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 contribitions: 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. Theses 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

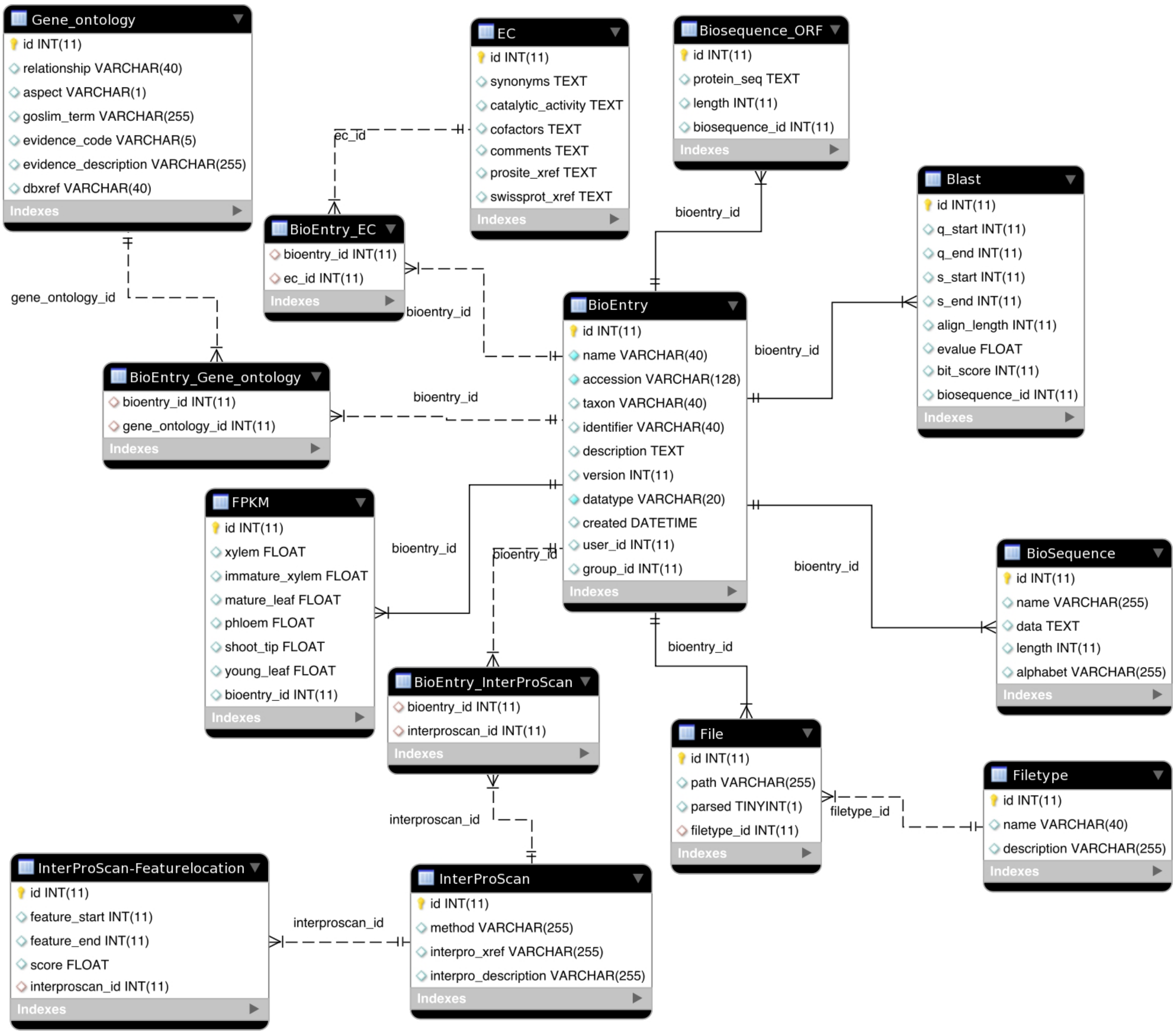
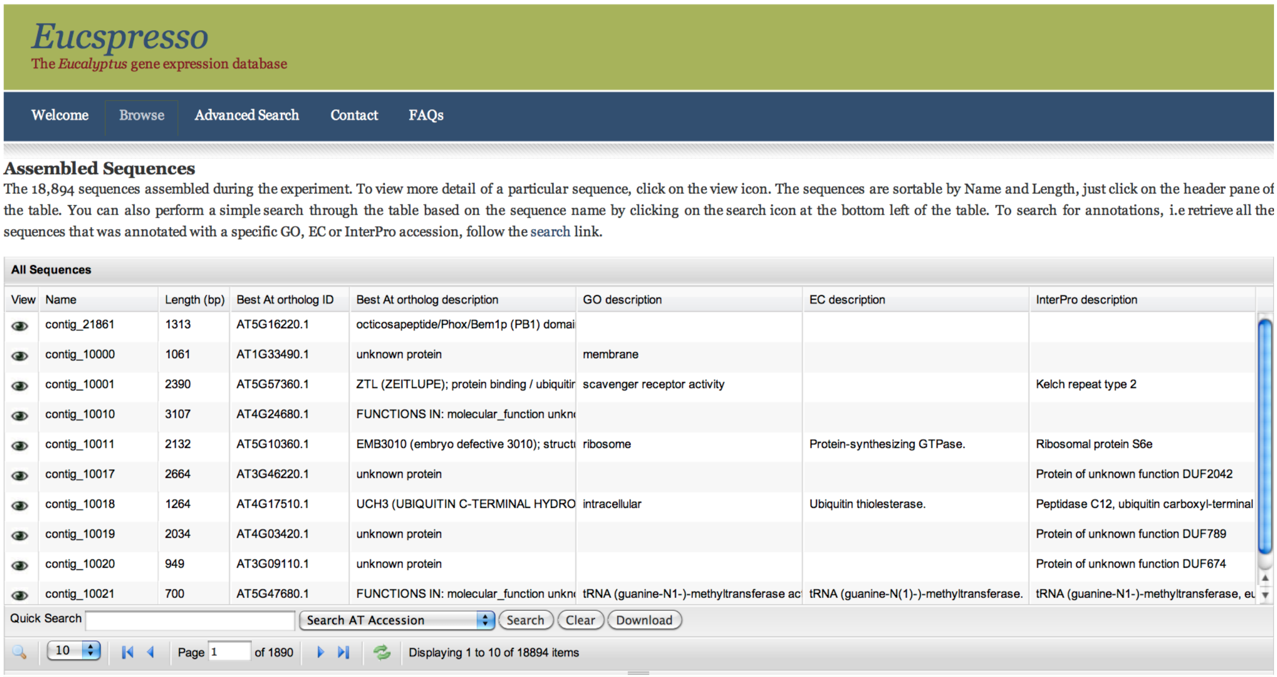


Figure 4.1: Entity relationship diagram of the main datatypes in *Eucspresso*. All the datatypes inherit attributes from the BioEntry table. The description and accession attributes of the BioEntry table are used for searching, and any link between different results occur through the BioEntry table.



