ROLE OF *ARABIDOPSIS* NUDIX HYDROLASE 7 (ATNUDT7) IN SEED AFTER-RIPENING & EXPLORING THE SWITCHGRASS TRANSCRIPTOME USING SECOND-GENERATION SEQUENCING TECHNOLOGY

By

XIN ZENG

Bachelor of Science in Biotechnology
Sichuan University
Chengdu, Sichuan, China
July, 2007

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2012

ROLE OF *ARABIDOPSIS* NUDIX HYDROLASE 7

(ATNUDT7) IN SEED AFTER-RIPENING

& EXPLORING THE SWITCHGRASS

TRANSCRIPTOME USING SECOND-GENERATION

SEQUENCING TECHNOLOGY

Dissertation Approved:

Dr. Ramamurthy Mahalingam (Advisor & Chair)

Dr. Andrew Mort (Committee Member)

Dr. Peter Hoyt (Committee Member)

Dr. Yanqi Wu (Outside Committee Member)

Dr. Sheryl Tucker **(**Dean of the Graduate College

TABLE OF CONTENTS

CHAPTER Page

I. INTRODUCTION 9

II. Role of *Arabidopsis* Nudix Hydrolase 7 (Atnudt7) in Seed After-Ripening 15

Materials and Methods.....15

Results.....23

Discussion.....45

III. REFERENCES 51

IV. Exploring the Switchgrass Transcriptome Using Second-Generation Sequencing Technology…….....60

V. APPENDIX……60

LIST OF FIGURES

Figure Page

FIGURE 1 24

FIGURE 2 26

FIGURE 3 28

FIGURE 4 30

FIGURE 5 31

FIGURE 6 33

FIGURE 7 35

FIGURE 8 38

FIGURE 9 40

FIGURE 10 41

FIGURE 11 43

FIGURE 12 47

LIST OF ABBREVIATIONS

ABA: Abscisic Acid

ADP: Adenosine Diphosphate

ADPR: Adenosine Diphosphate Ribose

CRC: Catabolic Redox Charge

DAI: Days After Imbibitions

EMSA: Electrophoretic Mobility Shift Assay

ERFs: Ethylene Response Element Binding Factors

EST: Expressed Sequence Tags

GAs: Gibberellins

GO: Gene Ontology

IAA: Indole-3-Acetic Acid

JA: Jasmonic Acid

JA-Ile: Jasmonate- Isoleucine

MBP: Maltose Binding Protein

NAD (P): Nicotinamide Adenine Dinucleotide (Phosphate)

NADH: Nicotinamide Adenine Dinucleotide, Reduced Form

NGS: Next Generation Sequencing

NMN: Nicotinamide Mononucleotide

NMNH: Nicotinamide Mononucleotide, Reduced Form

OPDA: 12-Oxo-Phytodienoic Acid

PAR: Poly Adenosine Diphosphate –Ribose

PN: Pyridine Nucleotide

RFU: Relative Fluorescence Units

ROS: Reactive Oxygen Species

RPKM: Reads Per Kilobase Per Million Mapped Reads

SA: Salicylic Acid

SSRs: Simple Sequence Repeats

CHAPTER I

INTRODUCTION

ROLE OF *ARABIDOPSIS* NUDIX HYDROLASE 7 (ATNUDT7) IN SEED AFTER-RIPENING

Seed dormancy is usually defined as failure of a viable seed to complete germination in conditions favorable for germination to proceed (Bewley, 1997). Seed dormancy is an important developmental checkpoint, allowing plants to regulate when and where they grow. In nature, germination of dormant seeds is triggered by environmental signals such as changes in temperature, light, soil hydration, that allow plants to restrict the timing of their establishment to certain seasons. One interesting feature of seed dormancy is that plants have evolved different mechanisms for inducing dormancy (Bentsink *et al.*, 2006; Finch-Savage & Leubner-Metzger, 2006).

In the laboratory conditions numerous chemical and physical treatments have been described that can reduce seed dormancy. For example, a period of dry after-ripening (several weeks to months), moist-chilling or cold stratification (for few days in dark) or application of gibberellic acid (GA), nitrate and nitric oxide reduce seed dormancy level (Debeaujon & Koornneef, 2000; Debeaujon *et al.*, 2000). Environmental factors such as high or low temperatures and water levels, and the hormone abscisic acid (ABA) are known to promote seed dormancy (Penfield & King, 2009).

One of the early biochemical events following seed imbibition are the changes in pyridine nucleotide (PNs) levels (Gallais et al., 1998). In the embryos of non-dormant caryopses of *Avena sativa* significant increases in NADH levels were observed within six hours of imbibition (absorption of water) while in the dormant caryopses their levels remained unchanged (Gallais et al., 1998). In the highly dormant Cape Verdes Island (Cvi) ecotype of Arabidopsis, levels of NAD were high and NADP levels were low when compared with Ws (Wassilewskija, Russia) ecotype that has intermediate dormancy or Col-0 (Columbia, USA) ecotype with least or no dormancy (Hunt & Gray, 2009). Measurement of PNs in fresh and after-ripened seeds of Cvi showed that breaking of dormancy was associated with a reduction in NAD levels, but not with an increase in NADP levels (Hunt & Gray, 2009). In *Arabidopsis* seeds that have reduced nicotinamidase activity (*nic2-1*), levels of NAD were high, and this was associated with increase in seed dormancy (Hunt et al., 2007). These studies clearly show that PNs homeostasis plays an important role in seed dormancy.

The majority of dormancy or germination related mutants identified to date are associated with hormone biosynthesis or hormone signaling (Koornneef et al., 2002). These include the well-characterized ABA mutants (Kucera et al., 2005), GA mutants (Steber *et al.*, 1998; Russell *et al.*, 2000) and ethylene mutants (Beaudoin et al., 2000; Ghassemian et al., 2000). Another class of dormancy or germination associated mutants is structural mutants wherein the seed coat or testa were shown to have an important role in regulating germination (Debeaujon et al., 2000). Recently, mutants associated with redox metabolism have been shown to play a key role in seed dormancy and includes mutants of the NADPH oxidase (*AtrbohB*) (Muller et al., 2009), nicotinamidase (Hunt et al., 2007).

Nudix (Nucleoside diphosphates linked to moiety X) hydrolases are pyrophosphohydrolases that work on a wide range of metabolites including PNs, ADP-ribose, coenzyme A, dinucleoside polyphosphates, ribo and deoxyribo-nucleoside triphosphates (Dunn et al., 1999). In a proteomics based analysis of seed dormancy, a nudix hydrolase, AtNUDT3 was shown to vary between dormant and nondormant seeds (Chibani et al., 2006). We and several other researchers have reported that *Arabidopsis* Nudix hydrolase 7 (AtNUDT7) is a NADH pyrophosphatase and also an ADP-ribose pyrophosphatase (Ogawa *et al.*, 2005; Olejnik & Kraszewska, 2005; Jambunathan & Mahalingam, 2006; Ge & Xia, 2008; Ishikawa *et al.*, 2009). Loss-of-function of *AtNudt7* leads to stunted plant growth, increased levels of ROS and NADH, and constitutive activation of stress response genes (Jambunathan & Mahalingam, 2006; Jambunathan *et al.*, 2010). In addition, previous study from our group has shown that *AtNudt7* transcript is very rapidly induced in response to biotic and abiotic stresses. Using polyclonal antibodies raised against recombinant AtNUDT7, rapid accumulation of the protein was seen in response to ozone, bacterial pathogens and hydrogen peroxide (Jambunathan *et al.*, 2010) Rapid induction of AtNUDT7 transcript and protein levels during stress indicates this protein is important for oxidative signaling (Jambunathan & Mahalingam, 2006; Jambunathan *et al.*, 2010).

In this study we report a reduced germination potential phenotype in the after-ripened *Atnudt7* mutant seeds. We examine the AtNUDT7 protein levels during germination and in ecotypes and mutants with varying levels of dormancy. We analyzed the changes in phytohormone levels, NADH and ADP-ribose pyrophosphohydrolase activity as well as ROS in WT and *Atnudt7*. Based on these observations we propose a model showing the inter connections between AtNUDT7, PNs, ROS, and phytohormones in regulating seed dormancy in *Arabidopsis*. We have identified a 16-bp unique sequence with GCC-box in the promoter region

of AtNUDT7 that is known to be a binding site for the ethylene responsive transcription factors. Using a combination of data mining and traditional enzyme mobility shift assays we demonstrate that ERF1(Ethylene-Responsive Element Binding Factors 1) is the transcription factor that binds to this site and may be one of the important TFs for regulating AtNUDT7 levels during stress and development.

EXPLORING THE SWITCHGRASS TRANSCRIPTOME USING SECOND-GENERATION SEQUENCING TECHNOLOGY

Even though genome sequencing technologies have become progressively efficient over the last few years, complete sequencing of complex genomes is still expensive. Identification of transcribed portions of the genome using expressed sequence tags (ESTs) technology provides a viable alternative for analyzing non-model systems and organisms with large genome sizes. ESTs have large amount of functional information and have been proven to be valuable for gene annotation and gene discovery (Andersen & Lubberstedt, 2003; Emrich *et al.,* 2007; Kaur *et al.,* 2011). ESTs have been useful for development of molecular markers (Mahalingam *et al.*, 2003; Barbazuk *et al.*, 2007; Novaes *et al.*, 2008; Puckette *et al.*, 2009; Sun *et al.*, 2010), comparative genomics (Tobias, 2008; Vera *et al.*, 2008) and for genetic analysis of adaptive traits (Namroud *et al.*, 2008; Parchman *et al.*, 2010) . Genes are expressed in particular tissues or cell types, developmental stages and vary in their expression levels by several orders of magnitude. Traditional EST projects require substantial investments in terms of library construction and sequencing, especially if the goal is to capture rare transcripts (Wall *et al.*, 2009).

Next generation sequencing technologies such as pyrosequencing, bypass lengthy cloning steps involved in Sanger sequencing and provide rapid and economical technologies for transcriptomics (Margulies *et al.*, 2005; Chi, 2008; Mardis, 2008; Morozova & Marra, 2008; Schuster, 2008; Wang *et al.*, 2009; Wang *et al.*, 2010). To date, the massively parallel DNA sequencing developed by Roche life Sciences called 454 pyrosequencing is the most widely used next-generation technology for *de novo* sequencing and analysis of transcriptomes of non-model systems. The first commercial NGS, 454 GS20 produced 200,000 reads with an average read of 100 bases per run (Chi, 2008; Schuster, 2008). Rapid improvements in emulsion PCR and sequencing chemistry have greatly improved the throughput, read-length and accuracy of 454 sequencing technology (Metzker, 2010). The newest 454-sequencing platform, GS FLX Titanium, can generate a million reads with an average read of 700 bases at 99.5% accuracy per run.

Switchgrass is a C4 (C4 carbon fixation) perennial grass that was selected in 1991 by the department of energy as a model herbaceous energy crop for the development of renewable feed stock resource to produce transportation fuel (Bouton, 2007). This choice has been attributed to several features of this plant native to North America. 1. Biomass – Switchgrass plants can grow 3-8 feet tall depending on ecotype 2. Low input – Switchgrass can thrive in marginal lands with minimal input of nutrients and water 3. Carbon sink – the large and fibrous root system of Switchgrass serves as a major reservoir of captured carbon (Bouton, 2007; Schmer *et al.*, 2008; Keshwani & Cheng, 2009). To further accelerate the pace of switchgrass breeding several groups have embarked on developing genomic resources including SSR

markers (Tobias CM, 2006; Okada *et al.*, 2010; Wang *et al.*, 2011; Zalapa *et al.*, 2012), ESTs (Tobias, 2008; Palmer NA, 2011) and miRNAs (Matts *et al.*, 2010).

In this study we describe 454 based transcriptome analysis in four different switchgrass tissues – dormant seeds, germinating seedlings, tillers and flowers. We describe the de novo assembly of these ESTs, and assembly and annotation of EST sequences using the foxtail millet genome as the reference. Second, we discuss the transcriptome coverage using proxy methods in the absence of the switchgrass genome sequence. Thirdly, we assessed the overlap in expression profiles from these four tissue samples using the reads per kilobase of million reads analysis. Fourthly, we utilize these ESTs for predicting more than 2000 SSRs that can be very useful for mapping and population genomic studies in switchgrass.

CHAPTER II

ROLE OF *ARABIDOPSIS* NUDIX HYDROLASE 7 (ATNUDT7) IN SEED AFTER-RIPENING

Material and methods:

Plant materials:

Seeds of *Arabidopsis thaliana* ecotypes Col-0, WS, Cvi, mutants *Atnudt7-1* and *aba 2-1*, *Atnudt7-2* and *AtNUDT7t* complementation transgenic plants in the *Atnudt7-1* background were used in this study. After-ripening was accomplished by storing the seeds for 4-6 months at room temperature. Dry after-ripened *Arabidopsis* seeds were placed on top of wet double layer whatman paper in petri-dish plates, sealed with parafilm and placed under light for 10 hours at 22 ℃ in a plant growth chamber. Samples were collected before placing the seeds in petri plates (0 days) and 1-, 2-, and 3-days after imbibition (DAI), frozen in liquid nitrogen and stored at -80 ℃ for future experimental use. The *Atnudt7-2* seeds were generously provided by Dr. Jane Parker (Max-Planck Institute Of Plant Breeding Research, Cologne, Germany).

AtNudt7-1 Complementation:

To establish the *AtNUDT7t* complementation transgenic lines, primers with *Kpn* I and *Hin* dIII restriction enzyme sites were designed to amplify about 2355bp that included the full length AtNudt7 gene sequence along with the promoter region. The PCR product was ligated into the

pGEM-T easy vector. And restriction enzyme *Kpn* I and *Hin*d III were used to digest the target clone, and release the insert from pGEM-T easy vector. Then it was fused into PZP121 vector. Then the AtNudt7 gene was transformed into DH5 alpha cells and the purified plasmid was obtained. This plasmid was then transfected into *Agrobacterium* and *Atnudt7-1* plants were transformed using the floral dip method (Clough & Bent, 1998; Clough, 2005). The plants are then allowed to set seeds and the seeds were screened, on MS agar plates containing 25 ug/mL of the antibiotic kanamycin. The screening and growing was repeated until a T3 generation and fresh seeds were collected from the mature T3 plants.

Germination Assays:

Dry *Arabidopsis* seeds were plated as described above. The numbers of seeds that germinated or that were dormant were recorded at each day for seven days.

GUS Staining:

GUS (beta-glucuronidase) reporter gene fusion system (Jefferson, 1989) with the full length AtNudt7 promoter clone was constructed. Primers with *Bam* HI and *Hin*dIII restriction enzyme sites were designed to amplify full-length AtNudt7 promoter, the PCR product was ligated into the pGEM-T easy vector. The same enzymes were used to digest the clone to release the insert from pGEM-T vector and released insert was fused into a GUS reporter gene in the pBI101 binary vector. Then the GUS reporter containing fusion construct was transformed into DH5 alpha cells and the purified plasmid was obtained. This plasmid was then transfected into *Agrobacterium* and WT Col-0 plants were transformed using the floral dip method (Clough &

Bent, 1998; Clough, 2005). The plants were then allowed to set seeds. After screening, T3 seeds were used for experiments.

Transgenic dry seeds and the seeds on wet whatman paper in petri-dish after 24 hours at 22 ℃ under light were collected and used for the GUS expression analysis. Seed coat was separated from each seed and submerged in 100 mmol NaPO4 buffer (pH 7.0) containing 10 mmol EDTA, 0.01% Triton X-100, 0.5 mmol Potassium ferricyanide, 0.5 mmol Potassium ferrocyanide and 1 mg/ml 5-bromo-4-chloro-3-indolyl glucuronide (X-Gluc). Then the seeds were incubated overnight at 37 ℃ in dark . Seeds were destained overnight by replacing GUS staining solution with 70% EtOH at 37 ℃ in dark. Microscopic analysis was performed on a Nikon eclipse TE2000-e epifluorescence microscope (Nikon Belux, Brussels, Belgium) coupled with a standard Nikon CCD camera.

Reverse-Transcription PCR:

Total RNA was isolated from Col-0 and *Atnudt7-1* 0, 1, 2 and 3 days seeds sample using RNeasy kit (Qiagen, Valencia, CA) and was diluted to 200ng/µL. Two µg of this total RNA was used for cDNA synthesis, using Superscript reverse transcriptase (Invitrogen), carried out according to the manufacturer's instructions. As a positive control, gene specific primers of constitutively active Actin 2 gene (Actin 2 F: ACAACAGCAGAGCGGGAAATTGT; Actin 2 R: TCTTCATGCTGCTTGGTGCAAGT) were used; For amplifying *AtNudt2*, gene specific primers (AtNudt2 F: AGAGATTGAGGCAGCTCAGTGGAT; AtNudt2 R: GAGAGGCAGCAAAGAAACCAGCTT) were used; For amplifying *AtNudt6*, gene specific primers (AtNudt6 F: CGTAAACCAACCCTGGAACCAGAA;

AtNudt6 R: TTTCAACCAGAGGTGGAGGCTAGG) were used; For amplifying *AtNudt10*, gene specific primers (AtNudt10 F: AGGCGAGATTAAGCCATCGGTACT;

AtNudt10 R: ACTGAGAGTACCGGTTCGATGCTA) were used; For amplifying *AtNudt7*, gene specific primers (AtNudt7 F: AATGGGCGAGGATATATGGAC; AtNudt7 R: GCGAATCCCAAGTATTCTTCC) were used. PCR was conducted for 25 cycles using an annealing temperature of 60 ℃.

Western Analysis

Total protein was extracted from seed samples with protein extraction buffer. Protein concentration was determined using Bradford assay (Bio-Rad). Fifteen μg of each sample was used for the western blot and hybridizations were conducted using the Atnudt7 polyclonal antibodies as described earlier (Jambunathan *et al.*, 2010)

ADP-Ribose and NADH pyrophosphohydrolase activity assays

The seed samples (0.2 g) of Arabidopsis were homogenized with 0.5 mL of 100 mM Tris-HCl (pH 8.0) containing 20% glycerol (Ishikawa *et al.*, 2009). After centrifugation (20,000g) for 20 min at 4 ℃, the supernatant was used for analysis of enzymatic activity. ADP-Ribose and NADH pyrophosphohydrolase activities were assayed by coupling to alkaline phosphatase and measuring colorimetrically the amount of inorganic phosphate formed at 37 ℃ (Ames, 1966; Ribeiro *et al.*, 2001). The standard assay mixture contained, in a volume of 0.1 mL, 50 mM Tris-HCl (pH 8.0), 5 mM MgCl2, 0.1 mM substrates, 0.7 units of alkaline phosphatase, 1 mg/mL21

bovine serum albumin, and crude extract (approximately 10.0 mg of protein). The solution was incubated at 37 ℃ for 15 min (Jambunathan & Mahalingam, 2006) and the reaction was stopped and color was developed by addition of 1 mL of standard inorganic phosphate reagent (6 volumes of 3.4 mM ammonium molybdate in 0.5 M H2SO4, 1 volume of 570 mM AsA, and 1 volume of 130 mM SDS) (Ribeiro *et al.*, 2001). After 20 min incubation at 45 ℃, Absorbance at 820um was measured. Blanks without enzyme and/or substrate were run in parallel. Enzyme activities were linear with time and amount of enzyme.

Pyridine Nucleotide Extraction and Measurement

Glycyl Glycine Method was used to measure the NAD and NADH levels in seeds (Hayashi *et al.*, 2005) (Jambunathan *et al.*, 2010). Seed tissues collected from Col-0 wild-type and *Atnudt7-1* mutant (0,1,2, and 3 days samples) were used. At least six technical replicates and three biological replicates of each sample were analyzed.

ABA treatment

Seeds of Col-0 and *Atnudt7-1* mutant were placed on MS plates containing 1 uM ABA. Plates were placed vertically in the growth chamber under light for 10 hours at 22 ℃ and were photographed after 13 days.

Reactive Oxygen Species (ROS) analysis

Seeds of Col-0 and *Atnudt7-1* mutant were set up on petriplates as described above and samples were collected 0-,1-,2-, and 3-days after transfer to plates. After-ripened dry seeds of Col, Ws, Cvi, *aba2-1* and *aba3-1* were ground to a fine powder with liquid nitrogen. ROS measurement was done using the H2-DCFDA (Sigma) as described previously (Joo *et al.*, 2005). The average fluorescence value obtained from three measurements was divided by the protein content and expressed as relative fluorescence units per milligram of protein. The analysis was repeated with three biological replicates.

PolyADP-ribose immunoassay

Levels of PAR in Col-0 and *Atnudt7-1* seeds during the course of germination (0, 1 ,2 and 3 days after imbibition) were measured by dot blotting three different concentrations of total protein extract (1, 1.5 and 2ug) and detected using polyclonal antibody to PAR (BD BioSciences) The autoradiograms were scanned using the ImageMaster Pro software to measure intensity and volume of signals. The experiment was replicated twice.

Phytohormone analysis

Seeds of Col-0 and *Atnudt7-1* mutant were set up on petriplates as described above and samples were collected 0-,1-,2-, and 3-days after transfer to plates. About 100 mg seed tissue from each sample were collected and the hormone extraction, LC-MS/MS analysis were conducted at Donald Danforth Plant Science Center Proteomics and Mass Spectrometry Facility using previously described methods (Chen *et al.*, 2009).

Seed samples from each time points were analyzed for plant hormone concentration including acid hormones: Abscisic acid (ABA), IAA-Asp (Indole-3-acetyl-aspartic acid), Indole-3-acetic

acid (IAA), Jasmonic acid (JA), JA-Ile (jasmonoyl-isoleucine), OPDA (12-oxo-phytodienoic acid), and Salicyclic acid (SA). The data was normalized based on the internal standards: D4SA, D6ABA, D5IAA and H2JA. For the analysis of the GA the method described in (Zhang *et al.*, 2011) was followed.

