Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

RNA-Seq: Yellow Nutsedge *(Cyperus esculentus)* transcriptome analysis of lipid-accumulating tubers from early to late developmental stages

Axel Thieffry

Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics

428

Uppsala 2014

Master Thesis, 30 hp

One-Year Master's
Programme in Biology
– Bioinformatics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

RNA-Seq: Yellow Nutsedge *(Cyperus esculentus)* transcriptome analysis of lipid-accumulating tubers from early to late developmental stages

Axel Thieffry

Supervisors:
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Anders S. Carlsson, SLU, Department of Plant Breeding
Examiner:
Göran Andersson, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Biology
Course code: EX0578
Programme: One-Year Master's Programme in Biology
- Bioinformatics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2014
Name of series: Examensarbete / Swedish University of Agricultural Sciences,
Department of Animal Breeding and Genetics, 428
On-line publication: http://epsilon.slu.se

Key words: Cyperus esculentus, transcriptome, tubers, RNA-seq, bioinformatics

ABSTRACT

Thanks to high amounts of starch and oil amassed in the parenchyma of its tubers, yellow nutsedge (*Cyperus esculentus*) stands as a unique plant species with regards to nutrient biosynthesis and accumulation in underground organs. In the last decades, understanding of enzymatic processes in lipid, starch and sugar pathways underwent great improvements. Nevertheless, the underlying mechanisms of carbon allocation in sink tissues are still obscure, and insights may be rendered through the study of yellow nutsedge. Furthermore, in the global context of a still rising need for vegetable oils, *Cyperus esculentus* appears as a promising candidate for the introduction of novel high-yield oil species.

Here is presented the first in-depth analysis of the yellow nutsedge tuber transcriptome, which was conducted using Roche 454 sequencing and targeted two developmental stages, coinciding with (i) the beginning of oil accumulation, but also (ii) an important increase of starch content, and finally (iii) a substantial drop in sugar amount. *Denovo* assembly led to a reference transcriptome of 37k transcripts, which underwent extensive functional and biological pathway annotation, leaving only 7 % of completely unknown sequences. A set of 186 differentially expressed genes (DEGs) was cross-confirmed by three different R packages. To cover the most important changes, top-30 rankings of up and down-regulated genes were investigated. Except a pronounced up-regulation of the WRI1 transcription factor (27-fold), no enzyme related to lipid, starch or sugar was found. Instead, massive changes in growth activity and stress response were observed. Analysis of expression at individual stages showed that several lipid, sugar and starch genes are actually abundant but would undergo changes of lower intensities, hence not visible in the top-30s. A private and user-friendly web-interface has been developed and compiles all the data and results generated through this study, providing with a convenient access for additional investigations, along with directives for further work.

Keywords: Cyperus esculentus, transcriptome, tubers, RNA-seq, bioinformatics

CONTENTS

ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
BP	Base Pair
cDNA	Complementary Deoxyribo-Nucleic Acid
CGI	Common Gateway Interface
CPU	Central Processing Unit
CSV	Comma Separated Values
DAI	Days After tuber Initiation
DEG	Differentially Expressed Gene
DNA	Deoxyribo-Nucleic Acid
dNTP	Deoxynucleoside triphosphate
emPCR	Emulsion polymerase chain reaction
ES	Early developmental Stage
FC	Fold Change
HTML	Hyper-Text Markup Language
LS	Late developmental Stage
NCBI	National Center for Biotechnology Information
NR	Non-Redundant
PPi	Inorganic pyrophosphate
RAM	Random Access Memory
RC	Read Count
RNA	Ribo-Nucleid Acid
RPKM	Reads Per Kilobase per Million reads
rRNA	Ribosomal RNA
SFF	Standard Flowgram Format
SLU	Swedish University of Agricultural Sciences
SQL	Simple Query Language
XML	eXtensible Markup Language

INTRODUCTION

In the active search for fossil-fuel alternatives, vegetable oils feature as a renewable and environmental-friendly solution. One of the challenges to meet the still growing demands of vegetable oils, i.e. for human food, animal feed, industrial use and biofuels, is to introduce and develop new high-yield oil species (Carlsson et al. 2011).

In this context, the cultivated yellow nutsedge (*Cyperus esculentus* L, also known as chufa) has proved to be a promising new oil-producing candidate thanks to a relatively large amount of lipids amassed in its underground tissues (Turesson et al. 2010). Moreover, a recent study indicated that chufa oil might be processed directly for biofuel production, as it originally meets a very large number of prerequisites such as viscosity, cetane number, density and so forth (Makareviciene et al. 2013). In addition to starch, this member of the sedge family accumulates high amounts of oil in the stem tubers, i.e. respectively 30 % DW (dry weight) and 25 % DW (Turesson et al. 2010). For this kind of organs, i.e. tubers and roots, the predominant forms of storage molecules are different kind of carbohydrates such as starch, best represented by the potato (*Solanum tuberosum* L), and sugars, such as in the sugar beet (*Beta vulgaris* L). Here, we investigated the transcriptome of yellow nutsedge, a unique species with regards to the build-up and accumulation of lipids in tubers.

Transcriptome sequencing (RNA-seq) is based on the deep, massively parallel and high-throughput characteristics of the second sequencing generation (Bähler et al. 2009). This recent technology, mainly featured by Illumina, Roche 454 and ABI Solid companies, allows retrieving an absolute digital abundance of mRNA (messenger RNA), without signal background noise (Wang et al. 2009). Moreover, as no prior genetic knowledge is required, RNA-seq can be applied in both known species and discovery-based studies, e.g. *de novo*. Ultimately, this cost-effective sequencing technology provides scientists with astonishing insights in transcriptomes, along with the discovery of genes, transcripts and isoforms, and stands as an ideal investigation technique to dive into the genetic of yellow nutsedge.

BACKGROUND

2.1 Yellow nutsedge

The perennial yellow nutsedge (*Cyperus esculentus* L. 1753) is a spermatophyte (Angiosperm division) belonging to the Cyperaceae family (sedge) and has C_4 photosynthetic characteristics (Defelice 2002). This herbaceous monocotyledon of the Poales order is found worldwide as a weed, a crop or as a wild plant (Bendixen & Nandihalli 2008). Although chufa, the cultivated variety (*C. esculentus* VAR. sativus - Boeck.), evolves best in mild climates, yellow nutsedge has been also reported to spread intensively in colder climates such as Alaska and North Canada (Bendixen & Nandihalli 2008; Stoller et al. 1972). This plant is particularly known for infesting crop fields and has been ranked the world's 16th worst weed (Holm et al. 1977). Surprisingly, chufa was uncovered as an essential consumable (perfume, medicine) back in ancient Egypt, where it was also cooked in diverse forms for the sweetness of its tubers (Negbi 2008). Currently, the nutsedge tubers are still widely used for food in Spain, particularly in a beverage known as horchata (Pascual et al. 2000).

Cyperus esculentus is a 25-65 cm tall plant with grass-like and non-branched leaves; a yellowish inflorescence can be found at the end of a triangular stem (Fig. 1a-c) (Stoller & Sweet 2013). Yellow nutsedge possesses a complex underground system composed of a central bulb from where rhizomes originate (Mulligan & Junkins 1976). At the end of the rhizomes, tubers eventually develops and grow up to ca. 1 - 2 cm long (Fig. 1a, c, d-g) (Turesson et al. 2010). They are recognized as the primary dispersal unit and stock nutrients prior to entering a dormancy phase. Later on, when environmental conditions are favorable, each tuber can give birth to a new shoot, taking benefit of nutrient stocks in order to develop (Stoller & Sweet 2013) (Fig. 1a).

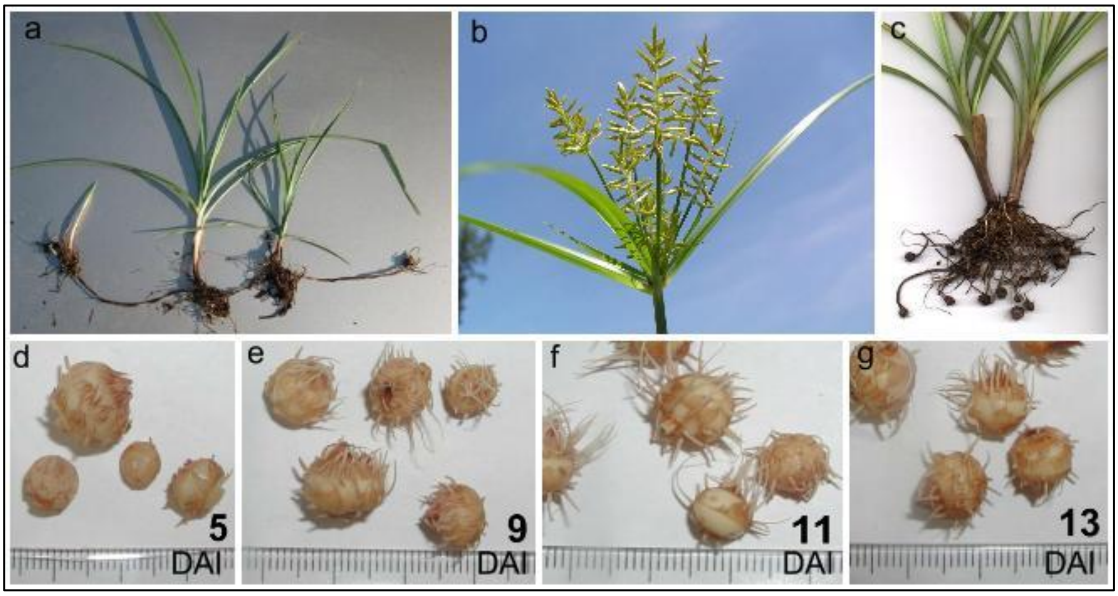


Figure 1 – Yellow nutsedge (*C. esculentus* VAR *sativus*), also known as chufa. The complex underground system allows new shoots to emerge (a). Inflorescences have a brown-yellowish color (b). Tubers can enter a dormancy phase (c). Tubers harvested at 5, 9, 11 and 13 days after initiation (DAI) at the apical end of rhizomes (d, e, f, g) (Turesson et al. 2010).

A few years ago, a study on the nutrient accumulation of yellow nutsedge tubers showed that oil, starch and sugars are build up in a sequential manner (Turesson et al. 2010), after the apex of rhizomes were profoundly reshaped into a storage organ, e.g. tubers (Fig. 2). The underlying idea is a switch in carbon flow from sugar to oil, as starch continues to accumulate and maintains later on an elevated dry weight percentage.

Image adapted from (Turesson et al. 2010).

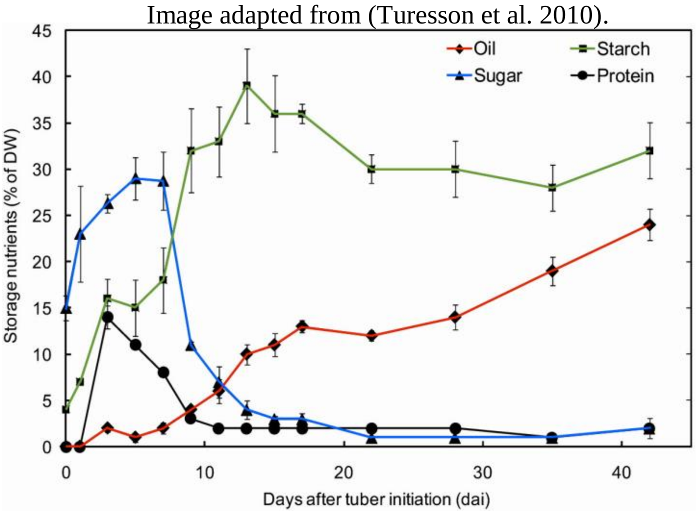


Figure 2 – Dry weight (DW) percentage of nutrients as a function of tuber development (DAI). Sugars and starch contents are rapidly growing up to 7 DAI, where both undergo a important and opposite evolution: a large sugar decrease accords with a dramatic starch expansion. In the meantime, oil accumulation slowly begins.

2.2 Roche 454 transcriptome sequencing

Transcriptome sequencing, often referred to as mRNA-seq, is defined as a snapshot-like, massively parallelized sequencing method targeting messenger RNA of a specimen (cell, cell-line or tissue) at a certain time and under specific conditions (Wang et al. 2010). The high-throughput characteristics of RNA-seq allow to discover gene products with an elevated sequence confidence, but also to get insights in the expression of genes *via* the mRNA abundances (Wang et al. 2010). Nevertheless, due to post-transcriptional regulation events, the abundance of a gene transcript does not always correlate with the expression of the related protein, neither does the activity of the underlying enzyme (Greenbaum et al. 2003).

A messenger RNA sequencing experiment begins with the extraction of total RNA from the sample of interest. This step is processed under highly negative temperatures in order to avoid RNA degradation. Selection of messenger RNAs relies on the complementarity between their 3' polyadenylated (poly-A) tail and poly-T oligonucleotides covalently bound to a substrate, such as magnetic beads (Fig. 3a-c). The enrichment allows to select for coding mRNAs and also to discard ribosomal RNAs (rRNA), which can account for up to 90% of the initial total RNA (Jarvie & Harkins 2008a). Because all sequencing platforms are limited in the read length that can be generated with sufficient quality, the generated cDNA has to be broken down into smaller pieces, fitting the machine's specific length ranges. The mRNA fraction is thus retro-transcribed and sheared into a smaller complementary DNA library (Wang et al. 2010). Prior to the library preparation method, cDNA fragments with the desired length are selected. Each selected cDNA fragment is then flanked with specific primers (Fig. 3d): the first is a ligation primer allowing each fragment to hybridize onto a bead (Fig. 3e), whereas the second permits an amplification by emPCR (emulsion polymerase chain reaction, Fig. 3f), in the case of Roche 454 sequencing platform (Jarvie & Harkins 2008b). This results in a collection of cDNA fragments, with the *one fragment = one bead = one read* principle.

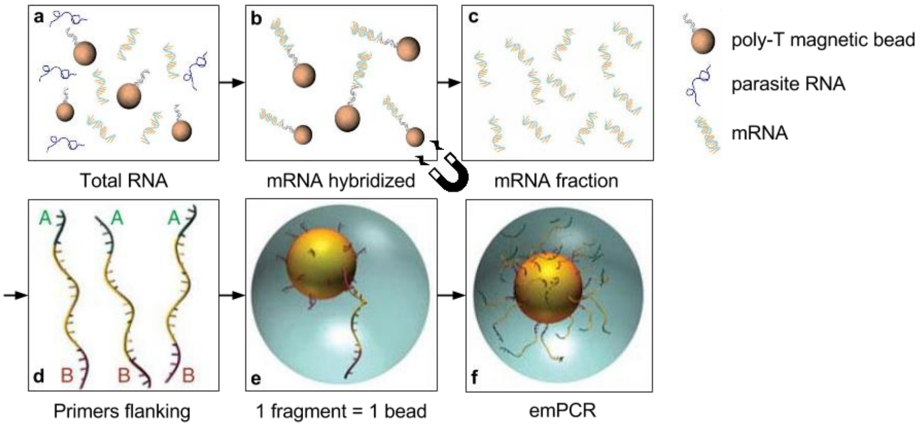


Figure 3 – Messenger RNA enrichment followed by clonal amplification by emPCR (image adapted from Wang et al. 2011; Jarvie & Harkins 2008a). Total RNA extraction involves different types of RNAs, such as messenger, ribosomal and non-coding RNAs (a). mRNAs are separated from parasite RNAs on the basis of hybridization between poly-A tails and poly-T oligos of magnetic beads (b). The result is the messenger RNA fraction (c), which then undergoes flanking of primers (d). Clonal amplification is processed (e and f), prior to the sequencing by synthesis step.

Roche 454 relies on a sequencing-by-synthesis method where each amplified fragment is a template for synthesizing the complementary stretch, by DNA polymerase enzymes (Steen & Cooper 2011) (Fig. 4). The required deoxynucleoside-triphosphates (dNTPs) are released sequentially so that only one type of nucleotide is available at a time. Incorporation of a dNTP is followed by the release of an inorganic pyrophosphate (PPi) in stoichiometric proportions. Finally, sulfurylase and luciferase enzymes use the PPi in order to emit light, which in turn allows to detect and monitor the dNTP addition (Owen-Hughes & Engeholm 2007).

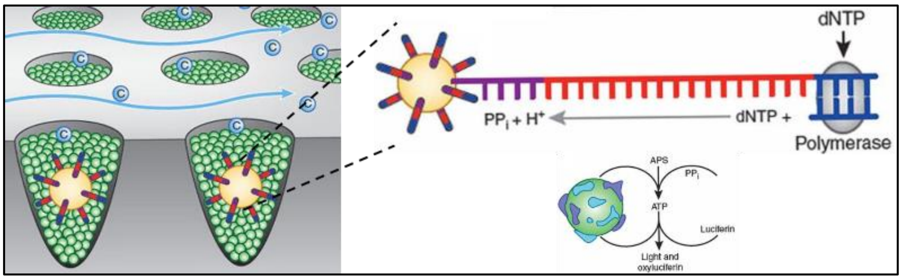


Figure 4 – Roche 454 pyrosequencing (image adapted from Steen & Cooper 2011). Beads are loaded into micro-reactor wells (left). A light-emitting cascade is triggered, involving luciferase and sulfurylase enzymes, when the DNA polymerase enzyme incorporates a dNTP.