Assembled Sequences

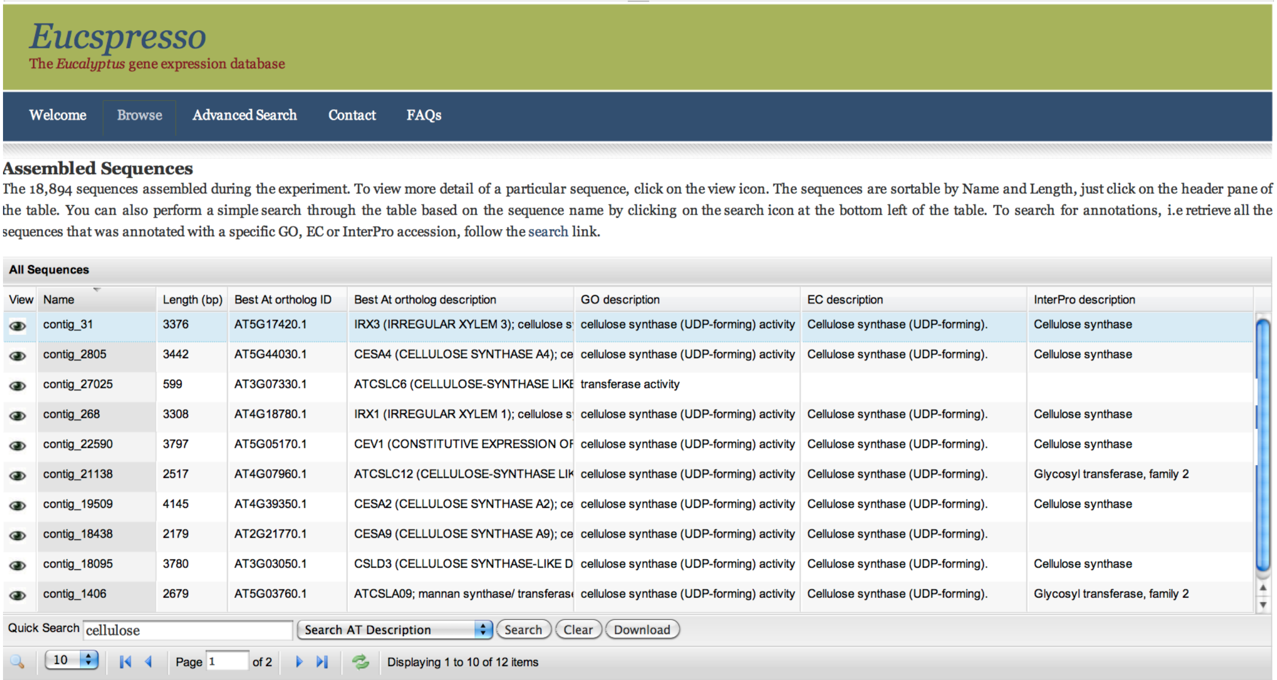
The 18,894 sequences assembled during the experiment. To view more detail of a particular sequence, click on the view icon. The sequences are sortable by Name and Length, just click on the header pane of the table. You can also perform a simple search through the table based on the sequence name by clicking 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) domai			
👁	contig_10000	1061	AT1G33490.1	unknown protein	membrane		
👁	contig_10001	2390	AT5G57360.1	ZTL (ZEITLUPE); protein binding / ubiquitir	scavenger receptor activity		Kelch repeat type 2
👁	contig_10010	3107	AT4G24680.1	FUNCTIONS IN: molecular_function unkn			
👁	contig_10011	2132	AT5G10360.1	EMB3010 (embryo defective 3010); structı	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 HYDRO	intracellular	Ubiquitin thiolesterase.	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 unkn	tRNA (guanine-N1-)-methyltransferase ac	tRNA (guanine-N(1)-)-methyltransferase.	tRNA (guanine-N1-)-methyltransferase, eu

Quick Search: Search AT Accession

Displaying 1 to 10 of 18894 items



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

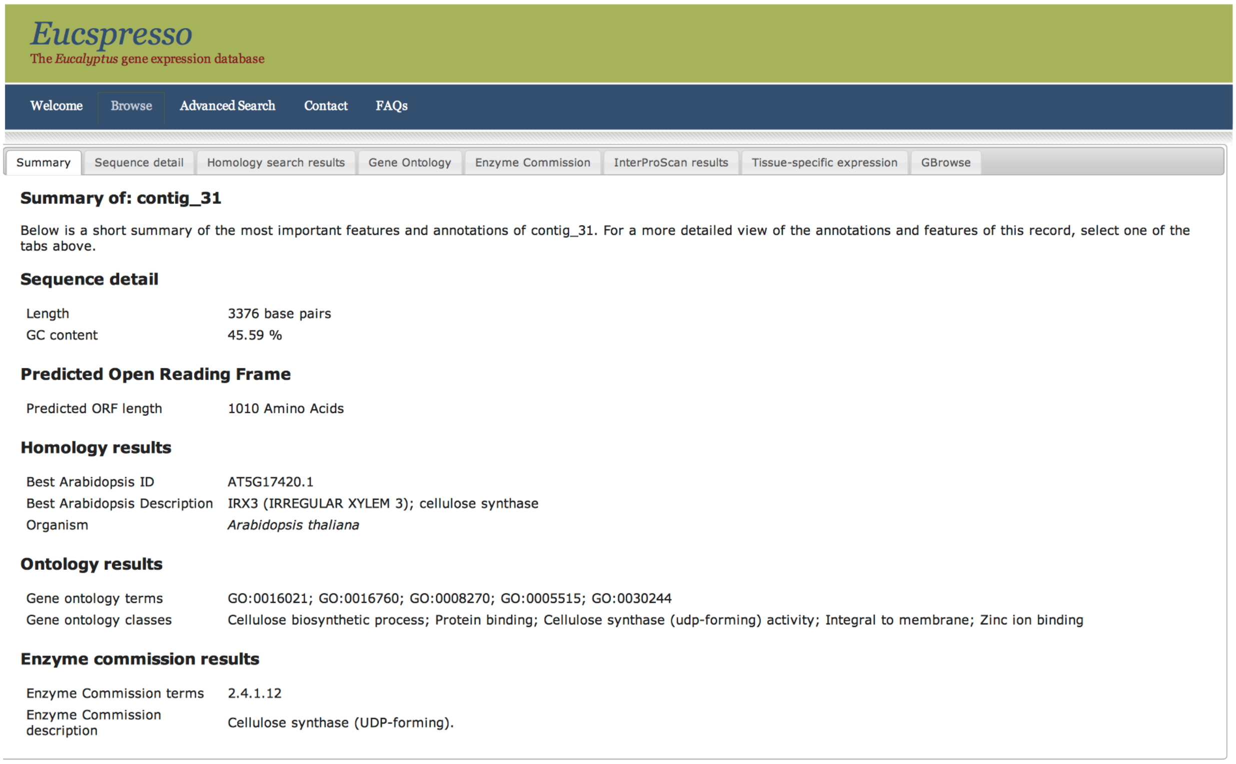
View	Name	Length (bp)	Best At ortholog ID	Best At ortholog description	GO description	EC description	InterPro description
👁	contig_31	3376	AT5G17420.1	IRX3 (IRREGULAR XYLEM 3); cellulose s	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
👁	contig_2805	3442	AT5G44030.1	CESA4 (CELLULOSE SYNTHASE A4); ce	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 s	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
👁	contig_22590	3797	AT5G05170.1	CEV1 (CONSTITUTIVE EXPRESSION OF	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
👁	contig_21138	2517	AT4G07960.1	ATCSLC12 (CELLULOSE-SYNTHASE LIK	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Glycosyl transferase, family 2
👁	contig_19509	4145	AT4G39350.1	CESA2 (CELLULOSE SYNTHASE A2); ce	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Cellulose synthase
👁	contig_18438	2179	AT2G21770.1	CESA9 (CELLULOSE SYNTHASE A9); ce	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	
👁	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/ transferas	cellulose synthase (UDP-forming) activity	Cellulose synthase (UDP-forming).	Glycosyl transferase, family 2