Electrophoretic Mobility Shift Assay

The coding region of AtERF1 was cloned into the pMBP vector (provided by Dr. Junpeng Deng, Oklahoma State University) and expressed in *Escherichia coli* BL21 cells. Maltose binding protein (MBP)-AtERF1 fusion proteins were purified using nickel resin, as specified by the manufacturer. Both strands of the following oligonucleotides were synthesized, biotin labeled and annealed: AtNUDT7 promoter with the unique 16bp sequence: (5'GCCGTTAGGCGTTGCCGCCTGTAGTAAT 3'); control from AtNUDT7 promoter without the 11bp unique sequence (5'AGAGTTTGCTGCTGCCGTTAGGCGTAAT 3').

Binding buffer contained 12 mM Tris–HCl (pH 7.5), 50 mM KCl, 2.5 mM MgCl2, 1 mg of poly(dA-dT), 1 mM DL-dithiothreitol (DTT), 0.5 mM EDTA, 5% glycerol, 0.05% NP-40 and 200 pmol biotin-labeled probes (Fujimoto *et al.*, 2000; Shen *et al.*, 2012). The samples were loaded and run on a 6% DNA retardant gel after the reactions had been incubated at room temperature for 20 min. The DNA was transferred onto nylon membranes and signal detected with a LightShift-Chemiluminescent EMSA Kit (Thermo Fisher Scientific Inc., Rockford, IL, USA) using standard protocols.

Statistical Analysis

Microsoft Excel was used for conducting student's t-test for determining the statistical significance of results comparing WT and *Atnudt7-1* mutant for differences in germination rates, gene expression, pyridine nucleotide levels, ROS levels and phytohormones.

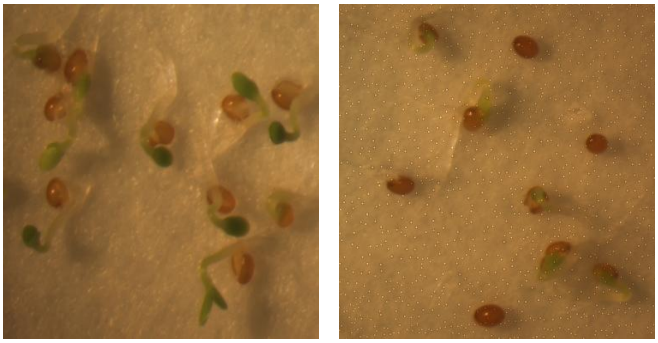
RESULTS

Loss of AtNUDT7 slows down germination pace and lowers germination potential of *Arabidopsis* seeds

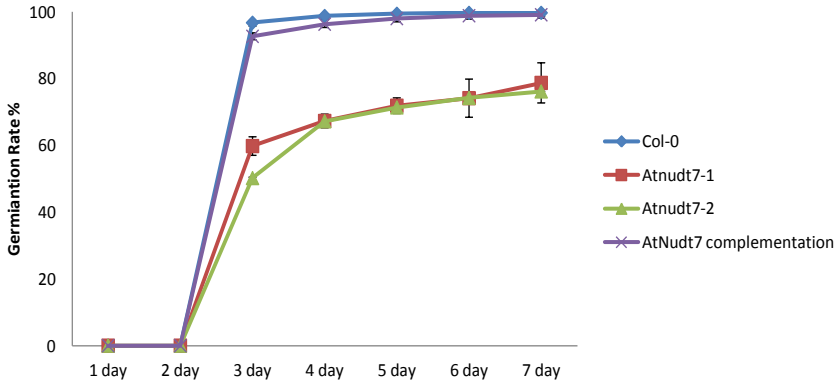
In our previous studies we have shown that 3-week old *Arabidopsis* nudix hydrolase 7 mutant plants (*Atnudt7-1*) are significantly smaller compared to age matched wild type plants (Jambunathan & Mahalingam, 2006; Jambunathan *et al.*, 2010). We examined if the differences in the size between the *Atnudt7-1* mutant and WT manifest at much earlier stages of development. We examined the developmental progression of after- ripened seeds of WT and *Atnudt7-1* mutants on petri-plates. As expected we saw that the *Atnudt7-1* seedlings were much smaller in size when compared with WT, as early as 3 days after imbibition (Fig. 1a). More interestingly, we observed that the germination potential of the *Atnudt7-1* seeds were significantly lower compared with WT plants （$p< 0.001$） (Fig. 1b). We also examined the germination potential of seeds of a second insertion line *Atnudt7-2* using the same method. It also showed the similar germination potential as *Atnudt7-1*, and was significantly lower compared with WT seeds ($p<0.001$) (Fig. 1b). Atnudt7t complementation lines in *Atnudt7-1* background carrying the WT version of the *AtNudt7* gene including the promoter region showed germination potentials, which was close to 100%, similar to WT seeds (Fig.1b) confirming that a functional AtNUDT7 is important for ensuring high germinability of the seeds.

Furthermore, cold-stratification treatment in 0.15% phytagel at 4 ℃ for 3 days could not restore the germination rate of *Atnudt7-1* seeds (Fig.1c). These results suggest that *AtNUDT7* gene may play an important role during seed after-ripening in *Arabidopsis*.

(a)



(b)



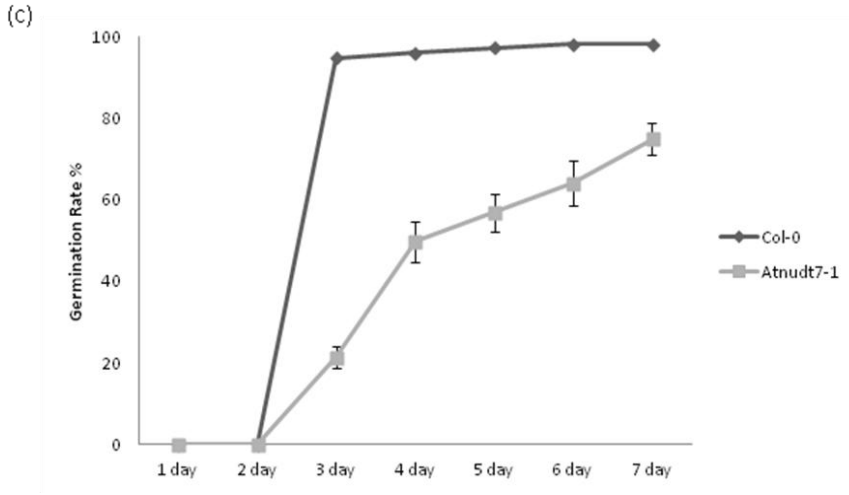


Fig. 1 Seed germination phenotype associated with *Atnudt7* mutant. (a) Col-0 and *Atnudt7-1* seeds on wet whatman paper in petri-dish during 7 days at 22 ℃. (b) Germination potential of after-ripened seeds of *Arabidopsis* Col-0, *Atnudt7-1, Atnudt7-2* knock-out mutant and AtNudt7t complemented line. (c) Germination potential of cold-stratified seeds of *Arabidopsis* Col-0 and *Atnudt7-1*. There seeds samples were kept in 0.15% phytagel at 4 ℃ for 3 days before the germination assays. Percentage germination was assessed by radicle emergence each day during the 7 days at 22 ℃. Ten independent experiments with approximately 100 seeds for each experiment were used. Bars represent standard error ($p<0.001$).

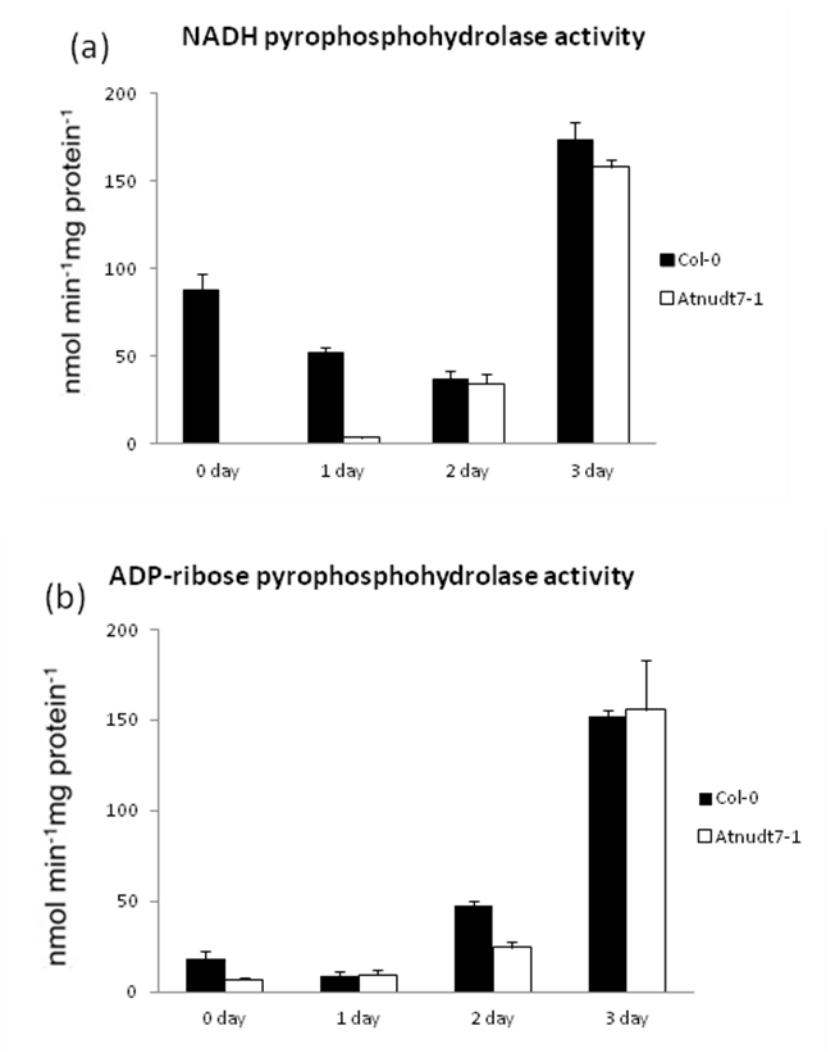


Fig. 2 ADP-ribose and NADH pyrophosphohydrolase activity was measured during 0 (dry seeds), 1, 2, 3 days after imbibition in Col-0 and *Atnudt7-1* mutant. (a): NADH pyrophosphohydrolase activity; (b): ADP-Ribose pyrophosphohydrolase activity. The data shown are means of three replicates for each sample. Bars represent standard error. ($P<0.05$)

AtNUDT7 pyrophosphohydrolase activity is important for seed germination

The enzyme assays were done to measure the ADP-ribose and NADH pyrophosphohydrolase activities during the first 3 days after imbibition in *Col-0 and Atnudt7-1* seeds. The NADH

pyrophosphohydrolase activity in Col-0 was significantly higher than in *Atnudt7-1* mutant in dry seeds and one day following imbibition ($p<0.001$) (Fig.2 a). The enzyme activities correlated well with the protein expression profile of AtNUDT7 in Col-0 sample during 0 and 1 day after imbibition. All these results once again confirmed the presence of NADH pyrophosphohydrolase activity of AtNUDT7 was very important for dry seeds and during 24 hrs of imbibition. On the other hand, the ADP-ribose pyrophosphohydrolase activity (Fig. 2 b) was not that dramatically different between Col-0 and *Atnudt7-1* mutant in 0 and 1 day sample. This might indicate again that NADH is the preferred substrate for AtNUDT7 rather than ADP-ribose during the early stages following imbibition.

In addition, the relatively high activity of both ADP-ribose and NADH pyrophosphohydrolase in both Col-0 and *Atnudt7-1* mutant in 2 and 3 days sample (Fig.2 a & b) indicates there might be some other AtNUDT enzymes playing a role here. Previous study has shown that AtNudt6 and AtNudt10 also utilize NADH and ADP-ribose as primary substrates (Dobrzanska *et al.*, 2002; Kraszewska, 2008). This prompted us to examine the gene expression level of *AtNudt 6, 7* and *10* during the 0, 1, 2 , 3 days in Col-0 as well as *Atnudt7-1* mutant. RT-PCR was done to examine the differences in steady-state levels of these transcripts. It turned out that the expression levels of *AtNudt6* and *AtNudt10* were higher in 2 and 3 day samples compared to in the dry seeds and 1 day following imbibition in both Col-0 and *Atnudt7-1* mutant (Fig. 3). On the other hand, the *AtNudt7* expression level in Col-0 during the 0, 1, 2, 3 days were nearly opposite to *AtNudt6* and *AtNudt10*, the AtNudt7 expression level in 0 and 1 day was higher than observed in the 2 and 3 day samples (Fig. 3), which is very similar to the protein expression pattern shown in Fig 4 a. Higher expression of *AtNudt6* and *AtNudt10* could confer the higher ADP-Ribose and NADH pyrophosphohydrolase activity observed in both the Col-0 and *Atnudt7-1* mutant.

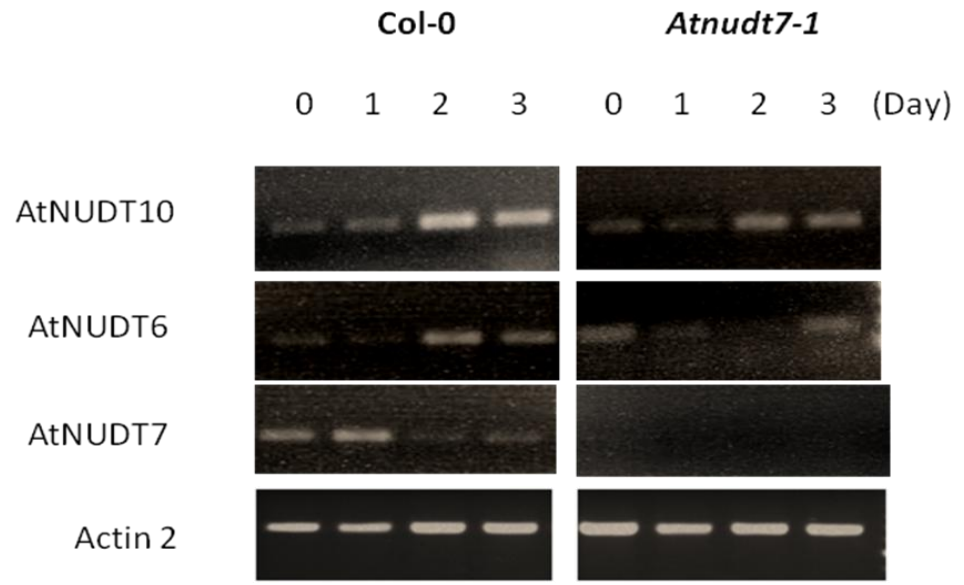


Fig. 3 RT-PCR analysis of *Arabidopsis* NUDT6 and 10. Primers were designed for: Actin 2 control; AtNUDT6; AtNUDT10 and AtNUDT7. Each lane represents amplifications from cDNA reverse-transcribed from Col-0 and *Atnudt7-1* RNA samples: 0= dry seeds, 1= 1 day, 2= 2 day, 3= 3 day after imbibition.

AtNUDT7 protein levels are down regulated during germination

Changes in the AtNUDT7 protein levels during germination were analyzed using AtNUDT7 polyclonal antibodies. AtNUDT7 protein levels were reduced to nearly 50% within 24 h after imbibition when compared with their levels in dry seeds, (Fig. 4a). Protein levels at the 2-day time point was extremely low and at the 3-day time point AtNUDT7 levels were not detectable.

A finer time-course experiment was conducted wherein samples were taken at 6-12 h intervals. Western analysis indicated that reduction in AtNUDT7 levels was apparent as early as six hours after imbibition and continued to decline steadily up to 36 hours and was completely phased out by 48 h (Fig. 4a, bottom panel).

We examined the AtNudt7 gene promoter activity during the early days of imbibition by testing GUS activity in the transgenic seeds with AtNudt7 full length promoter GUS reporter transgenic lines. Based on the AtNUDT7 protein profile during the first 3 days after imbibitions, we examined the 0 and 1 day sample of the GUS reporter harboring transgenic seeds. As indicated in Fig.5, the significant darker staining color in the GUS stained samples compared to the water controls reflected the AtNudt7 promoter is active in the dry seeds and one day after imbibition. And this result confirmed the protein pattern we have observed in Col-0 seeds during the 0 and 1 day imbibition. In addition, it also proved the AtNUDT7 expression is transcriptionally regulated during the early stages of seed imbibition.

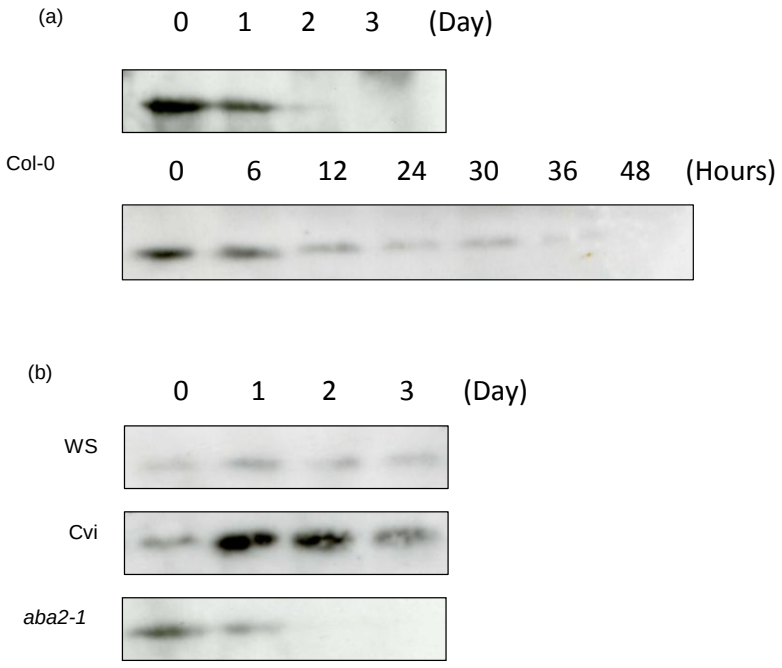


Fig. 4 Protein expression profile of AtNUDT7 during seed germination in *Arabidopsis*. (a) Western blot analysis of Col-0 wild-type seeds. Fifteen micrograms of total protein extracts from Col-0 loaded in each lane; In the top panel, proteins were extracted from dry seeds (0), 1, 2 and 3 days after imbibition. For the bottom panel, proteins were extracted from dry seeds (0), 6, 12, 24, 36 and 48 hours after imbibition. (b) Fifteen micrograms of total protein extracts from WS, Cvi and *aba2-1* mutant samples were loaded. Proteins were extracted from dry seeds (0), 1, 2 and 3 days after imbibition.

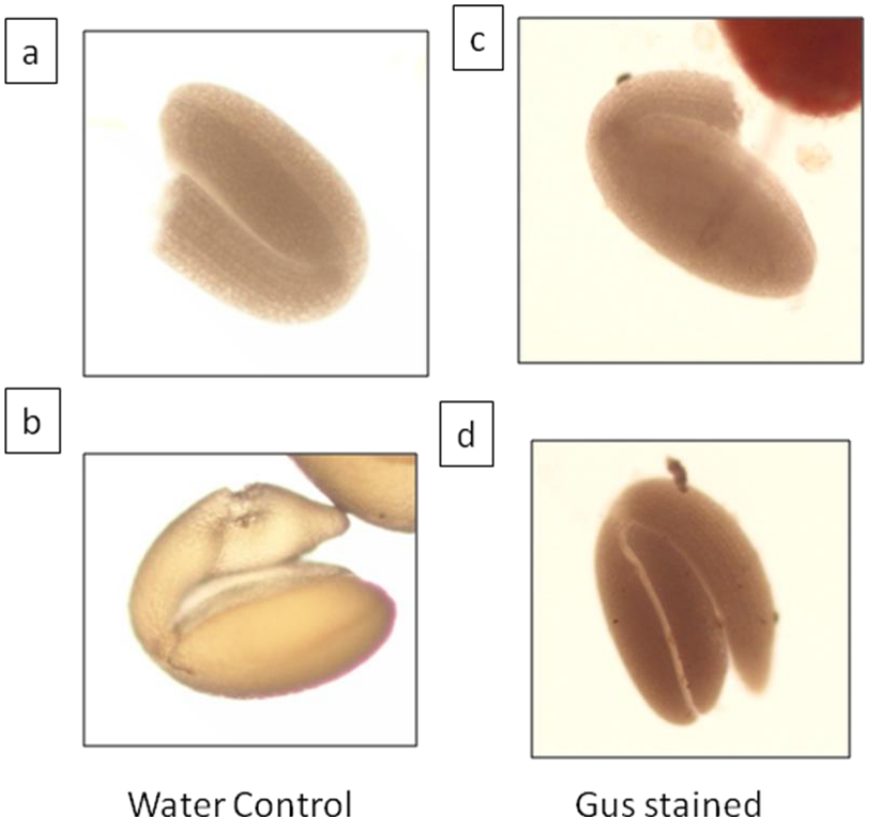


Fig. 5 GUS expression analysis pattern of AtNudt7 promoter in Col-0 seeds. **a**: Dry seeds; **b**: Seeds in water after 24 hours imbibition as control; **c**: Dry seeds after GUS staining; **d**: GUS staining in seeds after 24 hours imbibition.

Significant differences in AtNUDT7 protein levels in Arabidopsis lines with different levels of seed dormancy

Arabidopsis ecotype Cvi exhibit high levels of dormancy, while Ws ecotype has intermediate levels of seed dormancy when compared with Col-0 ecotype that has low levels of dormancy (Hunt & Gray, 2009). We examined the changes in AtNUDT7 protein over a three-day time period in these two ecotypes with varying levels of dormancy. AtNUDT7 levels in dry after-

ripened seeds of Cvi and Ws were much lower compared with Col-0 (Fig. 4b). Furthermore, the protein profiles in these two ecotypes were completely different when compared with Col-0 over the three-day time period following imbibition. In the Cvi ecotype, levels of AtNUDT7 increased more than 8-fold within 1-day after imbibition. At the end of day-2, levels of AtNUDT7 in Cvi showed a decline compared to the 1-day time point, but was still 4-5 fold higher compared to the fresh seeds. At the end of the third day, levels of AtNUDT7 were comparable to the dry seed sample. On the contrary, in Ws ecotype, levels of AtNUDT7 increased by 2-fold at the end of day-1 and were slightly reduced on days-2 and -3.