The outcome of any second generation sequencing experiment is a list of reads, e.g. a stretch of nucleotides corresponding to a unique sequenced cDNA fragment, which was read from the light emissions during the sequential synthesis reactions (hence the name). After quality control and filtering, reads are assembled into contigs in order to retrieve the complete sequence of each transcript (Grabherr et al. 2011).

2.3 RNA-seq and expression

In order to retrieve the abundance/expression of each transcript in the sample, all reads are aligned onto the assembled transcripts. The number of reads mapped per transcript constitutes the read count (RC), a direct digital measure of the abundance (Robinson et al. 2010).

Two sources of systematic variability were introduced during the sequencing experiment, which avoid comparison of expression based on the read counts. First, the fragmentation step gave longer transcripts to generate more reads, compared to shorter transcripts (Oshlack & Wakefield 2009). The second variability originates from the number of reads produced by the each sequencing experiment, when one wishes to compare expression between samples from different sequencing runs (Garber et al. 2011). Normalization of read count using the RPKM method (Reads Per Kilobase per Million mapped reads) allows to bypass these concerns (Mortazavi et al. 2008) (Fig. 5). The RPKM unit allows a direct comparison of different transcripts, possibly from different sequencings.

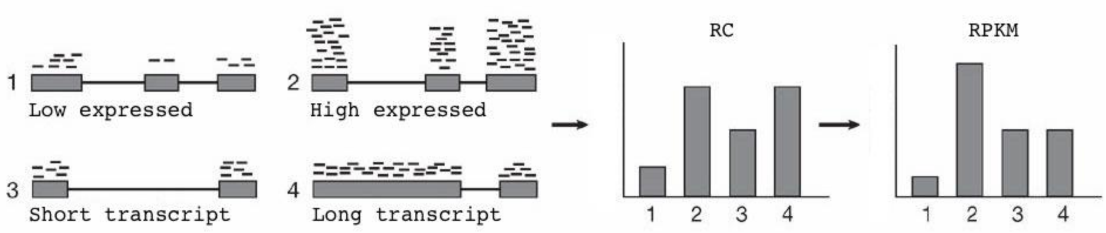


Figure 5 – Systematic variability introduced in transcriptome sequencing (image adapted from Garber et al. 2011). The read count (RC) digital expression unit displays inconsistencies due to transcript length (3 and 4). RPKM normalization method allows un-biased comparison of expressions.

Given *RC* as the read count, *N* as the total number of mappable reads generated by the sequencing and *L* the transcript length in base pairs, the RPKM normalization method is applied as follows (Mortazavi et al. 2008):

$$RPKM = \frac{10^9 . RC}{N . L}$$

AIM

The large-scale project aims at extending knowledge of molecular mechanisms and genetic switches responsible for carbon allocation (between starch, sugars and lipids), thanks to the unique nutrient composition in *C. esculentus* tubers.

To begin with, this work is conducted to gain information on the genetic level of yellow nutsedge, with the first high-throughput dive into its tuber transcriptome and its evolution through two developmental stages. To do so, the three following objectives were drawn:

- To establish the first reference transcriptome of *Cyperus esculentus* tubers, e.g. to assemble the sequencing reads into contigs prior to functionally annotate them.
- To investigate highest expressed genes in each individual developmental stage, as well as the differentially expressed genes, with focus on lipid biosynthesis and storage related-genes.
- To compile all results in a user-friendly environment, allowing for convenient access and further exploit of the data.

MATERIAL & METHODS

4.1 Biological Sampling

Yellow nutsedge plants (10-15 plants) were grown in an aeroponic system (as described in Turesson et al. 2010). At the onset of their development on rhizomes, tubers were individually labeled with the date. In this study, two replicated developmental stages of tubers were considered: the early developmental stage is 5 days after initiation of tubers (DAI), whereas the late developmental stage is a pooling of 9, 11 and 13 DAI. Tubers labeled with the pre-cited DAIs were harvested from the aeroponic growth system and straightaway frozen in liquid nitrogen prior to their storage at -80 °C.

To facilitate reading, the 5 DAI sample and its replicate will be further referred to as *ES1* and *ES2* (Early Stage), whereas the pooled 9-11-13 DAIs sample and its replicate will be referred to as *LS*1 and *LS2* (Later Stage).

4.2 Library Preparations and Pyrosequencing

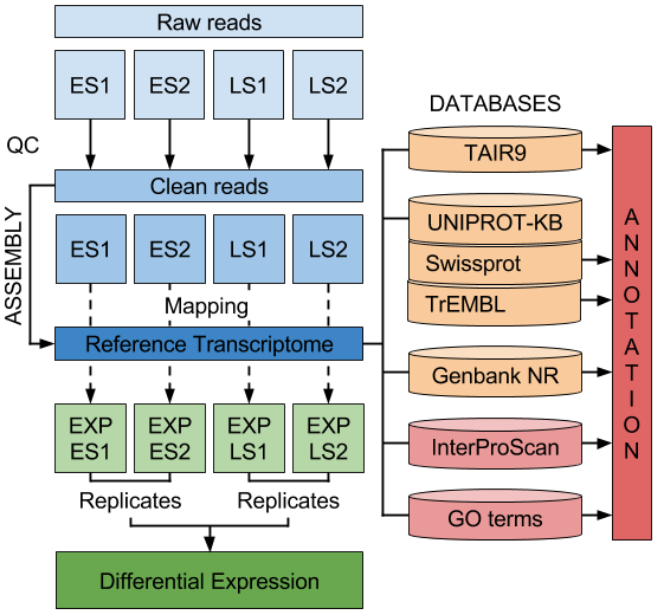
For the sequencing analyses, 5 to 10 tubers of each biological sample were removed from -80°C stocks, frozen again in liquid nitrogen and crushed down into powder, using the MM400 Mixer Mill (Retsch GmbH). Tuber total RNA was isolated using PureLink® RNA reagent (Invitrogen, cat. #12322-012). Isolation of polyadenylated messenger RNA from ribosomal and transfer RNAs was processed with the GenElute mRNA Miniprep Kit (SIGMA-ALDRICH, cat. #DMN10). Complementary DNA (cDNA) was generated with SuperScript® Double-Stranded cDNA Synthesis Kit (Invitrogen, cat. #11917-010) and underwent size-fractionation with CHROMA SPIN-400 Column (Clontech, cat. #636076). A modified protocol of the SMART™ cDNA Library Construction Kit (Clontech, cat. #634901) was used (Appendix 1). Roche 454 GS FLX+ pyrosequencing of cDNA libraries (no strand-specificity) was conducted at the Michigan State University Research Technology Support Facility (MSU-RTSF, http://rtsf.msu.edu). Finally, ES1, LS1 and their replicates ES2 and LS2 constitute the 4 sequenced libraries of nutsedge tubers.

4.3 Bioinformatic Workflow

A schematic overview of the bioinformatic workflow is shown at Figure 6, where (i) quality filtering and assembly are depicted in blue, (ii) annotation steps are colored in red, and finally (iii) green groups point for detection of differential expression.

Most of bioinformatic processes and analyses were conducted on the PlantLink server. The latter is an Ubuntu Server (12.04.03 LTS, Long Term Support) with a total of 24-cores (2 x Intel® Xeon® E5-2620), 100Gb RAM (Random Access Memory) and a hard drive capacity of 10Tb. PlantLink (http://www.plantlink.slu.se) is a Swedish capacity resource in plant sciences, formed cooperatively by Lund University and the Swedish University of Agricultural Sciences (SLU).

Figure 6 – Bioinformatic workflow. After quality control (QC), reads were merged for denovo assembly step (blue). To retrieve expressions (EXP), reads were mapped back onto the reference transcripts (Mapping). The reference transcriptome underwent functional annotation against several databases (red). Finally, differentially expressed genes (DEG's) were investigated (green). ES: Early developmental Stage; LS: Late developmental Stage; TAIR9: The Arabidopsis Information Resource version 9; TrEMBL: Translated European Molecular Biology Laboratory database; Genbank NR: Non-Redundant; GO: Gene Ontology.



4.3.1 Quality Control and Assembly

Nucleotide sequences were provided in binary files, using the SFF format (Standard Flowgram Format, quality score v1.1.03). Extraction of nucleotide sequences and their qualities was processed with sff_extract package. Sff_extract takes advantage of supplementary information hosted in the SFF file to process an initial quality filtering: trimming 5' adaptor sequences and low-quality 3' ends. Removal of low-complexity regions and hypothetic vector contaminants was accomplished with the SeqClean pipeline (The Gene Index, http://compbio.dfci.harvard.edu), combined with the NCBI UniVec database (National Center for Biotechnology Information, v7.1). Clean reads of all sequencings were merged together for clustering and de novo assembly of the tuber reference transcriptome, with The Gene Indices Clustering Tool pipeline (TGICL v2.1) (Pertea et al. 2003).

4.3.2 Annotations

To retrieve functional annotation, BLASTx (Altschul et al. 1990) served to align the reference contigs against several protein databases. The expectation threshold was set to 1e^{-10}. Each contig was searched against the three following databases: The Arabidopsis Information Resource (TAIR, release #9 from 5th Oct. 2013), The Universal Protein Knowledge Base (UniProt-KB, release #2013-03 from 3rd Apr. 2013), composed of SwissProt and TrEMBL (Translated European Molecular Biology Laboratory database), and finally Genbank-NR (Non-Redundant, release #195.0 from 16th Apr. 2013). For alignments against Genbank-NR, the output format was set to XML (eXtensible Markup Language, `-out_fmt 5`), and a maximum of 10 hits were allowed per contig `(-max_target_seqs 10)`, e.g. the 10 best hits based on the expectation value.

Blast2GO pipeline (Conesa & Götz 2008) was fed with the tuber reference contigs along with their BLASTx results from Genbank-NR in XML format. This allowed Blast2GO to associate GO terms, from The Gene Ontology database (Ashburner et al. 2000), to each of the contigs. After the GO mapping step, contigs underwent InterProScan analysis (Quevillon et al. 2005), which screened for proteic signatures and domains in order to associate more GO terms. Finally, GO terms permitted to retrieve EC numbers (enzymatic codes), which in turn made possible the associations between contigs and biological pathways from the KEGG database, the Kyoto Encyclopedia of Genes and Genome (Kanehisa & Goto 2000).

Plotting of the GO terms classification was processed with WEGO (Web Gene Ontology Annotation Plotting, Beijing Genomic Institute) (Ye et al. 2006). Contigs and their GO terms were separated between early and later stages. Both lists were fed in the standard WEGO format in order to compare developmental stages.

4.3.3 Read Count Retrieval

Clean reads of each sequencing library (ES1, ES2, LS1 and LS2) were mapped back onto the reference contigs using Bowtie2 (v2.1.0) (Langmead & Salzberg 2012), with local alignment parameter, `--local`. The number of mapped reads to each contig provided the read count (RC). Considering lengths of the reference transcript (in base pairs), along with RC and the sequencing output at each stage, expressions of contigs were normalized following the RPKM method (Reads Per Kilobase per Million reads) (Mortazavi et al. 2008).

4.3.4 Differential Expression Analysis

DEB Web Interface for RNA-seq Data Analysis (Yao & Yu 2011) was used to detect differentially expressed genes (DEGs). This web interface implements three different R packages, namely DESeq (Anders & Huber 2010), edgeR (Robinson et al. 2010) and BaySeq (Hardcastle & Kelly 2010). The necessary CSV (Comma Separated Values) input file was constructed, consisting of each of the 37,585 reference contigs along with their expressions. Although a stringent FDR (False Discovery Rate) is advised to lesser type I and II errors (Kvam et al. 2012), a more loosened FDR of 10% was applied in this study, and still yields a very low number of DEGs (Fig. 9a). Top-30 down-regulated and up-regulated contigs (from early to late developmental stage) were selected on the basis of differential expression intensity (Fold Change).

4.4 The Nutsedge 454 Database

A private online resource of nutsedge RNA-seq data and analyses has been build, using a HTML form (HyperText Markup Language), a Python CGI script (Common Gateway Interface) and a MySQL database (Simple Query Language). These three components are further succinctly described. The website has been developed and is currently hosted on the PlantLink server (plantlink.ltj.slu.se). An overview of the website structure is shown in Figure 7.

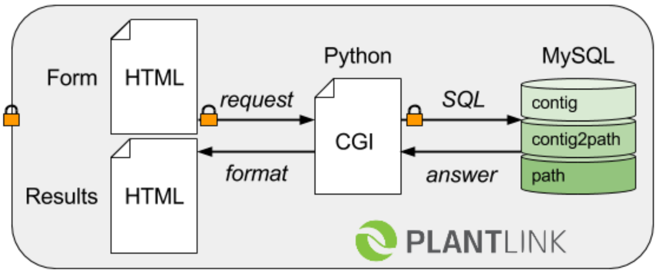


Figure 7 – Representation of The Nutsedge 454 Database structure. A Python script digests user's request from a HTML form, and query the MySQL database. The answer is returned to the user as a HTML table. PlantLink server hosts the database and user interface. Padlocks represent security-related procedures.

The MySQL database was intended to remain simple, with only three different tables:

1. *contig :* statistics, expression data, annotations
2. *path :* biological pathways and IDs of associated contigs
3. *contig2path :* links table *contig* and table *path* (N:N)

The many-to-many relationship (N:N) linking *contig* and *path* tables is provided thanks to *contig2path* table. Indeed, one contig can be associated to several biological pathways, and vice versa. Nutsedge RNA-seq results were joined in CSV format, to further be imported in each of the tables by SQL scripts. Descriptions of tables are available in Appendix 2.

A HTML form was developed and optimized for recent versions of Google Chrome, Safari and Mozilla Firefox browsers. Users can set and combine several filters, as well as select numerous features to be displayed or not. Once submitted, the form is sent to a Python CGI script. To reach the PlantLink server, a connection to SLU network, locally or via VPN (Virtual Private Network), is required. Moreover, to access the website itself, a supplementary login and password combination is needed. Besides the form, the website also includes an extensive *Help* page, describing options, filters and their use; a *Pathway* page, listing all biological pathways characterized in the dataset; and finally an *Expression* page with succinct explanations of differential expression packages and their respective metrics.

The Python CGI script digests the user's request and submits it to the MySQL database. First, the Python script runs through all users' inputs and validates each entry for forbidden characters. This verification is performed by RE (Regular Expressions) and was implemented in order to avoid most of the basic SQL injections, e.g. attempts to hack in the MySQL database. Next, the SQL query is build up and applied to the MySQL database using *MySQLdb* Python module. At this step, the Python script connects to the database as a MySQL user with very limited permission, adding another level of security. Finally, the CGI script obtains response from the database, formats and yields results to the user in a new HTML page.

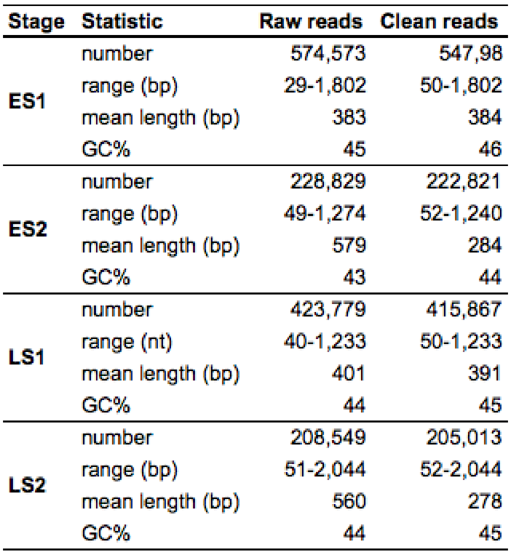
RESULTS

5.1 Sequencing and Quality Control

Roche 454 pyrosequencing yielded a total of 1,435,730 raw reads: 574,573 reads for ES1 and 228,829 reads for its replicate (ES2); when it comes to the late developmental stage, respectively 423,779 and 208,549 reads were obtained for LS1 and LS2. All raw libraries considered, read lengths range from 29 bp up to 2,044 bp, with a mean length of 480 bp. Almost 97% (1,391,681) of reads passed sff_extract and SeqClean quality filters. This lead to a clean dataset of 547,980 reads for ES1 222,821 reads for ES2, 415,867 reads for LS1 and finally 205,013 reads for LS2. Table 1 gathers descriptive statistics for all sequencing libraries.

5.2 Reference Transcriptome Assembly

The tuber reference transcriptome contains a total of 37,585 unique sequences, ranging from 52 to 8,776 bp (mean length = 862 bp, N50 = 1,027 bp) with a GC content of 44.8%. Distribution of reference contig lengths is visible in Figure 8. The total number of bases is 32,403,582 bp. As the first RNA-seq study for yellow nutsedge, this reference transcriptome increases genetic knowledge by more than a 400-fold difference for this species. Indeed, there are 10 sequences publicly available for *Cyperus esculentus*, totalizing 7,283 bp (data from NCBI, November 2013).



Stage	Statistic	Raw reads	Clean reads
ES1	number	574,573	547,98
	range (bp)	29-1,802	50-1,802
	mean length (bp)	383	384
	GC%	45	46
ES2	number	228,829	222,821
	range (bp)	49-1,274	52-1,240
	mean length (bp)	579	284
	GC%	43	44
LS1	number	423,779	415,867
	range (nt)	40-1,233	50-1,233
	mean length (bp)	401	391
	GC%	44	45
LS2	number	208,549	205,013
	range (bp)	51-2,044	52-2,044
	mean length (bp)	560	278
	GC%	44	45

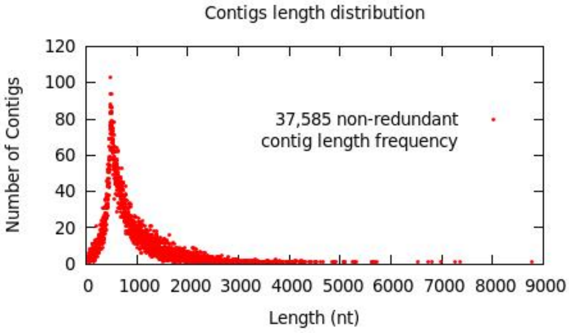


Figure 8 – Contig length distribution of the reference transcriptome composed of 37,585 different transcripts (up).

