UNLV | University Libraries
University of Nevada, Las Vegas

UNLV Theses, Dissertations, Professional Papers, and Capstones

May 2016

A Bioinformatics Approach to Addressing the Responses of Rice Aleurone Cells to Hormones Abscisic Acid and Gibberellic Acid

Kenneth Arthur Watanabe
University of Nevada, Las Vegas, watana43@unlv.nevada.edu

Follow this and additional works at: https://digitalscholarship.unlv.edu/thesesdissertations

Part of the Bioinformatics Commons, Biology Commons, and the Plant Sciences Commons

Repository Citation
Watanabe, Kenneth Arthur, "A Bioinformatics Approach to Addressing the Responses of Rice Aleurone Cells to Hormones Abscisic Acid and Gibberellic Acid" (2016). *UNLV Theses, Dissertations, Professional Papers, and Capstones*. 2759.
https://digitalscholarship.unlv.edu/thesesdissertations/2759

This Dissertation is protected by copyright and/or related rights. It has been brought to you by Digital Scholarship@UNLV with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself.

This Dissertation has been accepted for inclusion in UNLV Theses, Dissertations, Professional Papers, and Capstones by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.

A BIOINFORMATICS APPROACH TO ADDRESSING THE RESPONSES OF RICE ALEURONE CELLS TO HORMONES ABSCISIC ACID AND GIBBERELLIC ACID

By

Kenneth A. Watanabe

Bachelor of Science - Chemical Engineering
Cornell University
1988

Masters of Engineering - Computer Science
Cornell University
1989

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Biological Sciences

School of Life Sciences
College of Sciences
The Graduate College

University of Nevada, Las Vegas
May 2016

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

April 14, 2016

This dissertation prepared by

Kenneth A. Watanabe

entitled

A Bioinformatics Approach to Addressing the Responses of Rice Aleurone Cells to Hormones Abscisic Acid and Gibberellic Acid

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Biological Sciences
School of Life Sciences

Jeffrey Q. Shen, Ph.D.
Examination Committee Chair

Laurel Raftery, Ph.D.
Examination Committee Member

Martin Schiller, Ph.D.
Examination Committee Member

Andrew Andres, Ph.D.
Examination Committee Member

Amei Amei, Ph.D.
Graduate College Faculty Representative

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

ABSTRACT

A BIOINFORMATICS APPROACH TO ADDRESSING THE RESPONSES OF RICE ALEURONE CELLS TO HORMONES ABSCISIC ACID AND GIBBERELLIC ACID

By

Kenneth A. Watanabe

Dr. Jeffery Q. Shen, Examination Committee Chair

Professor of the School of Life Sciences

University of Nevada Las Vegas

The hormone abscisic acid (ABA) is biosynthesized by higher plants in response to various abiotic stresses such as drought and antagonizes the growth and germination-promoting hormone gibberellic acid (GA). The seed is a model system for studying desiccation tolerance and germination. The thin layer of cells surrounding the seed, the aleurone layer, plays a direct role in seed germination and an indirect role in desiccation tolerance. The goal of my research is to address the molecular mechanism underlying the responses of rice aleurone cells to ABA and GA, by taking a genomics approach. An accurate and complete annotation of the rice genome would greatly expand the results for such an approach. Without a complete annotation, key genes involved in the hormone signaling pathways may be missing. Since the sequencing of the rice (*Oryza sativa*) genome in 2002, over 57,000 putative and confirmed genes have been annotated; yet annotation of the rice genome is far from complete. Invention of the RNA-sequencing (RNA-seq) technologies provided a new highthroughput approach to genome annotation. To analyze RNA-seq data derived from rice aleurone cells treated with and without ABA, GA and a combination of these two hormones, I developed a software package, called Clustering Algorithm (CA). In combination with the popular transcript assembly software Cufflinks, I identified

hundreds of potential novel genes in rice. Thorough filters were applied to minimize the number of false positives resulting in 553 high confidence novel genes. A subset of these novel genes were experimentally validated via qRT-PCR, and analysis of these genes indicated that these genes encode for proteins and/or microRNAs.

The CA software had some limitations: it requires a partially annotated reference genome, and it cannot identify exon/intron boundaries of the genes. To overcome these problems, I developed a novel algorithm dubbed the Tiling Assembly (TA). TA is reference annotation independent, and accurately identified exon/intron boundaries compared to popular transcript assembly software Cufflinks. Analyses of RNA-seq data with TA and Cufflinks, followed by application of the same thorough filters aforementioned, led to identification of 767 high confidence novel genes, far surpassing the previous number of novel genes previously identified. TA was applied to other organisms as well and was able to identify hundreds of high confidence novel genes in Drosophila, yeast, *C. elegans* and Arabidopsis. Therefore, for many organisms, TA is an invaluable tool for genome annotation based on RNA-seq data.

Defining the transcriptomes of rice aleurone cells treated with ABA, GA and both hormones helped address the crosstalk of ABA and GA signaling. There were 2,443 gene upregulated by ABA, 5,138 genes upregulated by GA and 4,273 genes upregulated by both ABA and GA in aleurone cells treated with the hormones for 4 hours. The 4 hour treatment was used because previous studies have shown that transcription of ABA induced some transcription factors reached a peak at 4 hours (Hoth et al., 2002). Out of the 2,443 ABA-inducible genes identified, 251 were induced by more than 4 fold. Using a bioinformatics approach, I identified a novel element that was overrepresented in the promoter regions of these 251 highly ABA-inducible genes. I named this element ABREN for ABA Responsive Element Novel. To determine whether or not this

element plays a role in ABA induction, transient expression analyses via particle bombardment were performed on rice aleurone cells. Constructs containing the β-glucuronidase (*GUS*) reporter gene driven by a promoter containing ABRENs in both the promoter and 5' UTR were introduced into the aleurone cells, then the cells were exposed to ABA and the reporter gene activity was measured. The results showed that when the ABREN in the promoter region was mutated, the level of ABA induction was significantly decreased, thus confirming ABREN's role in ABA signaling. The discovery of this novel element will greatly help elucidate stress response pathways in plants. Overall, my research has advanced our understanding of the signaling network underlying plant responses to stresses and may help develop more stress resistant crops to meet the ever increasing food demands by the growing world population.

ACKNOWLEDGEMENTS

I would like to thank Dr. Jeffery Q. Shen for his guidance in the lab and with life. He has helped me more than most other professors would have done and I thank him for his patience in putting up with me. Without his support I would not have been able to achieve my goals. I would also like to thank my committee members, Dr. Amei Amei, Dr. Andrew Andres, Dr. Laurel Raftery and Dr. Martin Schiller for their time, effort and support as well as their constructive criticisms, for without them I would not have been able to complete my degree. I would also like to thank my father Dr. Kyoichi A. Watanabe for supporting me through my studies. He has passed away this past year due to cancer, may he rest in peace and I apologize that he was unable to be here for my defense. I would like to thank my wife Thea Gee and my sisters Shauna Leigh Esposito, Kate Watanabe, Kim Cohen, Kelly Bedford, and Katharine Watanabe and my amazing newborn son Kyoichi Amadeus Watanabe.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

GENERAL INTRODUCTION

The world population is growing at a very rapid rate. As of 2016, there are 7.3 billion people on this world and the population is predicted to grow to 9.6 billion people by 2050 (Figure 1.1). With less agricultural land available and climate change resulting in drought conditions, the question arises how are we going to feed all these people.

To solve this problem, we first have to look at the current food supply to see what people are consuming today. As of 2009, rice and wheat tie for the grain with the highest calories consumed per capita in the world. Rice has the highest per capita consumption in Asia, while wheat has the highest per capita consumption in all other areas (Figure 1.2). Improving yields of either of these grains would help secure the world's food supply for our growing world population.

Between rice and wheat, rice was the clear choice of the grains to study because of its relatively small genome. The genome of rice (*Oryza sativa*) has been sequenced in 2002 and it contains 12 chromosomes and about 57 thousand predicted and confirmed genes. The genome contains about 380 million base pairs, which is about 1/8th the size of the human genome (Goff et al., 2002; Yu et al., 2002). On the other hand, the genome of wheat (*Triticum aestivum*) has yet to be sequenced but is estimated to be about 17 billion base pairs, about 5 times the size of the human genome. Wheat has 21 chromosomes, there are 124,201 genes identified so far and it has a hexaploid genome composed of three subgenomes from three different progenitor species (Eversole et al., 2014). This makes wheat a substantially more difficult crop plant to study than rice.

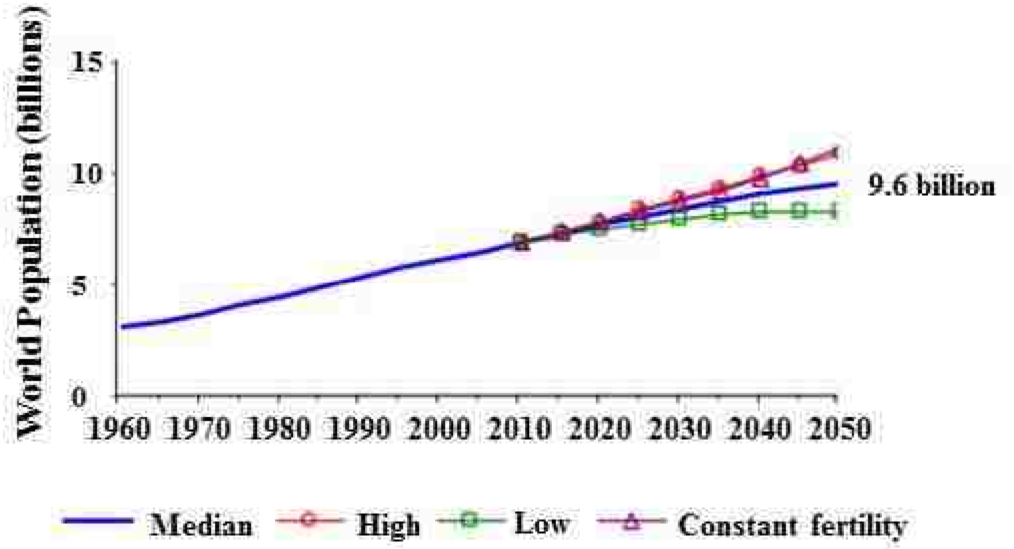


Figure 1.1: Projected world population from 1960 to 2050.

The solid line is the median projected world population. The line with open circles is the high projection. The line with open boxes is the low projection. The line with open triangles is the projection assuming constant fertility.
Data is from the Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat (2013). *World Population Prospects: The 2012 Revision*.

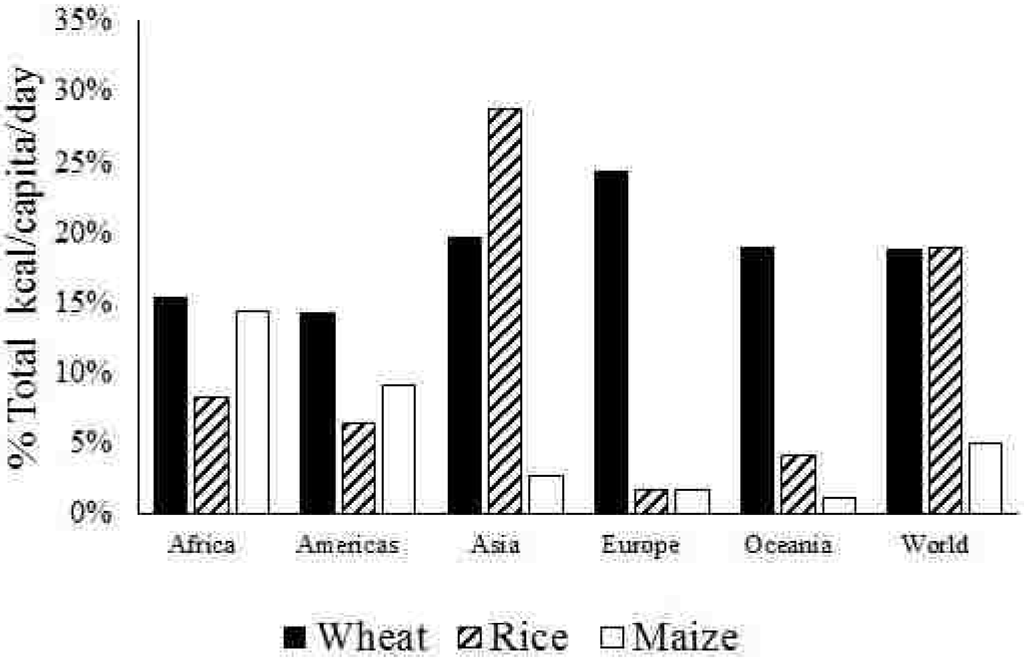


Figure 1.2: World grain consumption in 2009.

Wheat and rice tie as the most consumed grain worldwide. Rice is the highest consumed grain in Asia while wheat is the highest consumed grain in all other regions.
Food and Agriculture Organization of the United Nations (FAO) Database

The most important limiting factor for the production of rainfed rice is drought. Severe drought could affect rice production by more than 20% and can result in starvation and impoverishment in affected areas (Mohanty et al., 2013). Understanding the mechanisms of how rice responds to stresses such as drought may help us develop a drought tolerant strain of rice.

Another limiting factor for the production of grains is pre-harvest sprouting (PHS) or vivipary. PHS occurs when physiologically mature cereal grains germinate while on the panicle before harvest. PHS occurs in many cereal crops such as wheat, barley, maize, and rice in most regions of the world. PHS not only causes reduction of grain yield, but also affects the quality of grains, resulting in significant economic losses (Fang and Chu, 2008). A further understanding of seed germination will enable us to reduce the incidence of PHS, resulting in better yields.

Understanding the mechanisms of both stress and germination will result in higher crop yields and thus enable us to secure the world's food supply.

Abscisic acid (ABA) biosynthesis from β-carotene

Of primary importance to the engineering of stress tolerant plants is the understanding of the signaling pathways involved in the plants stress response. One key hormone that regulates plant stress response is abscisic acid (ABA). ABA is a 15-carbon terpenoid plant hormone that accumulates rapidly in response to stresses and mediates many biotic stresses, such as pathogen attack, and abiotic stresses, such as drought and salinity, that allow plants to survive under less than optimal conditions (Finkelstein and Rock, 2002). ABA plays important roles in other physiological processes such as promotion of stomatal closure (Bright et al., 2006), synthesis of storage proteins and lipids, leaf senescence and defense against pathogens (Finkelstein et al., 2002).

Abscisic acid is synthesized from β-carotene as shown in Figure 1.3. In plants, carotenoids are essential components of the photosynthetic apparatus, and thus, early ABA biosynthesis reactions occur in the plastids of the plant cells. First β-carotene is converted into zeaxanthin by β-carotene hydroxylase (BCH). Then zeaxanthin is converted into violaxanthin by zeaxanthin epoxidase (ZEP). Then violaxanthin is converted into xanthoxin by one of two pathways. Violaxanthin can be converted into 9-*cis*-violaxanthin by an undetermined isomerase, then to xanthoxin by 9-cis epoxicarotenoid dioxygenase (NCED). Alternatively, violaxanthin can be converted to neoxanthin by neoxanthan synthase (NSY), then to 9-*cis*-neoxanthin by an undetermined isomerase and then to xanthoxin by NCED. Xanthoxin is then transited to the cytoplasm where it undergoes conversion to abscisic aldehyde by a short-chain dehydrogenase/reductase (SDR) (Cheng et al., 2002; Ye et al., 2012). Then abscisic aldehyde is converted to ABA by abscisic aldehyde oxidase (AAO3). For a more detailed review see (Seo and Koshiba, 2002).

ABA signaling pathway

The current understanding of the ABA signaling pathway is depicted in Figure 1.4 below. The ABA receptors pyrabactin resistance1/PYR1-like/regulatory components of ABA receptors (PYR/PYL/RCAR) and are located in both the nucleus and cytosol (Raghavendra et al., 2010). These receptors have an ABA binding pocket and a gating loop. In the absence of ABA, the PYR/PYL/RCAR receptors do not interact with other proteins (Figure 1.4A). A constitutive negative regulator of ABA, protein phosphatase type 2C (PP2C), inhibits sucrose non-fermenting related kinase 2 (SnRK2) from autophosphorylating and activating its downstream targets. In the presence of ABA (Figure 1.4B), the ABA molecule binds to the pocket on the ABA receptor, which initiates an allosteric loop closure which locks the ABA molecule in place. This creates a

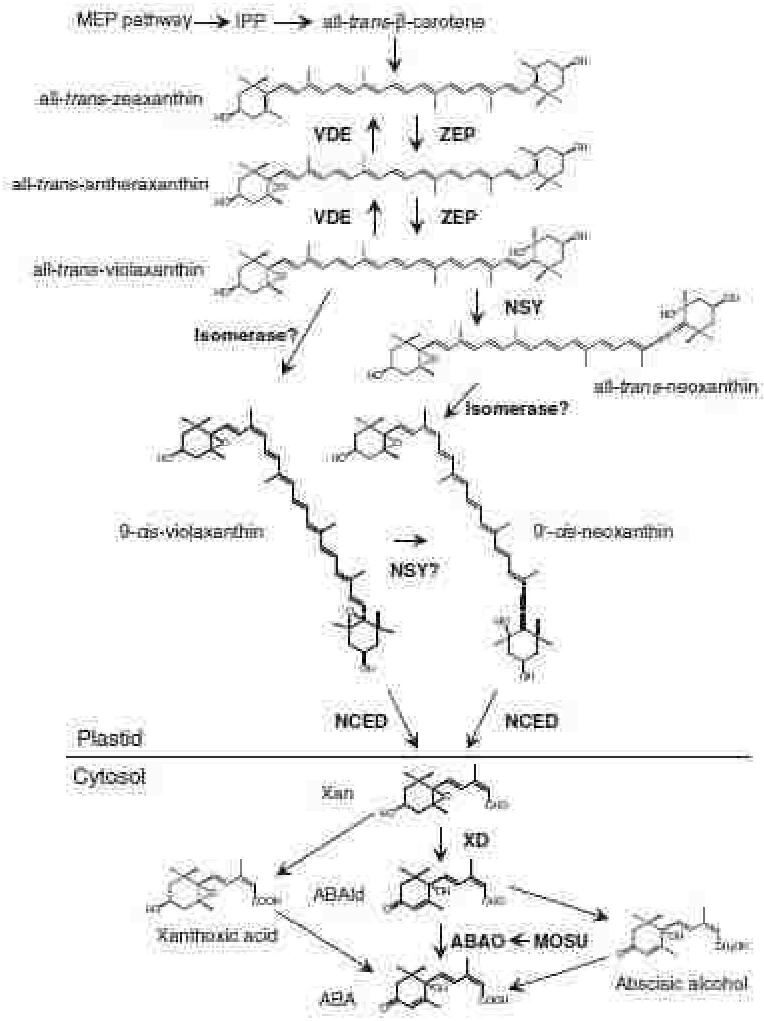
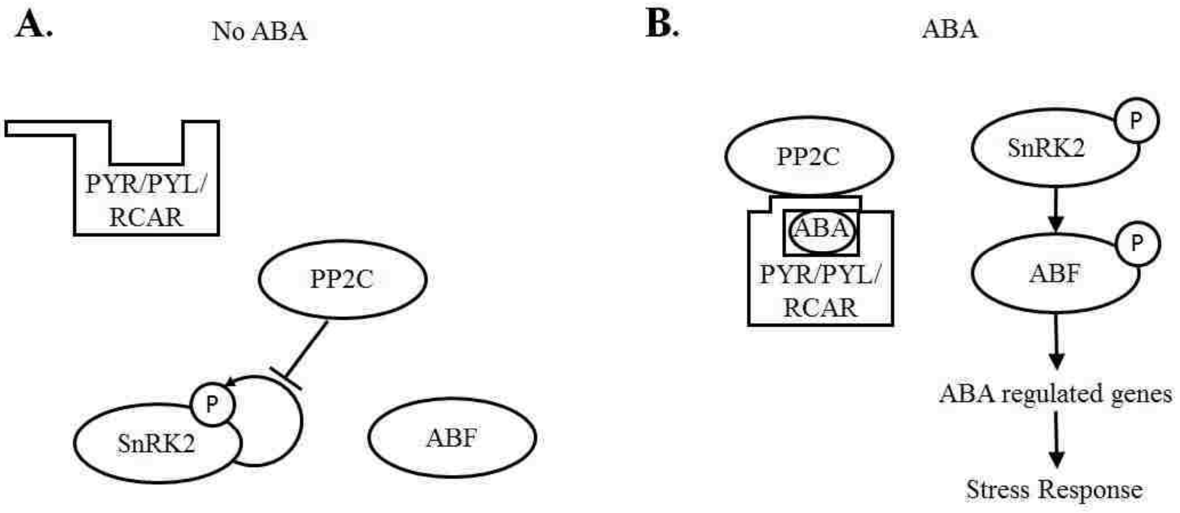


Figure 1.3: ABA biosynthetic pathway in higher plants.

The ABA precursor, a C40 carotenoid, is synthesized from IPP originating from the MEP pathway. Enzyme names are given in bold. IPP – isopentenyl pyrophosphate, ZEP – zeaxanthin epoxidase, VDE – violaxantin de-epoxidase, NSY - neoxanthin synthase, NCED – 9-*cis*-expoxycarotenoid dioxygenase, XD – xanthoxin dehydrogenase, ABAO – abscisic aldehyde oxidase, MOSU – MoCo sulfurase, Xan – xanthoxin, ABAld – abscisic aldehyde. Figure from (Endo et al., 2012).



C.

Transcription factor	cis-regulatory element	cis-regulatory element sequence
bZip (AtABI5)	ABRE	ACGTG(G/T)C
bZip (AtTRAB1)	CE3	GCGTGTC
AP2/EREBP (AtABI4)	CE1	TGCCACCGG
MYB	Myb binding element	YAAC(G/T)G
MYC	Myc binding element	CANNTG
VP1 (AtABI3)	Sph motif	CGTGTCGTCCATGCAT
Unknown	TT motif	TTTCGTGT
CpR18	CDeT27-45	AAGCCCAAATTTCACAGCCCGATTAACCG
AP2/EREBP	DRE1	CGAGAAGAACCGAGA
AP2/EREBP (ZmDBF1/2)	DRE2	CCGGGCCACCGAC GCACCG

Figure 1.4: **ABA signaling pathway.**

A.) In the absence of ABA, the ABA receptors (PYR/PYL/RCAR) do not interact with other proteins while PP2C inhibits SnRK2 auto activation. B.) In the presence of ABA, ABA is perceived by the ABA receptors and interact with PP2C, preventing PP2C from inhibiting SnRK2. SnRK2 can then activate downstream ABA transcription factors such as bZips C.) ABA transcription factors and their binding elements. PP2C – protein phosphatase type 2C, SnRK2 – sucrose non-fermenting related kinase, ABF – ABA transcription factor, P – phosphate.

binding site where PP2C binds thus preventing PP2C from interacting with the SnRK2 kinases. This allows the SnRK2 kinases to activate by autophosphorylation and then phosphorylate downstream ABA transcription factors (ABFs). The ABFs then induce transcription of ABA responsive genes which lead to the plant stress response (Sheard and Zheng, 2009). Many ABFs and their corresponding binding elements have been identified (Figure 1.4C) (Srivastava, 2002).

Gibberellic acid (GA) biosynthesis from geranyl-geranyl pyrophosphate

Another important hormone is the germination and growth promotion hormone gibberellic acid (GA). When seeds are exposed to water, the seed resumes metabolic and respiratory activity. Protein synthesis also resumes and enzymes involved in GA biosynthesis are produced (Bewley, 1997). GA is a tetracyclic diterpene compound which is biosynthesized by the seed embryo. GA diffuses to the aleurone layer of cells promoting transcription of hydrolases to break down the starchy endosperm to provide energy for germination (Wang et al., 1996). In addition to its role in germination, GA also stimulates stem (Marth et al., 1956) and root (Ohkawa et al., 1989) growth,and plays an important role in flower, fruit and seed development (Thomas and Sun, 2004).

There are many isoforms of GA, only a few of which are biologically active, GA1, GA3, GA4 and GA7 (Olszewski et al., 2002). Geranyl-geranyl pyrophosphate (GGPP) is a precursor from which all gibberellins are biosynthesized (Figure 1.5). During GA biosynthesis, GPP is converted into ent-copalyl diphosphate (ent-CPP) by ent-copalyl diphosphate synthase (CPS) and then into ent-kaurene by ent-kaurene synthase (KS) in the plastid. Then ent-kaurene is converted into ent-kaurenoic acid (ent-KA) by ent-kaurene oxidase (KO) and then into ent-7-hydroxy kaurenoic acid (ent-7HKA) by ent-kaurenoic acid oxidase (KAO) on the ER membrane. Then ent-7HKA is

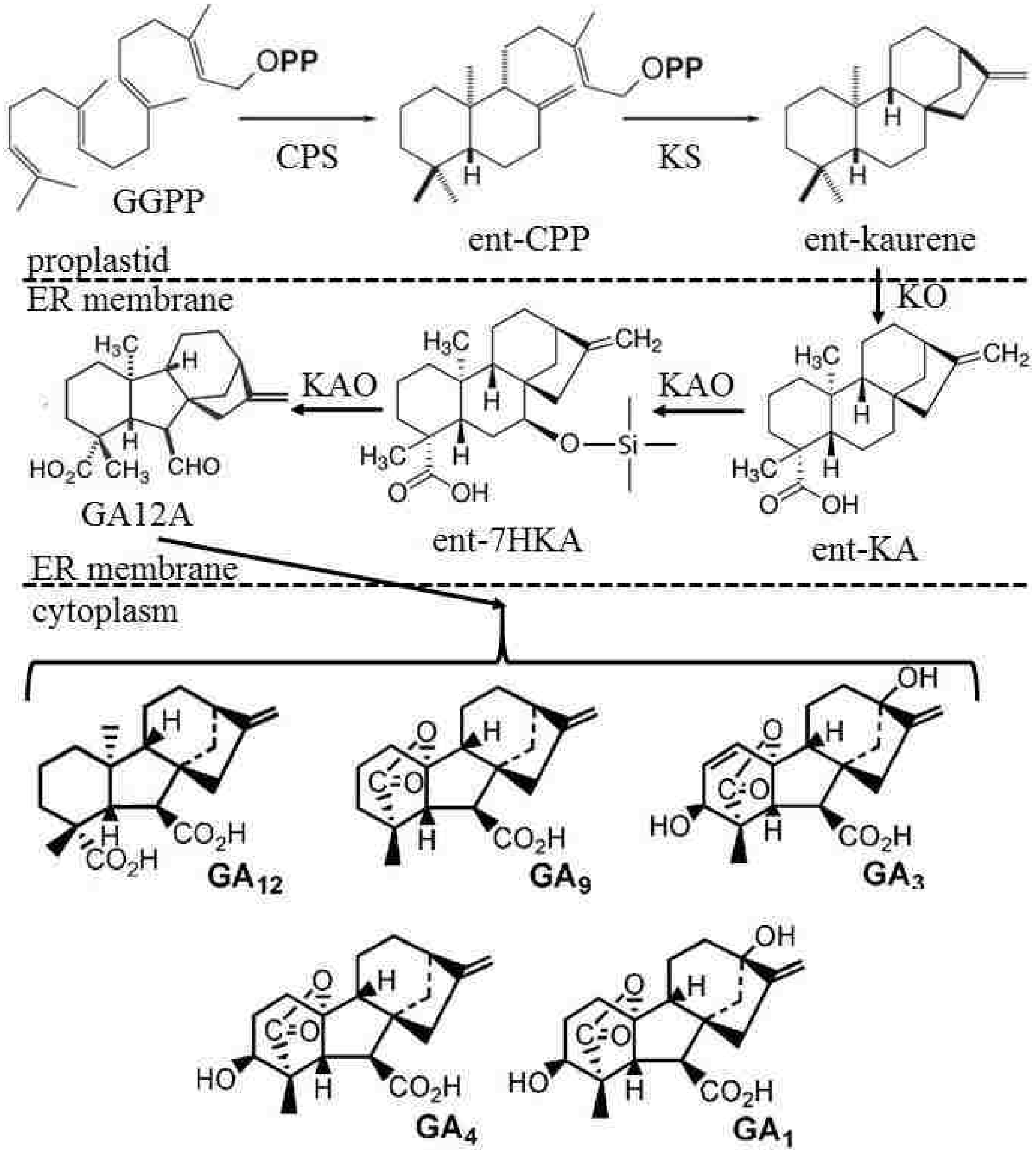


Figure 1.5: GA biosynthesis from geranyl-geranyl pyrophosphate

Geranyl-geranyl pyrophosphate (GPP) is converted into ent-copalyl (ent-CPP) diphosphate by ent-copalyl diphosphate synthase (CPS) and then into ent-kaurene by ent-kaurene synthase (KS). Then ent-kaurene is converted into ent-kaurenoic acid (ent-KA) by ent-kaurene oxidase (KO) and then into ent-7-hydroxy kaurenoic acid (ent-7HKA) by ent-kaurenoic acid oxidase (KAO). Then ent-7HKA is converted into gibberellic acid 12 aldehyde (GA12A) by KAO. GA12A can then be converted into GA12 or other forms of GA by enzyme catalyzed oxidation. This figure in part was contributed by (Hedden and Thomas, 2012).

converted into gibberellic acid 12 aldehyde (GA12A) by KAO. GA12A then enters into the cytoplasm where it can be converted into GA12 or other forms of GA by enzyme catalyzed oxidation. For a detailed review see (Hedden and Thomas, 2012).

GA signaling pathway

The current understanding of the GA signaling pathway is depicted in the Figure 1.6 below. The GA receptor, GA-insensitive dwarf 1 (GID1), is preferentially located in the nucleus (Ueguchi-Tanaka et al., 2005). In the absence of GA, GID1 does not interact with other proteins (Figure 1.6A). Phytochrome interacting factors (PIFs) interact with DELLA proteins which are negative regulators of GA signaling. In the presence of GA, the GID1 binds to GA and the N-terminal extension folds over to cover the GA bound pocket (Figure 1.6B). The folded N-terminal extension creates a binding surface for the DELLA protein (Murase et al., 2008). Then the skip-cullen-F-box ubiquitin E3 ligase (SCF) is recruited and unbiquitinates the DELLA protein tagging it for destruction by the 26S proteasome. This allows the PIFs to activate genes involved in germination and growth.

The seed is a great model system to study desiccation tolerance, hormone crosstalk and germination.

One approach of physiological research in dehydration tolerance has been to use specific plant structures that can withstand severe desiccation (Ingram and Bartels, 1996). During the final maturation stage of the development of seeds, as much as 90% of the original water is removed to attain a state of dormancy with unmeasurable metabolism (Leprince et al., 1993). This desiccated state allows survival under extreme environmental conditions and makes the seed the ideal model for the study of desiccation tolerance. The seeds of many species have been used to isolate the

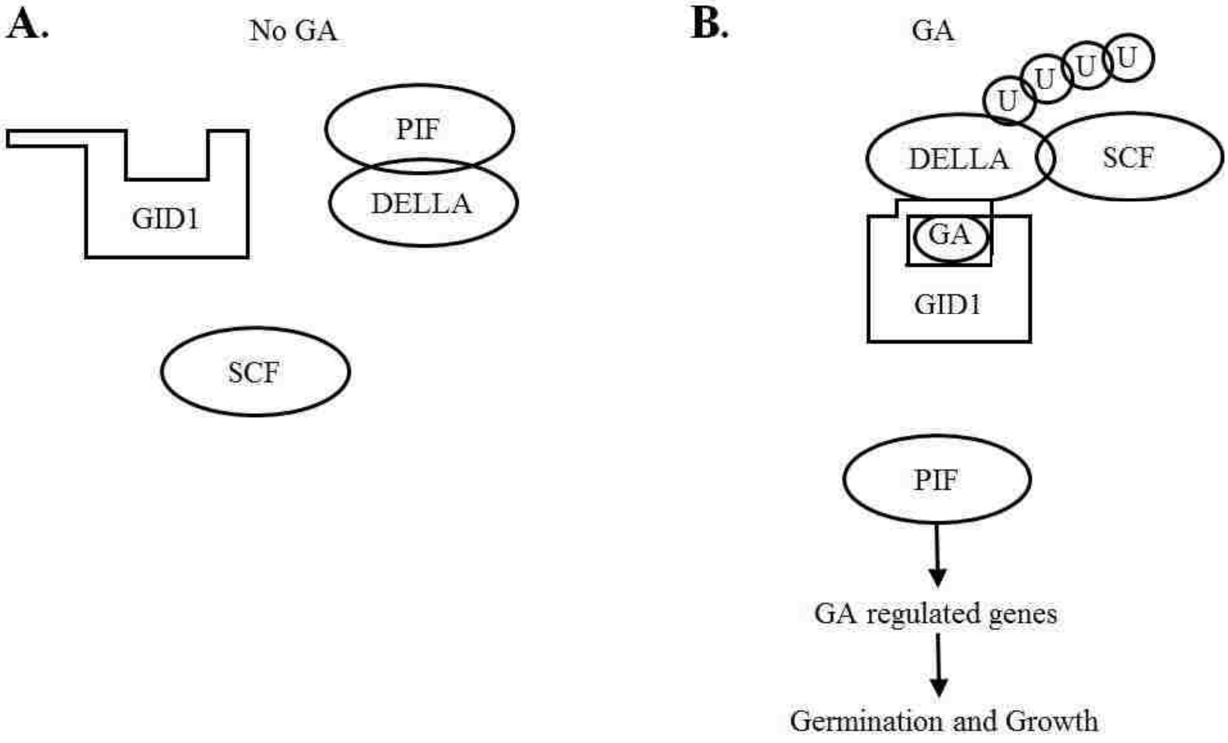


Figure 1.6: GA signaling network.

A.) In the absence of GA, the GA receptor GID1 does not interact with other proteins while DELLAs interact and inhibit PIF factors. B.) In the presence of GA, GA is perceived by GID1 and GID1 recruits SCF to ubiquitinate DELLA for degradation, leaving PIFs to activate downstream germination and growth genes. GID1 – GA insensitive dwarf 1, PIF – phytochrome interacting factor, SCF - skip-cullen-F-box E3 ubiquitin ligase, U – ubiquitin, DELLA – DELLA domain containing protein.

mRNA and proteins related to the desiccation-tolerance response, including, in particular, those of *Arabidopsis thaliana* and of food crop species such as barley (*Hordeum vulgare*) (Bartels et al., 1988), maize (*Zea mays*) (Pages et al., 1993), and rice (*Oryza sativa*) (Mundy and Chua, 1988).

The thin layer of cells that surround the endosperm of a seed is the aleurone layer of cells. These cells are terminally differentiated, easy to obtain and homogenous. The aleurone cells of seeds also make a great model for the study of germination and hormone crosstalk since these cells respond to hormones such as GA and ABA produced by the embryo but are not known to produce hormones themselves (Smith, 1977; Wang et al., 1996) . The aleurone cells play a key role in seed germination (Figure 1.7). When a seed is imbibed in water, the water triggers the biosynthesis of GA in the embryo. GA diffuses to the aleurone layer of cells where transcription of hydrolases such as alpha-amylase occur (Obata, 1975). These hydrolases then break down the starches in the endosperm to produce sugars, which provide energy for the embryo to grow. The hormone abscisic acid (ABA), which acts antagonistically to GA, blocks the GA induction of a germination response and promotes seed dormancy. Thus, studying this cell type will give us a better understanding of hormone crosstalk. It will also lead to a further understanding of seed germination and this knowledge may prevent pre-harvest sprouting from occurring.

RNA-seq for transcriptome analysis of rice aleurone cells

To study desiccation tolerance, hormone crosstalk and germination of rice on a transcriptome-wide scale, the transcriptome of hormone treated rice aleurone would be of great help. To obtain the necessary transcriptomic data, RNA-seq can be utilized. RNA-seq is performed by extracting high quality mRNA from a biological source such as the rice aleurone cells. Then the mRNA is fragmented and size selected to lengths of about 300 bp, the Illumina recommended optimal length

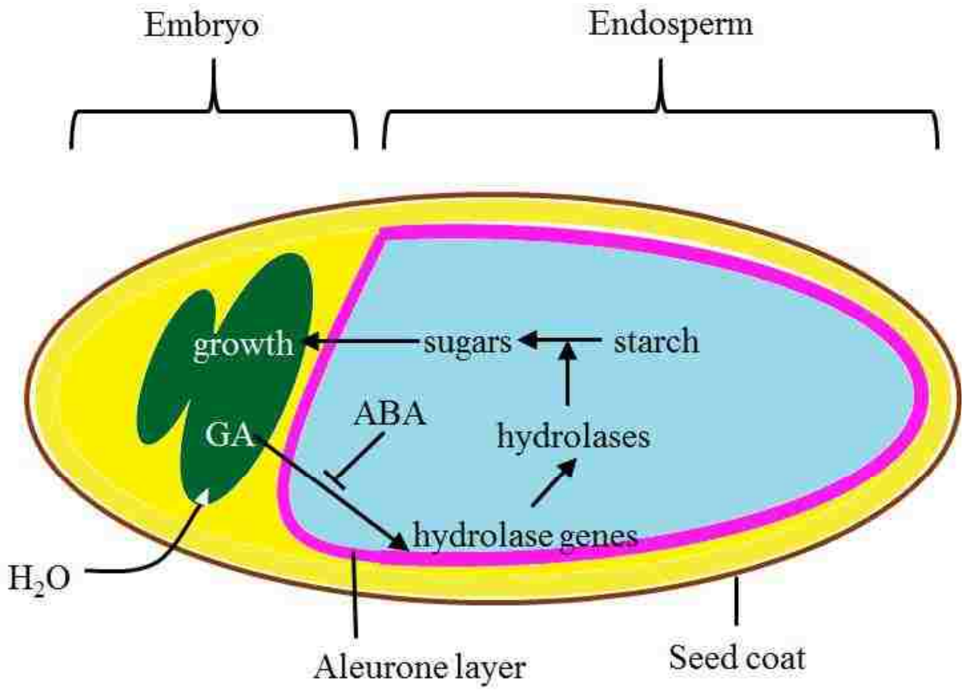


Figure 1.7: Hormone signaling pathways involved in seed germination.

Water is taken up by the embryo and triggers the biosynthesis of GA. GA diffuses to the aleurone layer of cells inducing the transcription of hydrolases such as alpha-amylase occur. These hydrolases then break down the starches in the endosperm to produce sugars which can then provide energy for the embryo to grow.

for single-end reads. The mRNA is then converted into complementary DNA (cDNA) via random primers. Adapters are appended to the ends of the fragments. The fragments are fixed onto a flow cell and amplified by PCR. Then the fragments are sequenced by using fluorescently tagged nucleotides producing “short reads” (Figure 1.8). The resulting short reads can be mapped to the rice genome via mapping software. For a more detailed review, see (Wang et al., 2009b).

Substantial amounts of information can be obtained from these short reads. Taken together, these reads represent the entire transcriptome of the biological source at a given point in time for a given treatment. For example, RNA-seq can be utilized to identify hormonal response of genes on a transcriptome wide scale. By performing RNA-seq on a control sample, and samples exposed to hormones ABA, GA or a combination of the two hormones, hormone induced differential expression of genes can be observed (Trapnell et al., 2010). Genes that have few reads aligned in the control sample, but have many reads aligned in the hormone treated sample (or vice versa) are considered differentially expressed. Identifying the genes that are differentially expressed by a hormone treatment will give clues to which genes are involved in the hormone signaling network. This will also allow one to build co-expression networks of genes and will help in understanding the hormone signaling pathway, as well as hormone crosstalk.

There are numerous other applications of RNA-seq, including identification of unannotated genes (Lu et al., 2010), identification of transcription start and transcription termination sites (Watanabe et al., 2015), identification of induced alternative splicing (Gulledge et al., 2012), identification of single nucleotide polymorphisms (SNPs) (Quinn et al., 2013), identification of quantitative trait loci (QTL) (Lionikas et al., 2012), plant organ specific transcriptomes and development of plant organs (Guan and Lu, 2013), and other applications in various species.

In this dissertation, RNA-seq was utilized for the identification of unannotated genes. For the rice genome, most genes are computationally predicted based on gene models. However, the use of models often leads to many false negative results, resulting in many real genes remaining undiscovered. Regions of the genome where many reads align but no known gene is present are indicative of an undiscovered gene (Roberts et al., 2011). However, read alignment to an unannotated region of the genome does not always mean that there is a gene present at that location. Past genome fragment duplication events have led to regions of the genome that show homology to other regions of the genome. During read mapping, if a read can map equally well to multiple locations, one of those locations is randomly assigned to the read. Because of this, there may be regions of the genome that are not expressed but have high read alignment due to high sequence homology to an expressed region. If this were to happen, the region that was actually expressed would show a decrease in measured expression over the actual expression level, and the unexpressed region would appear to be expressed. Thus, precautions must be made to ensure that regions that show homology to another genomic region are not classified as novel unannotated genes (Watanabe et al., 2014).

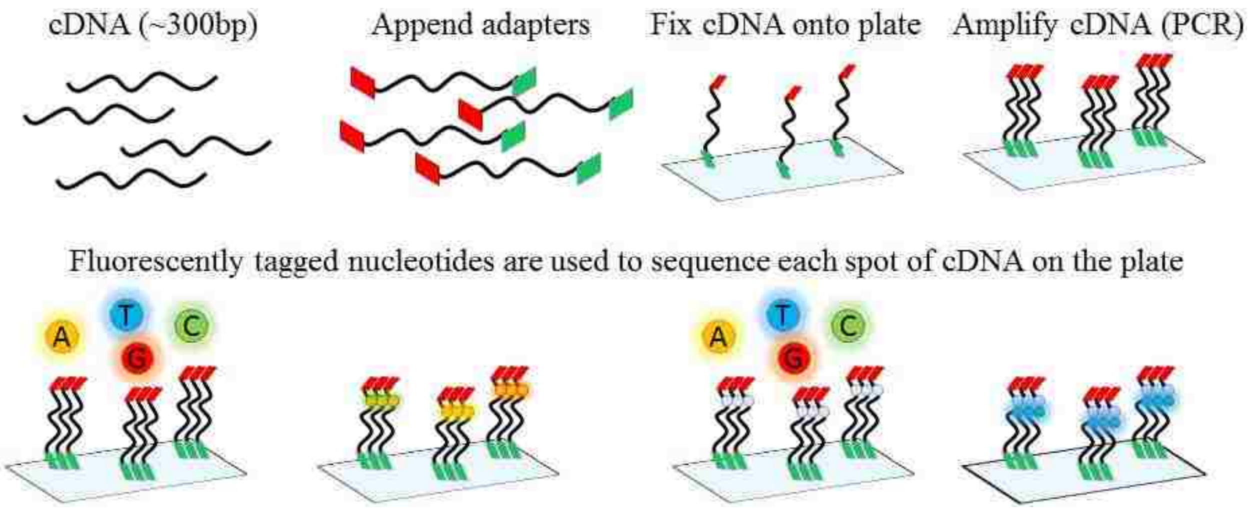

cDNA (~300bp)
Append adapters
Fix cDNA onto plate
Amplify cDNA (PCR)
Fluorescently tagged nucleotides are used to sequence each spot of cDNA on the plate
A
T
G
C
A
T
G
C

Figure 1.8: RNA sequencing work flow and sequencing of short reads.

Adapters are appended to the ends of the cDNA fragments. The fragments are fixed onto a flow cell and amplified by PCR. Then the fragments are sequenced by using fluorescently tagged nucleotides producing short reads

Bioinformatic Identification of cis-regulatory elements involved in hormone signaling

cis-regulatory elements (CREs) are specific short sequences (typically 8 to 14 bp in length) which regulate the transcription of nearby genes and are usually located in the non-coding regions of DNA. They are concentrated near the promoter region upstream of the gene they regulate. CREs typically regulate gene transcription by functioning as binding sites for transcription factors. For example, the ABA response element (ABRE) has the sequence ACGTG(T/G)C and is the binding site for bZip transcription factors. bZip transcription factors are upregulated in the presence of ABA, and thus upregulate the genes containing the ABRE.

In the rice genome, only 39.5% of the genes upregulated by 4-hour treatment of ABA contain an ABRE in their promoters Though some of these genes may be indirectly induced by ABA by downstream ABA induced transcription factors, the short hormone treatment reduces this effect. Thus, the low frequency of the ABRE is an indication that there may be other elements involved in the immediate ABA response. To identify additional *cis*-regulatory elements involved in the ABA signaling pathway, all the ABA induced genes were identified and the promoter regions were analyzed for enriched elements.

One method for identification of *cis*-regulatory elements is by exhaustive bioinformatic search. This is accomplished by examining all possible 8 – 14 bp sequences and determining their level of enrichment among the promoter regions of ABA inducible genes (2,443 genes). Unfortunately, this method is computationally expensive and takes substantial amounts of time.

There are algorithmic strategies to avoid an exhaustive search. One such strategy is Multiple Em for Motif Elicitation (MEME) (Bailey and Elkan, 1994). This method identifies multiple motifs by utilizing expectation maximization (Em). This process uses an iterative process that grows exponentially with the number of sequences. Since the number of ABA inducible genes is quite high, this method is also quite time consuming.

Gibbs sampling was the ultimate strategy chosen to identify enriched elements amongst the promoters of ABA inducible genes. Gibbs sampling is a very fast algorithm and produces accurate results. Gibbs sampling works by selecting a random motif, then querying the motif against the set of DNA sequences, in this case the promoter regions of hormone induced genes, to generate a score. The higher the score, the more abundantly the motif exists in the promoter regions. Then the selected motif is slightly modified in an attempt to produce a higher score. The score of the modified motif is calculated and if the new score is higher than the previous score, the new motif replaces the previously selected motif. The process is iterated until the program converges to a maximally scoring motif. This maximally scoring motif is termed a local maximum. There may be several local maxima for a given set of promoter regions. The sampling is performed many times with random seeds to identify several local maxima. The highest scoring local maxima are the most enriched elements and thus are the motifs that are most likely involved in hormonal response.

Particle bombardment to confirm the role of cis-acting elements in the plant stress response

To determine whether the elements identified in the promoter regions of hormone inducible genes are responsible for hormone induction, these elements can be inserted into a plasmid construct upstream of a reporter gene. Then the construct can be introduced into plant cells, the cells can be treated with hormones, then the expression level of the reporter gene can be measured to determine the level of hormone induction. There are several methods to introduce constructs into cells, these include heat shock transformation and electroporation. However, unlike bacterial cells, plant cells have a thick cell wall which prevents constructs from entering the plant cells by these methods.

To overcome these issues, particle bombardment via a gene gun is the transformation method of choice. The results have been proven to be reliable and the turn-around time is quick. For example, to determine whether elements in the promoter regions of hormone inducible genes confer hormone induction, these elements can be inserted into the minimal promoter region upstream of a *GUS* reporter gene within a DNA construct. The constructs can be introduced into plant cells by accelerating tungsten nanoparticles coated with the construct and bombarding them into the plant cells. The gene gun utilizes a polyethylene projectile containing a droplet of tungsten nanoparticles coated with the construct DNA (Figure 1.9A). A .22 caliber nail gun cartridge was used to propel the projectile down the barrel of the gun at high velocity (Figure 1.9B). When the projectile reaches the end of the barrel, it impacts a annulus shaped pellet stopper and is stopped (Figure 1.9C). The DNA passes through the hole of the annulus spraying the DNA coated particles at high velocity onto the plant cells below. The plant cells are then imbibed in a solution containing hormones. If the reporter gene of the hormone treated sample was expressed at a level higher than the control, then this demonstrates the promoter elements in the construct is responsible for the induction.

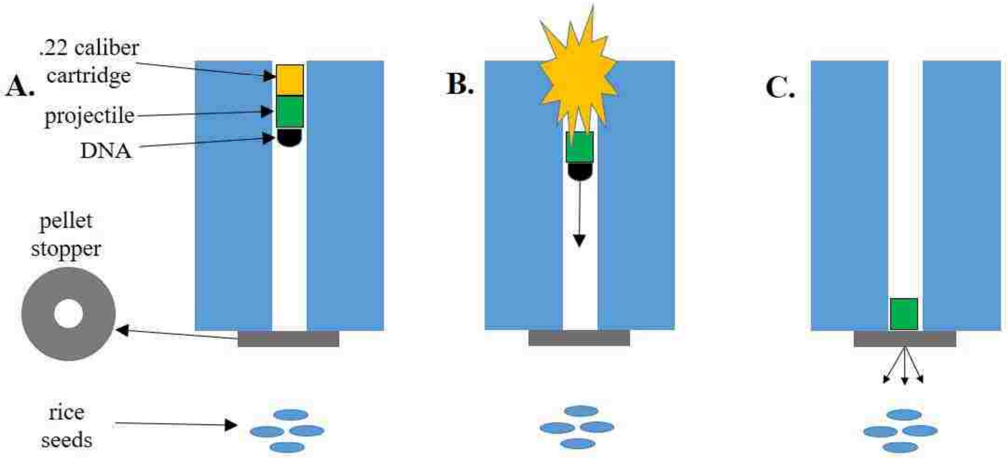


Figure 1.9: Particle bombardment via gene gun.
A.) A .22 caliber nail gun cartridge and projectile with a drop of tungsten coated DNA is packed into the barrel of the gun. B.) the cartridge is ignited accelerating the projectile down the barrel of the gun. C.) The projectile hits the doughnut shaped pellet stopper and stops while the DNA passes through the doughnut hole onto the aleurone cells below.