We also examined expression of AtNUDT7 in *aba 2-1*, an ABA deficient mutant that has reduced seed dormancy (Leon-Kloosterziel *et al.*, 1996). Similar to the WT, AtNUDT7 protein levels in the *aba2-1* mutant were not detectable within 48 h after imbibition (Fig. 4b).

These observations showed significant intrinsic differences in AtNUDT7 protein levels in *Arabidopsis* ecotypes and mutants under experimental conditions. These innate differences in AtNUDT7 may play an important role during the early hours following imbibition that may aid in overcoming seed dormancy. Lack of AtNUDT7 as in the *Atnudt7-1* mutant, or presence of this protein during later stages of imbibition was associated with lower seed germination potential.

Atnudt7-1 mutant is hyper-sensitive to ABA

The plant hormone abscisic acid (ABA) is a positive regulator of dormancy, while gibberellins (GAs) release dormancy and promote the completion of germination, counteracting the effects of

ABA (Sarath *et al.*, 2007; Holdsworth *et al.*, 2008). We observed that 1uM of exogenous ABA, in *Atnudt7-1* strongly inhibited germination and normal seedling development (Fig. 6a), while the WT plants appeared normal under these treatment conditions. This result suggested that *AtNudt7-1* mutants may have higher amount of endogenous ABA and this may be responsible for its lower germination potential.

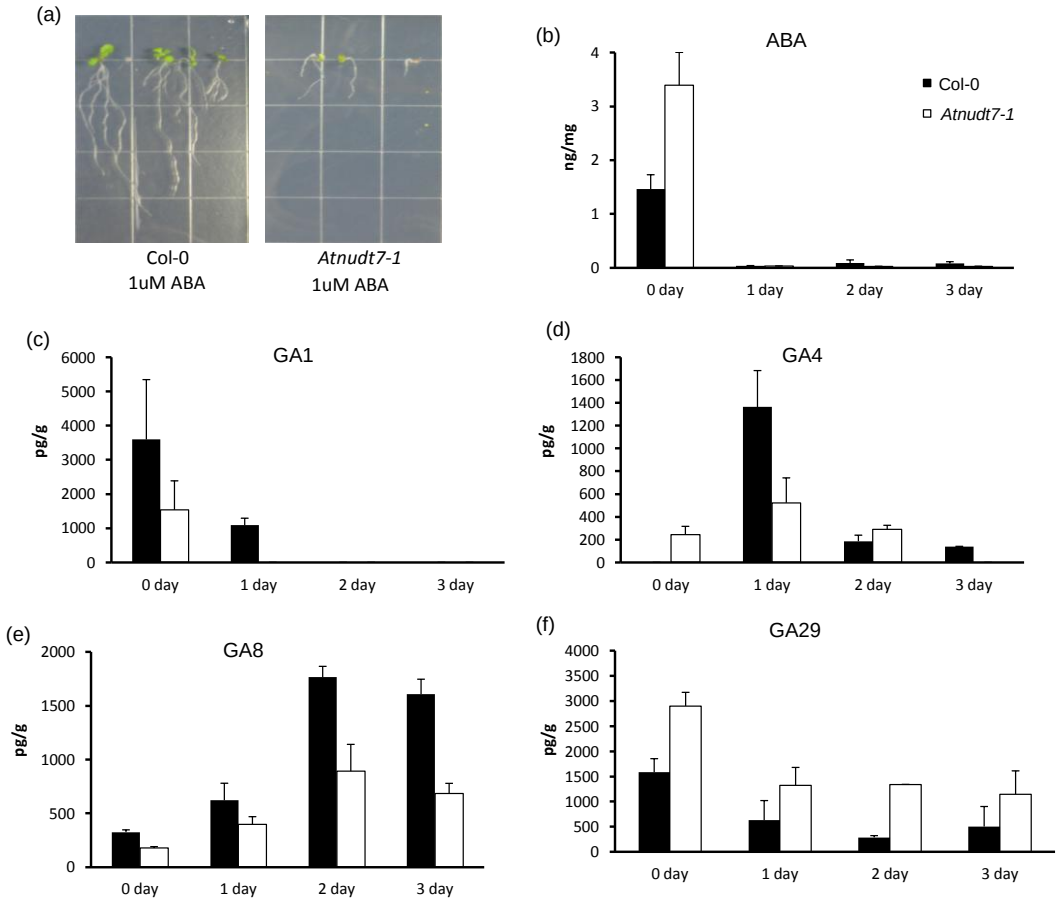


Fig. 6 (a) Growth of *Atnudt7-1* mutant and wildtype in the presence of ABA. Representative plate of Col-0 and *Atnudt7-1* mutant plants after 13 days with exogenous 1uM ABA added in the MS medium. ABA and GA levels as measured by LC-MS/MS during 0 (dry seeds), 1, 2, and 3 days after imbibition in Col-0 and *Atnudt7-1* mutant. (b) ABA (c) GA1 (d) GA4 (e) GA8 and (f)

GA29. The data shown are means of three replicates for each sample. Bars represent standard error.

ABA and GA levels are significantly altered in *Atnudt7-1* mutant

The ABA hypersensitive phenotype of *Atnudt7-1* prompted us to examine the phytohormone levels in the dry seeds and during the course of germination. We observed more than 2-fold higher levels of ABA in dry seeds of *Atnudt7-1* mutant than Col-0 seeds ($p < 0.02$). ABA levels were hardly detectable in both WT and *Atnudt7-1* mutant at 1, 2 and 3-days after imbibition (Fig. 6b).

GA1 was found in higher levels in dry seeds of WT compared with *Atnudt7-1* (Fig. 6c). Levels of GA1 decreased by about 4-fold within 24 h after imbibition in WT ($p<0.01$) and in the *Atnudt7-1* it was not detectable. GA4 is one of the major active GAs that stimulate seed germination (Ogawa *et al.*, 2003). GA4 levels were undetectable in the dry seeds of WT while the mutant contained low levels of this bioactive GA (Fig. 6d). Within 24 h after imbibition the levels of GA4 surged by a nearly 1000-fold in WT, while in the mutant the increase was a modest 2.5 fold ($p< 0.05$). GA8 that usually is formed by the oxidation of the GA1 leading to the inactivation of the latter was indeed found to increase 2 and 3 DAI, when the levels of GA1 were not detectable in both WT and *Atnudt7-1* mutant (Fig. 6e). Furthermore, the levels of GA29, the inactive form of GA was about 2-fold higher in the *Atnudt7-1* seeds compared with the WT ($p<0.02$) (Fig. 6e). Though the levels of this inactive form of GA dropped by nearly 60% following imbibition in the *Atnudt7-1*, these levels were still nearly 2-3-fold higher compared with the WT ($p<0.03$) at the corresponding time points (Fig. 6f). These results suggest that the

higher levels of ABA and inactive forms of GA in the *Atnudt7-1* may be lowering its germination potential.

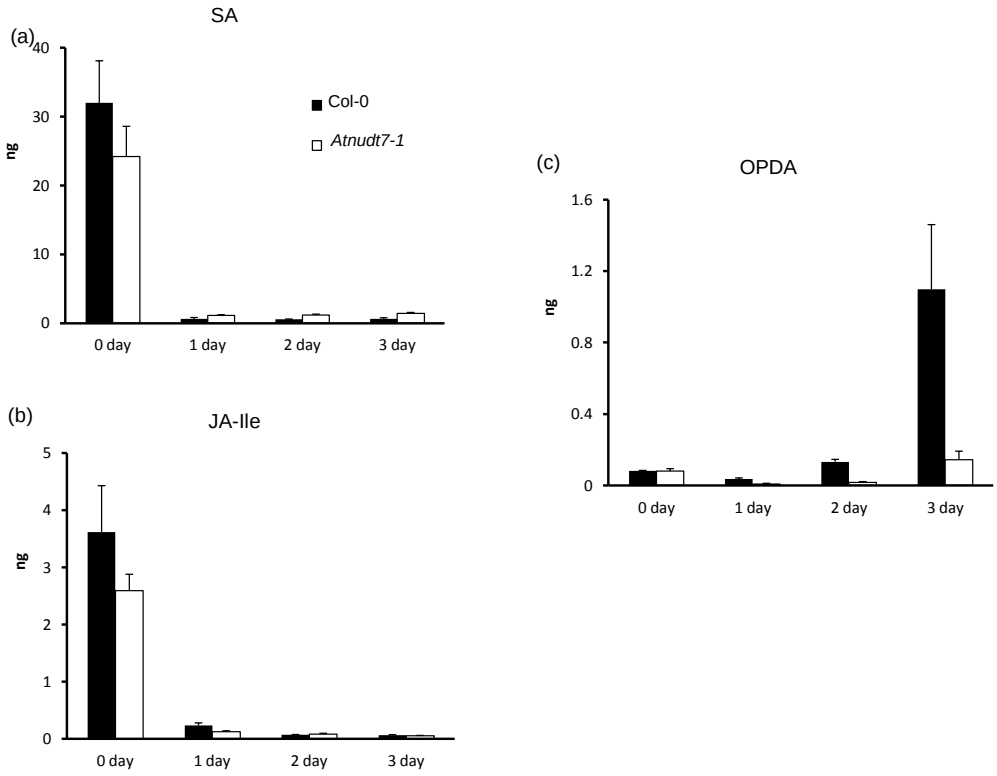


Fig. 7 Other endogenous plant horomones as measured by LC-MS/MS during 0 (dry seeds), 1, 2, and 3 days after imbibition in Col-0 and *Atnudt7-1* mutant. (a) Salicylic acid (b) Jasmonic acid-conjugated with amino acid Isoleucine (c) 12-Oxo-phytodienoic acid. The data shown are means of three replicates for each sample. Bars represent standard error.

Loss of *AtNudt7* also alters the levels of several other plant hormones

In the *Atnudt7-1* rosette leaves the levels of SA was reported to be 4-5 fold higher when compared with WT plants (Bartsch *et al.*, 2006). Interestingly, we found that levels of SA in the dry seeds of *Atnudt7-1* were lower than the WT plants (Fig. 7a). SA levels significantly reduced following imbibition. JA-Ile, which is the functional JA derivative needed for JA signaling (Katsir *et al.*, 2008) follows the pattern of SA profile (Fig. 7b). Interestingly, oxophytodieonic acid (OPDA) - the important upstream hormone needed for JA biosynthesis (Katsir *et al.*, 2008) showed totally different pattern from any other hormones (Fig. 7c). It started with relative low level in dry seeds in both mutant and WT. But in Col-0 3-day sample, the OPDA level showed a 10-fold increase when compared to the day-0 sample ($p< 0.04$), while in the mutant the changes in OPDA levels were not significant (Fig. 7c). These analyses demonstrate the changes in various phytohormones during the imbibition and germination phases. More importantly these results demonstrate that loss of *AtNUDT7* gene function perturbs the phytohormone balance in dry seeds and during germination, and the differences in phytohormone levels could contribute to the observed defects in seed after-ripening and germination potential.

Rapid changes in NAD and NADH levels during seed germination

AtNUDT7 has been shown to hydrolyze NADH by several independent groups (Ogawa *et al.*, 2005; Olejnik & Kraszewska, 2005; Jambunathan & Mahalingam, 2006; Ge & Xia, 2008; Ishikawa *et al.*, 2009). A relationship between pyridine nucleotides and seed dormancy was reported recently (Hunt & Gray, 2009). The NADH levels in WT seeds (0 day samples) were nearly 60% higher than *Atnudt7-1* mutant seeds ($p<0.002$) (Fig. 8a). Within 24 h after imbibition levels of NADH were lowered by more than 60% in both WT and *Atnudt7-1* ($p<0.001$) (Fig 8a)

On the contrary, NAD^+ levels remained stable in WT while their levels in *Atnudt7-1* seeds were lowered by nearly 50%, 24 h after imbibition. Interestingly, on day-2 and -3 the levels of NAD^+ increased by nearly 30% in both WT and *Atnudt7-1* mutant compared with the 1-day seed sample (Col-0: $p<0.001$, *Atnudt7-1:* $p< 0.00001$) (Fig. 8b). These rapid changes in NAD^+ and NADH levels leads to significant changes in the seed redox levels during the course of seed germination (Fig. 8c). Catabolic redox charge (CRC) defined as $NADH/NAD^+$+NADH, showed significant differences in the dry seeds of WT and *Atnudt7-1* with the values for the latter being lower ($p<0.05$). CRC values were at their lowest levels in WT within 48 h after imbibition. We speculate these intrinsic differences in CRC contribute to the reduced seed germination potential in *Atnudt7-1*. Changes in redox brought about during the imbibition phase will impact the biochemical milieu for the seeds to transition from quiescence phase to germination phase, and NADH pyrophosphohydrolase activity of AtNUDT7 may play a crucial role in this phase transition.

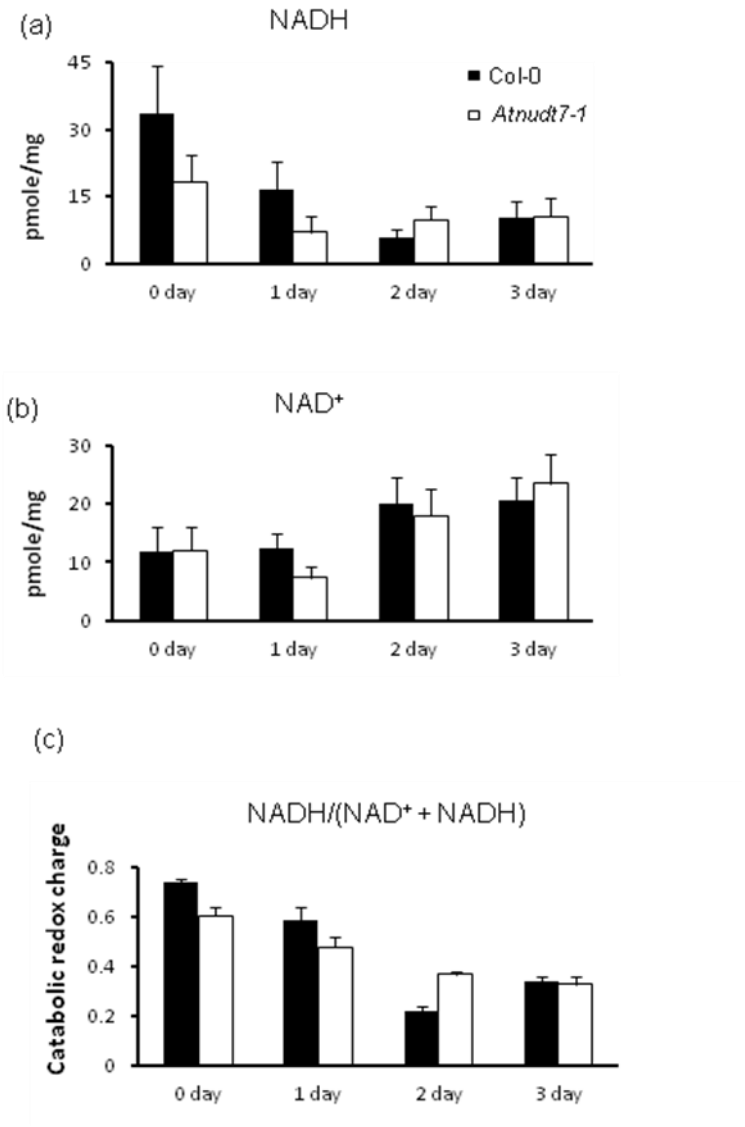


Fig. 8 Pyridine nucleotide analysis during 0 (dry seeds), 1, 2, 3 days after imbibition in Col-0 and *Atnudt7-1* mutant. (a) NADH; (b) NAD^+; (c) Catabolic redox charge which is defined as ratio of NADH to NAD^++NADH. Values shown are the means of eight independent experiments. Bars represent standard errors.

ROS profiles of *Atnudt7-1* and WT during imbibition and germination are very different

Several studies have highlighted the importance of ROS in breaking dormancy and during seed germination (Sarath *et al.*, 2007; El-Maarouf-Boteau & Bailly, 2008). In the *Atnudt7-1* mutant leaves, 2-3 fold higher levels of ROS have been reported (Jambunathan & Mahalingam, 2006). In contrast to the ROS levels in the leaves, we found that in the dry seeds levels of ROS in the *Atnudt7-1* were 6-8 fold lower compared with the WT seeds ($p<0.02$) (Fig. 9a). Interestingly, within 24 hours following imbibition levels of ROS increased more than 20-fold in the *Atnudt7-1* mutant compared to WT ($p<0.02$) (Fig. 9a). On day-2, ROS levels in the mutant were lowered by about 50% compared to their levels on day-1. These levels were still about 4-fold higher than in the dry seeds ($p<0.05$). ROS levels in WT showed a second peak of slightly higher magnitude at the 3-day time point.

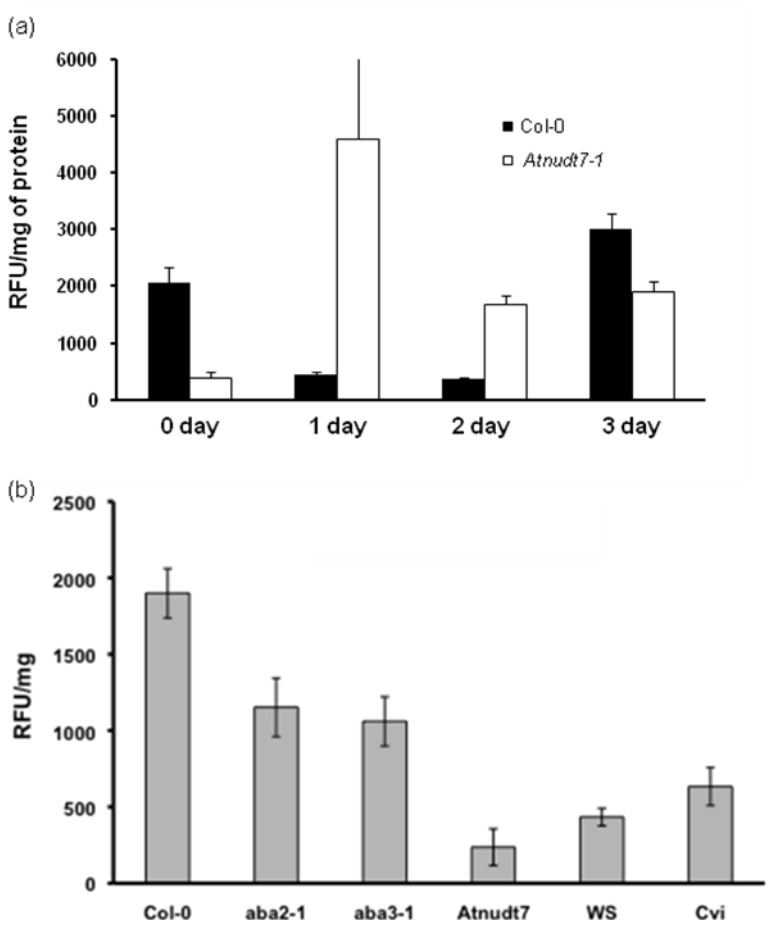


Fig. 9 Analysis of total reactive oxygen species during *Arabidopsis* seed germination. (a) ROS measurement in Col-0 and *Atnudt7-1* mutant 0 (dry seeds), 1, 2, 3 days after imbibition. Data represents average of three independent experiments with standard error. On the Y-axis, RFU/mg represents relative fluorescence units per mg of protein extract. (b) ROS analysis in after-ripened seeds of *Arabidopsis* ecotypes Ws and Cvi and mutants deficient in ABA biosynthesis - *aba2-1* and *aba3-1*. Average of three independent experiments are shown with standard errors.

The significant differences in the seed ROS levels of WT and *Atnudt7-1* prompted us to examine the levels of ROS in seeds of other ecotypes and mutants with differing levels of seed dormancy. ROS levels in the highly dormant Cvi seeds and moderately dormant Ws ecotypes

were 4-5 fold lower than the WT and were comparable to the levels seen in the *Atnudt7-1* mutant (*p<0.02*) (Fig. 9b). In seeds of *aba2-1* and *aba3-1* with low levels of dormancy, ROS levels were 2-3 fold higher than that in *Atnudt7-1* (*p<0.01*), but still lower than the WT. These results suggest that seed ROS levels may be a useful biochemical marker for predicting seed germination potential.

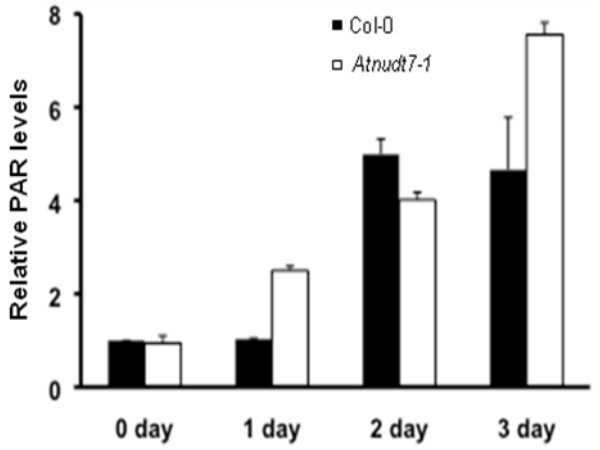


Fig. 10 Analysis of poly (ADP-ribose) (PAR) levels during *Arabidopsis* seed germination. Protein extracts (three concentrations) from Col-0 and *Atnudt7-1* mutant collected from 0 (dry seeds), 1, 2, and 3 days after imbibition were used for making dot-blots. PolyADP ribosylation levels were assayed with anti-PAR antibodies. Hybridization intensities were extracted using the Imagene software. Data from two independent experiments were averaged and bars represent standard error.