Quick Search: cellulose
Search AT Description

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 descriptiom of the contig in the table is provided by clicking on the "View" column in the table.

A)



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	*Arabidopsis thaliana*

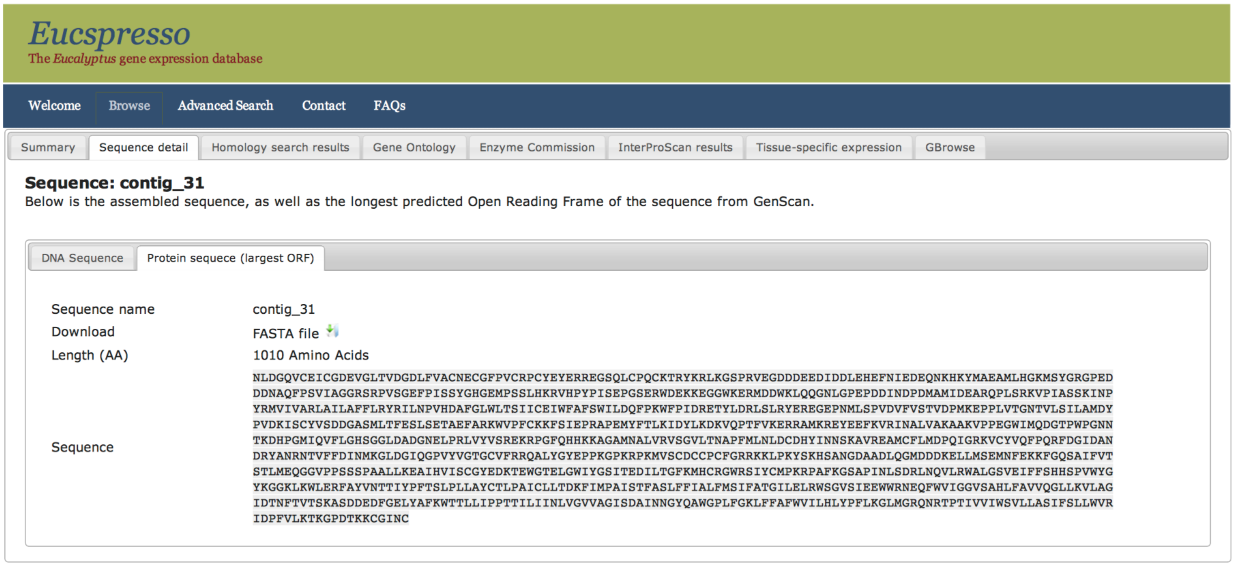
Ontology results

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)



Sequence: contig_31

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

Sequence name	contig_31
Download	FASTA file
Length (AA)	1010 Amino Acids
Sequence	NLDGQVCEICGDEVGLTVDGDLFVACNECGFPVCRPCYEYERREGSQLCPQCKTRYKRLKGSPRVEGDDDEEDIDDLEHEFNIEDEQNKHKYMAEAMLHGKMSYGRGPED DDNAQFPSVIAGGRSRPVSGEFPISSYGHGEMPSSLHKRVHPYPISEPGSERWDEKKEGGWKERMDDWKLQQGNLGPEPDDINDPDMAMIDEARQPLSRKVPIASSKINP YRMVIVARLAILAFFLRYRILNPVHDAFGLWLTSIICEIWFAFSWILDQFPKWFPIDRETYLDRLSLRYEREGEPNMLSPVDVFVSTVDPMKEPPLVTGNTVLSILAMDY PVDKISCYVSDDGASMLTFESLSETAEFARKWVPFCKKFSIEPRAPEMYFTLKIDYLKDKVQPTFVKERRAMKREYEEFKVRINALVAKAAKVPPEGWIMQDGTPWPGNN TKDHPGMIQVFLGHSGGLDADGNELPRLVYVSREKRPGFQHHKKAGAMNALVRVSGVLTNAPFMLNLDCDHYINNSKAVREAMCFLMDPQIGRKVCYVQFPQRFDGIDAN DRYANRNTVFFDINMKGLDGIQGPVYVGTGCVFRRQALYGYEPPKGPKRPKMVSCDCCPCFGRRKKLPKYSKHSANGDAADLQGMDDDKELLMSEMNFEKKFGQSAIFVT STLMEQGGVPPSSSPAALLKEAIHVISCGYEDKTEWGTELGWIYGSITEDILTGFKMHCRGWRSIYCMPKRPAFKGSAPINLSDRLNQVLRWALGSVEIFFSHHSPVWYG YKGGKLKWLERFAYVNTTIYPFTSLPLLAYCTLPAICLLTDKFIMPAISTFASLFFIALFMSIFATGILELRWSGVSIEEWWRNEQFWVIGGVSAHLFAVVQGLLKVLAG IDTNFTVTSKASDDEDFGELYAFKWTTLLIPPTTILIINLVGVVAGISDAINNGYQAWGPLFGKLFFAFWVILHLYPFLKGLMGRQNRTPTIVVIWSVLLASIFSLLWVR IDPFVLKTKGPDTKKCGINC

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).