Dissertation Scope

The ultimate aim of this research is to understand the signaling pathways involved in stress response and seed germination in rice. If the genome is not fully annotated, then it may be impossible to build a complete pathway if key genes in the signaling pathway have not been identified and annotated. Therefore, it is important that as many genes as possible of the rice genome need to be annotated. From the research presented in Chapter 2, it was demonstrated by RNA-seq analysis, that as many as 8% of the rice genes may be unannotated. Hundreds of high confidence novel genes were identified and a subset of them were experimentally confirmed by RT-PCR demonstrating that the rice genome annotation is still far from complete.

Current bioinformatics tools to identify genes are lacking and may be a contributing factor for the shortcomings of the current rice genome annotation. Better tools and pipelines to identify genes needed to be developed. In Chapter 3, the Tiling Assembly was developed and compared to a popular transcript assembly software Cufflinks. The Tiling Assembly was shown to have superior gene finding capabilities to Cufflinks and hence is an invaluable tool to annotate genomes of many organisms.

To further understand the hormone signaling networks in rice, the genes that were responsive to the hormone abscisic acid (ABA) were identified by RNA-seq. Elements that are enriched in the promoters of the ABA induced genes may be *cis*-regulatory elements involved in the plant hormone response. In Chapter 4, a bioinformatics approach was used to identify an enriched element in the promoters of ABA induced genes which we named ABREN for ABA Responsive Element Novel. Transient expression experiments via particle bombardment confirmed that the ABREN plays a role in the ABA signaling pathway.

The need for accurate high quality transcript structure diagrams is of great importance to researchers. Currently available software have limitations on number of transcripts that can be displayed, font selection, customization, speed and lack of publication quality images. Other software require substantial user input such as GFF files, protein sequences and cDNA sequences, not all of which may be available to the user. In Chapter 5, the Transcript Structure and Domain Display (TSDD) software is described that resolves all these issues. TSDD is an online program so no software installation is required, and it is fast, easy to use, requires only a GFF file and produces publication quality transcript structure diagrams.

In Chapter 6, a brief summary of the previous chapters, discussion of the results and possible future directions for the research presented in this dissertation. These include further identification of novel genes, determination of rice gene coexpression networks, study of the crosstalk of sugar and ABA signaling, production of stable transgenic plant containing the ABREN, identification of the protein binding partner of the ABREN, crosstalk between ABA and GA signaling, and RNA-seq analyses of non-poly adenylated mRNA.

CHAPTER 2

IDENTIFICATION OF NOVEL GENES IN RICE AND THEIR DIFFERNTIAL EXPRESSION IN RESPONSE TO HORMONES

Previously published as:

RNA-sequencing reveals previously unannotated protein- and microRNA-coding genes expressed in aleurone cells of rice seeds.

Kenneth A. Watanabe, Patricia Ringler, Ling Kun Gu, Qingxi J. Shen.

Genomics. 2014 Jan;103(1):122-34. doi: 10.1016/j.ygeno.2013.10.007

Disclaimer:

KAW participated in RNA extraction, experiments, performed the sequence alignment, made most of the figures and composed most of the manuscript. PR participated in RNA extraction, composed some of the abstract, introduction and discussion and participated in reviewing the manuscript. LKG participated in RNA extraction and experiments. JQS conceived the study and supervised the design and coordination of the experiments. All authors read and approved the final manuscript.

Abstract

The rice genome annotation has been greatly improved in recent years, largely due to the availability of full length cDNA sequences derived from many tissues. Among those yet to be studied is the aleurone layer, which produces hydrolases for mobilization of seed storage reserves during seed germination and post germination growth. Herein, we report transcriptomes of aleurone cells treated with the hormones abscisic acid, gibberellic acid, or both. Using a comprehensive approach, we identified hundreds of novel genes. To minimize the number of false positives, only transcripts that did not overlap with existing annotations, had a high level of expression, and showed a high level of uniqueness within the rice genome were considered to be novel genes. This approach led to the identification of 553 novel genes that encode proteins and/or microRNAs. The transcriptome data reported here will help to further improve the annotation of the rice genome.

Introduction

Nearly half of the world population relies on rice as a staple food source (Mohanty et al., 2013). With climate change affecting the amount of agricultural land available, in conjunction with the growing population, increasing food production plays a vital role in reducing hunger worldwide (Zhang and Cai, 2011). In 2002, a draft of the rice genome was released, and rice became the first crop genome to be sequenced (Goff et al., 2002; Yu et al., 2002). The rice genome is about 380 million base pairs, and there are over 57,000 putative and confirmed genes in the japonica rice genome (Ouyang et al., 2007). The rice genome has been instrumental for research leading to further understanding of other monocot species, such as maize (Alexandrov et al., 2009) and barley (Thiel et al., 2009) , as well as grass evolution in general (Campbell et al., 2007). The annotation of the rice genome is continually being refined with the discovery of new genes and transcription units, and improved understanding of known genes.

The phytohormone ABA plays a central role in the plant stress response system. ABA accumulates rapidly in response to stresses and mediates many biotic (Mauch-Mani and Mauch, 2005) and abiotic (Zhang et al., 2006) stress responses that allow plants to survive under less than optimal conditions. ABA is synthesized in roots in response to decreased soil water potential and plays a role in root and shoot growth (Sharp and LeNoble, 2002). In leaves, ABA alters the osmotic potential of stomatal guard cells, causing the stomata to close, and preventing water loss through transpiration (Bright et al., 2006). ABA also inhibits seed germination through the antagonism of the signaling pathway of the germination-promoting hormone, GA (Gomez-Cadenas et al., 2001; Wang et al., 1996). In addition to its role in germination, GA can also stimulate rapid stem (Marth et al., 1956) and root (Ohkawa et al., 1989) growth, induce mitotic division in the leaves of some plants, and plays an important role in flower, fruit and seed development (Thomas and Sun, 2004).

The cells of the seed aleurone layer are involved in the germination pathway and are responsive to both ABA and GA generated by the embryo (Wang et al., 1996). In addition, the aleurone cells are terminally differentiated and have a high degree of homogeneity (Smith, 1977). However, RNA-sequencing (RNA-seq) analysis on this cell type has not been published before. Therefore, the transcriptome of aleurone specific genes remains to be revealed.

In this study, we investigated novel rice genes that are expressed in aleurone cells by RNA-seq. RNA-seq allows the analysis of gene expression on a transcriptome-wide scale. This is performed through extraction of mRNA, fragmenting and sequencing of cDNA, and mapping the resulting short reads to the rice genome (Wang et al., 2009b). Since RNA-seq does not depend on genome annotation for prior probe selection, and avoids biases introduced during hybridization of microarrays, RNA-seq is the method of choice for transcript discovery and genome annotation (Baginsky et al., 2010). As such, RNA-seq has been used for transcript discovery in several species including planaria (Blythe et al., 2010), mosquitoes (Bonizzoni et al., 2012), *Drosophila* (Palmieri et al., 2012), *Xenopus* (Tan et al., 2013), mice (Trapnell et al., 2010), humans (Roberts et al., 2011), and rice (Lu et al., 2010; Zhang et al., 2010a). Here, the aleurone cells were treated with the hormones ABA, GA, and a combination of the two hormones. Short reads from the RNA-seq data resulting from these experiments were aligned to the genome using the Bowtie and Tophat alignment software (Langmead et al., 2009; Trapnell et al., 2009). Results were compared to previously published data to confirm the quality of the data. Novel genes within the unannotated regions of the rice genome were identified using a combination of two methods, the well accepted Cufflinks assembly algorithm (Trapnell et al., 2010) in conjunction with a novel algorithm developed in-house. The results were filtered via BLAST search (Altschul et al., 1990) to remove potential novel genes that had high similarity to another genomic region to eliminate false positive results. Finally, the novel genes were analyzed for potential protein coding sequences, similarity to known primary microRNAs, and hormone regulation as detected by the RNA-seq analysis. The flowchart depicting the steps taken to identify these novel genes is shown in Figure 2.1.

The results of our study identified 553 potential novel genes, which do not overlap with an annotated gene, show less than 25% similarity to another genomic region and have an expression level greater than 1.0 RPKM (reads per kilobase of exon model per million mapped reads). Of these transcripts, 302 showed homology to known protein sequences, based on BLASTX queries, and 124 showed homology to known plant primary microRNAs.

A large proportion of RNA-seq reads aligned to unannotated regions of the rice genome, indicating the presence of unidentified novel genes

RNA-seq was performed on four samples: a control sample, and samples treated with ABA, GA, and a mixture of the two hormones. The results yielded a total of about 158 million reads across the four samples (Supplemental Table S1). The read count ranged from 37.5 million to

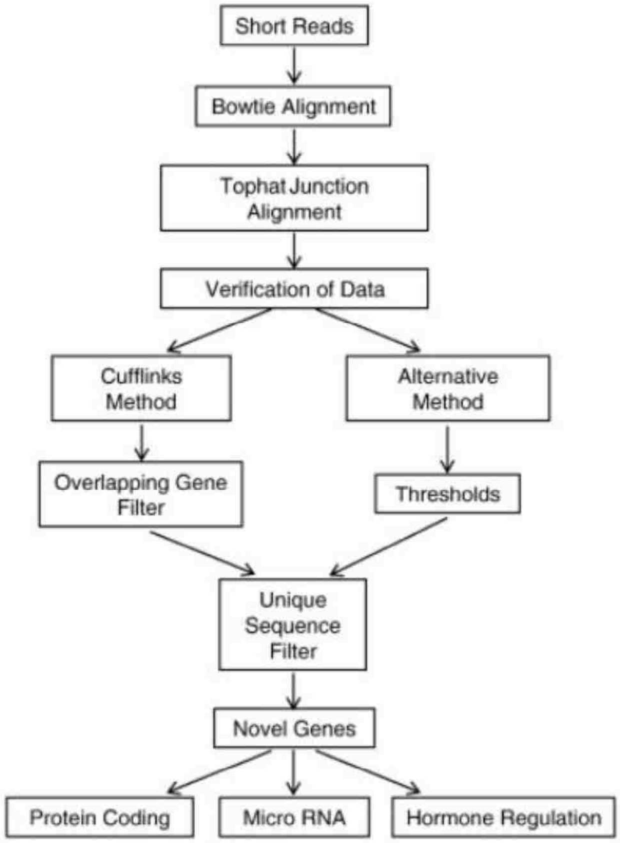


Figure 2.1: Flowchart depicting the workflow of this study.

This flowchart shows the logical sequence of this study, starting from the RNA-seq short read data, through the discovery of novel genes, prediction of protein coding sequences, microRNA prediction, and determination of differential expression by hormone treatment.

40.3 million reads per sample. In total, about 132 million reads (83.4%) were mapped to the rice genome via the Bowtie software. About 10 million reads (8.4%) mapped to unannotated regions of the rice genome. There are several possible reasons why this may have occurred. These reads may have mapped to unannotated exons of neighboring transcripts. These regions may also belong to pseudogenes or inactive genes created by past gene duplication events and which share homology with actively expressed genes. It is also possible that reads aligned to these areas due to mapping errors or DNA contamination. Finally, these reads may be derived from transcripts of undiscovered genes within these regions.

A combination of Cufflinks and a novel algorithm were used to identify novel genes in unannotated regions of the rice genome

Running Cufflinks on the pooled short read data of all our samples resulted in the identification of 95,620 transcripts. These transcripts included 57,617 currently annotated transcripts and 38,003 previously unannotated transcripts. To prevent novel exons of annotated genes from being classified as novel transcripts, the transcripts that were identified by Cufflinks that overlapped with currently annotated genes were not considered as novel genes. In order to exclude results which might be due to noise or trace amounts of genomic DNA contamination, only those genes with an expression level greater than or equal to 1.0 RPKM were included, leaving 1,409 potential novel genes as detected by Cufflinks.

While Cufflinks has been well demonstrated for its capacity in detecting novel transcripts from an RNA-seq data set, we found that the software did not identify many unannotated regions of the transcriptome which were found to contain large numbers of reads. Read alignment to an unannotated region is indicative of an expressed novel gene. A novel algorithm for gene discovery was developed in-house to confirm the accuracy of the Cufflinks software and to identify additional novel genes. This novel program scans the rice genome for clusters of mapped reads in unannotated regions of the rice genome. If the standard deviation of the cluster is beyond a threshold from the nearest neighboring gene, then this cluster is considered as a potential novel gene (Figure 2.2: A). Using this novel Clustering algorithm, 569 regions were identified as potential novel genes, ranging in size from 76 to 20,271 base pairs (bp).

While examining the potential novel genes identified through the Clustering algorithm, we found a number of identified regions that showed a small number of narrow peaks with a high number of reads aligning on those peaks. A number of unannotated regions were found to contain large numbers of reads aligned to a region that was equivalent to less than three reads in width (Figure 2.2B). It is likely that alignment of reads to these regions was due to the presence of a highly expressed gene or genes with a similar sequence, and that this region is not a novel gene. To eliminate inclusion of this and similar regions in the list of potential novel genes, we chose to use a footprint equivalent to the footprint of this region, 140 bp, as the minimum cutoff for consideration as a novel gene. One of the remaining genes was greater than 10 kb (20,271 bp) and may contain several novel genes. Separate analysis needs to be performed to identify potential individual genes within this region.

In order to eliminate the possibility that the identified novel genes are unannotated exons of neighboring genes, PERL scripts were written to scan for flanking junctions of the remaining novel genes. After eliminating possible false-positive results through these methods, a total of 356 potential novel genes were predicted by the novel algorithm.

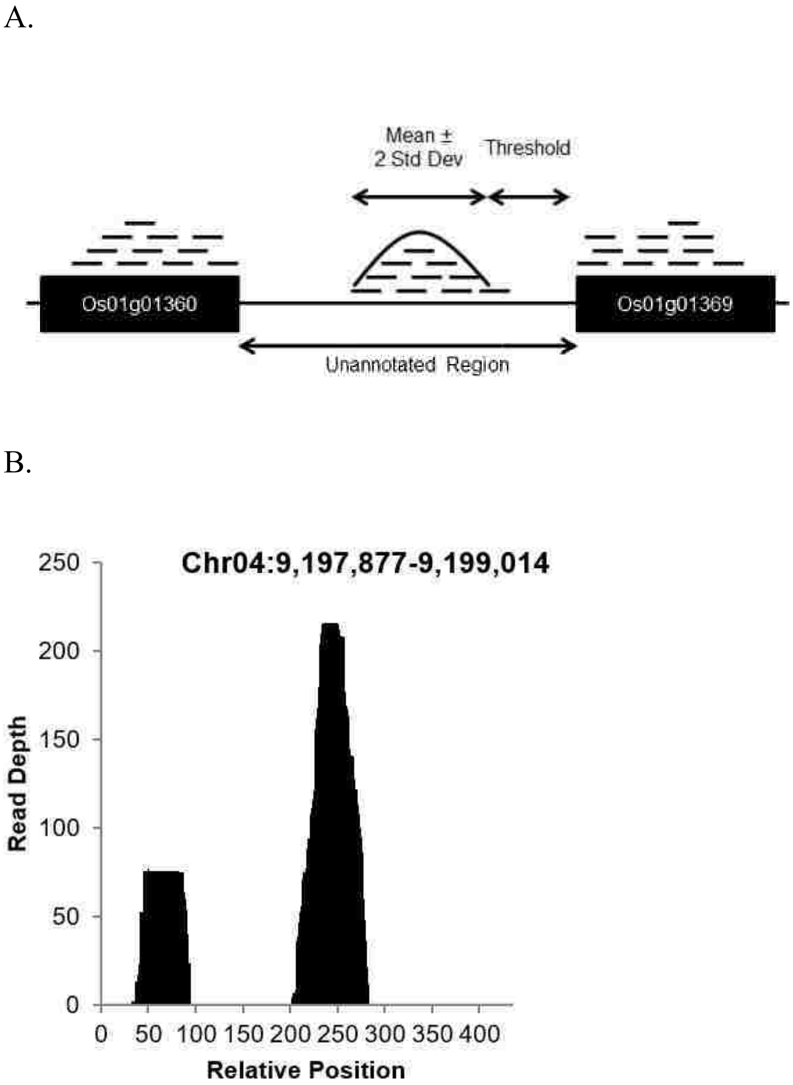


Figure 2.2: Methodology for identifying novel genes.

(A) The novel gene finding algorithm calculated the mean and standard deviation of the midpoint of the reads mapping to unannotated regions. For regions with more than 200 reads, if ±2 standard deviations from the mean was completely within the unannotated region, and the distance from the nearest gene was greater than the threshold, the region was selected as a possible novel gene. The threshold was the larger of 5% of the size of the unannotated region or 100 bp. (B) The footprint of the two peaks shown totals 139 bp. Reads most likely aligned to this region due to similarity to one or more other regions of the rice genome. The footprint of this example was used as the minimum length for inclusion as a potential novel gene. Read depth is determined from the cumulative reads of all samples.

Regions with high sequence similarity to another genomic region were excluded as potential novel genes

Since Bowtie (Langmead et al., 2009) aligns reads that match more than one position in the genome randomly to one of these positions, we recognized the possibility that many of the novel genes identified by Cufflinks may be corresponding to regions of the genome that are not expressed, but show high similarity to an expressed region. If this were to happen, the region that was actually expressed would show a decrease in measured expression over the actual expression level, and the region that was not expressed would appear to be expressed.

One predicted novel gene, OsNT10-38 (chr10:18,181,961-18,185,429) which was identified by both Cufflinks and the novel algorithm, shares 99.1% similarity to the annotated gene LOC_Os12g05700, a putative transposon protein. More than twice the number of reads align to OsNT10-38 than LOC_Os12g05700, which suggests that OsNT10-38 is expressed. Further experimentation must be performed to determine whether OsNT10-38, LOC_Os12g05700, or both are expressed.

The predicted novel genes OsNT07-44 (chr07:26,386,827-26,389,003) and OsNT03-38 (chr03:21,945,597-21,947,756), which were also identified by both Cufflinks and the novel algorithm, are 99.5% similar to each other but share less than 10% similarity to another genomic region. Both genes have about the same number of reads aligned. While one or both of these genes may be an expressed novel gene, we have excluded both to prevent the possibility that either is a false positive result.

The predicted novel gene OsNT02-38 (chr02:33,647,022-33,648,205), identified by the novel algorithm, is 99% identical to an intronic region of LOC_Os09g26770 (chr09:16,259,839-

16,261,025), a ribosomal L18p/L5e family protein (Figure 2.4). The alignment data suggest that the intron region of LOC_Os09g26770 is retained, since there are more reads aligning to the intron than to the identified potential novel gene. In addition, where there are differences in sequence between the two regions, few or no reads align to OsNT02-38, as can be seen by the locations of the black triangles in Figure 2.3. Further experiments need to be performed to confirm whether the intron is expressed or OsNT02-38 is expressed.

In all three of the above mentioned examples, further analysis must be performed to determine whether the predicted novel gene or its identical counterpart, which can be an annotated gene, another novel gene, or an intron, is expressed. Thus, we eliminated the genes that have a high similarity to another region of the rice genome as potential novel genes to reduce the rate of false positive results. This does not exclude the possibility that some of these genes are expressed novel genes.

Chromosome nine contains a large section with two annotated uncharacterized genes within a region of rRNA repeats

The novel gene OsNT09-01a (chr09:1137-9103) has 1,413,301 reads aligned and corresponds with the location of the 17S and 25S rRNA. A BLAST search shows that OsNT09-01a is 100% identical to the regions corresponding to chr09:9084-17,031 (OsNT09-01b), chr09:17,012-24,959 (OsNT09-01c) and chr09:24,940-32,887 (OsNT09-01d) and part of the region at chr09:32868-36815 (OsNT09-01e) (Figure 2.4). The four regions (OsNT09-01a, OsNT09-01b, OsNT09-01c and OsNT09-01d) are tandem repeats of each other. Since the rice genome was assembled via random-fragment shotgun sequencing, which uses overlapping fragments to generate the genomic sequence, the exact number of repeats of this rRNA region is not known and is potentially variable.

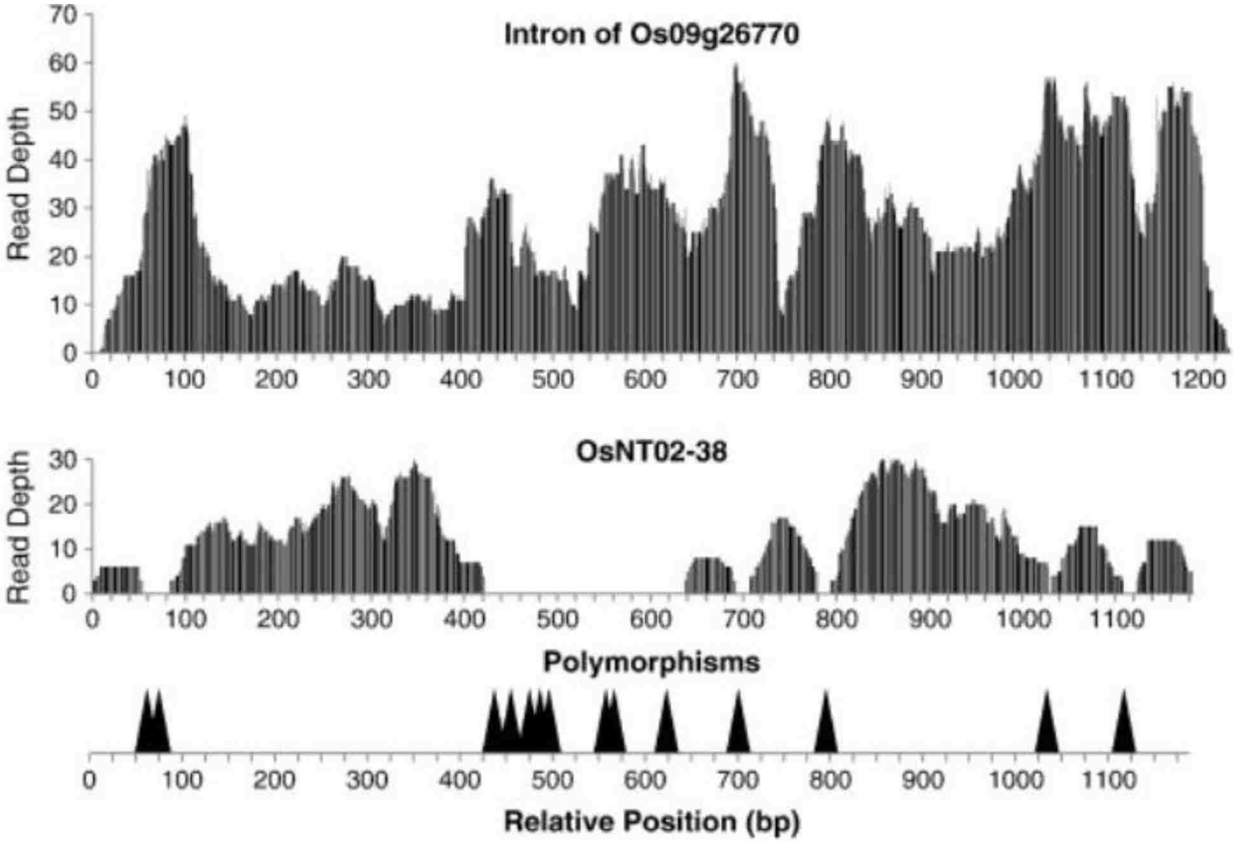


Figure 2.3: Initially identified novel genes shared high similarity to another genomic region and hence may be false positive results.

An intron of Os09g26770 showed high similarity to novel gene OsNT02-38. The black triangles in the bottom panel indicate differences in the DNA sequences between the two transcripts.

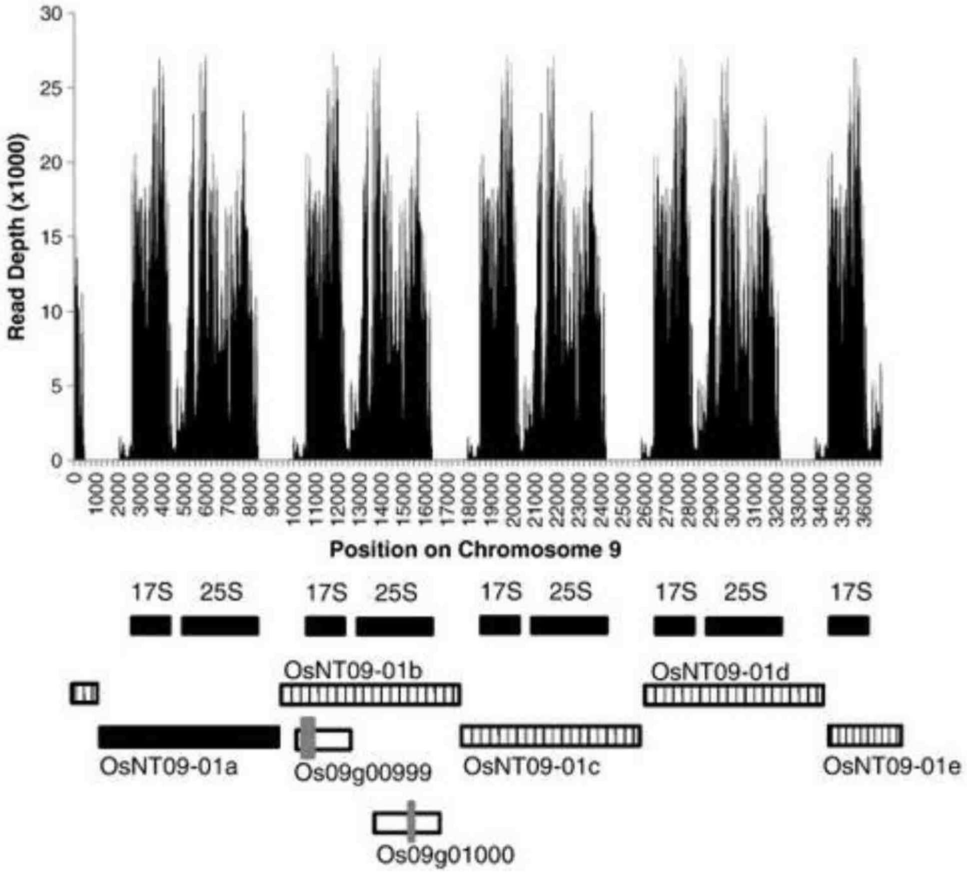


Figure 2.4: The novel algorithm helped eliminate previously annotated genes within the region of rDNA repeats.

OsNT09-01a (solid black bar) is identical to the three tandem upstream regions, labeled OsNT09-01b, OsNT09-01c and OsNT09-01d (vertically lined bars). Each of these regions contains a repeat corresponding to the sequences of the 17S and 25S rRNAs. OsNT09-01e, a partial repeat, contains a sequence corresponding to the 17S rRNA. Two annotated genes, Os09g00999 and Os09g01000, are uncharacterized proteins. Read depth is determined from the cumulative reads of all samples.

The region OsNT09-01b contains two annotated genes, LOC_Os09g00999 and LOC_Os09g01000, both of which are putative expressed genes of uncharacterized proteins (Ouyang et al., 2007). Since these are part of the rRNA tandem repeat area, it is likely that they are not protein coding genes.

OsNT09-01a is also highly similar to a region on chromosome 2. The region we named OsNT02-34 (chr02:28,709,481-28,716,116), has two distinct regions within it (chr02:28,709,481-28,711,172 and chr02:28,712,494-28,716,116) (Figure 2.5). These two regions show 99% and 98% similarity to two distinct regions within OsNT09-01a, chr09:6767-8449 and chr09:1631-5254, respectively. Interestingly, these regions appear to be reversed; the 5′ region of OsNT02-34 shows similarity to the 3′ region of OsNT09-01a and the 3′ region of OsNT02-34 shows similarity to the 5′ region of OsNT09-01a. The black triangles indicate positions where the DNA sequences differ between the two genes. These differences correspond to low read depth in OsNT02-34. This indicates that the reads aligning to OsNT02-34 most likely have arisen from RNA transcribed from the rRNA region of chromosome 9. The similarities between these two regions of the rice genome are indicative of a past gene duplication and rearrangement event.

BLAST searches helped eliminate potential novel genes that shared sequence similarity with other regions

In order to eliminate potential novel genes with high similarity to another genomic region, the novel genes predicted by Cufflinks and the novel algorithm were subjected to a BLAST search against the rice genome. To calculate the percent similarity to another genomic region, the length of the longest BLAST match to the potential novel gene, other than itself, was divided by the length of the potential novel gene.

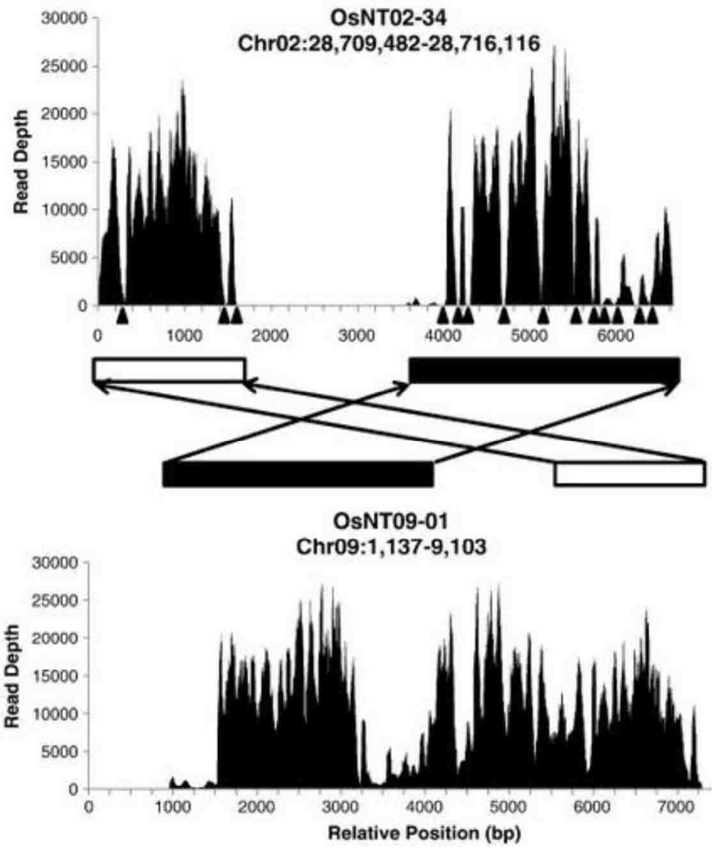


Figure 2.5: The novel algorithm identified a gene duplication and rearrangement event.

Comparison of the sequences of OsNT09-01 and OsNT02-34 showed that the two sequences were highly similar. The region marked by a white bar on OsNT02-34 was found to be 99% identical to the region marked by the white bar on OsNT09-01. The region marked by a black bar on OsNT02-34 was found to be 98% identical to the region similarly marked on OsNT09-01. Black triangles underneath the read depth graph of OsNT02-34 indicate differences in the DNA sequences between the two transcripts. These differences correspond to low read depth in OsNT02-34. Read depth is determined from the cumulative reads of all samples.

BLAST searches of the DNA sequences of the 1,409 non-overlapping potential novel genes identified via Cuffflinks were performed to identify the level of uniqueness of the genes in the rice genome (Figure 2.6A). More than a third of the genes identified (537 genes) showed less than 25% similarity to another genomic region. These 537 genes were considered to be potential novel genes determined by Cufflinks due to their high expression level and uniqueness within the rice genome.

BLAST searches of the DNA sequences of the 356 non-overlapping potential novel genes identified via the novel algorithm were also performed. Of the 356 novel genes, 124 showed less than 25% similarity to another region of the rice genome Figure 2.6B. The 124 genes which showed less than 25% similarity to another region, and passed all of the previous criteria, were considered to be potential novel genes determined by the novel algorithm.

Comparison of genes identified by each method reveals that the novel algorithm in conjunction with Cufflinks increases the reliability of gene detection

When comparing the regions identified by Cufflinks (537 genes) to those identified by the novel algorithm (124 genes), the novel algorithm identified 16 genes that Cufflinks did not identify. In comparison, Cufflinks identified 429 genes that the novel algorithm did not identify. Combining the number of genes identified by both methods gives a total of 553 novel genes (Supplemental Table S2). The size distribution pattern of the novel genes is generally similar to that of the annotated genes. The novel genes ranged from 140 bp to 22,181 bp, with a large proportion of the genes ranging between 0.5 and 1 kb (Supplemental Figure S1). Similarly, there is a large peak in size of annotated genes within this range. As the lengths of the genes increases, their abundance declines in both the annotated and the novel genes.

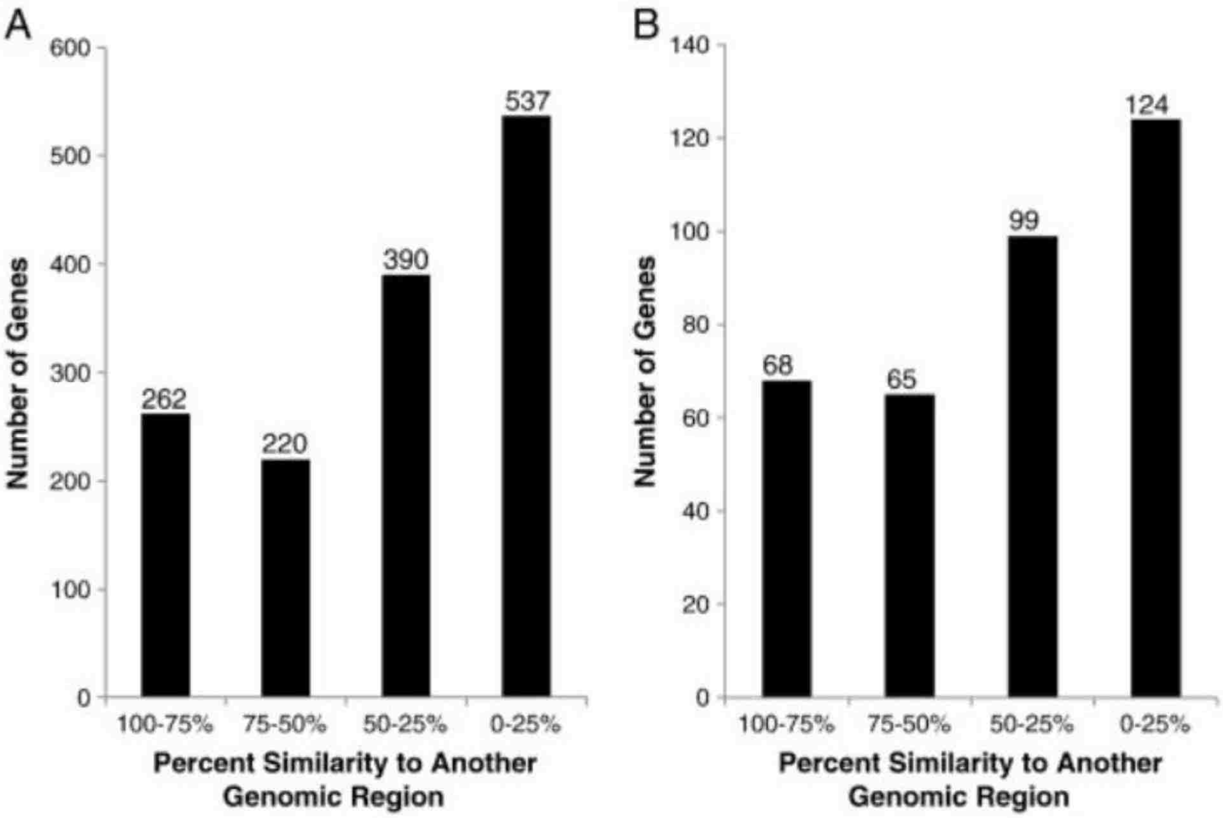


Figure 2.6: BLAST searches of the novel genes identified by Cufflinks and the novel algorithm were used to determine the similarity of the potential novel genes to other regions within the rice genome.

BLAST searches were performed to identify the level of uniqueness of the potential novel genes. Genes with high similarity to another genomic region may be remnants of past gene duplication events. (A) Of the 1409 novel genes identified by Cufflinks, 537 genes showed less than 25% similarity to another genomic region. (B) Of the 356 potential novel genes identified by the novel algorithm, 124 showed less than 25% similarity to another genomic region.

Of the 16 potential novel genes identified by the novel algorithm and not by Cufflinks, 4 genes had a single read junction alignment that associated the potential novel gene to the adjacent gene. Cufflinks may have used this single read alignment to associate the gene with the adjacent gene. Using a single read to link a highly expressed region to neighboring gene may be premature, since a single read may be a result of mapping or sequencing errors or a trace amount of genomic DNA contamination. The remaining 12 genes had no flanking junctions or had a single read junction that did not extend to the adjacent gene. Three of these genes also showed no homology to another region in the rice genome. It is not clear why Cufflinks did not identify these regions as novel genes.

Close examination of 25 of the genes identified by Cufflinks but not by the novel algorithm reveals that the stringent requirements of the novel algorithm were the limiting factors. The novel algorithm is very conservative to avoid false positive gene identification, requiring several criteria: the distance from the nearest annotated gene must be greater than 5% of the unannotated region, the minimum read count must be 200 reads, and there must be no flanking junctions. These criteria greatly limit the number of genes identified by the novel algorithm. One of the 25 genes had a flanking junction and was eliminated by the novel algorithm. It is uncertain why Cufflinks considered this gene a novel gene. Another gene lay adjacent to a highly expressed gene that had read alignments that extended beyond the annotated region. The intergenic region analyzed by the novel algorithm included these reads, and the resulting novel gene was discarded due to overlap with the annotated gene. Four of the genes were within intergenic regions that contained multiple transcripts. The increased standard deviation of these intergenic reads prevented the novel algorithm from considering them. Seven genes had fewer than 200 reads and did not meet the minimum read requirement for the novel algorithm, but due to the small size of the genes, the level of expression calculated by Cufflinks was greater than 1.0 RPKM. The most common factor that prevented the novel algorithm from identifying genes that Cufflinks found was the requirement for a minimal distance to the nearest annotated gene. The remaining twelve genes were too close to the adjacent annotated gene to be considered as a novel gene.

Both the novel algorithm and Cufflinks identified novel genes that were not identified by the other method. The novel algorithm identified fewer genes than Cufflinks due to its stringent requirements. If the requirements of the novel algorithm were adjusted, it may be able to identify more novel genes. On the other hand, these requirements were put in place to prevent false positive results. Further improvement is planned for future versions of the novel algorithm to improve the gene finding capability without increasing false positive results.

RT-PCR and sequencing of a subset of novel genes confirmed the presence of transcripts from the genes

Experimental verification of the novel genes was carried out by performing RT-PCR on a subset of ten genes that were shown to be highly expressed in the control sample. As shown in Figure 2.7, bands with expected sizes were detected for all ten genes. The sequences of the PCR products matched those of the predicted novel genes. In some cases, the size of a PCR product was much smaller than that of the corresponding novel genes (Supplemental Table S2). This is because the length of the predicted gene listed in Supplemental Table S2 includes both exons and introns. Also, there appears to be instances of alternative splicing within the tested novel genes. For example, comparison of RNA-seq data and the sequence of the cloned cDNA indicates that the PCR product for OsNG03-04 represents a short isoform of its transcripts. For the six genes that contain introns,

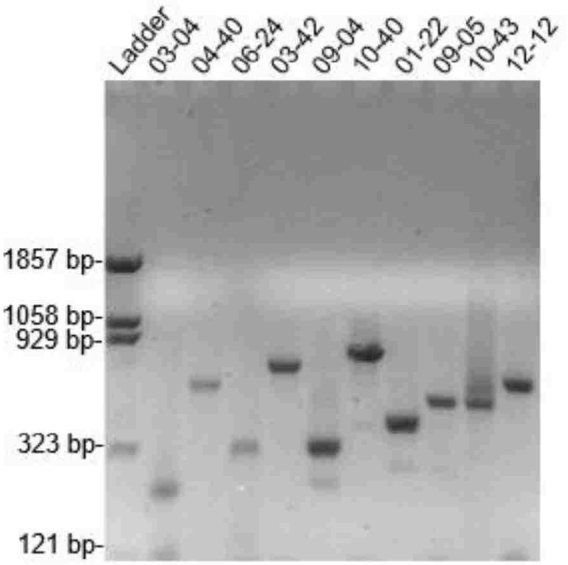


Figure 2.7: RT-PCR of 10 highly expressed novel genes.

All selected novel genes revealed a band that corresponded to the predicted size.

the sequencing confirmed the exon–intron boundaries, and the cDNA sequences corresponded closely with the expression regions revealed by the RNA-seq data. This supports the validity of the novel genes as transcribed units.

Confirmation of published hormone-induced genes shows the reliable quality of our RNA-seq dataset

While the statistics of the RNA-seq data show it to be of high-quality, to verify that the data followed expected hormone induction patterns, we compared the expression patterns of known genes in our dataset to previously published results. We examined the fold change of nine known ABA inducible genes (Ross and Shen, 2006) and eight known GA inducible genes (Chen et al., 2006), and compared the results to published expression levels. All nine ABA-inducible genes show ABA induction in our data set, ranging from 63 to 204 fold induction (Supplemental Table S3). Comparatively, of eight α-amylase genes that had been shown to be induced in rice endosperm 5 days after imbibition, seven were induced by GA in our data set by greater than 2 fold (Supplemental Table S4). The remaining α-amylase gene, *RAmy3A*, had a very low expression level in our data set and an accurate fold change could not be determined. This gene also showed a very low level of expression in the endosperm study. Agreement of the RNA-seq data with the previously published data confirms its reliability for the identification of hormone induced genes.

The majority of the novel genes have predicted protein coding sequences

In order to determine if the novel genes are likely to contain protein coding sequences, BLASTX was used to detect open reading frames that produced similar sequences to those of known proteins in the NCBI non-redundant protein database. A maximum e-value of 1×10^{-4} was used for protein predictions. NCBI suggests an e-value cutoff of 1×10^{-3} for protein BLAST queries, but a more stringent cutoff was used to ensure fewer false positive alignments. For regions that had overlapping exons, steps were taken to select the exon with the highest probability of being the correct exon. Of the 553 novel genes, 302 had at least one predicted amino acid sequence. The predicted amino acid sequences ranged from 21 to 3761 amino acids. This indicates that the majority of the novel genes encode proteins with homology to known proteins. However, this analysis does not exclude the remaining 249 transcripts as non-protein-coding. These transcripts may encode proteins but they have no homology to the proteins in the NCBI non-redundant protein database.

While the functions of the predicted proteins are unknown, the roles played by their homologous protein matches may give a clue to these functions. While GA and ABA are not known to be produced in aleurone cells, the predicted protein products of two novel genes, OsNG10-31 and OsNG10-36 showed homology to proteins that may play a role in ABA or GA biosynthesis. The predicted protein product of OsNG07-08 shows homology to a spliceosome subunit, indicating a potential role in aleurone- or hormone-specific alternative splicing. In addition, there are several homologous proteins that play roles in cellular respiration and the electron transport chain, cell growth, signaling, and transcriptional control.

The novel genes contain probable primary microRNA sequences

Some primary microRNA (pri-miRNA) genes are transcribed and the transcripts are polyadenylated, and can potentially be included with the pull down of polyadenylated mRNAs. To determine if our novel genes include potential primary microRNAs, we performed a BLAST search of the novel genes against the Plant MicroRNA Database (PMRD) (Zhang et al., 2010c), a database of the known plant pri-miRNAs from 121 different species. There were a total of 10,597 plant miRNA sequences in the database, which included 2773 rice miRNAs. We compared our

553 novel genes, via BLAST search, to the PMRD database and found that 124 of the novel genes (22%) showed similarity to at least one known pri-miRNA with an e-value of 1.0×10^{-5} or less. Of the 124 novel genes that showed similarity to a pri-miRNA, 93 of them matched pri-miRNA from *Oryza sativa*. Of these 93 genes, 11 contained a sequence identical to a known pri-miRNA in PMRD, most of which existed within the intronic regions of the novel genes.

Of the 124 novel genes that contained possible pri-miRNAs, 70 transcripts were also predicted to code for proteins. There did not appear to be a preference for the direction of the predicted pri-miRNA in relation to the direction of the predicted protein coding sequence. Of the 70 transcripts, 32 pri-miRNAs were oriented in the same direction as the predicted protein coding sequence and 38 were oriented in the opposite direction. These latter transcripts may code for either a protein or a pri-miRNA. Alternatively, it may be possible that those pri-miRNAs that are oriented in the opposite direction of the protein coding sequence may produce miRNAs targeting the mRNAs produced from the opposite strand. There were 54 transcripts that did not have a reliable predicted protein product but did have a match to a pri-miRNA. Overall, there were 356 transcripts, out of 553, that can code for either a protein or a pri-miRNA or both.

Many of the novel genes show differential expression between hormone-treated samples

While the novel genes were detected by pooling data from all four RNA-seq treatments, many of them showed differential expression between the samples. By counting the number of reads that aligned to a particular gene on the control sample and comparing it to the number of reads that aligned to a hormone treated sample, the level of differential expression can be calculated. Version 6.1 of the MSU japonica rice genome contains 57,624 annotated genes. Our experimental data showed that 18,152 of these genes had a calculated RPKM greater than 1.0. Of these genes, there were 8684 genes that were induced at least 2 fold by one or more of the hormone treatments (Supplemental Figure S2). Of the induced genes, 3893 genes were induced by ABA, 6512 genes were GA induced, and 5552 genes were ABA + GA induced. There were 1928 genes that were induced by all of the hormone treatments. In comparison, 3992 genes were repressed at least 2 fold by one or more of the hormone treatments. Of the repressed genes, 1604 were repressed by ABA, 2142 were repressed by GA and 2840 were repressed by ABA + GA. There were 600 genes that were repressed by all of the hormone treatments.

Many of the novel genes also showed differential expression between hormone treated samples. For example, the putative novel gene at locus OsNG02-37 (chr02:25,530,515-25,532,089) had a modest expression level of 7.2 RPKM in the control sample (Figure 2.8A). This potential novel gene is 90% unique in the rice genome and has a single putative exon. When treated with GA, the expression level of OsNG02-37 increased nearly 10 fold to 70 RPKM (Figure 2.8C). However, when treated with ABA, a modest repression was observed (4.8 RPKM) (Figure 2.8B). When both hormones were applied, the induction level was mediated to 4 fold over the control (28 RPKM) (Figure 2.8D) demonstrating an antagonistic response between ABA and GA.

Not all of the hormone-responsive novel genes showed antagonism between ABA and GA. Novel gene OsNG12-12 (chr12: 6,491,569-6,492,428), identified by both methods, is only 18.8% similarity to another region in the rice genome. This potential novel gene consists of a single exon and was very highly expressed, with an RPKM of 1277 in the control sample (Figure 2.9A). However, when treated with hormones, the expression level of OsNG12-12 severely dropped. The samples treated with ABA and GA showed 2.8 fold and 7.2 fold repression respectively

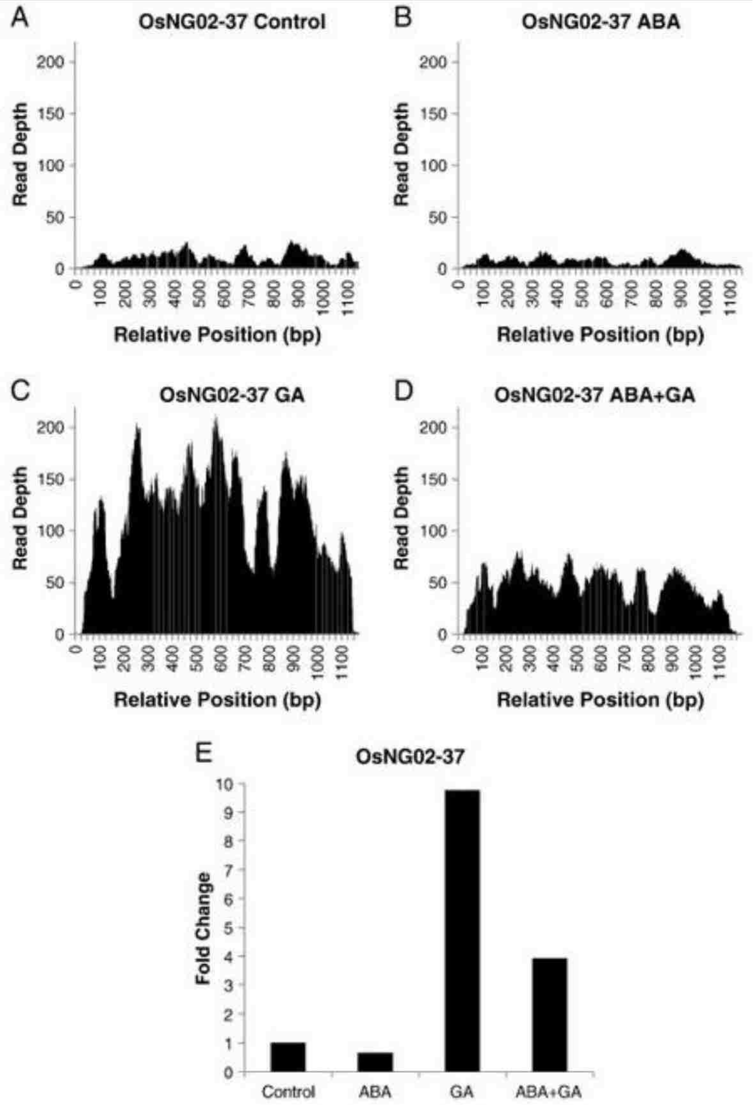


Figure 2.8: Some novel genes showed antagonistic differential expression in response to the hormones ABA and GA.