PAR levels are altered during seed germination

Previously a link between seed dormancy, NAD and PAR levels were demonstrated using the null mutant of nicotinamidase (*nic2-1*) expressed in seeds (Hunt *et al.*, 2007). Furthermore,

Atnudt7-1 mutant plants were shown to have lower levels of PAR (Ishikawa *et al.*, 2009). We observed that PAR levels in the *Atnudt7-1* seeds were comparable with the levels in WT (Fig. 10). PAR levels increased on days-2 and -3 by 4-fold in the WT ($p<0.02$). In *Atnudt7-1* mutant PAR levels steadily increased by 2 ($p<0.01$), 4 ($p<0.02$) and 8-fold ($p<0.005$) on day-1, -2 and -3, respectively, compared to their levels in dry seeds. Significant increases in the PAR levels during day-2 and -3 are consistent with the observed increase in NAD^+ levels in both WT and *Atnudt7-1* mutant, since it is well established that increased PARP activity leads to NAD^+ depletion (Mahalingam *et al.*, 2007; Ishikawa *et al.*, 2009). It is also possible that higher PAR levels in *Atnudt7-1* mutant may be triggered by higher ROS levels 1, 2 and 3 days after imbibition that could lead to oxidative damage to DNA.

ERF1 is the putative transcription factor that regulates AtNudt7 gene expression

Comparing the AtNudt7 promoter sequences between Ws and Col-0, we have identified a 16-bp unique sequence (GGCGTTGCCGCCTGTA) with GCC-box in the WS promoter sequence (Fig. 11a). And GCC box is known to be a binding site for the ethylene response element binding factors (Fujimoto *et al.*, 2000). Genevestigator analysis indicated that ethylene response element binding factor 1, AtERF1 expression level significantly decreased in the *Atnudt7-1* knock-out mutant (Sup Fig. 1). These information together prompted us to test the in vitro binding of AtERF1 protein to the unique 16 bp sequence in AtNUDT7 promoter region, thus to prove that AtERF1 is the transcription factor that binds and regulates the AtNudt7 gene expression during development and under stress.

To test if AtERF1 could bind to the 16 bp unique AtNudt7 promoter sequence with GCC-box, we overexpressed the entire coding region of AtERF1 as a maltose binding protein (MBP) fusion in *Escherichia coli* and performed electrophoretic mobility shift assays (EMSA). As shown in Fig. 11b, AtERF1-MBP fusion protein could bind 16 bp unique AtNudt7 promoter sequence with GCC-box and showed the gel shift on the top of the gel. While the oligos without the 16bp unique sequence from the AtNudt7 promoter do not exhibit any binding to AtERF1 fusion protein. This result confirmed our hypothesis that AtERF1could be the transcriptional factor that binds and regulates the AtNudt7 gene expression during development and under stress.

(a)

```
Ws      361  ACTTCAACAATATGCGGTAGAGTTTGCTGCTGCCGTTAGGCGTTGCCGCCTGTAGTAATG  420
             ||||||||||||||||||||||||||||||||||||||||||        | | |    ||||||
Col-0   361  ACTTCAACAATATGCGGTAGAGTTTGCTGCTGCCGTTA------G--G-C----GTAATG  407
```

(b)

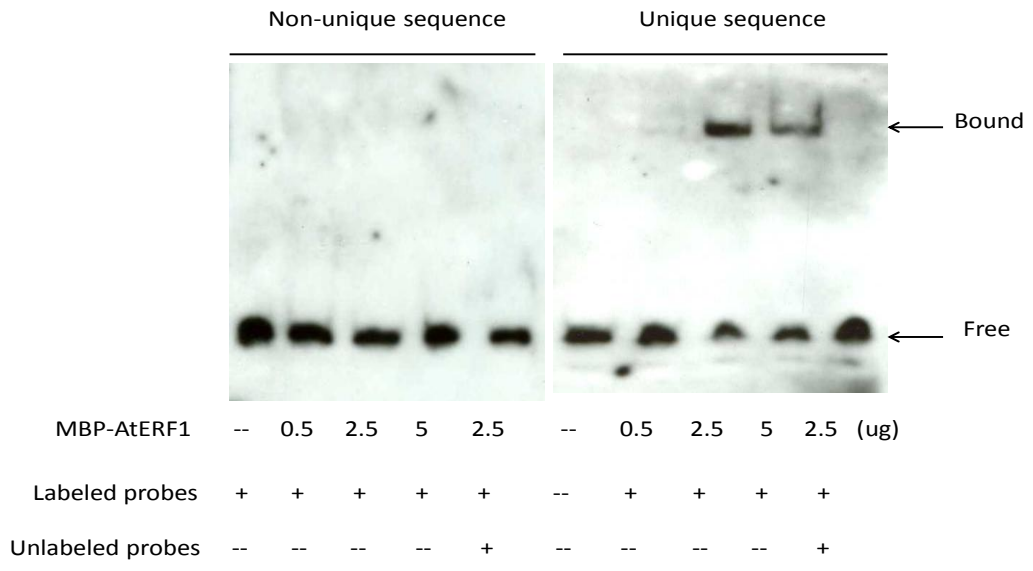


Fig. 11 AtERF1 binds to the 16 bp unique AtNudt7 promoter sequence with GCC-box. (a) The sequence alignment of Ws and Col-0 AtNudt7 promoter region showing the 16 bp unique sequence, red color indicates the GCC box element; (b)Electrophoretic mobility shift assays was performed using MBP–AtERF1 fusion protein and biotin labeled double stranded specific oligonucleotides as a probe.

DISCUSSION

We have identified a reduced seed germination phenotype in the *Arabidopsis Atnudt7-1* mutant (Fig. 1). The reduced germination potential, slower germination rates and hypersensitivity to ABA of *Atnudt7*-1 mutant are very similar to the characteristics described for the nicotinamidase deficient *nic2-1* mutant (Hunt *et al.*, 2007). Unlike *nic2-1* mutant, moist-chilling of *Atnudt7-1* seeds failed to restore the germination potential to WT levels indicating that a functional AtNUDT7 protein is important for normal seed after-ripening in *Arabidopsis*.

It has been suggested that germination potential is programmed during the seed maturation phase (Holdsworth *et al.*, 2008). Based on the information in the *Arabidopsis* eFP browser (http://bbc.botany.utoronto.ca/efp/cgi-bin/efpWeb.cgi?primaryGene=AT4G12720&modeInput=Absolute) (Winter *et al.*, 2007) AtNudt7 gene is expressed during the early stages of embryo development, from the globular to the walking-stick stage (Supplementary Figure 2). It has been shown that stored proteins and translation of pre-existing mRNAs are key players during the germination process (Rajjou *et al.*, 2004). Analysis of the AtNUDT7 protein levels in after-ripened dry seeds over a three-day period following imbibition showed significant variation in Arabidopsis ecotypes (Fig. 4). Based on the reduced germination potential observed in *Atnudt7-1* mutant, gene expression data during seed imbibition, and our observations of AtNUDT7 protein levels in different ecotypes (Fig 4), we speculate that this is an important stored protein that aids in normal seed after-ripening.

What is the role of AtNUDT7 in seed after-ripening? After-ripening of the severely dormant Cvi ecotype showed a significant flux of NMN derived from NAD that may be catalyzed by nudix hydrolases (Hunt & Gray, 2009). Furthermore, majority of the NADH pyrophosphohydrolase activity (53.1%) in *Arabidopsis* is attributed to AtNUDT7 (Ishikawa *et al.*, 2009) and our results indicate that *AtNUDT7* is the nudix hydrolase which could be involved in this process. The NAD^+:NADH ratio showed significant reduction during after-ripening in Cvi (Hunt & Gray, 2009) and in the Col-0 seeds in our studies (Fig. 8). This suggests that the high levels of NADH in seeds lead to activation of the AtNUDT7's NADH pyrophosphatase activity following imbibition, and this activity is vital in order to maintain NAD^+:NADH redox homeostasis.

This begs the question what processes could lead to NADH buildup during seed germination? Upon imbibition the quiescent dry seed rapidly absorbs water but the resumption of metabolic activity is more gradual. One of the earliest changes is the resumption of energy metabolism, especially respiration that can be detected within minutes (Bewley, 1997; Nonogaki *et al.*, 2010). Glycolytic and oxidative pentose phosphate pathway resume during the imbibition phase (Aldasoro & Nicolas, 1979; Salon *et al.*, 1988) and these processes can lead to NADH accumulation. Secondly, following imbibition many seeds experience temporary anaerobic conditions leading to ethanol production (Kennedy *et al.*, 1992). This ethanol is converted to acetaldehyde by alcohol dehydrogenase. The highly reactive acetaldehyde is further converted to acetate by aldehyde dehydrogenase. Each of these reactions also leads to production of NADH. The low oxygen availability in imbibed seeds could restrict mitochondrial ATP production and favor the glycolytic pathway that in turn leads to an increase in the NADH levels.

The high levels of NADH leads to higher catabolic redox charge that favors the accumulation of ROS. We have developed a model where to prevent an excessive buildup of NADH in the cytosol, AtNUDT7 is maintained in the dry seeds and during the early stages of imbibition. The observed increase in the NAD^+:NADH ratio favoring the oxidant by 48 h after imbibition and the concomitant disappearance of AtNUDT7 may be important phase transition markers for the imbibed seeds to move from phase II (*senso stricto* germination) to phase III stages of germination (Nonogaki *et al.*, 2010).

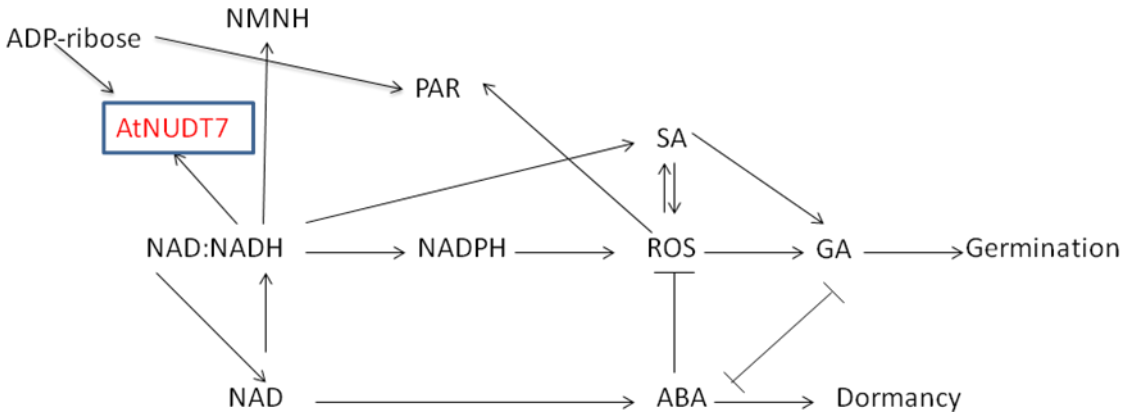


Fig. 12 Proposed model for AtNUDT7 in the process of seed germination. AtNUDT7 regulates NAD: NADH balance in after-ripened seeds and during early stages of seed imbibition. Imbalance in NAD^+ or NADH favors ABA biosynthesis that in turn suppresses ROS, SA and GA and eventually promotes dormancy. Lack of AtNUDT7 also leads to excess ROS that leads to oxidative stress as evidenced by increased PAR activity that eventually reduces germination potential of the seeds.

NMN: nicotinamide mononucleotide: NMNH: nicotinamide mononucleotide, reduced form;

Apart from their traditional roles as redox carriers, recent studies have shown that pyridine nucleotides participate in the biosynthesis of phytohormones such as ABA (Gonzalez-Guzman *et al.*, 2002) and SA (Zhang & Mou, 2009). SA has been shown to be important for priming seed metabolism, synthesis of antioxidant enzymes, and mobilization of seed storage proteins (Rajjou *et al.*, 2006). SA also has been shown to be important in a feed-forward self-amplifying loop resulting in the production of ROS that in turn could lead to oxidative stress and growth retardation (Rao *et al.*, 1997; Ruef *et al.*, 1997; Overmyer *et al.*, 2003). We speculate that the higher levels of SA in the dry seeds of WT (Fig. 7) could prime the higher ROS levels (Fig. 9) and this may be responsible for lowering the levels of ABA (Fig. 6), since SA and ABA antagonism during seed germination has been reported earlier (Kanno *et al.*, 2010). On the same grounds, in the *Atnudt7-1* seeds the lower levels of SA (Fig. 7) and in turn ROS (Fig. 9) could lead to a higher level of ABA (Fig 6) and this in turn leads to the higher level of dormancy. SA also has been shown to stimulate GA biosynthesis especially in seeds under salt stress (Rajjou *et al.*, 2006; Alonso-Ramirez *et al.*, 2009a; Alonso-Ramirez *et al.*, 2009b) and this may also contribute to the higher germination rates of WT compared to *Atnudt7-1* with lower levels of GA.

The significantly higher levels of ROS within 24 h after imbibition in *Atnudt7-1* suggests a functional AtNUDT7 may be involved in maintaining ROS homeostasis during seed maturation and early phases of seed imbibition. Higher levels of ABA have been shown to inhibit ROS and enhance the antioxidant enzyme activity (El-Maarouf-Boteau & Bailly, 2008). ROS, especially H_2O_2 has been shown to break dormancy in several different plant systems and this was attributed to a decrease in the ABA levels (Wang *et al.*, 1995; Naredo *et al.*, 1998) Consistent with these reports, the higher ABA levels of *Atnudt7-1* in fresh seeds are accompanied by lower levels of ROS when compared with the WT seeds. Removal of ABA

during the early hours of imbibition (approximately 6 hours) (Preston *et al.*, 2009) may negatively impact the antioxidant enzyme activity leading to massive increase in ROS levels in the *Atnudt7-1* mutant by 24 h. The lack of NADH pyrophosphohydrolase activity of AtNUDT7 may lead to superoxide generation by the sequential actions of NADH kinase and NADPH oxidase (Torres *et al.*, 1998; Murata *et al.*, 2001; Kwak *et al.*, 2003; Hunt *et al.*, 2004). This vast excess of ROS will cause an oxidative stress as evidenced by the significant increases in the PAR levels (Fig. 10) and eventually lead to growth retardation phenotype observed in the slowly germinating *Atnudt7-1* mutant. In fact, AtNUDT7 was shown to have a significant impact on the DNA repair pathways via modulation of PAR reaction under oxidative stress conditions (Ishikawa *et al.*, 2009).

PNs, especially NAD that can participate in hydride transfer as well as ADP-ribose transfer reactions serves as a vital link connecting cellular metabolism and signaling (Ziegler, 2000; Hunt *et al.*, 2004). Our study shows that AtNUDT7 protein plays a very important role during seed after-ripening and early stages of imbibition by regulating PN homeostasis that in turn impacts hormone biosynthesis and ROS metabolism (Fig. 12). The fine balance of SA, ABA, GA and ROS in seeds and during early phases of imbibition plays a key role in determining the seed germination potential.

Understanding the regulation of AtNUDT7 gene will further aid in dissecting the complex interplay between redox homeostasis, phytohormones and seed germination. AtERF1 (*Arabidopsis* Ethylene Response Factor1) are triggered by stress signals such as: salt stress, SA and viral pathogen infection (Guo & Ecker, 2004). AtERF genes may be up regulated via an ethylene-dependent pathway or an ethylene-independent pathway (Fujimoto *et al.*, 2000).

AtERF1 was induced by ethylene and can interact with GCC box–containing stress response genes (Solano *et al.*, 1998). GCC box has been reported to act as a *cis*-regulatory element for biotic and abiotic stress signal transduction in upregulating the expression of GCC box–containing stress responsive genes dependent on or independent of the ethylene signal (Solano *et al.*, 1998). In beech seeds ERF1 expression was undetectable in dormant seeds but accumulated in non-dormant seeds during germination (Jiminez *et al.*, 2005) Several other studies have shown the participation of ERF in the regulation of germination (Leubner-Metzger *et al.*, 1998; Song *et al.*, 2005; Pirrello *et al.*, 2006) Besides ethylene, jasmonate signaling pathways could also regulate AtERF1 expression (Lorenzo *et al.*, 2003). The higher levels of jasmonates observed in the dry arabidopsis seeds could be a potent inducer of AtERF1 and in turn AtNudt7. ROS generating compounds such as cyanide and methyl viologen have been shown to induce the ERF expression in the seeds (Oracz *et al.*, 2009). Thus a complex interaction involving phytohormones and ROS may be involved in regulating ERF1. Identifying other trans-acting factors that are important for AtNUDT7 regulation will provide a better understanding of the early changes occurring in the seed following imbibition.

CHAPTER III

REFERENCES

Aldasoro J, Nicolas G. 1979. Change in the concentrations of glycolytic intermediates and adenosine phosphates during germination of seeds of *Cicer arietinum L*. *Int. J. Biochem* **10**: 947-950.

Alonso-Ramirez A, Rodriguez D, Reyes D, Jimenez JA, Nicolas G, Lopez-Climent M, Gomez-Cadenas A, Nicolas C. 2009a. Cross-talk between gibberellins and salicylic acid in early stress responses in Arabidopsis thaliana seeds. *Plant Signal Behav* **4**(8): 750-751.

Alonso-Ramirez A, Rodriguez D, Reyes D, Jimenez JA, Nicolas G, Lopez-Climent M, Gomez-Cadenas A, Nicolas C. 2009b. Evidence for a role of gibberellins in salicylic acid-modulated early plant responses to abiotic stress in Arabidopsis seeds. *Plant Physiol* **150**(3): 1335-1344.

Ames. 1966. Assay of inorganic phosphate, total phosphate and phosphatases. *Methods Enzymol* **1**: 115-118.

Andersen JR, Lubberstedt T. 2003. Functional markers in plants. *Trends Plant Sci* **8**(11): 554-560.

Barbazuk WB, Emrich S, Schnable PS. 2007. SNP Mining from Maize 454 EST Sequences. *CSH Protoc* **2007**: pdb prot4786.

Bartsch M, Gobbato E, Bednarek P, Debey S, Schultze JL, Bautor J, Parker JE. 2006. Salicylic acid-independent ENHANCED DISEASE SUSCEPTIBILITY1 signaling in Arabidopsis immunity and cell death is regulated by the monooxygenase FMO1 and the Nudix hydrolase NUDT7. *Plant Cell* **18**(4): 1038-1051.

Beaudoin N, Serizet C, Gosti F, Giraudat J. 2000. Interactions between abscisic acid and ethylene signaling cascades. *Plant Cell* **12**(7): 1103-1115.

Bentsink L, Jowett J, Hanhart CJ, Koornneef M. 2006. Cloning of DOG1, a quantitative trait locus controlling seed dormancy in Arabidopsis. *Proc Natl Acad Sci U S A* **103**(45): 17042-17047.

Bewley JD. 1997. Seed Germination and Dormancy. *Plant Cell* **9**(7): 1055-1066.

Bouton JH. 2007. Molecular breeding of switchgrass for use as a biofuel crop. *Curr Opin Genet Dev* **17**(6): 553-558.

Chen Q, Zhang B, Hicks LM, Wang S, Jez JM. 2009. A liquid chromatography-tandem mass spectrometry-based assay for indole-3-acetic acid-amido synthetase. *Anal Biochem* **390**(2): 149-154.

Chi KR. 2008. The year of sequencing. *Nat Methods* **5**(1): 11-14.

Chibani K, Ali-Rachedi S, Job C, Job D, Jullien M, Grappin P. 2006. Proteomic analysis of seed dormancy in Arabidopsis. *Plant Physiol* **142**(4): 1493-1510.

Clough SJ. 2005. Floral dip: agrobacterium-mediated germ line transformation. *Methods Mol Biol* **286**: 91-102.

Clough SJ, Bent AF. 1998. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J* **16**(6): 735-743.

Debeaujon I, Koornneef M. 2000. Gibberellin requirement for Arabidopsis seed germination is determined both by testa characteristics and embryonic abscisic acid. *Plant Physiol* **122**(2): 415-424.

Debeaujon I, Leon-Kloosterziel KM, Koornneef M. 2000. Influence of the testa on seed dormancy, germination, and longevity in Arabidopsis. *Plant Physiol* **122**(2): 403-414.

Dobrzanska M, Szurmak B, Wyslouch-Cieszynska A, Kraszewska E. 2002. Cloning and characterization of the first member of the Nudix family from Arabidopsis thaliana. *J Biol Chem* **277**(52): 50482-50486.

Dunn CA, O'Handley SF, Frick DN, Bessman MJ. 1999. Studies on the ADP-ribose pyrophosphatase subfamily of the nudix hydrolases and tentative identification of trgB, a gene associated with tellurite resistance. *J Biol Chem* **274**(45): 32318-32324.

El-Maarouf-Boteau H, Bailly C. 2008. Oxidative signaling in seed germination and dormancy. *Plant Signal Behav* **3**(3): 175-182.

Emrich SJ, Barbazuk WB, Li L, Schnable PS. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* **17**(1): 69-73.

Finch-Savage WE, Leubner-Metzger G. 2006. Seed dormancy and the control of germination. *New Phytol* **171**(3): 501-523.

Fujimoto SY, Ohta M, Usui A, Shinshi H, Ohme-Takagi M. 2000. Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. *Plant Cell* **12**(3): 393-404.

Gallais S, Crescenzo M-APD, Laval-Martin DL. 1998. Pyridine Nucleotides and Redox Charges during Germination of Non-dormant and Dormant Caryopses of *Avena sativa* L. *Journal of Plant Physiology* **153**: 664-669.

Ge X, Xia Y. 2008. The role of AtNUDT7, a Nudix hydrolase, in the plant defense response. *Plant Signal Behav* **3**(2): 119-120.

