

- Fedurco, M., Romieu, A., Williams, S., Lawrence, I. and Turcatti, G. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* **34**, 3, e22 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 4, 6
- Ferragina, P. and Manzini, G. (2000) Opportunistic data structures with applications in *FOCS 2000: Proceedings of the 41st Annual Symposium on Foundations of Computer Science* p390 IEEE Computer Society, Washington, DC, USA. 32
- Ferragina, P. and Manzini, G. (2001) An experimental study of an opportunistic index in *SODA 2001: Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms* p269–279 Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. 32
- Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., Wong, W.-K. and Mockler, T. C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res* **20**, 1, 45–58 ISSN 1549-5469 (Electronic); 1088-9051 (Linking). 23, 83
- Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. and Bateman, A. (2010) The Pfam protein families database *Nucleic Acids Res* **38**, Database issue, D211–22. 48
- Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J. and Turner, S. W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 6, 461–465 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 12, 141
- Forment, J., Gilibert, F., Robles, A., Conejero, V., Nuez, F. and Blanca, J. M. (2008) EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration *BMC Bioinformatics* **9**, 5. 42
- Forrest, M. and Moore, T. (2008) *Eucalyptus gunnii*: A possible source of bioenergy? *Biomass and Bioenergy* **32**, 10, 978–980. 1
- Frey, B. J., Mohammad, N., Morris, Q. D., Zhang, W., Robinson, M. D., Mnaimneh, S., Chang, R., Pan, Q., Sat, E., Rossant, J., Bruneau, B. G., Aubin, J. E., Blencowe, B. J. and Hughes, T. R. (2005) Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. *Nat Genet* **37**,

- 9, 991–996 ISSN 1061-4036 (Print); 1061-4036 (Linking). 20
- Fullwood, M. J., Wei, C.-L., Liu, E. T. and Ruan, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res* **19**, 4, 521–532 ISSN 1088-9051 (Print); 1088-9051 (Linking). 14, 19
- García-Alcalde, F., García-Lopez, F., Dopazo, J. and Conesa, A. (2010) Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data *Bioinformatics*. 89
- Garcia-Hernandez, M., Berardini, T. Z., Chen, G., Crist, D., Doyle, A., Huala, E., Knee, E., Lambrecht, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Rhee, S. Y., Scholl, R., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. (2002) TAIR: a resource for integrated Arabidopsis data *Funct Integr Genomics* **2**, 6, 239–53. 123
- Garg, R., Patel, R. K., Tyagi, A. K. and Jain, M. (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification *DNA Res* **18**, 1, 53–63. 144
- Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., Panhuis, T. M., Mieczkowski, P., Secchi, A., Bosco, D., Berney, T., Montanya, E., Mohlke, K. L., Lieb, J. D. and Ferrer, J. (2010) A map of open chromatin in human pancreatic islets *Nat Genet* **42**, 3, 255–9. 38
- Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation *Genome Res* **11**, 8, 1425–33. 38, 48
- Ghadessy, F. J., Ong, J. L. and Holliger, P. (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A* **98**, 8, 4552–4557 ISSN 0027-8424 (Print); 0027-8424 (Linking). 5
- Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J. and Nekrutenko, A. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 10, 1451–1455 ISSN 1088-9051 (Print); 1088-9051 (Linking). 44
- Gilad, Y., Pritchard, J. K. and Thornton, K. (2009) Characterizing natural variation using next-generation sequencing technologies *Trends Genet* **25**, 10, 463–71. 145

- Goecks, J., Nekrutenko, A., Taylor, J. and Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences *Genome Biol* **11**, 8, R86. 35, 38, 142
- Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.-l., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*) *Science* **296**, 5565, 92–100. 16
- Goren, A., Ozsolak, F., Shores, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P. M. and Bernstein, B. E. (2010) Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nat Methods* **7**, 1, 47–49 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 13, 141
- Goto, N., Nakao, M., Kawashima, S., Katayama, T. and Kanehisa, M. (2003) BioRuby: open-source bioinformatics library *GENOME INFORMATICS SERIES* 629–630. 43
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome *Nat Biotechnol.* **79**, 118, 142
- Grattapaglia, D. and Kirst, M. (2008) Eucalyptus applied genomics: from gene sequences to breeding tools *New Phytol* **179**, 4, 911–29. 1
- Graveley, B. R. (2008) Molecular biology: Power sequencing *Nature* **453**, 7199, 1197–1198. 3
- Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1992) Prediction of gene structure *Journal of Molecular Biology* **226**, 1, 141–157. 87

- Hansen, K. D., Brenner, S. E. and Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming *Nucleic Acids Res* **38**, 12, e131. 80
- Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H. and Xie, Z. (2008) Single-molecule DNA sequencing of a viral genome. *Science* **320**, 5872, 106–109 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 13
- Hashimoto, S., Qu, W., Ahsan, B., Ogoshi, K., Sasaki, A., Nakatani, Y., Lee, Y., Ogawa, M., Ametani, A., Suzuki, Y., Sugano, S., Lee, C. C., Nutter, R. C., Morishita, S. and Matsushima, K. (2009) High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. *PLoS ONE* **4**, 1, e4108. 141
- Hibino, T. (2009) "Post-genomics" research in Eucalyptus in the near future *Plant Biotechnology* **26**, 1, 109–113. 82
- Hiller, D., Jiang, H., Xu, W. and Wong, W. H. (2009) Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics* **25**, 23, 3056–3059 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 21, 79
- Hinchee, M., Rottmann, W., Mullinax, L., Zhang, C., Chang, S., Cunningham, M., Pearson, L. and Nehra, N. (2009) Short-rotation woody crops for bioenergy and biofuels applications. *In Vitro Cell Dev Biol Plant* **45**, 6, 619–629 ISSN 1054-5476 (Print); 1054-5476 (Linking). 2, 82
- Hofreuter, D., Tsai, J., Watson, R. O., Novik, V., Altman, B., Benitez, M., Clark, C., Perbost, C., Jarvie, T., Du, L. and Galan, J. E. (2006) Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect Immun* **74**, 8, 4694–4707 ISSN 0019-9567 (Print); 0019-9567 (Linking). 15, 16, 141
- Holland, R. C. G., Down, T. A., Pocock, M., Prlić, A., Huen, D., James, K., Foisy, S., Dräger, A., Yates, A., Heuer, M. and Schreiber, M. J. (2008) BioJava: an open-source framework for bioinformatics *Bioinformatics* **24**, 18, 2096–7. 43
- Homer, N., Merriman, B. and Nelson, S. F. (2009a) BFAST: an alignment tool for large scale genome