The potential novel gene OsNG02-37 showed the typical antagonism expected between the ABA treated samples and the GA treated samples. (A) No hormone treatment. (B) ABA treatment. (C) GA treatment. (D) Combined ABA and GA treatment. (E) Fold change of each treatment, normalized to the total reads of the sample.

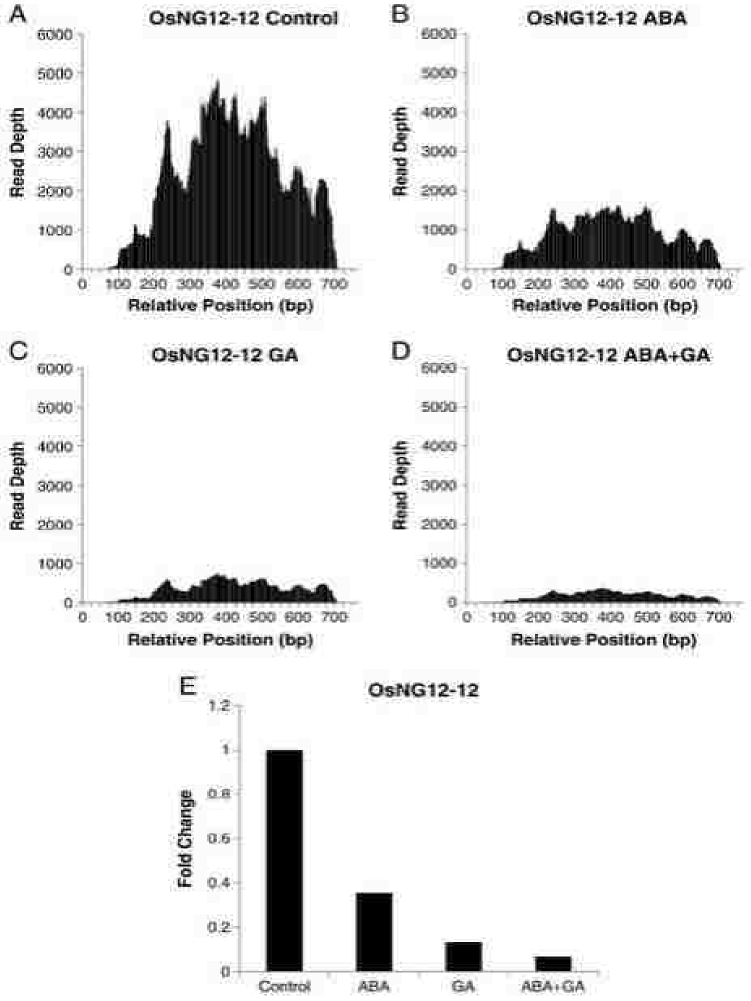


Figure 2.9: Some novel genes showed cumulative differential expression in response to the hormones ABA and GA.

Novel gene OsNG12-12 showed repression when treated with ABA or GA alone. When treated with both hormones simultaneously, the repression was cumulative. (A) No hormone treatment. (B) ABA treatment. (C) GA treatment. (D) Combined ABA and GA treatment. (E) The fold change of each treatment, normalized to the total reads of the sample.

(Figure 2.9B and Figure 2.9C). When both hormones were applied simultaneously, repression was cumulative, resulting in a 14.4 fold repression (Figure 2.9D). This is unusual, since ABA is well known for its antagonistic effects on GA signaling.

Out of 553 putative novel genes identified, 440 transcripts (80%) showed at least two fold induction or repression in at least one hormone treatment. Of these genes, 273 (49%) were induced two fold or greater by at least one of the hormone treatments and 67 showed induction in all three hormone treatments (Figure 2.10A). On the other hand, 209 (38%) of the novel genes showed two fold or greater repression in at least one of the hormone treatments (Figure 2.10B). There were 50 genes that were repressed in all three hormone treatments. While our RNA-seq data showed that many more annotated genes were induced than repressed, the novel genes showed a much less drastic increase in induced genes over repressed. This is surprising, as repressed genes are less likely to be detected by either gene finding algorithm due to lower pooled read count.

To confirm that these novel genes are indeed transcribed, we randomly selected 10 highly expressed novel genes and performed reverse-transcriptase polymerase chain reaction (RT-PCR) (Figure 2.7). The PCR products were run on a 3% agarose gel and all 10 novel genes produced a band. The locations of each band corresponded exactly to the predicted band location.

The PCR product of each of these 10 genes were sequenced at the UNLV Genomics Core facility and the sequence was compared to the predicted DNA sequence of the corresponding novel gene. All 10 novel genes matched identically in sequence.

Discussion

To detect novel genes potentially involved in the response of rice aleurone to the hormones ABA and GA, we performed RNA-seq analysis on data from control and hormone-treated RNA samples.

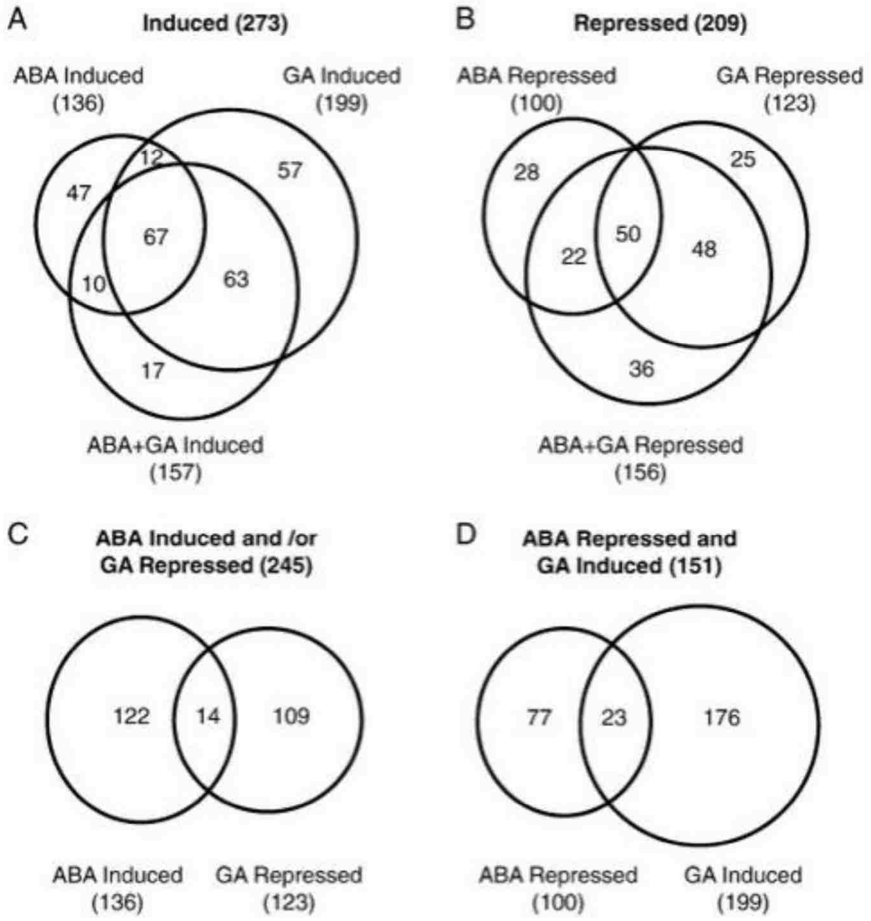


Figure 2.10: Venn diagrams showing overlap of induced and repressed novel genes.

Out of the 553 novel genes identified, (A) 273 were induced by at least one hormone treatment and (B) 209 were repressed by at least one hormone treatment. (C) Induction by ABA and/or repression by GA was seen in 245 novel genes. (D) Repression by ABA and/or induction by GA was seen in 151 novel genes. Only transcripts that had at least two fold induction or repression, compared to the control treatment, were used in constructing this figure.

Gene annotation was performed using Cufflinks and a novel algorithm. Cufflinks, in combination with Tophat, has been shown to give more accurate transcript annotation than other software packages (Palmieri et al., 2012; Trapnell et al., 2010). Despite the demonstrated usefulness of Cufflinks for novel gene annotation, large numbers of reads aligned to several unannotated regions of the rice genome that were not identified as novel genes by Cufflinks. Thus, a novel gene detection algorithm was developed. This novel algorithm was used to detect genes within unannotated regions based on the standard deviation of the read distribution within the unannotated region, the distance of the endpoints of the standard deviations from an annotated gene, and the lack of spanning reads mapped by Tophat. The novel genes identified through the two methods showed a high level of expression (RPKM ≥ 1.0), eliminating the possibility that the reads aligned to a region due to mapping or sequencing errors or a trace amount of genomic DNA contamination. The novel genes did not overlap with currently annotated genes and thus were not likely to be alternative splice variants of existing genes. In order to verify that they are not alternative exons of neighboring genes that failed to accumulate Tophat junction-alignments, we compared the hormone expression patterns of the novel genes with their nearest neighbor in all four treatments. A small number of the neighboring genes (58 genes) were induced or repressed under the same hormone treatments as the adjacent novel gene. However, the majority of the neighboring genes had either differing expression patterns (229 genes) or no detectable level of expression (259 genes). The remaining 7 novel genes did not have reads aligned to the control, however, all of the neighboring genes did have reads in the control sample, indicating a difference in expression patterns.

The novel genes showed less than 25% similarity to another genomic region, as determined by BLAST searches, eliminating the possibility of false positive results due to similarity to an actively expressed gene. Combining both methods of novel gene identification with BLAST search analysis greatly reduced the possibility of false positive results. BLAST can be used to determine the level of uniqueness of a transcript within the genome. For regions that are not unique due to past gene duplication and rearrangement events, current alignment software is incapable of determining the correct region where the reads originated, leading to false positive identification of novel genes. Gene duplication events may lead to inactive genes or pseudo genes. Pseudo genes may be transcribed to produce mRNA but may not be translated into functional proteins. However, for regions that are unique in the genome and have been identified as potentially novel genes, it is difficult to argue that these regions are not transcribed. While we discounted those possible novel genes that showed high similarity to another area of the rice genome, it is possible that some of these are still valid expressed genes.

Identification of these novel genes presents a question as to why they have not been annotated previously. First of all, many genes of the rice genome have been detected by extracting and sequencing full length cDNA from various tissues (Rice Full-Length c et al., 2003). However, not all genes may be expressed at a detectable level in a given tissue, and some genes may only be detected in a given tissue under a specific condition. This leaves the possibility that many genes may go undiscovered. Novel gene detection in the aleurone layer of rice treated with germination regulating hormones has not been performed prior to this study. Because of this, we were able to identify many new genes. Second, software programs using gene models have also been used to identify genes in the rice genome. Various gene finding software programs, such as Eukaryotic GeneMark (Lomsadze et al., 2005), GENESCAN 1.0 (Burge and Karlin, 1997) and FGENESH 2.0 (Salamov and Solovyev, 2000), have been used to identify transcripts within a genome. These programs identify patterns in the known genes and use these patterns to identify genes in unannotated regions. However, genes that don't follow traditional gene models may elude detection by these programs. For example, the novel gene OSNT03-06 (chr03:4,346,095-4,347,135) is highly unique in the rice genome and was identified by both the Cufflinks software and the novel algorithm. Visual inspection of this transcript showed an intron between two exons, based on read depth and the presence of an intron branching sequence (Supplemental Fig. S7). The sequence of the gene was entered into the gene prediction programs mentioned above to determine if a predicted gene could be found. Only GeneMark.hmm predicted a gene within this region, however, the predicted mRNA consisted of two small exons. In addition, these exons did not coincide with the RNA-seq data, and the direction of the predicted mRNA was opposite that of the intron predicted from the branching sequence. Analysis of the highly expressed novel genes OsNT01-37, OsNT06-31, and OsNT06-34 showed similar results. There was no agreement between the deep sequencing results and the gene prediction programs.

Previous works have used RNA-seq to find novel rice genes using different methods. Zhang et al. sequenced the mRNA of eight organs in the 93-11 cultivar of the indica rice subspecies and found 7,232 novel transcriptional units (Zhang et al., 2010a). Lu et al. performed RNA-seq on two-week old seedlings of the Nipponbare japonica as well as the Guangluai-4 and 93-11 indica rice cultivars (Lu et al., 2010). Their analysis found 14,300 novel Transcriptionally Active Regions (nTARs) in japonica rice, 9043 of which were comprised of a single fragment. Fragments consisted of paired-end reads, 42 or 75nt in length, with continuous mapping and read depth of at least five times per base. Comparison of the nTAR fragment sequences to the sequences of the novel genes discussed in this paper revealed that, of the 553 novel genes, 315 contained sequences similar to those of at least one nTAR. Of these, most nTARs were much smaller than the novel gene; nearly half were less than 25% of the size of the novel gene. The proportion of novel genes that overlapped with nTARs and contained predicted protein or miRNA sequences, compared to all overlapping novel genes, was similar to the proportion of all novel genes that contained predicted protein miRNA sequences. Approximately 60% of the genes overlapping with nTARs contained predicted protein sequences, and approximately 24% contained sequences similar to those of pri-miRNAs.

During alignment of reads to the rice genome, Lu et al. ignored reads that had the capability of mapping to more than one position. While this removed the possibility that genomic regions might erroneously be assigned reads due to sequence similarity to an expressed region, it also may have led to an underrepresentation of expression in regions of genes that are similar to other genes, such as in highly conserved protein domain-coding sequences. As such, it is unsurprising that many of the nTARs are much smaller than the novel genes that are represented in this paper. Nonetheless, discovery of at least portions of these genes through multiple methods from independent sources supports the presence of transcriptionally active units within these areas.

In addition to potential protein products, many of the novel genes contained matched pri-miRNAs from PMRD. The ratio of pri-miRNAs to known genes in rice is 4.8%. Thus, the novel genes contain a larger proportion of possible microRNAs than the annotated genes in the rice genome. While the roles of the novel genes are unknown, insight into these potential roles may be gained by investigating the targets of the predicted pri-miRNAs and the roles of proteins partially homologous to the predicted protein products. Overrepresented ontologies of the predicted pri-miRNA targets include transcriptional regulation, cell signaling, temperature stress response, defense response, and apoptosis. Partially homologous proteins include those that play roles in ABA or GA biosynthesis and metabolism, alternative splicing, cellular respiration, transcriptional control, including control of transcription in relation to energy availability, cell growth, and signaling. Together, these hint at roles of the novel genes in gene regulation, post-transcriptional and post-translational modification, growth, and ABA and GA signaling.

The hormones ABA and GA are extensively studied, and their antagonism is well documented during many developmental stages. Many of the novel genes did show antagonistic responses, as shown in Figure 2.8, however, we discovered that many of these genes had cumulative induction or repression levels upon treatment of a combination of ABA and GA, as shown in Figure 2.9. While this goes against the conventional wisdom of ABA-GA antagonism, it has been shown that there are proteins, such as WRKY24, that repress the signaling pathways of both hormones (Zhang et al., 2009). Of the 553 putative novel genes identified, 273 of the potential novel genes are induced two fold or greater by at least one of the hormone treatments and 209 of the novel genes showed two fold or greater repression in at least one of the hormone treatments. There were 67 genes induced by all hormone treatments and 50 genes repressed by all hormone treatments (Figure 2.10).

Materials and methods

RNA preparation and RNA-seq data analysis

Rice (*O. sativa* L. ssp. *japonica*, cv. Nipponbare) seeds were obtained from the USDA ARS, Dale Bumpers National Rice Research Center. Rice seeds were cut to remove the embryo. The remaining half-seeds were treated with a 10% α-amylase (Tokyo Chemical Industry Co) solution overnight to aid in the removal of the starch and isolate the aleurone layer. The resulting aleurone cells were treated for 4 h with 20 μM ABA, 1 μM GA, or a combination of 20 μM ABA and 1 μM GA. A mock treatment was used as a control. The rice aleurone was then flash frozen with liquid nitrogen and the total RNA was extracted via guanidium sulfate phenol chloroform extraction

(Chomczynski and Sacchi, 2006). The RNA was then treated with Ambion's TURBO DNA-*free*™ DNase to remove any genomic DNA contamination. The TruSeq™ RNA Sample Preparation kit v2 was used to prepare the cDNA libraries. In brief, poly-T coated magnetic beads were used to capture mature polyadenylated mRNA. The mRNA was then fragmented using divalent cations under elevated temperatures. The fragmented mRNA was reverse-transcribed to cDNA using random primers and then amplified by PCR. The cDNA was then sequenced on an Illumina HiSeq 2000 system at the Huntsman Cancer Institute, University of Utah. The sequencing parameter of 50 base pair single-end was used. The Tophat (Trapnell et al., 2009) and Bowtie (Langmead et al., 2009) software were used to align the short reads to the MSU rice genome Version 6.1 (Ouyang et al., 2007).

Analysis for tRNA and rRNA contamination

The mRNA molecules were pulled-down from the total RNA by poly-T-coated magnetic beads, thus eliminating rRNA and tRNA. However, it is possible that some rRNA and tRNA molecules were pulled down along with the mRNAs since they may have stretches of poly-A or may adhere to the mRNA. To determine the level of tRNA contamination, the number of reads aligning to the tRNA genes was counted. Out of 323 tRNA sequences within the rice genome, 297 tRNA sequences had zero or one read aligning to them. Of the remaining 26 tRNA sequences, six tRNA sequences contained > 10 reads. Of these, five were identical to a region of an expressed annotated protein coding gene, and the one remaining tRNA sequence showed 93% similarity to an annotated gene. The reads aligning to these tRNAs are likely derived from these genes, rather than the tRNAs themselves. The small number of reads aligning to the remaining tRNAs support a lack of tRNA contamination in our RNA-seq data.

For a given RNA sample, over 80% of the total RNA is rRNA, thus the retention and sequencing of even a small percentage of rRNA might represent a large proportion of the RNA sequenced. The genes for rRNA exist as an array of tandem repeats, and in plants there may be 150 to 26,000 repeats (Richard et al., 2008). BLAST analysis shows that chromosome nine contains the coding regions for the 17S (Takaiwa et al., 1984) and 25S (Takaiwa et al., 1985) rRNA genes. There is a region on chromosome two that also shows high homology to the 17S and 25S rRNA genes. Counting the reads that map to these regions reveals that about 6.8 million (5.2%) of the 131.6 million mapped reads align to these regions.

Ribosomal RNA is not poly-adenylated and therefore should not be present in the samples. However, because there was such a high concentration of reads aligning to rDNA in the extracted sample, it appears that some rRNA was retained during the RNA-sequencing. This may be due to partial binding of the adenine rich regions of the rRNA to the oligo-T used to pull down the mRNA, or the rRNA partially adhering to the poly-adenylated mRNA fragments. The 17S and 25S rRNA regions show very low homology to the rest of the genome. The 6.0 million reads that aligned to the rRNA region of chromosome nine were remapped to annotated protein coding genes. Only 32 reads of the 6.0 million reads mapped to an annotated protein-coding gene. Therefore, though there is substantial rRNA in the samples, the effect on calculation of differential gene expression and novel gene detection is minimal.

Analysis for possible genomic DNA contamination

While the RNA samples used for RNA-seq were treated with DNAse to prevent DNA contamination, it is arguable that the presence of reads in unannotated regions may be due to such contamination. To rule out this possibility, we examined the reads aligning to the rice chloroplast genome and compared them to the reads aligning to the rice mitochondrial genome. Since aleurone cells are non-photosynthetic, they do not contain the high densities of chloroplasts one might expect from a photosynthetic cell, such as a leaf cell. However, they do contain plastids in the form of leucoplasts , which also contain the chloroplast genome. Leucoplast gene expression is at a relatively low rate in rice aleurone (Bethke et al., 2006). In addition, RNAs produced by the leucoplasts are not expected to be picked up in our RNA-seq experiment. While evidence indicates that plastid RNAs undergo post-transcriptional processing (Asano et al., 2013), most of these RNAs are not polyadenylated (Stern et al., 2010). Thus, DNA contamination in our samples should show up in the presence of large numbers of reads preferentially aligning to the chloroplast genome. Of the 33,889,264 reads that mapped to the genome, only 323 reads mapped to the chloroplast genome. This represents a ratio of about 1×10^{-5} chloroplast reads per read mapped. Thus, we conclude that genomic DNA contamination does not explain the large proportion of reads aligning to the unannotated regions of the rice genome.

Detection of novel genes through Cufflinks software

Cufflinks software (Trapnell et al., 2010) was used to assemble the reads into transcripts. The transcripts were compared to annotated transcripts to identify transcripts that did not map to current annotations. Transcripts with an RPKM expression value less than one were not considered due to potential sequencing error and statistical mapping error or a trace amount of genomic DNA contamination (Mao et al., 2012).

Development of a novel algorithm to further detect novel genes

An alternative threshold analysis was used to identify potential novel genes to compare to the Tuxedo suite software. PERL scripts were written to identify clusters of reads that mapped to unannotated regions of the rice genome. The script calculated the mean and standard deviation of the midpoint of the reads. For unannotated regions containing at least 200 reads, if the region that included the mean ± 2 standard deviations was completely within the unannotated region, and the distance from the nearest gene was greater than the threshold, the program selected the region as a possible novel gene (Figure 2.2: A). The threshold was the larger of 5% of the size of the unannotated region or 100 bp. This 5% threshold was chosen based on observation. A threshold less than 5% tends to select regions that are part of adjacent annotations. A threshold greater than 5% eliminates many potentially novel genes.

Elimination of potential false positive results from both gene detection methods

Potential novel genes detected by both methods were eliminated if the expression of the gene was below 1.0 RPKM. Previous research shows that cumulative genome coverage plateaus at 100 million reads (Mizuno et al., 2010), and that additional reads do not significantly increase the genome coverage. To calculate the expression level of genes within 5% accuracy for genes with low expression (3 to 30 RPKM), about 80 million reads are sufficient (Mizuno et al., 2010); therefore, 132 million reads is sufficient for accurate determination of the expression level of genes to as low as 3.0 RPKM. However, the expression level necessary to detect the presence of a gene can be lower than 3.0 RPKM. Previous studies have used the minimum expression level of 1.0 RPKM for gene detection (Lu et al., 2013b; Mao et al., 2012). Expression levels below 1.0 RPKM may be affected by sequencing error, statistical mapping error, or a trace amount of genomic DNA contamination (Mao et al., 2012). Therefore, we used a minimum expression level of 1.0 RPKM for novel gene identification.

The potential novel genes identified both by the Tuxedo suite software and the novel algorithm were compared to the rice genome using BLAST (Altschul et al., 1997) to determine the similarity of the novel genes to other regions of the rice genome. The following parameters were used in the BLAST search: maximum expectation value (e-value) 1.0×10^{-5}, word size 11, no filtering. As occurrences of "TATA" repeat sequences are overabundant in the rice genome, sequences of four or more tandem "TATA" repeats were removed from the novel genes and replaced by N's prior to performing the BLAST search. The UCSC Genome browser was used to visualize the results of the RNA read alignment (Kent et al., 2002).

RT-PCR and sequencing of a subset of novel genes

Fifteen novel genes were selected for RT-PCR verification based on a high level of expression in the control treatment. Primers were designed from unique regions of the expressed region of the predicted novel genes (Supplemental Table S5). Where introns were present, the primers were designed to span at least one intron. RNA was prepared from rice aleurone prepared in the same manner as that of the RNA produced for RNA-seq. The Qiagen Omniscript RT kit was used to reverse transcribe the RNA, according to the manufacturer's protocol. The novel genes were amplified from the resulting complementary DNA using KlenTaq LA, using the recommended typical PCR reaction mixture, with the optional addition of betaine. The cDNA was amplified using the following thermocycle: denaturing at 95 °C for 8 min, followed by 40 cycles of denaturing at 95 °C for 30 s, annealing at 68 °C for 40 s, decreasing by 2 °C per cycle for the first five cycles, and amplifying at 68 °C for 2 min. Prior to sequencing, the amplified DNA was purified using the Qiagen MinElute PCR purification kit, using the manufacturer's protocol. The purified DNA was sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), followed by capillary electrophoresis on an ABI 3130 Genetic Analyzer (Applied Biosystems). The sequencing results were compared to the predicted novel gene sequence via BLAT.

Determination of potential protein coding sequences within novel genes

To determine the potential protein sequences of the novel genes, the DNA sequences were compared to the NCBI non-redundant protein database via the BLASTX search tool. BLASTX is a BLAST program that compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database. The protein sequence with the highest percent match and lowest e-value was considered as the most likely protein product. Exons that had fewer than 10 reads or e values greater than 1.0×10^{-4} were also excluded. If an exon was completely contained within another exon, the smaller exon was excluded to give better coverage of the novel gene. If exons overlapped but one exon was not completely contained within another, then the exon with the higher e-value was excluded. The exon with the lower e-value has the greater probability of being the correct translated sequence. If both exons had low e-values ($< 1.0 \times 10^{-4}$), the exon with more reads was selected.

MicroRNA analysis of novel genes

The primary microRNA database in FASTA format was downloaded from the Plant MicroRNA Database (Zhang et al., 2010c). This database included all the experimentally confirmed and computationally predicted microRNAs from 121 different plant species. The 537 potential novel genes identified by Cufflinks were then subject to a BLAST search against the PMRD database and matches with an e-value less than 1.0×10^{-5} were accepted.

The psRNATarget: A Plant Small RNA Target Analysis Server (Dai and Zhao, 2011), was used to identify the target genes of the novel miRNA genes. Target search was performed on the MSU Rice Genome Annotation, version 7. The target genes were queried against the Rice Oligonucleotide Array Database to determine their gene ontology category.

Differential expression calculations

Because each sample had a different number of mapped reads, the differential expression calculation was normalized based on the number of mapped reads in the sample. For example, the ABA fold change for novel gene i was calculated via the following equation:

$F_i=(A_i/C_i)\times(C_r/A_r)Fi=Ai/Ci\times Cr/Ar$

where: F_i is the ABA fold change for novel gene i; A_i is the number of reads that aligned to novel gene i on the ABA treated sample; C_i is the number of reads that aligned to novel gene i on the control sample; C_r is the total number of reads in the control sample that mapped to the genome (31,186,982 reads); A_r is the total number of reads in the ABA treated sample that mapped to the genome (32,961,116 reads).

CHAPTER 3

TILING ASSEMBLY ALGORITHM WAS CREATED TO IDENTIFY NOVEL GENES IN THE RICE GENOME

Previously published as:

Tiling Assembly: a new tool for reference annotation-independent transcript assembly and novel gene identification by RNA-sequencing.

Kenneth A. Watanabe, Arielle Homayouni, Tara Tufano, Jennifer Lopez, Patricia Ringler, Paul Rushton, Qingxi J. Shen.

DNA Res. 2015 Oct;22(5):319-29. doi: 10.1093/dnares/dsv015.

Disclaimer:

KAW participated in RNA extraction, performed the sequence alignment, made most of the figures and composed most of the manuscript. AH, TT and JL participated in the data analysis and reviewing the manuscript. PR participated in RNA extraction, composed some of the abstract, introduction and discussion and participated in reviewing the manuscript. LKG participated in RNA extraction and experiments. PR participated in reviewing the manuscript. JQS conceived the study and supervised the design and coordination of the experiments. All authors read and approved the final manuscript.

Abstract

Annotation of the rice (*Oryza sativa*) genome has evolved significantly since release of its draft sequence, but it is far from complete. Several published transcript assembly programs were tested on RNA-sequencing (RNA-seq) data to determine their effectiveness in identifying novel genes to improve the rice genome annotation. Cufflinks, a popular assembly software, did not identify all transcripts suggested by the RNA-seq data. Other assembly software were CPU intensive, lacked documentation, or lacked software updates. To overcome these shortcomings, a heuristic *ab initio* transcript assembly algorithm, Tiling Assembly, was developed to identify genes based on short read and junction alignment. Tiling Assembly was compared with Cufflinks to evaluate its gene-finding capabilities. Additionally, a pipeline was developed to eliminate false-positive gene identification due to noise or repetitive regions in the genome. By combining Tiling Assembly and Cufflinks, 767 unannotated genes were identified in the rice genome, demonstrating that combining both programs proved highly efficient for novel gene identification. We also demonstrated that Tiling Assembly can accurately determine transcription start sites by comparing the Tiling Assembly genes with their corresponding full-length cDNA. We applied our pipeline to additional organisms and identified numerous unannotated genes, demonstrating that Tiling Assembly is an organism-independent tool for genome annotation.

Introduction

RNA-sequencing (RNA-seq) technology enables whole-transcriptome profiling via the collection and mapping of short cDNA fragments (reads) to a reference genome (Nagalakshmi et al., 2010; Wang et al., 2009b). Regions of the genome where many reads align indicate such regions are highly expressed (Oshlack et al., 2010). Regions where no known gene has been annotated and a large number of reads align are indicative of an undiscovered gene (Bertone et al., 2004).

Reconstruction of transcripts can be obtained through a variety of computational strategies, each of which has its own benefits and drawbacks. These assembly algorithms fall into two general classes, *ab initio* assembly and *de novo* assembly. *Ab initio*, mapping-first approaches, rely on the availability of a reference genome to which the short reads can be aligned (Grabherr et al., 2011). The major drawback of this method stems from the dependency of accurate transcript identification on the presence of a high-quality reference genome (Steijger et al., 2013). The main benefit of *ab initio* assembly is the maximum sensitivity exhibited for gene identification (Grabherr et al., 2011; Yandell and Ence, 2012), though higher sensitivity tends to result in a lower accuracy due to a higher number of false positive genes being reported (Yandell and Ence, 2012). On the other hand, *de novo* transcript assembly, or assembly-first approaches, are independent of a reference genome and directly determine the transcripts of a genome through the short reads (Grabherr et al., 2011). Since this avenue is dependent solely on short read data, genes with low read coverage due to low expression levels can result in inaccurate determination of full-length transcripts (Lu et al., 2013a). The main benefit provided by *de novo* assembly methods is the lack of reliance on a reference genome, which makes it a vital tool for gene identification in organisms that lack a reference genome. Thus, both genome biology as well as the availability of an accurate and complete genome play major factors in the decision to use *ab initio* assembly versus *de novo* assembly (Yandell and Ence, 2012).

Recently, we published a clustering algorithm to identify novel protein- and microRNA-coding genes by searching only the unannotated regions of the rice genome (Watanabe et al., 2014). However, this method relied on the genome being partially annotated. Analysis of the RNA-seq data with other transcript assembly software revealed that they failed to identify genes, were CPU intensive, or lacked documentation or recent software updates. Here, we report an improved algorithm for *ab initio* transcript assembly and novel gene identification, Tiling Assembly, to compensate for such shortfalls.

By combining Tiling Assembly with Cufflinks (Langmead et al., 2009), thousands of potential novel genes were identified in the rice genome. To reduce the possibility of false-positive novel gene identification, stringent filters on minimum gene length, minimum gene expression level, and percent similarity of the potential novel genes to another region in the genome were included. Utilizing this pipeline, 767 high-confidence, unannotated genes were identified in the rice genome. By applying Tiling Assembly to other model organisms, we identified 200 potential novel genes in *Arabidopsis thaliana*, 126 in *Caenorhabditis elegans*, 361 in *Drosophila melanogaster*, and 460 in *Saccharomyces cerevisiae*. This study demonstrated that, by utilizing Tiling Assembly, many potential novel genes can be identified in even the most well-annotated genomes.

Short Read Alignment

The short-read data from our previous publication were used in this study since we obtained the data and can vouch for its quality (Watanabe et al., 2014). In addition, using the same data allowed for comparison between our previously published results and the results in this study. The short reads were aligned to MSU R7 using the latest available version of Bowtie and Tophat. Of the 157,773,782 reads, 151,492,182 reads (96.0%) were aligned to the rice genome. Of the reads that aligned, over 10% aligned to currently unannotated regions, indicating that there are potentially many unidentified novel genes in the rice genome, similar to what was found for the human genome (Bertone et al., 2004).

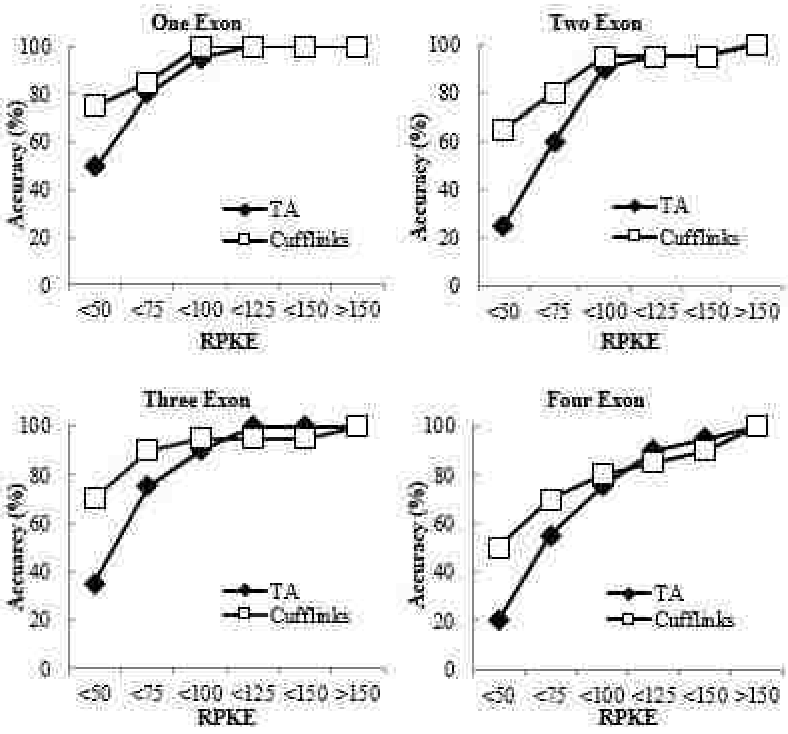
Though the focus on this study is on transcript assembly rather than mapping, we compared Tophat with the splice-sensitive mapping tool OLego (Wu et al., 2013). Both programs identified junctions that the other did not, indicating that neither program was 100% accurate (Supplemental Figure S7). Of the 158,314 junctions identified by OLego, 124,594 junctions (78.7%) matched identically with a junction identified by Tophat. Although OLego identified more junctions, most of the junctions uniquely identified by OLego (71.3%) were determined from a single read. Since nearly 80% of the OLego junctions matched identically with Tophat and most of the remaining were derived from a single read, we chose to keep within the Tuxedo suite software since Cufflinks was designed to work with Bowtie and Tophat. OLego may be used as an alternative mapping tool if so desired.

Benchmark Analysis of the Tiling Assembly Algorithm

To determine the minimum level of read alignment required for reliable detection of transcripts, 80 highly expressed single-isoform genes with evenly distributed reads across the exons were inserted into a randomly generated test genome. A control sample containing about 40 million reads was aligned to the test genome and not a single read mapped. This demonstrated that the test genome was an ideal method for excluding noise. Single isoform genes were used to simplify the problem of correctly identifying all exons, while also allowing us to accurately determine the baseline for exon detection. Four categories were used, containing 20 genes each: one-exon, two-exons, three-exons, and four-exons. Tophat was used to align RNA-seq data from the control sample to the genes. Prior to analyzing its ability to identify exons at a specific expression level, Tiling Assembly was run on the data to ensure the genes were properly identified as single isoform genes with the appropriate number of exons. The minimum expression level used by Tiling Assembly to detect exons for these experiments was set to 50 RPKE, based on our analysis which we will discuss in the following section, which indicated that expression levels below this threshold led to the identification of noise reads as exons. To test the ability of Tiling Assembly to correctly identify the genes at specific expression levels, 1,000 reads were randomly selected from the pool of reads that aligned to each gene and run through Tiling Assembly. The number of randomly selected reads was incrementally reduced, and Tiling Assembly was run again. This process was repeated several times until 50 reads per gene were selected. To compare the performance of Tiling Assembly to a well-established assembly algorithm, the same procedure was performed using Cufflinks. The accuracy with which Tiling Assembly and Cufflinks identified the genes was determined using a point of first failure method, which assumed that the highest RPKE where a gene was first misidentified was the point where identification becomes unreliable. As can be seen in Figure 3.1, at an expression level of 100 RPKE, both Tiling Assembly and Cufflinks were able to identify single-exon and multi-exon genes with an 85% or higher accuracy rate. At a read depth lower than 100 RPKE the accuracy rate dropped dramatically for both Cufflinks and Tiling Assembly, though Cufflinks' accuracy declined at a slower rate. These data indicate that Cufflinks and Tiling Assembly can accurately identify gene transcripts at a read depth of 100 RPKE or greater.

Application of Tiling Assembly Predicted 40,491 Genes Expressed in Rice Aleurone Cells

To determine the gene identification capabilities of Tiling Assembly in an actual genome, Tiling Assembly was applied to rice aleurone short-read data composited from four samples and aligned to MSU R7. Prior to application of Tiling Assembly to the rice genome, it was necessary to determine the minimum gene expression required for accurate detection of exons. This was determined through analysis of several genes, with *LOC_Os01g01010* used as a representative example. Tiling Assembly was run multiple times on these genes, with varied expression thresholds, to determine at what point the exons were accurately recognized. The threshold was incrementally reduced from 100 RPKE. At 50 RPKE, Tiling Assembly identified exons e3 and e4


One Exon

Two Exon

Three Exon

Four Exon

Figure 3.1: Tiling Assembly and Cufflinks show similar accuracy predicting known genes inserted in a random test genome.

A random test genome containing 80 highly expressed, single isoform genes was created. Four categories were used, containing 20 genes each: A) one-exon, B) two-exons, C) three-exons, and D) four-exons. The number of reads aligned to each gene was varied to determine the read depth at which each program failed to accurately predict the genes. If the program identifies the gene with the correct number of exons with the correct exon lengths, then the identification is considered accurate.

of *LOC_Os01g01010* (Supplemental Figure S8); however, lower expression thresholds resulted in identification of false exons due to noise read alignment. A threshold of 50 RPKE was thus used to ensure accurate exon identification.

Exons were identified from MSU R7 using several steps. First, identification of exons was achieved through analysis of overlapping reads aligning to the same region of the genome, with minimum expression of 50 RPKE required for these regions to be considered an exon. This step led to an initial identification of 207,908 potential exons.

Identification of exons via tiling of read alignments limits the size of the exons found to the length of a single read. Tiling Assembly gets around this limitation via analysis of junction alignments produced by Tophat, as described in Section 2.1. In this second step of exon identification, an additional 1,397 potential exons were identified, bringing the total number to 209,305 potential exons.

When a gene has low read coverage, there may be exons which contain gaps in read alignment. Since Tiling Assembly depends on overlapping reads to identify exons, such gaps lead to artificial fragmenting of exons. To avoid this issue, exons that were within 50 nt of each other, a space which could be closed by the length of a single read in our data set, were merged together. To determine whether this merging of exons was likely to result in erroneous merging of genes, the annotated genes in MSU R7 were investigated. It was found that only 0.36% of the non-overlapping annotated genes reside within 50 nt of one another, so linking exons 50 nt apart should not cause a significant misrepresentation of the number of genes obtained from the overall RNA-seq analysis. However, linking closely spaced exons together may result in false merging of exons of the same gene. Of the 209,305 potential exons found in our RNA-seq data, 50,895 fragments were within 50 nt of each other. After linking these closely spaced fragments together, there were 158,410 potential exons.

To fix any exons that may have been mistakenly merged, either by noise or linking of exons, Tiling Assembly utilized the junction alignments produced by Tophat. Tiling Assembly searched for exons that contained a Tophat junction to identify any exons that may have been mistakenly merged. Our goal in this analysis with Tiling Assembly was to find the most common isoform of a gene where intron retention is a possibility. Thus, if the density of reads aligning on the Tophat junction was <50%, compared with those aligned to the adjacent regions, then the junction was considered to be an intron (Supplemental Figure S9). This splitting of exons increased the total number of potential exons to 185,445 exons.

Once the exons were identified, the Tophat junction alignments were used to join exons together to form transcripts. Occasionally, Tophat maps false junctions across large distances due to sequence similarities to the actual junction elsewhere in the genome. To avoid considering these false junctions, Tiling Assembly was set to disregard junctions that skipped exons and spanned distances greater than 50k nt. In addition, in our previous study (Watanabe et al., 2014), it was found that there were numerous areas of the genome where large numbers of reads mapped to small regions, often <140 nt in length, due to high sequence similarity to other highly expressed regions of the genome. Tiling Assembly was thus set to disregard potential genes that had <140 nt. After linking all of the exons together and removing these very small genes, 40,491 genes were identified, containing 136,164 exons. This number does not represent the entire rice genome because Tiling Assembly relies on transcriptome data for gene identification and not all genes are expected to be expressed in all tissues, such as aleurone.

The genes found by Tiling Assembly were compared with the annotated genes in MSU R7. Of the 40,491 genes identified, 28,019 overlapped with an annotated gene by at least 75%, and 10,129 genes by <5% (Figure 3.2). Thus, 94% of the genes identified by Tiling Assembly either corresponded well to an annotated gene or by a minimal amount. The 10,129 minimally overlapping genes were considered as potential novel genes.

The 2,343 Tiling Assembly genes that overlapped with an annotated gene between 5 and 75% may be the result of undiscovered alternatively spliced forms of known genes. Since these transcripts were identified using RNA-seq data from rice aleurone cells, a tissue that has not previously been used for gene identification, the presence of potential unannotated alternative splice variants of genes is not surprising.

Application of Cufflinks Identified 38,175 Genes Expressed in Rice Aleurone Cells

To compare the gene-finding capabilities of Tiling Assembly to a well-established assembly program, Cufflinks was run on the same RNA-seq data. Of the 38,175 transcripts identified by Cufflinks, 32,969 overlapped with an MSU R7 annotation by at least 5%. The remaining 5,206 transcripts included multiple isoforms of the same genes. To reduce overrepresentation of the same gene, transcripts that were completely contained within another transcript were eliminated. This left 4,051 potential novel genes identified by Cufflinks. Of these, 48 genes were below the minimum 140 nt requirement used in this analysis and 18 genes were on unknown chromosomes. Therefore, 3,985 potential novel genes were identified by Cufflinks.

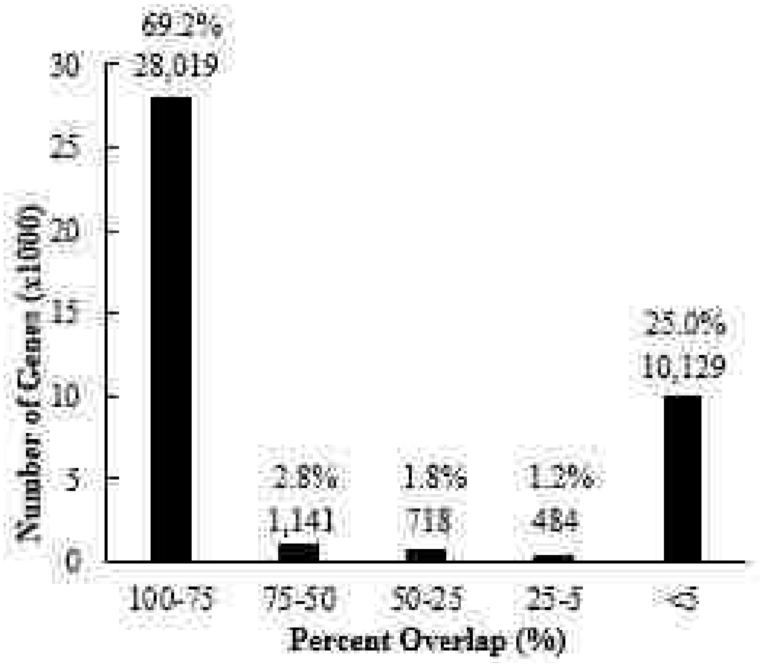


Figure 3.2: Genes identified by Tiling Assembly overlapped either well or minimally with annotated genes.

The start and termination positions of genes identified by Tiling Assembly were compared with those of the MSU R7 genes to determine the amount each Tiling Assembly gene corresponded with an MSU R7 gene. Of the 40,491 genes, 28,019 corresponded to an annotated gene by more than 75%, and 10,129 by less than 5%. This 5% category represents potential novel genes.

Comparison of the Novel Genes Identified by Tiling Assembly and Cufflinks

Among the potential novel genes identified by Tiling Assembly and Cufflinks, 1,316 genes were identified exclusively by Cufflinks, 7,460 by Tiling Assembly, and 2,669 by both (Figure 3.3A). After eliminating potential novel genes with low expression, using 100 RPKE as a threshold for gene identification based on our benchmark analysis, 4,690 genes were identified as potential novel genes. Of these, 3,473 genes were identified by Tiling Assembly, 52 by Cufflinks, and 1,166 by both (Figure 3.3B).

BLAST Searches Were Used to Eliminate Potential Novel Genes with High Sequence Similarity to Other Genomic Regions, Resulting in 767 High-Confidence Unannotated Novel Genes

During read alignment, if a read can map to multiple locations within a genome, the read is randomly assigned to one of the locations (Van Verk et al., 2013; Wang et al., 2009b). Because of this, potential novel genes that have a high similarity to another region in the rice genome may be false-positive genes. In addition, the rice genome contains a number of transcriptionally active gene fragments with high levels of sequence identity to annotated protein-coding genes and genes which may be involved in regulation of those genes rather than functioning as protein-coding genes themselves (Wang et al., 2009a). BLAST (Altschul et al., 1997) searches revealed that 774 genes showed less than 25% sequence similarity to another region within the genome. Of these genes, seven genes had a footprint of less than 140 nt and were filtered out, bringing the total number of potential novel genes with less than 25% similarity to 767 genes (Figure 3.4). These 767 genes were considered to be high-confidence novel genes based on the following criteria: they were unannotated, highly expressed, and contained relatively unique sequences (Supplemental Table S6). Of these high confidence novel genes, 151 genes were uniquely identified by Tiling Assembly

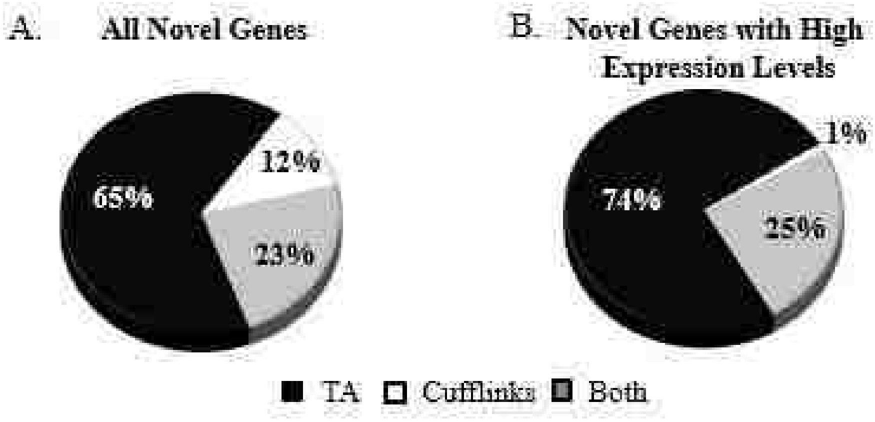


Figure 3.3: Tiling Assembly identified over 2.5 times more novel genes than Cufflinks.

Novel genes identified by Tiling Assembly and Cufflinks were compared to determine how much they overlapped. A) When all novel genes were considered, 7,460 genes were exclusively identified by Tiling Assembly, 1,316 by Cufflinks, and 2,669 by both. B) When only novel genes with high expression levels (≥100 RPKE) were considered, 3,473 genes were exclusively identified by Tiling Assembly, 52 by Cufflinks 1,166 by both.