Ghassemian M, Nambara E, Cutler S, Kawaide H, Kamiya Y, McCourt P. 2000. Regulation of abscisic acid signaling by the ethylene response pathway in Arabidopsis. *Plant Cell* **12**(7): 1117-1126.

Gonzalez-Guzman M, Apostolova N, Belles JM, Barrero JM, Piqueras P, Ponce MR, Micol JL, Serrano R, Rodriguez PL. 2002. The short-chain alcohol dehydrogenase ABA2 catalyzes the conversion of xanthoxin to abscisic aldehyde. *Plant Cell* **14**(8): 1833-1846.

Guo H, Ecker JR. 2004. The ethylene signaling pathway: new insights. *Curr Opin Plant Biol* **7**(1): 40-49.

Hayashi M, Takahashi H, Tamura K, Huang J, Yu LH, Kawai-Yamada M, Tezuka T, Uchimiya H. 2005. Enhanced dihydroflavonol-4-reductase activity and NAD homeostasis leading to cell death tolerance in transgenic rice. *Proc Natl Acad Sci U S A* **102**(19): 7020-7025.

Holdsworth MJ, Bentsink L, Soppe WJ. 2008. Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. *New Phytol* **179**(1): 33-54.

Hunt L, Gray JE. 2009. The relationship between pyridine nucleotides and seed dormancy. *New Phytol* **181**(1): 62-70.

Hunt L, Holdsworth MJ, Gray JE. 2007. Nicotinamidase activity is important for germination. *Plant J* **51**(3): 341-351.

Hunt L, Lerner F, Ziegler M. 2004. NAD-New roles in signalling and gene regulation in plants. *New Phytol* **163**: 31-44.

Ishikawa K, Ogawa T, Hirosue E, Nakayama Y, Harada K, Fukusaki E, Yoshimura K, Shigeoka S. 2009. Modulation of the poly(ADP-ribosyl)ation reaction via the Arabidopsis ADP-ribose/NADH pyrophosphohydrolase, AtNUDX7, is involved in the response to oxidative stress. *Plant Physiol* **151**(2): 741-754.

Jambunathan N, Mahalingam R. 2006. Analysis of Arabidopsis growth factor gene 1 (GFG1) encoding a nudix hydrolase during oxidative signaling. *Planta* **224**(1): 1-11.

Jambunathan N, Penaganti A, Tang Y, Mahalingam R. 2010. Modulation of redox homeostasis under suboptimal conditions by Arabidopsis nudix hydrolase 7. *BMC Plant Biol* **10**: 173.

Jefferson RA. 1989. The GUS reporter gene system. *Nature* **342**(6251): 837-838.

Jiminez HR, Pacheco O, Blanco H, Chamorro DR, Tricarico JM. 2005. Effects of yeast culture and natural saponin sources on ruminal microbial populations and tropical forage digestion in vitro. *Journal of Animal Science* **83**: 310-310.

Joo JH, Wang S, Chen JG, Jones AM, Fedoroff NV. 2005. Different signaling and cell death roles of heterotrimeric G protein alpha and beta subunits in the Arabidopsis oxidative stress response to ozone. *Plant Cell* **17**(3): 957-970.

Kanno Y, Jikumaru Y, Hanada A, Nambara E, Abrams SR, Kamiya Y, Seo M. 2010. Comprehensive hormone profiling in developing Arabidopsis seeds: examination of the site of ABA biosynthesis, ABA transport and hormone interactions. *Plant Cell Physiol* **51**(12): 1988-2001.

Katsir L, Chung HS, Koo AJ, Howe GA. 2008. Jasmonate signaling: a conserved mechanism of hormone sensing. *Curr Opin Plant Biol* **11**(4): 428-435.

Kaur S, Cogan NO, Pembleton LW, Shinozuka M, Savin KW, Materne M, Forster JW. 2011. Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics* **12**: 265.

Kennedy RA, Rumpho ME, Fox TC. 1992. Anaerobic metabolism in plants. *Plant Physiol* **100**(1): 1-6.

Keshwani DR, Cheng JJ. 2009. Switchgrass for bioethanol and other value-added applications: a review. *Bioresour Technol* **100**(4): 1515-1523.

Koornneef M, Bentsink L, Hilhorst H. 2002. Seed dormancy and germination. *Curr Opin Plant Biol* **5**(1): 33-36.

Kraszewska E. 2008. The plant Nudix hydrolase family. *Acta Biochim Pol* **55**(4): 663-671.

Kucera B, Cohn MA, Leubner-Metzger G. 2005. Plant hormone interactions during seed dormancy release and germination. *Seed Sci Res* **15**: 281-307.

Kwak JM, Mori IC, Pei ZM, Leonhardt N, Torres MA, Dangl JL, Bloom RE, Bodde S, Jones JD, Schroeder JI. 2003. NADPH oxidase AtrbohD and AtrbohF genes function in ROS-dependent ABA signaling in Arabidopsis. *EMBO J* **22**(11): 2623-2633.

Leon-Kloosterziel KM, Gil MA, Ruijs GJ, Jacobsen SE, Olszewski NE, Schwartz SH, Zeevaart JA, Koornneef M. 1996. Isolation and characterization of abscisic acid-deficient Arabidopsis mutants at two new loci. *Plant J* **10**(4): 655-661.

Leubner-Metzger G, Petruzzelli L, Waldvogel R, Vogeli-Lange R, Meins F. 1998. Ethylene-responsive element binding protein (EREBP) expression and the transcriptional regulation of class I beta-1,3-glucanase during tobacco seed germination. *Plant Molecular Biology* **38**(5): 785-795.

Lorenzo O, Piqueras R, Sanchez-Serrano JJ, Solano R. 2003. ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense. *Plant Cell* **15**(1): 165-178.

Mahalingam R, Gomez-Buitrago A, Eckardt N, Shah N, Guevara-Garcia A, Day P, Raina R, Fedoroff NV. 2003. Characterizing the stress/defense transcriptome of Arabidopsis. *Genome Biol* **4**(3): R20.

Mahalingam R, Jambunathan N, Penaganti A. 2007. Pyridine nucleotide homeostasis in plant development and stress. *Int J Plant Dev Biol* **1**: 194-201.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**(3): 133-141.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Matts J, Jagadeeswaran G, Roe BA, Sunkar R. 2010. Identification of microRNAs and their targets in switchgrass, a model biofuel plant species. *Journal of Plant Physiology* **167**(11): 896-904.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11**(1): 31-46.

Morozova O, Marra MA. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**(5): 255-264.

Muller K, Carstens AC, Linkies A, Torres MA, Leubner-Metzger G. 2009. The NADPH-oxidase AtrbohB plays a role in Arabidopsis seed after-ripening. *New Phytol* **184**(4): 885-897.

Murata Y, Pei ZM, Mori IC, Schroeder J. 2001. Abscisic acid activation of plasma membrane Ca(2+) channels in guard cells requires cytosolic NAD(P)H and is differentially disrupted upstream and downstream of reactive oxygen species production in abi1-1 and abi2-1 protein phosphatase 2C mutants. *Plant Cell* **13**(11): 2513-2523.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J. 2008. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Mol Ecol* **17**(16): 3599-3613.

Naredo MEB, Juliano AB, Lu BR, De Guzman F, Jackson MT. 1998. Responses to seed dormancy-breaking treatments in rice species (*Oryza L*.). *Seed Sci Tech* **26**: 675-689.

Nonogaki H, Bassel GW, Bewley JD. 2010. Germination- Still a mystery. *Plant Sci*.

Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr., Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics* **9**: 312.

Ogawa M, Hanada A, Yamauchi Y, Kuwahara A, Kamiya Y, Yamaguchi S. 2003. Gibberellin biosynthesis and response during Arabidopsis seed germination. *Plant Cell* **15**(7): 1591-1604.

Ogawa T, Ueda Y, Yoshimura K, Shigeoka S. 2005. Comprehensive analysis of cytosolic Nudix hydrolases in Arabidopsis thaliana. *J Biol Chem* **280**(26): 25277-25283.

Okada M, Lanzatella C, Saha MC, Bouton J, Wu R, Tobias CM. 2010. Complete switchgrass genetic maps reveal subgenome collinearity, preferential pairing and multilocus interactions. *Genetics* **185**(3): 745-760.

Olejnik K, Kraszewska E. 2005. Cloning and characterization of an Arabidopsis thaliana Nudix hydrolase homologous to the mammalian GFG protein. *Biochim Biophys Acta* **1752**(2): 133-141.

Oracz K, El-Maarouf-Bouteau H, Kranner I, Bogatek R, Corbineau F, Bailly C. 2009. The mechanisms involved in seed dormancy alleviation by hydrogen cyanide unravel the role of reactive oxygen species as key factors of cellular signaling during germination. *Plant Physiol* **150**(1): 494-505.

Overmyer K, Brosche M, Kangasjarvi J. 2003. Reactive oxygen species and hormonal control of cell death. *Trends Plant Sci* **8**(7): 335-342.

Palmer NA SA, Kim J, Benson A, Tobias CM, et al. 2011. Next-generation sequencing of crown and rhizome transcriptome from an upland, tetraploid switchgrass. *Bioenergy Research DOI* **10**(1007): s12155-12011-19171-12151.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA. 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**: 180.

Penfield S, King J. 2009. Towards a systems biology approach to understanding seed dormancy and germination. *Proc Biol Sci* **276**(1673): 3561-3569.

Pirrello J, Jaimes-Miranda F, Sanchez-Ballesta MT, Tournier B, Khalil-Ahmad Q, Regad F, Latche A, Pech JC, Bouzayen M. 2006. Sl-ERF2, a tomato ethylene response factor involved in ethylene response and seed germination. *Plant Cell Physiol* **47**(9): 1195-1205.

Preston J, Tatematsu K, Kanno Y, Hobo T, Kimura M, Jikumaru Y, Yano R, Kamiya Y, Nambara E. 2009. Temporal expression patterns of hormone metabolism genes during imbibition of Arabidopsis thaliana seeds: a comparative study on dormant and non-dormant accessions. *Plant Cell Physiol* **50**(10): 1786-1800.

Puckette M, Peal L, Steele J, Tang Y, Mahalingam R. 2009. Ozone responsive genes in Medicago truncatula: analysis by suppression subtraction hybridization. *J Plant Physiol* **166**(12): 1284-1295.

Rajjou L, Belghazi M, Huguet R, Robin C, Moreau A, Job C, Job D. 2006. Proteomic investigation of the effect of salicylic acid on Arabidopsis seed germination and establishment of early defense mechanisms. *Plant Physiol* **141**(3): 910-923.

Rajjou L, Gallardo K, Debeaujon I, Vandekerckhove J, Job C, Job D. 2004. The effect of alpha-amanitin on the Arabidopsis seed proteome highlights the distinct roles of stored and neosynthesized mRNAs during germination. *Plant Physiol* **134**(4): 1598-1613.

Rao MV, Paliyath G, Ormrod DP, Murr DP, Watkins CB. 1997. Influence of salicylic acid on H2O2 production, oxidative stress, and H2O2-metabolizing enzymes. Salicylic acid-mediated oxidative damage requires H2O2. *Plant Physiol* **115**(1): 137-149.

Ribeiro JM, Carloto A, Costas MJ, Cameselle JC. 2001. Human placenta hydrolases active on free ADP-ribose: an ADP-sugar pyrophosphatase and a specific ADP-ribose pyrophosphatase. *Biochim Biophys Acta* **1526**(1): 86-94.

Ruef J, Hu ZY, Yin LY, Wu Y, Hanson SR, Kelly AB, Harker LA, Rao GN, Runge MS, Patterson C. 1997. Induction of vascular endothelial growth factor in balloon-injured baboon arteries. A novel role for reactive oxygen species in atherosclerosis. *Circ Res* **81**(1): 24-33.

Russell L, Larner V, Kurup S, Bougourd S, Holdsworth M. 2000. The Arabidopsis COMATOSE locus regulates germination potential. *Development* **127**(17): 3759-3767.

Salon C, Raymond P, Pradet A. 1988. Quantification of carbon fluxes through the tricarboxylic acid cycle in early germinating lettuce embryos. *J Biol Chem* **263**(25): 12278-12287.

Sarath G, Hou G, Baird LM, Mitchell RB. 2007. Reactive oxygen species, ABA and nitric oxide interactions on the germination of warm-season C4-grasses. *Planta* **226**(3): 697-708.

Schmer MR, Vogel KP, Mitchell RB, Perrin RK. 2008. Net energy of cellulosic ethanol from switchgrass. *Proc Natl Acad Sci U S A* **105**(2): 464-469.

Schuster SC. 2008. Next-generation sequencing transforms today's biology. *Nat Methods* **5**(1): 16-18.

Shen H, He X, Poovaiah CR, Wuddineh WA, Ma J, Mann DG, Wang H, Jackson L, Tang Y, Stewart CN, Jr., Chen F, Dixon RA. 2012. Functional characterization of the switchgrass (Panicum virgatum) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. *New Phytol* **193**(1): 121-136.

Solano R, Stepanova A, Chao Q, Ecker JR. 1998. Nuclear events in ethylene signaling: a transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1. *Genes Dev* **12**(23): 3703-3714.

Song LP, Zhao W, Huang JJ, Zhu SZ. 2005. Synthesis, characterization of neutral nickel complexes bearing N-fluorophenylsalicylaidimine chelate ligands and their catalytic activity to ethylene oligomerization. *Chinese Journal of Chemistry* **23**(6): 669-672.

Steber CM, Cooney SE, McCourt P. 1998. Isolation of the GA-response mutant sly1 as a suppressor of ABI1-1 in Arabidopsis thaliana. *Genetics* **149**(2): 509-521.

Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S. 2010. De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics* **11**: 262.

Tobias CM HD, Twigg P, Sarath G. 2006. Genic microsatellite markers derived from EST sequences of switchgrass (Panicum virgatum L.). *Molecular Ecology Notes* **6**: 185--187.

Tobias CM, Sarath, G., Twigg, P., Lindquist, E., Pangilinan, J., Penning, P.W., Barry, K., McCann, M.C., Carpita, N.C., Lazo, G.R. 2008. Comparative genomics in switchgrass using 61,585 high-quality expressed sequence tags. *The Plant Genome* **1**: 111-124.

Torres MA, Onouchi H, Hamada S, Machida C, Hammond-Kosack KE, Jones JD. 1998. Six Arabidopsis thaliana homologues of the human respiratory burst oxidase (gp91phox). *Plant J* **14**(3): 365-370.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* **17**(7): 1636-1647.

Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H, Landherr L, Tomsho LP, Hu Y, Carlson JE, Ma H, Schuster SC, Soltis DE, Soltis PS, Altman N, dePamphilis CW. 2009. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* **10**: 347.

Wang L, Li P, Brutnell TP. 2010. Exploring plant transcriptomes using ultra high-throughput sequencing. *Brief Funct Genomics* **9**(2): 118-128.

Wang M, Heimovaara-Dijkstra S, Van Duijn B. 1995. Modulation of germination of embryos isolated from dormant and nondormant barley grains by manipulation of endogenous abscisic levels. *Planta* **195**: 586-592.

Wang W, Wang Y, Zhang Q, Qi Y, Guo D. 2009. Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* **10**: 465.

Wang YW, Samuels TD, Wu YQ. 2011. Development of 1,030 genomic SSR markers in switchgrass. *Theor Appl Genet* **122**(4): 677-686.

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. 2007. An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS One* **2**(8): e718.

Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P. 2012. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am J Bot* **99**(2): 193-208.

Zhang X, Mou Z. 2009. Extracellular pyridine nucleotides induce PR gene expression and disease resistance in Arabidopsis. *Plant J* **57**: 302-312.

Zhang Y, Zhang B, Yan D, Dong W, Yang W, Li Q, Zeng L, Wang J, Wang L, Hicks LM, He Z. 2011. Two Arabidopsis cytochrome P450 monooxygenases, CYP714A1 and CYP714A2, function redundantly in plant development through gibberellin deactivation. *Plant J* **67**(2): 342-353.

Ziegler M. 2000. New functions of a long-known molecule. Emerging roles of NAD in cellular signaling. *Eur J Biochem* **267**(6): 1550-1564.

CHAPTER IV

EXPLORING THE SWITCHGRASS TRANSCRIPTOME USING SECOND-GENERATION SEQUENCING TECHNOLOGY

Appendix

OPEN ACCESS Freely available online

PLoS one

Exploring the Switchgrass Transcriptome Using Second-Generation Sequencing Technology

Yixing Wang[1☯], Xin Zeng[1☯], Niranjani J. Iyer[1¤], Douglas W. Bryant[2], Todd C. Mockler[2], Ramamurthy Mahalingam[1*]

1 Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, Oklahoma, United States of America, 2 Donald Danforth Plant Science Center, St. Louis, Missouri, United States of America

Abstract

Background: Switchgrass (*Panicum virgatum* L.) is a C4 perennial grass and widely popular as an important bioenergy crop. To accelerate the pace of developing high yielding switchgrass cultivars adapted to diverse environmental niches, the generation of genomic resources for this plant is necessary. The large genome size and polyploid nature of switchgrass makes whole genome sequencing a daunting task even with current technologies. Exploring the transcriptional landscape using next generation sequencing technologies provides a viable alternative to whole genome sequencing in switchgrass.

Principal Findings: Switchgrass cDNA libraries from germinating seedlings, emerging tillers, flowers, and dormant seeds were sequenced using Roche 454 GS-FLX Titanium technology, generating 980,000 reads with an average read length of 367 bp. *De novo* assembly generated 243,600 contigs with an average length of 535 bp. Using the foxtail millet genome as a reference greatly improved the assembly and annotation of switchgrass ESTs. Comparative analysis of the 454-derived switchgrass EST reads with other sequenced monocots including Brachypodium, sorghum, rice and maize indicated a 70–80% overlap. RPKM analysis demonstrated unique transcriptional signatures of the four tissues analyzed in this study. More than 24,000 ESTs were identified in the dormant seed library. *In silico* analysis indicated that there are more than 2000 EST-SSRs in this collection. Expression of several orphan ESTs was confirmed by RT-PCR.

Significance: We estimate that about 90% of the switchgrass gene space has been covered in this analysis. This study nearly doubles the amount of EST information for switchgrass currently in the public domain. The celerity and economical nature of second-generation sequencing technologies provide an in-depth view of the gene space of complex genomes like switchgrass. Sequence analysis of closely related members of the NAD$^+$-malic enzyme type C4 grasses such as the model system *Setaria viridis* can serve as a viable proxy for the switchgrass genome.

Citation: Wang Y, Zeng X, Iyer NJ, Bryant DW, Mockler TC, et al. (2012) Exploring the Switchgrass Transcriptome Using Second-Generation Sequencing Technology. PLoS ONE 7(3): e34225. doi:10.1371/journal.pone.0034225

Editor: Nicholas James Provart, University of Toronto, Canada

Received November 14, 2011; **Accepted** February 27, 2012; **Published** March 29, 2012

Funding: This material is based upon work supported by the National Science Foundation under Grant No. EPS-0814361. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Doug Bryant is affiliated to Intuitive Genomics. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: ramamurthy.mahalingam@okstate.edu

☯ These authors contributed equally to this work.

¤ Current address: Monsanto, Chesterfield, Missouri, United States of America

Introduction

Even though genome sequencing technologies have become progressively more efficient over the last few years, complete sequencing of complex plant genomes is still technically challenging and cost prohibitive. Identification of transcribed portions of the genome using expressed sequence tags (ESTs) provides a viable alternative for analyzing non-model systems and organisms with large genome sizes, wherein whole genome sequencing is daunting. ESTs have high functional information and have been proven to be valuable for gene annotation and gene discovery [1,2,3]. ESTs have been useful for development of molecular markers [4,5,6,7,8], comparative genomics [9,10] and for genetic analysis of adaptive traits [11,12]. *Ipso facto*, genes are expressed in particular tissues or cell types, developmental stages and vary in their expression levels by several orders of magnitude. Traditional EST projects require substantial investments in terms of library construction and sequencing, especially if the goal is to capture rare transcripts [13].

Next generation sequencing (NGS) technologies such as pyrosequencing, bypass lengthy and relatively low throughput steps involved in Sanger sequencing and provide rapid and economical technologies for transcriptomics [14,15,16,17,18,19,20]. To date, the massively parallel DNA sequencing developed by Roche Life Sciences called 454 pyrosequencing is the most widely used next-generation technology for *de novo* sequencing and analysis of transcriptomes of non-model systems. The first commercial NGS platform, the 454 GS20, produced 200,000 reads with an average read length of 100 bases per run [14,18]. Rapid improvements in emulsion PCR and sequencing chemistry

have greatly improved the throughput, read-length and accuracy of 454 sequencing technology [21]. The newest 454-sequencing platform, GS FLX Titanium, can generate a million reads with an average read length of 400 bases at 99.5% accuracy per run.

Switchgrass (*Panicum virgatum*) is a C4 perennial grass selected in 1991 by the Department of Energy as a model herbaceous energy crop for the development of a renewable feedstock resource to produce transportation fuel [22]. This choice has been attributed to several features of this plant native to North America: 1. Biomass – switchgrass plants can grow 3–8 feet tall depending on ecotype; 2. Low input – switchgrass can thrive on marginal lands with minimal input of nutrients and water; and 3. Carbon sink – the large and fibrous root system of switchgrass serves as a major reservoir of captured carbon [22,23,24]. To further accelerate the pace of switchgrass breeding several groups have embarked on developing genomic resources including SSR markers [25,26,27,28], ESTs [9,29] and miRNAs [30].

In this study we conducted 454 based transcriptome analysis in four different switchgrass tissues that are under-represented in the current EST collections – dormant seeds, germinating seedlings, emerging tillers and flowers. We describe the *de novo* assembly of these ESTs, and assembly and annotation of ESTs using the foxtail millet draft genome as a reference. Second, we discuss the transcriptome coverage using proxy methods in the absence of the switchgrass genome sequence. Thirdly, we assessed the expression profiles from these four tissue samples. Fourthly, we examined these ESTs for predicting more than 2000 SSRs that can be very useful for mapping agronomic traits and population genetic studies in switchgrass.