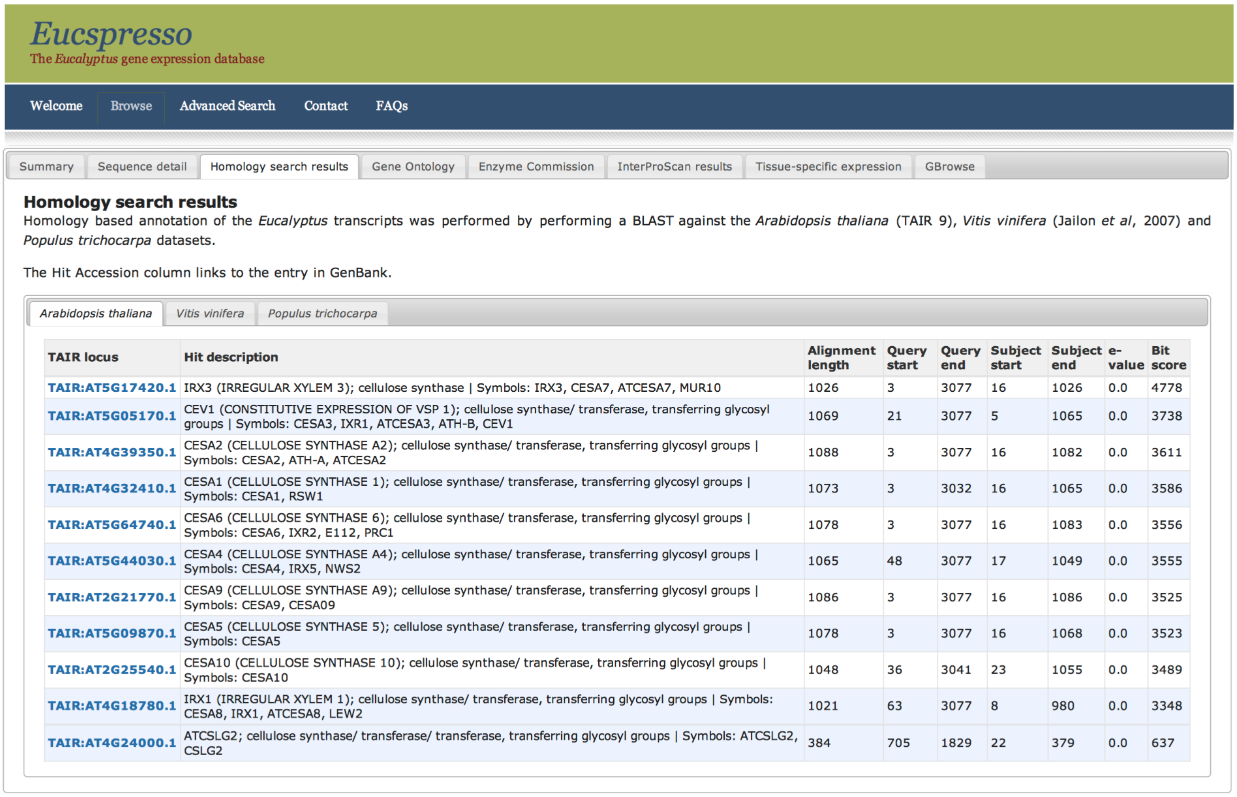
(Figure 4.4B) contains a description of the GO category that the sequence was annotated with, and links to a graph based representation of the ontology term, as rendered by the AmiGo server (Figure 4.5). The gene ontology page (Figure 4.5A) contains a link to download all the contigs in that GO category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



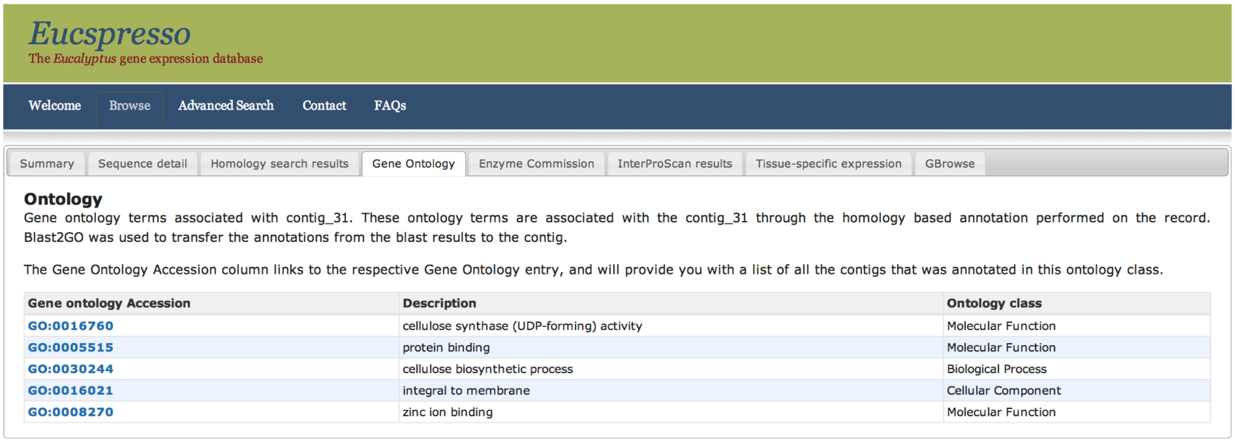
Homology search results

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.

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)



Ontology

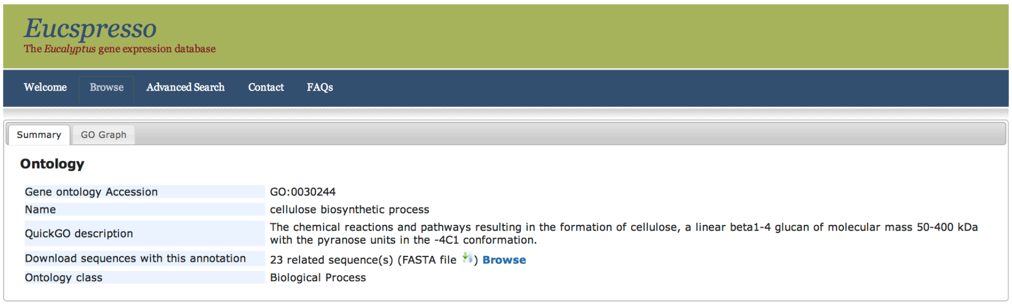
Gene ontology terms associated with contig_31. These ontology terms are associated with the contig_31 through the homology based annotation performed on the record. Blast2GO was used to transfer the annotations from the blast results to the contig.

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).