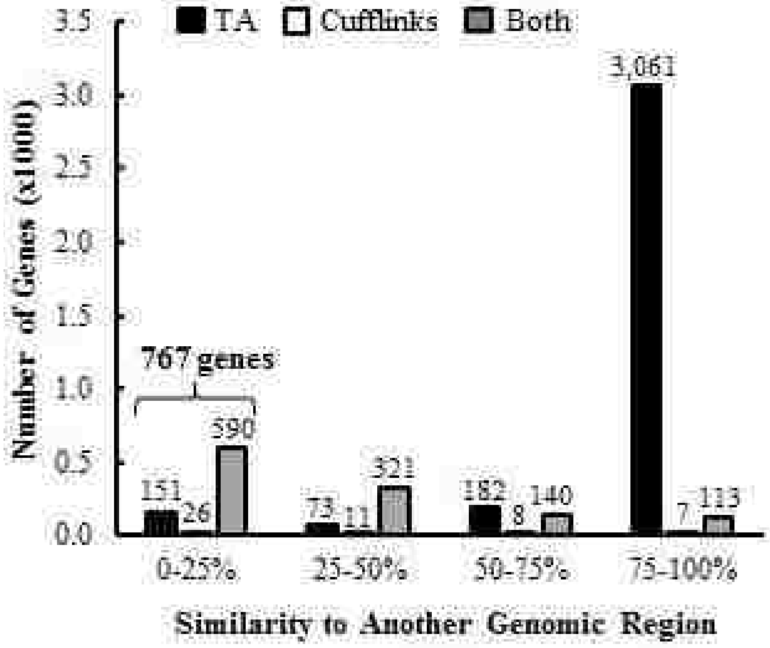


Figure 3.4: BLAST queries were preformed to determine percent similarity of novel genes to another region of the genome.

After filtering the novel genes that showed high percent similarity, 767 novel genes were identified by Cufflinks and the Tiling Assembly algorithm as high-confidence novel genes.

and 26 by Cufflinks. The remaining 590 genes were identified by both. Tiling Assembly was not only capable of finding 97% of the high-confidence novel genes found by Cufflinks, but it also found an additional 151 genes.

Comparison to Previously Published Results

In our previous publication, we identified 553 novel genes using a combination of Cufflinks and a custom Clustering Algorithm. Clustering Algorithm was developed to identify novel genes based on the presence of reads aligning to unannotated regions of the rice genome. Comparing the 767 potential novel genes identified by Tiling Assembly and Cufflinks to the 553 novel genes identified in our previous publication, there were 461 genes that coincided (Figure 3.5A). There were 306 genes that were not identified in our previous publication. These additional 306 genes demonstrate that our new pipeline is superior to that previously reported.

There were 92 genes identified in our previous publication that were not considered as novel genes in this study. Most of these 92 genes were identified by Tiling Assembly in the initial steps but, due to slight differences in gene length, were filtered out as a result of low RPKE, high similarity to another region of the genome or overlap with an annotated gene. In addition, an older version of the rice genome and older versions of Cufflinks, Tophat and Bowtie were used to identify genes in our previous publication. These factors resulted in a difference in the number of genes identified by Cufflinks in our previous publication, as compared to those reported in this study.

Of the 553 genes identified in our previous study, 124 were identified by Clustering Algorithm. These Clustering Algorithm genes were compared to Tiling Assembly and Cufflinks to determine.

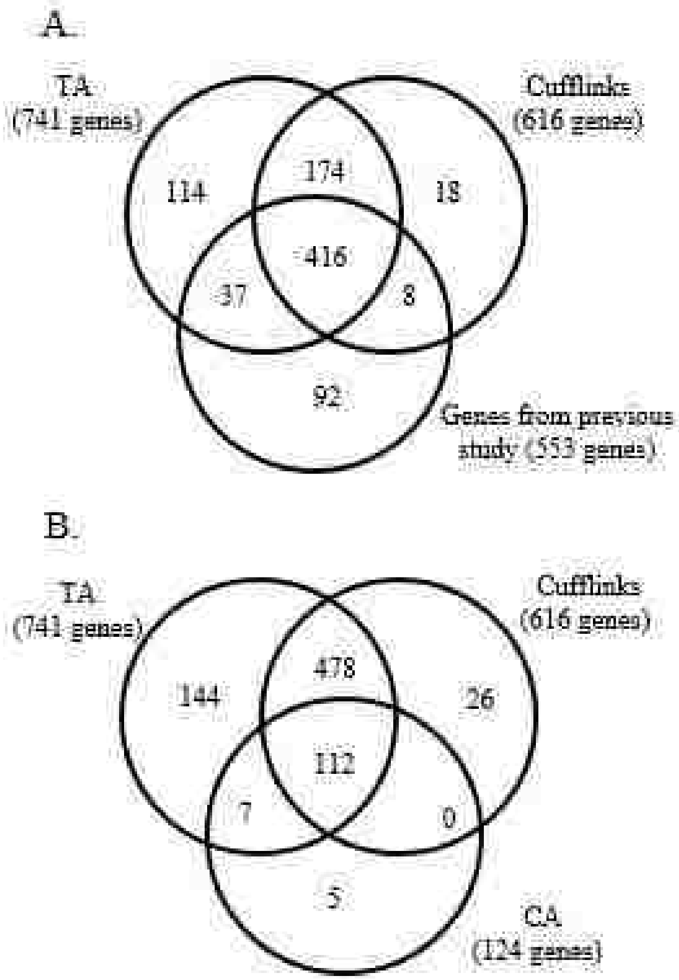


Figure 3.5: Comparison of Tiling Assembly and Cufflinks identified potential novel genes with those published in our previous study.

(A) There were 92 genes identified in our previous study that were not classified as novel genes by Tiling Assembly or Cufflinks. While all of them were identified by Tiling Assembly or Cufflinks, slight changes in their length disqualified them from fitting into the category of potential novel genes. Of the 767 unannotated genes identified in this study, 306 genes were not identified in our previous study, demonstrating that our new pipeline is superior to that previously reported. (B) While all of the Clustering Algorithm genes were also found by Tiling Assembly, slight changes in their identification disqualified five from fitting into the category of potential novel genes.

the efficiencies of the algorithms (Figure 3.5B). Of the Clustering Algorithm genes, 112 genes were found by both Tiling Assembly and Cufflinks. There were 5 Clustering Algorithm genes that were not identified by Tiling Assembly or Cufflinks

Upon closer inspection of these 5 genes, it appeared Tiling Assembly correctly identified genes within the same location, however, there were differences in how the genes were identified. Two genes identified by Clustering Algorithm were associated with previously annotated genes by Tiling Assembly and were not considered novel genes. One Clustering Algorithm gene was identified as two genes by Tiling Assembly, each of which was eliminated as a potential novel gene based on similarity to another genomic region. Slight differences in the boundaries of the remaining two genes, as identified by Tiling Assembly, resulted in shorter genes. This resulted in an increased percent similarity to another genomic region, thus eliminating them as potential novel genes. Though these five genes may be potential novel genes, they do not satisfy our requirements for consideration as high-confidence potential novel genes.

Open reading frame identification

Many regions of the genome are actively transcribed, but do not produce protein products. To determine whether Tiling Assembly and Cufflinks identified novel protein-coding genes, the introns were removed and the longest possible open reading frame (ORF) associated with each of the 767 potential novel genes was determined. The predicted peptide lengths ranged from 23 to 4,737 codons. The average ORF length was 155 codons and more than half of the genes were 80–160 codons in length (Figure 3.6). Since random DNA sequences are statistically unlikely to be more than 50 codons long without containing a stop codon (Brown, 2002), the fact that most of

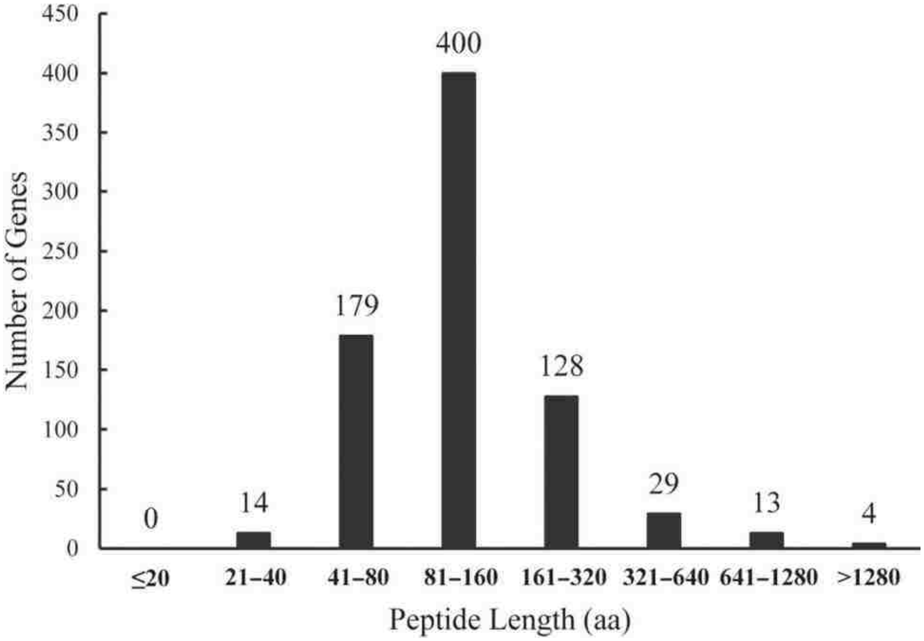


Figure 3.6: The peptide length distribution of potential novel genes is similar to that of annotated genes.

The longest ORF was determined for each of the genes using an internally developed program. The ORFs ranged from 23 to 4,737 codons in length, with an average length of 155 codons.

the ORFs found code for longer sequences indicates that they are likely protein-coding genes. Though proteins as small as 20 amino acids have been discovered in other organisms, they are uncommon (Yang et al., 2011c), and no ORFs with fewer than 23 codons were found in our data set. Only 14 novel genes (1.8%) had predicted ORFs <40 codons and are probably not protein-coding genes. These genes may encode micro-RNAs or other non-coding RNAs (Ulitsky and Bartel, 2013; Yang et al., 2011c). In addition, during the development of MSU R7, a 50 codon threshold was used. These data indicate that the majority of high-confidence novel genes identified by Tiling Assembly and Cufflinks are likely protein-coding genes, and it is not likely that they are genes that merely failed to meet the MSU R7 50 codon threshold.

The 767 potential novel genes identified by Tiling Assembly were further analyzed by performing a protein BLAST on each gene to determine whether they exhibited any sequence homology to known proteins. There were 641 genes that showed some level of sequence homology, with 99 genes having an *E*-value ≤0.0001. Of these, 97 genes were homologous to predicted proteins, one to a hypothetical protein, and one to a bacterial heat-shock protein. The remaining 126 genes did not exhibit any sequence homology to known proteins. These genes may encode lincRNAs, or other long non-coding RNAs (Ulitsky and Bartel, 2013).

Comparing Tiling Assembly and Cufflinks genes to FL-cDNAs

MSU R7 compiles annotation data from multiple sources, including FL-cDNA sequences, ESTs, and gene prediction software (Kawahara et al., 2013). As such, many of the annotated genes are hypothetical and not known to be expressed. To further evaluate the accuracy of Tiling Assembly in the identification of expressed genes, the genes identified by Tiling Assembly were compared to over 28,000 published FL-cDNAs, collected and sequenced by Kikuchi et al (Rice Full-Length c et al., 2003). These FL-cDNAs represent mRNA transcripts obtained from the rice plant and are thus more reliable than the computationally predicted transcripts in MSU R7. Of the 26,302 genes identified by Tiling Assembly with an expression level of at least 100 RPKE, 7,104 overlapped with a published FL-cDNA by more than 90% of their sequences. If multiple FL-cDNA variants overlapped with the same Tiling Assembly gene, the FL-cDNA variant with the same number of exons as the Tiling Assembly gene was selected for comparison. There were 5,767 genes that matched in exon number with their corresponding FL-cDNAs. The remaining 1,337 genes (18.8%) are herein referred to as discrepant genes. To determine the source of these discrepancies between Tiling Assembly and the FL-cDNA, seven different categories of classification were used: extra exon, missing exon, extra intron, missing intron, missing junction, gap or multiple discrepancies. In the instance Tiling Assembly recognized an exon where the corresponding FL-cDNA did not, the discrepancy was categorized as an extra exon (Supplemental Figure S10A). In the instance Tiling Assembly did not identify an exon where the corresponding FL-cDNA did, it was categorized as a missing exon (Supplemental Figure S10B). In the instance Tiling Assembly recognized an intron within the corresponding exon of the FL-cDNA, it was categorized as an extra intron (Supplemental Figure S10C). In the instance Tiling Assembly recognized a single exon where the corresponding FL-cDNA recognized two exons, it was categorized to be a missing intron (Supplemental Figure S10D).

The analysis was also performed on the Cufflinks genes. Of the 26,876 genes identified by Cufflinks that had an expression level of at least 100 RPKE, 7,690 overlapped with a published FL-cDNA by more than 90% of their sequences. Of these, 5,970 genes matched in exon number with their corresponding FL-cDNA and the remaining 1,720 genes (22.4%) were considered discrepant genes. Though Cufflinks identified more matching genes, the percentage and the number of discrepant genes was greater than Tiling Assembly.

Because it is time-consuming to analyze the causes of discrepancies for the 1,337 Tiling Assembly genes and the 1,720 Cufflinks genes, a portion of the gene pool was sampled for detailed analysis. The appropriate sample size for this comparison was calculated to be 290 genes, based on a 95% confidence level and a 5% margin of error (see sample size calculation section in Materials and Methods) and was rounded up to 300 genes. FL-cDNAs that had both a corresponding Tiling Assembly and Cufflinks gene that were discrepant were chosen for manual analysis. If a gene exhibited numerous reads, but showed a difference in the number of exons between the two data sets, alternative splicing was considered a possible cause. More than 60% of multi-exonic genes in plants are alternatively spliced (Syed et al., 2012), with intron retention being the most common form of alternative splicing (Keren et al., 2010). It is unlikely that the FL-cDNA dataset (Rice Full-Length c et al., 2003) contains all alternative splice variants of transcripts. In addition, Tiling Assembly was designed to identify a single splice variant. Therefore, it was expected that the majority of the discrepancies may be due to alternative splicing. Indeed, of the 300 discrepant Tiling Assembly genes analyzed, 96.7% could be attributed to alternative splicing. Extra or missing introns were the most abundant cause of the discrepancy as would be expected for plants. The remaining 3.3% were attributed to missing junctions or gaps. Similar results were obtained for the Cufflinks dataset, with 96.3% due to possible alternative splicing and the remaining 3.7% attributed to missing junctions or gaps.

Applying the results from the sample of 300 discrepant Tiling Assembly genes, out of the 1,337 discrepant genes, it was expected that about 1,293 discrepancies (96.7%) may be due to alternative splicing events. The remaining 44 genes (3.3%) were expected to be the result of missing junctions, gaps, or other discrepancies. These 44 genes represented 0.6% of the 7,104 genes that overlapped with the FL-cDNAs. Therefore, it was expected that Tiling Assembly may be as high as 99.4% accurate in the identification of genes with at least 90% overlap.

Appling the results from the sample of 300 discrepant Cufflinks genes, out of the 1,720 discrepant genes, it was expected that 1,656 discrepancies were possible alternative splicing events. The remaining 64 genes were expected to be the result of missing junctions, gaps, or other discrepancies. These 64 genes represented 0.8% of the 7,690 genes that overlapped with the FL-cDNAs. Therefore, it was expected that Cufflinks may be as high as 99.2% accurate in the identification of genes with at least 90% overlap.

Identification of Transcription Start and Termination Sites by Tiling Assembly

The transcription start and termination sites were compared between genes identified by Tiling Assembly with those identified by the FL-cDNAs (Figure 3.7). Only Tiling Assembly genes that overlapped with an FL-cDNA by at least 90% were considered. Of the 7,174 transcription start sites that satisfied the specified overlap threshold, about 83% differed by less than or equal to 100 nt (Figure 3.7A). The transcription start sites predicted by Tiling Assembly were on average 30 nt upstream of the FL-cDNAs. This data demonstrated that Tiling Assembly is a reasonably reliable tool for the identification of transcription start sites.

Of the 7,174 transcription termination sites that satisfied the specified overlap threshold, about 69% differed by less than or equal to 100 nt (Figure 3.7B). The transcription termination sites predicted by Tiling Assembly were on average 71 nt downstream of the FL-cDNAs. These data demonstrate that Tiling Assembly is less reliable at predicting the transcription termination sites.

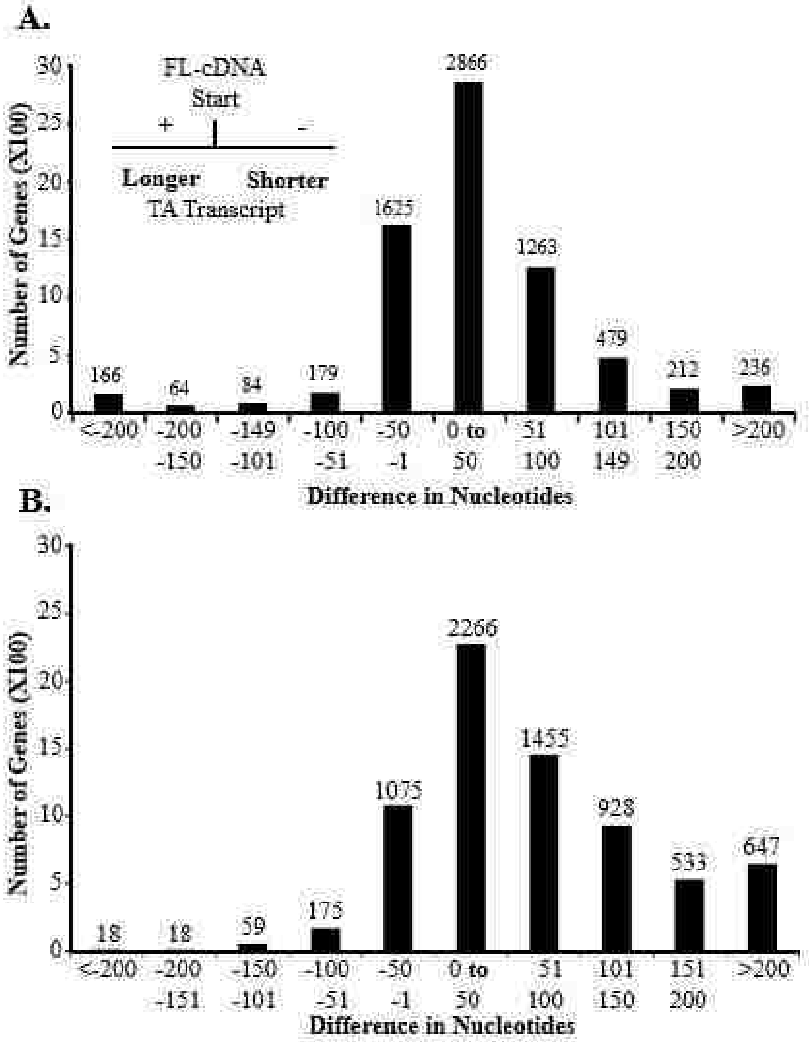


Figure 3.7: Nucleotide differences between transcriptional start (A) and termination (B) sites of FL-cDNAs and Tiling Assembly genes.

The transcriptional start and termination sites were compared to those of FL-cDNAs previously published. The distribution of the difference between the Tiling Assembly and FL-cDNA start and stop sites is presented here. A negative value indicates the Tiling Assembly gene is shorter than the FL-cDNA and a positive value indicates the Tiling Assembly gene is longer.

Overall, Tiling Assembly overestimated the length of the transcripts. This overestimation may be due to noise reads near the start and termination sites. Alternatively, past research has indicated that termination sites are variable (Richard and Manley, 2009). Hence the accuracy rates of Tiling Assembly in predicting transcription termination sites may be underestimated.

Application of Tiling Assembly to the Genomes of Model Organisms

Having demonstrated the effect of Tiling Assembly on detecting novel genes in rice, its performance was evaluated on other model organisms. Loraine et al. (Loraine et al., 2013) discovered 5,312 transcriptionally active regions (TARs) in the unannotated regions of the *A. thaliana* genome (Loraine et al., 2013), however, few filters were used to remove false-positive TARs. For instance, of the 5,312 TARs reported, 3,490 were the length of a single read (75 nt). However, this large number of TARs indicated that there may still be undiscovered genes in Arabidopsis. Using the same Tiling Assembly parameters used for identifying potential novel genes in *O. sativa,* 218 potential novel genes were identified in Arabidopsis, representing nearly 1% of all annotated genes. Of these 218 potential novel genes, 99 genes (45%) contained at least part of one or more TARs, and 35 genes (16%) had at least one TAR completely contained within the gene. It is likely that most of the TARs reported by Lorraine et al. did not correlate with a Tiling Assembly novel gene because the majority of them were identified based on individual reads. These individual reads may have resulted from genomic DNA contamination or noise caused by statistical mapping error.

To determine if the ability of Tiling Assembly to find large numbers of unannotated genes was applicable to non-plant species, the algorithm was applied to several additional model organisms (Table 1). Surprisingly, 458 novel genes were identified in *S. cerevisiae*, representing almost 7%

of the known genes. Even though this model organism has been more intensively studied than Arabidopsis, this large number of potential novel genes may be due to the fact that only a single RNA-seq replicate was used in this study. Similar analysis was performed on *D. melanogaster* and *C. elegans*. The number of potential novel genes identified in each of these additional organisms may be improved by using parameters specific to the organism.

Table 1. Potential novel genes identified in other organisms by TA.

Species	Ch	GS (Mbp)	No. of Genes	No. of Novel Genes	% Annot. Genes
O. sativa	12	381	55,986	767	1.37
A. thaliana	5	125	25,498	218	0.85
S. cerevisiae	16	12	6,603	458	6.94
C. elegans	6	97	47,060	126	0.27
D. melanogaster	4	120	17,294	361	2.09

Notes: RNA-seq data downloaded from the SRA were used to investigate the ability of TA to find unannotated genes in additional model organisms. For each of the organisms, the same criteria were used for identifying potential novel genes as those used for rice. Ch: number of chromosomes; GS: genome size; Annot. Genes: annotated genes.

Discussion

There are relatively few publicly available transcript assembly programs despite the vast increase in the use of RNA-seq. In this study, a novel assembly algorithm, Tiling Assembly, was developed to address the lack of established algorithms to identify transcribed regions as genes. This algorithm was compared to Cufflinks transcript assembly software to evaluate its gene-finding capabilities in relation to established assembly software (Table 2). It was concluded that Tiling Assembly found substantially more genes and found a lower number of false-positive genes, though Cufflinks ran somewhat faster and was able to identify multiple transcripts for a given gene. It was also determined that Tiling Assembly and Cufflinks had similar accuracy at predicting single and multi-exonic genes down to 100 RPKE (Fig. 2). When the Tiling Assembly genes were compared to FL-cDNAs, the vast majority of discrepancies could be attributed to alternative splicing. Excluding those genes, Tiling Assembly appeared to be as high as 99.4% accurate in identification of genes.

Table 2.

Comparison of Tiling Assembly to Cufflinks

Category	**Tiling Assembly**	**Cufflinks**
Novel gene finding[a]	3,473 genes	52 genes
False-positive rate[b]	17 out of 100 genes	25 out of 100 genes
Ease of use	User interface	Command line
Algorithm	Read tiling	Bipartite graph
Minimum expression	100 RPKE	100 RPKE
Run time[c]	3–4 h	1.5 h
Transcripts identified	Single transcript	Multiple transcripts

[a] These figures represent genes found exclusively by either TA or Cufflinks that have an expression level greater than 100 RPKE, prior to percent similarity filter.

[b] Data from random genome with 100 known genes inserted. No filtering performed.

[c] Excludes set-up time. Set-up for Tiling Assembly and Cufflinks is about the same time.

Before applying filters for potential novel genes, 28,019 genes that overlapped at least 75% with an annotated gene (Fig. 3) were found from the RNA-seq data using Tiling Assembly. MSU R7 annotation, contains 55,986 genes which would seem to imply that half the annotated genes in the rice genome were expressed in the aleurone cells. In our previous publication, we reported that 18,152 annotated genes were expressed in the aleurone cells (Watanabe et al., 2014). The criteria used for expression in our previous publication required an expression level of at least 1.0 read per kilobase of exon model per million mapped reads (RPKM) which is equivalent to about 150 RPKE for this data. At 150 RPKE, Tiling Assembly identified 17,971 annotated genes which was in line with our previous publication. Since 100 RPKE was determined as the minimum expression for gene detection, Tiling Assembly identified 20,230 expressed annotated genes in rice aleurone.

Comparison of the novel genes identified by Tiling Assembly to Cufflinks shows that Tiling Assembly identifies up to 74% more genes that Cufflinks (Fig. 4). After eliminating genes that showed high similarity to another genomic region, the Tiling Assembly identified more 151 more novel genes than Cufflinks (Fig. 5). We analyzed the data to determine possible causes for the large discrepancy in genes identified between the two programs. One possible reason for this discrepancy was that Tiling Assembly was calling regions that were highly similar to another region, whereas Cufflinks disregarded them. BLAST search filters revealed that this was not the case. Another possible reason was that Cufflinks and Tiling Assembly found a gene within the same region, but the start and stop locations of those genes were different. Investigation of the positional overlap of the high-confidence novel genes found by each of the programs verified that those genes unique to each program did not overlap. A final possibility considered was that some novel genes detected by Cufflinks were longer that the corresponding Tiling Assembly gene and thus overlapped with an adjacent annotated gene. This overlap eliminated the Cufflinks gene as a potential novel gene. Analysis showed that there were 101 Cufflinks genes that overlapped with an annotated gene. These occurrences were attributed to two causes; the Cufflinks gene included regions with low read depth, which were considered noise reads by Tiling Assembly, or Cufflinks exons were merged based on junction alignments that were disregarded by Tiling Assembly because they were very long and skipped exons. In five cases, one of these long junctions skipped over an expressed region, which was called a novel gene by Tiling Assembly but an intron by Cufflinks. In 19 additional cases, these long junction alignments led to extremely long genes identified by Cufflinks which spanned multiple MSU R7 annotated and Tiling Assembly genes. There were nine expressed regions detected as a gene by Tiling Assembly where there was no corresponding Cufflinks gene, for unknown reasons.

Tiling Assembly is a heuristic, *ab initio* transcript assembly algorithm which uses a read tiling approach to identify transcripts. Unlike *de novo* assembly algorithms such as Trinity (Grabherr et al., 2011), Tiling Assembly takes advantage of a sequenced genome to improve the accuracy of transcript assembly while decreasing CPU requirements. Tiling Assembly does not require an annotated genome, so it may be used for organisms where the genome is sequenced but the annotation is naïve. Many of the current transcript assembly algorithms attempt to reproduce each of the isoforms available to a gene using a bipartite graph approach, which can lead to reporting of statistically probable, but non-real isoforms and dilution of expression levels of real isoforms as reads are assigned to the non-real isoforms. Tiling Assembly instead produces the longest possible isoform of a gene. Comparison of Tiling Assembly to a well-established transcript assembly program, Cufflinks, revealed that Tiling Assembly's strengths lie in the accurate prediction of exons in the presence of noise and improved discovery of high-confidence novel genes.

In conclusion, we describe a heuristic approach to novel gene identification. Using this approach in combination with Cufflinks, 767 high-confidence unannotated genes were identified in rice. These genes contained predicted ORFs ranging from 40 to over 4,000 codons, with the majority showing sequence homology to known and predicted proteins. The accuracy of the genes identified by Tiling Assembly was validated through comparison with their corresponding FL-cDNAs, which implied Tiling Assembly may be as accurate as 99.4%. Tiling Assembly accurately predicted the transcription start sites to within 100 nt of the corresponding FL-cDNA, but was less accurate at predicting the transcription termination sites. Application of Tiling Assembly on *A. thaliana*, *D. melanogaster*, *C. elegans*, and *S. cerevisiae* identified hundreds of high-confidence novel genes, demonstrating that even in the most well-studied model organisms there are still undiscovered genes. This pipeline proves to be an effective way to identify novel genes in a diverse array of organisms. The novel genes identified here should be further studied to determine their functions and roles in their organisms.

Materials and methods

Tiling Assembly Pipeline

The Tiling Assembly pipeline, depicted in Figure 3.8, begins with alignment of RNA-seq short-read data to the appropriate genome using the alignment software, Bowtie (Langmead et al., 2009) and Tophat (Trapnell et al., 2009). Exons of potential genes are identified based on the presence of regions containing overlapping reads. Exons too short to be identified by read alignment are identified by Tophat junction alignments. Gaps in the read coverage can cause single exons to be identified as two or more exons; thus, exons that are very closely spaced are linked together. To prevent accidental merging of exons, exons containing one or more junction alignments are

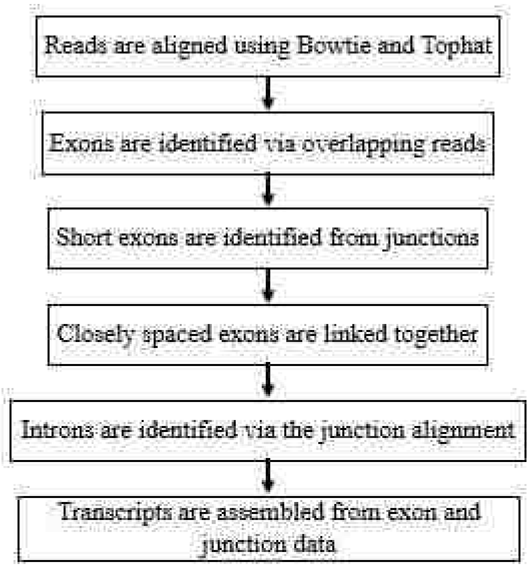

Reads are aligned using Bowtie and Tophat
Exons are identified via overlapping reads
Short exons are identified from junctions
Closely spaced exons are linked together
Introns are identified via the junction alignment
Transcripts are assembled from exon and junction data

Figure 3.8: Flowchart of the Tiling Assembly algorithm to identify novel genes in rice

separated into multiple exons. Finally, all of the exons identified are joined together by the junction alignments to form the assembled transcripts.

Gene identification begins with the detection of exons based on the short-read data. The aligned reads are loaded into a MySQL database, which is queried by Tiling Assembly to identify exons based on the presence of overlapping reads. To prevent misidentification of exons due to noise, a threshold may be set by the user to determine the minimum reads per kilobase of exon (RPKE) required to identifying an exon.

Exons shorter than the length of a read (50 nt) tend to have few to no reads aligned, regardless of gene expression, preventing them from being detected by considering only read alignment. To identify these short exons, Tiling Assembly relies on the partial read alignments from junction mapping. Tophat takes the reads that cannot be directly mapped to the genome and breaks them into two parts for independent alignment. The junctions that these reads map across are indicative of an intron. Tiling Assembly relies on these partial read alignments to detect short exons (Supplemental Figure S3).

Low read coverage from genes with low expression often leads to gaps in the read coverage, causing alignment algorithms to mistakenly identify multiple exons where only a single exon is present. To avoid such false gaps, Tiling Assembly merges exons that are within a user-specified distance of each other. Linking closely spaced exons together, however, may result in an incorrect merging of exons.

Other factors, such as intron retention, pre-spliced mRNA, and noise, may also contribute to incorrectly merged exons because they result in reads mapping to intronic regions (Supplemental

Figure S4). To prevent mistakenly merged exons, Tiling Assembly searches for junction alignments within the identified exons. Exons containing both sides of a junction alignment are separated and trimmed based on the boundaries of the junction to ensure accurate exon–intron boundaries (Supplemental Figure S5). The beginning of the first exon and the end of the last exon of a transcript cannot be determined by junction alignment.

Once these high-confidence exons are produced, Tiling Assembly assembles the exons into specific genes using junction alignments. To avoid false junctions between similar genomic regions, the user can specify a maximum length of a junction that skips over one or more exons (Supplemental Figure S6). In addition, the user can specify the size of very large junctions to be disregarded to avoid invalid junctions due to mapping errors.

RNA-seq and Genome Data

The RNA-seq data used for rice were obtained from RNA-extraction of rice aleurone performed in our lab, followed by library preparation and sequencing on the Illumina Hi-seq 2000 platform by the Huntsman Cancer Institute, University of Utah. The data was submitted to the Sequence Read Archive (SRA) (Leinonen et al., 2011) and is publicly accessible under the accession number SRP028376. The SRA accession numbers that were used for the analysis of other species were SRP022162 (*A. thaliana*), SRR590802-4 (*D. melanogaster*), SRR650494-5 (*C. elegans*), and SRR1019759 (*S. cerevisiae*).

The rice genome and annotation were downloaded from the MSU Rice Genome Annotation Project Release 7.0 (MSU R7) (Kawahara et al., 2013) for *O. sativa* (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/).

The *A. thaliana* data was downloaded from PhytozomeV10 (Goodstein et al., 2011) (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10).

The *C. elegans* data was downloaded from WormBase (Harris et al., 2013) (ftp://ftp.ensembl.org/pub/release-75/fasta/caenorhabditis_elegans/dna/).

The yeast data was downloaded from *Saccharomyces* Genome Database (Engel et al., 2014) (http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/).

The *D. melanogaster* data was downloaded from FlyBase (St. Pierre et al., 2014) (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.56_FB2014_02/fasta/).

Alignment of full-length cDNA to the rice genome

The full-length cDNAs (FL-cDNAs) (Rice Full-Length c et al., 2003) were aligned to the rice genome using the Exonerate alignment software with the following parameters: model est2genome, geneseed 250, and bestn 1 (Slater and Birney, 2005). The FL-cDNAs and short-read data were loaded into the University of California, Santa Cruz (UCSC) Genome Browser (Kent et al., 2002), for visualization at the following address: http://shenlab.sols.unlv.edu/cgi-bin/hgGateway.

Short Read Alignment

Short reads were aligned to MSU R7 via Bowtie version 2.1.0 and Tophat version 2.0.9 software and OLego (Wu et al., 2013). The maximum junction length was set to 50,000 nt. Default values were used for all other parameters. The short read data for all of the samples were merged. Transcript assembly was then performed using Cufflinks version 2.0.2 on the composite data to generate GTF files. The mapped short reads and junctions were loaded into a MySQL database. Tiling Assembly queried the database to identify exons and genes.

Sample Size Calculation for Comparison of Tiling Assembly and Cufflinks genes to FL-cDNA

The minimum sample size (*n*) needed to ensure the genes identified by Tiling Assembly and Cufflinks coincided to FL-cDNAs, was calculated using the following equation: $n = \frac{Nx}{[(N-1)E^2+x]}$, where $x = Z_{(\frac{c}{100})}^{2}\, r(100 - r)$ (Gutierrez et al., 2013); *N* represents the total number of discrepant genes; *E* is the margin of error, which was set to 5%; the critical value $Z_{(\frac{c}{100})}$ was set to 1.96 based on a 95% confidence level (*c*); and *r* was set to 60% since this value was the expected discrepancy rate contributed to alternative splicing.

CHAPTER 4

IN SILICO IDENTIFICATION AND EXPERIMENTAL VERIFICATION OF A NOVEL ABSCISIC ACID RESPONSIVE ELEMENT IN RICE

Abstract

Abscisic acid (ABA) is a central plant hormone involved in many developmental and physiological processes. For the study of ABA signaling in cereal crops, aleurone cells have served as a model system because of they are not known to produce hormones but respond to hormones such as ABA. By performing RNA-seq on hormone treated rice aleurone cells, 2,443 ABA-inducible genes were identified. The ABRE consensus sequence (ACGTG(G/T)C), known as the key central *cis*-acting element of ABA signaling, was found to be present in only 39.5% of these genes, suggesting other ABREs may exist. In this study, we used a bioinformatics approach to identify a novel ABA Response Element, designated as ABREN. This element is enriched up to 45 fold in the highly ABA-inducible rice genes, but not in ABA-inducible genes of *Arabidopsis thaliana*. Particle bombardment-mediated transient expression studies confirmed that ABREN indeed mediates ABA signaling in rice aleurone cells. Gene ontology analyses suggest that many of the ABREN-containing genes are involved in stress responses. We have also shown that α-amylase treatment of rice aleurone increases the sensitivity of rice aleurone cells to ABA. Collectively, this study advanced our understanding of diverse *cis*-regulatory sequences underlying ABA responses and of the transcriptomes underlying the crosstalk between ABA and sugar signaling.

Introduction

Abscisic acid (ABA) mediates plant responses to many biotic and abiotic stresses such as high salinity, temperature extremes and exposure to UV light, which allows survival under less than optimal conditions (Finkelstein and Rock, 2002; Zhang et al., 2006). The physiological responses to ABA include stomatal closure (Schroeder et al., 2001), maintenance of shoot and root growth (Sharp and LeNoble, 2002) and delayed flowering and promotion of senescence (Finkelstein, 2013). ABA also promotes the dormancy of seeds (Koornneef and Reuling G, 1984) and buds (McWha and Langer, 1979) and inhibits germination . In addition, ABA antagonizes the effects of development and growth stimulating hormones such as auxins, gibberellins and cytokinins (Bethke et al., 1997; Khan and Downing, 1968; Tanaka et al., 2006).

Research on ABA signal transduction pathways has led to a more thorough understanding of plant ABA response. ABA evokes a plant response via an interaction with PYRABACTIN RESISTANCE1 (PYR1)/PYR1-LIKE (PYL)/REGULATORY COMPONENTS OF ABA RECEPTORS (RCAR) PYR/PYL/RCAR receptors (Ma et al., 2009; Park et al., 2009). This interaction between ABA and the receptors results in subsequent binding of the type 2C protein phosphatase (PP2C) to the receptor. Consequently, protein kinase SnRK2 is activated via autophosphorylation, which in turn activates bZIP transcription factors and leads to an ABA response (for the latest review, see (Yoshida et al., 2015)). In rice there have been reported 12 PYR/PYL/RCAR receptors (Kim et al., 2012), 78 PP2Cs (Xue et al., 2008), 10 SnRK2s (Kobayashi et al., 2004), and 89 bZip transcription factors (Nijhawan et al., 2008).

In addition to PYR/PYL/RCAR, two GPCR-type G proteins (GTG1 and GTG2) were identified as ABA receptors in Arabidopsis. These proteins bind to ABA and have GTPase activity (Pandey et al., 2009). It has also been shown that Rho GTPases play an inhibitory role in the ABA signaling pathway and that mutations in these elements resulted in enhanced ABA sensitivity (Miyawaki and Yang, 2014).

Many ABA-inducible genes in various species contain a conserved *cis*-regulatory ABA responsive element (ABRE). The ABRE contains the core sequence, ACGT, also known as the G-box and was first identified in wheat (Marcotte et al., 1989). The bZIP type transcription factor family of genes were identified as the binding factor to the G-box with various degrees of affinity depending on the bZIP protein and the flanking nucleotides to the ACGT core (Izawa et al., 1993). In rice, the G-box containing ABRE consensus sequence ACGTG(G/T)C has been reported (Hobo et al., 1999; Yamaguchi-Shinozaki et al., 1990). However, in order for ABA responsive transcription to occur, a single copy of the ABRE is not sufficient. In barley, the combination of an ABRE and one of two known coupling elements CE1 (TGCCACCGG) (Shen and Ho, 1995) and CE3 (GCGTGTC) (Shen et al., 1996), constitute an ABA responsive complex (ABRC) in the regulation of the ABA-inducible genes *HVA1* and *HVA22* (Zhang et al., 2004). It was also shown that a pair of ABREs can function as an ABRC with the second ABRE playing the role of the coupling element in barley (Shen et al., 1996) and in Arabidopsis (Nakashima and Yasunari Fujita, 2006). It has also been shown that two CE3s can couple to form an ABRC in rice (Hobo et al., 1999). In Arabidopsis, the CE3 element is practically absent; thus, Arabidopsis relies on paired ABREs to form ABRCs (Gomez-Porras et al., 2007).

There have been several other reported *cis*-acting elements that are responsive to ABA that do not contain the G-box. These include the drought response elements (DRE1:CGAGAAGAACCGAGA and DRE2:CCGGGCCACCGACGCACGG) in maize (Kizis and Pages, 2002) and (TACCGACAT) in Arabidopsis (Narusaka et al., 2003); the Myb (YAAC(G/T)G)) and Myc (CANNTG) elements in Arabidopsis (Iwasaki et al., 1995); the TT motif (TTTCGTGT) in carrot (Chung et al., 2005); the Sph motif (CGTGTCGTCCATGCAT) in maize (Kao et al., 1996); and the motif (AAGCCCAAATTTCAC-AGCCCGATTAACCG) in the

C. plantagineum (resurrection plant) (Hilbricht et al., 2002) (for review see Srivastava, 2002). This wide variety of ABA-responsive *cis*-regulatory elements suggest that there may be other undiscovered ABA-responsive elements.

RNA sequencing (RNA-Seq) analysis revealed 2,443 genes that were ABA-inducible in rice aleurone cells. Of these genes, 39.5% contained the ABRE consensus sequence)ACGTG(G/T)C) published previously (Shen et al., 2004). This disparity between the presence of the ABRE in ABA-inducible genes and the fact that a variety of other ABA-responsive *cis*-regulatory element have been identified in other species suggest that other elements are involved in ABA signaling. To identify such elements, we performed Gibbs sampling analyses and identified a novel element that was enriched 45 fold in the 1,000 bp upstream region of the highly ABA-inducible rice genes. Transient expression studies confirmed that this element was indeed involved in ABA response in rice aleurone cells.

The high quality of the RNA-seq data was demonstrated by several analyses.

The results of the sequencing gave about 120.6 million short reads on the control and 130.2 million short reads on the ABA-treated sample. Over 89% of the reads were mapped using Bowtie alignment software (Langmead et al., 2009) (Supplemental Table S7). This high alignment rate indicated that the quality of the RNA-seq was very high.

The alignment software showed that about 22.5 million reads of the control sample mapped to more than one location. This equated to about 18.6% of the reads, which was very close to a previous RNA-seq report (Toung et al., 2011), suggesting that the alignment data were in line with expectations.

To further confirm the quality of the RNA-seq data, we evaluated the expression patterns of several known ABA-inducible genes (Joshee et al., 1998; Moons et al., 1997; Morris et al., 1990; Ross and Shen, 2006; Yamaguchi-Shinozaki et al., 1990; Zhang et al., 2004). The RNA-seq data showed that all nine previously reported ABA-inducible genes were in fact ABA-inducible, with fold changes ranging from 17 to 55 (Supplemental Table S8).

The α-amylase treatment increased the sensitivity of aleurone cells to ABA

RNA-seq data on ABA-treated rice aleurone cells were already reported in our previous publication (Watanabe et al., 2014). However, there were some issues with the experiment. The enzyme α-amylase was used to break down the starchy endosperm to facilitate the isolation of aleurone cells. Since α-amylase hydrolyzes starch into sugars, sugar signaling pathways may be triggered upon α-amylase treatment that may interfere with the ABA response. Also, only a single RNA-seq analysis was performed on each treatment preventing adequate statistical analyses. Since our last RNA-seq analysis, our techniques have improved to the point where α-amylase treatment was not necessary to isolate high quality mRNA from rice aleurone cells. Also, three biological replicates were performed for each sample so that statistical data analyses can be performed.

Comparison of the single-replicate α-amylase treated (AAT) RNA-seq data showed substantially more genes were responsive to ABA treatment compared to the three-replicate non-α-amylase treated (NAAT) data. In all, 3,408 genes were significantly ABA inducible in the AAT dataset, but only 2,443 genes in the NAAT dataset. There were 1,476 genes that were significantly ABA inducible in both samples (Figure 4.1A). Of the 1,476 genes, 1,333 genes (90.3%) had a higher level of ABA induction in the AAT sample than the NAAT sample (Figure 4.1B).

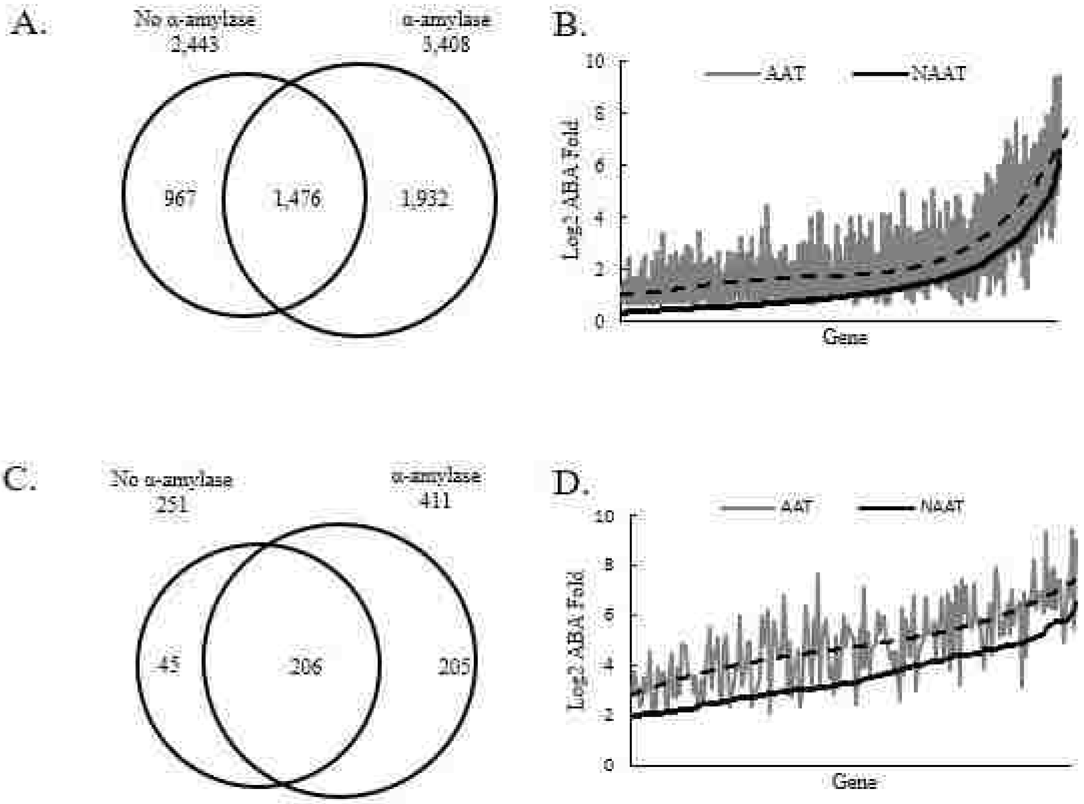


Figure 4.1: Alpha-amylase increases the sensitivity of ABA response.

A.) The AAT sample shows substantially more genes (3,408 genes) responsive to ABA compared with the NAAT samples (2,443 genes). B.) The ABA induction level of genes in the AAT sample were higher than in the NAAT sample. C.) When considering only highly ABA induced genes (>4-fold induction, >10 FPKM), almost all the genes (206 out of 251 genes) induced in the NAAT samples were also induced in the AAT sample. D.) The ABA induction level of highly induced genes in the AAT sample were higher than in the NAAT sample. The dashed line is a polynomial trend line that approximates the running average of the AAT ABA fold induction.

Since there is always some fluctuation in gene expression, the NAAT data should be more reliable than the AAT data since three replicates were used to determine whether a gene was ABA inducible or not. However, there were 967 genes that were ABA inducible in the NAAT data that were not ABA-inducible in the AAT data. Though a test statistic was used to determine the significance of ABA induction, there still may be genes that were falsely identified as ABA inducible due to random fluctuations in gene expression, low read counts and background noise. To reduce the number of false positive ABA-inducible genes, we considered genes that were at least four-fold ABA induced and have an expression level of at least 10 FPKM. These high thresholds reduce the possibility of mistakenly calling a gene ABA inducible.

This substantially pared down the number of ABA-inducible genes in our NAAT data to 251 genes, and in the AAT data to 411 genes (Figure 4.1C). There were 206 genes that were ABA-inducible in both data sets. There were 45 genes that were ABA inducible in the NAAT data that were not inducible in the AAT data. Of these 45 genes, 25 genes were ABA-inducible by the AAT data but the level of induction was less than four-fold. The remaining 20 may have been missed for various reasons: Some genes may be ABA inducible in the AAT data but were not statistically significant; some may have been repressed by the α-amylase treatment; or some may have been missed due to random fluctuations in gene expression or noise. It has also been shown that the rice annotation has some inaccuracies that my also contribute to this discrepancy (Watanabe et al., 2014).

There were 205 genes that were ABA-inducible by more than four-fold in the single-replicate AAT data but not in the NAAT data. Almost all of these 205 genes (191 genes, 93.2%) were significantly ABA inducible in the NAAT data but the level of ABA induction was less than four-fold.

Of the 206 genes that were ABA-inducible by more than four-fold in both the AAT and NAAT data, almost all of these 206 genes (181 genes, 88%) had a higher fold change in the AAT data

(Figure 4.1D). This demonstrates that the α-amylase treatment enhanced the sensitivity of aleurone cells to ABA

The ABA induction levels of the nine previously reported ABA inducible genes were compared between the AAT and the NAAT datasets. All nine genes in the AAT dataset had a higher ABA induction fold than the NAAT dataset.

To determine whether the added α-amylase enzyme caused the change in gene expression or the sugars produced results in changes in gene expression, we looked at two genes that have been previously analyzed for sugar sensitivity. These are αAmy3 (LOC_Os08g36910) and αAmy8 (LOC_Os08g36900) (Chen et al., 2006). Both genes contain the sugar response complex in their promoter regions, but it was demonstrated that only αAmy3 shows sugar sensitivity in the endosperm. αAmy8 does not show sugar sensitivity in the endosperm but shows sensitivity in the embryo. Comparing the read counts of the AAT control sample to those of the NAAT sample for αAmy3, the number of reads on the AAT sample were very low at 1.05 FPKM (Supplemental table S9) while the reads on the NAAT sample were 228.25 FPKM resulting in a log2 fold change of 7.77. For αAmy3, the read count on the AAT sample was 18.95 FPKM on the AAT sample and 23.70 FPKM on the NAAT sample resulting in a statistically insignificant fold change. This data corresponds well with previously published work and supports that the observed phenomena is due to the sugar signaling.