Results

454 sequencing

Four normalized cDNA pools using RNA extracted from dormant seeds, seedlings, tillers and flowers of switchgrass were created. Pyrosequencing of these cDNA pools on a 454 Life Sciences FLX Titanium platform produced approximately 360 million base pairs (Mbp) of sequence data, in the form of 979,903 reads. The cDNA library from dormant seeds had the lowest number of reads (Table 1). The longest read (695 bp) and largest number of total bases sequenced were from the flower sample.

Filtering and de novo assembly

Filtering was done to remove poor quality sequence reads, ESTs that were less than 100 bp after trimming the adapters and polyA/T, and ESTs that matched the NCBI prokaryote sequences. Following these filtering parameters, 69,506 reads were removed. The average read length of cleaned reads was 367 bp (Fig. 1). Pre-clustering using a custom BLAST-like alignment tool [31]-based pipeline was conducted with 910,397 454 EST sequences and 545,894 switchgrass ESTs obtained from Genbank, totaling 1,456,291 sequences. Pre-clustering created relatively smaller (than the entire dataset) groups of overlapping/partially-overlapping reads that were then *de novo* assembled. To accommodate the potential for multiple homologs given the polyploidy in switchgrass [32], the clustering and assembly approach allowed for individual ESTs to exist in more than one cluster. The ESTs were assembled into 243,601 contigs, while 215,923 ESTs remained unassembled. The assembled contigs had a mean length of 535 bp (Figure 2A). About 65% of the reads contributed to contigs that were between 200–600 bp long and nearly 87% of the assembled contigs had between 2–50 EST reads (Figure 2B). This highly left-skewed distribution of assembled contigs is typical for normalized libraries and confirmed that cDNA normalization was effective [33].

Table 1. Raw metrics of 454 sequencing of switchgrass transcriptome.

	Total	Seeds	Seedling	Tiller	Flower
FltrPassWells	979,903	193,511	270,778	246,073	269,541
Total Bases	359,009,624	74,023,325	85,638,441	97,098,266	102,249,592
Avg. length	367	383	317	395	380
Median Len	395	415	331	432	421
Longest Read	695	643	624	652	695

doi:10.1371/journal.pone.0034225.t001

BLAST analysis was undertaken to estimate the proportions of grass genes represented in the switchgrass transcriptome data. This analysis showed that the switchgrass transcriptome data represents up to 24,675 Brachypodium transcripts (79.5% of the 31,029 Brachypodium v1.2 transcripts analyzed), 28,375 sorghum transcripts (78% of the 36,338 sorghum transcripts analyzed), 41,625 rice transcripts (73% of the 56,797 rice transcripts analyzed) and 69,794 maize transcripts (71.5% of the 97,522 maize genes analyzed). These results are consistent with the broad sampling of the switchgrass transcriptome expected using the 454 pyrosequencing technology, assuming that although switchgrass is polyploid the gene space is not very different from other grasses.

Reference based assembly

Availability of the foxtail millet genome sequence (http://www.phytozome.net) prompted a reference guided-sequence assembly of the switchgrass ESTs. Foxtail millet is the closest member of the Panicoideae subfamily with a sequenced genome and a shared common ancestry with switchgrass between 7–10 million years ago [34,35,36]. This analysis yielded 98,086 assembled contigs and 82,902 unassembled ESTs. In the Setaria genome assembly the nine scaffolds corresponding to the nine chromosome pseudomolecules comprise nearly 99% of the total sequence length (http://www.phytozome.net). The assembled contigs and the unassembled ESTs were assigned to the foxtail millet genome in 0.5-Mb intervals based on BLAST similarity scores. Biases in the distribution of the ESTs in the central portions of each chromosome in the pericentric regions were obvious (Figure 3). Curiously, the representation of the assembled contigs as well as

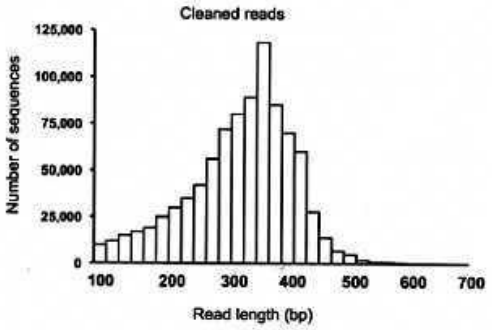


Figure 1. Frequency distribution of 454 sequencing read lengths. Histogram of Roche 454 GS-FLX Titanium read lengths after filtering and trimming adapters.
doi:10.1371/journal.pone.0034225.g001

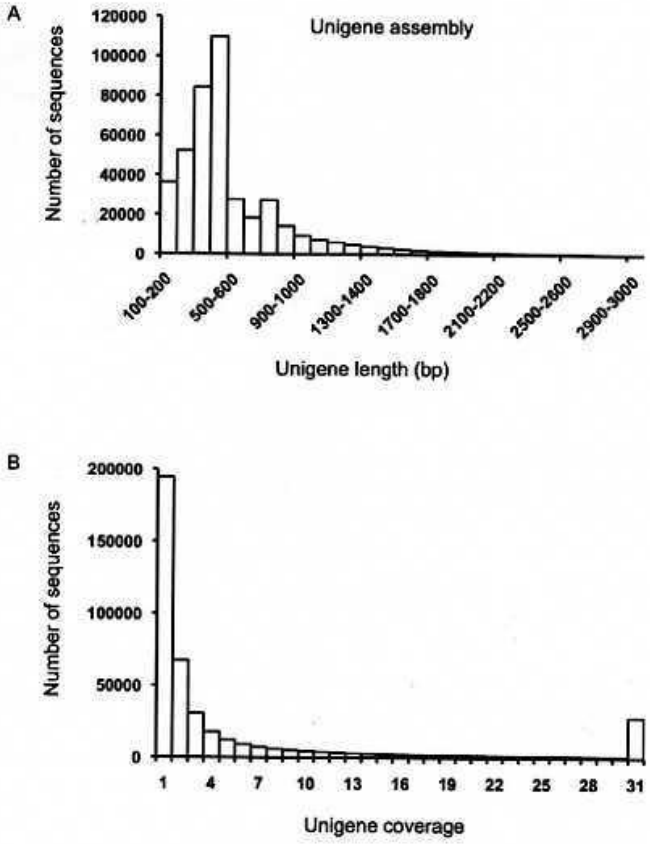


Figure 2. Unigene assembly features of switchgrass transcriptome. (A) Histogram of contig lengths following the 2-step *de novo* assembly process. The x-axis has been truncated at 3 kb. The longest contig is 12,437 base pairs. (B) Histogram of the average read-depth coverage for assembled contigs. Coverage values greater than 30× have been binned together.
doi:10.1371/journal.pone.0034225.g002

unassembled ESTs on the foxtail millet chromosomes 7 and 8 was extremely low.

Gene Ontology (GO) Annotations

Plant specific GO slim terms associated with 36,080 (36.7%) of the 98,086 assembled EST contigs were available. Of these, assignment of contigs to the cellular component made up the majority (49,028), followed by biological process category (12,075) and molecular function category (3916). The GO categories represented in the switchgrass transcriptome did not show any significant biases and showed similar distribution patterns reported in other plant species [12,37,38]. The majority of contigs with annotations for cellular component category were in plastids or mitochondria, but a large number were also associated with vesicles and membrane (Figure 4). Predominant contig annotations for the biological processes category were reproduction, followed by translation, transport and response to stress (Figure 4). DNA binding, RNA binding and protein binding were the major GOs associated with the molecular function (Figure 4).

Transcriptome coverage

Since the sequence of the switchgrass genome is unavailable, the actual size and composition of the transcriptome is not known. We used a simulation-based tool, ESTcalc [13], to approximate the coverage of switchgrass transcriptome using 454-pyrosequencing data. Based on this simulation analysis, our dataset covers 92% of the transcriptome, with every gene represented by at least one read (Table 2). We compared the NCBI switchgrass Unigene set containing 20,973 genes with the ESTs from our 454 analyses. BLAST analysis indicated that 20,963 of the 20,973 (99.95%) Genbank unigenes are represented by 125,341 contigs derived from the current 454 assemblies. It should be noted that our assembly process allows individual ESTs to belong to more than one EST cluster. This approach is intended to reduce erroneous assembly of homologs, paralogs, and splice variants, but results in

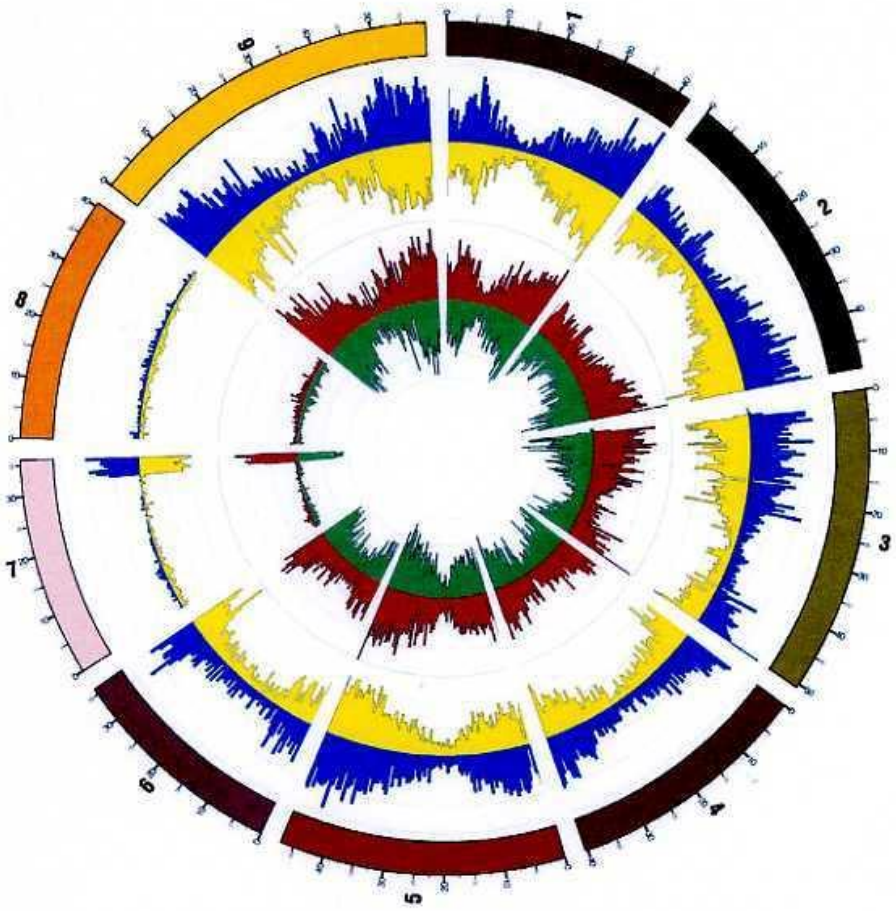


Figure 3. Switchgrass 454-based sequencing reads mapped to the foxtail millet genome. The number of contigs and unassembled ESTs that produced significant alignments to the foxtail millet genome are plotted for each 0.5 megabase interval. Radial axis line represents one log interval. Numbers on the circumference represent the Foxtail millet chromosomes (1–9) and each chromosome is given a different color. Assembled contigs aligned to the forward strand of the *Setaria italica* reference genome assembly are shown in blue for forward strand alignments and in yellow for reverse strand alignments. Singleton reads aligned to the forward strand are shown on the same figure in red, while singleton reads aligned to the reverse strand are shown in green. Diagram was prepared using Circos (http://mkweb.bcgsc.ca/circos).
doi:10.1371/journal.pone.0034225.g003

more contigs than the approach used by dbEST for generating the unigenes.

Ultraconserved orthologs (UCOs) and APVO (*Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera* and *Oryza sativa*) sequences represent a highly conserved set of genes expected to be present in eukaryotic and plant genomes, respectively, and has been used as a proxy for gene detection and sampling breadth [38]. We identified all the 357 (100%) UCOs in assembled switchgrass contigs. We detected 878 (91.5%) of the 959 shared single copy tribes represented in the PlantTribes database [39,40]. [38] Based on these estimations and comparisons we estimate that the set of switchgrass ESTs identified in this study has covered more than 90% of the switchgrass gene space.

Assessment of repetitive sequences in the switchgrass transcriptome

i. Retrotransposon abundance. Given that switchgrass has a polyploid genome we examined the abundance of retrotransposon sequences in the EST collection. It has been reported that retrotransposons constituted nearly 7.5% of the sequenced genome of sorghum [41] and a large number of copia-like and gypsy-like retrotransposons were actively transcribed in sorghum protoplasts derived from embryogenic callus tissues [42]. We retrieved the sorghum sequences for the 24 copia-like and 48 gypsy-like elements to develop a query database. We identified 8826 assembled switchgrass EST contigs (4.06%) and 6990 unassembled EST sequences (4.24%) that showed significant

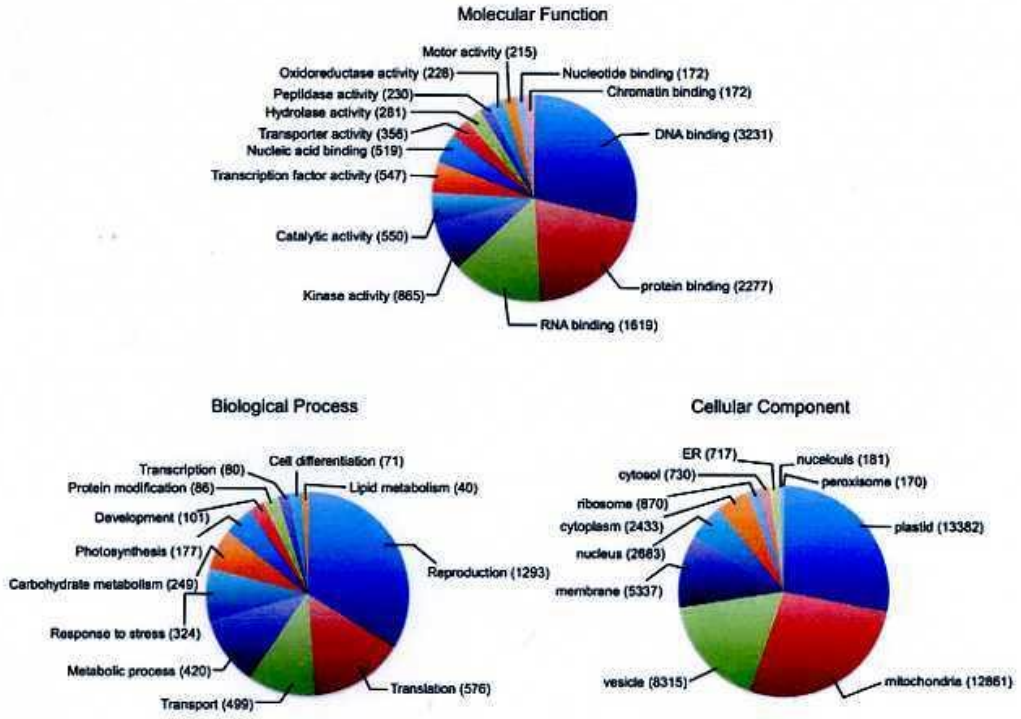


Figure 4. Plant GO-slim terms associated with switchgrass. Venn diagram of the distribution of plant GO-slim terms associated with switchgrass contigs represented in molecular function, biological process and cellular component categories.
doi:10.1371/journal.pone.0034225.g004

Table 2. ESTcalc-based transcriptome coverage estimates.

Input parameters	ESTcalc	Actual
Number of technologies	1	1
Technology	454 GS-FLX	454 GS-FLX (Titanium)
Library type	normalized	normalized
Reads/plate	979,903	979,903
Mean read length (bp)	367	366.56
Predicted assembly		
Total assembled sequence (MB)	359.6	279.9
Unigene count	28665	523228
Mean contig length (bp)	963	535
Mean contig length (longest contig per gene, bp)	1274	-
Singleton yield (%)	13	15
Percent transcriptome (%)	92	-
Percent of genes tagged (%)	100	-
Percent of genes with 90% coverage (%)	80.3	-
Percent of genes with 90% coverage by largest contig (%)	68	-
Percent of genes with 100% coverage (%)	31.7	-
Percent of genes with 100% coverage by largest contig (%)	30.1	-

doi:10.1371/journal.pone.0034225.t002

homologies to the 74 sorghum retrotransposon sequences. Retrotransposon abundance from large EST collections from 10 different plant taxa ranged between 0.03–0.1% [12]. In a recent study in the *Pinus contorta* transcriptome using 454 pyrosequencing, retrotransposons constituted 3.89% of the ESTs [12]. This clearly suggested that as in *P. contorta*, in switchgrass there is a significant over-representation of retrotransposons and the fact that these were identified from RNA samples indicates that they are actively transcribed in the tissues analyzed in our experiments.

ii. Simple Sequence Repeats (SSRs). The assembled switchgrass contigs with annotations were used for identifying SSRs. The distribution of di-, tri-, tetra-, penta- and hexa-nucleotide SSRs in these assembled contigs are shown (Figure 5). This analysis using the 243,600 assembled switchgrass contigs with annotations, identified 21,437 contigs that contained SSRs between 2–6 nucleotides and greater than 15 bp in length using PHOBOS program. This indicated that nearly 8.8% of the switchgrass ESTs contained SSRs and this is 2.8 times more than the average number of EST-SSRs reported in other grasses [43]. The results from the PHOBOS output were filtered in excel to identify only the perfectly matching SSRs. A total of 5840 perfect di-, tri-, tetra-, penta-, and hexa-nucleotide SSRs longer than 8, 6, 4, 3, and 3 repeat units, respectively, were identified (Figure 5). The tri-nucleotide repeats were the most abundant (48.6%), which is consistent with the findings in other grasses including switchgrass [9,43]. The identified SSRs were GC rich (Table 3). The CCG tri-nucleotide repeats were the most abundant (22%) class of repeats that was identified in this study (Figure 5) and has been reported in earlier reports on other grass species [43]. Among di-nucleotide SSRs, AG repeats were the most abundant while the CG repeats were the least abundant. Interestingly, we identified that the frequency of penta-nucleotide repeats (20%) were more abundant than di-nucleotide repeats (11%) among these EST sequences.

Expression profiling

Reads per kilobase per million mapped reads (RPKM) values were obtained for 240,981 assembled contigs. In this analysis 1 RPKM corresponds to ~25 mapped reads per kilobase of target transcript sequence. A total of 44,279 contigs had RPKM values of less than 1 and hence were not considered for further analysis. Of the 196,702 contigs, 120,193 (61%) were from flower library, 102,184 (51.9%) from germinating seedlings library, 86,128 (43.8%) from tiller library and 80,923 (41.1%) from dormant seed library. We chose to examine contigs whose RPKMs were greater than ten in at least one of the four tissues analyzed. This resulted in 136,612 contigs used for clustering analysis. The RPKM values were log 2 transformed and subjected to average linkage hierarchical clustering using the Genesis software [44]. This analysis showed that the transcriptional landscapes of the four tissues that were examined were unique (Figure 6). A majority of transcripts showed maximal expression in the tiller tissue, while in the seedling tissues the expression of these genes was at an intermediate level. In contrast to the tiller tissues, expression levels of a majority of dormant seed ESTs were low. Surprisingly, we identified several clusters of genes that had their highest expression in dormant seeds. In the flowers, a majority of the ESTs showed low to intermediate levels of gene expression when compared with tillers and seedlings.

RT-PCR analysis

We examined the gene expression patterns of 22 genes with annotations derived from homology searches and also a set of 20 genes that did not show any homologies to any of the sequences in the various databases (Figure 7). Primers were designed from the EST sequences to amplify 200–300 bp products for most of the selected genes (except lanes 29 and 39 wherein the EST sequence was about 125 bp). Most of the amplifications resulted in single discrete product and two of them resulted in multiple bands in

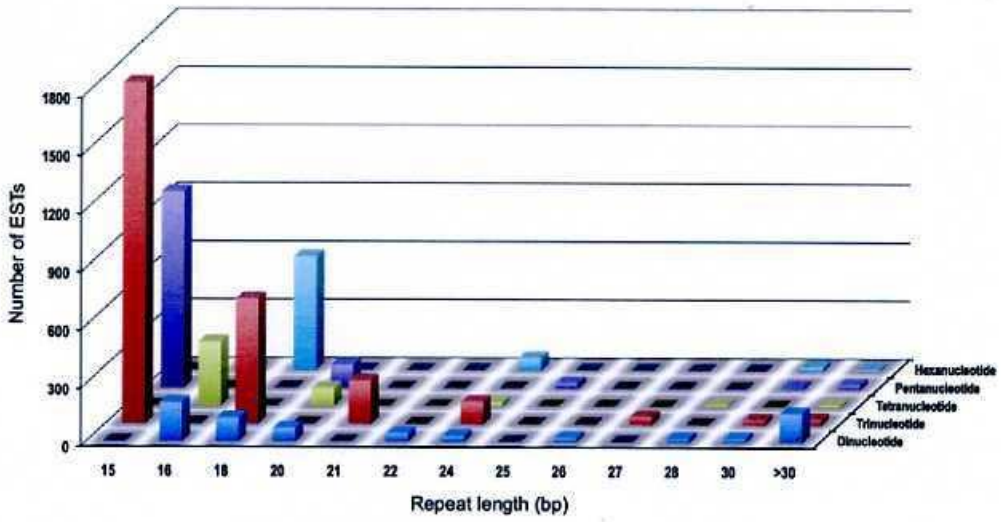


Figure 5. Distribution of simple sequence repeats in switchgrass ESTs. Di-, tri-, tetra-, penta- and hexa-nucleotide repeats were analyzed and their frequency plotted as a function of the repeat number.
doi:10.1371/journal.pone.0034225.g005

Table 3. Commonly found SSR repeat units in switchgrass ESTs.