A)



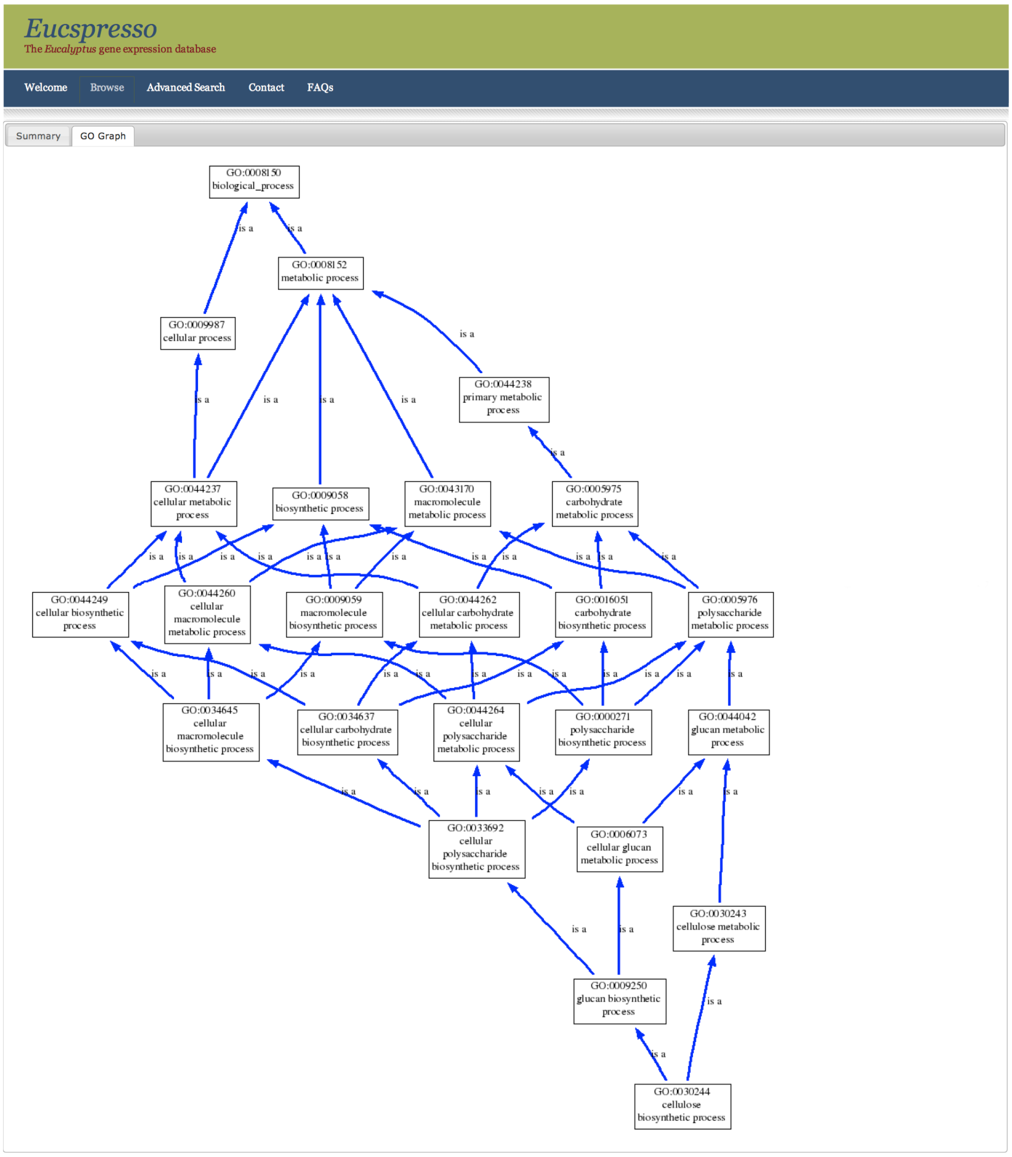
Eucspresso

The *Eucalyptus* gene expression database

Ontology

Gene ontology Accession	GO:0030244
Name	cellulose biosynthetic process
QuickGO description	The chemical reactions and pathways resulting in the formation of cellulose, a linear beta1-4 glucan of molecular mass 50-400 kDa with the pyranose units in the -4C1 conformation.
Download sequences with this annotation	23 related sequence(s) (FASTA file) **Browse**
Ontology class	Biological Process

B)