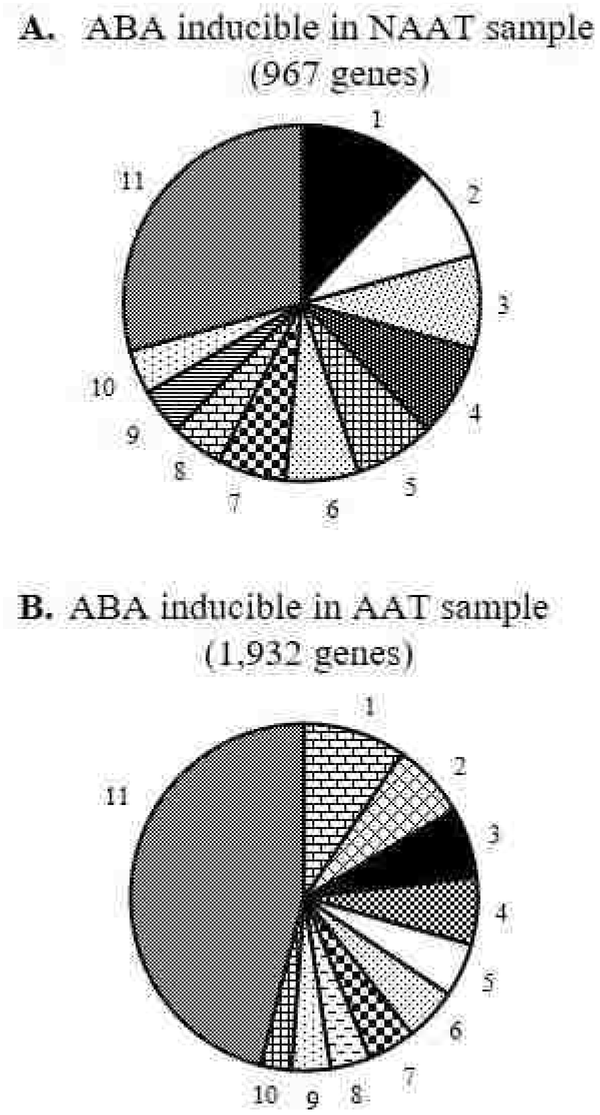
Sugar transport genes were examined for their sensitivity to α-amylase treatment. These genes are expected to be repressed by α-amylase treatment because the excess sugar should reduce the necessity of sugar transporters. Of 10 sugar transport genes examined (OsMST1-8, LOC_Os09g12590 and LOC_Os09g24924), three genes were significantly differentially expressed (OsMST3, OsMST4 and OsMST6) and all three where sensitive to the α-amylase treatment as expected. This is additional support that the sugar generated from the α-amylase is responsible for the changes in expression between the AAT and NAAT samples.

ABA synthesis genes have been shown to be induced by sugar in Arabidopsis (Cheng et al., 2002). By performing protein BLAST queries of the Arabidopsis ABA synthesis genes (SDR1, ABA3, AAO3, NCED2, and NCED3), the rice homologues were identified. All the rice homologues showed low expression in both the AAT and NAAT samples and only ABA3 showed sensitivity to sugar. This is an indication that ABA synthesis does not take place in rice aleurone cells.

Gene ontology enrichment analysis showed substantial difference between the genes induced by ABA in the AAT dataset versus those in the NAAT dataset

To understand the difference between the genes induced in the AAT dataset versus those in the NAAT dataset, gene ontology analysis was performed. The analysis was performed using the UC Davis Rice Array Database GO Enrichment software (Cao et al., 2012).

Gene ontology enrichment analysis was performed on the genes induced by ABA in the NAAT sample but not the AAT sample (Figure 4.2A) and on the genes induced by ABA in the AAT sample but not the NAAT sample (Figure 4.2B). When comparing the gene ontology categories of the ABA inducible genes in the NAAT against those of the AAT sample, a substantial difference is observed. For instance, in the NAAT sample, the largest category is metabolic process which consists of 12% of the genes, followed by regulation of transcription (9%), homiothermy (8%) and response to freezing (8%). The ontology categories homiothermy and response to freezing are not even in the top 10 categories of the AAT genes. Regulation of transcription is the 5th category in the AAT sample (5%).



A. ABA inducible in NAAT sample (967 genes)

B. ABA inducible in AAT sample (1,932 genes)

Figure 4.2: Gene ontology analysis of the genes induced by ABA in the AAT and NAAT samples show different distribution of categories.

A.) Genes that are induced by ABA in the NAAT sample but not the AAT sample **1)** metabolic process **2)** regulation of transcription **3)** homiothermy **4)** response to freezng **5)** oxidation reduction **6)** regulation of transcription, DNA dependent, **7)** transport **8)** protein amino acid phosphorylation **9)** carbohydrate metabolic process **10)** transcription **11)** other. **B.)** Genes that are induced by ABA in the AAT sample but not the NAAT sample **1)** protein amino acid phosphorylation **2)** defense response**3)** metabolic process **4)** apoptosis **5)** regulation of transcription **6)** regulation of transcription, DNA dependent, **7)** transport **8)** proteolysis **9)** transcription **10) 11)** other. Only gene ontology categories with a hypergeometric p value less than 0.05 were used.

In the AAT sample, the largest category is protein amino acid phosphorylation which consists of 10% of the genes, followed by defense response (7%), metabolic process (7%) and apoptosis (6%). The protein amino acid phosphorylation, is the 8th category in the NAAT genes. The defense response and apoptosis ontology categories of the AAT genes are not even in the top 10 categories of the NAAT genes. Only one category is common in the top four categories of both samples and that is metabolic process which is the first category amongst the NAAT genes and the third category amongst the AAT genes.

A novel cis-acting element was discovered in ABA-inducible rice genes

To identify novel ABA responsive elements, we obtained the 251 highly ABA-inducible genes (ABA induced by ≥ 4 fold) and had high level of expression (>10 FPKM) from our three-replicate NAAT dataset. The DNA sequences 1,000 bp upstream from the ATG start codon of these genes were obtained from the MSU R7 rice dataset. Many genes in the MSU rice dataset do not have an annotated 5' UTR, so the upstream from the ATG start codon was chosen. This region includes the 5' UTR which is typically about 127 bp (Umezawa et al., 2008). The 1,000 bp length was chosen since the majority of transcription factors binding sites (74%) are within 500bp of the TSS (Harbison et al., 2004). The Bioprospector software (Liu et al., 2001) was used to identify enriched DNA motifs in the upstream region of these ABA-inducible genes. Various lengths of motifs, from 6 to 12 bp, were queried. An eight base pair sequence (GATCGATC) was identified as enriched. Other potential ABA response elements were also identified (Table 1) but their frequencies were substantially less than the GATCGATC sequence. This sequence was identified a total of 99 times, with some genes containing multiple copies of the motif. About 20% of highly ABA-inducible genes (51 of the 251) contained this motif at least once. Queries of the Plant *Cis*-acting Regulatory DNA Elements (PLACE) database (Higo and Ugawa Y, 1999) and the PlantCARE database

(Lescot et al., 2002) found no match to previously reported *cis*-elements in plants. We named this motif the ABREN for ABA Responsive Element Novel.

The high correlation between ABA induction and the presence of the ABREN suggested that the ABREN might play a role in ABA response

The region 1,000 bp upstream from the start codon of the 2,443 ABA-inducible genes were compiled and scanned for the presence of the ABREN. As can be seen in Figure 4.3B, the number of occurrences of the ABREN, as well as number of genes that contained an ABREN, was higher in genes that showed high ABA induction. The correlation between ABA induction and the ABREN showed the same trend as that of the known ABRE (Figure 4.3A). The data was normalized by dividing the number of the elements found by the calculated expected number. This normalization allowed direct comparison of the frequencies of the ABREN and ABRE frequencies (Figure 4.3C). The expected value was calculated based on the formula in the Materials and Methods section. A value of 1.0 means the element occurs at a frequency expected from a random sequence. As can be seen in Figure 4.3C, both the ABREN and ABRE are enriched in all categories and their enrichment increases with increasing ABA induction. The ABREN has a higher level of enrichment than the ABRE in all categories of ABA induction. Genes that were infinitely induced by ABA, i.e. no reads mapped on the control sample, were not included in this analysis since an accurate ABA fold induction could not be calculated.

To ensure the results were not due to chance, we also scanned for the presence of AATTCCGG as a control sequence. This control sequence was chosen because it has the same length and number of each base as the ABREN, and thus has the same probability of appearing in a random sequence as the ABREN. This sequence is non-palindromic and not known to be a DNA binding site for

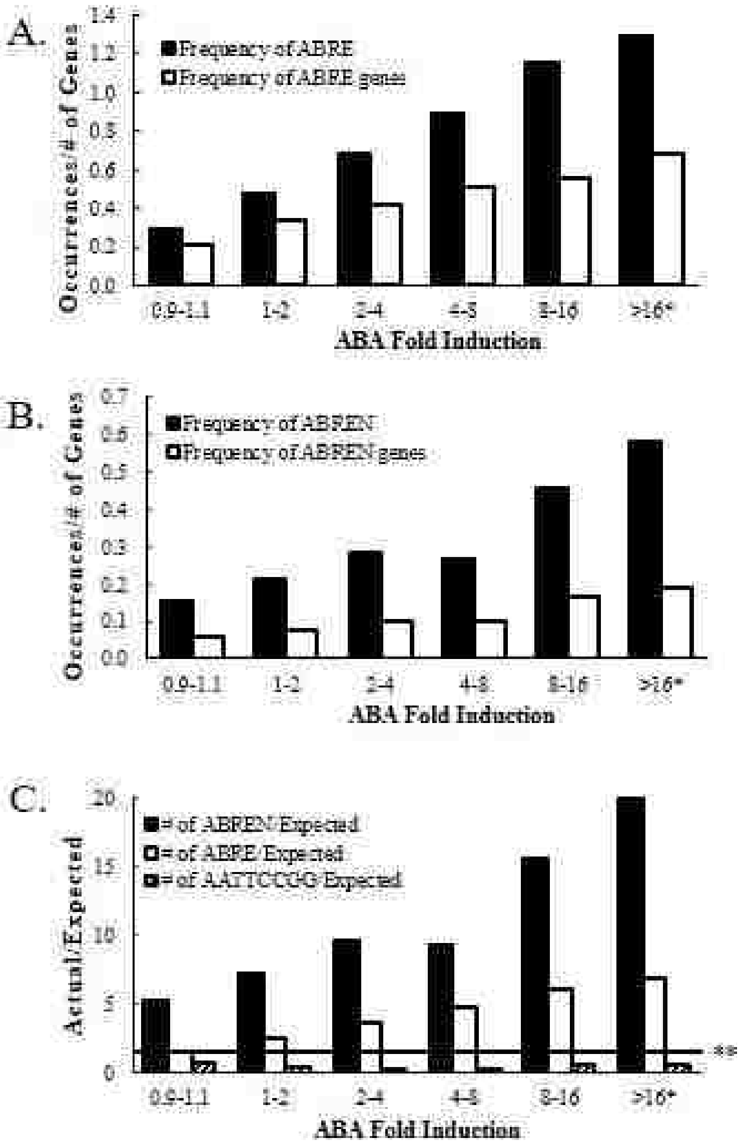


Figure 4.3: There is a strong correlation between the frequency of the ABREN and the ABRE elements, and the level of ABA induction.

The frequency of the A) ABRE, and B) ABREN in the 1,000 bp region upstream of the ATG start codon increases with increasing ABA induction. Solid black bars represent number of occurrences of the ABRE or ABREN divided by the number of genes in the category. White bars represent the number of genes that contain the ABRE or ABREN divided by the number of genes in the category. The results were normalized C) so that the ABREN and ABRE could be directly compared. The analysis was also performed on AATTCCGG as a control. The AATTCCGG showed no correlation with ABA induction. *Genes that had infinite ABA induction (i.e. no reads aligned on the control sample), were not included in the analysis. ** If the number of occurrences divided by the calculated expected number of occurrences is 1.0, then the motif is appearing at a frequency equal to that expected from a random sequence.

transcription factors. The control sequence showed an underrepresentation in all categories and there was no correlation with frequency of the AATTCCGG and ABA induction Figure 4.3C. Other control sequences were also tested with similar results.

We scanned the various regions of genes for occurrences of the ABREN, ABRE and AATTCCGG (Figure 4.4A) to determine the overrepresentation of these elements in other parts of the gene. The number of occurrences of the ABREN was 6.4-fold higher than expected in the 1,000 bp upstream regions of all rice genes, 14.7-fold in the 5' UTR region and over 5.7-fold the 3' UTR region. In all other regions, the occurrence was less than four-fold higher than the expected. On the other hand, the ABRE was enriched by 1.5-fold in the 1,000 bp upstream promoter region. In all other regions, the ABRE was underrepresented with fewer occurrences than expected. The number of occurrences of the control sequence AATTCCGG appeared at a frequency below what was expected (0.5 to 0.7) in all regions.

Further analysis was performed on the ABA-inducible genes to determine if the enrichment of the ABREN correlated with ABA induction. The number of occurrences of the ABREN was 14.7-fold greater than that expected in the 5' UTR when all genes are considered. When considering the 2,443 genes that were significantly ABA induced, the frequency of the ABREN increased to 21.4-fold greater than expected in the 5' UTR (Figure 4.4B). When considering only the 251 highly ABA induced genes, the frequency of the ABREN in the 5' UTR was 46.1-fold greater than expected. In the 1,000 bp region upstream of the start codon, which included the 5' UTR, the ABREN frequency was 6.4 fold higher than expected in all genes. In the ABA-induced genes, the

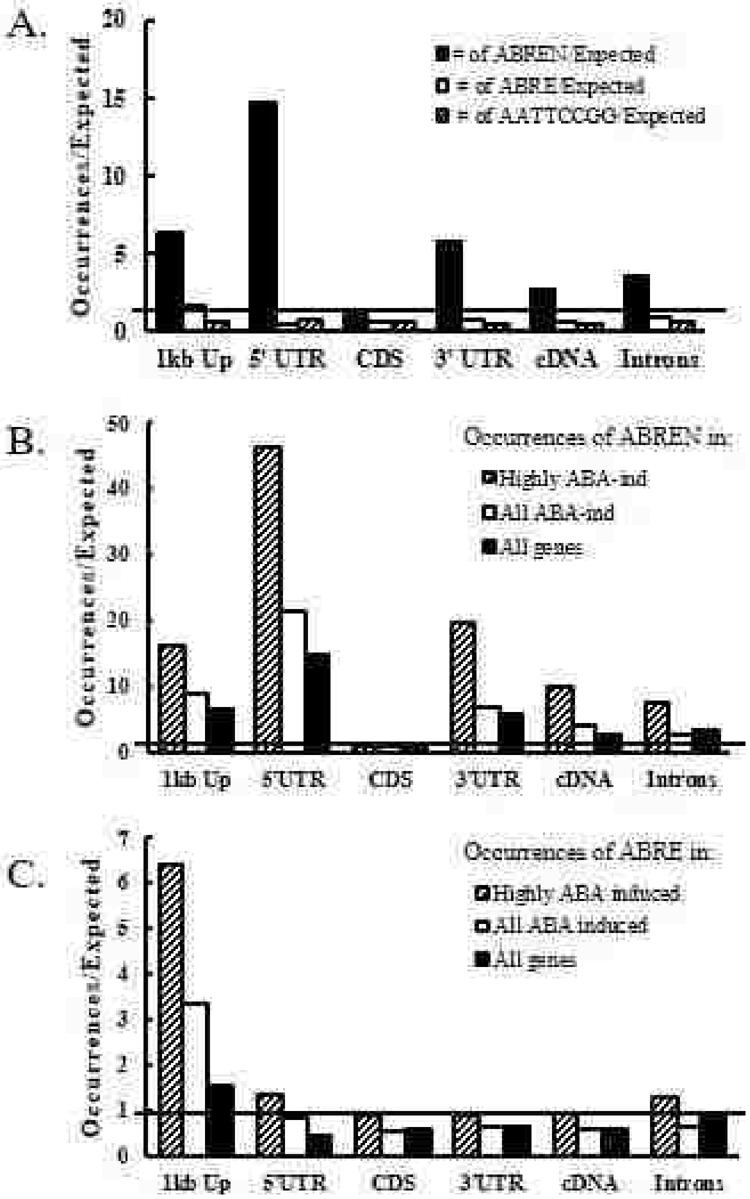


Figure 4.4: Compared to the ABRE, the ABREN appears at a very high frequency in the 1000 bp upstream and 5'UTR regions of all genes and even higher in ABA-inducible genes.

A.) The number of occurrences of the ABREN, ABRE and the AATTCCGG control sequence was determined in each of the regions of rice genes and normalized by dividing by the calculated expected number of occurrences. The number of occurrences of B) the ABREN and C) the ABRE was determined in each region for all genes, ABA inducible genes, and highly ABA inducible genes (> four-fold ABA-induced). The 1kb upstream region is 1000 bp upstream of the ATG start codon. CDS excludes introns, 5' UTR and 3' UTR. cDNA excludes introns but includes 5' UTR and 3' UTR.

ABREN was 9.0-fold greater than expected and in the highly ABA-induced genes the ABREN was 16.1-fold greater than expected. This correlation strongly suggests that the ABREN played a significant role in the ABA response.

The ABREN was also enriched in the 3' UTR and intronic regions of the highly ABA-inducible genes, by 7.6-fold and 5.5-fold, respectively. This data indicated that the ABREN may be able to induce gene expression from locations other than the promoter region of genes. The ABREN was enriched in the cDNA by 5.4 fold, however, the cDNA included the 5' UTR and 3' UTR. In contrast, the ABREN was underrepresented by 0.41 fold in the coding DNA sequences (CDS) of the ABA-inducible genes. Therefore, the enrichment of the ABREN in the cDNA was due to the enrichment in the 5' UTR and 3' UTR. The underpresentation of the ABREN in the CDS demonstrates that the ABREN does not have a function in the translated regions of genes, as generally expected for *cis*-regulatory elements. The ABRE showed enrichment only in the 1,000 bp upstream region of ABA induced genes and highly ABA induced genes Figure 4.4C.

The ABREN was enriched in the 100 bp region upstream of the start codon

We analyzed the distribution of the ABREN within the 1,000 bp upstream region of all rice genes that contained the ABREN in order to determine if there was a preferential distance between the ABREN and the start codon. This distribution analysis showed an increase in the frequency of the ABREN within the region 100 bp upstream of the ATG start codon (Figure 4.5A). Of all the ABREN elements, 35.8% of the time it was within 100 bp of the start codon. When the 100 bp region was further analyzed, the ABREN continued to increase down to 25 bp of the start codon.

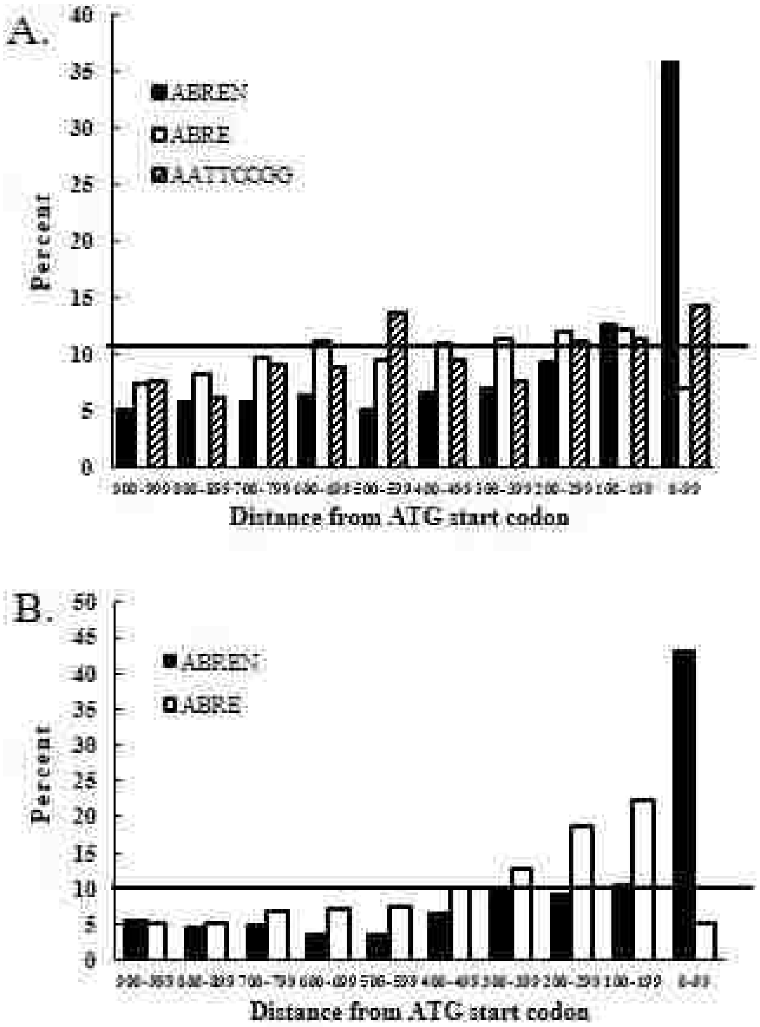


Figure 4.5: The ABREN was preferentially located near the start codon.

A) Distribution analysis of the ABREN in the 1000 bp upstream region of all rice genes showed that the ABREN occurred more frequently within 100 bp of the ATG start codon. The ABRE and the AATTCCGG control sequence were relatively flat across the region. B) In the ABA-inducible genes, the frequency of the ABREN near the start codon was more pronounced and the ABRE also showed preference toward the start codon. The AATTCCGG did not occur frequently enough in ABA-inducible genes for comparison

Both the ABRE and the AATTCCGG control sequence did not show any preference with respect to the distance between the sequence and the start codon. In the ABA-inducible genes, however, 46% of the ABREN elements found were within 100 bp of the start codon (Figure 4.5B). The ABRE also showed an increased presence in the 100-199 bp and 200-299 bp regions upstream of the start codon confirming that the ABRE plays a role in ABA induction.

The role of ABREN in mediating ABA response was experimentally verified.

To experimentally verify that the ABREN played a role in the ABA response pathway of plants, the 1,000 bp upstream region of all the ABA-inducible genes were scanned to identify genes that contain the ABREN but not the canonical ABRE (ACGTG(G/T)C) (Shen et al., 2004) and CE elements (GCGTGGC and TGCCACCGG) (Shen and Ho, 1995; Shen et al., 1996). One such gene was identified, the oleosin gene (LOC_Os09g15520) that was highly ABA-inducible but did not contain the ABRE or CE elements defined in the previous studies. Oleosins are amphiphilic proteins that work with phospholipids to stabilize the oil bodies. Oleosins further prevent the oil bodies from coalescing by providing steric hindrance (Tzen and Huang, 1992) and are also known to be ABA-inducible (Zou et al., 1995). The 1,000 bp upstream region of the oleosin gene contained two copies of the ABREN, one in the 5' UTR and one in the promoter region. The oleosin gene was over 12 fold ABA-induced and the expression level on the ABA treated sample was 449 FPKM based on our RNA-seq data.

To determine if the ABREN was indeed involved in the ABA response, a 322 bp region upstream of the oleosin gene was cloned into a plasmid upstream of a *GUS* reporter gene. This region contained two ABRENs separated by a TATA box (Supplemental Figure S11). The ABREN in the 5' UTR and promoter regions were mutated respectively or in combination. The constructs were then introduced into rice seeds by particle bombardment. Student's T tests demonstrated that the ABA induction level was significantly reduced when the ABREN in the promoter was mutated or both ABRENs were mutated (Figure 4.6). The mutation of the ABREN in the 5' UTR showed a reduction in ABA induction, however, it was not statistically significant. This suggests that the ABREN in the promoter plays a more significant role in ABA induction. The particle bombardment was performed three times to confirm the difference between the WT oleosin promoter and the double mutant and the results were consistent.

Upon close examination of the oleosin promoter, an ABRE-like element (ACGTGCC) and CE3-like element (GCGTGGC) were identified upstream of the TATA box in the promoter region. Mutation of the ABRE-like and CE3-like elements reduced ABA induction to levels similar to the double mutant ABREN. This demonstrates that these elements can also couple to form an ABRC to regulate ABA-induced expression.

Gene ontology enrichment analysis of rice genes containing the ABREN showed a correlation with stress response and transcription

Gene ontology enrichment analysis (Cao et al., 2012) of the promoter regions of ABREN containing genes were analyzed to determine if there was a relationship between the presence of an ABREN and the potential functions of genes. The resulting pie chart showed the number of genes found in each gene ontology category based on biological process (Figure 4.7A). Though there were many categories, the largest category (15%) of genes belonged to the category "response to stresses" which included response to freezing, response to oxidative stress, defense response and others. This was an indication that the ABREN might play a role in the response of rice plants to stresses.

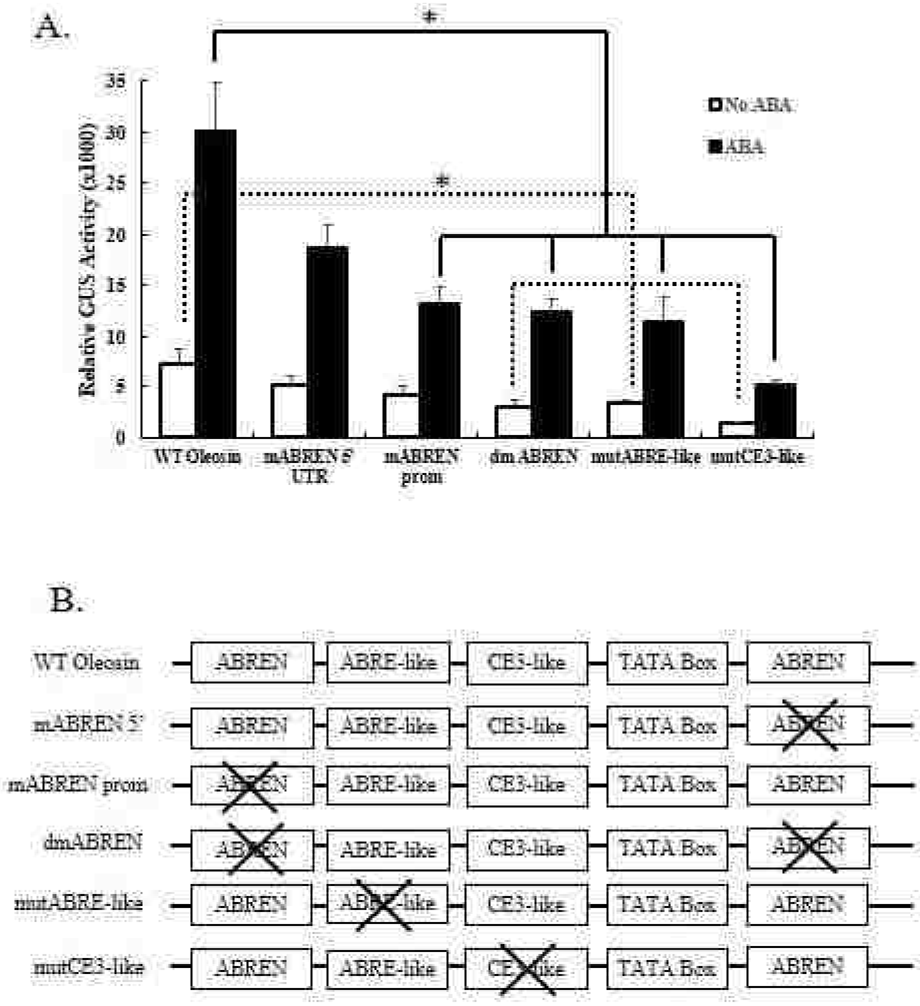


Figure 4.6: When both ABREN in the 5' UTR and promoter regions of the oleosin gene are mutated, the level of GUS expression upon ABA treatment is reduced.

The upstream region of the oleosin gene contains two ABREN motifs, one in the 5' UTR and one in the promoter. The wild-type upstream region of the oleosin gene (WT oleosin) was highly ABA-inducible. Pairwise Student's T test of the ABA treated samples show that the mutated ABREN in the 5' UTR is not statistically significant from the WT (p=0.10), however, the mutated ABREN in the promoter, the double mutant ABREN, the mutated ABRE-like and the mutated CE3-like are significantly different from the WT. In the non-ABA treated samples, the WT is significantly different from the double mutant ABREN, the mutant ABRE-like and the mutant CE3-like.

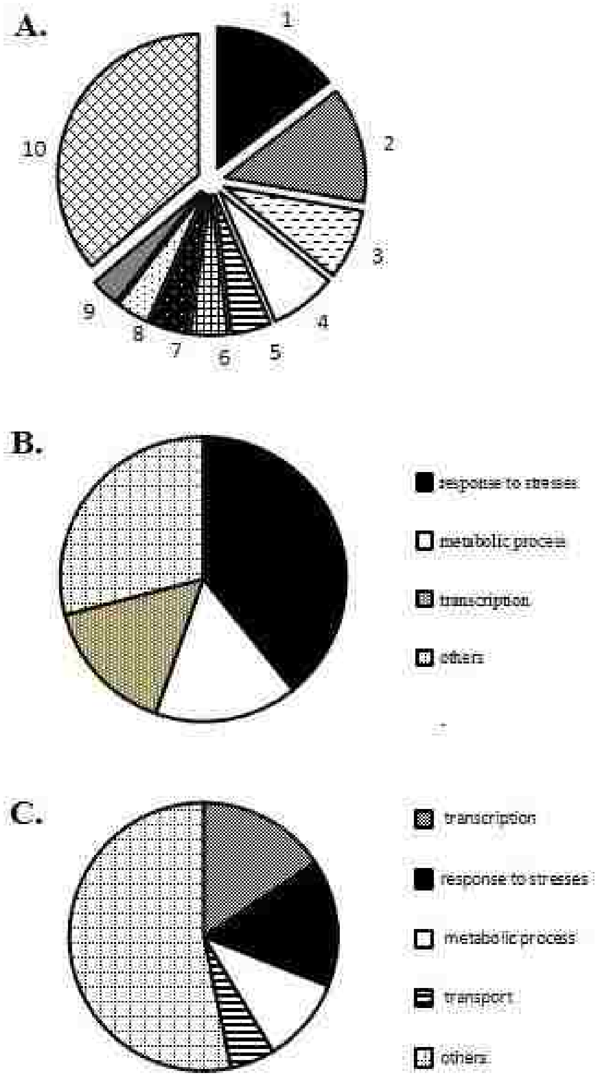


Figure 4.7: Gene ontology analysis shows that many of the genes that contain the ABREN in the 1000bp upstream region were related to stress and transcription.

A.) 1) About 15% of genes that contain the ABREN in the 1000 bp upstream region were related to stress (freezing, homoiothermy, oxidative stress, defense response). **2)** About 13% of the genes that contain the ABREN were related to transcription. Both **3)** proteolysis and **4)** metabolic process genes made up 8% of the genes. **5)** Transport genes made up 5% of the genes. The remaining categories include **6)** DNA-integration (4%), **7)** oxidation-reduction (4%) **8)** RNA-dependent DNA replication (4%) **9)** protein phosphorylation (3%). **10)** The categories that were composed of fewer than 3% of the genes were combined into the category labeled “other”. **B.)** For the highly ABA inducible genes, the largest category is response to stress, followed by metabolic process, transcription and response to water. **C.)** For all ABA inducible genes, the number of genes involved in transcription is higher than those involved in the response to stress.

The next largest category (13%) was transcription which included transcription factors, regulation of transcription and DNA-dependent transcription. This was also an indication that the ABREN may play a role in gene regulation through transcription.

The ABREN was not enriched in the dicot species Arabidopsis thaliana

We found that the ABREN was highly enriched in the 1,000 bp region upstream of the ATG start codon of ABA-inducible genes in rice. To determine whether the ABREN plays a role in ABA responses in other plant species, we analyzed two sets of ABA-inducible Arabidopsis genes. The first set includes 102 genes that were induced by at least seven-fold after ABA treatment, based on the microarray data (Seki et al., 2002). Of these genes, not one gene contained the ABREN in the promoter region, and only one occurrence of the ABREN was found in the 5' UTR (Supplemental Table S10). The second set contains 141 ABA-inducible genes, also as determined by microarray analyses (Li et al., 2006). Only six occurrences of the motif were found in the promoter region, and one occurrence was found in the 5' UTR region (Supplemental Table S10). The combined total of the results from both sets of ABA-inducible genes was approximately the same as that expected from a random sequence of nucleotides. This data showed that the ABREN is not enriched in ABA-inducible genes in Arabidopsis.

Discussion

In this study, it was shown that the change in ABA sensitivity was due to sugar rather than the application of α-amylase by the comparing the expression levels of two genes containing SRC elements in their promoters (Supplemental Table S9). We have demonstrated that sugar signaling and ABA signaling are more closely related than previously thought. Almost all the genes that were highly ABA inducible in the absence of α-amylase (88%) had an even higher level of ABA-induction in the presence of α-amylase (Figure 4.1). From the highly ABA-inducible genes identified by our RNA-seq data, we have identified a novel ABA-inducible *cis*-regulatory element we named ABREN. The ABREN has a strong correlation between its frequency and the level of ABA induction (Figure 4.3) as does the ABRE. The ABREN was enriched in the promoter region, 5’ UTR and 3’ UTR (Figure 4.4), although it is preferentially located near the start codon (Figure 4.5). Gene ontology analysis shows that many of the genes that contain the ABREN in the 1000bp upstream region were related to stress and transcription (Figure 4.7). Transient expression experiments confirmed that mutation of the ABREN did indeed reduce the response to ABA treatment (Figure 4.6).

Role of sugar in ABA response

Sugar signaling has been well-studied and was reported to be involved in ABA biosynthesis. Glucose is perceived by hexokinase which leads to increased expression of ABA synthesis proteins such as SDR1, ABA3, and AAO3 in Arabidopsis (Cheng et al., 2002). Glucose also mimics ABA-induced seedling developmental arrest, and both glucose and ABA induce ABI3, ABI5 and LEA genes demonstrating overlap in signaling pathways (Dekkers et al., 2008). It has been shown that ABA and glucose synergistically regulate the expression of some genes. For example, microarray analyses showed that the expression levels of 95 out of 692 ABA-inducible genes (14%) were enhanced by glucose in Arabidopsis (Li et al., 2006). This was generally supported by our RNA-seq data. However, our data also showed that up to 88% of ABA-inducible genes were enhanced by sugars. This data confirmed the overlap of ABA and sugar response networks and demonstrates that the relationship between ABA and sugar is much higher than previously thought.

Identification of the ABREN and experimental verification of its role in ABA induction

There have been several reported *cis*-acting elements that are responsive to ABA. The fact that only 39.5% of ABA inducible genes in rice aleurone contain an ABRE, and that other motifs have been shown to be ABA inducible suggests that there may be undiscovered ABA responsive elements. From a set of 251 highly ABA-inducible genes identified by RNA-seq, a novel ABA responsive element, the ABREN, was identified which showed a very strong correlation between the frequency and the level of ABA induction that was higher than the known ABRE (Figure 4.3A). The ABREN was enriched in the promoter region and is preferentially located near the start codon (Figure 4.5). All these lines of evidence suggest that the ABREN is a *cis*-regulatory element involved in the ABA signaling pathway.

The promoter region of the highly ABA inducible oleosin gene was used as the subject for analysis since it contains two copies of the ABREN, one in the promoter and one in the 5' UTR. Transient expression experiments confirmed that mutation of the ABREN in the promoter region significantly reduced the response to ABA treatment (Figure 4.6). Mutation of the ABREN in the 5' UTR region also showed a decrease in GUS activity but the result was not statistically significant (p value = 0.10). The double mutation of the ABREN in the 5' UTR and the promoter showed a reduction of GUS activity similar to the mutation of the ABREN in the promoter. This data seems to indicate that the ABREN in the promoter plays a bigger role in ABA induction than the ABREN in the 5' UTR. This seems contrary to the bioinformatics analysis which suggests that the ABREN closer to the start codon plays a bigger role in ABA induction since it is more abundant near the start codon. There are reasons that may explain why the ABREN in the promoter plays a greater role in the ABA response.

The flanking regions to the ABREN may play a role in the ability for the ABREN to regulate ABA response. Flanking regions around the ACGT core of the ABRE have been shown to play a role in bZip binding to the ABRE (Shen et al., 2004; Williams et al., 1992). The flanking regions of the ABREN in the promoter are different than those in the 5' UTR. It is possible that there are several ABREN binding proteins that selectively bind to the ABREN based on the flanking regions. The ABREN in the promoter may have a flanking region that is more appropriate for the ABREN binding protein than the flanking region to the ABREN in the 5' UTR.

It has also been well demonstrated that promoter elements may not function alone but require another element to couple with in order to regulate the ABA response. The ABRE has been demonstrated to couple with coupling elements CE1 and CE3 (Shen and Ho, 1995). It has also been demonstrated that the distance between elements could play a role in ABA induction. A longer distance between elements correspond to a lower level of ABA response when ABRE couples with CE3 and distances that are multiples of 10 bp tend to have higher ABA response when ABRE couples with CE1 (Shen et al., 2004). Multiples of 10 bp suggest that the ABRE and CE1 must be on the same side of the DNA double helix. It may be possible that the ABREN can couple with the ABRE-like or CE3-like elements to cause ABA induction. The ABREN in the promoter is closer to the ABRE-like element (79 nt) than the ABREN in the 5' UTR (134 nt). Also the ABREN in the promoter is one base pair from being a multiple of 10 bp from the ABRE.

Also, there is the possibility that there are other elements within the 322 nt region upstream of the oleosin gene. The 79 nt region between the ABREN in the promoter and the ABRE-like element as well as a 55 nt upstream of the ABREN in the promoter may contain other undiscovered ABA responsive elements. One or more elements in these regions may couple with the ABREN in the promoter to cause an ABA response. A coupling element in these regions would be closer to the

ABREN in the promoter than the ABREN in the 5' UTR, thus making the ABREN in the promoter the dominant element in the ABA response.

Redefinition of the ABRE and CE elements in rice

In addition, mutations of the ABREN in the promoter and 5' UTR, mutations of the ABRE-like (ACGTGCC) and CE3-like (GCGTGGC) elements also showed reduction in ABA response demonstrating that these two elements may couple each other and possibly to the ABREN to form an ABRC to regulate ABA induction. Previous studies by linker-scan have shown the ABRE-like element to have lower level of ABA induction and a lower level of reporter gene expression compared to the currently accepted ABRE consensus ACGTG(G/T)C (Hattori et al., 2002; Shen et al., 2004). Thus, the ABRE-like element was not considered an ABRE. A similar conclusion was made with the CE3-like element (Shen et al., 2004). The reduction in ABA-induction and expression levels may be due to the different flanking regions to the ABRE and CE3 elements compared to those in the oleosin promoter. Also, the absence of the ABREN in the constructs may also play a role. Since mutation of the ABRE-like and CE3-like elements in the oleosin promoter show substantial reduction in ABA induction, we propose that the new consensus sequence for the ABRE and CE3 elements in rice are ACGTG(G/T/C)C and GCGTG(G/T)C respectively.

Enrichment of the ABREN in non-coding regions suggests the ABREN has enhancer-like gene regulation

The increasing enrichment of the ABREN with increasing ABA induction in non-coding regions of genes was similar to the increasing enrichment of the ABRE with increasing ABA induction in the 1,000 bp upstream region. This was strong evidence that the ABREN plays a regulatory role in the ABA response. The fact that the ABREN enrichment increased in the 5' UTR and 3' UTR indicated that the ABREN can regulate gene expression from regions other than the promoter. The analysis of the AATTCCGG and other control sequence showed no enrichment in all regions of the gene as expected by a motif with no regulatory significance.

Regulation of gene expression from regions other than the promoter are typical of enhancer elements, however, enhancer elements are typically a few hundred base pairs in length and are usually a long distance from the gene they regulate (1 kb to 1 Mbp) (Spitz and Furlong, 2012). Since the ABREN sequence was only eight base pairs in length and it was located very close to the gene that it regulated, calling it an enhancer may be a misnomer. However, the possibility that the ABREN was a posttranscriptional regulatory element was not excluded. Regulatory RNA binding proteins that bind to the 5' UTR of mRNA molecules have been reported (Stripecke et al., 1994). Also, regulatory motifs have been identified in the 3' UTR of mammalian species (Xie et al., 2005) and were thought to be target sites of miRNAs. These previously identified motifs have a strong directional bias and have a strong peak length of 8 nucleotides, the same length as the ABREN. However, the ABREN was an inverted-repeat-palindrome, so all ABRENs have the same directionality by default.

The ABREN is a target for restriction enzymes

Although there is no known transcription factor that binds to the ABREN, the GATC sequence of the ABREN is identical to the target sequence of several restriction enzymes. Also, the methylation status of the restriction site determines whether or not the restriction enzyme can cleave the DNA (Supplemental Table S11). For instance, DpnI will only cleave the GATC if the adenosine base is methylated and the cytosine base is not methylated. DpnII will only cleave if the adenosine is not methylated. MboI will cleave only if both the adenosine and cytosine are not methylated. BufCI will cleave if the cytosine is not methylated. Though these restriction enzymes were derived from bacteria, it is conceivable that regulation of ABA inducible genes in rice could be regulated via epigenetic DNA methylation of the ABREN and methylation sensitive DNA binding factors. Further research needs to be performed to identify if there are DNA binding factors that bind to the ABREN and if they are methylation sensitive.

In summary, RNA-seq has shown that sugar upregulates ABA responsive genes and that sugar and ABA signaling pathways overlap more than previously thought. A novel ABA responsive element, the ABREN, was identified as enriched in ABA inducible genes in rice and experimentally confirmed via particle bombardment to play a role in ABA signaling. The ABREN could be further explored for use in biotechnological applications as a molecular switch to control genes to enhance plant stress tolerance. The ABREN was underrepresented in Arabidopsis and thus may be a monocot specific element. Redefinition of the ABRE and CE3 elements consensus sequences in rice have been proposed. These findings enrich the current knowledge of ABA and sugar signaling in rice and may lead to the development of more robust strains of rice and other crops.

Materials and methods

Software and data used to identify enriched cis-acting elements

In order to identify novel *cis*-regulatory elements that might be involved in the ABA signaling pathway, the ABA-inducible genes were identified by RNA-seq. Of the 55,986 genes in the MSU Rice Genome Release 7.0 dataset (Kawahara et al., 2013), 2,443 genes were significantly induced by ABA as determined by our three biological replicates of RNA-seq data of rice aleurone cells. The sequences consisting of 1,000 bp upstream of the ATG start codon of the 2,443 ABA-inducible genes were extracted from the MSU rice dataset. Bioprospector software (Liu et al., 2001), a C program that utilizes Gibbs sampling strategy, was used to identify sequences that were enriched in these genes. Bioprospector was run using default parameters and varying the searched motif length.

PERL scripts were written to identify the relative abundance of the ABREN in the various regions of all known genes in the genome of *Oryza sativa* ssp. japonica genome. The relative abundance of the ABREN was calculated by dividing the actual number of occurrences of the ABREN by the number of occurrences expected by random chance. The following equation was used to calculate the number expected by random chance:

$$E = X \prod_{i=1}^{n} P_i$$

where E is the expected number, X is the total number of bases in the region of interest, n is the number of bases in the motif of interest, and P_i is the probability of the ith base of the motif occurring. P_i is calculated by the number of occurrences of the ith base in the region of interest divided by X.

Search for the ABREN within various regions of genes

A PERL script was written to search the forward and reverse strands of DNA for motifs within a FASTA file. This script was used to search for the ABREN, ABRE and control sequences within the promoter region, 5’ UTR, 3’ UTR, intron, cDNA, and CDS’s of the rice genome.

Preparation of constructs

A fragment upstream of the start codon of the ABA-inducible oleosin gene (LOC_Os09g15520) was amplified via PCR using the forward and reverse primers (Supplemental Table S12). The PCR product was then inserted upstream of the minimal promoter in the negative control. Two ABRENs were present in the cloned fragment, one in the promoter and the other in the 5’ UTR. These elements were mutated individually or in combination via site-directed mutagenesis (Kunkel, 1985), as detailed in Supplemental Table S12. The mutations were confirmed by DNA sequencing.

The pAHC18 (*Ubi1-Luciferase*) construct, containing the luciferase reporter gene driven by the constitutive maize ubiquitin promoter (Bruce et al., 1989), was used as an internal control construct to normalize GUS activity of the reporter construct as previously described (Shen et al., 1993)..

Particle bombardment of constructs into rice aleurone cells

Rice aleurone was prepared and transiently transformed by particle bombardment. Briefly, embryoless seeds were surface sterilized by 1.5% sodium hypochlorite for 60 minutes, then washed with sterile water 10 times. These embryoless seeds were imbibed on vermiculite and a filter paper saturated with 1/2 MS plus 100 mM glucose medium for 72 hours. The pericarp was peeled from the seeds and the seeds were longitudinally cut in half. The starch was scraped from the half-seeds using a sterile razor blade leaving the thin layer of aleurone. The aleurone layers from 16 half seeds were arranged on a Whatmann filter paper saturated with shooting buffer in a square plate.

DNA for each reporter was mixed with the DNA of the internal control construct and cobombarded in the molar ratio indicated for each experiment. After bombardment, the aleurone layers were treated with or without 20uM ABA for 24 hours. The bombarded aleurone layers were ground and assayed for luciferase and GUS activities as previously described (Wang et al., 2007).

Total RNA isolation from rice by guanidinium-sulfate-phenol-chloroform extraction

Rice seeds were obtained from the USDA ARS, Dale Bumpers National Rice Research Center. The rice seeds were washed then soaked in imbibing solution for 24 hours at 25° C. The seeds were vertically cut in half with a sterile scalpel and soaked in imbibing solution for another 24 hours after which the starchy endosperm was scraped away from the aleurone layer. The aleurone was treated with 20uM ABA for four hours. RNA was extracted from the treated aleurone with slight modifications to the RNA extraction technique described previously (Li and Trick, 2005). In short, the aleurone was ground in liquid nitrogen and lysed in a lysis buffer. The nucleic acids were separated from the protein via a phenol-chloroform extraction. A guanidinium-sulfate extraction buffer was used to extract the total RNA. The RNA was precipitated out of solution via isopropanol-sodium chloride precipitation. RNA quality was confirmed by agarose-formaldehyde gel electrophoresis and BioRad Experion Automated Electrophoresis System. The RNA was sent to the Huntsman Cancer Institute, University of Utah for RNA-seq on an Illumina HiSeq 2000 Sequencing System. Three biological replicates were performed for each sample type.

Short-read alignment

Short reads generated from the RNA-seq analysis were single-ended reads, 50 nucleotides in length. Short read data was aligned to the MSU rice genome release 7.0 (MSU R7) by the Bowtie software (Langmead et al., 2009) using the –best, –n 2, –l 28, and –e 70 options. Reads with more than two mismatches in the first 28 base pairs or reads with a total Phred quality score of mismatched bases exceeding 70 were excluded. For reads that aligned to more than one location, the location with the fewest mismatches was chosen. The UCSC Genome browser (Kent et al., 2002) was used for viewing of the short read alignment data (http://shenlab.sols.unlv.edu/cgi-bin/hgGateway).

Gene ontology enrichment analysis

Gene ontology analysis was performed using the Rice Oligonucleotide Array Database (http://ricearray.org/analysis/go_enrichment.shtml). Only ontology categories with a hyper p value less than 0.05 were used.

CHAPTER 5

TRANSCRIPT STRUCTURE AND DOMAIN DISPLAY: A CUSTOMIZABLE TRANSCRIPT VISUALIZATION TOOL

Previously published as:

Transcript structure and domain display: a customizable transcript visualization tool

Kenneth A. Watanabe, Kaiwang Ma, Arielle Homayouni, Paul J. Rushton and Qingxi J. Shen

Bioinformatics *(2016) doi: 10.1093/bioinformatics/btw095*

Disclaimer:

KAW created the MySQL database, wrote most of the PHP code and participated in the HTML and JavaScript coding. KM wrote most of the HTML and JavaScript code. AH participated in the HTML coding and reviewing of the manuscript. PR participated in reviewing the manuscript. JQS supervised the design and coordination of the software development. All authors read and approved the final manuscript.

Abstract

Transcript Structure and Domain Display (TSDD) is a publicly available, web-based program that provides publication quality images of transcript structures and domains. TSDD is capable of producing transcript structures from GFF/GFF3 and BED files. Alternatively, the GFF files of several model organisms have been pre-loaded so that users only needs to enter the locus IDs of the transcripts to be displayed. Visualization of transcripts provides many benefits to researchers, ranging from evolutionary analysis of DNA-binding domains to predictive function modeling.