Table 1 – Statistics of sequencing libraries and outcome of quality control steps (left).

5.3 Functional Annotations

The reference transcriptome underwent annotation against TAIR9, UniProt-KB and Genbank-NR databases. As a result, 93% (34,823) of contigs found a significant hit in at least one of the pre-cited databases, leaving 7% of completely un-annotated contigs (Fig. 9a). By ranking plant species according to the number of hits they yielded (Fig. 9b), a prevalence of monocots over dicots is observable, such as rice (*Oryza sativa*), maize (*Zea mays*) and sorghum (*Sorghum bicolor*). Pearson's correlation coefficient between the number of hits per species and number of Genbank-NR entries is r^2=0.18, indicating that top-species are completely unlinked to an actual over-representation in public databases of these very species (Appendix 3). Distribution of the number of annotated contigs according to cellular component, molecular function and biological process is visible in Appendix 4, with a distinction between early and late stages. The latter graph shows no particular differences between both developmental phases.

After BLASTx annotations, the whole set of reference contigs was submitted to Blast2GO-pipeline and screened for protein signatures with InterProScan (IPS). 24,365 contigs (65%) were found to have at least one hit in IPS search. BLASTx and IPS hits allowed a total of 11,399 contigs (30%) to be associated to one or several GO terms. Exactly 140 different biological pathways are represented in tuber sequencings, with 3,818 sequences (10%) associated. Most represented pathways are *purine metabolism* (541 transcripts), *starch and sucrose metabolism* (447 transcripts) and *glycolysis/gluconeogenesis* with 280 transcripts.

When it comes to the number of enzymes detected, the *purine metabolism* is still first with 54 different enzymes, next comes the *amino-sugar and nucleotide-sugar metabolism* (42 different enzymes) and also the *cysteine and methionine metabolism* with 35 enzymes (Tab. 2).

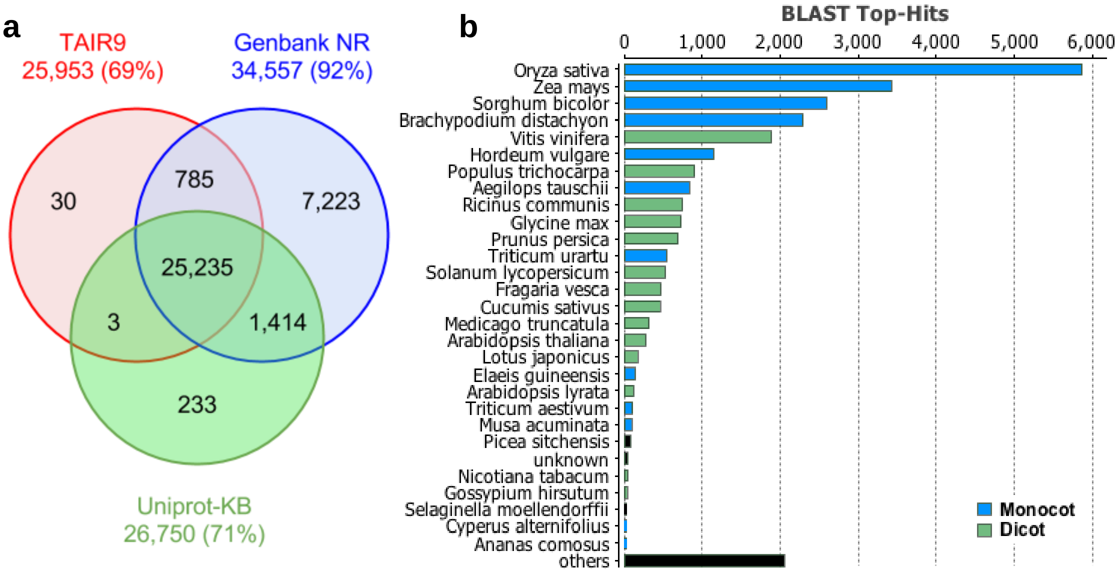
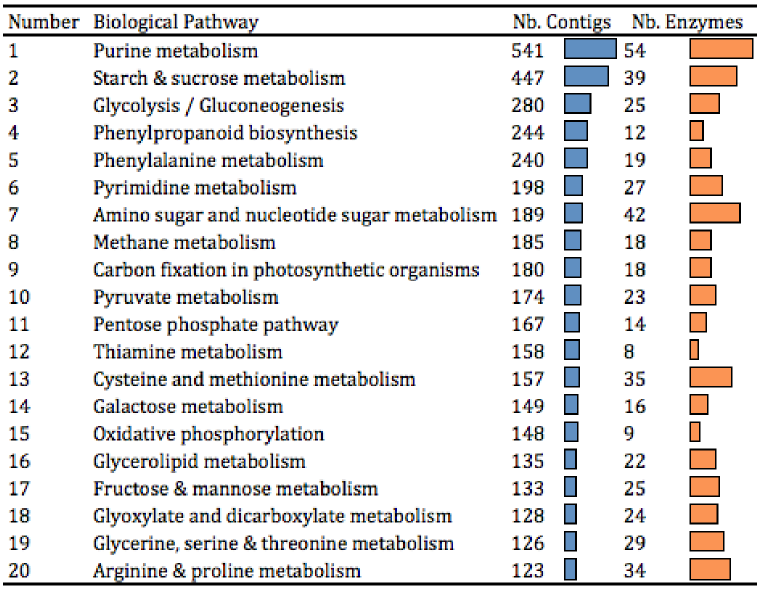


Figure 9 – Results of annotation steps: **a)** Out of a total of 37,585 reference transcripts, 7% remains without any hit, whereas 93% were annotated in at least one of the three databases and 67% are annotated unilaterally. Uniprot-KB is divided in two databases, namely Swissprot (50% of hits) and TrEMBL (21% of hits). **b)** Ranking of species according to the number of hits they provided. Black boxes are taxa different of monocots or dicots, such as *Picea sitchensis* (gymnosperm) and *Selaginella moellendorffii* (lycopod). The three first species are rice, maize and sorghum, adding up to 31% of hits (respectively 5,872, 3,429 and 2,600 hits).



Number	Biological Pathway	Nb. Contigs	Nb. Enzymes
1	Purine metabolism	541	54
2	Starch & sucrose metabolism	447	39
3	Glycolysis / Gluconeogenesis	280	25
4	Phenylpropanoid biosynthesis	244	12
5	Phenylalanine metabolism	240	19
6	Pyrimidine metabolism	198	27
7	Amino sugar and nucleotide sugar metabolism	189	42
8	Methane metabolism	185	18
9	Carbon fixation in photosynthetic organisms	180	18
10	Pyruvate metabolism	174	23
11	Pentose phosphate pathway	167	14
12	Thiamine metabolism	158	8
13	Cysteine and methionine metabolism	157	35
14	Galactose metabolism	149	16
15	Oxidative phosphorylation	148	9
16	Glycerolipid metabolism	135	22
17	Fructose & mannose metabolism	133	25
18	Glyoxylate and dicarboxylate metabolism	128	24
19	Glycerine, serine & threonine metabolism	126	29
20	Arginine & proline metabolism	123	34

Table 2 – Top 20 biological pathways according to the number of associated contigs (blue boxes). The number of enzymes for each pathway is also available (orange) and was counted on the basis of EC codes.

5.4 Expression Analyses

Bowtie2 mapping of clean reads back onto the reference transcriptome allowed retrieving of read count for each contig, at each of the sequencing stage. For ES1, 98.1% of reads were mapped onto the reference and 98.4% for ES2. When it comes to the late developmental stages, 97.3% of LS1 reads were mapped and 97.7% for LS2. To account for differences in sequencing yields and basepair length of each contig, readcounts were normalized using the RPKM method. Figure 10 shows the distributions of readcounts and RPKMs for each of the sequencings.

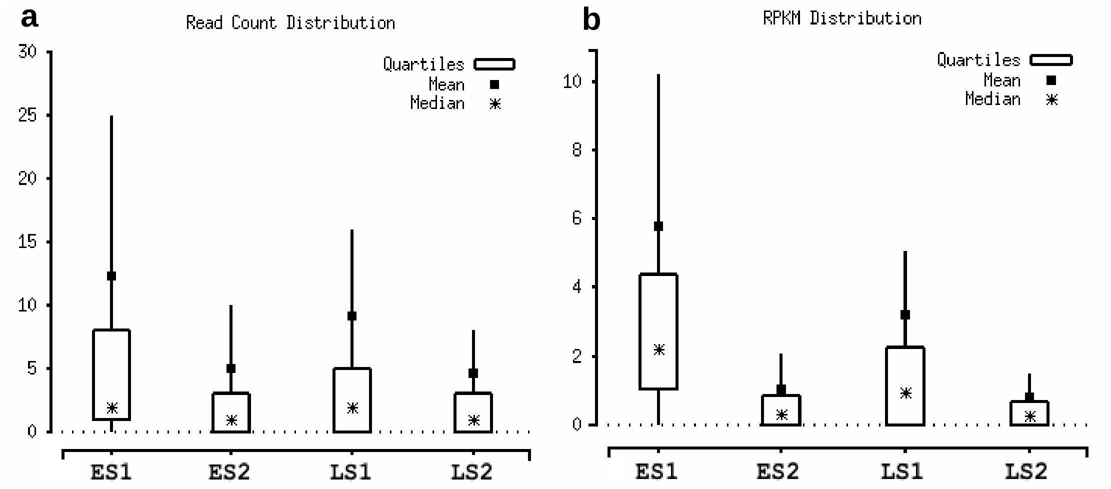
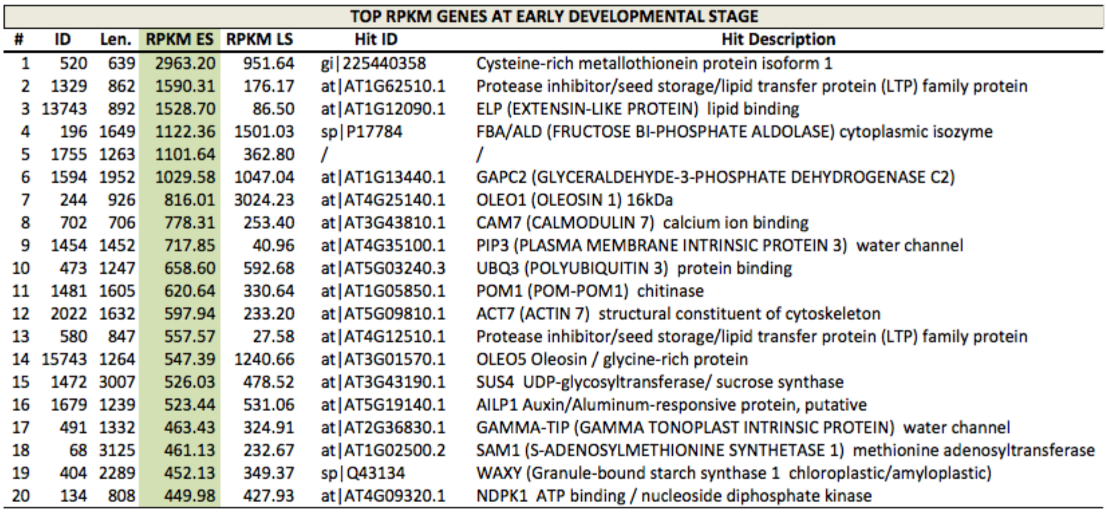


Figure 10 – Distribution of expression: Read count (a), and normalized RPKM values (b) for each library. Boxes are 25-75% quartiles and whiskers are the 10-90% percentiles. Median and mean values are respectively depicted with asterisks and black squares. All sequencings have the great majority of transcript expressions crushed down, but a very few of them reaches high level of expression (not displayed).

5.4.1 Individual stages: most expressed transcripts

Before looking at the differential expressions, e.g. the changes between developmental stages, expressions of genes at individual time points are considered. To do so, the 20-most expressed contigs for the early and the late developmental step are respectively presented in Tab. 3 and Tab. 4. Ranking was processed on the basis of RPKM expression, which allows direct comparison between different contigs.

Tab. 3 provides details of genes with highest expression at the early stage. The most expressed gene, by far, is a cysteine-rich metallothionein protein (MT), which outranges all the other contigs of the top-20. MTs have a high affinity for essential cations such as zinc (Zn^{2+}) and copper (Cu^{2+}). Their high abundance is frequently associated to a high metabolic activity (Moffatt & Denizeau 1997). A member of the actin gene family, ACT7, occupies the 12th place. ACT7 is commonly considered as a housekeeping gene, constitutively expressed, but has also been found to be highly expressed in vegetative tissues and organs undergoing rapid growth (McDowell et al. 1996). As members of the CTL family are strongly assumed to be essential for cell walls building (Zhang et al. 2004), POM1 (also known as CTL1, chitinase-like protein) illustrates again the intense growth and development at early stage. Despite manual researches, the 5th highest expressed contig remained completely un-annotated. Water channels are also present with PIP3 and GAMMA-TIP, facilitating the transport of water molecules through intracellular membranes.



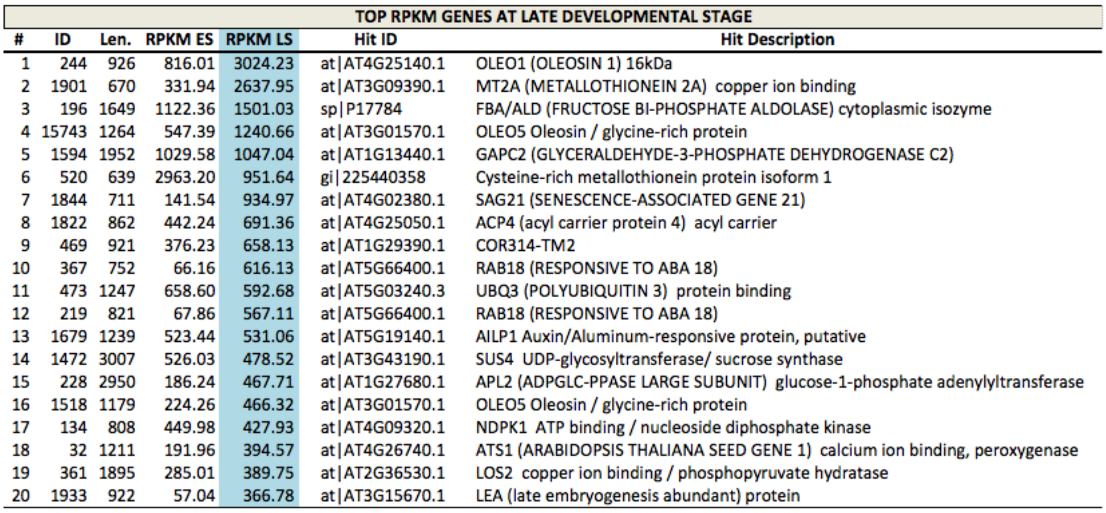
TOP RPKM GENES AT EARLY DEVELOPMENTAL STAGE						
#	ID	Len.	RPKM ES	RPKM LS	Hit ID	Hit Description
1	520	639	2963.20	951.64	gi|225440358	Cysteine-rich metallothionein protein isoform 1
2	1329	862	1590.31	176.17	at|AT1G62510.1	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
3	13743	892	1528.70	86.50	at|AT1G12090.1	ELP (EXTENSIN-LIKE PROTEIN) lipid binding
4	196	1649	1122.36	1501.03	sp|P17784	FBA/ALD (FRUCTOSE BI-PHOSPHATE ALDOLASE) cytoplasmic isozyme
5	1755	1263	1101.64	362.80	/	/
6	1594	1952	1029.58	1047.04	at|AT1G13440.1	GAPC2 (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C2)
7	244	926	816.01	3024.23	at|AT4G25140.1	OLEO1 (OLEOSIN 1) 16kDa
8	702	706	778.31	253.40	at|AT3G43810.1	CAM7 (CALMODULIN 7) calcium ion binding
9	1454	1452	717.85	40.96	at|AT4G35100.1	PIP3 (PLASMA MEMBRANE INTRINSIC PROTEIN 3) water channel
10	473	1247	658.60	592.68	at|AT5G03240.3	UBQ3 (POLYUBIQUITIN 3) protein binding
11	1481	1605	620.64	330.64	at|AT1G05850.1	POM1 (POM-POM1) chitinase
12	2022	1632	597.94	233.20	at|AT5G09810.1	ACT7 (ACTIN 7) structural constituent of cytoskeleton
13	580	847	557.57	27.58	at|AT4G12510.1	Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein
14	15743	1264	547.39	1240.66	at|AT3G01570.1	OLEO5 Oleosin / glycine-rich protein
15	1472	3007	526.03	478.52	at|AT3G43190.1	SUS4 UDP-glycosyltransferase/ sucrose synthase
16	1679	1239	523.44	531.06	at|AT5G19140.1	AILP1 Auxin/Aluminum-responsive protein, putative
17	491	1332	463.43	324.91	at|AT2G36830.1	GAMMA-TIP (GAMMA TONOPLAST INTRINSIC PROTEIN) water channel
18	68	3125	461.13	232.67	at|AT1G02500.2	SAM1 (S-ADENOSYLMETHIONINE SYNTHETASE 1) methionine adenosyltransferase
19	404	2289	452.13	349.37	sp|Q43134	WAXY (Granule-bound starch synthase 1 chloroplastic/amyloplastic)
20	134	808	449.98	427.93	at|AT4G09320.1	NDPK1 ATP binding / nucleoside diphosphate kinase

Table 3 – Details of the 20 highest expressed transcripts at early developmental stage based on RPKM value (green).