- resequencing *PLoS One* **4**, 11, e7767. 31
- Homer, N., Merriman, B. and Nelson, S. F. (2009b) Local alignment of two-base encoded DNA sequence *BMC Bioinformatics* **10**, 175. 31
- Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L. A., Bhattacharyya, D., Bhaya, D., Sobral, B. W., Beavis, W., Meinke, D. W., Town, C. D., Somerville, C. and Rhee, S. Y. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* **29**, 1, 102–105 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 88
- Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W. J., Wang, X., Xie, B., Ni, P., Ren, Y., Zhu, H., Li, J., Lin, K., Jin, W., Fei, Z., Li, G., Staub, J., Kilian, A., van der Vossen, E. A. G., Wu, Y., Guo, J., He, J., Jia, Z., Ren, Y., Tian, G., Lu, Y., Ruan, J., Qian, W., Wang, M., Huang, Q., Li, B., Xuan, Z., Cao, J., Asan, Wu, Z., Zhang, J., Cai, Q., Bai, Y., Zhao, B., Han, Y., Li, Y., Li, X., Wang, S., Shi, Q., Liu, S., Cho, W. K., Kim, J.-Y., Xu, Y., Heller-Uszynska, K., Miao, H., Cheng, Z., Zhang, S., Wu, J., Yang, Y., Kang, H., Li, M., Liang, H., Ren, X., Shi, Z., Wen, M., Jian, M., Yang, H., Zhang, G., Yang, Z., Chen, R., Liu, S., Li, J., Ma, L., Liu, H., Zhou, Y., Zhao, J., Fang, X., Li, G., Fang, L., Li, Y., Liu, D., Zheng, H., Zhang, Y., Qin, N., Li, Z., Yang, G., Yang, S., Bolund, L., Kristiansen, K., Zheng, H., Li, S., Zhang, X., Yang, H., Wang, J., Sun, R., Zhang, B., Jiang, S., Wang, J., Du, Y. and Li, S. (2009a) The genome of the cucumber, *Cucumis sativus* L *Nat Genet* **41**, 12, 1275–81. 16
- Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T. and Han, B. (2009b) High-throughput genotyping by whole-genome resequencing. *Genome Res* **19**, 6, 1068–1076 ISSN 1088-9051 (Print); 1088-9051 (Linking). 17
- Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry,

- J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H. and Yeats, C. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, Database issue, D211–5 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 88
- Hyten, D. L., Cannon, S. B., Song, Q., Weeks, N., Fickus, E. W., Shoemaker, R. C., Specht, J. E., Farmer, A. D., May, G. D. and Cregan, P. B. (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11**, 38 ISSN 1471-2164 (Electronic); 1471-2164 (Linking). 17
- Hyten, D. L., Song, Q., Choi, I.-Y., Yoon, M.-S., Specht, J. E., Matukumalli, L. K., Nelson, R. L., Shoemaker, R. C., Young, N. D. and Cregan, P. B. (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean *Theor Appl Genet* **116**, 7, 945–52. 18
- Illumina (2008) Sequence analysis Software User Guide: For Pipeline Version 1.3 and CASAVA Version 1.0 Technical report Illumina, Inc. 26
- Imelfort, M. and Edwards, D. (2009) De novo sequencing of plant genomes using second-generation technologies *Briefings in bioinformatics* **10**, 6, 609. 16
- Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Huguency, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.-F., Weissenbach, J., Quetier, F. and Wincker, P. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 7161, 463–467 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 88
- Jiang, Z., Tang, H., Ventura, M., Cardone, M. F., Marques-Bonet, T., She, X., Pevzner, P. A. and

- Eichler, E. E. (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution *Nat Genet* **39**, 11, 1361–8. 28
- Johnson, J. M., Edwards, S., Shoemaker, D. and Schadt, E. E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments *Trends Genet* **21**, 2, 93–102. 21
- Kapur, K., Xing, Y., Ouyang, Z. and Wong, W. H. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biol* **8**, 5, R82 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 20
- Katagiri, T., Ishiyama, K., Kato, T., Tabata, S., Kobayashi, M. and Shinozaki, K. (2005) An important role of phosphatidic acid in ABA signaling during germination in *Arabidopsis thaliana* *Plant J* **43**, 1, 107–17. 108
- Kent, W. J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res* **12**, 4, 656–664 ISSN 1088-9051 (Print); 1088-9051 (Linking). 30
- Kim, J.-I., Ju, Y. S., Park, H., Kim, S., Lee, S., Yi, J.-H., Mudge, J., Miller, N. A., Hong, D., Bell, C. J., Kim, H.-S., Chung, I.-S., Lee, W.-C., Lee, J.-S., Seo, S.-H., Yun, J.-Y., Woo, H. N., Lee, H., Suh, D., Lee, S., Kim, H.-J., Yavartanoo, M., Kwak, M., Zheng, Y., Lee, M. K., Park, H., Kim, J. Y., Gokcumen, O., Mills, R. E., Zaranek, A. W., Thakuria, J., Wu, X., Kim, R. W., Huntley, J. J., Luo, S., Schroth, G. P., Wu, T. D., Kim, H., Yang, K.-S., Park, W.-Y., Kim, H., Church, G. M., Lee, C., Kingsmore, S. F. and Seo, J.-S. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 7258, 1011–1015 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18, 19
- Kislyuk, A., Katz, L., Agrawal, S. and Hagen, M. (2005) A computational genomics pipeline for microbial sequencing projects. 4
- Ko, J.-H., Beers, E. P. and Han, K.-H. (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana* *Mol Genet Genomics* **276**, 6, 517–31. 108
- Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y. and Carninci, P. (2006) CAGE: cap analysis of gene expression.