Repeat unit	Reads	Repeat unit	Reads
AG	337	CCCCG	28
AT	134	ACACG	27
CG	11	ACCGG	26
CCG	1279	AATGC	25
AGC	417	ACCGG	26
AGG	282	AATGC	25
AAG	244	AAGAG	22
ACC	165	AGCTC	19
AAC	142	AGGCG	19
ACG	140	CCCGG	19
ATC	94	AAGGG	18
AAT	44	ACCCG	18
ACT	30	ACAGC	16
ACAT	45	AGATC	16
AAAT	34	ACACC	15
AGGG	34	AATGG	14
CCGG	30	AGCGG	14
CCCG	28	ATCCG	14
AAAG	25	AAAGG	13
AGCT	25	AAATC	13
AGCG	21	ACGGC	13
ACGC	20	AGATG	13
AAAC	19	AAAGC	12
AGGC	19	AACCG	12
ATCC	17	ACTCC	12
AAGC	16	AATCC	11
AGCC	15	ACGCC	10
AGAT	14	ATCCC	10
AATC	11	ATCGC	10
ACCC	11	CCGGCG	47
AAGG	10	ACGGCG	24
ATGC	10	AAGCCG	23
AAAAG	106	AGGCGG	20
CCGCG	65	AGCTCC	17
AAACC	44	AAGAGG	16
AAAAC	43	ACCGCC	16
AAAAT	41	AGAGGG	15
AAATT	35	CCCCCG	15
AGGGG	32	AAGGAG	12
AGCCG	31	AGCGGG	12
AGAGC	30	AGCGGC	11
AGAGG	29	ACGAGG	10

doi:10.1371/journal.pone.0034225.t003

only certain tissues. Even though this was not a quantitative PCR analysis we observed that the gene expression patterns were significantly different for several genes among the four tissues tested here. All of the amplifications from the EST sequences with no homologies gave the expected size amplification product using the switchgrass cDNAs from the four different tissues. This analysis confirmed that these novel EST sequences were indeed expressed in switchgrass tissues.

Gene inventories

We focused our analysis of the switchgrass transcriptome on genes associated with C4 photosynthesis, an attribute that is extremely important for biomass accumulation.

C4 photosynthesis. Based on GO annotations derived from homology searches we have identified all major enzymes associated with C4 photosynthesis (Figure 8). Among the C4 pathway genes in our collection, EST's coding for carbonic anhydrases formed the largest group. Multiple sequence alignments juxtaposed with the GO annotations suggested that there are probably five different genes encoding carbonic anhydrases in switchgrass. This estimate of the number of carbonic anhydrase genes is consistent with an earlier report in switchgrass [9].

Phosphoenolpyruvate carboxylase (PEPC) was the second most abundant enzyme of the C4 pathway based on EST representation. Based on multiple sequence alignment analysis we speculate that there are three PEPC gene families in switchgrass with reference to their putative localization – cytoplasm, mitochondria or plastids. Recently, rice PEPC gene localized to the chloroplasts has been shown to be important for ammonium assimilation [45]. It will be important to examine the PEPC gene expression in switchgrass in the context of its higher nitrogen use efficiency compared with other grasses.

Pyruvate is an important metabolite especially with reference to the CO_2 concentrating mechanisms in C4 plants like switchgrass [46,47]. Recently it was shown that sodium-dependent pyruvate transporters in plastids were encoded by BILE ACID: SODIUM SYMPORTER FAMILY PROTEIN (BASS) [48]. Sixteen ESTs in our collection were identified showing various levels of homologies to BASS protein. Four of them are possibly localized to plastids based on their GO annotations and showed closest homologies to BASS2 and BASS4 of C4 *Flaveria* species identified in the above study. Interestingly, four ESTs were annotated as being localized to mitochondria and eight other ESTs were predicted to be associated with both plastids and mitochondria. Multiple sequence alignment analysis indicated that ESTs with dual localization were unique and did not cluster together with plastidial or mitochondrial pyruvate transporters (Figure S1). Further detailed analysis of expression patterns and precise *in situ* localization of putative mitochondrial and plastidial pyruvate transporters warrants attention.

In all the C4 plants with NAD^+- malic enzyme (NAD^+-ME) this enzyme is localized to the mitochondria. Consistent with these observations we found that all eight ESTs with strong homologies to the NAD^+-ME were annotated as being localized to the mitochondria. Curiously, we identified four ESTs that were annotated as having malic enzyme activity but were localized to cytoplasmic membrane bound vesicles. Multiple sequence alignments indicated that these four ESTs were unrelated to the canonical NAD^+-ME ESTs (Figure S2).

Bundle sheath mitochondria and mesophyll cell cytosol are the major locations for the Aspartate aminotransferases in C4 plants [49]. A majority of Aspartate aminotransferases identified in this study was predicted to be cytosolic while only one was identified as being localized to mitochondria. Alanine aminotransferases that lead to reversible conversion of pyruvate to alanine were abundantly expressed in the tissues analyzed in this study. In our collection, 20 ESTs predicted to represent genes with alanine transaminase activity were identified. Based on their consensus sequences we estimate that this gene family may be represented by

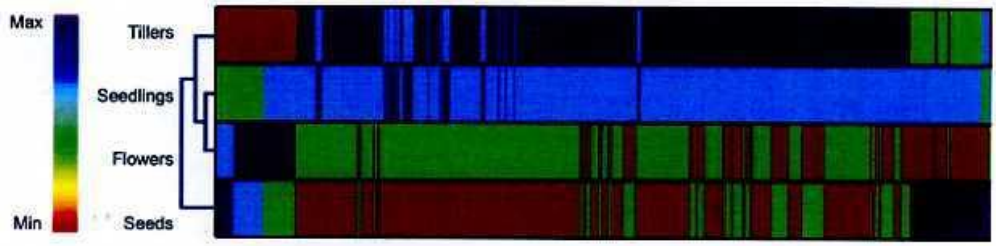


Figure 6. Heat map of switchgrass gene expression in four different tissues. Individual columns represent the four tissues used in this study while each row represents a unique contig. The dendrogram represents the similarity of expression profiles between the tissue samples based on average linkage clustering. The color key represents the log2 transformed RPKM values. Red indicates a low level of expression, green and blue are intermediate levels, while purple indicates maximal levels of gene expression. The highest value for RPKM after log 2 transformation was 17 and the lowest value was 0.1.
doi:10.1371/journal.pone.0034225.g006

5–6 unique genes. More than 20 malate dehydrogenase ESTs were identified in our collection. Only six of these were localized to mitochondrion while six were localized to both mitochondria and chloroplasts, and 10 were localized to the cytoplasm based on their GO annotations. Even the six localized to the mitochondrion were quite different and could represent four unique genes based on the multiple sequence alignment analysis.

Seed dormancy associated genes. Most of the large-scale transcriptome and proteome studies in seed dormancy have been conducted in the model plant *Arabidopsis thaliana*. Previous studies using DNA microarrays indicated that there are about 12,000 stored mRNAs that were detectable in dry seeds of Arabidopsis [50] and barley [51], and 17,000 genes in rice [52]. Using the 454 platform we have identified 24,095 contigs expressed in the dormant seeds of switchgrass [50,51,52]. Genes from all major GO categories were identified in the dormant switchgrass seeds.

We carefully analyzed one study wherein the highly dormant Arabidopsis ecotype Cvi was used for microarray analysis [53]. In a comparison of the dormant Cvi seeds versus ripened seeds, 442 genes were identified as being differentially expressed and up regulated in the former. We compared this list of 442 genes with the genes showing highest expression in the dormant switchgrass seed sample. We identified 170 genes in the switchgrass dormant seed samples that were closely related to the 442 Arabidopsis genes associated with dormancy (Table S1). The more than 38% overlap in the switchgrass and Arabidopsis genes suggest that the genetic mechanisms leading to dormancy are comparable between monocots and dicots. Nearly one fourth of the genes in the Arabidopsis data set are still annotated as unknown proteins or expressed proteins with unknown functions. We speculate that as these annotations are updated, the extent of overlap will improve significantly. It will be interesting to examine which of these stored mRNAs are actually translated and are crucial for maintaining dormancy. Using captured polysome-associated RNAs in combination with high-throughput transcript sequencing will be a useful strategy to explore this issue further.

Of the 53 different transcription factors (TFs) in the Arabidopsis dormancy related gene set [53], homologs to 30 were identified in the switchgrass 454 EST data. This included members of a DREB subfamily, AP2 domain containing TFs, several different classes of

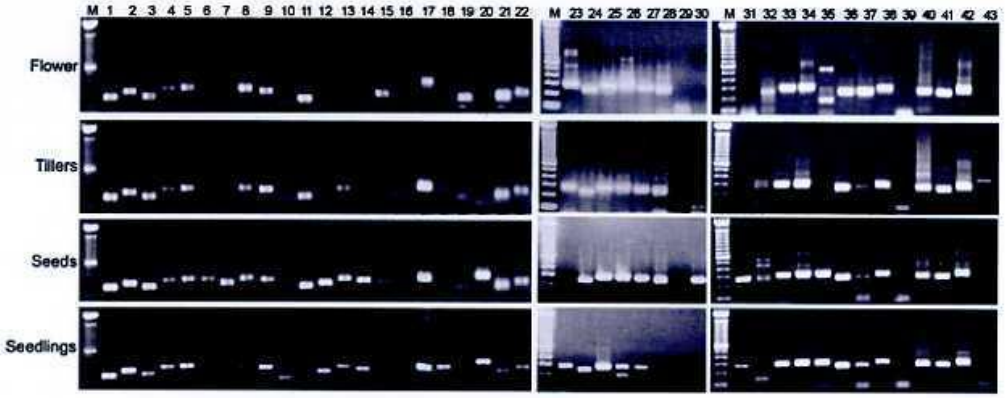


Figure 7. Switchgrass gene expression analysis by RT-PCR. The four panels represents the amplifications from the cDNA derived from the four different tissues – flowers, tillers, seeds and seedlings. The lanes labeled 1–22 are amplifications of EST contigs with functional annotations. Lanes 23–42 contains amplifications of EST contigs that did not show any homologies to sequences in the databases. Lane 43 is the amplification of the teosinte branched 1 gene that shows tiller specific expression. M indicates the 100 bp DNA size ladder.
doi:10.1371/journal.pone.0034225.g007

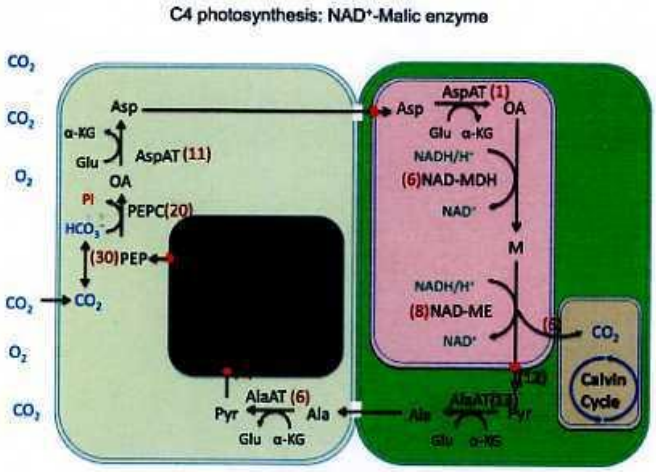


Figure 8. Overview of the C4 NAD⁺-Malic enzyme photosynthesis. The numbers shown in red font in the figure correspond to the ESTs that were identified in the 454 sequencing analysis. Ala-Alanine; AlaAT – Alanine aminotransferase; Alpha-KG-alpha ketoglutarate; Asp-aspartate; AspAT-aspartate aminotransferase; CA-carbonic anhydrase; Glu-glutamate; OA-oxaloacetate; PEP-phosphoenol pyruvate; PEPC-phosphoenolpyruvate carboxylase; PPDK –pyruvate phosphate dikinase; Pyr-pyruvate; NAD⁺-MDH- NAD⁺ dependent malate dehydrogenase; NAD⁺-ME- NAD⁺ dependent malic enzyme.
doi:10.1371/journal.pone.0034225.g008

zinc finger TFs such as C2H2, C3H, C5HC2, NF-YA, ethylene responsive factors such as EIN3, ERF, auxin responsive TFs such as ARF, Myb family TFs, Scarecrow proteins, NAC, NAM, Aintegumenta, bZIPs, and MADS box TFs.

Ten different heat shock proteins including two heat shock transcription factors were also identified in this set. Another set of interesting genes was the RNA binding proteins implicated in post-transcriptional gene regulation and included Mei2, pumilio, RRM domain containing proteins, and the decapping enzyme. Genes encoding various protective proteins such as late embryogenesis abundant proteins, GSTs, Mn SOD, glutathione gamma-glutamylcysteinyltransferase important for GSH biosynthesis, peroxiredoxin, thioredoxin, and PMSR were also identified in the dormant seeds.

Discussion

Expressed sequence tag analysis is one of the most popular techniques for gene discovery. Traditional EST analysis by Sanger sequencing is still very time-consuming, labor and cost intensive. The advent of the next generation sequencing (NGS) technologies has circumvented many of the pitfalls associated with the conventional EST analysis. Apart from the speed and the cost, NGS eliminates the bacterial cloning step that can bias the composition of the cDNA libraries. The Roche GS FLX NGS platform has proven to be valuable for non-model plant systems such as olive [54], chestnut [55], *Artemisia annua* [20], ginseng [8], strawberry [56], bracken fern [38] and recently, in switchgrass [29].

In this study four different tissues were analyzed based on their agronomic importance and/or their under-representation in existing EST collections. Dormancy is one of the major agronomic problems with reference to large-scale production of switchgrass directly from seeds [57]. There are no studies to date that examine the switchgrass genes associated with dormancy. Flower tissue is under-represented in the public switchgrass EST databases. Tillering is an important trait that has a direct bearing on the biomass yield in switchgrass [58]. Furthermore, young tiller tissues are under-represented in existing EST collections.

Assembly quality

The 3.8×10^{8} bp of sequence data here represents a substantial sequence resource and nearly doubles the expressed sequence data available for switchgrass in Genbank (NCBI Genbank dbEST). The increased read lengths (average of 367 bp) from the 454 GS Titanium instrument helped to assemble contigs that were approximately 535 bp that is much larger than the studies that used previous version of the 454 technologies [6,10,59]. *De novo* assembly of the switchgrass ESTs indicated that about 85% of the ESTs could be assembled into contigs while 15% remained as singletons. The higher success rate with the *de novo* assembly may be due to the iterative assembly process used in this study. Using the foxtail millet genome as a reference reduced the unassembled ESTs to approximately 8%. This clearly demonstrates the value of using a closely related species as a surrogate reference in the absence of a whole genome sequence for switchgrass.

Transcriptome coverage

Estimating the number of genes and the extent of gene coverage is an important metric for transcriptome sequencing projects. In the absence of a genome assembly it is only possible to make an approximation of the extent of coverage. Reciprocal BLAST analysis between switchgrass ESTs from this study and four other monocots (Brachpodium, rice, sorghum and maize genomes) indicate that homologs for 70–80% of the genes in those species are represented in this collection. Even though the sorghum genome is about 75% larger than the rice genome, it has been reported that these two grasses have similar quantities of euchromatin (252 Mb and 309 Mb, respectively) [41]. On the

same grounds, we speculate that a significant proportion of the switchgrass gene space has been covered by this current study and based on the ESTcalc estimations this may be as high as 90%.

In a previous EST analysis in switchgrass using the conventional Sanger sequencing technology, sorghum was used as the reference genome. Switchgrass ESTs were evenly distributed across the sorghum pseudomolecules [9]. In this current study using the foxtail millet genome as reference we observe that representation of switchgrass ESTs on the pseudomolecules 7 and 8 of the *Setaria italica* genome is conspicuously low. We speculate this may be due to vast stretches of repetitive sequences on these chromosomes that may have been masked during the assembly process and in turn led to this skewed distribution. It is also possible that chromosomes 7 and 8 of the *S. italica* genome may have diverged significantly. A complete analysis of the *S. italica* genome sequence will provide more insight in this regard.

Marker development

EST-based SSR markers are advantageous when compared with genomic SSRs owing to their higher PCR amplification rates and cross-species transferability [60]. In a previous EST analysis study in switchgrass 830 SSRs were successfully amplified and about 38% of these were reported to be polymorphic between the parents of a mapping population [9]. We conducted a BLAST analysis with our EST data sets and the primer sequences reported from the above study. We found that less than 5% of the primer sequences showed perfect matches to the sequences in our collection (Wu and Mahalingam, unpublished data). It is possible that the use of different cultivars in the two studies may be one of the reasons for this low degree of overlap. Our ongoing work is assessing amplification efficiency and polymorphism rates for more than 1000 EST-SSRs using the parents of a mapping population. The frequency of SSRs identified in Next-Gen sequencing projects depends on template used, criteria for defining SSRs, and the software used for identifying SSRs in the sequences, and has shown extensive variation in recently reported studies [12,61,62]. Despite these variations, the above-mentioned work and the current study illustrate the speed and cost effectiveness of identifying SSRs in non-model systems using the 454 technology.

In summary a highly significant improvement in the switchgrass EST assembly was facilitated by the availability of the foxtail millet draft genome. The 180,000 unique sequences (98,000 assembled contigs and 82,000 unassembled ESTs) identified in this 454-based EST collection represent a major genomic resource for switchgrass. We estimate that more than 90% of the gene space of switchgrass is represented in this analysis. Identification of more than 24,000 unique sequences in the dormant seeds of switchgrass was unexpected and provides an important resource to further investigate this important agronomic trait. The large number of EST-SSR markers identified in this study will provide valuable resources for marker-assisted breeding programs. Sequencing the transcriptomes and genomes of closely related members of the NAD^+-malic enzyme- type C4 grasses such as the model system *Setaria viridis* is extremely important and will be a viable proxy for the switchgrass genome [63].

Materials and Methods

Sample collection

'Cimarron', a high biomass yielding switchgrass cultivar released by Oklahoma State University was used for this analysis. Seeds were first placed on a Whatman filter paper pre-soaked in 50% Benomyl solution, a fungicide. Seeds were surface sterilized in Falcon tubes containing 30 ml isopropanol. Then seeds were placed in 50% bleach solution for 5–10 mins. Seeds were washed in distilled water for 4–5 times to remove the bleach. Sterilized seeds were placed on top of a Whatman paper pre-soaked with 50% Benomyl in a petri dish. Petri dishes were placed in a growth chamber with 10 hours/day and 14 hours/night for six days. Seeds that failed to germinate were harvested separately and labeled as dormant seeds while those that had sprouted were harvested as germinating seedlings. Samples were frozen in liquid nitrogen and stored at −80°C. Seeds of SL93-7×15 were planted in cones containing SUNGRO metromix 200 series soil in the Plant and Soil Sciences greenhouse facility at Oklahoma State University. Plants were maintained at 24°C with 16 h day and 8 h night regime. Six weeks following germination, the young emerging tillers were harvested, frozen in liquid nitrogen and stored in −80°C. Flowers were harvested from field-grown SL93-7×15 plants in the Oklahoma State University Agricultural station agronomy plots, Stillwater, OK.

RNA isolations

About 100 mg of frozen switchgrass seed tissue was ground to fine powder with liquid nitrogen and sterile quartz powder. The powder was transferred to a 2-ml tube with 2 ml of extraction buffer (8 M LiCl, 2% b-mercaptoethanol, pre-cooled to −20°C) and was incubated overnight at 4°C. The mixture was centrifuged at 13,000 rpm for 30 min at 4°C. The resulting supernatant was decanted, and the pellet was washed with cold (4°C) 70% ethanol, briefly air-dried, and dissolved in 1 ml of solubilization buffer (0.5% SDS, 100 mM NaCl, 25 mM EDTA, 10 mM Tris-HCl, pH 7.6, 2% 2-mercaptoethanol). Following this, RNA was further purified, once with phenol: chloroform: isoamyl alcohol (25:24:1), and twice with chloroform: isoamyl alcohol (24:1). The aqueous phase was precipitated with 0.1 volumes of 3 M sodium acetate and 1.5 volumes of ethanol.

The tubes were then centrifuged at 13,000 rpm for 30 min at 4°C, the supernatant was poured off, and 0.5 ml of 3 M sodium acetate was added. The pellet was vortexed for 1 min and centrifuged at 13,000 rpm for 10 min at 4°C, washed with 70% ethanol, and dissolved in 50 ul of RNase free water. Total RNA from switchgrass flowers and tillers was extracted using the RNEasy Plant Mini kit (Qiagen). The RiboMinus™ Plant kit for RNA-Seq (Invitrogen) was used to remove rRNA from all the four samples. All steps were performed according to the manufacturer's instructions. The RNA integrity was assessed by agarose gel electrophoresis.

cDNA library construction and 454 sequencing

Approximately 200 ng of Ribominus RNA was used for first and second strand cDNA synthesis as described in the cDNA Rapid Library Preparation Method manual (Roche Life Sciences, Inc.) with slight modification. The double stranded (ds) cDNA was nebulized using 30 psi nitrogen for 30 seconds to generate library fragments of the correct size followed by purification in a Qiagen MinElute column. The ds cDNA was eluted in 16 ul Tris-HCl, pH 7.5. The nebulized ds cDNA was used for fragment end repair followed by adaptor ligation. AMPure beads were used for removing the small fragments. Library quantitation was done using TBS fluorimeter and Roche Rapid Libary (RL) standards that are control ds DNA fragments with an attached FAM moiety. The units of the RL standards were "molecules/ul". The A adaptor on all ds cDNA fragments contained a FAM moiety as well. The RFU value of the library (x) was applied to the RFU values of the standard curve, $Y = mx+b$, solved for y to yield molecules/ul. Small volume emulsion (em) PCR was set up for each ds cDNA library based on each library's optimal "library

molecules/DNA capture bead" that was calculated as described in the emPCR Method Manual- Lib-L SV (Roche Life Sciences, Inc.). After the emPCR, all reactions were pooled followed by capture of the "enriched beads". Lastly, 790,000 of the enriched beads from each library were loaded into a four-region 454 Life Sciences Picotiter plate and sequenced with "454 Life Sciences FLX Titanium Chemistry".