Several contigs are enzymes belonging to the glycolysis pathway. The first one is the fructose biphosphate aldolase (FBA or ALD, #4), which uses zinc as cofactors to catalyze a reversible reaction in the glycolysis pathway (EC:4.1.2.13). This reaction involves fructose-1,6-biphosphate as substrate and can lead to PGAL (phosphoglyceraldehyde) or DHAP (dihydroxiacetone phosphate) (Schaftingen et al. 1982). The next enzyme is the cytosolic GAPC2 (glyceraldehyde-3-phosphate dehydrogenase C2, #), continuing the glycolysis reaction from PGAL to DPGA, e.g. 1,3-diphosphoglycerate (EC:1.2.1.12).

One also notices some lipid-related annotations, such as the two seed-storage LTPs (lipid transfer proteins) at the second and 13th position, shuttling phospholipids between membranes (Kader 1996). Oleosin family is also represented with high mRNA abundances of OLEO1 and OLEO5. The hairpin structure of the oleosin proteins holds a hydrophobic central domain that interacts with TAGs in order to form a single-membrane oil body (Huang 1996). Finally, the extensin-like protein (ELP, #3) has a very broad annotation but contains a conserved domain identified as HSP (hydrophobic protein from Soybean). A lipid-binding activity has been assumed from its structure, but no further function is currently known (Baud et al. 1993).

When it comes to the starch pathway, the WAXY gene (also known as GBSS1, EC:2.4.1.242) has a crucial role because it synthesizes amylose (Murai et al. 1999). The sucrose synthase gene (SUS4) also belongs to the starch and sucrose metabolism, as a member of the glycosyltransferase family. SUS4 preferentially catalyzes the production of UDP-glucose and fructose from sucrose (EC:2.4.1.13) (Geigenberger & Stitt 1993). This latter enzyme is found with the 15th highest expression in this early developmental stage.



TOP RPKM GENES AT LATE DEVELOPMENTAL STAGE						
#	ID	Len.	RPKM ES	RPKM LS	Hit ID	Hit Description
1	244	926	816.01	3024.23	at|AT4G25140.1	OLEO1 (OLEOSIN 1) 16kDa
2	1901	670	331.94	2637.95	at|AT3G09390.1	MT2A (METALLOTHIONEIN 2A) copper ion binding
3	196	1649	1122.36	1501.03	sp|P17784	FBA/ALD (FRUCTOSE BI-PHOSPHATE ALDOLASE) cytoplasmic isozyme
4	15743	1264	547.39	1240.66	at|AT3G01570.1	OLEO5 Oleosin / glycine-rich protein
5	1594	1952	1029.58	1047.04	at|AT1G13440.1	GAPC2 (GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C2)
6	520	639	2963.20	951.64	gi|225440358	Cysteine-rich metallothionein protein isoform 1
7	1844	711	141.54	934.97	at|AT4G02380.1	SAG21 (SENESCENCE-ASSOCIATED GENE 21)
8	1822	862	442.24	691.36	at|AT4G25050.1	ACP4 (acyl carrier protein 4) acyl carrier
9	469	921	376.23	658.13	at|AT1G29390.1	COR314-TM2
10	367	752	66.16	616.13	at|AT5G66400.1	RAB18 (RESPONSIVE TO ABA 18)
11	473	1247	658.60	592.68	at|AT5G03240.3	UBQ3 (POLYUBIQUITIN 3) protein binding
12	219	821	67.86	567.11	at|AT5G66400.1	RAB18 (RESPONSIVE TO ABA 18)
13	1679	1239	523.44	531.06	at|AT5G19140.1	AILP1 Auxin/Aluminum-responsive protein, putative
14	1472	3007	526.03	478.52	at|AT3G43190.1	SUS4 UDP-glycosyltransferase/ sucrose synthase
15	228	2950	186.24	467.71	at|AT1G27680.1	APL2 (ADPGLC-PPASE LARGE SUBUNIT) glucose-1-phosphate adenylyltransferase
16	1518	1179	224.26	466.32	at|AT3G01570.1	OLEO5 Oleosin / glycine-rich protein
17	134	808	449.98	427.93	at|AT4G09320.1	NDPK1 ATP binding / nucleoside diphosphate kinase
18	32	1211	191.96	394.57	at|AT4G26740.1	ATS1 (ARABIDOPSIS THALIANA SEED GENE 1) calcium ion binding, peroxygenase
19	361	1895	285.01	389.75	at|AT2G36530.1	LOS2 copper ion binding / phosphopyruvate hydratase
20	1933	922	57.04	366.78	at|AT3G15670.1	LEA (late embryogenesis abundant) protein

Table 4 – Details of the 20 highest expressed transcripts at late developmental stage based on RPKM value (blue).

Tab. 4 provides details of genes with highest expression at the late stage. This top-20 is very similar to the early stage top-20, given that more than 50% of the contigs have the same annotation as previously observed.

The highest expression is shown by OLEO1, a protein member of the oleosin family. As in the early stage top-20, another member of the family is present, e.g. OLEO5, but this time with two different contigs (#4 and 16). A highly conserved calcium-binding domain located in Arabidopsis thaliana Seed Gene 1 (ATSG1) classifies this gene in the caleosin family. Caleosins are oleosin-like proteins, highly expressed in *A. thaliana* mature seeds, where they are largely associated with storage lipid bodies (Murphy et al. 2000). The only other gene related to oil, in this top-20, is the acyl carrier protein (ACP isoform 4, #8), binding activated fatty acids in the plastid (Ohlrogge et al. 1979).

As opposed to the early-stage top-20, metallothioneins are here illustrated with a new member: MT2A, which occupies the second place in expression ranking, and is specialized in binding copper cations. The MT isoform 1, previously seen, has been relayed to the 6th position. Another Cu^{2+} binding protein is LOS2 (#19), better known as 2-phosphoglycerate dehydratase (EC:4.2.1.11): a multifunctional enzyme involved in the 9th step of carbohydrate degradation, e.g. synthesis of phosphoenolpyruvate (PEP), and activation of cold-responsive genes transcription (Lee et al. 2002).

The 7th position is displayed with a senescence-associated gene (SAG21) belonging to the late embryogenesis-associated protein family (LEA), from which another member is found at position 20. These late embryogenesis abundant proteins (LEA) were first discovered in late phases of cottonseed growth (Dure et al. 1981), then linked to water and cold stress responses (Bray 1997; Thomashow 1999). APL2 is the large subunit of the chloroplastic glucose-1-phosphate adenylyltransferase (AGPase, EC:2.7.7.27), which catalyzes the first reaction of starch synthesis, hence playing a regulatory key-role (Iglesias 1994; Copeland & Preiss 1981). Finally, both RAB18 genes (responsive to abscisic acid gene 18) and COR314-TM2 (cold regulated 314 thylakoid membrane 2) are linked to a cold-treatment response (Mantyla et al. 1995).

5.4.2 Differentially Expressed Genes from Early to Late stage

Differential expression analysis from early to late developmental stages was processed using three different R packages, namely BaySeq, DESeq and edgeR. A total of 872 different contigs were detected as differentially expressed gene (DEGs) by at least one of the packages, but only 186 were detected unilaterally (Fig. 11a).

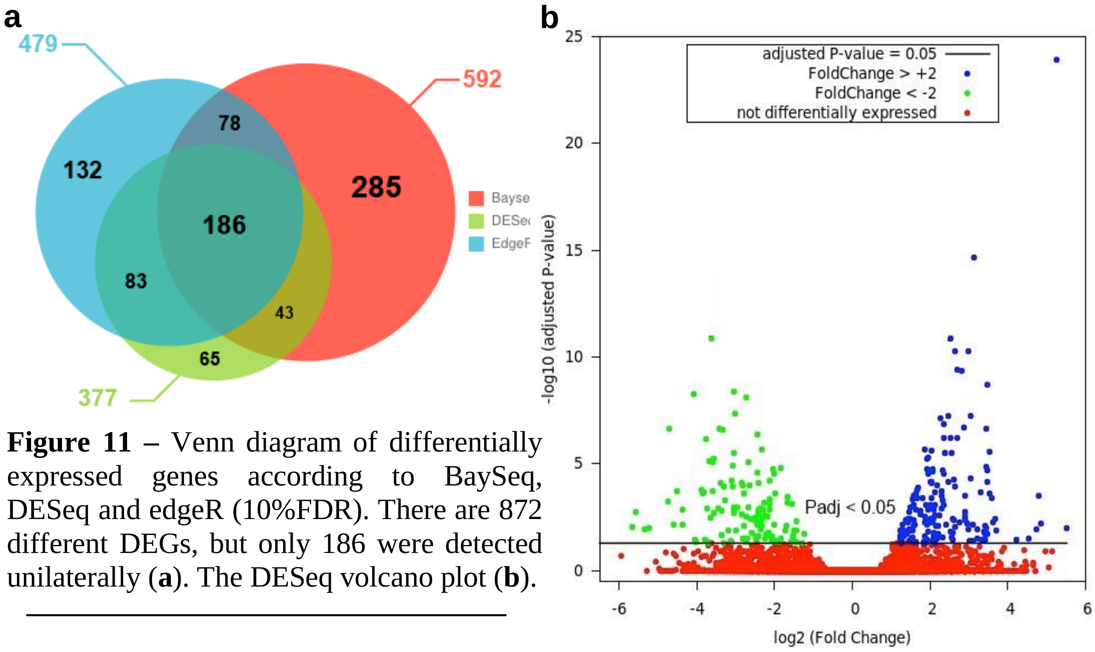


Figure 11 – Venn diagram of differentially expressed genes according to BaySeq, DESeq and edgeR (10%FDR). There are 872 different DEGs, but only 186 were detected unilaterally (**a**). The DESeq volcano plot (**b**).

Regardless of the intensities, 65% (564) of all DEGs were down-regulated, whereas 35% (308) are up-regulated. Difference of the same magnitude is found when looking at the cross-confirmed DEGs, with 60% (111) down-regulated genes compared to 40% (75) up-regulated.

Details of top-30 up-regulated and top-30 down-regulated genes are respectively available in Table 5 and Table 6, and will be presented in the two next points. For the sake of length and clarity, only metrics from DESeq package are exhibited. Moreover, from all packages used in this study, DESeq appears to be the most popular software with up to 916 citations, compared to BaySeq (133) and edgeR (626) (data: Google Scholar, 16/11/2013). The DESeq volcano plot (Fig. 11b) shows distribution of contigs according to the intensity of differential expression (fold change) and direction (up or down expressed, from early to later stage), along with the statistical significance (adjusted P-value). The upper left and right corners are the most interesting regions as they display contigs with both high expression change and significance (Cui & Churchill 2003).

Top-30 up-regulated

The 30 up-regulated transcripts (Tab. 5) display 6 entries with an uninformative annotation (#1, 6, 8, 19, 27 and 30 - yellow). Included is the first transcript with the highest FC (Fold Change), which found a hit in *Drosophila* species but no further detail has been found about this sequence.

#	ID	FC	P-adj	Hit ID	Hit Description
1	20690	46.11	9.74E-03	gi|194759382	GF14692 [Drosophila ananassae]
2	13306	38.17	1.27E-24	at|AT5G13930.1	TT4 (TRANSPARENT TESTA 4) naringenin-chalcone synthase
3	24275	28.91	6.08E-03	at|AT4G10250.1	HSP22 heat shock protein
4	11531	27.72	3.23E-04	at|AT3G54320.1	WRI1 (WRINKLED 1) DNA binding TF
5	11553	26.71	1.13E-02	at|AT1G18100.1	MFT (E12A11 ORF) phosphatidylethanolamine binding
6	6768	23.05	2.95E-02	at|AT3G12960.1	Putative uncharacterized protein
7	1777	18.69	3.47E-02	at|ATCG00020.1	Encodes chlorophyll binding protein D1
8	6785	14.87	8.50E-02	at|AT3G21190.1	Unknown protein
9	9849	13.06	1.23E-02	sp|Q0WRC9	F-box protein SKIP17
10	34401	12.28	5.53E-03	at|AT3G04080.1	APY1 (APYRASE 1) ATPase/ calmodulin binding / nucleotide diphosphatase
11	34404	12.15	3.99E-04	at|AT1G18100.1	E12A11 phosphatidylethanolamine binding
12	27355	11.63	2.28E-04	at|AT1G01470.1	LEA14 (LATE EMBRYOGENESIS ABUNDANT 14)
13	3906	11.48	2.77E-06	at|AT3G50360.1	CEN2 (CENTRIN2) Caltractin - calcium ion binding
14	11550	11.21	7.26E-05	at|AT3G04080.1	APY1 (APYRASE 1) ATPase/ calmodulin binding / nucleotide diphosphatase
15	2150	11.05	1.86E-02	at|AT3G45090.1	2-phosphoglycerate kinase-related
16	13246	11.05	2.11E-05	at|AT1G66340.1	ETR1 (ETHYLENE RESPONSE 1) ethylene binding / protein histidine kinase
17	15587	11.05	1.93E-09	at|AT4G21020.1	Late embryogenesis abundant (LEA) domain-containing protein
18	11324	10.89	2.36E-07	at|AT1G17810.1	BETA-TIP (BETA-TONOPLAST INTRINSIC PROTEIN) water channel
19	8391	10.79	1.09E-02	gi|147783381	Hypothetical protein
20	17118	10.78	1.85E-02	tr|D2JX40	ASR1 Abscisic stress ripening
21	213	10.77	7.89E-03	at|AT1G67230.1	LINC1 (LITTLE NUCLEI1)
22	233	10.76	3.52E-02	sp|A1EA46	50S ribosomal protein L16 chloroplastic
23	3139	10.52	1.42E-02	gi|115486507	Extensin-like protein
24	20545	10.05	3.70E-02	sp|P15214	Glutathione S-transferase GST-6.0
25	452	9.92	2.20E-04	at|AT1G66340.1	ETR1 (ETHYLENE RESPONSE 1) ethylene binding
26	6446	9.76	7.15E-02	at|AT5G23750.2	Remorin family protein
27	20137	9.21	1.14E-02	gi|195097629	Unknown protein
28	2044	8.53	6.86E-03	at|ATCG00770.1	Chloroplast 30S ribosomal protein S8
29	1337	8.52	4.12E-02	at|AT4G35160.1	Trans-resveratrol di-O-methyltransferase
30	13288	8.18	4.71E-05	at|AT1G56660.1	Unknown protein

Table 5 – Top-30 up-regulated genes from early to late tuber developmental stage, detected by all three packages (BaySeq, DESeq and edgeR). ID: contig identification number; FC: Fold Change; P-adj: adjusted P-value. All metrics are from DESeq package. Hit ID comes with the following priority: TAIR9 (at|), Uniprot-KB (sp|: Swissprot, tr|: TrEMBL), Genbank-NR (gi|). Here, no contig was associated to any biological pathway, hence not displayed.

In a 30-gene space, the fold change drops from 46-fold to a plateau of ca. 10-fold, showing that up-regulation is very rapidly decreasing, and only a few genes undergo an outstanding up-regulation between early and late stage (all adjusted p-values <0.05) (Appendix 4). Also, several contigs have the same or closely related annotations, such as #16 and 25 (ER1 genes, ethylene response 1 - green), #5 and 11 (MFT genes, mother of flowering locus T - blue), #10 and 14 (APY1 genes, apyrase - grey) or again #12 and 17 (LEA genes, late embryogenesis abundant protein - red). Despite confirmation is required one can assume these sets of contigs to be isoforms of the same transcripts, or mis-assemblies that should be merged together.

Several up-regulated genes are related to stress. Both ETR1 genes (Ethylene Response 1) show a fold change close to 10-fold. LEA (Late Embryogenesis Abundant) proteins are in the same magnitude with a 11-fold up-regulation. In Arabidospsis but also in other non-plant organisms, LEA proteins are synthesized in response to stress conditions such as cold, dessication and salinity (K et al. 2005; Gal et al. 2004; Hundertmark & Hincha 2008). Again, Heat Shock Protein (HSP22, #3) and abscisic stress (ASR1, #20) are highly up-regulated (28-fold and 11-fold) and enforce the assumption of a stress response between early and late developmental stages.

Transparent Testa 4 gene (TT4) displays the second highest up-regulation with a 38-fold change. This gene codes for a chalcone synthase (CHS) involved in the biosynthesis of flavonoids (Tohge et al. 2007). Flavonoids are, between many other things, the most important pigment for plant flowers and its very name originates from the Latin word '*flavus*': yellow (Winkel-Shirley 2001).