*Nat Methods* **3**, 3, 211–222 ISSN 1548-7091 (Print); 1548-7091 (Linking). 20

- Kosakovsky Pond, S., Wadhawan, S., Chiaromonte, F., Ananda, G., Chung, W.-Y., Taylor, J., Nekrutenko, A. and Galaxy Team (2009) Windshield splatter analysis with the Galaxy metagenomic pipeline *Genome Res* **19**, 11, 2144–53. 38
- Külheim, C., Yeoh, S. H., Maintz, J., Foley, W. J. and Moran, G. F. (2009) Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways *BMC Genomics* **10**, 452. 120
- Kuznetsov, V. A. (2009) Relative avidity, specificity, and sensitivity of transcription factor-DNA binding in genome-scale experiments *Methods Mol Biol* **563**, 15–50. 141
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczyk, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F.,

- Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S. and Chen, Y. J. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 6822, 860–921  
ISSN 0028-0836 (Print); 0028-0836 (Linking). 18
- Langmead, B., Hansen, K. D. and Leek, J. T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna *Genome Biol* **11**, 8, R83. 80
- Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, 3, R25 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 30, 31, 32, 33, 89, 90
- Lawrence, C. J., Dong, Q., Polacco, M. L., Seigfried, T. E. and Brendel, V. (2004) MaizeGDB, the community database for maize genetics and genomics *Nucleic Acids Res* **32**, Database issue, D393–7.

- Lee, E. A. (2009) Finite State Machines and Modal Models in Ptolemy II Technical Report UCB/EECS-2009-151 EECS Department, University of California, Berkeley. 37
- Lee, E. A. and Zheng, H. (2005) Operational Semantics of Hybrid Systems in *HSCC* 25–53. 37
- Leung, M.-K., 0002, T. M., Lee, E. A., Latronico, E., Shelton, C. P., Tripakis, S. and Lickly, B. (2009) Scalable Semantic Annotation Using Lattice-Based Ontologies in *MoDELS* 393–407. 37
- Lewis, B. P., Green, R. E. and Brenner, S. E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100**, 1, 189–192.
- 23
- Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D. E., Koboldt, D. C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L. W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J. R., Minx, P., Gordon, D., Chinwalla, A., Zhao, Y., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W. D., Nagarajan, R., Walter, M. J., Link, D. C., Graubert, T. A., DiPersio, J. F. and Wilson, R. K. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 7218, 66–72 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17
- Li, D., Guo, Y., Shao, H., Tellier, L. C., Wang, J., Xiang, Z. and Xia, Q. (2010a) Genetic diversity, molecular phylogeny and selection evidence of the silkworm mitochondria implicated by complete resequencing of 41 genomes. *BMC Evol Biol* **10**, 81 ISSN 1471-2148 (Electronic); 1471-2148 (Linking). 17
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 14, 1754–1760 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31, 33, 87, 89
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009a) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 16,

2078–2079 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 89, 90

Li, H., Lovci, M. T., Kwon, Y.-S., Rosenfeld, M. G., Fu, X.-D. and Yeo, G. W. (2008a) Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci U S A* **105**, 51, 20179–20184 ISSN 1091-6490 (Electronic); 0027-8424 (Linking). 21, 79

Li, H., Ruan, J. and Durbin, R. (2008b) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 11, 1851–1858 ISSN 1088-9051 (Print); 1088-9051 (Linking). 30, 33, 90

Li, H., Ruan, J. and Durbin, R. (2008c) Mapping short DNA sequencing reads and calling variants using mapping quality scores *Genome Res* **18**, 11, 1851–8. 31

Li, R., Fan, W., Tian, G., Zhu, L., H. and He, Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O., Leung, F.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C., Lam, T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J. and Wang, J. (2010b) The sequence and de novo assembly of the giant panda genome *Nature* **463**, 7279, 311–317. 15, 141

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008d) SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 5, 713–714 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. and Wang, J. (2009b) SOAP2: an



- improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 15, 1966–1967 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31, 33
- Lin, H., Zhang, Z., Zhang, M. Q., Ma, B. and Li, M. (2008) ZOOM! Zillions of oligos mapped. *Bioinformatics* **24**, 21, 2431–2437 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30
- Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., Thompson, J. F., Letovsky, S., Milos, P. and Causey, M. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* **27**, 7, 652–658 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 13
- Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. and Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis *Cell* **133**, 3, 523–36. 141
- Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J. and Zhao, Y. (2005) Scientific Workflow Management and the Kepler System *Concurrency and Computation: Practice and Experience* 1–19. 35, 37, 43, 44
- Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks *Bioinformatics* **21**, 16, 3448–9. 89
- Malhis, N., Butterfield, Y. S. N., Ester, M. and Jones, S. J. M. (2009) Slider–maximum use of probability information for alignment of short sequence reads and SNP detection *Bioinformatics* **25**, 1, 6–13. 31
- Manfield, I. W., Jen, C.-H., Pinney, J. W., Michalopoulos, I., Bradford, J. R., Gilmartin, P. M. and Westhead, D. R. (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis *Nucleic Acids Res* **34**, Web Server issue, W504–9. 139
- Mardis, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 3, 133–141 ISSN 0168-9525 (Print). 3
- Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D., Fulton, L. A., Locke, D. P., Magrini, V. J., Abbott, R. M., Vickery, T. L., Reed, J. S., Robinson, J. S., Wylie, T., Smith, S. M., Carmichael, L., Eldred, J. M., Harris, C. C., Walker, J., Peck, J. B., Du, F., Dukes, A. F., Sanderson, G. E., Brummett, A. M.,