Introduction

Due to high demand for publication quality transcript visualization software, there have been several software solutions available for researchers. Each of these solutions has advantages and drawbacks. For instance, GECA (Fawal et al., 2012), which can be used online or downloaded to a server for offline usage, has a relatively quick runtime, and has an appealing user interface. However, it is difficult to use and lacks customizable features and custom motifs. GECA also requires the DNA sequence, protein sequence, and a GFF3 file which could pose a problem for those who may not have this information available. FeatureStack's (Frech et al., 2012) advantages include relatively high quality images and the capacity to display custom domains. However, FeatureStack lacks an online version, requiring users to download and install a program onto their server. FancyGene (Rambaldi and Ciccarelli, 2009) is highly customizable, allowing selection of feature colors, feature sizes and custom motifs. However, it can only display a single gene at a time. GSDraw (Wang et al., 2013) produces customizable, high quality images with the capability of custom motifs. In addition, GSDraw provides users with a phylogenetic tree. However, GSDraw has a long run time, requires both the genomic and CDS sequences, and cannot utilize a GFF3 file. GSDS 2.0 (Hu, et al., 2015), is a fast, easy to use, web-based program that requires only a GFF3 file and produces customizable results that include custom motifs and a phylogenetic tree. However, GSDS has some limitations. GSDS can only process a maximum of 50 genes at a time, which may pose a problem when studying large gene families. For example, the WRKY gene family contains well over 100 genes in many species (Eulgem et al., 2000; Zhang and Wang, 2005). Other limitations of GSDS include the inability to select a font or font size. GSDS also lacks the capability to vertically space genes so when the size of gene features is enlarged, the genes overlap. Herein, we report an alternative and enhanced transcript display software that is fast, easy to use, web-based and resolves issues with other transcript visualization software. TSDD is capable of providing researchers with publication quality images that can be customized and downloaded.

Usage and Implementation

Transcript Structure and Domain Display (TSDD) can be accessed from our website at: http://shenlab.sols.unlv.edu/shenlab/software/TSD/transcript_display.html. TSDD relies on GFF/GFF3 files to draw the transcript structures and domains. GFF/GFF3 files were used as a data source to produce the transcript structures since almost all annotated genomes use this file format. The GFF files and protein sequences of 13 model organisms were preloaded into the TSDD database. To generate transcript structures of one of these organisms, users only need to select the organism from the pull-down menu and then enter the locus IDs of the genes they wish to display into the text box. Alternatively, users can load a ".txt" file of the locus IDs from their PC. For other organisms, users can select custom GFF3 or BED file from the pull-down menu and then can either paste the gene structure data into the textbox or load a file from their PC with the extension ".gff" or ".bed" respectively. Users can click the "Demo Data" button to load example data of the selected organism or data file so the format of the data can be viewed.

TSDD will then parse the data and generate the transcript structures (Figure 5.1A). TSDD can display at least 300 transcripts in a single run depending on the memory of users' PC.

TSDD is highly customizable to meet the needs of users. They can select from a variety of fonts and font sizes, and can also display an arrow indicating the orientation of the transcript (Figure 5.1B). The vertical spacing between transcripts can also be adjusted, as well as the color and height of the genomic features: UTR, CDS or Introns. Users can also select the color and height of up to 20 custom domains. If the user selects one of the preloaded organisms, they can specify the pattern of the domains they wish to display. In addition, they may choose to display either all of the domains that fit the pattern, only the first or last, or ones within a specified distance from the N- or C-terminus. Since some transcripts have long introns making the transcript difficult to view, TSDD also has the option to compress the introns by a specified percentage or even omit the introns or UTRs for easier viewing (Figure 5.1C). TSDD also has a preview feature on the main input screen which allows users to preview any changes to the default settings prior to generating the transcript structures. This feature will save users time if a large number of transcripts are being generated. TSDD also gives users the option to save their current configuration to a text file. This will allow users to quickly reload their data so that results can easily be replicated at a future time. The generated transcript structures can be saved as one of several file formats (PNG, PDF, JPG, GIF, BMP, TIFF), which can be downloaded to users' PC for viewing and editing. TSSD proves easy to use, fast, and customizable, in addition to providing features that none of the current alternatives provide.

Future direction

We plan on building a database of domains and their consensus sequences using popular protein domain databases, such as Pfam (Finn et al., 2014), so that users do not need to populate the pattern field manually. We also plan on having TSDD scan the entered loci for any domains within our database. Users will then be able to select which of the identified domains to display. The list of pre-loaded organisms will grow over time as users request them to be added to our database. We will listen to the suggestions and feedback from the users so that we can continuously improve TSDD to meet the growing needs of the scientific community.

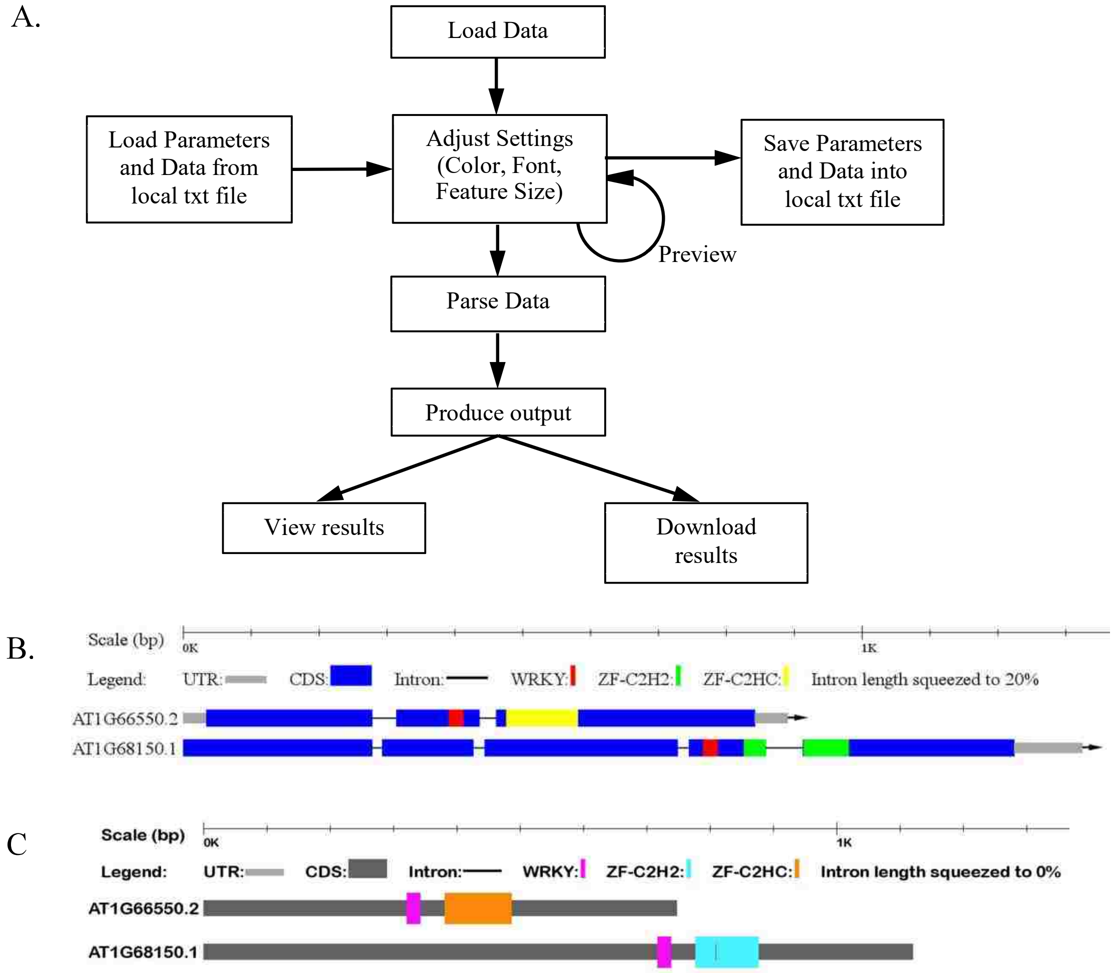


Figure 5.1: Transcript Structure and Domain Display (TSDD).

A) The workflow of TSDD for producing transcript structures. B) Example output showing transcript structures of two Arabidopsis genes using default settings. C) Alternative output showing customized output of the same two Arabidopsis genes.

Availability of software

Project name: Transcript Structure and Domain Display

Project home page: http://shenlab.sols.unlv.edu/shenlab/software/TSD/ transcript_display.html

Programming language: HTML, Javascript, PHP

License: Open Source license GNU General Public License version 2.0

Restrictions to use by non-academics: license needed

Acknowledgement

This work also received support from China Scholarship Council (Grant No.:201308410108)..

Funding

This work is supported by a USDA grant (2008-35100-04519).

Conflict of Interest: none declared.

CHAPTER 6

GENERAL DISCUSSION

With the growing world population and the decreasing amount of agricultural land due to climate change, there is a dire need to develop more robust cultivars of food crops to feed the world. Rice ties with wheat for the grain with the highest calories consumed per capita in the world with the highest per capita consumption in Asia (Figure 1.2). Rice was the choice grain to study because of its sequenced genome, smaller genome size and fewer genes than wheat. In this dissertation, bioinformatic approaches followed by experimental verification were performed to identify unannotated genes in the rice genome as well as identification of a novel ABA responsive element. Improvement of the rice genome annotation and improving our understanding of the hormone signaling network will undoubtedly aid in our ultimate goal of developing cultivars of rice, and possibly other crops, that have higher yield and are more robust to abiotic and biotic stresses, and thus, secure the world's food supply for our growing world population.

Brief summary of chapters

Figure 6.1 is a flowchart that summarizes the research performed throughout this dissertation. In brief, the mRNA was extracted from rice aleurone tissue. Then RNA-seq was performed to generate short reads. Then the short reads were mapped to the rice genome using Tophat software to define the transcriptome. From here, two pathways were taken. In one path, novel genes were identified via custom written programs, Tiling Assembly and Clustering Algorithm. To determine the function of the novel genes, hormone regulation, domain search and BLAST queries were performed. In the other path, the hormone induced genes were identified via the Cuffdiff software, and from the ABA induced genes, Gibbs Sampling was used to identify a novel *cis*-regulatory

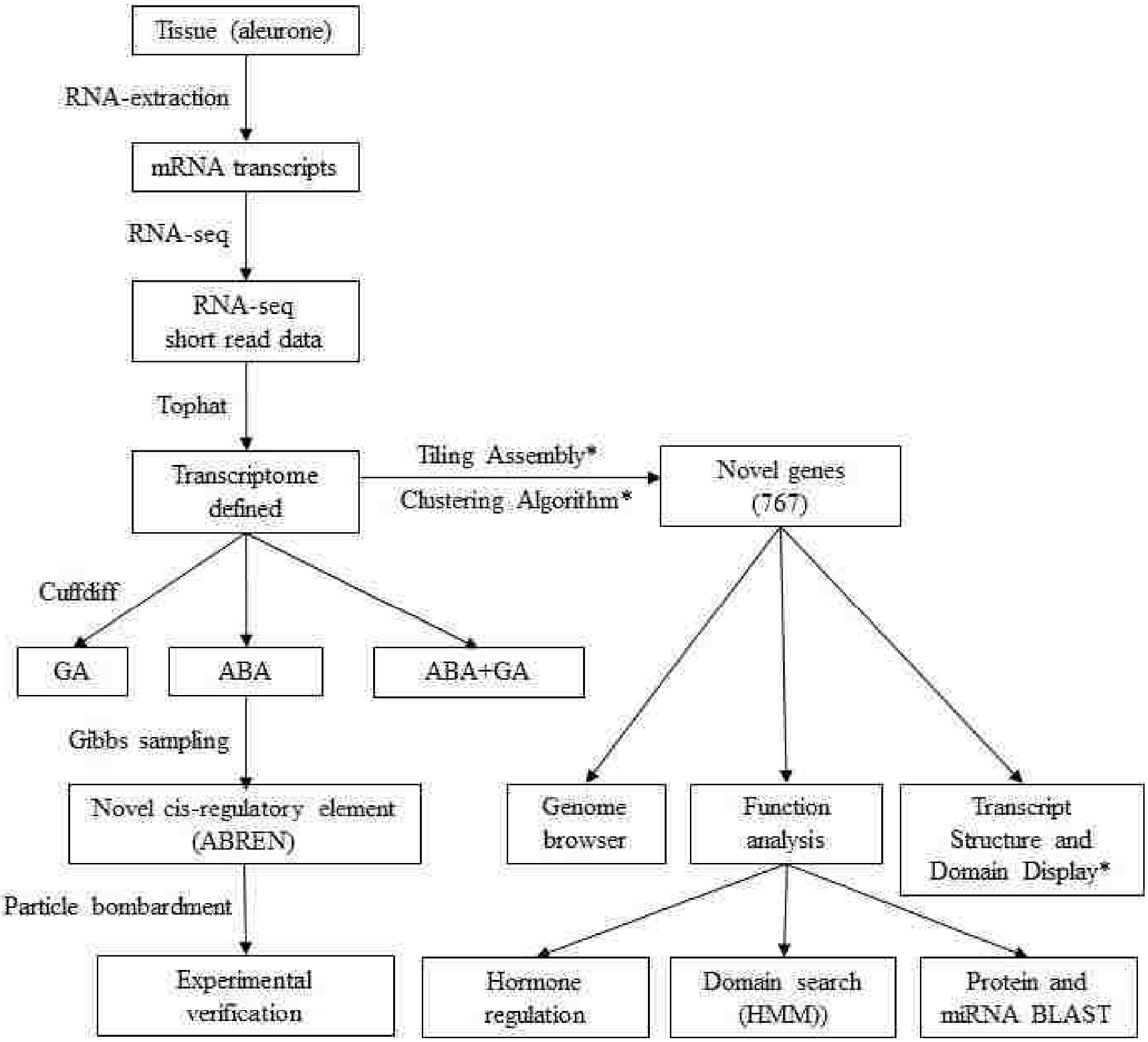


Figure 6.1: Flowchart summarizing the research performed in this dissertation

mRNA was extracted from rice aleurone cells and RNA-seq was performed. The resulting short reads were aligned to the genome via Tophat to define the transcriptome. Differential expression analysis via Cuffdiff determined the genes that are up- or downregulated by the various hormone treatments. Gibbs sampling was used to identify a novel *cis*-regulatory element (ABREN) which was experimentally verified by particle bombardment. Using custom gene finding software, Tiling Assembly and Clustering Algorithm, 767 novel genes were identified. The data were loaded into a genome browser and Transcript Structure and Domain Display for visual display of gene features. To determine gene function, hormone regulation, domain search and BLAST queries were performed. Items marked by an asterisk (*) were custom written software developed by myself.

element, ABREN. The ABREN was then experimentally confirmed to be involved in the ABA response by particle bombardment.

In chapter 2, I demonstrated by RNA-seq that the rice genome annotation is missing as many as 8% of the genes (Supplemental Table S1). Further bioinformatics analysis via the Cufflinks and a custom Clustering Algorithm has identified 553 high confidence novel genes in the rice genome (Supplemental Table S2). These novel genes are considered high confidence since they do not overlap with known annotated genes, they have sufficient expression, and they share low similarity with other regions of the genome. Experimental verification by RT-PCR on a subset of these genes (Figure 2.7) confirmed that mRNA transcripts are being transcribed making it difficult to deny these are real novel genes. Further analysis demonstrated that many of these novel genes show homology to protein- and/or microRNA-coding genes and many of the novel genes showed differential expression when treated with ABA, GA or both (Figure 6.2). This data demonstrates that these genes may play important roles in the hormone signaling pathways.

Having demonstrated that the rice genome is not complete, there is a need for improved gene finding software to analyze RNA-seq data. In chapter 3, I describe a novel gene finding algorithm, the Tiling Assembly algorithm and compared it to the popular Cufflinks software on a randomly generated genome with known genes inserted. It was demonstrated that Cufflinks and Tiling Assembly are equal in their capacity to identify the correct number of exons of genes at 100 RPKE or higher (Figure 3.1). However, the Tiling Assembly was more capable of identifying genes than Cufflinks, finding virtually all the highly expressed novel genes that Cufflinks identified, plus an additional 3,373 genes (Figure 3.3B). After filtering out the genes that show high similarity to another genomic region, 767 high confidence novel genes were identified (Supplemental Table

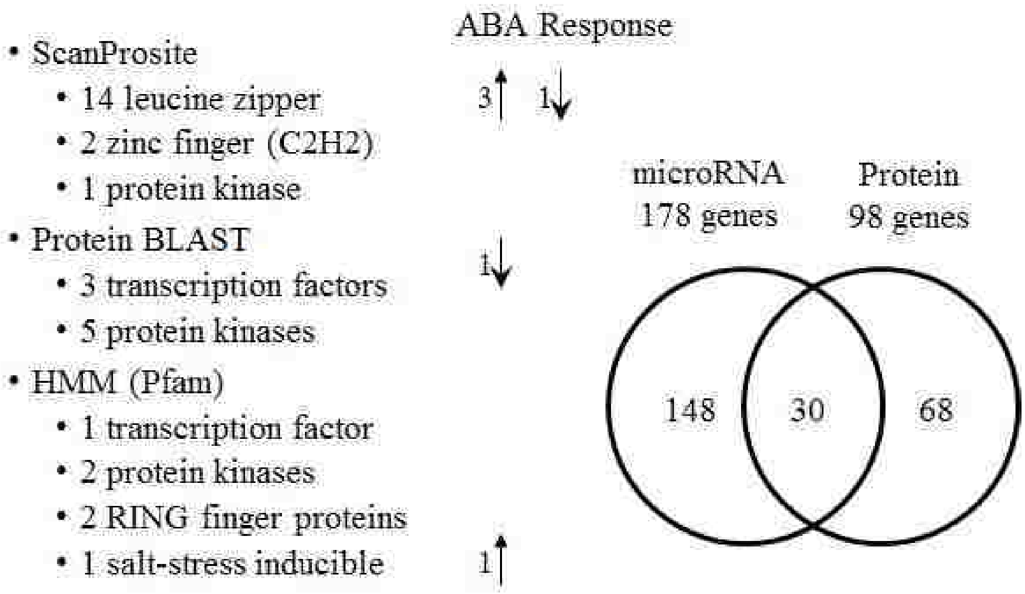


Figure 6.2: Summary of the analysis of novel genes identified

Analysis of novel genes was performed by scanning for known domains (ScanProsite), protein BLAST queries and Hidden Markov Models (HMM) of the Pfam domain database. RNA-seq data revealed some of the domain-containing novel genes were regulated by ABA. Nucleotide BLAST queries revealed many genes showed homology to protein-coding and microRNA genes.

S6). Comparison to our previously published results demonstrates that the Tiling Assembly is superior (Figure 3.5), however, there are still genes that have been identified by the other algorithms suggesting that further improvements are still possible for the Tiling Assembly.

In chapter 4, RNA-seq was used to identify 2,443 ABA inducible genes in rice aleurone cells (Figure 4.1). When comparing the RNA-seq data with the previous RNA-seq data of α-amylase treated rice aleurone, we have shown that α-amylase treatment increases the sensitivity of ABA inducible genes to ABA. Of the ABA inducible genes identified in both the α-amylase treated sample (AAT) and the non-α-amylase treated sample (NAAT), almost all of these had a higher ABA fold induction in the AAT data than the NAAT data. Analysis of the known sugar responsive genes showed a sensitivity to α-amylase treatment (Supplemental Figure S9), demonstrating that the sugar produced by the hydrolysis of starches by α-amylase is the causative factor enhancing the sensitivity of cells to ABA.

By use of Gibbs sampling, an enriched element was identified in the promoter region of the ABA inducible genes. We dubbed this element the ABREN for ABA Responsive Element Novel. To validate that the ABREN was involved in the ABA response pathway, particle bombardment was performed on rice aleurone cells (Figure 1.9). The promoter region of an ABA inducible oleosin gene, which contained two copies of the ABREN, was cloned into a construct upstream of a *GUS* reporter gene. Additional constructs containing mutated versions of the ABREN were also made. The constructs were introduced into rice aleurone cells via particle bombardment, then the cells were exposed to ABA and the level of GUS activity was measured. The constructs containing the mutated ABREN had reduced GUS activity (Figure 4.6), thus experimentally confirming the ABREN plays a role in ABA signaling.

In chapter 5, I introduce the Transcript Structure and Domain Display (TSDD). There is a need for software to quickly generate accurate proportional transcript structures and domains. Current software solutions have many limitations. Some programs must be downloaded and installed on the user's server, other programs require substantial data to generate a structure (CDS sequences, genomic sequences, GFF files and protein sequences) not all of which may be available to the user. Transcript Structure and Domain Display is web-based, so no installation is required, and it only requires a GFF or BED file to generate transcript structures. In addition, the GFF file of several known species have been pre-loaded into a database so the user only needs to enter locus IDs for these species to generate gene structures. In addition, an HMM scan feature was added so for the pre-loaded species, the user has the option to scan their transcripts for known protein motifs. TSDD is web-based, easy to use and fast software solution for generating transcript structures.

Further identification genes within the rice genome

As mentioned in chapter 1, the ultimate aim of this research is to understand the signaling pathways involved in stress response and seed germination in rice. If the genome is not fully annotated, then it may be impossible to build a complete pathway if key genes in the signaling pathway have not been identified and annotated. Therefore, it is important that as many genes in the rice genome be properly annotated. In chapter 2, it was demonstrated that more than 8% of the genes in the rice genome might have been unannotated, this equates to about 4,560 genes. By identifying hundreds of high confidence novel genes and experimentally verifying a subset of them via RT-PCR, we have confirmed that the rice genome annotation is in need of improvement. Not only does the rice genome annotation need improvement, but further analysis of the function of these novel genes will be required to more completely understand their role in the hormone signaling networks of rice as well as understanding the crosstalk between ABA and GA.

By combining TA and Cufflinks, a total of 4,691 highly expressed potential novel genes were identified (Figure 3.3). Filtering out potential novel genes that showed high sequence similarity to another region of the genome eliminated the vast majority of potential novel genes bringing the total to 767 potential novel genes (Figure 3.4). This sequence similarity filter was performed since it is not possible by RNA-seq analysis to determine if a potential novel gene is truly expressed if there is an alternative region of the genome with similar sequence. Since it is well established that redundant genes exist in many species (Nowak et al., 1997), elimination of potential novel genes due to their high sequence similarity to another genomic region may exclude many real novel genes. Further analysis should be performed to identify these potential novel genes with high similarity to other regions of the genome (NGHSORG). At the time the RNA-seq experiments of this dissertation were performed, single-end 50 nt read length was typical. Since then, the Illumina read lengths have increased to 300 nt (Schirmer et al., 2015). Unless the nucleotide differences between the NGHSORG and its similar genomic counterpart, are spaced beyond 300 nt apart, longer read lengths should aid in confirmation of NGHSORGs.

There are platforms that can produce even longer read lengths. For example, Pacific Biosciences (PacBio) claims that their Single Molecule, Real-Time (SMRT®) DNA sequencing technology, averages a read length of 10,000 nt. However, with a low read coverage, the SMRT platform may not identify the NGHSORGs of interest. Also, the very high raw error rate, between 11 and 15 % (Rhoads and Au, 2015), would make it difficult to identify the location on the genome where the NGHSORG transcripts originated.

The Ion torrent PGM platform can produce 400 nt read lengths, however, the reported error rate is between 1.4 and 1.5 %, compared to the 0.9% error rate on Illumina's MiSeq platform (Salipante et al., 2014) and 0.26% error rate on Illumina's HiSeq2000 platform (Quail et al., 2012). Though

Ion Torrent has a longer continuous read length than Illumina, the higher error rate is still the limiting factor. When PGM sequencing was performed on the AT-rich *Plasmodium falciparum*, approximately 30% of the genome had no coverage. PGM also proved very poor in correctly identifying the correct number nucleotides in regions where the same nucleotide is repeated many times (Loman et al., 2012).

Overall, high-throughput sequencing technologies are advancing at an exponential rate with increasing read lengths. Though PacBio and Ion Torrent can produce longer read lengths than Illumina, the increased error rate combined with limitation on same nucleotide repeats, Illumina still appears the be the best choice at this time. There are 3,891 NGHSORG identified by TA that have greater than 25% similarity to another genomic region which were excluded in our most recent study (Figure 3.4). Of these, 1,865 NGHSORG show greater than 99% similarity to another genomic region. These genes may be difficult to identify even with increased read lengths due to their nearly identical sequence similarity to other genomic regions. This leaves 2,051 NGHSORGs that may be identifiable with a longer read length. Thus, RNA-seq should be performed again on the same hormone treated samples but with 300 nt read length to identify these 2,051 NGHSORGs.

Of the 1,865 NGHSORGs that have greater than 99% similarity to anther genomic region, 895 show 100% similarity to another genomic region. For these genes, an increased read length would not aid in determining whether or not they are transcribed. Analysis of these NGHSORGs must be performed on an individual basis. This can be done by knocking out the annotated gene that is homologous to the NGHSORG via clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated protein 9 (Cas9) targeted genome editing (for a review see (Deveau et al., 2010; Sander and Joung, 2014). Cas9 is a DNA endonuclease that cleaves double-stranded DNA. The enzyme is guided to the target DNA site by a guide RNA molecule (gRNA) that contains a sequence that matches the sequence to be cleaved. This gRNA can be designed to target the DNA at the location where the gene to be knocked out resides. The RNA-guided Cas9 then creates a site-specific double-stranded DNA break, which when repaired by the cell by non-homologous end-joining (NHEJ), often leads to a mutation at the cut site (Figure 6.3).

A mutant rice plant can be created by mutating a gene with high similarity to a potential novel gene. Good candidates would be genes that are highly expressed so that detection of gene expression by methods, such as RT-PCR, are possible. Since this is a low throughput method, NGHSORGs that are inducible by hormones such as ABA or GA are good candidates since these genes may play a role in the hormone signaling network. For example, potential novel gene TAOs01g00880 is a good candidate because it is highly expressed and ABA inducible (Figure 6.4A). Only 59 reads aligned on the control sample, but 4,093 reads aligned on the ABA treated sample. TAOs01g00880 also shows 98.8% sequence similarity to the annotated uncharacterized gene LOC_Os05g06399 (Figure 6.4B). The read alignment pattern of LOC_Os05g06399 was almost identical to TAOs01g00880 with 57 reads aligned on the control sample and 4,319 reads aligned on the ABA treated sample. All the differences between these two genes are single nucleotide differences and they are sporadically spread out across the gene. The coded protein of these two genes differs by two amino acids. Since the sequence similarity between LOC_Os05g06399 and TAOs01g00880 is virtually identical, it is not possible to design a gRNA to cut LOC_Os05g06399 and not cut TAOs01g00880. Not only are the two genes virtually identical, but the regions 1,431 nt upstream and 251 nt downstream of LOC_Os05g06399 are 99.6% and 100% identical to the corresponding regions of TAOs01g00880 respectively. The similarity of the upstream promoter regions of these genes is an added indication that both genes are actively being transcribed.

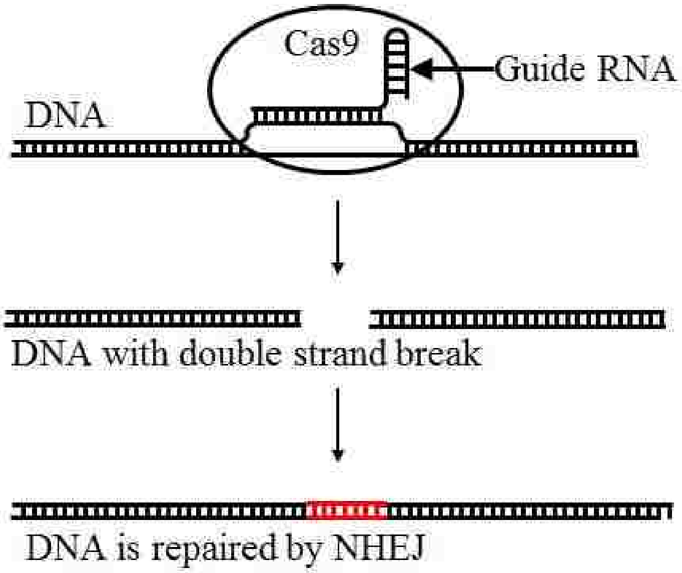


Figure 6.3: Cas9 is a DNA endonuclease that cleaves double-stranded DNA.

The Cas9 enzyme is guided to the target DNA by a gRNA molecule that contains a sequence that matches the sequence to be cleaved. Cas9 activity creates site-specific double-stranded DNA break, which when repaired by the cell by NHEJ, often leads to a mutation at the target site.

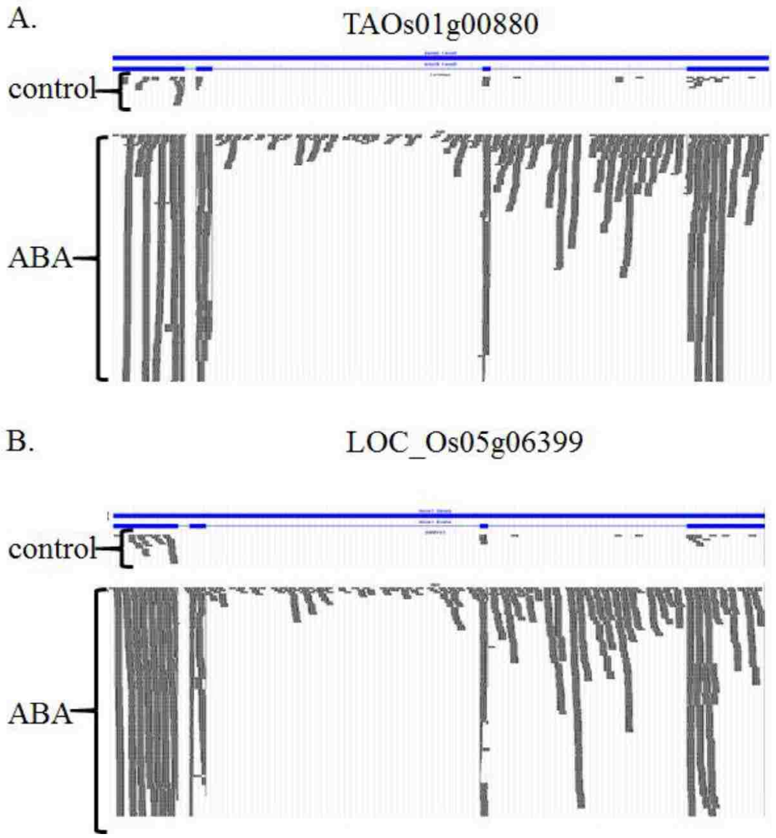


Figure 6.4: Potential novel gene TAOs01g00880 shows high similarity to LOC_Os05g06399.

Potential novel genes TAOs01g00880 shows 98.8 % similarity to annotated uncharacterized gene LOC_Os05g06399 and is also highly ABA inducible. TAOx01g00880 has 59 control reads and 4093 ABA reads, LOC_Os05g06399 has 57 control reads and 4319 ABA reads.

To eliminate LOC_Os05g06399 via the CRISPR/Cas9 system, two gRNAs must be designed that flank the LOC_Os05g06399 outside the regions similar to TAOs01g00880. By performing double-stranded DNA cuts around LOC_Os05g06399, we can effectively cut out the genes. To do so, a construct harboring the Cas9 gene, the two gRNA genes that flank LOC_Os05g06399, and a hygromycin resistance gene for selection, can be introduced into rice callus tissue, undifferentiated rice cells, via particle bombardment. The calli can then plated on media containing hygromycin. The DNA of the surviving calli can be extracted and sequenced to confirm that LOC_Os05g06399 was spliced out. If LOC_Os05g06399 is an essential gene, then knocking out this gene would result in a non-viable plant unless the potential novel gene TAOs01g00880 was an actual gene with redundant functions to LOC_Os05g06399. TAOs01g00880 may also be a polygenic gene of LOC_Os05g06399 resulting in a viable plant with a polygenic or variable phenotype (Richa et al., 2016). A double knockout of both genes may result in a stronger phenotype and give clues to the function of the genes. If the mutant plant survives, RT-PCR can be performed on TAOs01g00880 to determine if it is expressed. If the mRNA of TAOs01g00880 is detectable, then it has been experimentally confirmed as a real novel gene.

Gene coexpression network to help elucidate functions of novel genes

It is a major challenge to elucidate the functions of large numbers of genes within a genome and to discover how these genes interact to perform specific biological processes. A gene co-expression network is one tool that can be used to help resolve this issue. A gene co-expression network is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them. Genes are co-expressed if their expression pattern across different experimental treatments are similar. For example, if gene A and gene B are both upregulated when treated with ABA, downregulated when treated with GA and neutral when treated with both hormones, then they are considered co-expressed. Gene co-expression networks are of biological interest in determination of gene function since clusters of co-expressed genes are often controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or protein complex (Stuart et al., 2003). Thus, gene co-expression networks can aid in the determination of the gene function of the novel genes as well as identified genes with unknown function.

Several methods have been developed for constructing gene co-expression networks. However, they are all based on a similar methodology. The co-expression measure between all pairs of genes is calculated. Then the pairs of genes whose coexpression value exceeds a threshold are considered to have significant co-expression relationship and are connected by an edge in the network. A common method for determining the threshold is by using a Fisher's Z-transformation which calculates a z-score for each correlation based on the number of samples. This z-score is then converted into a p-value for each correlation and a cutoff is set based on the p-value (e.g. p <= 0.05) (Makashir et al., 2015).

By utilizing the differential expression results from our RNA-seq data, a co-expression network can be determined based on the differential expression of ABA, GA and ABA+GA treatments. Determination of the gene clusters within a network is an NP-hard problem, meaning that an exhaustive search is necessary to identify the gene clusters. Since the number of operations necessary to solve this problem grows exponentially with the number of genes and thousands of genes are involved, an exhaustive search is not feasible. The walktrap algorithm (Pons and Latapy, 2005) is a popular algorithm to identify clusters or communities of genes that may share common function. Walktrap is an approach based on random walks. The general idea is that if you perform random walks on the graph, then the walks are more likely to stay within the same community because there are only a few edges that lead outside a given community. Walktrap runs short random walks of 3 to5 steps (depending on user's parameters) and uses the results of these random walks to define communities in a bottom-up manner. There are several other algorithms such as fastgreedy (Kumar et al., 2015) and edgebetweenness (Girvan and Newman, 2002), but walktrap is a fast, and popular algorithm and is thus the algorithm of choice. Using the walktrap algorithm, subnetworks in the rice transcriptome can been identified. Further research can be performed to identify the functions of these subnetworks. By adding the novel genes identified in this dissertation to the network, gene function of these novel genes can be better understood.

Sugar response and ABA

When comparing the RNA-seq results of our most recent NAAT experiment with our previous AAT results, we determined that the number of genes induced by ABA was higher in the presence of α-amylase (Figure 4.1). Not only were the number of ABA induced genes higher, but the level of ABA induction was also higher in the AAT sample compared to the NAAT sample (Figure 4.1). Over 90% of the genes that were ABA inducible in both AAT and NAAT had a higher level of ABA induction in the AAT sample.

The α-amylase treatment was used to break down the starchy endosperm to facilitate isolation of the aleurone cells in the AAT sample. Starch consists of two components, amylose and amylopectin, both of these components can be hydrolyzed into monosaccharides maltose and glucose by α-amylase (Figure 6.5). We believe that the sugar was generated from the α-amylase treatment and that sugar signaling may be the cause of the higher level of ABA response. There is some evidence to support this hypothesis.

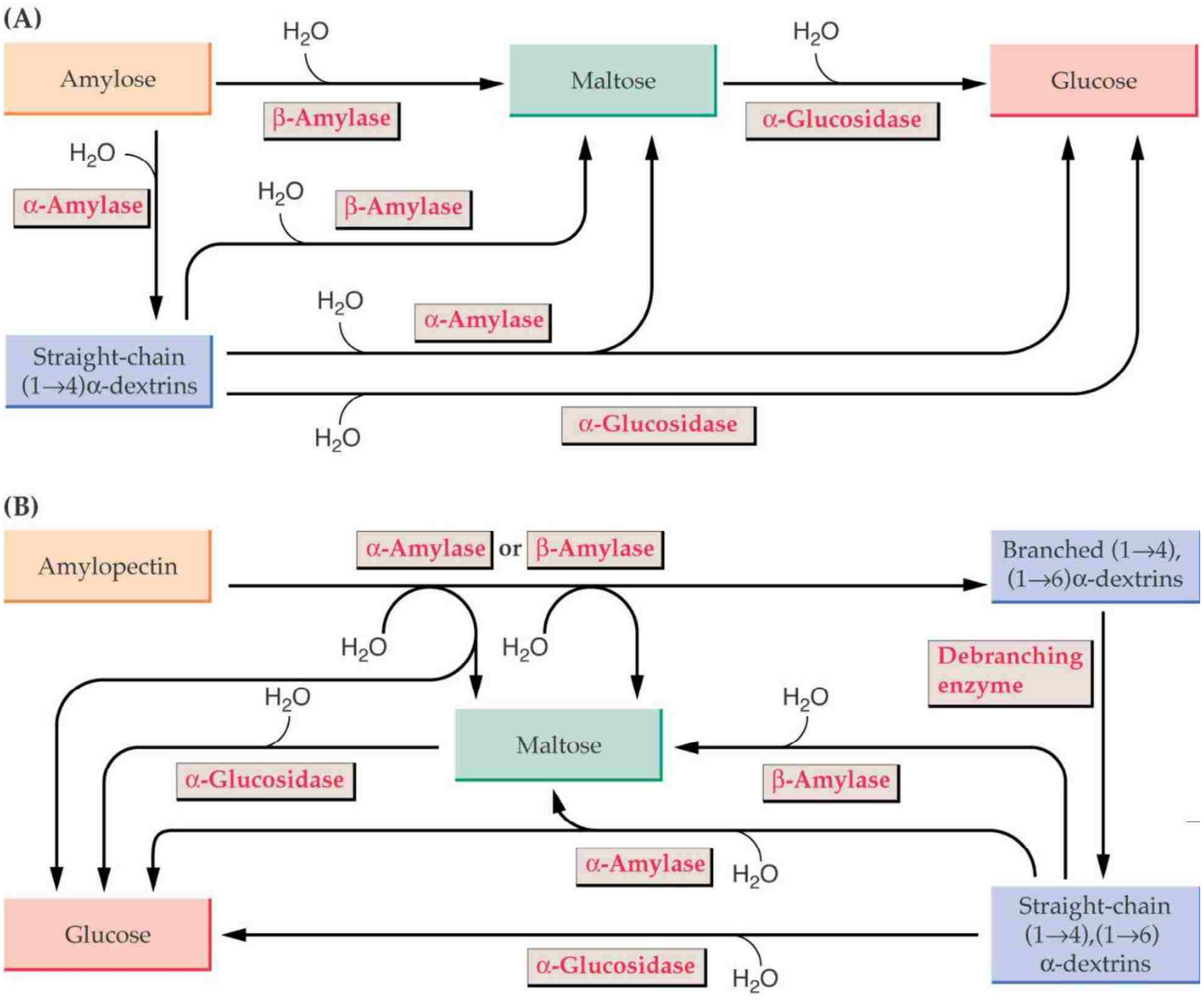


Figure 6.5: Hydrolysis of starch to sugar.

Starch consists of two types of polysaccharides, A.) amylose which makes up 20-30% of starch and B.) amylopectin which makes up 70-80% of starch. Both of these starch components can be broken down to monosaccharides (maltose or glucose) by hydrolases such as α-amylase or β-amylase. Figure taken from (Zeeman, 2002).

The α-amylase gene αAmy3, which has previously been shown to be down regulated in the presence of sugar (Chen et al., 2006), also shows significant downregulation in the AAT sample (Supplemental Table S9).

Also, sugar transporter genes (OsMST3, OsMT4 and OsMT6) were significantly downregulated in the AAT sample (Supplemental Table S9). These sugar transporters transport sugar across the cell membrane. With the abundance of sugars surrounding the aleurone cells, the necessity for sugar transporters is reduced and this could explain the downregulation of the sugar transporters.

Further experiments should be performed to validate the presence of sugars in the samples prior to RNA-seq. A qualitative test for the presence of sugar can be performed via Benedict's reagent. This can be done by grinding deembryonated rice seeds, exposing one sample to water and another sample to an α-amylase solution and incubating the samples overnight. Then by applying Benedict's reagent, the presence of sugar can be determined by a change in the color of the solution.

Once it is established that sugar is present in the samples, further RNA-seq should be performed to confirm the original RNA-seq data. Four experimental treatments should be performed: a control sample with no application of sugar or hormones, a glucose treated sample, and ABA treated sample and a sample with both glucose and ABA. Other experimental treatments with other sugars such as maltose can also be peformed.

With direct application of sugars to the aleurone cells, this should eliminate the possibility that the α-amylase protein may be causing the change in ABA induction. By comparing the read counts of the different treatments to the control sample, the genes inducible by glucose and ABA can be identified and the effect of these genes upon both treatments will confirm whether or not the combination of the ABA and glucose is an additive effect or whether glucose is simply making the ABA inducible genes more sensitive to ABA treatment.

Stable transgenic plant containing the ABREN

Particle bombardment was used to confirm that the ABREN plays a role in the ABA signaling pathway. This method of transformation was performed because it returns quick and reliable results. However, since the reporter gene is in a plasmid and not part of the genome, chromatin effects are not observed. There also may be epigenetic factors such as DNA methylation or histone methylation that may also effect the results. To solve these issues, a transgenic plant must be created that contains the ABREN promoter driving a reporter gene integrated within the genome. To make a stable plant, *Agrobacterium*-mediated rice transformation can be performed on immature rice embryos. *Agrobacterium tumefaciens* is gram negative bacteria that has the ability to transfer DNA into plants (Gelvin, 2000). The constructs containing the ABREN promoter driving a reporter gene can be inserted into *A. tumefaciens* via heat shock. Constructs containing a mutated ABREN can also be inserted into *A. tumefaciens* as a negative control Then the transformed *A. tumefaciens* can be used to infect rice embryos and integrate the ABREN contruct into the rice genome. The infected rice embryos can then be grown into full transgenic rice plants. For more detailed protocol see (Slamet-Loedin et al., 2014).

It may also be possible to transform the reporter gene into the rice genome via CRISPR/Cas9 genome editing. In a recent publication, herbicide-resistant rice plants were engineered through CRISPR/Cas9-mediated homologous recombination (Sun et al., 2016). This was performed by using CRISPR/Cas9 to cleave a region out of the rice genome and utilizing homology directed repair (HDR) to insert a modified acetolactate synthase gene in its place (Figure 6.6). The modified acetolactate synthase gene confers herbicide resistance to chlorsulfuron and bispyribac sodium. In

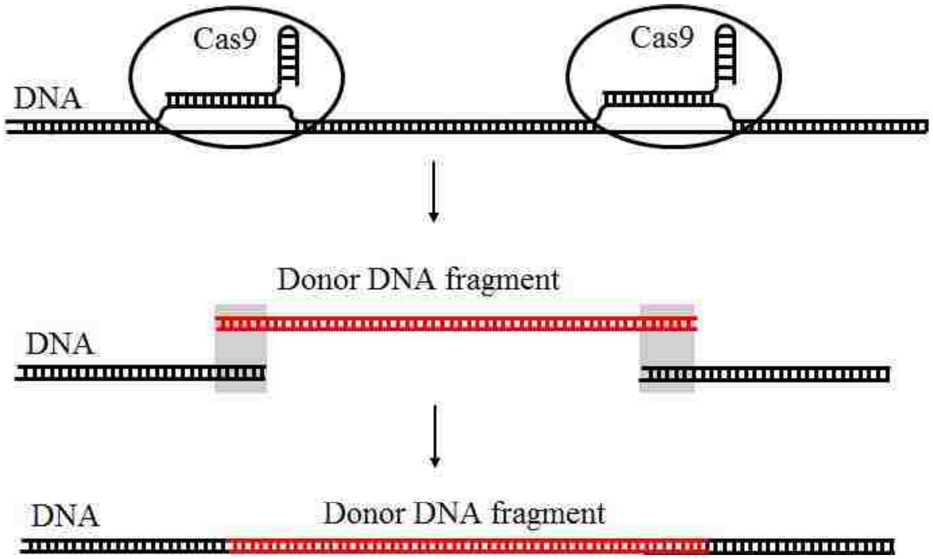


Figure 6.6: CRISPR/Cas9-Mediated Homologous Recombination

A construct containing the Cas9 gene and two gRNAs were introduced into rice callus tissue. The Cas9/gRNAs created double stranded DNA breaks in the rice genome at the two locations corresponding to the gRNAs. Then a donor DNA fragment containing regions homologous to the region of the double stranded break (gray regions) is inserted into the gap via homology directed repair.

brief, a construct harboring the Cas9 gene, the two gRNA genes that flank the acetlactate synthase gene, the modified acetolactate synthase gene and a hygromycin resistance gene for selection, is introduced into rice callus tissue via particle bombardment. The calli are then plated on media containing hygromycin. The DNA of the surviving calli are extracted and sequenced to confirm that the gene substitution took place. CRISPR/Cas9 can be used to splice out the oleosin gene (LOC_Os09g15520) and introduce a *GUS* reporter gene in its place leaving the oleosin ABREN-containing promoter to drive the *GUS* reporter gene.

Once a stable transgenic plant is produced, either by agrobacterium or CRISPR/Cas9, the plant can be grown to maturity and the transgenic seeds can be obtained. The reporter gene expression in the aleurone cells in the presence of ABA can be measured to confirm that the ABREN is indeed ABA responsive in a stable plant.

Identification of the protein binding partner of the ABREN

The ABREN was experimentally confirmed to plays a role in ABA induction by particle bombardment of an ABREN containing construct into rice aleurone cells. To further explain the role of the ABREN in the ABA signaling pathway, the ABREN binding protein must be identified. Yeast one-hybrid screening is an assay that identifies proteins that can bind to specific sequences of DNA (Singh et al., 1988). Yeast one-hybrid can be performed to identify the protein that binds to the ABREN. In brief, a construct can be created with the ABREN sequence inserted upstream of a reporter gene, such as an aureobasidin A resistance gene. The construct can then be inserted into the genome of a special strain of yeast. Then the yeast can be transformed with a library of fusion constructs that produce proteins fused to an activation domain. If one of the gene fusion products can bind to the ABREN, then the activation domain will transcribe the aureobasidin A resistance reporter gene. If there is a yeast colony that can survive on plates containing the aureobasidin A antibiotic, then the construct of the fusion gene can be extracted from the yeast and sequenced to identify the protein that binds to the ABREN.

Once the ABREN binding protein (ABBP) has been identified, confirmation by another assay should be performed since protein interactions that occur in yeast may not necessarily occur in plants. ChIP-seq analysis can be performed to validate the ABBP is indeed binding to the ABREN containing promoters. This is performed by cross-linking the proteins to their DNA binding sites, and sonicating to break the DNA fragments to 300nt fragments. Then ABBP specific antibodies can be used to pull down the ABBP along with its crosslinked DNA. Then the ABBP can be removed from the DNA and the DNA fragments can be sequenced. If there is an abundance of ABREN containing DNA fragments in resulting sequencing, this will not only confirm that the ABBP binds to the ABREN, but it will also identify the genes that are regulated by ABBP.

Assays such as yeast two-hybrid (Young, 1998) can be performed to identify the proteins that ABBP interacts with. Bimolecular fluorescence complementation (BiFC) (Hu et al., 2002) can be performed to not only confirm the protein interaction with ABBP but identify the location within the cell where the interaction takes place. This may help identify the location of the ABBP within the ABA signaling pathway. For example, if the ABBP interacts with the SnRK2 kinase, then this suggests that SnRK2 may activate the ABBP by phosphorylation, thus enabling the ABBP to bind to the ABREN and induce gene transcription.

To determine the role that the ABBP plays in the plant stress response, a knock out of the ABBP can be performed via the CRISPR/cas9 gene editing system. The phenotype of the mutant can be observed. If this protein is indeed a key player in the ABA signaling pathway, then the mutant plant should be ABA insensitive or show an aberrant response to ABA. RNA-seq analysis can be performed to obtain the transcriptome of the mutant ABBP when treated with ABA and the transcriptome can be compared to the WT transcriptome to identify the genes affected by the mutation.

Crosstalk between ABA and GA

The antagonism between the stress and seed dormancy hormone ABA and the growth and germination hormone GA has been well documented. Understanding how these hormones interact with each other is of great importance in understanding the mechanisms of germination, seed dormancy and stress response. To get a better understanding, we can observe the genes that were upregulated by these hormones or a combination of these hormones from our RNA-seq data of rice aleurone cells. Overall, there were substantially more genes affected by GA than ABA. There were 2,443 genes that were induced by ABA and 4,113 genes that were induced by GA. There were 2,822 genes repressed by ABA and 4,577 genes repressed by GA (Figure 6.7). This indicates that there are more genes involved in seed germination than in seed dormancy which intuitively makes sense since more biological activity takes place during seed germination than seed dormancy.