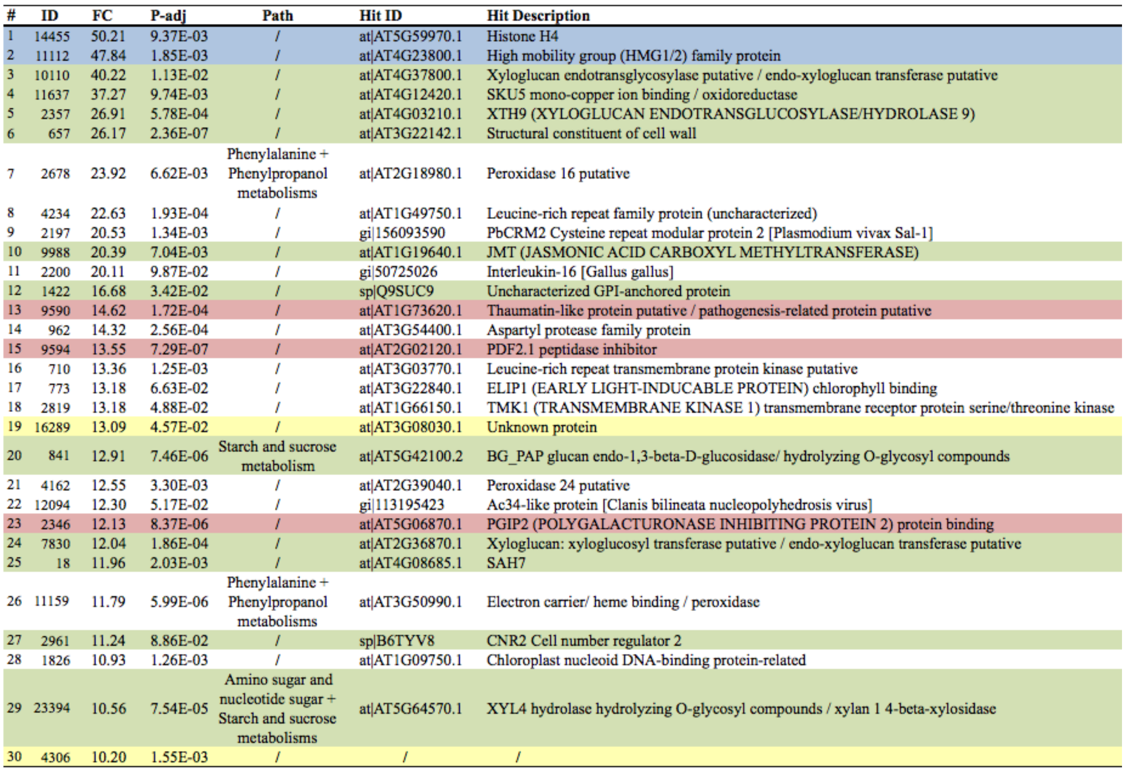
Wrinkled 1 (WRI1) gene occupies the 4th place with a strong 28X up-regulation. This ethylene-responsive transcription factor (TF) positively orchestrates the fatty acid biosynthesis in Arabidopsis seeds (Cernac & Benning 2004; Baud et al. 2007). The important up-regulation of WRI1 is consistent with the lipid accumulation (in the form of triacylglycerols, TAGs) of nutsedge tubers (Turesson et al. 2010). Interestingly, WRI1 and ER1 genes are linked with ethylene response mechanism and both show critical arise of abundance between time points.

Top-30 down-regulated

The top-30 down-regulated transcripts (Tab. 6) displays several genes with uninformative or missing hit such as #19 and #30 (yellow), but also 5 entries with a putative hit (#3, 7, 13, 16, and 24). All adjusted p-values are significant (p-adj < 0.05) and fold changes range from 50 to 10X. Comparatively to the top-30 up-regulated genes, here a handful of genes were linked to a biological pathway, such as *phenylalanine and phenylpropanol metabolisms* (#7 and 26), and *starch and sucrose metabolism* (#20 and 29).

First two genes, histone H4 and HMG1/2 (high mobility group) are related to chromatin structure (Grasser 1998), and show the highest down-regulation of 50-fold and 48-fold, respectively (blue).

Plant growth and especially genes related to cell wall development are greatly down-regulated and largely represented among this top-30, with at least 9 genes (#3-6, 10,12, 20, 24, 25, 27 and 29 - green). The great majority are related to cell wall, through synthesis and modification of xyloglucans, such as SKU5, XTH9, XYL4, BG_PAP, SAH7 and the GPI-anchored protein (Sedbrook 2002; Baumann et al. 2007; Hyodo et al. 2003; Borner et al. 2003). Equally, jasmonic o-methyltransferase (JMT, #10) underwent a 20-fold down-regulation. JMT catalyzes the formation of methyl jasmonate which is, among others, an important actor in plant growth and development, including tuberization process (Creelman & Mullet 1995; Seo et al. 2001).



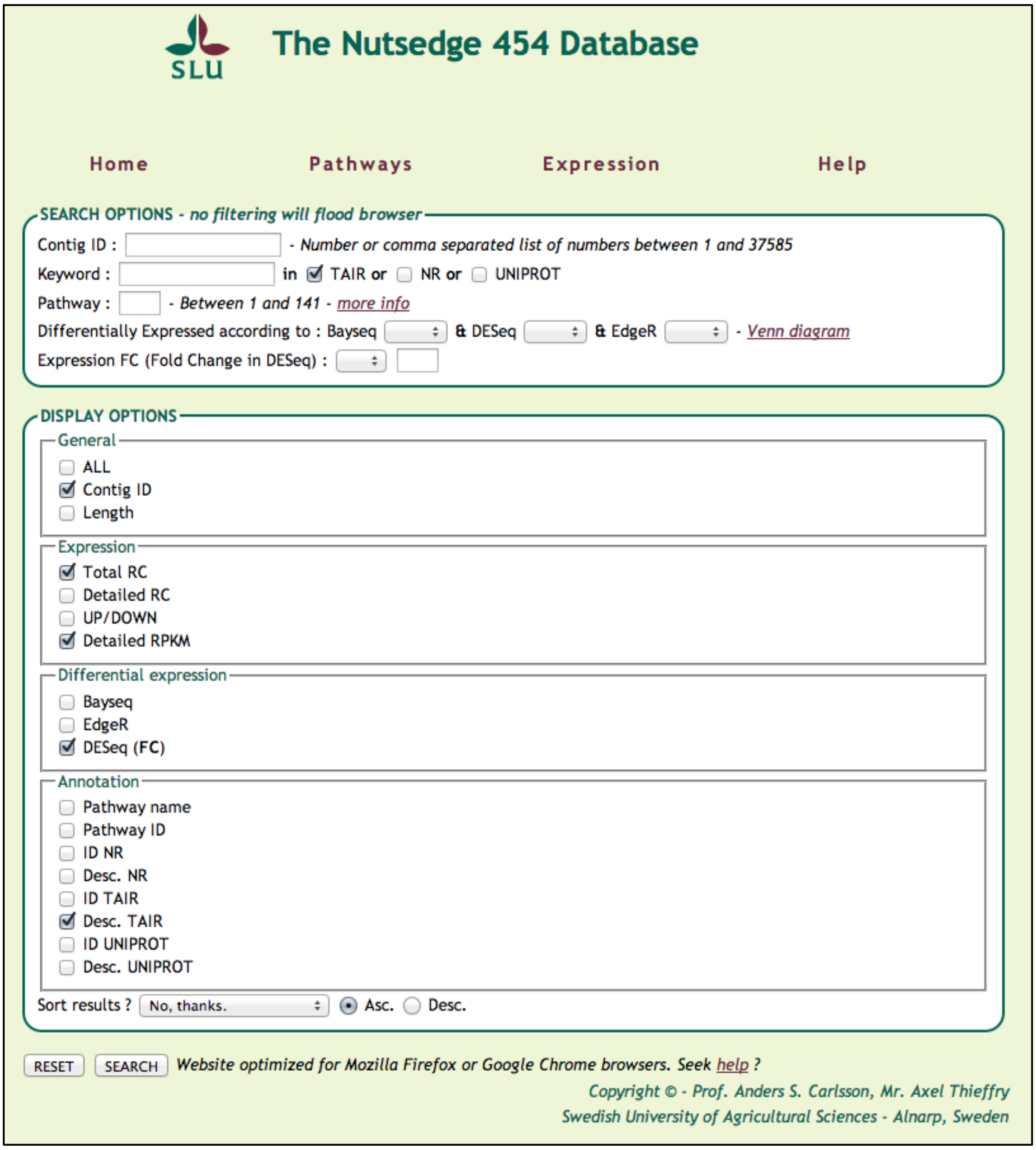
#	ID	FC	P-adj	Path	Hit ID	Hit Description
1	14455	50.21	9.37E-03	/	at|AT5G59970.1	Histone H4
2	11112	47.84	1.85E-03	/	at|AT4G23800.1	High mobility group (HMG1/2) family protein
3	10110	40.22	1.13E-02	/	at|AT4G37800.1	Xyloglucan endotransglycosylase putative / endo-xyloglucan transferase putative
4	11637	37.27	9.74E-03	/	at|AT4G12420.1	SKU5 mono-copper ion binding / oxidoreductase
5	2357	26.91	5.78E-04	/	at|AT4G03210.1	XTH9 (XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 9)
6	657	26.17	2.36E-07	/	at|AT3G22142.1	Structural constituent of cell wall
7	2678	23.92	6.62E-03	Phenylalanine + Phenylpropanol metabolisms	at|AT2G18980.1	Peroxidase 16 putative
8	4234	22.63	1.93E-04	/	at|AT1G49750.1	Leucine-rich repeat family protein (uncharacterized)
9	2197	20.53	1.34E-03	/	gi|156093590	PbCRM2 Cysteine repeat modular protein 2 [Plasmodium vivax Sal-1]
10	9988	20.39	7.04E-03	/	at|AT1G19640.1	JMT (JASMONIC ACID CARBOXYL METHYLTRANSFERASE)
11	2200	20.11	9.87E-02	/	gi|50725026	Interleukin-16 [Gallus gallus]
12	1422	16.68	3.42E-02	/	sp|Q9SUC9	Uncharacterized GPI-anchored protein
13	9590	14.62	1.72E-04	/	at|AT1G73620.1	Thaumatin-like protein putative / pathogenesis-related protein putative
14	962	14.32	2.56E-04	/	at|AT3G54400.1	Aspartyl protease family protein
15	9594	13.55	7.29E-07	/	at|AT2G02120.1	PDF2.1 peptidase inhibitor
16	710	13.36	1.25E-03	/	at|AT3G03770.1	Leucine-rich repeat transmembrane protein kinase putative
17	773	13.18	6.63E-02	/	at|AT3G22840.1	ELIP1 (EARLY LIGHT-INDUCABLE PROTEIN) chlorophyll binding
18	2819	13.18	4.88E-02	/	at|AT1G66150.1	TMK1 (TRANSMEMBRANE KINASE 1) transmembrane receptor protein serine/threonine kinase
19	16289	13.09	4.57E-02	/	at|AT3G08030.1	Unknown protein
20	841	12.91	7.46E-06	Starch and sucrose metabolism	at|AT5G42100.2	BG_PAP glucan endo-1,3-beta-D-glucosidase/ hydrolyzing O-glycosyl compounds
21	4162	12.55	3.30E-03	/	at|AT2G39040.1	Peroxidase 24 putative
22	12094	12.30	5.17E-02	/	gi|113195423	Ac34-like protein [Clanis bilineata nucleopolyhedrosis virus]
23	2346	12.13	8.37E-06	/	at|AT5G06870.1	PGIP2 (POLYGALACTURONASE INHIBITING PROTEIN 2) protein binding
24	7830	12.04	1.86E-04	/	at|AT2G36870.1	Xyloglucan: xyloglucosyl transferase putative / endo-xyloglucan transferase putative
25	18	11.96	2.03E-03	/	at|AT4G08685.1	SAH7
26	11159	11.79	5.99E-06	Phenylalanine + Phenylpropanol metabolisms	at|AT3G50990.1	Electron carrier/ heme binding / peroxidase
27	2961	11.24	8.86E-02	/	sp|B6TYV8	CNR2 Cell number regulator 2
28	1826	10.93	1.26E-03	/	at|AT1G09750.1	Chloroplast nucleoid DNA-binding protein-related
29	23394	10.56	7.54E-05	Amino sugar and nucleotide sugar + Starch and sucrose metabolisms	at|AT5G64570.1	XYL4 hydrolase hydrolyzing O-glycosyl compounds / xylan 1 4-beta-xylosidase
30	4306	10.20	1.55E-03	/	/	/

Table 6 – Top-30 down-regulated genes from early to late tuber developmental stage, detected by all three packages (BaySeq, DESeq and edgeR). ID: contig identification number; FC: Fold Change; P-adj: adjusted P-value. All metrics are from DESeq package but FC. Hit ID comes with the following priority: TAIR9 (at|), Uniprot-KB (sp|: Swissprot, tr|: TrEMBL), Genbank-NR (gi|). Only a few contigs were associated to a biological pathway.

Resistance and defense activities are also down-regulated and represented, though in a lesser extent, by 3 genes (#13, 15 and 23 - reddish). Thaumatins are pathogenesis-related proteins, usually triggered by viroid and fungi infection (Ruiz-Medrano et al. 1992), but also known for conferring a sweetness degree 10^5 times higher than sucrose on a molar basis (Wel & Loeve 1972). The PDF2.1 is a member of the cysteine-rich plant defensin family (PDF) (Sels et al. 2008), and finally PGIP2, the polygalacturonase inhibiting protein 2, an cell wall protein against phytopathogenic fungi (Toubart et al. 1992).

5.5 The Nutsedge 454 Database

An online website has been developed, gathering most of results from this work for in-house use. A MySQL database of the reference tuber transcriptome sequences, along with their annotation and the differential expression analyses, is questionable through a user-friendly web form (Fig. 12). The website has been predominantly optimized for Google Chrome, Safari and Mozilla Firefox browsers, other systems may experience layout problems. The Nutsedge 454 Database is available at www.plantlink.se/nutsedge.



SEARCH OPTIONS - *no filtering will flood browser*

Contig ID : - *Number or comma separated list of numbers between 1 and 37585*

Keyword : in ☑ TAIR **or** □ NR **or** □ UNIPROT

Pathway : - *Between 1 and 141* - *more info*

Differentially Expressed according to : Bayseq & DESeq & EdgeR - *Venn diagram*

Expression FC (Fold Change in DESeq) :

DISPLAY OPTIONS

General
- □ ALL
- ☑ Contig ID
- □ Length

Expression
- ☑ Total RC
- □ Detailed RC
- □ UP/DOWN
- ☑ Detailed RPKM

Differential expression
- □ Bayseq
- □ EdgeR
- ☑ DESeq (**FC**)

Annotation
- □ Pathway name
- □ Pathway ID
- □ ID NR
- □ Desc. NR
- □ ID TAIR
- ☑ Desc. TAIR
- □ ID UNIPROT
- □ Desc. UNIPROT

Sort results ? No, thanks. ● Asc. ○ Desc.

Website optimized for Mozilla Firefox or Google Chrome browsers. Seek help ?

Copyright © - Prof. Anders S. Carlsson, Mr. Axel Thieffry
Swedish University of Agricultural Sciences - Alnarp, Sweden

Figure 12 – Main page of The Nutsedge 454 Database website. The first part of this web form allows the user to graphically build a request (search options). The second part permits to choose the features that will be displayed (display options). Other pages of the website are not depicted, but reachable through the menu, and will provide with help, further details and succinct explanations, through the *Expression*, *Pathway* and *Help* pages.

DISCUSSION

Sampling – Harvesting of tubers from 10-15 different plants, prior pooling them according to DAI label, allowed to account for biological variation within the populations. Effects of hypothetical outliers are greatly diminished, as the mRNAs add-up and balance possible isolated differences.

Sequencings – A high sequencing depth is preferable for many aspects in RNA-seq. First, it will give a better coverage and allow detection of rare transcripts (Tarazona et al. 2011; Cai et al. 2012). Moreover, it raises confidence in the nature of assembled contigs for building a high quality reference transcriptome. It also improves accuracy in expressions measurements, especially between replicates (Cai et al. 2012; Toung et al. 2011). Read count distribution exhibits a very low average: the great majority of contigs have less than 5 reads mapped. Normalization using the RPKM method also exhibits a largely decreased expression for most of contigs. Throughout sequencing statistics to RC and RPKM distributions, one can observe a clear difference between replicates: first sequencings yielded twice more reads, which is automatically reverberated in RC and their RPKM normalization (Fig. 10). There is no significance in the number of reads generated in the first *vs*. second sequencing run as the number of usable reads was quite variable from the 454 GS FLX instrument.

Blast2GO Annotation – Blast2GO analysis is a time consuming step, especially with more than 37k contigs, and requires intensive online querying. Indeed, a free version of the software has a limited Internet bandwidth, which slows down the performances. Moreover, all free users are in competition for reaching the Blast2GO servers. However, to bypass these bottlenecks, it is possible to install the pipeline locally, as well as the large set of required database. Numerous hours were spent trying to reach this step. Eventually, incessant bugs and update incompatibilities between versions of MySQL, Java, InterProScan and The Gene Ontology made the free, online and slow Blast2GO to be the only solution. A dozen days was required for this analysis, even though BLASTx annotations were run locally and fed into the program, hence already saving this central step from being run by Blast2GO.