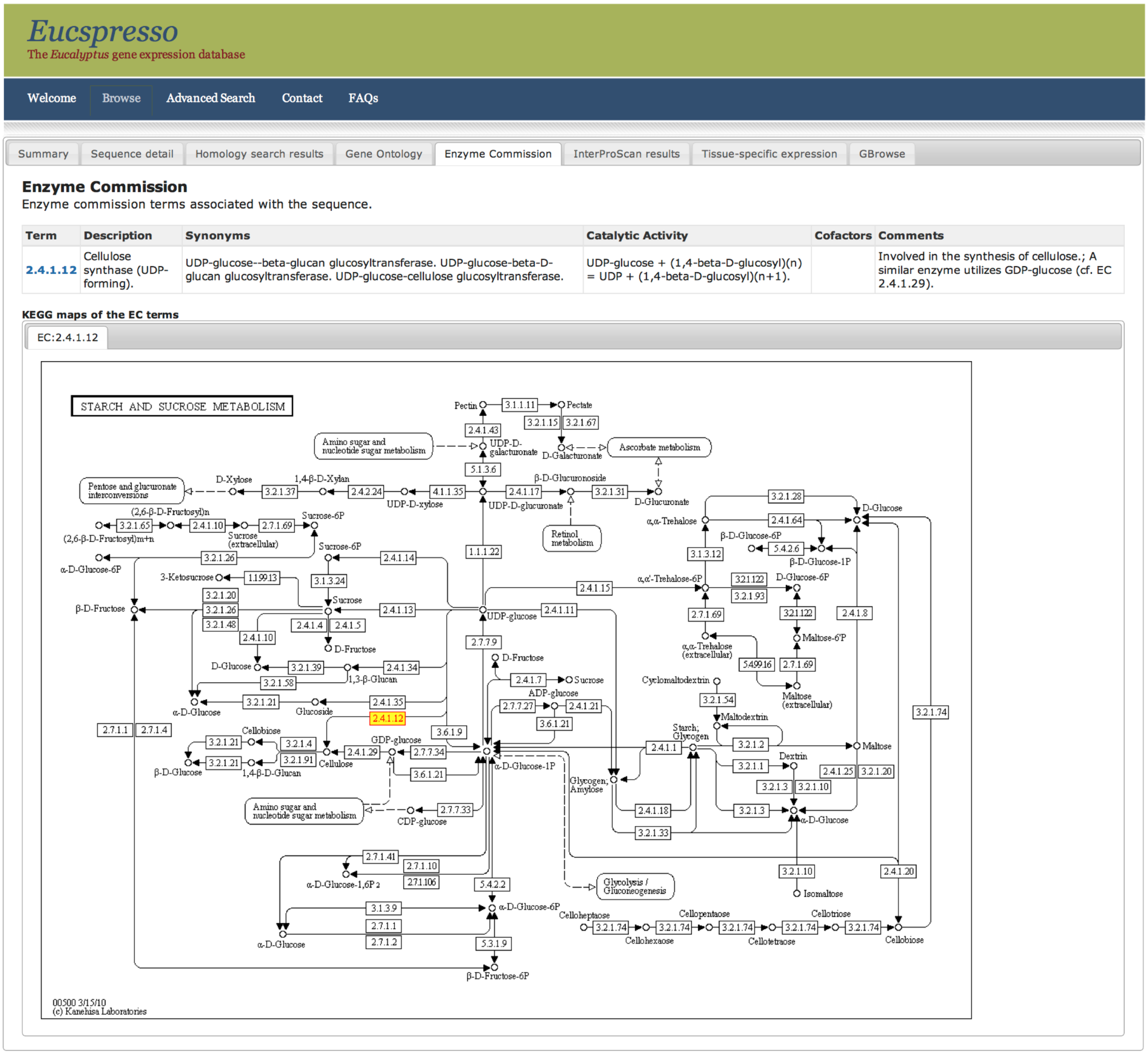


Eucspresso

The *Eucalyptus* gene expression database

Welcome
Browse
Advanced Search
Contact
FAQs
Summary
GO Graph
GO:0008150 biological_process
GO:0008152 metabolic process
GO:0009987 cellular process
GO:0044238 primary metabolic process
GO:0044237 cellular metabolic process
GO:0009058 biosynthetic process
GO:0043170 macromolecule metabolic process
GO:0005975 carbohydrate metabolic process
GO:0044249 cellular biosynthetic process
GO:0044260 cellular macromolecule metabolic process
GO:0009059 macromolecule biosynthetic process
GO:0044262 cellular carbohydrate metabolic process
GO:0016051 carbohydrate biosynthetic process
GO:0005976 polysaccharide metabolic process
GO:0034645 cellular macromolecule biosynthetic process
GO:0034637 cellular carbohydrate biosynthetic process
GO:0044264 cellular polysaccharide metabolic process
GO:0000271 polysaccharide biosynthetic process
GO:0044042 glucan metabolic process
GO:0033692 cellular polysaccharide biosynthetic process
GO:0006073 cellular glucan metabolic process
GO:0030243 cellulose metabolic process
GO:0009250 glucan biosynthetic process
GO:0030244 cellulose biosynthetic process
is a

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).



Eucspresso

The *Eucalyptus* gene expression database

Enzyme Commission

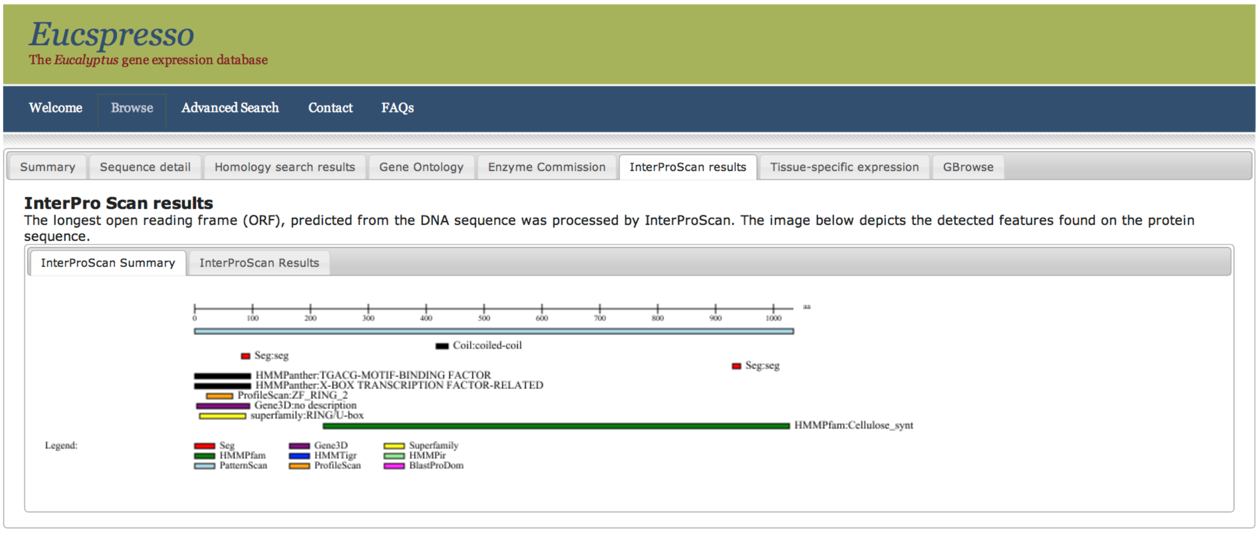
Enzyme commission terms associated with the sequence.

Term	Description	Synonyms	Catalytic Activity	Cofactors	Comments
2.4.1.12	Cellulose synthase (UDP-forming).	UDP-glucose--beta-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.

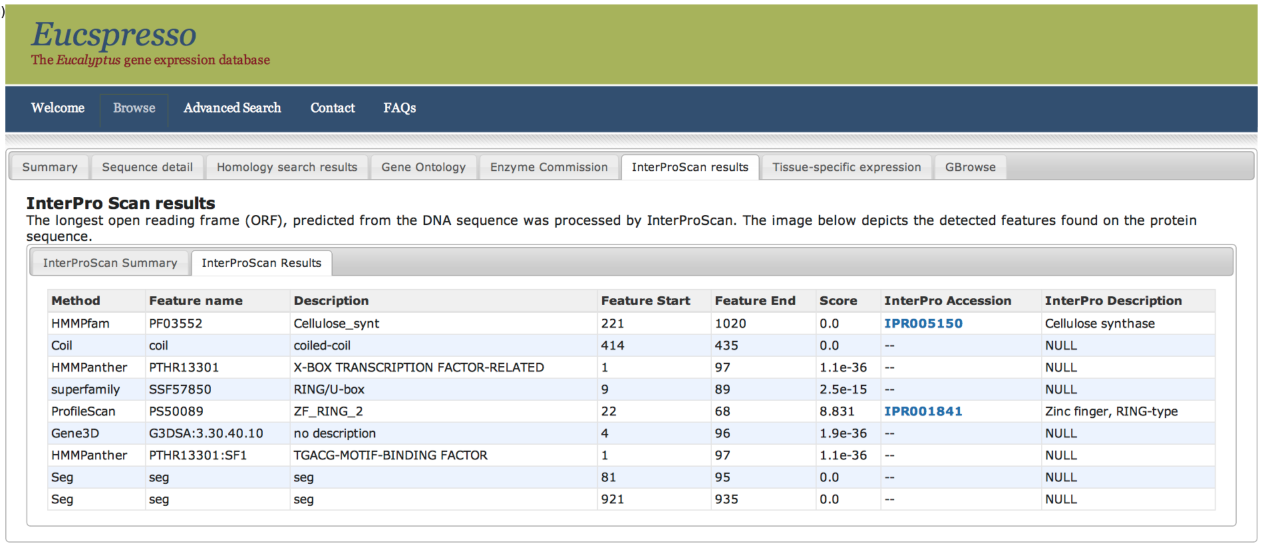
A)



InterPro Scan results

The longest open reading frame (ORF), predicted from the DNA sequence was processed by InterProScan. The image below depicts the detected features found on the protein sequence.