- Clark, E., McMichael, J. F., Meyer, R. J., Schindler, J. K., Pohl, C. S., Wallis, J. W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M. E., Ivy, J. V., Kalicki, J., Elliott, G., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M. A., Baty, J., Heath, S., Shannon, W. D., Nagarajan, R., Link, D. C., Walter, M. J., Graubert, T. A., DiPersio, J. F., Wilson, R. K. and Ley, T. J. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 11, 1058–1066 ISSN 1533-4406 (Electronic); 0028-4793 (Linking). 17
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bembien, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 7057, 376–380 ISSN 1476-4687 (Electronic). 2, 4, 5, 6, 15, 16, 141
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 9, 1509–1517 ISSN 1088-9051 (Print); 1088-9051 (Linking). 21, 79, 80, 139
- McKernan, K., Blanchard, A., Kotler, L. and Costa, G. (2006) Reagents, methods and libraries for bead-based sequencing. Technical report filed as US patent application 20080003571. 8
- McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D.,

- Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M. and Blanchard, A. P. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 9, 1527–1541 ISSN 1549-5469 (Electronic); 1088-9051 (Linking). 18, 19, 141
- Mir, K. U. (2009) Sequencing genomes: from individuals to populations. *Brief Funct Genomic Proteomic* **8**, 5, 367–378 ISSN 1477-4062 (Electronic); 1473-9550 (Linking). 17
- Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., Williams, A., Oinn, T. and Goble, C. (2009) Taverna, reloaded. unpublished technical documentation. 37
- Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F. and Myburg, A. A. (2010) De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq *BMC Genomics* **11**, 1, 681. 84, 120, 123, 155
- Mondego, J. M., Vidal, R. O., Carazzolle, M. F., Tokuda, E. K., Parizzi, L. P., Costa, G. G., Pereira, L. F., Andrade, A. C., Colombo, C. A., Vieira, L. G., Pereira, G. A. and Brazilian Coffee Genome Project Consortium (2011) An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora* *BMC Plant Biol* **11**, 30. 79
- Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E. T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 7289, 773–777 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 24
- Moore, B., Fan, G. and Eilbeck, K. (2010) SOBA: sequence ontology bioinformatics analysis *Nucleic Acids Res* **38 Suppl**, W161–4. 82
- Moore, P., Ming, R. and Delmer, D. (2008) *Genomics of tropical crop plants* volume 1 of *Plant genetics and genomics: Crops and models* chapter Genomics of Eucalyptus, a global tree for energy, paper and wood, 259–298 Springer New York. 2
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying

- mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 7, 621–628 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 3, 14, 20, 22, 31, 34, 58, 73, 82, 83, 89, 141
- Mungall, C. J., Emmert, D. B. and FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information *Bioinformatics* **23**, 13, i337–46. 38
- Mungall, C. J., Misra, S., Berman, B. P., Carlson, J., Frise, E., Harris, N., Marshall, B., Shu, S., Kaminker, J. S., Prochnik, S. E., Smith, C. D., Smith, E., Tupy, J. L., Wiel, C., Rubin, G. M. and Lewis, S. E. (2002) An integrated computational pipeline and database to support whole-genome sequence annotation *Genome Biol* **3**, 12, RESEARCH0081. 42
- Mutwil, M., Klie, S., Tohge, T., Giorgi, F. M., Wilkins, O., Campbell, M. M., Fernie, A. R., Usadel, B., Nikoloski, Z. and Persson, S. (2011) PlaNet: Combined Sequence and Expression Comparisons across Plant Networks Derived from Seven Species *Plant Cell* **23**, 3, 895–910. 139
- Myburg, A. A., Potts, B., Marques, C., Kirst, M., Gion, J., Grattapaglia, D. and Grima-Pettenati, J. (2005) *The genomes: a series on genome mapping, molecular breeding and genomics of economic species*. chapter Genome mapping and molecular breeding in Eucalyptus: molecular domestication of a major fiber crop. Enfield, NH, USA; Plymouth, UK: Science Publishers Inc. 2
- Myers, E. W. (2005) The fragment assembly string graph *Bioinformatics* **21 Suppl 2**, ii79–85. 28
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 5881, 1344–1349 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 3, 14, 20
- Ng, P., Tan, J. J. S., Ooi, H. S., Lee, Y. L., Chiu, K. P., Fullwood, M. J., Srinivasan, K. G., Perbost, C., Du, L., Sung, W.-K., Wei, C.-L. and Ruan, Y. (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes *Nucleic Acids Res* **34**, 12, e84. 19
- Ng, P., Wei, C.-L., Sung, W.-K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., Liu, E. T. and Ruan, Y. (2005) Gene identification signature (GIS) analysis for