Pathway annotation – Association of contigs to KEGG biological pathways is based on enzyme codes (Conesa & Götz 2008). Consequently, other type of gene products such as transcription factors and structural proteins are not considered for this step. The best example to illustrate this principle is the WRI1 transcription factor, which is highly up-regulated in nutsedge tubers, extensively annotated through BLASTx, InterProScan and The Gene Ontology but, being not an enzyme, this transcription factor undoubtedly linked to the lipid biosynthesis is not associated to a KEGG pathway (Tab. 5). Also, contigs were stated as annotated when found a significant hit in any public database. Nevertheless, the term 'annotated' remains ambiguous because, even with a very high sequence similarity, the hit does necessarily provide useful information on its very nature, due to a *hypothetical*, *predicted*, *unknown* or again an *uncharacterized* protein. Actually, the only information given with such annotation is that this very transcript has a highly similar hit, which has already been sequenced and found in another organism, but nothing is known on its function. For those cases, the great majority of contigs is not linked to any biological pathway (Appendix 3). These points, along with other factors, provide the essential answers with regards to the low number of pathway-annotated contigs, e.g. 10% (3,818 out of 37,585).

In top-20 biological pathways (Tab. 2), the number of contigs associated to a biological pathway is often greater than the number of enzymes associated to the same pathway. For example, 135 transcripts were associated to the *Glycerolipid metabolism* pathway but, of them, only 22 enzymes were found. This may be justified by at least two events: (i) First, several contigs are actually the same gene (coding for an enzyme) but were dissociated by the assembly software due to sequencing errors, unsolved repetitive regions or unidentified artifacts. In that specific case, contigs would have to be merged and will probably benefit greater expression, but this requires manual curation. This hypothesis is in line with the fact that all reads did not align onto the reference contigs, when mapped back for the readcount retrieval step. Furthermore, (ii) a gene may have several truly different transcript isoforms because of alternative splicing events (Trapnell et al. 2010). For that case, each isoform has to be considered individually but, during the pathway annotation step, they are ultimately seen as originating from the same gene.

The GO-term graph (Appendix 4) shows the great majority of classes to be under-represented, e.g. less than 5% of the total number of transcripts. Nevertheless, there are clear outliers around 20%, such as *binding* and *catalytic* molecular functions, and *cellular* and *metabolic* biological processes. Still, one can barely observe differences between early stage (grey) and late stage (blue), all categories included. Despite the number of genes in a category does not allow to compare expressions, the fact that very few differences are observed between stage indicate that we may expect a limited number of switched-on/off genes.

Highest expressed genes – According to the annotations, genes highly expressed at the early stage are the witnesses of a high metabolic activity resulting from intense development and growth. Indeed, nutsedge tubers are undertaking heavy structural re-organization till ca. 7DAI (Turesson et al. 2010). Hence, elevated abundances for genes like metallothioneins (MTs), actin 7 and POM1 (cell wall building) seem legitimate in early stage, e.g. 5DAI. Moreover, it is assumable that the highly abundant MTs provide the fructose biphosphate aldolase (ALD) with Zn^{2+} cations as the necessary cofactor. The rapid accumulation of starch at 5DAI (Fig. 2) may be explained with the combined high abundances of the sucrose synthase (SUS4) and starch synthase (GBSS1).

The late stage does not display any important growth or development activity, but rather stress responses *via* the late embryogenesis proteins (LEAs), RAB18 and COR314-TM2. Still in the late phase, the already present OLEO genes gained in abundance significantly, reaching the first position, and even adding more members of the oleosin family, as well as closely-related caleosins. Also, ACP (acyl-carrier-proteins) appeared as sign of fatty acid biosynthesis activity. The fact that lipid-related genes arise at the late stage is in line with the onset of oil accumulation in the tubers, increasing at ca. 10 DAI (Fig. 2). Finally, the starch synthesis is still very present with a high expression of AGPase.

Differential Expression Analysis – Three differential expression methods were processed for the mRNA sequencing of nutsedge tubers: DESeq, edgeR and BaySeq. All of them assume an over-dispersed Poisson distribution of expressions: the negative binomial model. While DESeq and edgeR are known to yield similar results (Kvam et al. 2012; Yao & Yu 2011), BaySeq contrasts by relying on a Bayesian approach. This difference probably explains the greater number of contigs exclusively identified by BaySeq (285) compared to the other packages individually did, e.g. 65 for edgeR and 132 for DESeq (Fig. 11a). BaySeq has been shown to be better at establishing significance of differential expression, and uses the data to obtain initial parameters of overdispersion (Kvam et al. 2012; Zhou et al. 2011). On the other hand, the *likelihood* metric used by BaySeq is often difficult to grasp, and no indication of expression change intensity has been found in BaySeq package, e.g. fold change. In an ideal world, all packages should rather agree on the same DE gene set, but the Venn diagram definitely shows there are considerable differences between packages that may be due, at least partially, to a too high FDR percentage.

The fold change method, where the expression of late stage is divided by the expression of early stage to get DE intensity, seems legitimate and is easily understood. Nevertheless, this somehow naïve approach can be challenging. For example, if one contig is not present or detected at the early stage, FC calculation will lead to a division by zero. The contigs in that case should be discarded for further analysis. If a too large proportion of genes are concerned, a way to bypass this is to add a small arbitrary constant value (such as 0.01) to zero-expression stage, such as seen in microarray experiments (Black & Doerge 2002). But this technique is highly criticized as it leads to extreme fold changes, which may not be representative: a very practical example would be an undetected contig at early stage but having a very small RC of 2 at late stage, providing a fold change of 200X.

Top-30 up and down regulated genes – Most of genes from the top-30 up/down regulated lists were expected from the analysis of highly expressed genes at individual stages. Up-regulations have been confirmed for responses related to water and cold stresses through LEA, ASR1, HSP22 and ETR1 genes. The initial and important activity of development and growth, also perceived individually at the early phase analysis, is strongly down-regulated (SKU5, XTH9, SAH7, JMT, CNR2, XYL4 genes). No lipid biosynthesis enzyme, either from fatty acid or triacylglycerol pathway, was found in any of the top-30 up/down-regulated transcripts. It is strongly assumable that enzymes related to oil, starch and sugar, will show up with lower fold changes, and are masked in these top-30s because of greater changes featured by growth, development and stress activities. Nevertheless, the WRI1 transcription factor governing fatty acid synthesis is intensely up-regulated (27X).

The Nutsedge 454 Database – As a first version, The Nutsedge 454 online database has reached a fully functional and efficient stage. Through this website, data and up-to-date results on the *Cyperus esculentus* project have been made more convenient to access, especially when compared to an obscure UNIX command-line framework.

CONCLUSIONS & FURTHER DIRECTIVES

This present work was the first in-depth analysis of the *Cyperus esculentus* tuber transcriptome, allowing to produce a large amount of brand new genetic information with regards to this unique species when it comes to nutrient biosynthesis and accumulation in the tubers. Using the Roche 454 sequencing platform, a reference transcriptome of more than 37k sequences was assembled and extensively annotated with biological functions, ontology, and pathway associations.

Differential expression analyses from early to late developmental stages were processed with different bioinformatic packages and allowed a cross-confirmed set of 186 transcripts to be investigated. Several ranking of top-DEGs showed intense growth and development activities of tubers at the early phase, later on strongly down-regulated in favor of nutrient building. Starch and sucrose metabolism featured a very reduced number of abundant enzymes, but playing a key role in their respective pathway. No lipid biosynthesis enzymes were found in the rankings, but structural proteins for lipid-storage (oleosins and caleosins) as well as lipid-binding and transport proteins, e.g. ACP and LTP, were highly expressed. Moreover, the WRI1 transcription factor governing fatty acids biosynthesis underwent a 27-fold up-regulation.

Results were compiled in a web-interfaced database (plantlink.ltj.slu.se/nutsedge) offering a user-friendly environment for further private consultations and exploits of the data and generated results.

It has to be noticed that analysis of top-ranking genes according to their expression is the necessary first step to point out tremendous changes on the transcriptome level. On the other hand, it is also a very limiting view when investigation on a focused subject, e.g. lipids, is desired.

To dig further into the changes related to oil in yellow nutsedge, examination of differentially expressed genes according to the relevant pathways may yield more pertinent results for enzymes. Such research will certainly involve the need for extensive manual curation of the investigated transcripts. Also, searching for transcription factors using binding site databases could give further insights on the regulatory level.

Finally, a recent sequencing of 8 different developmental stages of nutsedge tubers was processed using Solexa/Illumina platform, along with a complete transcriptome sample. Illumina sequencing provides shorter read lengths but the number of reads is considerably greater than Roche 454, hence allowing expression analyses of higher accuracy. This 8-phases time series will certainly provide with an increased number of results and investigation opportunities.

ACKNOWLEDGEMENTS

I sincerely thank my supervisors Erik Bongcam-Rudloff and Anders Carlsson for providing me with the opportunity to integrate their groups at the Swedish University of Agricultural Sciences, as well as the numerous and highly-valuable discussions with regards to this project, but also efficient solutions to side-concerns. Equally, I thank my examiner Göran Andersson for interesting discussions and lectures throughout this Master studies.

I would like to communicate my gratitude to Estelle Wera, first on a professional level with help on the bioinformatics and the Plantlink server, but also as a friend with devoted support and trust.

Also, I address my highest respects to David Coornaert for being the cornerstone of my education and motivation towards bioinformatics and providing enlightening support and encouragements.

Last but certainly not least, I deeply thank my family and my beloved girlfriend, supporting and waiting for me throughout this adventure in Sweden.

REFERENCES

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2231712 [Accessed November 7, 2013].

Anders, S. & Huber, W., 2010. Differential expression analysis for sequence count data. *Genome biology*, 11(10), p.R106. Available at: http://genomebiology.com/2010/11/10/R106 [Accessed November 6, 2013].

Ashburner, M. et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1), pp.25–9. Available at: http://dx.doi.org/10.1038/75556 [Accessed November 11, 2013].

Bähler, J., Wilhelm, B.T. & Landry, J.-R., 2009. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3), pp.249–257. Available at: http://www.sciencedirect.com/science/article/pii/S1046202309000632 [Accessed December 3, 2013].

Baud, F. et al., 1993. Crystal Structure of Hydrophobic Protein from Soybean; a Member of a New Cysteine-rich Family. *Journal of Molecular Biology*, 231(3), pp.877–887. Available at: http://www.sciencedirect.com/science/article/pii/S0022283683713343 [Accessed December 8, 2013].

Baud, S. et al., 2007. WRINKLED1 specifies the regulatory action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis. *The Plant journal : for cell and molecular biology*, 50(5), pp.825–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17419836 [Accessed November 11, 2013].

Baumann, M.J. et al., 2007. Structural evidence for the evolution of xyloglucanase activity from xyloglucan endo-transglycosylases: biological implications for cell wall metabolism. *The Plant cell*, 19(6), pp.1947–63. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1955714&tool=pmcentrez&rendertype=abstract [Accessed November 25, 2013].

Bendixen, L.E. & Nandihalli, U.B., 2008. Worldwide Distribution of Purple and Yellow Nutsedge (Cyperus rotundus and C. esculentus). Available at: http://www.jstor.org/stable/info/3986985 [Accessed December 9, 2013].

Black, M.A. & Doerge, R.W., 2002. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, 18(12), pp.1609–1616. Available at: http://bioinformatics.oxfordjournals.org/content/18/12/1609.short [Accessed December 9, 2013].

Borner, G.H.H. et al., 2003. Identification of glycosylphosphatidylinositol-anchored proteins in Arabidopsis. A proteomic and genomic analysis. *Plant physiology*, 132(2), pp.568–77. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=166998&tool=pmcentrez&rendertype=abstract [Accessed November 25, 2013].

Bray, E.A., 1997. Plant responses to water deficit. *Trends in Plant Science*, 2(2), pp.48–54. Available at: http://www.sciencedirect.com/science/article/pii/S1360138597825629 [Accessed December 9, 2013].

Cai, G. et al., 2012. Accuracy of RNA-Seq and its dependence on sequencing depth. *BMC bioinformatics*, 13 Suppl 1(Suppl 13), p.S5. Available at: http://www.biomedcentral.com/1471-2105/13/S13/S5 [Accessed November 8, 2013].

Carlsson, A.S. et al., 2011. Replacing fossil oil with fresh oil - with what and for what? *European journal of lipid science and technology : EJLST*, 113(7), pp.812–831. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3210827&tool=pmcentrez&rendertype=abstract [Accessed November 14, 2013].

Cernac, A. & Benning, C., 2004. WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in Arabidopsis. *The Plant journal : for cell and molecular biology*, 40(4), pp.575–85. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15500472 [Accessed November 11, 2013].

Conesa, A. & Götz, S., 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, 2008, p.619832. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2375974&tool=pmcentrez&rendertype=abstract [Accessed November 7, 2013].

Copeland, L. & Preiss, J., 1981. Purification of Spinach Leaf ADPglucose Pyrophosphorylase. *PLANT PHYSIOLOGY*, 68(5), pp.996–1001. Available at: http://www.plantphysiol.org/content/68/5/996 [Accessed December 9, 2013].

Creelman, R. a & Mullet, J.E., 1995. Jasmonic acid distribution and action in plants: regulation during development and response to biotic and abiotic stress. *Proceedings of the National Academy of Sciences of the United States of America*, 92(10), pp.4114–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=41895&tool=pmcentrez&rendertype=abstract.

Cui, X. & Churchill, G.A., 2003. Statistical tests for differential expression in cDNA microarray experiments. *Genome biology*, 4(4), p.210. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=154570&tool=pmcentrez&rendertype=abstract [Accessed December 1, 2013].

Defelice, M.S., 2002. Yellow Nutsedge Cyperus esculentus L.—Snack Food of the Gods 1. *Weed Technology*, 16(4), pp.901–907. Available at: http://dx.doi.org/10.1614/0890-037X(2002)016[0901:YNCELS]2.0.CO;2 [Accessed December 2, 2013].

Dure, L., Greenway, S.C. & Galau, G.A., 1981. Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by in vitro and in vivo protein synthesis. *Biochemistry*, 20(14), pp.4162–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7284317 [Accessed December 9, 2013].

Gal, T.Z., Glazer, I. & Koltai, H., 2004. An LEA group 3 family member is involved in survival of C. elegans during exposure to stress. *FEBS Letters*, 577(1), pp.21–26. Available at: http://www.sciencedirect.com/science/article/pii/S0014579304011779 [Accessed November 24, 2013].

Garber, M. et al., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods*, 8(6), pp.469–77. Available at: http://dx.doi.org/10.1038/nmeth.1613 [Accessed November 7, 2013].

Geigenberger, P. & Stitt, M., 1993. Sucrose synthase catalyses a readily reversible reaction in vivo in developing potato tubers and other plant tissues. *Planta*, 189(3), pp.329–39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24178489 [Accessed December 12, 2013].

Grabherr, M.G. et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7), pp.644–52. Available at: http://dx.doi.org/10.1038/nbt.1883 [Accessed November 6, 2013].

Grasser, K.D., 1998. HMG1 and HU proteins: architectural elements in plant chromatin. *Trends in Plant Science*, 3(7), pp.260–265. Available at: http://www.sciencedirect.com/science/article/pii/S136013859801259X [Accessed November 25, 2013].

Greenbaum, D. et al., 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology*, 4(9), p.117. Available at: http://genomebiology.com/2003/4/9/117 [Accessed November 23, 2013].

Hardcastle, T.J. & Kelly, K.A., 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1), p.422. Available at: http://www.biomedcentral.com/1471-2105/11/422 [Accessed November 15, 2013].

Holm, L.G. et al., 1977. The world's worst weeds. Available at: http://www.cabdirect.org/abstracts/19770359792.html;jsessionid=914633C5393E0EBB45F87C340C2876D7?freeview=true [Accessed December 2, 2013].

Huang, A.H., 1996. Oleosins and oil bodies in seeds and other organs. *Plant physiology*, 110(4), pp.1055–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=160879&tool=pmcentrez&rendertype=abstract [Accessed December 8, 2013].

Hundertmark, M. & Hincha, D.K., 2008. LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana. *BMC genomics*, 9(1), p.118. Available at: http://www.biomedcentral.com/1471-2164/9/118 [Accessed November 8, 2013].

Hyodo, H. et al., 2003. Active gene expression of a xyloglucan endotransglucosylase/hydrolase gene, XTH9, in inflorescence apices is related to cell elongation in Arabidopsis thaliana. *Plant Molecular Biology*, 52(2), pp.473–482. Available at: http://link.springer.com/article/10.1023/A:1023904217641 [Accessed November 25, 2013].

Iglesias, A., 1994. Characterization of the kinetic, regulatory, and structural properties of ADP-glucose pyrophosphorylase from Chlamydomonas reinhardtii. *PLANT PHYSIOLOGY*, 104(4), pp.1287–1294. Available at: http://www.plantphysiol.org/content/104/4/1287 [Accessed December 9, 2013].

Jarvie, T. & Harkins, T., 2008a. Transcriptome sequencing with the Genome Sequencer FLX system. , 5(9).

Jarvie, T. & Harkins, T., 2008b. Transcriptome sequencing with the Genome Sequencer FLX system. , 5(9).

K, G., L, W. & A, T., 2005. LEA proteins prevent protein aggregation due to water stress. Available at: http://www.biochemj.org/bj/388/bj3880151.htm [Accessed November 24, 2013].

Kader, J.-C., 1996. LIPID-TRANSFER PROTEINS IN PLANTS. *Annual review of plant physiology and plant molecular biology*, 47, pp.627–654. Available at: http://www.annualreviews.org/doi/abs/10.1146/annurev.arplant.47.1.627?journalCode=arplant.2 [Accessed December 8, 2013].

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp.27–30. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract [Accessed November 7, 2013].