There are 2,443 genes that were induced by ABA, but when both hormones ABA and GA are applied, about half of the ABA inducible genes (1,274 genes, 52.1%) were not induced and an additional 3,104 genes were induced (Figure 6.7A). A similar effect was observed for ABA repressed genes. There were 2,822 genes that were repressed by ABA, but when both hormones were applied, about half of the genes (1,403 genes. 49.7%) were no longer repressed and an

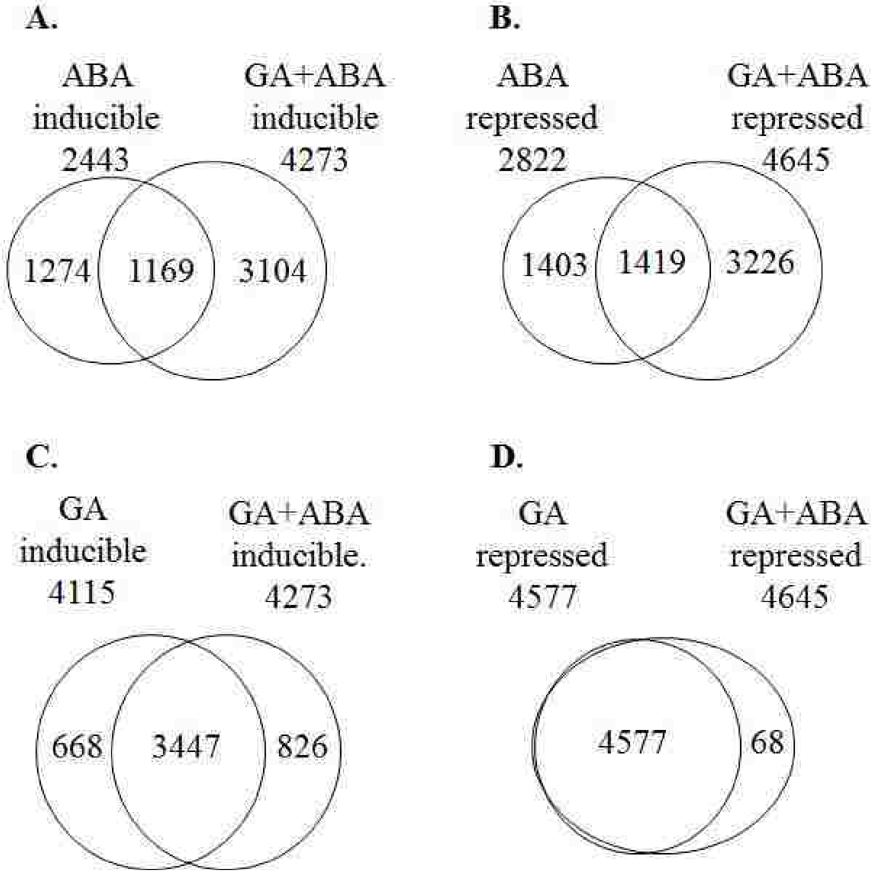


Figure 6.7: Crosstalk between ABA and GA.

A) Of the 2,443 genes induced by ABA, 1,274 genes (52.1%) are not induced when both hormones ABA and GA were applied. B) Of the 2,822 genes that are repressed by ABA, 1,403 (49.7%) are not repressed when both hormones were applied. C) Of the 4,115 genes induced by GA, only 668 genes (1.7%) were not induced when both ABA and GA were applied. D) All of the 4577 genes repressed by GA were also repressed when both ABA and GA were applied.

additional 3,226 genes were repressed (Figure 6.7B). This demonstrates that the antagonism between these two hormones is not 100% as previously thought. Further analysis should be performed to understand the function of the genes that were ABA inducible/repressed in the presence of ABA, but lost their induction/repression in the presence of both hormones. The genes that showed the largest change in induction/repression would be good starting candidates.

There are 4,115 genes that were induced by GA. Of these genes, there were only 668 genes (16 %) that were not induced in the presence of both ABA and GA (Figure 6.7C). This demonstrates that ABA does not antagonize GA nearly as much as previously believed. When looking at the GA repressed genes, all of the 4,577 genes (100%) that were repressed by GA were also repressed in the presence of both ABA and GA (Figure 6.7D). Only 68 genes (1.5 %) were additionally repressed by both hormones. Not one gene that was repressed by GA, became un-repressed in the presence of ABA. This demonstrates that ABA has little effect on the genes that are repressed by GA. In both the GA induced and GA repressed genes, ABA was far less antagonistic to GA than one would expect.

RNA-seq of non-poly adenylated mRNA

During the standard protocol for RNA-preparation for RNA-sequencing, the mRNA molecules that are poly-adenylated at the 3' end (poly(A)+) are isolated form the total RNA via annealing the mRNA to beads coated with poly-T oligos (oligo(dT)). This eliminates non-poly-A mRNA molecules (poly(A)-) such as ribosomal RNAs, histone mRNAs, some long non-coding RNAs (lincRNA) derived from introns (Yang et al., 2011a) and possibly many other transcripts that may function as coding or non-coding RNAs. There are also many bimorphic transcripts that can exist as either poly(A)+ or pol(A)-. Though most transcripts are poly(A)+ in humans (Table 3), the abundance of poly(A)- transcripts in plants has not been studied. These poly(A)- transcripts may play important roles in hormone signaling as well as other cellular processes. Thus these transcripts should be identified and characterized.

To identify the poly(A)- transcripts on a transcriptome-wide scale, RNA-seq can be utilized. In brief, the mRNA is extracted from a biological sample. Then the poly(A)+ transcripts are pulled out of the sample via oligo(dT) coated beads. Then the poly(A)- sample is rRNA depleted via one of many available kits (RiboMinus by Thermo Fisher, NEBNext by New England rRNA depletion by BioLabs, GeneRead rRNA depletion by Qiagen, and others). Then the poly(A)- RNA-seq library can be prepared and sequenced similar to standard RNA-seq library preparation protocols. For a more detailed protocol, see (Yang et al., 2011a).

Table 3

	H9	**HeLa**
Poly(A)- (transcripts)	278	324
Bimorphic (transcripts)	2550	1587
Poly(A)+ (transcripts)	8133	10183
Total (transcripts)	10961	12094
% poly(A)-	2.54	2.68
% bimorphic	23.26	13.12
% poly(A)+	74.20	84.20

Most transcripts are poly(A)+ in H9 human embryonic stem cells and HeLa cells.

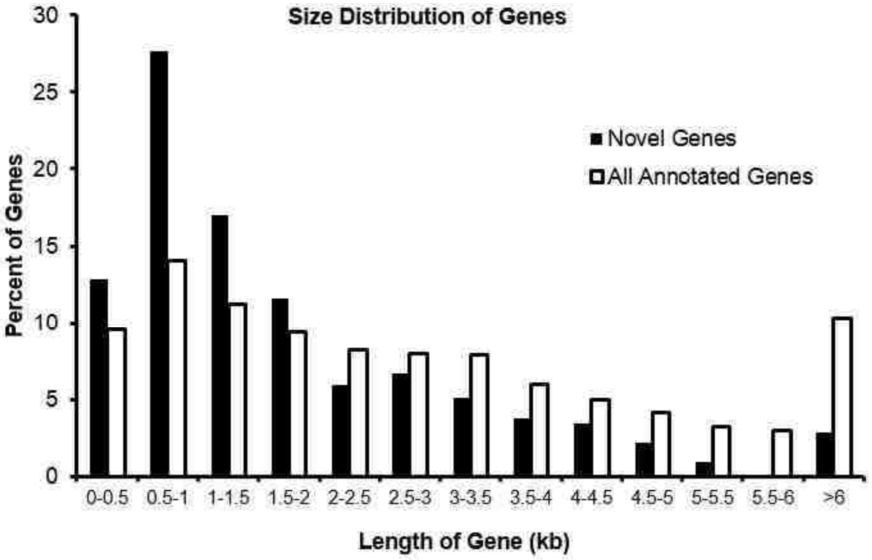
Concluding remarks

In this report, it was demonstrated that the rice genome annotated is far from complete with as much as 8% of the genes are unannotated. Hundreds of novel rice genes were identified which will improve the rice genome annotation. In addition, a new tool was developed to identify genes in the rice genome, and other genomes, and will greatly aid in annotation of genomes. An accurate rice genome annotation will be an invaluable tool for researchers for the study of signaling pathways. Also, a novel *cis*-acting element involved in the rice ABA response was identified via bioinformatics analysis of the promoter regions of ABA inducible genes. This element was experimentally confirmed to play a role in the ABA response via particle bombardment. Understanding the role of this element in the ABA response will greatly aid the scientific community in understanding the stress response pathway in rice. Understanding the stress response pathway may help lead to the development of more robust strains of rice and possibly other food crops and thus secure the world's food supply.

APPENDIX

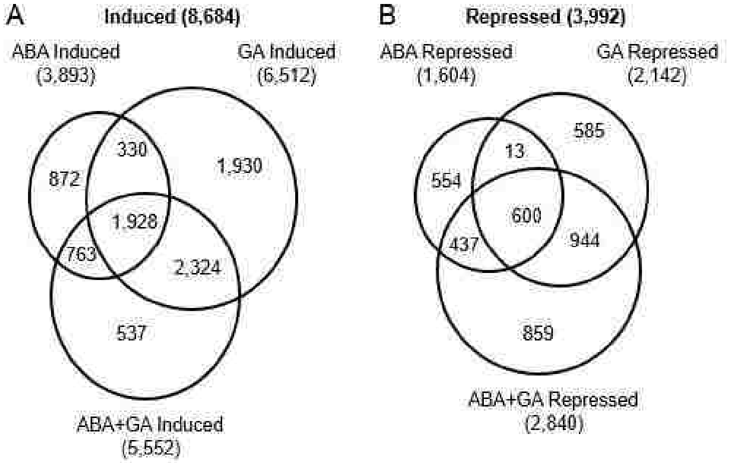
Supplemental figures

Supplemental Figure S1



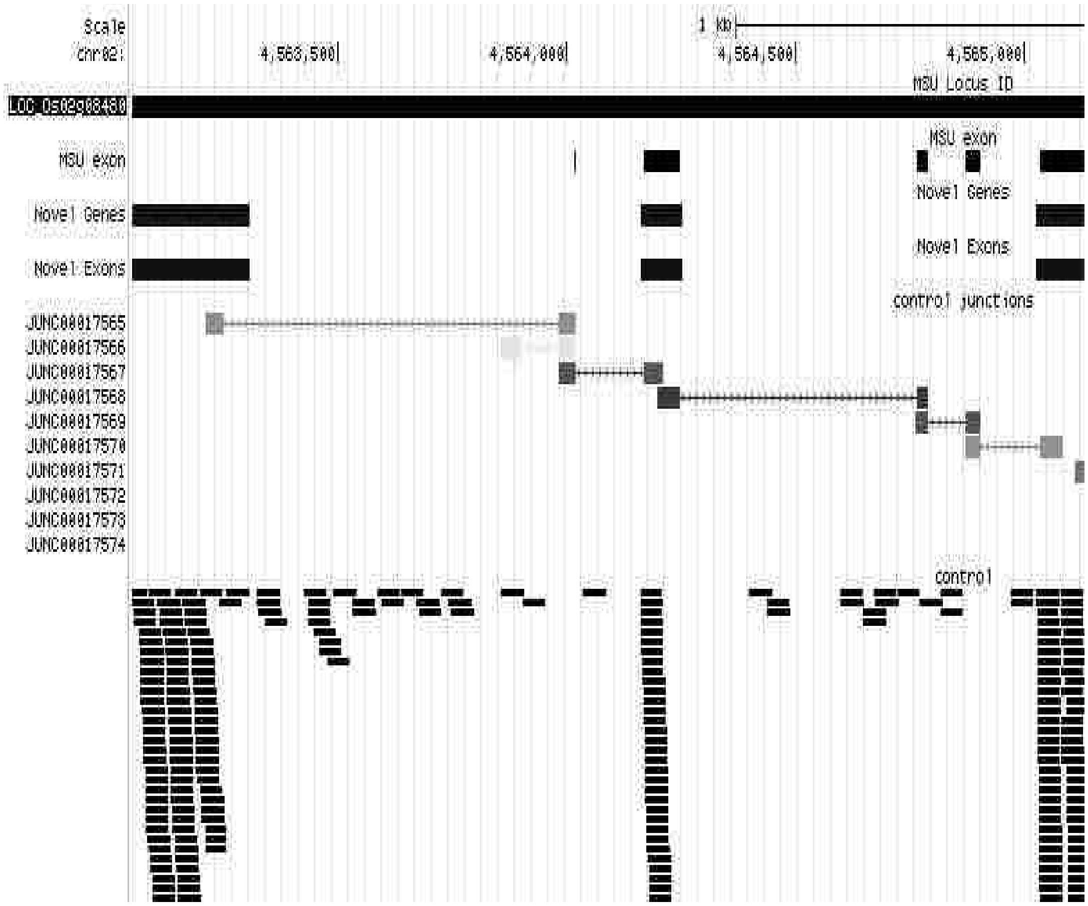
Supplemental Figure S1: Size distribution comparison of the novel genes to annotated genes. Both the novel genes and annotated genes have a maximum at a length of 0.5 to 1kb, and decline steadily as the lengths of the genes increase. The size of all genes includes introns and exons.

Supplemental Figure S2



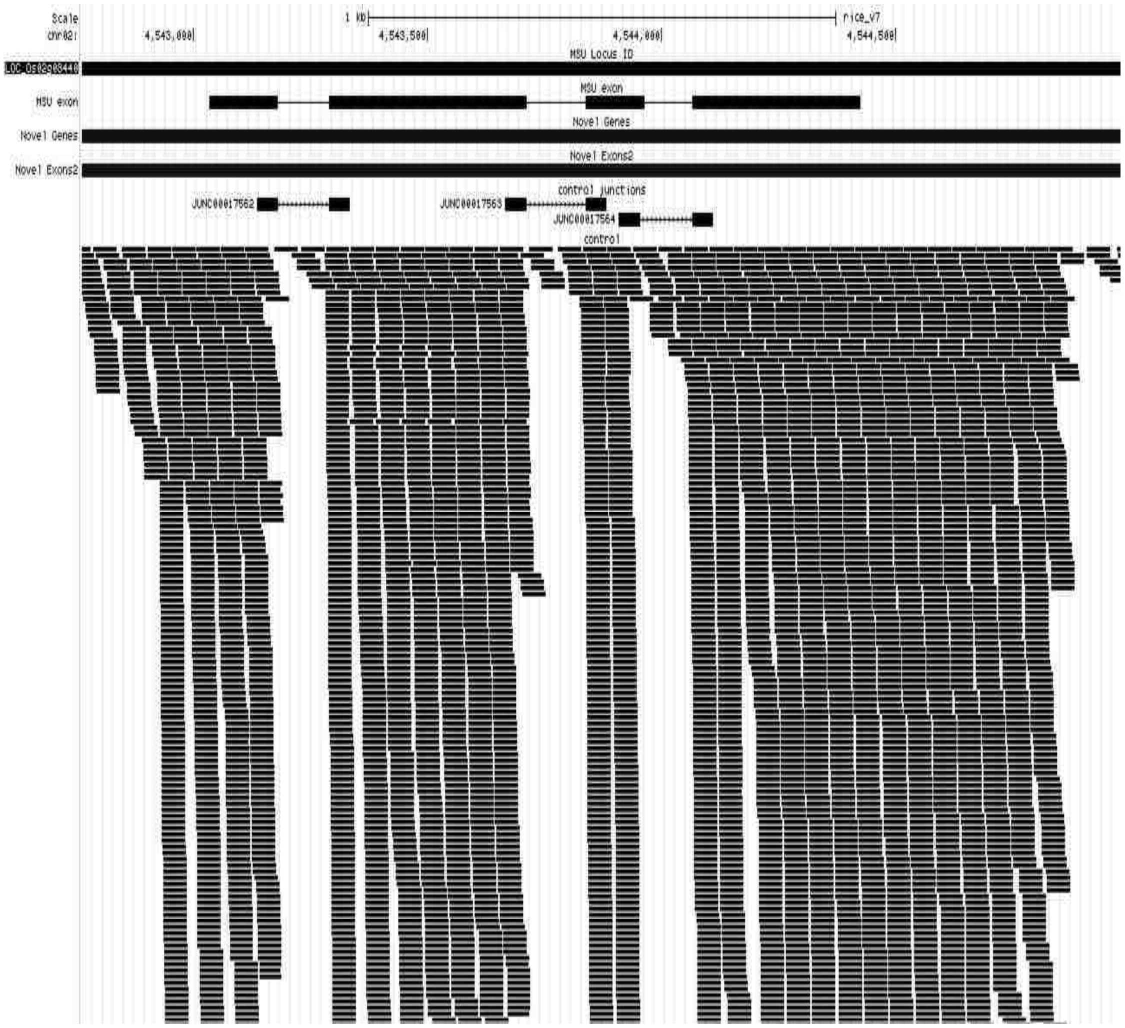
Supplemental Figure. S2: Hormone response of annotated genes. (**A**) RNA-Seq data of annotated genes showed that 8,684 genes were induced under at least one treatment. (**B**) In comparison, 3,992 were repressed under at least one treatment. As with the novel genes, approximately half of the genes induced or repressed by at least one treatment were induced or repressed in multiple treatments.

Supplemental Figure S3



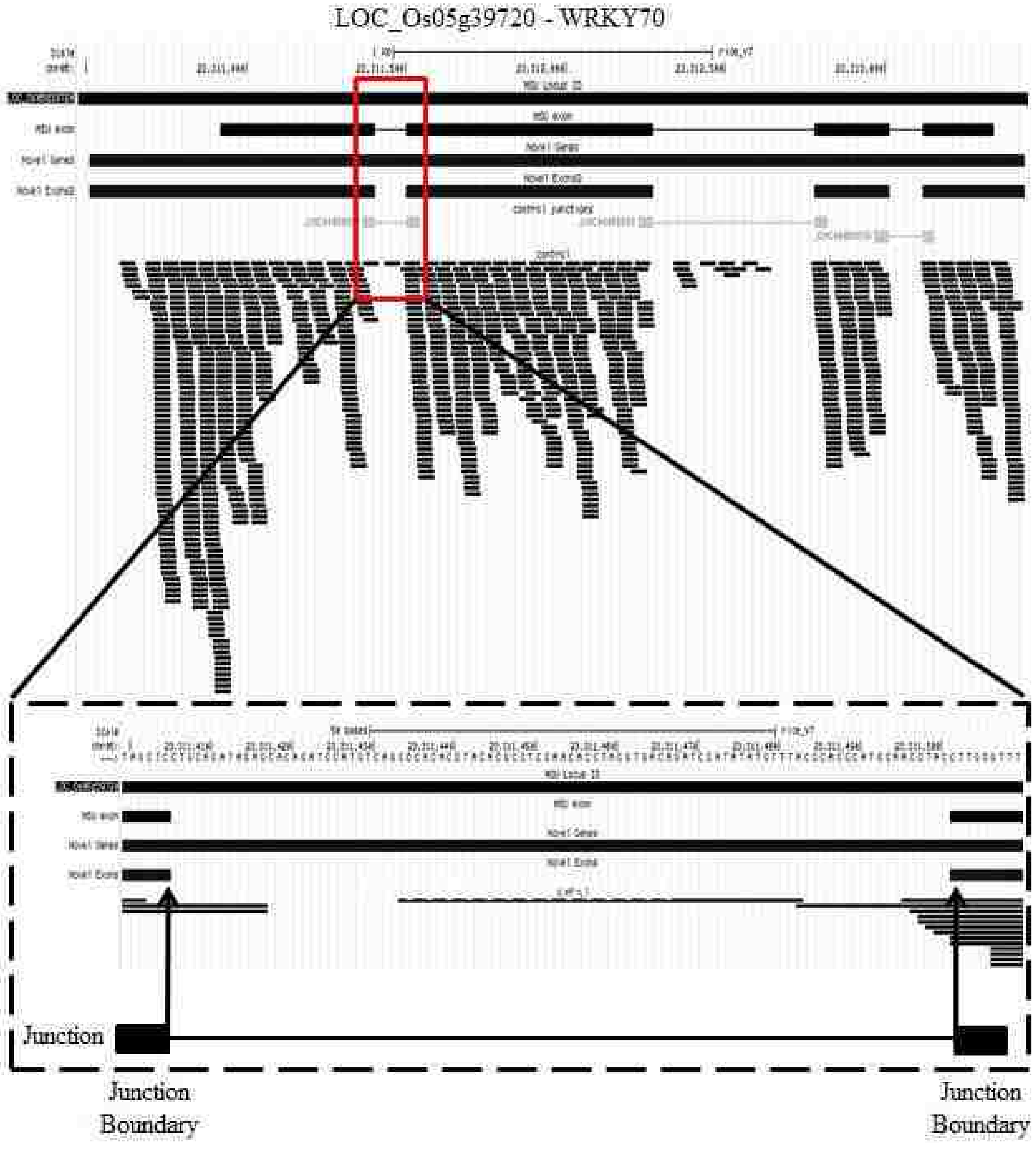
Supplemental Figure S3: Exons shorter than a read length have few or no reads aligned. The gene at LOC_Os02g08040 contains exons shorter than 50nt in length. Because these exons are shorter than a single read, full-length reads from spliced transcripts will not align to the genome at the location of the exons. By taking advantage of junction alignments by Tophat, though, the exons can be identified. The exons inside the red boxes are less than 50 nt in length and cannot be detected by Tiling Assembly based solely on the read alignment. The shade of a junction in the figure indicates the number of junctions at that position, with black bars indicating many junctions and light grey bars indicating fewer junctions.

Supplemental Figure S4



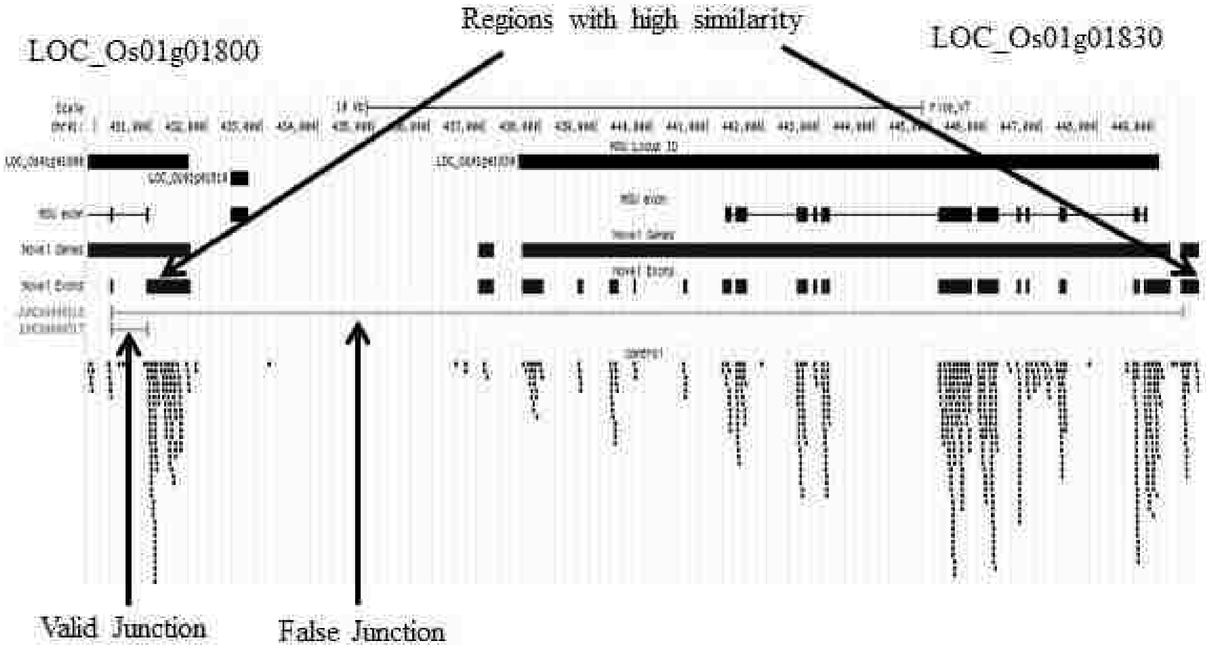
Supplemental Figure S4: Initial steps of Tiling Assembly show genes with intron retention or noise as single exon genes. Small numbers of reads aligning across a junction lead to identification of multiple exons as a single exon. The gene at LOC_Os02g08440 was initially identified as a single exon gene due to noise reads aligning to the introns (red boxes). If there is a junction with low read coverage, Tiling Assembly identifies this region as an intron.

Supplemental Figure S5



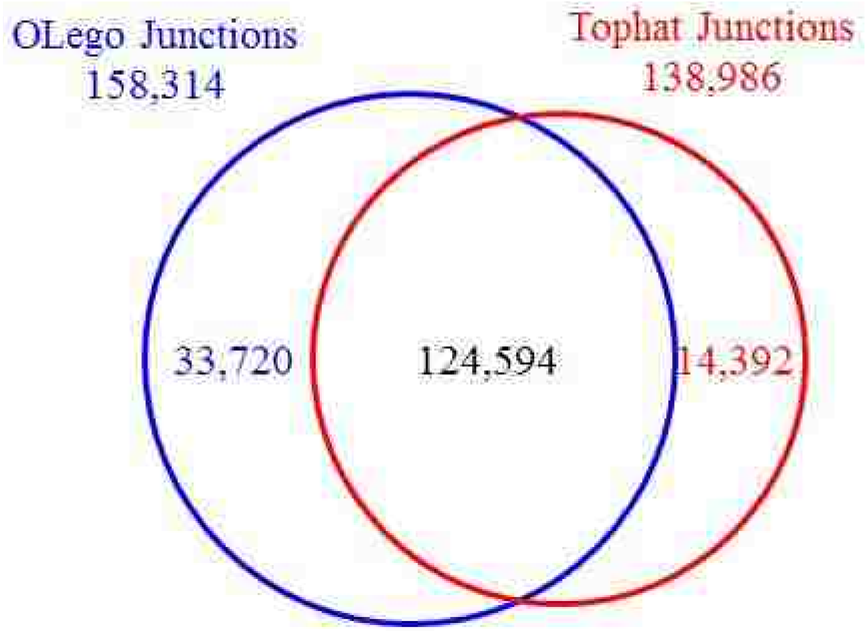
Supplemental Figure S5: Junction boundaries were used to identify exon boundaries and eliminate noise reads. Occasionally, noise reads align across a junction or reads overlap the junction. The boundaries specified by Tophat junction alignments were used to fine-tune exon boundaries to within one nucleotide. The portion of the upper figure surrounded by the red box is magnified in the lower figure to better show the exon boundaries.

Supplemental Figure S6



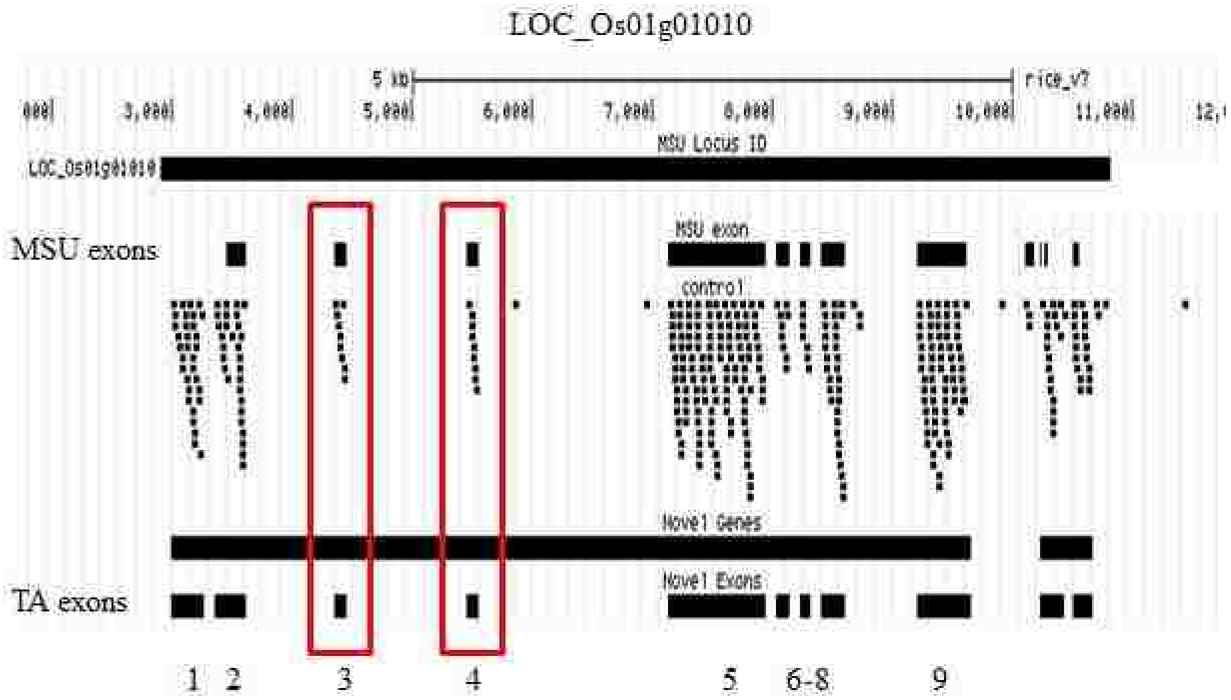
Supplemental Figure S6: Similar sequences can lead to invalid junction mapping. When two regions are highly similar to each other, junction alignments may erroneously lead to the alignment of a junction between two genes, as is seen with LOC_Os01g01800 and LOC_Os01g01830. In order to prevent two genes from being erroneously merged based on these junction alignments, Tiling Assembly allows the user to specify a maximum length for a junction that skips exons.

Supplemental Figure S7



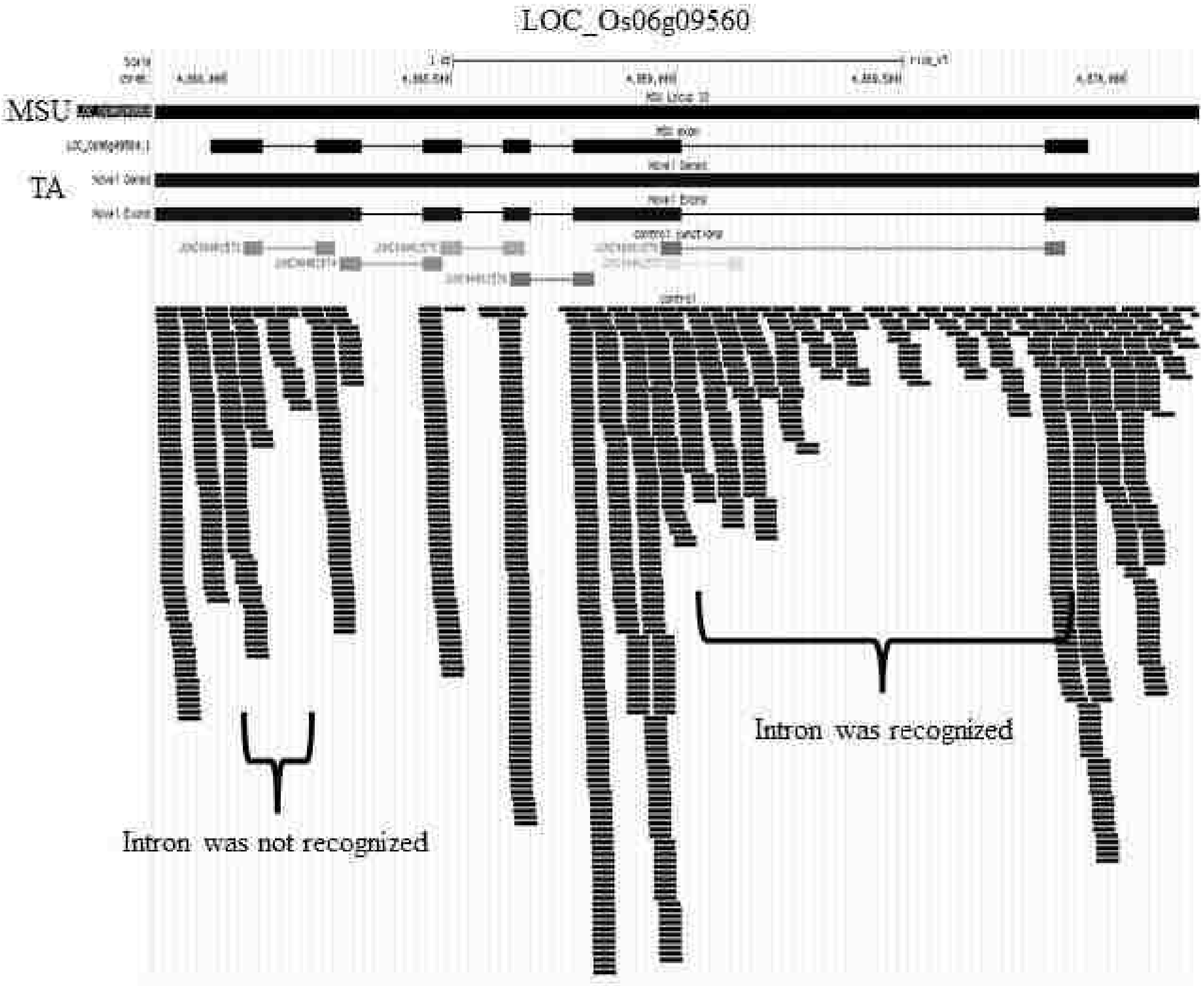
Supplemental Figure S7: OLego identified more junctions than Tophat. Of the 158,314 junctions identified by OLego, 124,594 junctions (78.7%) matched identically to a junction identified by Tophat. Of the remaining 33,720 junctions identified by OLego, 71.3% were determined from a single read.

Supplemental Figure S8



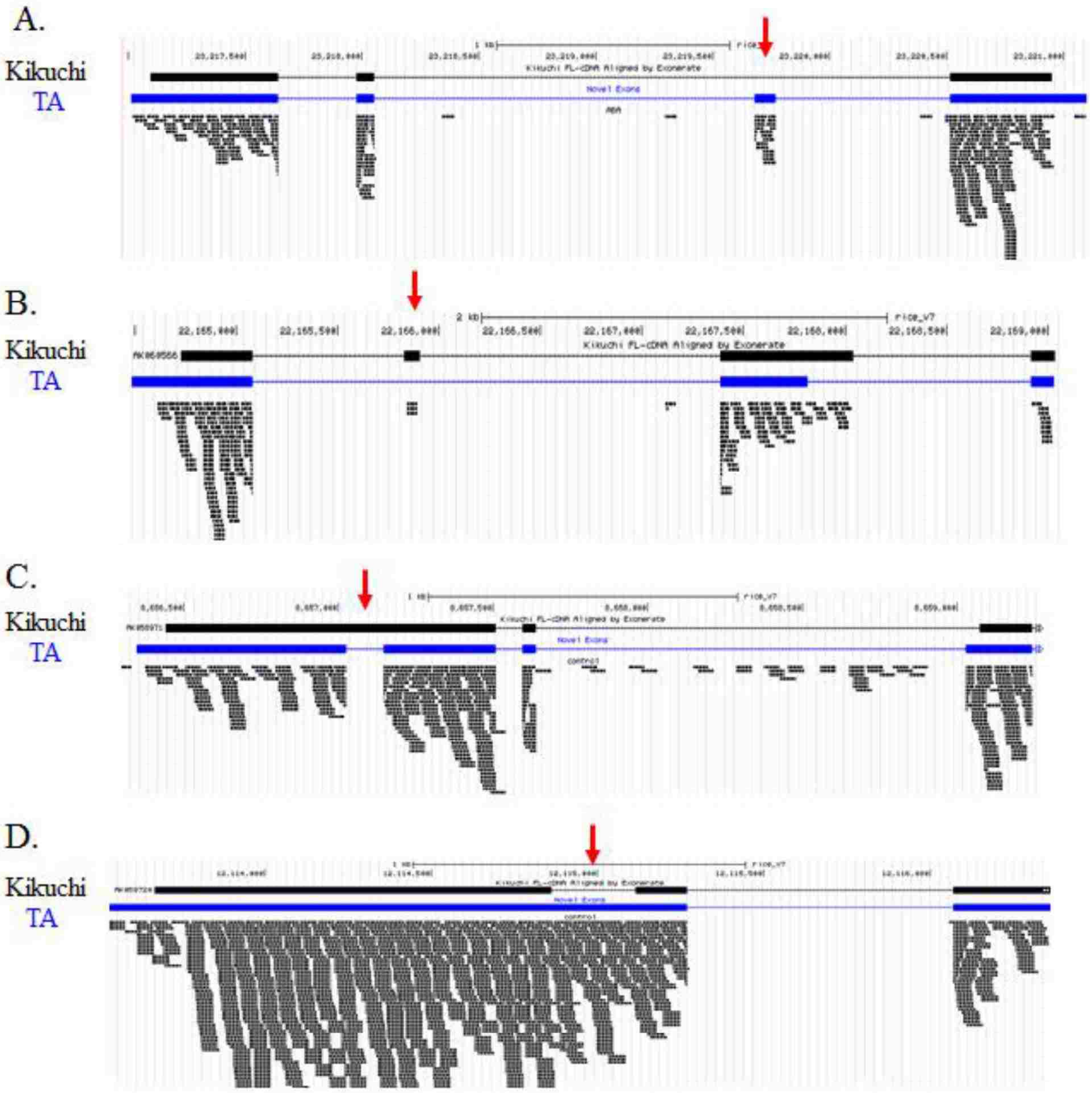
Supplementary Figure S8: Tiling Assembly can detect exons with an expression as low as 50 RPKE. In order to determine the point at which Tiling Assembly fails to correctly identify exons, reads aligning to LOC_Os01g01010 were reiteratively decreased and Tiling Assembly was run on the gene. All exons of the gene were correctly identified at expression levels of 50 RPKE, as can be seen with exons e3 and e4 in the red boxes. Below 50 RPKE, exons began to be misidentified. The user is able to specify the minimum expression level required for exon identification by Tiling Assembly.

Supplemental Figure S9



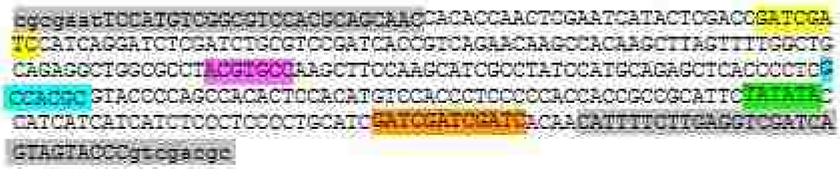
Supplemental Figure S9: Genes where introns are retained at less than 50% were recognized as introns by Tiling Assembly. In order to identify the most common isoform of a gene where intron retention is a possibility, a 50% read-depth threshold was used. Tophat junction alignments were recognized as introns if the read depth across the junction was less than 50% of the read depth of the exons on either side of the junction. This threshold is user-adjustable.

Supplemental Figure S10


A.
Kikuchi
TA
Kikuchi FL-cDNA Aligned by Exonerate
Novel Exons
B.
Kikuchi
TA
2 kb
rice_v7
Kikuchi FL-cDNA Aligned by Exonerate
C.
Kikuchi
TA
rice_v7
Kikuchi FL-cDNA Aligned by Exonerate
Novel Exons
control
D.
Kikuchi
TA
rice_v7
Kikuchi FL-cDNA Aligned by Exonerate
Novel Exons
control

Supplemental Figure S10: Differences between Tiling Assembly and FL-cDNAs may be attributed to alternative splicing. A large number of FL-cDNAs agreed with Tiling Assembly-identified genes, however, there were some areas where the exon number differed between Tiling Assembly and its corresponding FL-cDNA. The red arrows in the above images indicate A) Tiling Assembly has an extra exon, B) Tiling Assembly is missing an exon, and C) Tiling Assembly has an extra intron, and D) Tiling Assembly is missing an intron.

Supplemental Figure S11



```
ggcgaatTCCATGTCGGCGTCCACGCAGCAACCACACCAACTGGAATCATACTCGACCGATCGA
TCCATCAGGATCTCGATCTGCGTCCGATCACCGTCAGAACAAGCCACAAGCTTAGTTTTGGCTG
CAGAGGCTGGCGCCTACGTGCCAAGCTTCCAAGCATCGCCTATCCATGCAGAGCTCACCCCTCG
CCACGCGTACCCCAGCCACACTCCACATGTCCACCCTCCCCCACCACCGCCGCATTCTATATAC
CATCATCATCATCTCCCTCCCCTGCATCGATCGATCGATCACAACATTTTCTTCAGGTCGATCA
GTAGTACCCgtcgagc
```

Primer

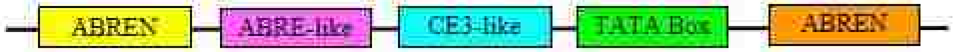
ABREN - promoter

ABREN - 5' UTR

TATA box

ABRE-like

CE3-like



Supplemental Figure 11: PCR product of oleosin promoter region

Primers are highlighted in gray, the lowercase portions of the primers do not base pair to the oleosin promoter. The ABREN in the promoter and the 5’ UTR are highlighted in yellow and orange respectively. The TATA box is highlighted in green. The ABRE-like is highlighted in fuchsia. The CE3-like is highlighted in cyan.

Supplemental tables

Supplemental Table S1

Treatment	**Reads Generated**	**Reads Mapped**	**% Mapped**	**Mapped to Locus**	**Unannotated**	**% Unannotated**
Control	37,488,762	31,186,982	83.2	28,047,420	3,139,562	10.07
ABA Treated	39,794,659	32,961,116	82.8	30,303,193	2,657,923	8.06
GA Treated	40,148,770	33,607,753	83.7	31,629,034	1,978,719	5.89
ABA + GA Treated	40,341,591	33,854,867	83.9	31,420,868	2,433,999	7.19
Total	157,773,782	131,610,718	83.4	121,400,515	10,210,203	8.41

Supplemental Table S2

Novel Gene ID	**Chrom**	**Start Position**	**End Position**	**Length (bp)**	**# of Reads**	**RPKM**	**ABA Fold**	**GA Fold**	**ABA+ GA Fold**	**% Match**	**Accession # of Homologous Protein**	**Accession # of Homologous miRNA**	**Cufflinks/ Novel Algorithm/Both**
OsNG01-01	chr01	184092	184839	747	597	6.1	0.6	0.5	0.4	15.3	BAD67708.1		cuff
OsNG01-02	chr01	223833	224722	889	808	6.9	0.9	0.4	1.2	11.4	ZP_06581228.1		both
OsNG01-03	chr01	2458474	2461452	2978	728	5.4	0.6	1.1	0.5	16.1	BAD61248.1		cuff
OsNG01-04	chr01	2480186	2482092	1906	196	1.3	1.6	1.1	0.9	11.7	ABC18336.1		cuff
OsNG01-05	chr01	2598361	2598726	365	332	6.9	3.2	2.6	2.1	0.0	EHB03565.1		cuff
OsNG01-06	chr01	3237048	3237877	829	169	1.5	0.3	0.6	0.5	24.7	EEC69991.1		cuff
OsNG01-07	chr01	5108106	5108678	572	120	1.8	0.7	0.5	0.7	18.5			cuff
OsNG01-08	chr01	5642860	5643437	577	92	1.7	0.7	0.5	0.5	0.0	NP_001172223.1		cuff
OsNG01-09	chr01	5642898	5644892	1994	282	3.4	0.8	0.5	0.4	7.2	NP_001054892.1		cuff
OsNG01-10	chr01	5644517	5645243	726	193	7.0	0.8	0.5	0.4	0.0	NP_001054892.1		cuff
OsNG01-11	chr01	7592544	7594554	2010	83	1.2	1.6	2.3	1.6	15.5	NP_001046462.1		cuff
OsNG01-12	chr01	7929467	7929616	149	21	1.1	999.0	999.0	999.0	0.0			cuff
OsNG01-13	chr01	8721662	8724197	2535	2460	7.4	1.4	108.3	31.8	11.4		osa-MIR2124a	cuff
OsNG01-14	chr01	8734425	8734925	500	85	1.3	999.0	999.0	999.0	0.0	EGO60756.1		cuff
OsNG01-15	chr01	9463902	9466847	2945	1319	4.1	2.8	2.3	1.7	1.7	EEE54310.1		cuff
OsNG01-16	chr01	9477245	9478130	885	349	3.0	0.9	3.9	2.4	6.9	AAX96243.1		both
OsNG01-17	chr01	9731634	9732573	939	198	1.6	0.6	2.7	2.3	23.4	BAD81121.1	osa-MIRf10254-akr	cuff
OsNG01-18	chr01	10397560	10398302	742	102	1.0	1.5	0.3	1.6	11.3	BAD09555.1	osa-MIRf11608-akr	cuff
OsNG01-19	chr01	11986820	11988006	1186	231	1.5	0.2	0.8	0.2	22.4	BAD45576.1		cuff
OsNG01-20	chr01	13197352	13198669	1317	309	1.8	0.8	0.4	0.9	13.2	NP_001172315.1		both

OsNG01-21	chr01	14801755	14803487	1732	362	2.2	1.2	0.4	1.5	8.0	EAY74027.1		cuff
OsNG01-22	chr01	14990747	14991289	541	3362	47.1	0.4	0.7	0.4	6.3	BAD88330.1		cuff
OsNG01-23	chr01	15546335	15547733	1398	202	1.1	9.6	3.0	7.5	20.0	EEC83524.1		cuff
OsNG01-24	chr01	17274490	17277339	2849	779	2.1	1.3	2.9	2.6	14.7	NP_001043130.1	sbi-MIR396c	cuff
OsNG01-25	chr01	18837935	18838577	642	275	3.3	1.5	0.1	1.5	8.6	NP_001172389.1		cuff
OsNG01-26	chr01	19000325	19001865	1540	102	1.8	1.9	5.8	11.7	10.3	AAX95708.1		cuff
OsNG01-27	chr01	22290684	22291300	616	91	1.1	0.4	2.7	2.5	0.0	BAD45117.1		cuff
OsNG01-28	chr01	24389628	24390767	1139	4405	47.6	286.2	54.8	1695.0	15.3	EAY74961.1		cuff
OsNG01-29	chr01	24833499	24837542	4043	893	3.9	1.9	4.0	2.4	11.0			cuff
OsNG01-30	chr01	25018141	25020579	2438	363	3.5	2.0	7.7	5.4	18.0	NP_001043629.1		cuff
OsNG01-31	chr01	26050715	26052697	1982	130	1.5	0.6	0.4	0.4	20.9	NP_001046306.2		cuff
OsNG01-32	chr01	26307737	26310799	3062	401	3.0	0.3	15.5	0.2	7.3	EAY75193.1		cuff
OsNG01-33	chr01	27932597	27934082	1485	1745	9.8	0.5	0.3	0.2	0.0	NP_001043862.1		both
OsNG01-34	chr01	27949189	27950410	1221	963	6.0	1.3	1.3	3.2	19.2		sbi-MIR396c	both
OsNG01-35	chr01	27956950	27958244	1294	359	2.1	3.4	1.1	0.6	0.0	EEQ86599.1		cuff
OsNG01-36	chr01	30871235	30872041	806	325	3.1	0.8	1.3	0.6	23.3	NP_001042828.1		cuff
OsNG01-37	chr01	31414092	31415994	1902	856	3.4	0.3	0.4	0.0	13.3	BAD09343.1	osa-MIRf10116-akr	both
OsNG01-38	chr01	31490799	31491990	1191	1032	6.6	1.3	0.8	3.1	6.8	NP_001044257.1		both
OsNG01-39	chr01	31556742	31557709	967	145	1.1	0.8	1.7	1.7	14.4	XP_001502621.3		cuff
OsNG01-40	chr01	32511446	32512024	578	135	1.8	0.6	0.6	0.5	0.0	BAD52720.1		cuff
OsNG01-41	chr01	32809908	32811717	1809	304	1.7	1.7	0.7	0.3	12.4	EEE55471.1		cuff
OsNG01-42	chr01	32811824	32812875	1051	147	1.5	2.0	0.9	0.2	5.5	EEE55471.1		cuff
OsNG01-43	chr01	33520317	33521448	1131	346	2.3	0.6	1.6	2.0	14.4	XP_002439577.1		cuff
OsNG01-44	chr01	33763504	33764436	932	237	1.9	1.5	0.4	0.6	0.0			cuff
OsNG01-45	chr01	34255912	34256620	708	161	1.7	1.1	1.1	0.9	0.0	BAD68064.1		both
OsNG01-46	chr01	34261303	34265131	3828	748	1.5	0.8	2.5	2.0	21.5		osa-MIR442	novel
OsNG01-47	chr01	34494665	34497472	2807	2481	25.4	0.4	0.5	0.2	17.8	NP_001172307.1	osa-MIRf11964-akr	both
OsNG01-48	chr01	34496781	34497485	704	1393	15.0	0.3	0.4	0.2	12.5	BAD29401.1	osa-MIRf10183-akr	cuff
OsNG01-49	chr01	34500721	34502501	1780	984	4.2	0.5	0.6	0.2	18.1	EEE57161.1	osa-MIR809b	cuff
OsNG01-50	chr01	34655846	34656979	1133	292	2.0	51.1	3.7	80.1	0.0	XP_002623652.1		cuff
OsNG01-51	chr01	34700879	34701722	843	128	1.2	0.3	0.0	0.1	4.0	BAD09604.1	osa-MIRf10735-akr	cuff
OsNG01-52	chr01	35744283	35748523	4240	661	1.2	1.5	3.2	3.4	7.3	ABA94062.2		both
OsNG01-53	chr01	35747943	35748552	609	85	1.3	0.5	1.0	0.8	9.2	ABA94062.2		cuff
OsNG01-54	chr01	35955559	35956544	985	582	4.9	0.7	1.0	0.7	0.0	NP_001044745.1		cuff
OsNG01-55	chr01	35955587	35958645	3058	1194	5.6	0.8	0.7	0.6	11.2	BAD81693.1	osa-MIR445a	cuff
OsNG01-56	chr01	35955616	35958670	3054	1173	5.9	0.8	0.7	0.6	11.2	BAD81693.1	osa-MIR445a	cuff
OsNG01-57	chr01	36092795	36093019	224	36	1.2	1.9	2.6	3.0	0.0			cuff
OsNG01-58	chr01	37542834	37550324	7490	699	3.2	2.1	1.5	1.3	3.3	NP_001172667.1	osa-MIRf10603-akr	cuff
OsNG01-59	chr01	38839141	38841380	2239	768	5.1	0.9	0.3	0.1	20.1	EAZ38029.1		cuff
OsNG01-60	chr01	39036311	39038836	2525	205	1.9	999.0	999.0	999.0	24.0	BAB86077.1		cuff
OsNG01-61	chr01	40911430	40913146	1716	319	2.9	0.8	1.1	0.9	21.0	NP_001172722.1		both
OsNG01-62	chr01	41498416	41503293	4877	3703	9.1	4.8	0.3	0.5	16.7	BAB63848.1		cuff
OsNG02-01	chr02	3135629	3136156	527	110	1.6	0.4	0.3	0.3	0.0	EEC72527.1		cuff
OsNG02-02	chr02	3271371	3272033	662	642	7.4	6.2	0.5	11.7	0.0	EEC72537.1		cuff
OsNG02-03	chr02	4019099	4023098	3999	867	9.8	0.7	0.8	0.5	12.6	EAZ31490.1		cuff
OsNG02-04	chr02	4524859	4526890	2031	122	2.2	1.6	1.5	0.9	1.9	BAD25133.1		cuff
OsNG02-05	chr02	4908077	4908695	618	569	7.0	1.9	7.3	6.4	0.0			novel
OsNG02-06	chr02	5938320	5939466	1146	770	5.1	4.3	0.1	1.2	16.1	AAM93712.1		both
OsNG02-07	chr02	6713386	6713597	211	29	1.0	2.2	0.8	0.6	0.0			cuff
OsNG02-08	chr02	6994115	6995904	1789	458	3.9	0.6	2.1	0.6	16.0	BAD17066.1		cuff
OsNG02-09	chr02	7236934	7238228	1294	172	2.2	0.6	0.4	0.4	0.0			cuff
OsNG02-10	chr02	7934657	7935444	787	308	3.0	0.4	0.1	0.1	24.9	NP_001176940.1		both
OsNG02-11	chr02	8383189	8384331	1142	526	3.5	0.1	0.1	0.0	11.6	NP_001172877.1		both
OsNG02-12	chr02	8533590	8533892	302	49	1.2	23.7	10.2	11.1	21.9	BAD25027.1	osa-MIRf11024-akr	cuff
OsNG02-13	chr02	8699380	8700179	799	185	2.0	0.7	0.8	0.4	5.1	EJK67546.1		cuff