- transcriptome characterization and genome annotation. *Nat Methods* **2**, 2, 105–111 ISSN 1548-7091 (Print); 1548-7091 (Linking). 5
- Nielsen, K. L., Høgh, A. L. and Emmersen, J. (2006) DeepSAGE–digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples *Nucleic Acids Res* **34**, 19, e133. 21
- Nishizawa, A., Yabuta, Y., Yoshida, E., Maruta, T., Yoshimura, K. and Shigeoka, S. (2006) Arabidopsis heat shock transcription factor A2 as a key regulator in response to several types of environmental stress *Plant J* **48**, 4, 535–47. 108
- Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Jr, Grattapaglia, D., Sederoff, R. R. and Kirst, M. (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome *BMC Genomics* **9**, 312. 3, 19, 20, 82, 88, 118, 144
- Nowrousian, M., Stajich, J. E., Chu, M., Engh, I., Espagne, E., Halliday, K., Kamerewerd, J., Kempken, F., Knab, B., Kuo, H.-C., Osiewicz, H. D., Poggeler, S., Read, N. D., Seiler, S., Smith, K. M., Zickler, D., Kuck, U. and Freitag, M. (2010) De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet* **6**, 4, e1000891 ISSN 1553-7404 (Electronic); 1553-7390 (Linking). 15, 141
- Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K. and Ohta, H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in *Arabidopsis* *Nucleic Acids Res* **35**, Database issue, D863–9. 139
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes *Nucleic Acids Res* **27**, 1, 29–34. 48
- Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. and Li, P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 17, 3045–3054 ISSN 1367-4803 (Print); 1367-4803 (Linking). 35, 36, 44
- Okoniewski, M. J. and Miller, C. J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics* **7**, 276 ISSN 1471-2105 (Electronic);

1471-2105 (Linking). 20

- Oracle (2009) A Comparison of Oracle Berkeley DB and Relational Database Management Systems  
White paper Oracle 500 Oracle Parkway, Redwood Shores, CA 94065, U.S.A. 85
- Orvis, J., Crabtree, J., Galens, K., Gussman, A., Inman, J., Lee, E., Nampally, S., Riley, D., Sundaram, J., Felix, V., Whitty, B., Mahurkar, A., Wortman, J., White, O. and Angiuoli, S. (2010) Ergatis: A web interface and scalable software system for bioinformatics workflows. *Bioinformatics* ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 35, 38, 44
- Oshlack, A. and Wakefield, M. J. (2009) Transcript length bias in RNA-seq data confounds systems biology *Biol Direct* 4, 14. 80
- Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N. and Weigel, D. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads *Genome Res* 18, 12, 2024–33. 19
- Ozsolak, F., Platt, A., Jones, D., Reifenberger, J., Sass, L., McInerney, P., Thompson, J., Bowers, J., Jarosz, M. and Milos, P. (2009) Direct RNA sequencing *Nature*. 13
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40, 12, 1413–1415 ISSN 1546-1718 (Electronic); 1061-4036 (Linking). 14, 23, 34, 83
- Pang, A., Macdonald, J., Pinto, D., Wei, J., Rafiq, M., Conrad, D., Park, H., Hurles, M., Lee, C., Venter, J., Kirkness, E., Levy, S., Feuk, L. and Scherer, S. (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* 11, 5, R52 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 14, 19
- Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitch, S., Lehrach, H. and Soldatov, A. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA *Nucleic Acids Res.* 23, 24
- Pavy, N., Pelgas, B., Beauseigle, S., Blais, S., Gagnon, F., Gosselin, I., Lamothe, M., Isabel, N. and Bousquet, J. (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce

and black spruce *BMC Genomics* **9**, 21. 18

- Peleg, S., Sananbenesi, F., Zovoilis, A., Burkhardt, S., Bahari-Javan, S., Agis-Balboa, R. C., Cota, P., Wittnam, J. L., Gogol-Doering, A., Opitz, L., Salinas-Riester, G., Dettenhofer, M., Kang, H., Farinelli, L., Chen, W. and Fischer, A. (2010) Altered histone acetylation is associated with age-dependent memory impairment in mice *Science* **328**, 5979, 753–6. 38
- Perkins, T. T., Kingsley, R. A., Fookes, M. C., Gardner, P. P., James, K. D., Yu, L., Assefa, S. A., He, M., Croucher, N. J., Pickard, D. J., Maskell, D. J., Parkhill, J., Choudhary, J., Thomson, N. R. and Dougan, G. (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet* **5**, 7, e1000569 ISSN 1553-7404 (Electronic). 23, 24
- Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly *Proc Natl Acad Sci U S A* **98**, 17, 9748–53. 28
- Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 7289, 768–772 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 24
- Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordonez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A. and Stratton, M. R. (2010a) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 7278, 191–196 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17, 141
- Pleasance, E. D., Stephens, P. J., O’Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordonez, G. R., Mudie, L. J.,



- Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A. and Campbell, P. J. (2010*b*) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 7278, 184–190 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17, 141
- Pushkarev, D., Neff, N. F. and Quake, S. R. (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27**, 9, 847–852 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 13, 18
- Ranik, M. and Myburg, A. A. (2006) Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis *Tree Physiol* **26**, 5, 545–56. 65
- Rasmussen-Poblete, S., Valdes, J., Gamboa, M. C., Valenzuela, P. D. and Krauskopf, E. (2008) Generation and analysis of an Eucalyptus globulus cDNA library constructed from seedlings subjected to low temperature conditions *Electronic Journal of Biotechnology* **11**, 2 ISSN 0717-3458. 82
- Reinhardt, J., Baltrus, D., Nishimura, M., Jeck, W., Jones, C. and Dangl, J. (2009) De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae* *Genome research* **19**, 2, 294. 15, 141
- Rengel, D., San Clemente, H., Servant, F., Ladouce, N., Paux, E., Wincker, P., Couloux, A., Sivadon, P. and Grima-Pettenati, J. (2009) A new genomic resource dedicated to wood formation in Eucalyptus *BMC Plant Biol* **9**, 36. 1, 82, 88
- Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 6, 276–277 ISSN 0168-9525 (Print); 0168-9525 (Linking). 87
- Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. and Pachter, L. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias *Genome Biol* **12**, 3, R22. 80
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y. S., Newsome, R., Chan, S. K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R. A.,