Kvam, V.M., Liu, P. & Si, Y., 2012. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American journal of botany*, 99(2), pp.248–56. Available at: http://www.amjbot.org/content/99/2/248.full#sec-17 [Accessed November 8, 2013].

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–9. Available at: http://dx.doi.org/10.1038/nmeth.1923 [Accessed November 7, 2013].

Lee, H. et al., 2002. LOS2, a genetic locus required for cold-responsive gene transcription encodes a bi-functional enolase. *The EMBO journal*, 21(11), pp.2692–702. Available at: http://dx.doi.org/10.1093/emboj/21.11.2692 [Accessed December 9, 2013].

Makareviciene, V. et al., 2013. Opportunities for the use of chufa sedge in biodiesel production. *Industrial Crops and Products*, 50, pp.633–637. Available at: http://www.sciencedirect.com/science/article/pii/S0926669013004536 [Accessed December 2, 2013].

Mantyla, E., Lang, V. & Palva, E.T., 1995. Role of Abscisic Acid in Drought-Induced Freezing Tolerance, Cold Acclimation, and Accumulation of LT178 and RAB18 Proteins in Arabidopsis thaliana. *Plant Physiology*, 107(1), pp.141–148. Available at: http://www.plantphysiol.org/content/107/1/141.short [Accessed December 9, 2013].

McDowell, J.M. et al., 1996. The arabidopsis ACT7 actin gene is expressed in rapidly developing tissues and responds to several external stimuli. *Plant physiology*, 111(3), pp.699–711. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=157885&tool=pmcentrez&rendertype=abstract.

Moffatt, P. & Denizeau, F., 1997. Metallothionein in physiological and physiopathological processes. *Drug metabolism reviews*, 29(1-2), pp.261–307. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9187522 [Accessed December 5, 2013].

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7), pp.621–628. Available at: http://dx.doi.org/10.1038/nmeth.1226.

Mulligan, G.A. & Junkins, B.E., 1976. THE BIOLOGY OF CANADIAN WEEDS. 17. Cyperus esculentus L. *Canadian Journal of Plant Science*, 56(2), pp.339–350. Available at: http://pubs.aic.ca/doi/abs/10.4141/cjps76-052 [Accessed December 9, 2013].

Murai, J., Taira, T. & Ohta, D., 1999. Isolation and characterization of the three Waxy genes encoding the granule-bound starch synthase in hexaploid wheat. *Gene*, 234(1), pp.71–79. Available at: http://www.sciencedirect.com/science/article/pii/S037811199900178X [Accessed December 8, 2013].

Murphy, D.J. et al., 2000. New insights into the mechanisms of lipid-body biogenesis in plants and other organisms. *Biochemical Society transactions*, 28(6), pp.710–1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11171180 [Accessed December 8, 2013].

Negbi, M., 2008. A Sweetmeat Plant, a Perfume Plant and Their Weedy Relatives: A Chapter in the History of Cyperus esculentus L. and C. rotundus L. Available at: http://www.jstor.org/stable/info/4255409 [Accessed December 9, 2013].

Ohlrogge, J.B., Kuhn, D.N. & Stumpf, P.K., 1979. Subcellular localization of acyl carrier protein in leaf protoplasts of Spinacia oleracea. *Proceedings of the National Academy of Sciences of the United States of America*, 76(3), pp.1194–8. Available at: http://www.pnas.org/content/76/3/1194.abstract [Accessed December 9, 2013].

Oshlack, A. & Wakefield, M.J., 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biology direct*, 4, p.14. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2678084&tool=pmcentrez&rendertype=abstract [Accessed November 7, 2013].

Owen-Hughes, T. & Engeholm, M., 2007. Pyrosequencing positions nucleosomes precisely. *Genome biology*, 8(6), p.217. Available at: http://genomebiology.com/2007/8/6/217 [Accessed December 10, 2013].

Pascual, B. et al., 2000. Chufa (Cyperus esculentus L. var. sativus boeck.): An unconventional crop. studies related to applications and cultivation. *Economic Botany*, 54(4), pp.439–448. Available at: http://link.springer.com/10.1007/BF02866543 [Accessed December 9, 2013].

Pertea, G. et al., 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, 19(5), pp.651–652. Available at: http://bioinformatics.oxfordjournals.org/content/19/5/651.short [Accessed November 11, 2013].

Quevillon, E. et al., 2005. InterProScan: protein domains identifier. *Nucleic acids research*, 33(Web Server issue), pp.W116–20. Available at: http://nar.oxfordjournals.org/content/33/suppl_2/W116.short [Accessed November 7, 2013].

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), pp.139–40. Available at: http://bioinformatics.oxfordjournals.org/content/26/1/139.long [Accessed November 6, 2013].

Ruiz-Medrano, R. et al., 1992. Nucleotide sequence of an osmotin-like cDNA induced in tomato during viroid infection. *Plant Molecular Biology*, 20(6), pp.1199–1202. Available at: http://link.springer.com/10.1007/BF00028909 [Accessed December 4, 2013].

Schaftingen, E. et al., 1982. A Kinetic Study of Pyrophosphate: Fructose-6-Phosphate Phosphotransferase from Potato Tubers. Application to a Microassay of Fructose 2,6-Bisphosphate. *European Journal of Biochemistry*, 129(1), pp.191–195. Available at: http://doi.wiley.com/10.1111/j.1432-1033.1982.tb07039.x [Accessed December 8, 2013].

Sedbrook, J.C., 2002. The Arabidopsis SKU5 Gene Encodes an Extracellular Glycosyl Phosphatidylinositol-Anchored Glycoprotein Involved in Directional Root Growth. *THE PLANT CELL ONLINE*, 14(7), pp.1635–1648. Available at: http://www.plantcell.org/content/14/7/1635.short [Accessed November 25, 2013].

Sels, J. et al., 2008. Plant pathogenesis-related (PR) proteins: A focus on PR peptides. *Plant Physiology and Biochemistry*, 46(11), pp.941–950. Available at: http://www.sciencedirect.com/science/article/pii/S0981942808001137 [Accessed December 4, 2013].

Seo, H.S. et al., 2001. Jasmonic acid carboxyl methyltransferase: a key enzyme for jasmonate-regulated plant responses. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8), pp.4788–93. Available at: http://www.pnas.org/content/98/8/4788.short [Accessed November 25, 2013].

Steen, J.A. & Cooper, M.A., 2011. Fluorogenic pyrosequencing in microreactors. *Nature methods*, 8(7), pp.548–9. Available at: http://dx.doi.org/10.1038/nmeth.1634 [Accessed December 10, 2013].

Stoller, E.W., Nema, D.P. & Bhan, V.M., 1972. Yellow Nutsedge Tuber Germination and Seedling Development Yellow Nutsedge Tuber Germination and Seedling Development ’. *Weed Science Society of American*, 20(1), pp.93–97.

Stoller, E.W. & Sweet, R.D., 2013. Biology and Life Cycle of Purple and Yellow Nutsedges (Cyperus rotundus and C . esculentus)’. , 1(1), pp.66–73.

Tarazona, S. et al., 2011. Differential expression in RNA-seq: a matter of depth. *Genome research*, 21(12), pp.2213–23. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3227109&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2013].

Thomashow, M.F., 1999. PLANT COLD ACCLIMATION: Freezing Tolerance Genes and Regulatory Mechanisms. *Annual review of plant physiology and plant molecular biology*, 50, pp.571–599. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15012220 [Accessed December 6, 2013].

Tohge, T. et al., 2007. Phytochemical genomics in Arabidopsis thaliana: A case study for functional identification of flavonoid biosynthesis genes. *Pure and Applied Chemistry*, 79(4), pp.811–823. Available at: http://iupac.org/publications/pac/79/4/0811/ [Accessed November 24, 2013].

Toubart, P. et al., 1992. Cloning and characterization of the gene encoding the endopolygalacturonase-inhibiting protein (PGIP) of Phaseolus vulgaris L. *The Plant Journal*, 2(3), pp.367–373. Available at: http://www.blackwell-synergy.com/links/doi/10.1046/j.1365-313X.1992.t01-35-00999.x [Accessed November 25, 2013].

Toung, J.M. et al., 2011. RNA-sequence analysis of human B-cells. *Genome research*, 21(6), pp.991–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3106332&tool=pmcentrez&rendertype=abstract [Accessed November 12, 2013].

Trapnell, C. et al., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5), pp.511–5. Available at: http://dx.doi.org/10.1038/nbt.1621 [Accessed November 6, 2013].

Turesson, H. et al., 2010. Characterization of oil and starch accumulation in tubers of Cyperus esculentus var. sativus (Cyperaceae): A novel model system to study oil reserves in nonseed tissues. *American journal of botany*, 97(11), pp.1884–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21616827 [Accessed November 22, 2013].

Wang, C.-H. et al., 2011. A magnetic bead-based assay for the rapid detection of methicillin-resistant Staphylococcus aureus by using a microfluidic system with integrated loop-mediated isothermal amplification. *Lab on a chip*, 11(8), pp.1521–31. Available at: http://pubs.rsc.org/en/content/articlehtml/2011/lc/c0lc00430h [Accessed December 10, 2013].

Wang, Z., Gerstein, M. & Snyder, M., 2010. RNA-Seq : a revolutionary tool for transcriptomics. , 10(1), pp.57–63.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), pp.57–63. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstract.

Wel, H. & Loeve, K., 1972. Isolation and Characterization of Thaumatin I and II, the Sweet-Tasting Proteins from Thaumatococcus daniellii Benth. *European Journal of Biochemistry*, 31(2), pp.221–225. Available at: http://doi.wiley.com/10.1111/j.1432-1033.1972.tb02522.x [Accessed December 4, 2013].

Winkel-Shirley, B., 2001. Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. *PLANT PHYSIOLOGY*, 126(2), pp.485–493. Available at: http://www.plantphysiol.org/content/126/2/485.short [Accessed November 11, 2013].

Yao, J.Q. & Yu, F., 2011. DEB: A web interface for RNA-seq digital gene expression analysis. *Bioinformation*, 7(1), pp.44–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3163933&tool=pmcentrez&rendertype=abstract [Accessed November 20, 2013].

Ye, J. et al., 2006. WEGO: a web tool for plotting GO annotations. *Nucleic acids research*, 34(Web Server issue), pp.W293–7. Available at: http://nar.oxfordjournals.org/content/34/suppl_2/W293.short [Accessed November 23, 2013].

Zhang, D. et al., 2004. Members of a new group of chitinase-like genes are expressed preferentially in cotton cells with secondary walls. *Plant molecular biology*, 54(3), pp.353–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15284492 [Accessed December 8, 2013].

Zhou, Y.-H., Xia, K. & Wright, F.A., 2011. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics (Oxford, England)*, 27(19), pp.2672–8. Available at: http://bioinformatics.oxfordjournals.org/content/27/19/2672.full [Accessed November 14, 2013].

APPENDIX 1 - Sequencing library preparation

Modified SMART cDNA Protocol for GS FLX Sequencing, following the CreatorTM SMARTTM cDNA Library Construction Kit (Clontech, cat. #K1053)

A. First-Strand cDNA Synthesis

(Clontech SMART cDNA library construction kit, cat. #634903)

1. Following reagents in 0.2 ml strip tubes:

- 1–3 µl RNA sample (0.025–0.5 µg poly A+ or 0.05–1.0 µg total RNA)
- 1 µl SMART IV oligo (10 µM)
- 1 µl MODIFIED CDS III/3' cDNA Synthesis Primer* (10 µM)
- If total volume is <5 µl, add deionized H_2O to bring volume up to 5 µl.

* MODIFIED CDS III/3' cDNA Synthesis Primer:
5' - TAG AGG CCG AGG CGG CCG ACA TGT TTT GTT TTT TTT TCT TTT TTT TTT VN - 3' (Order from IDT with PAGE purification)

2. Mix contents and spin the tube briefly in a microcentrifuge.
3. Incubate the tube at 72°C for 2 min.
4. Cool the tube on ice for 2 min.
5. Spin the tube briefly to collect the contents at the bottom.
6. Add the following to each reaction tube:

- 2.0 µl 5X First-Strand Buffer (Clontech)
- 1 µl DTT (20 mM)
- 1 µl dNTP Mix (10 mM)
- 1 µl SuperScriptII Reverse Transcriptase (Invitrogen)

Total volume: 10.0 µl

7. Mix the contents of the tube by gently pipetting and briefly spinning the tube.
8. Incubate the tube at 42°C for 1 hr. in an air incubator.
9. Place the tube on ice to terminate first-strand synthesis.
10. If you plan to proceed directly to the PCR step, take a 2-µl aliquot from the first-strand synthesis and place it in a clean, pre-chilled, 0.2-ml tube. Place tube on ice, and proceed to B.
11. Any first-strand reaction mixture that is not used right away should be placed at -20°C. First-strand cDNA can be stored at –20°C for up to three months.

B. cDNA Amplification by LD-PCR

(Clontech Advantage 2 Polymerase mix, cat. #639201 for 100 runs)

1. Combine the following components in the reaction tube:

- 2 µl First-Strand cDNA (from Step A.10)
- 80 µl Deionized H2O
- 10 µl 10X Advantage 2 PCR Buffer
- 2 µl dNTP Mix (10 mM)
- 2 µl 5' PCR Primer (10 µM)
- 2 µl MODIFIED-CDS III / 3' PCR Primer* (10 µM)
- 2 µl 50X Advantage 2 Polymerase Mix

Total volume: 100 µl

2. Mix contents by gently flicking the tube. Centrifuge briefly to collect the contents at the bottom of the tube.
3. Cap the tube and place it in a preheated (95°C) thermal cycler.
4. Commence thermal cycling using one of the following programs:

- 95°C 1 min
- 95°C 15 sec
- 68°C 6 min
- x cycles:

The SMART protocol refers to a table indicating how many PCR cycles are required depending on the initial amount of starting RNA material used for the 1st strand reaction:

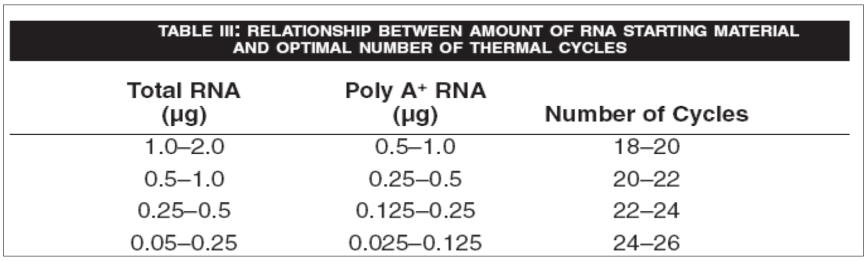


TABLE III: RELATIONSHIP BETWEEN AMOUNT OF RNA STARTING MATERIAL AND OPTIMAL NUMBER OF THERMAL CYCLES

Total RNA (µg)	Poly A$^+$ RNA (µg)	Number of Cycles
1.0–2.0	0.5–1.0	18–20
0.5–1.0	0.25–0.5	20–22
0.25–0.5	0.125–0.25	22–24
0.05–0.25	0.025–0.125	24–26

An alternate way to achieve this is by titration of the number of PCR cycles. Ten µl of 1st strand cDNA reaction allows for five 2nd strand reactions. We recommend doing 15, 18, 21, 24 and 27 cycles. Please be advised that over-cycling favors amplification of shorter fragments. A side-by-side cycle titration offers the chance to select the conditions that maximize the yield and the size distribution of the cDNA library.

5. When the cycling is complete, analyze a 5µl sample of the PCR product, alongside 0.1 µg of 1kb DNA size markers, on a 1.1% agarose/EtBr gel. The ds cDNA should appear as a 0.1–4kb smear on the gel, with some bright bands corresponding to the abundant mRNAs for that tissue or cell source. (cDNA prepared from some tissues may not have distinct bright bands, especially if the mRNA is highly complex.).