454 EST filtering and assembly

Roche/454 EST sequences were prepared for assembly by removal of library adapter sequences using estclean (https://sourceforge.net/projects/estclean/) and a custom perl script. Contaminating vector and microbial sequences and poly A/T stretches were removed using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/). Public switchgrass ESTs (546,245) were downloaded from Genbank and filtered for adapters, contaminating vector sequences and poly A/T stretches using estclean and SeqClean. ESTs were pre-clustered using a custom BLAT-based pipeline. Briefly, an all-versus-all comparison was performed using BLAT [31]. A custom perl script was used to generate clusters of overlapping ESTs from the resulting BLAT alignments, allowing for individual ESTs to exist in more than one cluster in order to accommodate polyploidy in switchgrass and avoid mis-assembly of close homologs and paralogs and splice variants. The resulting EST clusters were assembled using iterative cycles of MIRA (http://www.chevreux.org/projects_mira.html) [64] and CAP3 (http://seq.cs.iastate.edu/cap3.html) [65]; four cycles of MIRA assembly were performed and then followed by one cycle of CAP3 assembly performed on the contigs generated by MIRA.

Descriptive annotations and GO classifications

Descriptive annotations and GO classifications were derived by comparing the assembled switchgrass EST contigs to public sequence databases using NCBI BLAST [66]. Databases used for sequence comparisons were as follows: Brachypodium distachyon v1.2 (downloaded from MIPS; ftp://ftpmips.helmholtz-muenchen.de/plants/brachypodium/v1.2/); Sorghum bicolor v1.4 (downloaded from Phytozome; ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v6.0/Sbicolor/); Zea mays ZmB73_4a.53 (downloaded from Gramene; http://www.gramene.org/info/data/ftp/index.html); Oryza sativa v6.1 (downloaded from MSU; ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/); Plant Gene Indexes (http://compbio.dfci.harvard.edu/tgi/plant.html); Genbank nr protein database (ftp://ftp.ncbi.nih.gov/blast/db/); and Genbank est_others database (ftp://ftp.ncbi.nih.gov/blast/db/). EST contigs were compared to the nucleotide sequence databases listed above using BLASTN with default settings. Unassembled ESTs were compared to the Plant Gene Indexes (blastn; http://compbio.dfci.harvard.edu/tgi/plant.html) using BLASTN with default settings. Putative orphan transcript contigs that had no matches to plant sequences were further compared to the Genbank nr protein database using BLASTX and Genbank est_others database using BLASTN. In all cases, a perl script was used to filter for the single best database match (according to the BLAST bit score) for each query EST or EST contig sequence, with no additional filtering. Gene ontology (GO) classifications for biological process, molecular function and cellular localization were derived from the Plant Gene Index BLAST matches, providing putative GO annotations for only those switchgrass EST contigs matching Plant Gene Index sequences that have GO annotations.

Reference guided assembly

Switchgrass EST contigs and singletons were aligned to the *Setaria italica* draft genome (v1.64; ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Sitalica/) using BLAT and best matches were defined as the alignment for each EST contig or singleton with the highest BLAST score. A custom perl script was used to generate clusters of overlapping ESTs from the resulting alignments. The resulting EST clusters were assembled using iterative cycles of MIRA (http://www.chevreux.org/projects_mira.html) [64] and CAP3 (http://seq.cs.iastate.edu/cap3.html) [65]; four cycles of MIRA assembly were performed and then followed by one cycle of CAP3 assembly performed on the contigs generated by MIRA.

Circos analysis

Assembled contigs and singleton reads were aligned to the Setaria reference genome assembly using BLAT. Contig alignments were mapped to the outermost set of axes, minimum average coverage −0.1, and maximum average coverage 0.1. EST singleton alignments were mapped to the innermost set of axes, minimum average coverage −0.045, and maximum average coverage 0.045. In all cases the per-base alignment depth was averaged over 500,000 base pairs.

EST expression level estimates

Relative expression levels of 454 ESTs from the four libraries were quantified by a reads per kilobase transcript per million reads (RPKM) analysis. A Perl script was used to shred the Roche/454 ESTs into simulated 40mer Illumina-like RNA-seq "reads". The 40-mer reads were mapped to the switchgrass EST contigs using SOAPaligner (http://soap.genomics.org.cn/soapaligner.html) [67], and match counts were converted to RPKM values using a perl script. ESTs with RPKM values greater than 10 in at least one of the four tissues were used for constructing heatmaps using GENESIS [44]. RPKM values were log2 transformed. Average linkage clustering was selected.

Assessment of transcriptome coverage

Three different assessment tools were used to estimate transcriptome coverage. A web based tool called as ESTcalc [13] was used to estimate the predicted transcriptome coverage. Input parameters for the ESTcalc were one for the 454 GSFLX sequencing technology used, 979,903 for the number of reads and 367 bp for the read length. To determine the number of eukaryotic ultra conserved orthologs (UCOs) in the switchgrass 454 transcriptome dataset we used tblastx to query the list of 357 UCO protein sequences from Arabidopsis (sequences available at http://compgenomics.ucdavis.edu/compositae_reference.php) with an e-value threshold of 1e-10. Blast results were parsed to determine the number of switchgrass ESTs that showed a positive hit to the UCO sequences with amino acid alignments of at least 30 residues. We assessed the transcriptome coverage by comparing the switchgrass ESTs with the PlantTribes database [39,40]. In this analysis 959 shared single copy tribes from *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera* and *Oryza sativa* were compared with the switchgrass EST reads using tblastx and an e-value cutoff of 1e-06.

Estimating retrotransposon abundance

We used BLASTn to search the 454 EST contigs and singletons for 24 families of *copia*-like elements (GenBank accession numbers: CG026188–CG026196 and CF417056–CF417070) and 48 families of *gypsy*-like elements (GenBank accession numbers:

CG425584–CG425599 and CF542191–CF542222) that were reported in the sorghum genome [42]. An *E* value threshold of 1e-06 was used for this analysis. Custom perl scripts were used to make these comparisons and to parse and filter the ESTs that contained the retrotransposon sequences.

RT-PCR analysis

Primers were designed for 57 switchgrass EST contigs (Table S2).

Total RNA (5 ug) was used to make cDNA using Super-Script III Reverse Transcriptase (Invitrogen). PCR amplification was performed using Taq Master Mix Kit (Qiagen). Following initial denaturation at 95°C for 3 min, the PCR reaction was carried out in a PTC-200 thermal cycler (MJ Research) under the following conditions: denaturation at 94°C for 30 s, annealing at 60°C for 30 s, and extension at 72°C for 1 min. The final extension was carried out at 72°C for 10 min. Reaction products and DNA size markers (100 bp DNA ladder, Invitrogen) were resolved on the 1% agarose gels and visualized under UV light following ethidium bromide staining.

SSR analysis using PHOBOS

The GUI-based PHOBOS software (Phobos 3.3.11, 2006–2010, http://www.rub.de/spezzoo/cm/cm_phobos.htm) was used for SSR analysis. The minimum repeats unit length was set to two and the maximum repeat unit length was set to six. The minimum length of SSR was set to 15 and only sequences with perfect matches were selected.

Data availability

The 454 EST data obtained in this study are available in the NCBI Sequence Read Archive under the accession SRA050067.

Supporting Information

Figure S1 Multiple sequence alignment of switchgrass pyruvate transporters. Switchgrass pyruvate transporters localized to mitochondria and plastids versus pyruvate transporters localized to plastids only (A) or to motochondria only (B). MPpyrT refers to pyruvate transporters localized to mitochondria and plastids. PpyrT refers to pyruvate transporters localized to plastids. MpyrT refers to transporters localized to mitochondria. Multiple sequence alignments were conducted using the MAFFT version 6 (http://mafft.cbrc.jp/alignment/server/index.html). Flavaria BASS2 and BASS4 sequences were included in this analysis.
(TIF)

Figure S2 Multiple sequence alignment of switchgrass ESTs encoding NAD-malic enzyme. Thirty ESTs annotated as NAD malic enzyme was used for this analysis using the MAFFT version 6 (http://mafft.cbrc.jp/alignment/server/index.html). exCS refers to extracellular space. Thlk meb refers to thylakoid membrane. Plst refers to plastids. Mito refers to mitochondria.
(TIF)

Table S1 Dormancy related genes in switchgrass and Arabidopsis. Switchgrass ESTs from dormant seed libraries compared with Arabidopsis dormancy related genes [53].
(XLS)

Table S2 Primer sequences of switchgrass ESTs selected for RT-PCR analysis.
(DOC)

Acknowledgments

We thank Dr. Yanqi Wu for providing the seed materials required for this study. We greatly appreciate the technical support of MOGene LC for the 454 sequencing. The foxtail millet genome sequence data were produced by the US Department of Energy Joint Genome Institute.

Author Contributions

Conceived and designed the experiments: RM NJI YW XZ. Performed the experiments: YW XZ. Analyzed the data: RM NJI TCM DWB. Contributed reagents/materials/analysis tools: RM TCM DWB. Wrote the paper: RM NJI YW TCM.

References

1. Andersen JR, Lubberstedt T (2003) Functional markers in plants. Trends in Plant Science 8: 554–560.
2. Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Research 17: 69–73.
3. Kaur S, Cogan NO, Pembleton LW, Shinozuka M, Savin KW, et al. (2011) Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genomics 12: 265.
4. Barbazuk WB, Emrich S, Schnable PS (2007) SNP Mining from Maize 454 EST Sequences. CSH Protoc 2007: pdb prot4786.
5. Mahalingam R, Gomez-Buitrago A, Eckardt N, Shah N, Guevara-Garcia A, et al. (2003) Characterizing the stress/defense transcriptome of Arabidopsis. Genome Biology 4: R20.
6. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, et al. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9: 312.
7. Puckette M, Peal L, Steele J, Tang Y, Mahalingam R (2009) Ozone responsive genes in Medicago truncatula: Analysis by suppression subtraction hybridization. J Plant Physiol 166: 1284–1295.
8. Sun C, Li Y, Wu Q, Luo H, Sun Y, et al. (2010) De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11: 262.
9. Tobias CM, Sarath G, Twigg P, Lindquist E, Pangilinan J, Penning PW, Barry K, McCann MC, Carpita NC, Lazo GR (2008) Comparative genomics in switchgrass using 61,585 high-quality expressed sequence tags. The Plant Genome 1: 111–124.
10. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology 17: 1636–1647.
11. Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. Molecular Ecology 17: 3599–3613.
12. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11: 180.
13. Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, et al. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics 10: 347.
14. Chi KR (2008) The year of sequencing. Nat Methods 5: 11–14.
15. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends in Genetics 24: 133–141.
16. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
17. Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92: 255–264.
18. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nature Methods 5: 16–18.
19. Wang L, Li P, Brutnell TP (2010) Exploring plant transcriptomes using ultra high-throughput sequencing. Briefings in Functional Genomics 9: 118–128.
20. Wang W, Wang Y, Zhang Q, Qi Y, Guo D (2009) Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing. BMC Genomics 10: 465.
21. Metzker ML (2010) Sequencing technologies - the next generation. Nature Reviews Genetics 11: 31–46.
22. Bouton JH (2007) Molecular breeding of switchgrass for use as a biofuel crop. Current Opinion in Genetics & Development 17: 553–558.
23. Keshwani DR, Cheng JJ (2009) Switchgrass for bioethanol and other value-added applications: a review. Bioresource Technology 100: 1515–1523.

24. Schmer MR, Vogel KP, Mitchell RB, Perrin RK (2008) Net energy of cellulosic ethanol from switchgrass. Proceedings of the National Academy of Sciences of the United States of America 105: 464–469.
25. Okada M, Lanzatella C, Saha MC, Bouton J, Wu R, et al. (2010) Complete switchgrass genetic maps reveal subgenome collinearity, preferential pairing and multilocus interactions. Genetics 185: 745–760.
26. Tobias CM, Hayden DM, Twigg P, Sarath G (2006) Genic microsatellite markers derived from EST sequences of switchgrass (Panicum virgatum L.). Molecular Ecology Notes 6: 185–187.
27. Wang YW, Samuels TD, Wu YQ (2011) Development of 1,030 genomic SSR markers in switchgrass. Theoretical and Applied Genetics 122: 677–686.
28. Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, et al. (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American Journal of Botany 99: 193–208.
29. Palmer NA, Saathoff AJ, Kim J, Benson A, Tobias CM, et al. (2011) Next-generation sequencing of crown and rhizome transcriptome from an upland, tetraploid switchgrass. Bioenergy Research DOI 10.1007/s12155-011-9171-1.
30. Matts J, Jagadeeswaran G, Roe BA, Sunkar R (2010) Identification of microRNAs and their targets in switchgrass, a model biofuel plant species. Journal of Plant Physiology 167: 896–904.
31. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Research 12: 656–664.
32. Hopkins AA, Taliaferro CM, Murphy CD, DAnn C (1996) Chromosome numbers and nuclear DNA content of several switchgrass populations. Crop Science 36: 1192–1195.
33. Hale MC, McCormick CR, Jackson JR, Dewoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery. BMC Genomics 10: 203.
34. Christin PA, Salamin N, Kellogg EA, Vicentini A, Besnard G (2009) Integrating phylogeny into studies of C4 variation in the grasses. Plant Physiology 149: 82–87.
35. Group GPW (2001) Phylogeny and subfamilial classification of the Poaceae. Annals of Missouri Botanical Gardens 88: 373–457.
36. Vicentini A, Barber JC, Aliscioni SA, Giussani LM, Kellogg EA (2008) The age of the grasses and clusters of origins of C4 photosynthesis. Global Change Biology 14: 1–15.
37. Bevan M, Bancroft I, Bent E, Love K, Goodman H, et al. (1998) Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana. Nature 391: 485–488.
38. Der JP, Barker MS, Wickett NJ, dePamphilis CW, Wolf PG (2011) De novo characterization of the gametophyte transcriptome in bracken fern, Pteridium aquilinum. BMC Genomics 12: 99.
39. Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, et al. (2010) Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evolutionary Biology 10: 61.
40. Wall PK, Leebens-Mack J, Muller KF, Field D, Altman NS, et al. (2008) PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Research 36: D970–976.
41. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457: 551–556.
42. Muthukumar B, Bennetzen JL (2004) Isolation and characterization of genomic and transcribed retrotransposon sequences from sorghum. Molecular Genetics and Genomics 271: 308–316.
43. Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Molecular Biology 48: 501–510.
44. Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. Bioinformatics 18: 207–208.
45. Masumoto C, Miyazawa S, Ohkawa H, Fukuda T, Taniguchi Y, et al. (2010) Phosphoenolpyruvate carboxylase intrinsically located in the chloroplast of rice plays a crucial role in ammonium assimilation. Proceedings of the National Academy of Sciences of the United States of America 107: 5226–5231.
46. Schulze-Siebert D, Heineke D, Scharf H, Schultz G (1984) Pyruvate-derived amino acids in spinach chloroplasts : Synthesis and regulation during photosynthetic carbon metabolism. Plant Physiology 76: 465–471.
47. Schwender J, Ohlrogge J, Shachar-Hill Y (2004) Understanding flux in plant metabolic networks. Current Opinion in Plant Biology 7: 309–317.
48. Furumoto T, Yamaguchi T, Ohshima-Ichie Y, Nakamura M, Tsuchida-Iwata Y, et al. (2011) A plastidial sodium-dependent pyruvate transporter. Nature 476: 472–475.
49. Taniguchi M, Kobe A, Kato M, Sugiyama T (1995) Aspartate aminotransferase isozymes in Panicum miliaceum L., an NAD-malic enzyme-type C4 plant: comparison of enzymatic properties primary structures, and expression patterns. Archives of Biochemistry and Biophysics 318: 295–306.
50. Nakabayashi K, Okamoto M, Koshiba T, Kamiya Y, Nambara E (2005) Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. Plant Journal 41: 697–709.
51. Sreenivasulu N, Usadel B, Winter A, Radchuk V, Scholz U, et al. (2008) Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. Plant Physiology 146: 1738–1758.
52. Howell KA, Narsai R, Carroll A, Ivanova A, Lohse M, et al. (2009) Mapping metabolic and transcript temporal switches during germination in rice highlights specific transcription factors and the role of RNA instability in the germination process. Plant Physiology 149: 961–980.
53. Cadman CS, Toorop PE, Hilhorst HW, Finch-Savage WE (2006) Gene expression profiles of Arabidopsis Cvi seeds during dormancy cycling indicate a common underlying dormancy control mechanism. Plant Journal 46: 805–822.
54. Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, et al. (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics 10: 399.
55. Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, et al. (2009) Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection. BMC Plant Biology 9: 51.
56. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, et al. (2011) The genome of woodland strawberry (Fragaria vesca). Nature Genetics 43: 109–116.
57. Bouton JH (2008) Improvement of switchgrass as bioenergy crop. In: Vermerris W, ed. Genetic improvement of bioenergy crops. New York: Springer. pp 295–308.
58. Boe A, Beck DL (2008) Yield components of biomass in switchgrass. Crop Science 48: 1306–1311.
59. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10: 219.
60. Barbara T, Palma-Silva C, Paggi GM, Bered F, Fay MF, et al. (2007) Cross-species transfer of nuclear microsatellite markers: potential and limitations. Molecular Ecology 16: 3759–3767.
61. Castoe TA, Poole AW, Gu W, Jason de Koning AP, Daza JM, et al. (2010) Rapid identification of thousands of copperhead snake (Agkistrodon contortrix) microsatellite loci from modest amounts of 454 shotgun genome sequence. Mol Ecol Resour 10: 341–347.
62. Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, et al. (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (Vigna radiata (L.) Wilczek). BMC Plant Biology 9: 137.
63. Brutnell TP, Wang L, Swartwood K, Goldschmidt A, Jackson D, et al. (2010) Setaria viridis: a model for C4 photosynthesis. Plant Cell 22: 2537–2544.
64. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Research 14: 1147–1159.
65. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Research 9: 868–877.
66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.
67. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.

VITA

Xin Zeng

Candidate for the Degree of

Doctor of Philosophy

Thesis: REGULATION OF *ARABIDOPSIS* NUDIX HYDROLASE 7 (ATNUDT7) AND ITS ROLE IN SEED DORMANCY & EXPLORING THE SWITCHGRASS TRANSCRIPTOME USING SECOND-GENERATION SEQUENCING TECHNOLOGY

Major Field: Biochemistry and Molecular Biology

Biographical:

Born in Chengdu, Sichuan Province, China at 10/23/1984, son of Birong Zeng and Lixia Li.

Education:

Received Bachelors of Science degree in Biotechnology from College of Life Sciences at Sichuan University in Chengdu, China in July 2007.
Completed the requirements for the Doctor of Philosophy in Biochemistry and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma in August, 2012.

Experience:

Graduate Research Assistant	January 2008—present
Oklahoma State University	Stillwater, OK
Graduate Mentor for Undergraduates	September 2009—September 2011
Oklahoma State University	Stillwater, OK
Undergraduate Research Assistant	September 2005—July 2007
Sichuan University	Chengdu, China
Intern	July 2003—September 2006
Jin Tian Agriculture Technology Cooperation	Chengdu, China

Name: Xin Zeng Date of Degree: December, 2012

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: REGULATION OF *ARABIDOPSIS* NUDIX HYDROLASE 7 (ATNUDT7) AND ITS ROLE IN SEED DORMANCY & EXPLORING THE SWITCHGRASS TRANSCRIPTOME USING SECOND-GENERATION SEQUENCING TECHNOLOGY

Pages in Study: 65 Candidate for the Degree of Doctor of Philosophy

Major Field: Biochemistry and Molecular Biology

Scope and Method of Study:

In this study the role of Arabidopsis nudix hydrolase 7 in seed after-ripening was analyzed. Techniques include cloning, transgenics, seed physiology assays, enzyme cycling assays, fluorometry, spectophotometry, mass spectrometry, GUS staining, microscopy, real-time PCR, western analysis, recombinant proteins, and electrophoretic mobility shift assay.
In another study 454 sequencing was used to examine switchgrass transcriptome. Techniques include RNA isolations, affinity depletion, cDNA library preparation, 454 pyrosequencing, de novo and reference genome based sequence assembly, comparative genomics, reverse transcription PCR, in silico prediction of EST-SSRs, proxy methods for estimating transcriptome coverage.

Findings and Conclusions:

Loss of *Arabidopsis* nudix hydrolase 7 (*AtNudt7-1*) reduces seed germination potential. AtNUDT7 protein levels are required in dry seeds and during early hours following imbibition to realize maximum seed germination potential. Loss of AtNUDT7 impairs seed after-ripening mainly through pertubations in NAD:NADH balance, that in turn affect phytohormones and oxidative signaling pathways. Manipulating levels of NUDT7 will provide a novel avenue for overcoming seed dormancy and maximizing germination potential of seeds.

Approximtely 980,000 EST sequences averaging 367 bp were generated from dormant seeds, seedlings, tillers and flowers of switchgrass using 454 based pyrosequencing. Using foxtail millet genome as a reference, greatly improved the assembly and annotation of switchgrass ESTs. It is estimated that nearly 90% of the gene space of switchgrass was identified based on ESTcalc, coverage of ultraconserved orthologs and planttribes, and genes in C4 photosynthesis pathway. This study more than doubles the current publicly available switchgrass ESTs. Next generation sequencing technologies provides a viable alternative to whole genome sequencing of polyploids like switchgrass.

ADVISER'S APPROVAL: Dr. Ramamurthy Mahalingam