- Hirst, M., Marra, M. A., Jones, S. J. M., Hoodless, P. A. and Birol, I. (2010) De novo assembly and analysis of RNA-seq data *Nat Methods*. 79
- Rothberg, J. M. and Leamon, J. H. (2008) The development and impact of 454 sequencing. *Nature Biotechnology* **26**, 10, 1117–1124. 6
- Royce, T. E., Rozowsky, J. S. and Gerstein, M. B. (2007) Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res* **35**, 15, e99 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 20
- Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallbook, F., Besnier, F., Carlborg, O., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K. and Andersson, L. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 7288, 587–591 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 17
- Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* **5**, 5, e1000386 ISSN 1553-7358 (Electronic). 30
- Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W. and Velculescu, V. E. (2002) Using the transcriptome to annotate the genome *Nat Biotechnol* **20**, 5, 508–12. 21
- Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. and Tettelin, H. (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 1, 24–31 ISSN 0888-7543 (Print); 0888-7543 (Linking). 87
- Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors *Proc Natl Acad Sci U S A* **74**, 12, 5463–7. 20
- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M.,

- Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C. and Jackson, S. A. (2010) Genome sequence of the palaeopolyploid soybean *Nature* **463**, 7278, 178–83. 16
- Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A. and Wilson, R. K. (2009) The B73 maize genome: complexity, diversity, and dynamics *Science* **326**, 5956, 1112–5. 16
- Schneeberger, K., Haggmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. and Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes *Genome Biol* **10**, 9, R98.

- Schulze, U., Hepp, B., Ong, C. S. and Ratsch, G. (2007) PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics* **23**, 15, 1892–1900 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 34
- Schuster, S. C. (2008) Next-generation sequencing transforms today's biology *Nat Methods* **5**, 1, 16–8. 4
- Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q., Burhans, R., Riemer, C., Wittekindt, N. E., Moorjani, P., Tindall, E. A., Danko, C. G., Teo, W. S., Buboltz, A. M., Zhang, Z., Ma, Q., Oosthuysen, A., Steenkamp, A. W., Oostuisen, H., Venter, P., Gajewski, J., Zhang, Y., Pugh, B. F., Makova, K. D., Nekrutenko, A., Mardis, E. R., Patterson, N., Pringle, T. H., Chiaromonte, F., Mullikin, J. C., Eichler, E. E., Hardison, R. C., Gibbs, R. A., Harkins, T. T. and Hayes, V. M. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 7283, 943–947 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18
- Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A. and Shinozaki, K. (2002) Functional annotation of a full-length Arabidopsis cDNA collection *Science* **296**, 5565, 141–5. 82
- Senger, M., Rice, P. and Oinn, T. (2003) SOAPlab - a unified Sesame door to analysis tools in *UK-eScience, All hands meeting* 509–513. 36
- Shah, M. K., Lee, H., Rogers, S. A. and Touchman, J. W. (2004) An Exhaustive Genome Assembly Algorithm Using K-Mers to Indirectly Perform N-Squared Comparisons in  $O(N)$  in *CSB* 740–741. 28
- Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**, 10, 1135–1145 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 8, 13
- Shendure, J., Mitra, R. D., Varma, C. and Church, G. M. (2004) Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**, 5, 335–344 ISSN 1471-0056 (Print); 1471-0056 (Linking). 4

- Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 5741, 1728–1732 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 4, 5, 8, 15, 141
- Sibout, R., Eudes, A., Mouille, G., Pollet, B., Lapierre, C., Jouanin, L. and Séguin, A. (2005) CIN-NAMYL ALCOHOL DEHYDROGENASE-C and -D are the primary genes involved in lignin biosynthesis in the floral stem of *Arabidopsis* *Plant Cell* **17**, 7, 2059–76. 116
- Simková, H., Safár, J., Suchánková, P., Kovárová, P., Bartos, J., Kubaláková, M., Janda, J., Cíhalíková, J., Mago, R., Lelley, T. and Dolezel, J. (2008a) A novel resource for genomics of Triticeae: BAC library specific for the short arm of rye (*Secale cereale* L.) chromosome 1R (1RS) *BMC Genomics* **9**, 237. 16
- Simková, H., Svensson, J. T., Condamine, P., Hribová, E., Suchánková, P., Bhat, P. R., Bartos, J., Safár, J., Close, T. J. and Dolezel, J. (2008b) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley *BMC Genomics* **9**, 294. 16
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data *Genome Res* **19**, 6, 1117–23. 29
- Singer, T., Fan, Y., Chang, H.-S., Zhu, T., Hazen, S. P. and Briggs, S. P. (2006) A high-resolution map of *Arabidopsis* recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet* **2**, 9, e144 ISSN 1553-7404 (Electronic). 20
- Sjödin, A., Street, N. R., Sandberg, G., Gustafsson, P. and Jansson, S. (2009) The Populus Genome Integrative Explorer (PopGenIE): a new resource for exploring the Populus genome *New Phytol* **182**, 4, 1013–25. 123, 144
- Slater, G. S. C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison *BMC Bioinformatics* **6**, 31. 46, 49, 74
- Smith, A. D., Xuan, Z. and Zhang, M. Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping *BMC Bioinformatics* **9**, 128. 31