6. Proteinase K digestions: mix in the PCR tubes:

- 95 µl PCR product
- 4 µl Proteinase K
- Mix, spin briefly
- Incubate at 45ºC for 20 min (in thermocycler)
- Transfer to 0.5 ml tube (if using 0.2 ml tubes for PCR)
- Add 100 µl H2O
- Add 200 µl phenol/chloroform/isoamyl-alcohol (25:24:1)
- Mix 1-2 min by inversion
- Spin @ 13,000 rpm, room temp for 5 min
- Transfer aqueous phase to new 0.5 ml tube
- Add 200 µl chloroform/isoamyl-alcohol (24:1)
- Mix by inversion
- Spin 13,000 rpm, RT for 5 min
- Transfer aqueous phase to new 1.5 ml tube
    - At this point I pooled the aqueous phase from two PCR reactions to reduce the number of sizing columns I had to run
    - Volume was ~ 250 µl
- Add 25 µl 3M sodium acetate, pH 4.8
- Add 3.1 µl glycogen (20 µg/µl)
- Add 650 µl ethanol (room temp)
- Mix, then immediately spin at 13,000 rpm at room temp for 20 min
- Remove supernatant with a pipette
- Wash pellet with 200 µl of 80% Ethanol
- Air dry the pellet
- Re-suspend in 79 µl H2O

7. SfiI digest

- Transfer the 79 µl to a PCR tube
- Add 10 µl 10X SfiI buffer
- Add 10 µl SfiI enzyme
- Add 1 µl 100X BSA
- Incubate at 50ºC for 2 hr. (in thermocycler)

8. cDNA size fractionation

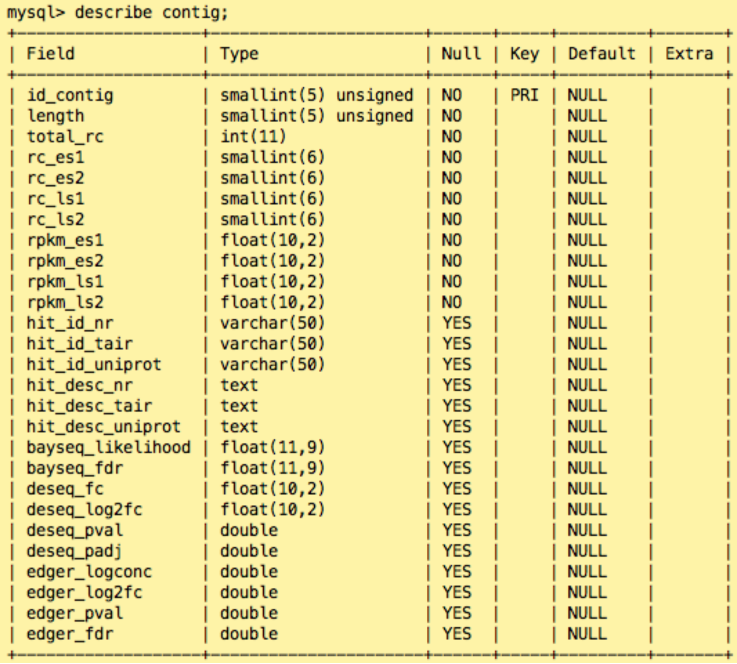
- While SfiI digest is going, prepare sizing columns
- Re-suspend column matrix, I used a Pasteur pipette
- Remove bottom cap from column and let column drain by gravity flow
    - If column matrix is much below the 1.0 ml mark, take some matrix from another column to bring the volume up to near 1.0 ml
    - Drops should be ~ 40ul each, if not, cap and re-suspend.
- Once column has drained, add 700 μl of column buffer and allow draining out. Column is ready for adding cDNA
- After SfiI digest is complete, add 2 μl of 1% xylene cyanol to cDNA
- Carefully pipette sample to the top-center of the matrix surface
- Allow the sample to fully enter the column matrix
- Add 100 μl of column buffer to wash out the cDNA sample tube, then pipette to the column and allow draining out. (The dye will have moved several mm into the column)
- Add 600 μl of column buffer to column and immediately begin collecting 1-drop fractions into labeled 1.5 ml tubes.
- Check 3 μl of each fraction on an Agarose gel and decide which fractions to pool based on the intensity and size distribution. (I usually pooled 4-5 fractions)
- If there is any column matrix in the pooled sample, spin and transfer pooled sample to new tube
- Precipitate pooled cDNA
    - Add 1/10 volume of 3M sodium acetate, pH4.8
    - Add 1.3 μl of glycogen for every 100 μl of sample
    - Add 2.5 volume of 95% ethanol (-20ºC)
    - Leave overnight at -20ºC
    - Spin at 13,000 rpm at room temp for 20 min
    - Remove supernatant with a pipette
    - Briefly spin again and remove any residual liquid without disturbing the pellet
    - Air dry pellet
    - Re-suspend in an appropriate amount of TE (I used 15 μl)

Important Notes:

1) It is important to use the Clontech 1st strand buffer in the reverse transcription reaction with Invitrogen Superscript II. The buffer that comes with Superscript II did not work in our hands, most likely because of the MgCl2 concentration difference (15 mM Invitrogen vs. 30 mM Clontech).
2) Use Superscript II, NOT Superscript III. Superscript III supposedly lacks the terminal transferase activity that is required to add the additional nucleotides to the 3'ends of the 1st strand cDNA to allow the cap-switching mechanism to work.
3) The SfiI digest and size fractionation is included because in our most recent run with the GS-FLX, we saw a significant number of reads with multiple 5'-primers on the 5' end as well as a significant number of reads that had both the 5'primer and the 3'primer (these seemed to mostly be just short fragments of transcripts). The SfiI digest and size fractionation will eliminate these issues.

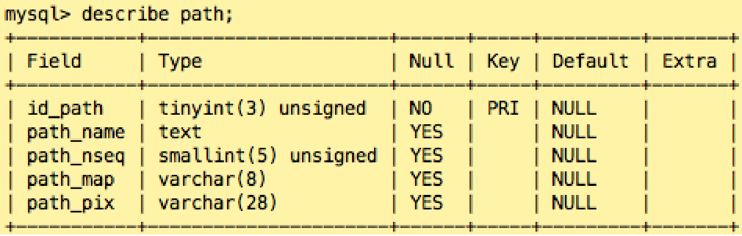
When re-suspending pellets after size fractionation, re-suspend one pellet with 15 μl of TE, then use the same 15 μl to re-suspend a second pellet to finally pool 4 PCR reactions in the same sample to have >5 μg in a volume that gives >300 ng/μl (based on estimating quantity by gel electrophoresis).

APPENDIX 2 – The Nutsedge 454 Database: MySQL tables



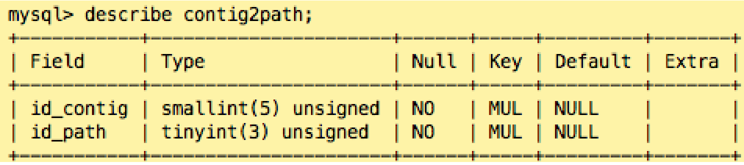
```
mysql> describe contig;
+-------------------+----------------------+------+-----+---------+-------+
| Field             | Type                 | Null | Key | Default | Extra |
+-------------------+----------------------+------+-----+---------+-------+
| id_contig         | smallint(5) unsigned | NO   | PRI | NULL    |       |
| length            | smallint(5) unsigned | NO   |     | NULL    |       |
| total_rc          | int(11)              | NO   |     | NULL    |       |
| rc_es1            | smallint(6)          | NO   |     | NULL    |       |
| rc_es2            | smallint(6)          | NO   |     | NULL    |       |
| rc_ls1            | smallint(6)          | NO   |     | NULL    |       |
| rc_ls2            | smallint(6)          | NO   |     | NULL    |       |
| rpkm_es1          | float(10,2)          | NO   |     | NULL    |       |
| rpkm_es2          | float(10,2)          | NO   |     | NULL    |       |
| rpkm_ls1          | float(10,2)          | NO   |     | NULL    |       |
| rpkm_ls2          | float(10,2)          | NO   |     | NULL    |       |
| hit_id_nr         | varchar(50)          | YES  |     | NULL    |       |
| hit_id_tair       | varchar(50)          | YES  |     | NULL    |       |
| hit_id_uniprot    | varchar(50)          | YES  |     | NULL    |       |
| hit_desc_nr       | text                 | YES  |     | NULL    |       |
| hit_desc_tair     | text                 | YES  |     | NULL    |       |
| hit_desc_uniprot  | text                 | YES  |     | NULL    |       |
| bayseq_likelihood | float(11,9)          | YES  |     | NULL    |       |
| bayseq_fdr        | float(11,9)          | YES  |     | NULL    |       |
| deseq_fc          | float(10,2)          | YES  |     | NULL    |       |
| deseq_log2fc      | float(10,2)          | YES  |     | NULL    |       |
| deseq_pval        | double               | YES  |     | NULL    |       |
| deseq_padj        | double               | YES  |     | NULL    |       |
| edger_logconc     | double               | YES  |     | NULL    |       |
| edger_log2fc      | double               | YES  |     | NULL    |       |
| edger_pval        | double               | YES  |     | NULL    |       |
| edger_fdr         | double               | YES  |     | NULL    |       |
+-------------------+----------------------+------+-----+---------+-------+
```

Appendix 2, Fig. 1: Details of table *contig*, gathering of basic descriptive statistics, as well as read count (RC) and normalized expression (RPKM) at each developmental stage (ES1, ES2, LS1, LS2). Annotations are stored for all three databases. Finally, differential expression analysis and metrics are available for the three R packages (DESeq, edgeR and BaySeq).



```
mysql> describe path;
+-----------+----------------------+------+-----+---------+-------+
| Field     | Type                 | Null | Key | Default | Extra |
+-----------+----------------------+------+-----+---------+-------+
| id_path   | tinyint(3) unsigned  | NO   | PRI | NULL    |       |
| path_name | text                 | YES  |     | NULL    |       |
| path_nseq | smallint(5) unsigned | YES  |     | NULL    |       |
| path_map  | varchar(8)           | YES  |     | NULL    |       |
| path_pix  | varchar(28)          | YES  |     | NULL    |       |
+-----------+----------------------+------+-----+---------+-------+
```

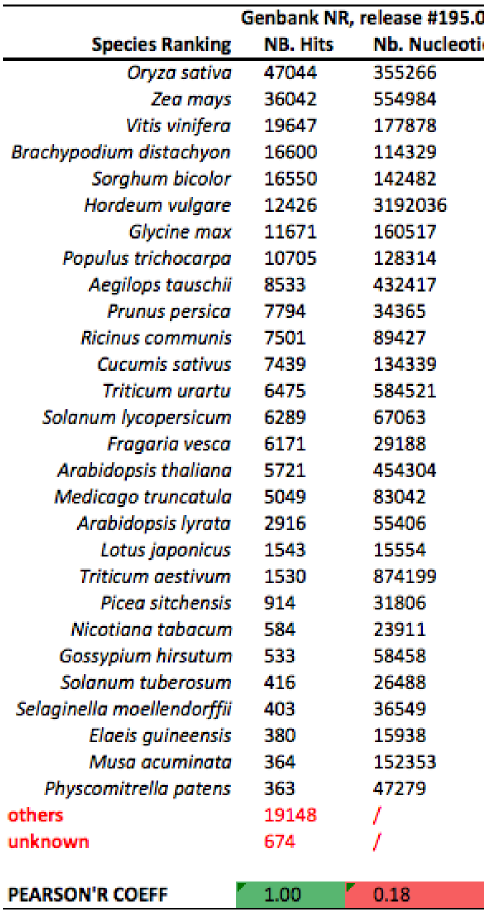
Appendix 2, Fig. 2: Details of table *path*, gathering information of each represented biological pathway in the tube reference transcriptome dataset.



```
mysql> describe contig2path;
+-----------+----------------------+------+-----+---------+-------+
| Field     | Type                 | Null | Key | Default | Extra |
+-----------+----------------------+------+-----+---------+-------+
| id_contig | smallint(5) unsigned | NO   | MUL | NULL    |       |
| id_path   | tinyint(3) unsigned  | NO   | MUL | NULL    |       |
+-----------+----------------------+------+-----+---------+-------+
```

Appendix 2, Fig. 3: Details of table *contig2path*, which links each contig to none, one, or several biological pathways, and *vice-versa* (many-to-many relationship).

APPENDIX 3 – Top-hit species and uninformative annotations



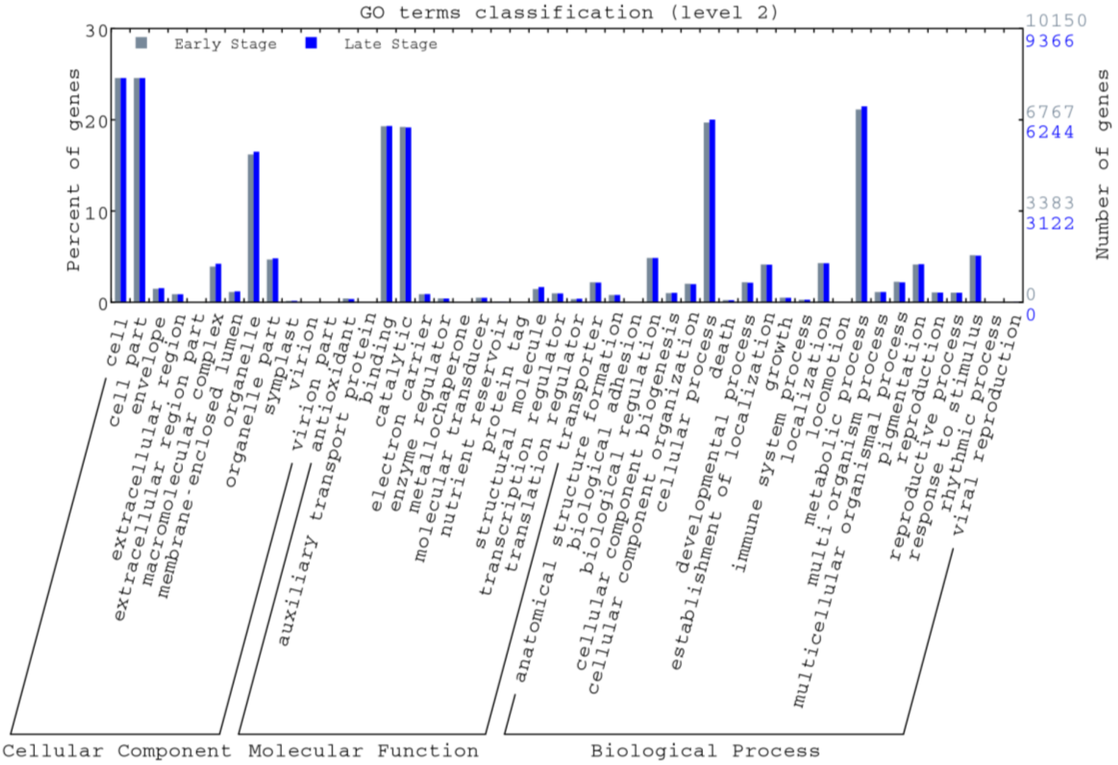
	Genbank NR, release #195.0	
Species Ranking	NB. Hits	Nb. Nucleoti
Oryza sativa	47044	355266
Zea mays	36042	554984
Vitis vinifera	19647	177878
Brachypodium distachyon	16600	114329
Sorghum bicolor	16550	142482
Hordeum vulgare	12426	3192036
Glycine max	11671	160517
Populus trichocarpa	10705	128314
Aegilops tauschii	8533	432417
Prunus persica	7794	34365
Ricinus communis	7501	89427
Cucumis sativus	7439	134339
Triticum urartu	6475	584521
Solanum lycopersicum	6289	67063
Fragaria vesca	6171	29188
Arabidopsis thaliana	5721	454304
Medicago truncatula	5049	83042
Arabidopsis lyrata	2916	55406
Lotus japonicus	1543	15554
Triticum aestivum	1530	874199
Picea sitchensis	914	31806
Nicotiana tabacum	584	23911
Gossypium hirsutum	533	58458
Solanum tuberosum	416	26488
Selaginella moellendorffii	403	36549
Elaeis guineensis	380	15938
Musa acuminata	364	152353
Physcomitrella patens	363	47279
others	19148	/
unknown	674	/
PEARSON'R COEFF	1.00	0.18

Appendix 3, Tab. 1: Correlation between the number of hits and the amount of nucleotidic sequences available in Genbank-NR (release 195) for each species. The very low coefficient (0.18) shows a complete independence between these two criteria.

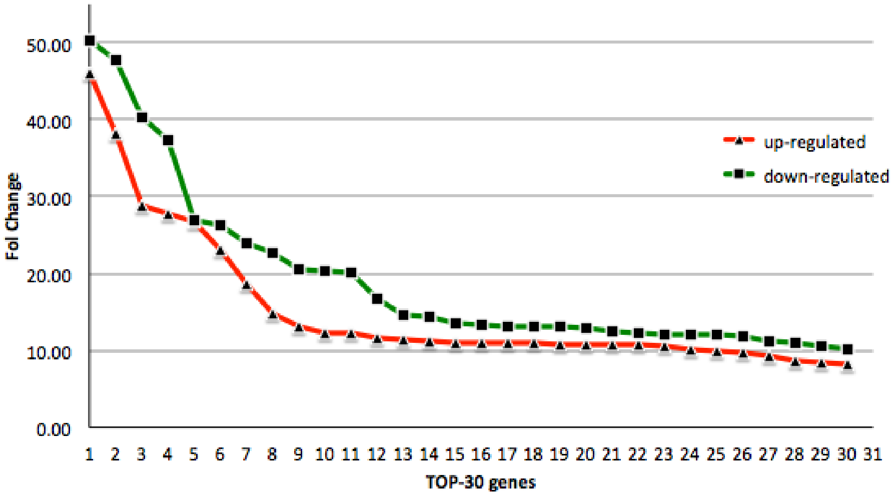
Keyword	Genbank-NR	TAIR9	Uniprot-KB
Hypothetical	14,998 (13,571)	15 (15)	3 (3)
Possible	6 (6)	0 (0)	1 (1)
Predicted	6,247 (5,573)	10 (10)	743 (729)
Putative	2,319 (1,989)	2,049 (1,518)	3,379 (3,229)
Uncharacterized	92 (92)	2 (2)	5,829 (5,651)
Unknown	557 (554)	3,843 (3,809)	4 (4)

Appendix 3, Tab. 2: Number of reference contigs having an uninformative hit (here referred to as ‘keyword’). The number of contigs with no association to a biological pathway is displayed between parentheses and reach up to 90% for NR and TAIR9, and 96% for UniProt-KB.

APPENDIX 4 – GO classification and FC distribution



Appendix 4, Fig. 1: Distribution of number of genes according to gene ontologies: cellular component, molecular function and biological process for both developmental stages. Grey boxes are early stage and blue boxes are late stage.



Appendix 4, Fig. 2: Distribution of fold-change for top-30 up and down-regulated genes. For both rankings, a plateau is reached rapidly showing only the ca. 13th first genes to have an outstanding fold-change (> 10X).