B)

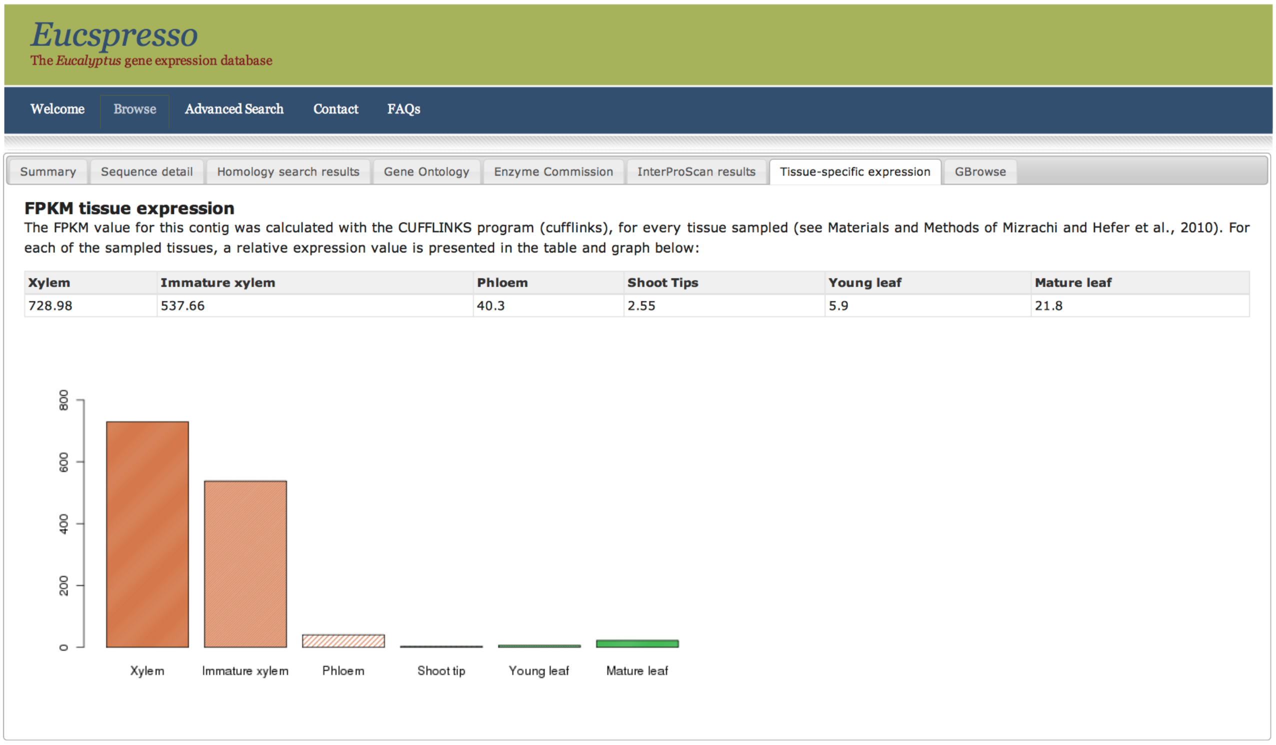


InterPro Scan results

The longest open reading frame (ORF), predicted from the DNA sequence was processed by InterProScan. The image below depicts the detected features found on the protein sequence.

Method	Feature name	Description	Feature Start	Feature End	Score	InterPro Accession	InterPro Description
HMMPfam	PF03552	Cellulose_synt	221	1020	0.0	IPR005150	Cellulose synthase
Coil	coil	coiled-coil	414	435	0.0	--	NULL
HMMPanther	PTHR13301	X-BOX TRANSCRIPTION FACTOR-RELATED	1	97	1.1e-36	--	NULL
superfamily	SSF57850	RING/U-box	9	89	2.5e-15	--	NULL
ProfileScan	PS50089	ZF_RING_2	22	68	8.831	IPR001841	Zinc finger, RING-type
Gene3D	G3DSA:3.30.40.10	no description	4	96	1.9e-36	--	NULL
HMMPanther	PTHR13301:SF1	TGACG-MOTIF-BINDING FACTOR	1	97	1.1e-36	--	NULL
Seg	seg	seg	81	95	0.0	--	NULL
Seg	seg	seg	921	935	0.0	--	NULL

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.



FPKM tissue expression

The FPKM value for this contig was calculated with the CUFFLINKS program (cufflinks), for every tissue sampled (see Materials and Methods of Mizrachi and Hefer et al., 2010). For each of the sampled tissues, a relative expression value is presented in the table and graph below:

Xylem	Immature xylem	Phloem	Shoot Tips	Young leaf	Mature leaf
728.98	537.66	40.3	2.55	5.9	21.8

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 abillity 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 Chaper 3. Although the resource curently 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 Eucpresso 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