- Srivastava, S. and Chen, L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data *Nucleic Acids Res* **38**, 17, e170. 80
- Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. and Birney, E. (2002) The Bioperl toolkit: Perl modules for the life sciences *Genome Res* **12**, 10, 1611–8. 43
- Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215–25 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 87
- Stein, L. (2002) Creating a bioinformatics nation *Nature* **417**, 119–120. 36
- Stein, L. D. (2010) The case for cloud computing in genome informatics *Genome Biol* **11**, 5, 207. 34, 36
- Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. and Lewis, S. (2002) The generic genome browser: a building block for a model organism system database *Genome Res* **12**, 10, 1599–610. 38
- Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., Van Montagu, M., Sandberg, G., Olsson, O., Teeri, T. T., Boerjan, W., Gustafsson, P., Uhlén, M., Sundberg, B. and Lundeberg, J. (1998) Gene discovery in the wood-forming tissues of poplar: analysis of 5, 692 expressed sequence tags *Proc Natl Acad Sci U S A* **95**, 22, 13330–5. 82
- Steur nagel, B., Taudien, S., Gundlach, H., Seidel, M., Ariyadasa, R., Schulte, D., Petzold, A., Felder, M., Graner, A., Scholz, U., Mayer, K. F. X., Platzer, M. and Stein, N. (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley *BMC Genomics* **10**, 547. 15, 16
- Stromberg, M. and Marth, G. (2008) MOSAIK: A reference-guided assembler for next-generation sequence data *Manuscript in preparation*. 30, 31, 85, 86
- Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina,

- T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O’Keeffe, S., Haas, S., Vingron, M., Lehrach, H. and Yaspo, M.-L. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 5891, 956–960 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 14, 23, 83
- Sun, Z., Asmann, Y. W., Kalari, K. R., Bot, B., Eckel-Passow, J. E., Baker, T. R., Carr, J. M., Khreb-tukova, I., Luo, S., Zhang, L., Schroth, G. P., Perez, E. A. and Thompson, E. A. (2011) Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing *PLoS One* **6**, 2, e17490. 141
- Swaminathan, K., Alabady, M. S., Varala, K., De Paoli, E., Ho, I., Rokhsar, D. S., Arumuganathan, A. K., Ming, R., Green, P. J., Meyers, B. C., Moose, S. P. and Hudson, M. E. (2010) Genomic and small RNA sequencing of *Miscanthus x giganteus* shows the utility of sorghum as a reference genome sequence for Andropogoneae grasses *Genome Biol* **11**, 2, R12. 16
- Tan, W., Foster, I. and Madduri, R. (2008) Combining the Power of Taverna and caGrid: Scientific Workflows that Enable Web-Scale Collaboration *IEEE Internet Computing* **12**, 61–68. 37
- Tan, W., Missier, P., Foster, I., Madduri, R., De Roure, D. and Goble, C. (2009) A comparison of using Taverna and BPEL in building scientific workflows: the case of caGrid *Concurrency and Computation: Practice and Experience*. 37
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. and Surani, M. A. (2009) mRNA-Seq whole-transcriptome analysis of a single cell *Nat Methods* **6**, 5, 377–82. 80
- Tauch, A., Trost, E., Tilker, A., Ludewig, U., Schneiker, S., Goesmann, A., Arnold, W., Bekel, T., Brinkrolf, K., Brune, I., Götker, S., Kalinowski, J., Kamp, P.-B., Lobo, F. P., Viehoyer, P., Weisshaar, B., Soriano, F., Dröge, M. and Pühler, A. (2008) The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing *J Biotechnol* **136**, 1-2, 11–21. 15, 141
- Tawfik, D. S. and Griffiths, A. D. (1998) Man-made cell-like compartments for molecular evolution. *Nat*



*Biotechnol* **16**, 7, 652–656 ISSN 1087-0156 (Print); 1087-0156 (Linking). 5

Taylor, J., Schenck, I., Blankenberg, D. and Nekrutenko, A. (2007) Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinformatics* **Chapter 10**, Unit 10.5 ISSN 1934-340X (Electronic); 1934-3396 (Linking). 43

Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 9, 1105–1111 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 31, 34, 48, 74, 80

Trapnell, C. and Salzberg, S. L. (2009) How to map billions of short reads onto genomes. *Nat Biotechnol* **27**, 5, 455–457 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 29, 104

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 5, 511–515 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 23, 49, 73, 80, 83, 89, 104, 130, 143

Travers, K., Chin, C., Rank, D., Eid, J. and Turner, S. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 11, 12

Turcatti, G., Romieu, A., Fedurco, M. and Tairi, A.-P. (2008) A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res* **36**, 4, e25 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 4, 6, 7

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehrling, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson,

- J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y. and Rokhsar, D. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 5793, 1596–1604 ISSN 1095-9203 (Electronic); 1095-9203 (Linking). 2, 16, 82, 88
- Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. and Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data *Nat Methods* **5**, 9, 829–34. 141
- van Bakel, H., Nislow, C., Blencowe, B. J. and Hughes, T. R. (2010) Most "dark matter" transcripts are associated with known genes *PLoS Biol* **8**, 5, e1000371. 22
- van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Versteegen, H. and van Eijk, M. J. T. (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes *PLoS One* **2**, 11, e1172. 19
- Van Tassel, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C. and Sonstegard, T. S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5**, 3, 247–252 ISSN 1548-7105 (Electronic); 1548-7091 (Linking). 16
- Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L. M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J. T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Demattè, L., Mraz, A., Battilana,

- J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J. A., Sterck, L., Vandepoele, K., Grando, S. M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S. K., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F. and Viola, R. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety *PLoS One* **2**, 12, e1326. 16
- Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science* **270**, 5235, 484–487 ISSN 0036-8075 (Print); 0036-8075 (Linking). 20
- Velicer, G. J., Raddatz, G., Keller, H., Deiss, S., Lanz, C., Dinkelacker, I. and Schuster, S. C. (2006) Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proc Natl Acad Sci U S A* **103**, 21, 8107–8112 ISSN 0027-8424 (Print); 0027-8424 (Linking). 15, 16
- Venter, J. and Smith, H. (1996) A new strategy for genome sequencing. *Nature*. 15
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F.,