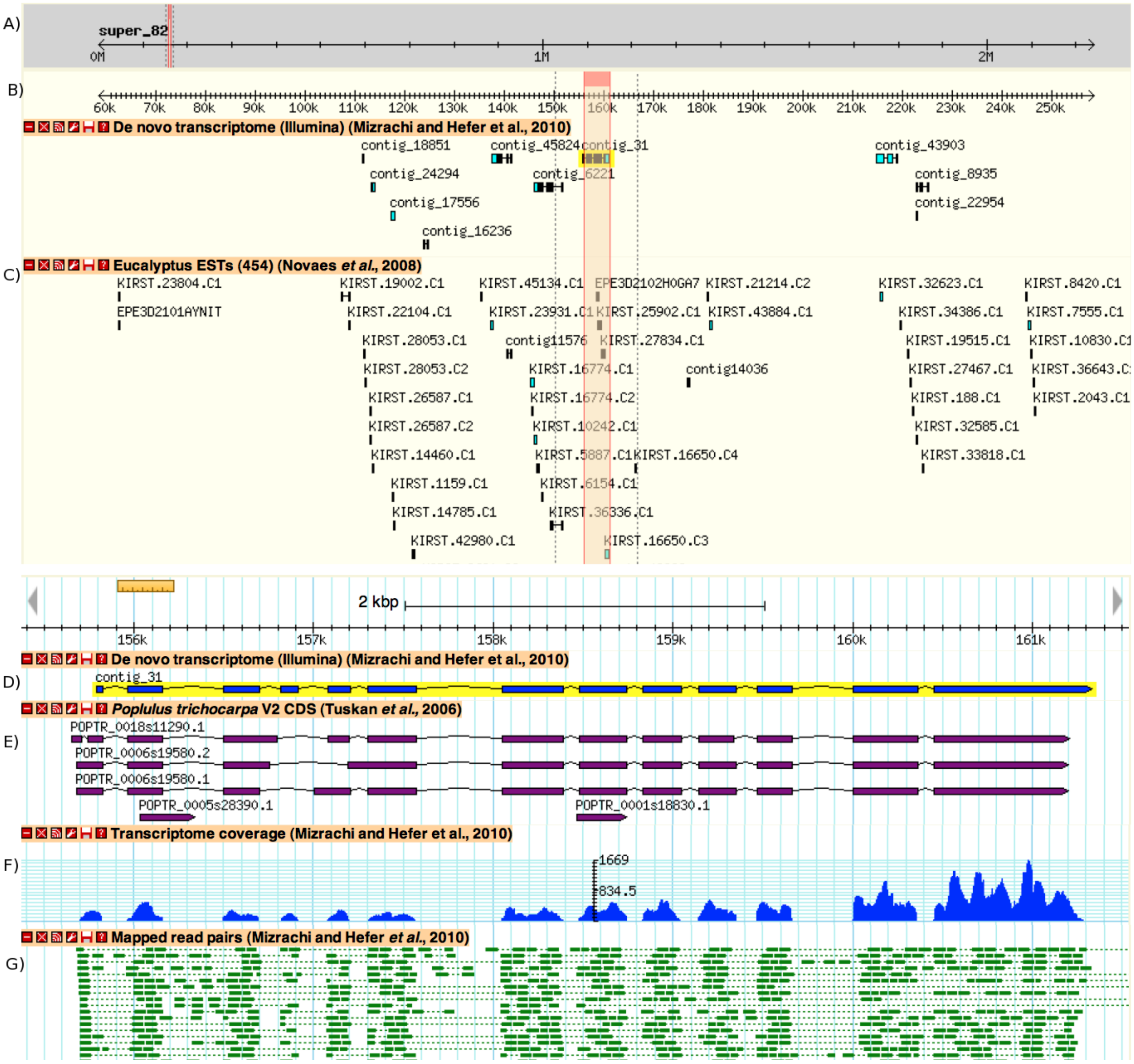
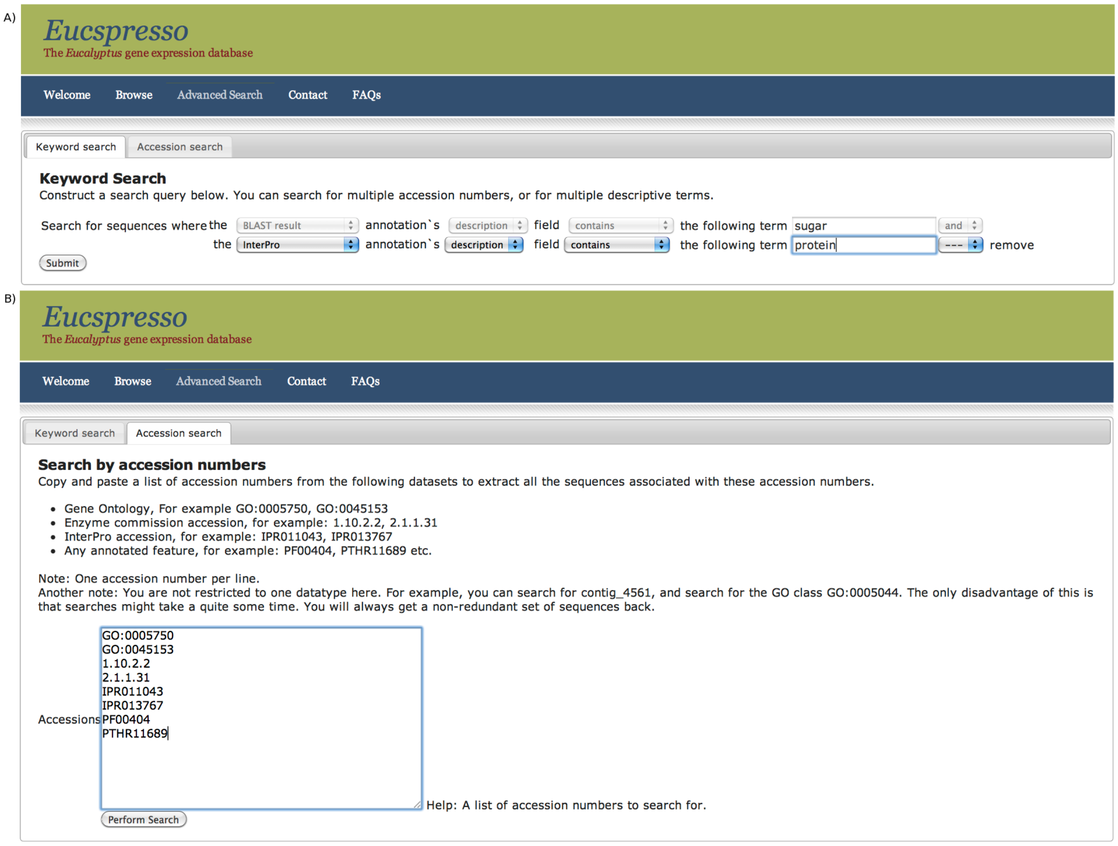


Figure 4.9: The `Eucspresso GBrowse` instance, indicating the position of contig_31 (IRX3) on the 8X *Eucalyptus* draft sequence (scaffold 82, A). The assembled contig is shown in relation to other assembled contigs (B) and some 454 EST data (C) from Novaes *et al.* (2008). When focussing on the highlighted 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).



Keyword Search

Construct a search query below. You can search for multiple accession numbers, or for multiple descriptive terms.

Search for sequences where the BLAST result annotation`s description field contains the following term sugar and
the InterPro annotation`s description field contains the following term protein --- remove

Search by accession numbers

Copy and paste a list of accession numbers from the following datasets to extract all the sequences associated with these accession numbers.

- Gene Ontology, For example GO:0005750, GO:0045153
- Enzyme commission accession, for example: 1.10.2.2, 2.1.1.31
- InterPro accession, for example: IPR011043, IPR013767
- Any annotated feature, for example: PF00404, PTHR11689 etc.

Note: One accession number per line.
Another note: You are not restricted to one datatype here. For example, you can search for contig_4561, and search for the GO class GO:0005044. The only disadvantage of this is that searches might take a quite some time. You will always get a non-redundant set of sequences back.

Accessions
GO:0005750
GO:0045153
1.10.2.2
2.1.1.31
IPR011043
IPR013767
PF00404
PTHR11689

Help: A list of accession numbers to search for.

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 becomming 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 multitute 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 greated 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 succesful *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 Innforamtion 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 extention 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 influences the results from uHTS analysis tools. The `Velvet (Zerbino and Birney, 2008)` assembler was shown to be a reliable transcript assembler, assembling reliable, long, contiguious contigs. One critical shortfall of the assembler is that 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 paramaters 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 highy 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 succesfully 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.*, 2010*a*) 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 succesfully 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 becomimg 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 conjuction 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 powerfull 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.