- An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001) The sequence of the human genome *Science* **291**, 5507, 1304–51. 3, 15, 18
- Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I. and Marden, J. H. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing *Mol Ecol* **17**, 7, 1636–47. 83
- Vizoso, P., Meisel, L. A., Tittarelli, A., Latorre, M., Saba, J., Caroca, R., Maldonado, J., Cambiazo, V., Campos-Vargas, R., Gonzalez, M., Orellana, A. and Silva, H. (2009) Comparative EST transcript profiling of peach fruits under different post-harvest conditions reveals candidate genes associated with

- peach fruit quality *BMC Genomics* **10**, 423. 79
- Wang, B.-B. and Brendel, V. (2006) Genomewide comparative analysis of alternative splicing in plants *Proc Natl Acad Sci U S A* **103**, 18, 7175–80. 23
- Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K.-S., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H. and Wang, J. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**, 7218, 60–65 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18
- Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010a) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 1, 136–138 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 46, 49, 76, 80, 143
- Wang, L., Li, P. and Brutnell, T. P. (2010b) Exploring plant transcriptomes using ultra high-throughput sequencing *Briefings in Functional Genomics* **9**, 2, 1–11. 141
- Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 1, 57–63 ISSN 1471-0064 (Electronic); 1471-0056 (Linking). 20
- Wanner, L. A., Li, G., Ware, D., Somssich, I. E. and Davis, K. R. (1995) The phenylalanine ammonia-lyase gene family in *Arabidopsis thaliana* *Plant Mol Biol* **27**, 2, 327–38. 108
- Wehmeyer, N. and Vierling, E. (2000) The expression of small heat shock proteins in seeds responds to discrete developmental signals and suggests a general protective role in desiccation tolerance *Plant Physiol* **122**, 4, 1099–108. 108
- Werner, T. (2010) Next generation sequencing in functional genomics. *Brief Bioinform* ISSN 1477-4054 (Electronic); 1467-5463 (Linking). 3

- Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A. and Rothberg, J. M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 7189, 872–876 ISSN 1476-4687 (Electronic); 0028-0836 (Linking). 18
- Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. and Stein, N. (2006) 454 sequencing put to the test using the complex genome of barley *BMC Genomics* **7**, 275. 15
- Wilhelm, B. T. and Landry, J.-R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 3, 249–257 ISSN 1095-9130 (Electronic); 1046-2023 (Linking). 21, 141
- Wilkinson, M. D. and Links, M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform* **3**, 4, 331–341 ISSN 1467-5463 (Print); 1467-5463 (Linking). 36
- Wittkopp, P. J., Haerum, B. K. and Clark, A. G. (2008) Independent effects of cis- and trans-regulatory variation on gene expression in *Drosophila melanogaster* *Genetics* **178**, 3, 1831–5. 145
- Wu, T., Qin, Z., Zhou, X., Feng, Z. and Du, Y. (2010) Transcriptome profile analysis of floral sex determination in cucumber *J Plant Physiol* **167**, 11, 905–13. 83
- Xia, Q., Guo, Y., Zhang, Z., Li, D., Xuan, Z., Li, Z., Dai, F., Li, Y., Cheng, D., Li, R., Cheng, T., Jiang, T., Becquet, C., Xu, X., Liu, C., Zha, X., Fan, W., Lin, Y., Shen, Y., Jiang, L., Jensen, J., Hellmann, I., Tang, S., Zhao, P., Xu, H., Yu, C., Zhang, G., Li, J., Cao, J., Liu, S., He, N., Zhou, Y., Liu, H., Zhao, J., Ye, C., Du, Z., Pan, G., Zhao, A., Shao, H., Zeng, W., Wu, P., Li, C., Pan, M., Li, J., Yin, X., Li, D., Wang, J., Zheng, H., Wang, W., Zhang, X., Li, S., Yang, H., Lu, C., Nielsen, R., Zhou, Z., Wang, J., Xiang, Z. and Wang, J. (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**, 5951, 433–436 ISSN 1095-9203 (Electronic); 0036-8075 (Linking). 17
- Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao,

- M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L. and Yang, H. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*) *Science* **296**, 5565, 79–92. 16
- Yuen, C. Y. L., Pearlman, R. S., Silo-Suh, L., Hilson, P., Carroll, K. L. and Masson, P. H. (2003) WVD2 and WDL1 modulate helical organ growth and anisotropic cell expansion in *Arabidopsis* *Plant Physiol* **131**, 2, 493–506. 108
- Zdobnov, E. M. and Apweiler, R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro *Bioinformatics* **17**, 9, 847–8. 46, 48, 70, 143
- Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 5, 821–829 ISSN 1088-9051 (Print). 28, 29, 46, 56, 64, 84, 142
- Zhao, W., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., Wei, S., Fu, J., Chen, Y., Ren, X., Zhang, Y., Ni, P., Zhang, J., Li, S., Wang, J., Wong, G. K.-S., Zhao, H., Yu, J., Yang, H. and Wang, J. (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics *Nucleic Acids Res* **32**, Database issue, D377–82. 123
- Zheng, M. S., Takahashi, H., Miyazaki, A., Hamamoto, H., Shah, J., Yamaguchi, I. and Kusano, T. (2004) Up-regulation of *Arabidopsis thaliana* NHL10 in the hypersensitive response to Cucumber mosaic virus infection and in senescing leaves is controlled by signalling pathways that differ in salicylate involvement *Planta* **218**, 5, 740–50. 108