OsNG02-14	chr02	12409551	12410682	1131	275	1.8	0.5	2.1	0.3	10.4	BAD15980.1		cuff
OsNG02-15	chr02	13849866	13853500	3634	2877	18.0	0.8	0.8	0.6	22.5	NP_001052085.1		cuff
OsNG02-16	chr02	13964033	13964390	357	49	1.0	4.3	10.7	6.9	22.7	XP_003784341.1	ptc-MIRf11913-akr	cuff
OsNG02-17	chr02	14971431	14972087	656	730	9.8	0.4	0.9	0.6	16.2	AAG13546.1		both
OsNG02-18	chr02	16196488	16198512	2024	195	3.1	3.5	26.0	14.7	6.7	AFW71440.1		cuff
OsNG02-19	chr02	16251521	16252345	824	201	1.9	0.2	0.2	0.1	0.0	YP_399103.1		cuff
OsNG02-20	chr02	16629110	16630729	1619	453	2.1	0.8	0.3	0.3	13.5	BAD22469.1		both
OsNG02-21	chr02	17436749	17437138	389	67	1.3	4.0	1.4	0.6	0.0	ACV32571.1		cuff
OsNG02-22	chr02	18373609	18374430	821	787	7.3	0.6	0.9	0.8	7.4	XP_002300326.1		cuff
OsNG02-23	chr02	18942818	18945627	2809	917	2.6	1.0	0.1	0.1	14.0	BAD26315.1		both
OsNG02-24	chr02	19844394	19852184	7790	306	2.1	1.0	4.3	2.4	11.7	BAD69364.1	osa-MIR812j	cuff
OsNG02-25	chr02	19946258	19947585	1327	220	1.3	1.4	13.8	4.2	17.6	BAD27921.1		cuff
OsNG02-26	chr02	20128918	20129442	524	91	1.3	0.2	14.2	5.8	15.1	NP_497811.1		cuff
OsNG02-27	chr02	20813397	20813829	432	66	1.2	5.7	7.4	16.6	0.0	XP_001841477.2		cuff
OsNG02-28	chr02	20814002	20814344	342	61	1.4	0.9	0.9	4.3	0.0	XP_002007964.1		cuff
OsNG02-29	chr02	21494528	21495777	1249	248	1.5	1.3	0.5	0.6	0.0	BAD15831.1		cuff
OsNG02-30	chr02	22410488	22411266	778	24	1.3	999.0	999.0	999.0	0.0	DAA34846.1		cuff
OsNG02-31	chr02	22429741	22429909	168	23	1.0	0.0	0.2	0.3	0.0			cuff
OsNG02-32	chr02	22520126	22521035	909	226	1.9	0.3	1.6	0.3	14.5	BAD29332.1		cuff
OsNG02-33	chr02	22892407	22893124	717	1010	10.7	0.8	1.5	1.4	0.0	ZP_11012553.1	osa-MIR5493	both
OsNG02-34	chr02	23026471	23027010	539	116	2.0	0.9	2.7	1.7	15.6	AAG13543.1		cuff
OsNG02-35	chr02	24558188	24559453	1265	190	1.1	999.0	999.0	999.0	6.8	NP_001047446.1		cuff
OsNG02-36	chr02	25404449	25405076	627	80	1.1	0.4	2.2	0.1	0.0	BAG95645.1		cuff
OsNG02-37	chr02	25530515	25532089	1147	4202	31.2	0.7	9.7	3.9	13.3	AAT44228.1		cuff
OsNG02-38	chr02	26124236	26126607	2371	660	2.1	2.0	41.9	16.3	0.0	XP_002147341.1	osa-MIR166d	both
OsNG02-39	chr02	26573152	26573932	780	209	2.0	0.6	6.3	2.9	22.2	XP_002327722.1		cuff
OsNG02-40	chr02	27267483	27273703	6220	1598	3.0	2.5	1.8	1.5	4.0	NP_001047701.1	bdi-MIR5169	both
OsNG02-41	chr02	27296205	27296950	745	147	1.7	0.3	0.3	0.3	0.0	ZP_10191562.1		cuff
OsNG02-42	chr02	28168372	28174939	6567	907	4.3	1.7	1.7	0.7	2.5	BAD07588.1	tae-MIR160	both
OsNG02-43	chr02	28423007	28424637	1630	919	4.3	0.5	0.4	0.1	12.7	BAD07894.1	osa-MIRf11634-akr	cuff
OsNG02-44	chr02	28550863	28551770	907	225	1.9	0.7	30.4	20.0	16.0	BAD07915.1	osa-MIRf10116-akr	both
OsNG02-45	chr02	29251287	29251572	285	63	1.7	4.4	0.4	0.2	21.8			cuff
OsNG02-46	chr02	29673639	29674639	1000	243	4.2	1.7	0.7	0.8	0.0	EKU20607.1		cuff
OsNG02-47	chr02	29709929	29711658	1729	593	6.9	1.8	3.2	3.1	5.1	EEE57685.1		cuff
OsNG02-48	chr02	30232429	30232878	449	304	5.1	0.2	61.9	7.4	0.0	BAD16086.1		both
OsNG02-49	chr02	30351191	30351824	633	112	1.3	2.2	1.1	0.7	5.8	EEE57734.1		cuff
OsNG02-50	chr02	30379627	30380488	861	1165	10.3	0.4	0.4	0.1	16.8	BAD29060.1	osa-MIRf11727-akr	both
OsNG02-51	chr02	31450793	31451559	766	107	1.1	2.2	0.5	1.1	0.0			cuff
OsNG02-52	chr02	33185180	33185786	606	134	1.7	3.2	2.0	1.2	11.9	XP_002513208.1		cuff
OsNG02-53	chr02	33967909	33969100	1191	241	3.0	2.9	0.8	0.3	14.1	YP_004222744.1		cuff
OsNG02-54	chr02	34898783	34899169	386	34	1.1	4.7	5.1	5.1	0.0	BAD21666.1		cuff
OsNG02-55	chr02	34963681	34964481	800	237	4.2	1.9	2.2	1.1	0.0			cuff
OsNG02-56	chr02	34964734	34966034	1300	205	1.2	2.1	3.4	1.6	0.0			cuff
OsNG02-57	chr02	35430220	35431319	1099	335	2.3	0.3	0.2	0.1	24.8		osa-MIRf11	novel
OsNG02-58	chr02	35470181	35471760	1579	429	7.4	1.2	0.8	0.8	11.7	XP_003980779.1	oru-MIR806	cuff
OsNG03-01	chr03	529244	530619	1375	66	1.0	1.0	0.8	1.1	0.0			cuff
OsNG03-02	chr03	529261	535285	6024	100	1.7	1.2	0.7	0.8	4.4	BAD69283.1	osa-MIRf11994-akr	cuff
OsNG03-03	chr03	711104	715191	4087	3732	12.4	2.7	1.5	1.2	4.0	EEC74379.1		cuff
OsNG03-04	chr03	727244	728844	1600	851	19.7	1.1	0.6	0.4	8.4	XP_001432268.1		cuff
OsNG03-05	chr03	836577	840760	4183	327	2.2	1.2	0.4	0.3	7.5	BAD25249.1	osa-MIR442	cuff
OsNG03-06	chr03	1744742	1747501	2759	838	5.4	0.9	0.9	0.8	0.0	EEC74447.1		cuff
OsNG03-07	chr03	2583798	2583940	142	28	1.5	2.6	1.4	1.6	0.0			cuff
OsNG03-08	chr03	3232752	3233224	472	68	1.3	4.1	3.1	2.5	0.0	EEE58361.1		cuff
OsNG03-09	chr03	3520002	3521747	1745	580	2.5	0.3	0.4	0.1	19.4			novel
OsNG03-10	chr03	4020017	4021407	1390	254	1.4	0.3	0.1	0.0	15.3	EEE62272.1	osa-MIRf10367-akr	cuff

OsNG03-11	chr03	4346052	4347291	1239	1788	15.4	0.4	0.1	0.0	3.5	YP_708918.1		cuff
OsNG03-12	chr03	4677586	4677997	411	206	3.8	1.1	1.4	2.2	0.0	AAN64149.1		cuff
OsNG03-13	chr03	4678890	4679041	151	91	4.6	1.2	1.3	1.2	0.0	EAY88850.1		cuff
OsNG03-14	chr03	5158059	5159250	1191	241	1.5	0.6	3.9	1.5	0.0	NP_001049276.1		both
OsNG03-15	chr03	5540857	5541852	995	242	1.8	0.1	0.0	0.0	7.2	XP_001691236.1		cuff
OsNG03-16	chr03	5693579	5694538	959	207	5.7	0.6	1.8	0.7	12.8	XP_003861052.1		cuff
OsNG03-17	chr03	6003436	6004004	568	101	1.4	1.7	10.2	5.9	8.1	NP_001063788.1	sbi-MIR396c	cuff
OsNG03-18	chr03	6174670	6176152	1482	78	1.6	0.9	2.9	1.3	14.9	NP_001043094.1		cuff
OsNG03-19	chr03	6178668	6181002	2334	355	2.1	1.1	3.7	1.2	17.2	NP_001049391.1	osa-MIRf10715-akr	cuff
OsNG03-20	chr03	7144911	7145915	1004	152	1.3	1.0	2.1	1.1	12.2	EEE64060.1		cuff
OsNG03-21	chr03	7234099	7234440	341	57	1.3	0.4	0.3	0.3	0.0	EEC74834.1		cuff
OsNG03-22	chr03	8205709	8206075	366	55	1.1	3.5	2.1	0.8	0.0	YP_001312850.1		cuff
OsNG03-23	chr03	11084210	11086392	2182	601	2.1	0.6	1.2	1.6	20.0	BAB63561.1	osa-MIRf10947-akr	both
OsNG03-24	chr03	11835308	11836957	1649	260	4.4	1.5	2.9	2.5	1.9	XP_003101923.1		cuff
OsNG03-25	chr03	11889292	11889517	225	42	1.4	3.4	0.4	3.1	0.0			cuff
OsNG03-26	chr03	11890759	11890967	208	42	1.5	4.0	1.9	1.1	0.0			cuff
OsNG03-27	chr03	12680612	12681404	792	210	2.0	0.6	0.6	0.3	0.0	ZP_04539105.1		both
OsNG03-28	chr03	12919072	12920426	1354	255	1.4	0.1	0.1	0.0	8.6	BAB63542.1		cuff
OsNG03-29	chr03	12936229	12937094	865	193	1.7	0.3	0.9	0.1	11.4	NP_001050092.1		cuff
OsNG03-30	chr03	13080612	13081642	1030	222	1.6	2.2	0.5	1.0	17.0			cuff
OsNG03-31	chr03	13081728	13083033	1305	642	3.7	1.4	0.3	0.7	23.2	EEC69983.1	osa-MIRf10773-akr	cuff
OsNG03-32	chr03	13130185	13130886	701	204	2.2	0.7	0.0	0.0	0.0	EEE57924.1		cuff
OsNG03-33	chr03	13706987	13707718	731	134	1.4	0.5	3.4	1.4	5.9	ZP_07306450.1		cuff
OsNG03-34	chr03	16196681	16200764	4083	485	1.2	1.4	1.4	1.5	5.8	EEE55732.1		both
OsNG03-35	chr03	17689604	17690887	1283	582	3.7	0.9	3.2	0.5	23.5	AAP20861.1	osa-MIRf10928-akr	cuff
OsNG03-36	chr03	18206252	18209164	2912	2106	14.1	4.9	3.9	8.2	1.8	EEE59315.1	osa-MIR435	cuff
OsNG03-37	chr03	19178026	19181555	3529	1431	5.0	2.3	3.1	1.6	12.5	NP_001143823.1	osa-MIR815b	cuff
OsNG03-38	chr03	19714255	19716226	1971	688	4.7	1.2	1.2	1.1	20.2	XP_001457417.1		both
OsNG03-39	chr03	21015314	21015913	599	90	1.1	1.1	0.1	0.0	11.0	AAP12964.1		cuff
OsNG03-40	chr03	21336815	21338816	2001	208	3.0	1.7	2.6	1.5	16.5	EEE63477.1		cuff
OsNG03-41	chr03	22105896	22106369	473	89	1.4	0.8	3.1	2.7	0.0			cuff
OsNG03-42	chr03	22339382	22344060	4678	3079	17.2	1.2	1.6	1.6	12.9	NP_001050613.2		cuff
OsNG03-43	chr03	22351018	22352143	1125	152	1.0	7.6	4.8	4.5	14.5	EEE60796.1	osa-MIRf10495-akr	cuff
OsNG03-44	chr03	22610908	22612967	2059	569	6.1	0.7	2.9	2.4	0.0	XP_003559690.1		cuff
OsNG03-45	chr03	22713808	22716977	3169	221	1.2	1.1	6.0	6.6	8.4	ABG22576.1		cuff
OsNG03-46	chr03	22747395	22748221	826	426	3.9	0.3	0.1	0.1	0.0	EEE68744.1		cuff
OsNG03-47	chr03	22872322	22875472	3150	1165	8.4	1.0	2.1	1.5	0.0	EAY90969.1		both
OsNG03-48	chr03	22873171	22875483	2312	584	3.8	1.4	2.6	2.6	0.0	XP_003446480.1		cuff
OsNG03-49	chr03	23629772	23630125	353	118	2.5	1.2	0.3	0.4	12.2			cuff
OsNG03-50	chr03	24002641	24004359	1718	205	2.0	0.6	1.6	0.4	10.6	AAP12955.1		cuff
OsNG03-51	chr03	24006732	24007300	568	160	2.1	0.3	0.4	0.2	0.0	EAZ27860.1		cuff
OsNG03-52	chr03	24582507	24582887	380	64	1.3	1.0	1.1	0.5	0.0	XP_002140519.1		cuff
OsNG03-53	chr03	26246965	26249112	2147	402	5.5	1.7	2.8	3.3	13.7	NP_001173577.1		cuff
OsNG03-54	chr03	26965727	26968460	2733	224	1.2	1.4	2.9	1.2	14.0	BAD69080.1		cuff
OsNG03-55	chr03	29006373	29007320	947	277	2.2	0.1	1.9	0.1	24.4	EEC70639.1	osa-MIRf10505-akr	cuff
OsNG03-56	chr03	29079749	29081382	1633	9185	105.9	0.2	0.0	0.0	16.5	NP_001173618.1		cuff
OsNG03-57	chr03	29088731	29089723	992	186	1.4	0.4	0.0	0.0	0.0	EEC76078.1		cuff
OsNG03-58	chr03	29160206	29161028	822	116	1.2	0.8	1.2	0.8	4.0			cuff
OsNG03-59	chr03	31096846	31100591	3745	949	2.0	2.9	0.8	1.1	19.9	EEE70170.1		cuff
OsNG03-60	chr03	31799753	31800577	824	253	2.3	0.1	0.0	0.0	0.0	CBJ26209.1		cuff
OsNG03-61	chr03	32249415	32250465	1050	690	5.4	5.0	3.3	2.5	0.0	AAM19033.1		cuff
OsNG03-62	chr03	32250521	32250834	313	84	2.0	5.4	3.0	1.8	11.5			cuff
OsNG03-63	chr03	34208495	34210945	2450	389	1.5	1.6	2.3	1.7	1.3	AAO60019.1		cuff
OsNG03-64	chr03	34294103	34297800	3697	213	1.3	1.4	3.6	2.9	8.0	AAO60029.1		cuff

OsNG03-65	chr03	34462159	34466087	3928	271	2.5	1.1	2.4	1.9	9.1	XP_002987131.1	osa-MIRf10001-akr	both
OsNG03-66	chr03	35558481	35560378	1897	678	2.7	1.0	1.5	1.1	12.9	EEC76521.1		cuff
OsNG03-67	chr03	35558558	35561150	2592	879	5.9	0.9	1.4	1.0	13.0	AAS07366.1		cuff
OsNG04-01	chr04	243204	243934	730	634	6.6	0.9	1.0	0.8	6.6	NP_001173728.1		both
OsNG04-02	chr04	1740363	1742035	1672	174	1.4	1.9	2.9	2.8	22.5	EEE60421.1		both
OsNG04-03	chr04	4310437	4311616	1179	96	1.8	0.5	0.8	0.5	6.4	YP_063566.1		cuff
OsNG04-04	chr04	4377655	4381184	3529	848	3.7	1.6	1.2	1.2	13.0	EAZ29707.1		both
OsNG04-05	chr04	5794950	5795667	717	100	1.5	0.5	2.7	2.5	0.0	EJK48265.1		cuff
OsNG04-06	chr04	7915340	7918371	3031	284	2.4	1.9	2.2	1.5	6.8	EEE60529.1	ssp-MIR437a	cuff
OsNG04-07	chr04	7934995	7939110	4115	468	1.9	2.8	2.8	2.7	8.0	EEE63277.1		cuff
OsNG04-08	chr04	8759866	8762411	2545	173	1.6	1.0	3.9	2.6	4.3	EAY74435.1		cuff
OsNG04-09	chr04	9668048	9669430	1382	998	6.4	1.1	0.5	0.5	15.7	NP_001052311.1		both
OsNG04-10	chr04	10461267	10463142	1875	332	4.3	1.0	0.4	0.5	9.7	NP_001052329.1		both
OsNG04-11	chr04	11234984	11235207	223	51	1.7	2.6	1.3	1.2	21.5		sbi-MIR396c	cuff
OsNG04-12	chr04	11235312	11235873	561	165	2.2	2.8	2.0	4.5	11.8	AAP12932.1		cuff
OsNG04-13	chr04	14715891	14718849	2958	159	1.2	0.9	1.3	2.2	17.7	CAD40056.3	ptc-MIRf11913-akr	cuff
OsNG04-14	chr04	15479336	15482139	2803	567	5.7	1.8	1.9	1.4	16.6	EEE59496.1		both
OsNG04-15	chr04	16668416	16670476	2060	324	1.2	0.7	0.9	1.0	10.3	EAZ30345.1		both
OsNG04-16	chr04	16915492	16920435	4943	857	3.3	1.3	0.7	0.9	7.2	CCC56932.1		both
OsNG04-17	chr04	18306692	18310721	4029	980	3.5	2.1	1.5	1.6	17.4	EAY87140.1		cuff
OsNG04-18	chr04	19784182	19785157	975	289	2.3	5.6	2.9	8.5	14.8	EAZ03406.1		cuff
OsNG04-19	chr04	19790973	19791988	1015	1843	13.8	0.6	8.4	7.8	7.1	NP_001052695.1	sbi-MIR396c	both
OsNG04-20	chr04	21440266	21448038	7772	548	2.3	3.5	1.6	2.1	2.5	CAD41022.1	osa-MIRf10377-akr	cuff
OsNG04-21	chr04	21472010	21486267	14257	829	8.2	1.7	1.1	1.0	4.3	NP_001145025.1	osa-MIR2118i	cuff
OsNG04-22	chr04	21556918	21559348	2430	197	1.3	1.7	0.9	0.9	6.6		osa-MIR812g	cuff
OsNG04-23	chr04	22071148	22071953	805	246	2.3	1.2	1.5	2.6	0.0	XP_003579820.1		cuff
OsNG04-24	chr04	22119536	22119945	409	116	2.2	2.2	3.9	10.9	0.0			cuff
OsNG04-25	chr04	22795474	22796367	893	821	7.0	0.5	0.4	0.2	0.0	ACG46097.1		cuff
OsNG04-26	chr04	22819519	22820252	733	1254	13.0	1.9	1.4	3.7	10.9	BAD69356.1		cuff
OsNG04-27	chr04	23310552	23311910	1358	386	2.2	0.7	0.6	0.6	10.9		tae-MIR113	novel
OsNG04-28	chr04	24031309	24032496	1187	845	6.7	0.4	1.6	0.6	0.0	EEE61213.1		cuff
OsNG04-29	chr04	24101399	24104401	3002	170	0.4	3.4	3.6	11.8	9.6		osa-MIRf11	novel
OsNG04-30	chr04	25055606	25055961	355	50	1.1	0.4	0.1	0.1	0.0	EFB15800.1		cuff
OsNG04-31	chr04	25085604	25087705	2101	942	3.4	0.7	2.1	2.4	20.4		osa-MIRf11	novel
OsNG04-32	chr04	26921442	26922706	1264	213	1.4	0.1	6.0	0.6	0.0	ZP_10621748.1		cuff
OsNG04-33	chr04	26980195	26984047	3852	763	1.6	0.6	0.3	0.1	19.3	ABA98771.1		cuff
OsNG04-34	chr04	27053589	27055521	1932	1385	5.4	0.9	4.9	3.5	6.1	NP_001174032.1		both
OsNG04-35	chr04	27351727	27352359	632	229	2.8	0.2	1.9	1.1	4.9	BAD03403.1		both
OsNG04-36	chr04	27989784	27990498	714	113	1.2	0.9	1.1	0.7	14.6	CAE03209.2		cuff
OsNG04-37	chr04	28704637	28706712	2075	1510	5.5	0.6	0.3	0.2	9.3	NP_001067268.1	osa-MIRf11171-akr	cuff
OsNG04-38	chr04	28868875	28870643	1768	2348	10.1	0.7	0.1	0.0	23.6	NP_001176640.1		cuff
OsNG04-39	chr04	28978197	28980024	1827	598	6.1	1.8	2.9	10.7	0.0	EAY95260.1		cuff
OsNG04-40	chr04	29221922	29223124	1202	4456	28.2	0.1	0.0	0.0	3.6	CAH67947.1		both
OsNG04-41	chr04	29531099	29531450	351	59	1.3	0.3	0.1	0.0	11.1	EAY95331.1		cuff
OsNG04-42	chr04	30329939	30330613	674	96	1.3	0.8	0.6	0.6	0.0			cuff
OsNG04-43	chr04	30896250	30897854	1604	255	1.3	0.9	1.0	0.8	6.8	XP_003509588.1		cuff
OsNG04-44	chr04	31175405	31176699	1294	2193	12.9	0.1	0.4	0.2	8.8	EAZ31996.1	osa-MIR818c	both
OsNG04-45	chr04	31527086	31528847	1761	290	1.3	1.3	1.5	0.7	8.2	AAU44297.1	osa-MIR171c	cuff
OsNG04-46	chr04	32550626	32551050	424	80	1.4	2.4	1.2	1.8	19.6	YP_006463912.1		cuff
OsNG04-47	chr04	32646661	32647677	1016	125	1.3	0.5	0.5	0.5	18.3	NP_001048190.1		cuff
OsNG04-48	chr04	34283614	34302998	19383	8653	4.5	4.1	5.2	5.6	0.8	EEE61898.1		cuff
OsNG05-01	chr05	153678	154709	1031	265	2.0	7.9	2.7	9.1	20.7	EEE62001.1		both
OsNG05-02	chr05	458768	459821	1053	532	4.3	0.5	0.2	0.4	21.8	AAK73133.1		both
OsNG05-03	chr05	483358	484594	1236	2646	28.2	0.9	0.3	0.2	17.6	NP_001054432.1		cuff
OsNG05-04	chr05	838794	839941	1147	100	2.0	1.2	1.3	0.7	3.9			cuff
OsNG05-05	chr05	887612	889362	1750	246	1.8	1.0	0.4	0.3	6.1	EEC78410.1	osa-MIR164f	cuff
OsNG05-06	chr05	1645644	1646047	403	128	2.4	2.1	15.6	5.2	0.0	XP_002462371.1		cuff
OsNG05-07	chr05	1646734	1646944	210	35	1.3	999.0	999.0	999.0	0.0			cuff
OsNG05-08	chr05	4165072	4166420	1348	262	2.1	1.0	1.2	0.9	0.0	XP_001030405.1		cuff

OsNG05-09	chr05	4169187	4170767	1580	118	1.3	2.7	1.7	1.1	10.7			cuff
OsNG05-10	chr05	4410267	4411875	1608	265	1.3	2.5	2.1	0.7	20.6	EEE60451.1		cuff
OsNG05-11	chr05	5818001	5821421	3420	537	5.8	0.8	1.1	0.9	4.6	NP_001174263.1		cuff
OsNG05-12	chr05	7848107	7849465	1358	1469	20.5	2.5	1.7	1.5	16.6	AFW71772.1		cuff
OsNG05-13	chr05	8258455	8266150	7695	181	4.3	2.6	1.1	1.0	16.2	Q0DJS1.3	osa-MIR5162	cuff
OsNG05-14	chr05	8271060	8272623	1563	238	3.7	3.0	1.3	0.8	3.2	Q0DJS1.3	osa-MIRf10603-akr	cuff
OsNG05-15	chr05	9445575	9447132	1557	131	1.0	0.7	0.7	0.7	17.6	ABA94669.2	osa-MIR827	cuff
OsNG05-16	chr05	12953947	12955636	1689	271	1.4	2.8	17.6	6.6	22.6	ABF97069.1		cuff
OsNG05-17	chr05	14252138	14252784	646	110	1.3	1.3	7.6	10.5	12.4	NP_001046963.1		cuff
OsNG05-18	chr05	16692114	16714295	22181	4651	6.0	2.4	2.5	2.4	1.2	EEE63377.1	sbi-MIR437r	cuff
OsNG05-19	chr05	17372629	17373215	586	65	1.0	0.7	0.6	0.3	17.1	EEC79075.1		cuff
OsNG05-20	chr05	17593574	17594611	1037	844	6.2	0.8	0.4	0.7	13.4	BAD07525.1		both
OsNG05-21	chr05	18323072	18325203	2131	234	1.5	0.9	0.6	0.6	5.3	EEC79119.1		cuff
OsNG05-22	chr05	20060714	20061377	663	98	1.1	0.1	0.2	0.0	16.3	BAD44997.1		cuff
OsNG05-23	chr05	20781961	20783076	1115	130	1.0	2.4	8.4	5.5	6.9	NP_001055597.1		cuff
OsNG05-24	chr05	21191836	21192792	956	261	2.1	1.5	0.6	0.8	11.3	AAP12932.1		both
OsNG05-25	chr05	21971436	21972862	1426	186	1.9	1.6	9.5	3.6	20.1	EEC70779.1		cuff
OsNG05-26	chr05	22722740	22723041	301	80	2.0	1.6	0.7	1.7	12.0	XP_003566174.1		cuff
OsNG05-27	chr05	23566029	23569402	3373	245	1.8	1.6	0.6	0.6	3.7	AFR43916.1		both
OsNG05-28	chr05	23661864	23662639	775	214	2.1	0.7	0.2	0.6	19.6	BAD20006.1		both
OsNG05-29	chr05	24131885	24134123	2238	481	2.4	1.1	1.0	0.8	10.2	BAD17473.1		cuff
OsNG05-30	chr05	24432340	24432788	448	68	1.2	0.5	0.1	0.1	13.4	XP_001963928.1		cuff
OsNG05-31	chr05	24473469	24474858	1389	304	1.7	0.4	2.2	0.7	19.9	EAY90939.1		both
OsNG05-32	chr05	26212187	26213716	1529	1443	7.2	2.4	9.3	6.2	3.5	AFW82549.1	osa-MIR1850	both
OsNG05-33	chr05	26539689	26539955	266	47	1.3	3.2	0.8	0.6	0.0	XP_003340527.1		cuff
OsNG05-34	chr05	27011422	27012083	661	119	1.4	0.6	0.7	0.4	0.0	XP_002097084.1		cuff
OsNG05-35	chr05	27148379	27148866	487	59	1.2	2.2	1.1	1.3	0.0	EFY99339.1		cuff
OsNG05-36	chr05	29271493	29272134	641	137	1.6	3.3	1.4	1.5	0.0			cuff
OsNG06-01	chr06	736384	738153	1769	1135	11.5	2.6	2.5	2.3	0.0	EEC79862.1		cuff
OsNG06-02	chr06	1660179	1661858	1679	874	5.8	1.0	2.1	1.4	18.6	EEE65029.1		both
OsNG06-03	chr06	2455087	2456155	1068	192	1.4	2.2	0.0	0.1	21.2	BAC84752.1		cuff
OsNG06-04	chr06	3292122	3292904	782	194	2.1	0.9	0.7	0.5	7.2	XP_501207.2	tae-MIR1132	cuff
OsNG06-05	chr06	7269372	7273647	4275	599	7.0	1.3	1.9	2.2	20.5	BAD37881.1	osa-MIRf12035-akr	cuff
OsNG06-06	chr06	7286794	7289623	2829	424	2.8	0.9	1.4	1.4	18.6	ABF96637.1	osa-MIRf11994-akr	both
OsNG06-07	chr06	7728949	7729641	692	223	2.4	0.9	0.7	2.6	0.0			both
OsNG06-08	chr06	7764012	7765365	1353	174	1.1	0.7	1.7	1.3	0.0	NP_001057291.2		cuff
OsNG06-09	chr06	8942652	8943635	983	203	1.6	1.0	0.1	0.2	0.0	EEE56432.1		cuff
OsNG06-10	chr06	9032420	9032697	277	52	1.4	6.6	1.9	14.7	0.0	ZP_01797769.1		cuff
OsNG06-11	chr06	9197419	9198450	1031	160	1.2	1.2	2.0	1.5	15.0	BAD68436.1	osa-MIR164f	cuff
OsNG06-12	chr06	12350539	12351007	468	69	1.5	1.1	1.0	0.6	0.0	YP_426185.1		cuff
OsNG06-13	chr06	13279867	13283634	3767	487	1.3	1.9	4.0	2.1	16.7	BAD33892.1		cuff
OsNG06-14	chr06	14926300	14927761	1461	235	1.2	1.9	1.5	1.4	24.8	AAU44254.3		cuff
OsNG06-15	chr06	15252701	15253514	813	998	9.3	0.8	1.4	1.0	0.0	ZP_07334613.1		both
OsNG06-16	chr06	16454976	16459836	4860	935	1.5	31.2	12.1	23.0	16.6	ABA97230.1		both
OsNG06-17	chr06	16787573	16788405	832	204	1.9	0.5	0.7	0.4	10.7	XP_004014431.1		both
OsNG06-18	chr06	16815157	16816907	1750	915	4.6	0.6	1.9	0.9	19.4	BAD37702.1	osa-MIRf10463-akr	both
OsNG06-19	chr06	17572050	17572572	522	72	1.0	0.9	0.7	0.7	0.0			cuff
OsNG06-20	chr06	17901034	17901726	692	298	3.3	0.2	0.6	0.3	17.2	BAD26102.1		both
OsNG06-21	chr06	17980847	17981383	536	787	11.2	0.8	0.2	0.4	13.4	EGB07034.1		both
OsNG06-22	chr06	18912207	18914772	2565	244	1.4	1.2	3.6	3.0	15.9	EEE62860.1		both
OsNG06-23	chr06	19568525	19570373	1848	532	2.2	1.5	1.0	1.4	23.3	NP_001057767.1		cuff
OsNG06-24	chr06	19729629	19734412	4783	4620	32.5	0.6	1.0	1.0	5.2	NP_001057777.1		cuff
OsNG06-25	chr06	20089457	20090043	586	86	1.1	0.8	1.9	1.7	6.7	YP_003395137.1		cuff
OsNG06-26	chr06	20976611	20981404	4793	171	2.1	0.6	0.3	0.3	5.1	EEC80797.1	sbi-MIR396c	cuff
OsNG06-27	chr06	22386796	22387348	552	123	1.7	0.5	0.3	0.1	15.6	EAZ20272.1		cuff
OsNG06-28	chr06	22555207	22555858	651	102	1.2	2.2	17.3	11.1	19.5	XP_003351267.1	osa-MIRf10218-akr	both
OsNG06-29	chr06	23727899	23730663	2764	1934	15.4	0.7	1.3	0.7	5.9	XP_003833089.1		both

OsNG06-30	chr06	24410883	24411125	242	51	1.6	10.9	7.0	5.1	0.0			cuff
OsNG06-31	chr06	24945206	24949724	4518	815	1.5	1.7	1.4	0.8	0.0	BAD35363.1		cuff
OsNG06-32	chr06	25903785	25905549	1764	402	1.7	1.3	0.8	1.6	17.9	AAM18729.1		cuff
OsNG06-33	chr06	26464121	26468654	4533	4606	7.7	1.2	0.5	0.4	6.9	NP_001174931.1		cuff
OsNG06-34	chr06	26484484	26484783	299	57	1.4	0.8	0.1	0.6	10.7		vun-MIR157b.1	cuff
OsNG06-35	chr06	26492198	26492992	794	106	1.0	2.2	0.4	0.6	6.4	NP_001058213.1		cuff
OsNG06-36	chr06	27296528	27296820	292	40	1.0	1.5	0.8	0.3	0.0			cuff
OsNG06-37	chr06	27642502	27645677	3175	449	1.1	4.2	3.8	5.0	1.2		osa-MIRf10569-akr	cuff
OsNG06-38	chr06	27651090	27655700	4610	2067	22.0	1.9	1.2	1.4	5.3	NP_001174946.1		cuff
OsNG06-39	chr06	27651566	27655803	4237	2022	3.6	2.0	1.2	1.6	5.8	NP_001174946.1		cuff
OsNG06-40	chr06	28445540	28446975	1435	1956	10.4	1.0	2.3	1.3	9.6	ACL52830.1		cuff
OsNG06-41	chr06	31089179	31092363	3184	250	1.2	1.0	0.8	1.4	6.9	Q5Z5A8.1	osa-MIR2918	cuff
OsNG06-42	chr06	31220873	31221391	518	100	1.5	1.6	1.9	1.8	0.0			cuff
OsNG07-01	chr07	566571	571888	5317	395	2.5	1.6	1.8	2.6	5.8	EEE66449.1		cuff
OsNG07-02	chr07	595773	597269	1496	327	1.7	0.4	0.6	0.4	12.7	BAC15840.1		cuff
OsNG07-03	chr07	1163266	1164078	812	1583	14.8	0.4	0.1	0.0	0.0	BAC83786.1		both
OsNG07-04	chr07	2028131	2030916	2785	340	4.6	0.7	2.4	1.6	9.7	BAC45042.1		cuff
OsNG07-05	chr07	2261168	2266571	5403	161	2.6	1.1	0.9	0.6	3.3	EAZ25827.1		cuff
OsNG07-06	chr07	3779802	3784004	4202	191	0.3	5.9	3.2	25.4	4.3			novel
OsNG07-07	chr07	5364943	5365634	691	312	3.4	0.6	1.7	1.1	15.6	ABA93603.1		cuff
OsNG07-08	chr07	5578804	5579561	757	255	2.6	0.6	0.4	0.4	0.0	BAC79787.1		cuff
OsNG07-09	chr07	5667485	5670153	2668	72	1.3	2.4	0.6	0.6	0.0			cuff
OsNG07-10	chr07	5669231	5670661	1430	89	1.0	1.9	0.5	0.3	9.2	CAE01957.2	osa-MIR819k	cuff
OsNG07-11	chr07	6374463	6375980	1517	1825	9.1	0.1	0.2	0.0	22.9	BAD35303.1		cuff
OsNG07-12	chr07	6571520	6572345	825	135	1.2	1.1	0.2	0.7	6.7	NP_001172521.1		cuff
OsNG07-13	chr07	6660058	6662783	2725	1133	3.7	2.2	3.7	5.2	12.0	EEE66812.1		both
OsNG07-14	chr07	6930472	6931650	1178	1827	11.8	0.7	0.1	0.1	9.6	BAD37707.1	tae-MIR160	both
OsNG07-15	chr07	7355121	7355857	736	77	1.1	0.2	0.4	0.1	14.5	NP_001175111.1		cuff
OsNG07-16	chr07	7712455	7716051	3596	1098	2.6	4.0	3.7	3.8	25.0	NP_001066455.1	osa-MIRf11841-akr	cuff
OsNG07-17	chr07	10111212	10113166	1954	1311	6.4	1.5	1.6	1.1	7.4	NP_001175132.1		cuff
OsNG07-18	chr07	11483768	11484539	771	173	1.7	2.4	1.3	0.9	0.0	EAZ11160.1		cuff
OsNG07-19	chr07	11842895	11843608	713	166	1.8	0.4	1.3	0.4	24.3	BAD73489.1		cuff
OsNG07-20	chr07	13192217	13193949	1732	437	1.9	0.2	1.6	0.4	18.5	BAC45169.1		both
OsNG07-21	chr07	14484312	14488252	3940	494	1.5	2.9	3.3	3.5	20.3	NP_001050054.1		both
OsNG07-22	chr07	15186617	15190344	3727	360	1.9	1.5	2.3	2.0	19.7	ABA97790.1		cuff
OsNG07-23	chr07	15285143	15285734	591	99	1.3	0.5	0.5	0.2	19.6	ZP_09788542.1		cuff
OsNG07-24	chr07	16034052	16037633	3581	456	1.9	1.6	2.0	2.0	17.8	EEC81974.1		cuff
OsNG07-25	chr07	16496999	16497409	410	59	1.1	0.5	0.7	0.4	17.1	BAC65969.1		cuff
OsNG07-26	chr07	16732275	16734686	2411	573	5.4	0.9	1.5	1.1	15.6	ABF94693.1		cuff
OsNG07-27	chr07	16733125	16734830	1705	546	2.4	0.8	1.4	1.0	15.9	ABF94693.1		cuff
OsNG07-28	chr07	17173127	17173664	537	72	1.0	0.3	3.0	1.8	0.0	BAC65929.1		cuff
OsNG07-29	chr07	18220212	18220733	521	108	1.6	0.6	0.2	0.1	0.0	ZP_02443189.1		cuff
OsNG07-30	chr07	18316490	18319882	3392	772	9.2	2.2	4.9	3.4	0.0	EEC82066.1		cuff
OsNG07-31	chr07	19192367	19195983	3616	748	2.8	1.4	2.7	1.2	9.3	BAD21744.1		cuff
OsNG07-32	chr07	19737183	19737921	738	47	1.1	1.7	2.6	3.5	14.0	NP_001059773.2		cuff
OsNG07-33	chr07	20948745	20949334	589	79	1.3	8.3	3.2	6.0	0.0	XP_003562986.1		cuff
OsNG07-34	chr07	21507317	21508545	1228	153	1.8	0.6	0.3	0.3	4.9			cuff
OsNG07-35	chr07	21934194	21934655	461	170	2.8	0.7	0.3	0.3	8.5	BAC80032.1		cuff
OsNG07-36	chr07	22719472	22720449	977	243	1.9	0.5	0.7	0.8	7.0	BAC22552.1		both
OsNG07-37	chr07	23449506	23450135	629	172	2.1	0.4	0.5	0.5	17.6	BAC79813.1		cuff
OsNG07-38	chr07	25259255	25260295	1040	997	7.3	1.4	0.0	0.1	4.8	EEE67580.1	csi-MIR396c	both
OsNG07-39	chr07	26666917	26668970	2053	254	2.8	2.2	3.5	2.2	11.7	BAC83450.1	osa-MIRf10207-akr	cuff
OsNG07-40	chr07	26938048	26938450	402	58	1.1	1.4	1.4	0.8	0.0	BAC84341.1		cuff
OsNG07-41	chr07	28238057	28238552	495	95	1.5	0.1	0.2	0.1	24.2			cuff
OsNG07-42	chr07	28522107	28522880	773	114	1.1	0.0	0.0	0.0	0.0	ZP_10540278.1	osa-MIR164a	cuff
OsNG07-43	chr07	28703112	28703777	665	67	1.0	0.8	0.4	0.3	0.0	XP_001748382.1		cuff
OsNG07-44	chr07	28834843	28836977	2134	201	2.1	3.7	2.3	1.6	0.0	EEE67827.1		cuff
OsNG07-45	chr07	29411679	29412838	1159	759	5.0	1.2	4.0	3.3	21.1	BAC75550.1	osa-MIR1846c	both
OsNG08-01	chr08	45676	52781	7105	2062	3.7	2.3	1.9	1.6	1.8	EEE67888.1		cuff
OsNG08-02	chr08	2580999	2581331	332	66	1.5	0.9	2.0	2.2	0.0	XP_739778.1		cuff

OsNG08-03	chr08	3116836	3117328	492	157	2.4	0.9	2.5	2.9	15.0	YP_006641726.1		cuff
OsNG08-04	chr08	3317475	3317996	521	150	2.2	1.2	1.6	1.4	15.9	ZP_08238010.1		cuff
OsNG08-05	chr08	3831542	3834082	2540	826	9.0	0.6	1.7	1.3	7.2	NP_001176757.1	osa-MIRf10787-akr	both
OsNG08-06	chr08	3857131	3857337	206	31	1.1	1.4	0.8	0.5	0.0			cuff
OsNG08-07	chr08	4307614	4308809	1195	186	1.7	0.4	0.4	0.2	13.1	AAM93704.1		cuff
OsNG08-08	chr08	4379331	4382112	2781	730	5.0	0.8	0.3	0.3	19.3	BAD05223.1		both
OsNG08-09	chr08	5689142	5690681	1539	929	5.0	2.0	1.0	1.5	9.6	NP_001050543.1		cuff
OsNG08-10	chr08	6089336	6094115	4779	217	0.3	2.0	0.8	0.7	23.7		osa-MIRf11	novel
OsNG08-11	chr08	8171796	8175788	3992	840	1.6	2.3	3.6	2.8	13.8	EEC69769.1		both
OsNG08-12	chr08	8181350	8181834	484	109	1.7	0.2	5.0	4.0	9.3	EEE68279.1		cuff
OsNG08-13	chr08	8334688	8338708	4020	466	2.2	1.3	1.6	1.4	8.9	NP_001175128.1		both
OsNG08-14	chr08	8337916	8338906	990	135	1.2	0.9	0.7	0.5	22.3	NP_001175128.1		cuff
OsNG08-15	chr08	9486717	9490812	4095	266	1.2	1.1	2.4	2.3	15.5	BAC98637.1	sbi-MIR396c	both
OsNG08-16	chr08	10577338	10578361	1023	1370	10.2	0.7	0.2	0.3	18.6	NP_001175490.1		cuff
OsNG08-17	chr08	10817088	10818549	1461	171	16.9	0.6	0.7	0.8	15.7	NP_001061922.1		cuff
OsNG08-18	chr08	11373334	11375207	1873	186	1.1	4.4	0.0	0.7	13.2	EEE68400.1		cuff
OsNG08-19	chr08	11547023	11549721	2698	341	1.9	2.3	3.7	3.7	0.0	EEC83282.1		cuff
OsNG08-20	chr08	12368954	12372654	3700	10469	38.2	2.3	0.8	1.3	4.1	BAD30716.1	osa-MIR809d	cuff
OsNG08-21	chr08	13907314	13908210	896	211	1.8	0.5	1.5	0.6	19.1	EEC67871.1		both
OsNG08-22	chr08	13956438	13957266	828	116	1.1	1.0	3.9	3.1	24.6	EAZ09606.1		cuff
OsNG08-23	chr08	14281124	14284638	3514	588	2.3	1.1	0.9	0.4	17.2	EEE68479.1		cuff
OsNG08-24	chr08	15095334	15099150	3816	185	1.3	16.1	10.0	16.1	4.9	EEE68513.1		cuff
OsNG08-25	chr08	16071573	16074408	2835	124	1.1	999.0	999.0	999.0	11.2	ZP_17502358.1	tae-MIR160	cuff
OsNG08-26	chr08	16870292	16870839	547	147	2.0	0.4	0.6	0.3	0.0	NP_001175538.1		cuff
OsNG08-27	chr08	18458034	18459582	1548	147	2.3	0.9	2.7	1.6	17.6	BAD05344.1		cuff
OsNG08-28	chr08	18612285	18614211	1926	282	1.1	1.0	0.4	1.3	15.9		osa-MIRf10	novel
OsNG08-29	chr08	18688765	18689808	1043	626	4.6	1.2	1.3	2.4	8.2	BAD05269.1		both
OsNG08-30	chr08	20878148	20880226	2078	142	1.4	0.9	0.9	0.8	0.0	NP_001175593.1		cuff
OsNG08-31	chr08	21877660	21878702	1042	183	1.3	0.4	4.8	0.7	8.5	EEE68778.1		both
OsNG08-32	chr08	21967600	21970288	2688	528	1.5	0.9	0.5	0.7	8.4	NP_001050553.1	osa-MIRf11360-akr	cuff
OsNG08-33	chr08	21967675	21970500	2825	534	2.7	0.9	0.5	0.8	8.0	NP_001050553.1	osa-MIRf11360-akr	cuff
OsNG08-34	chr08	22621698	22624994	3296	1529	5.3	1.4	1.7	1.2	10.4	NP_001061978.1		cuff
OsNG08-35	chr08	23231337	23231588	251	142	4.3	3.7	0.5	1.5	19.1			cuff
OsNG08-36	chr08	24793039	24793674	635	143	1.7	0.5	28.8	3.0	5.5		osa-MIR164f	cuff
OsNG08-37	chr08	25050801	25053578	2777	96	1.6	0.9	1.0	1.4	5.8	NP_001175650.1	osa-MIRf10757-akr	cuff
OsNG08-38	chr08	25390923	25391681	758	446	4.5	0.1	0.5	0.1	0.0	ZP_03972808.1		cuff
OsNG08-39	chr08	25660248	25660620	372	100	2.9	2.2	2.7	2.7	16.4	BAD09972.1	sbi-MIR396c	cuff
OsNG08-40	chr08	26015026	26016336	1310	263	1.5	0.5	2.0	1.5	16.9	EEE69018.1	osa-MIRf11208-akr	both
OsNG08-41	chr08	26186880	26189333	2453	730	4.3	1.0	1.7	1.9	0.0	BAD08744.1		both
OsNG08-42	chr08	26753627	26754144	517	104	1.5	0.4	0.2	0.1	0.0	BAD01223.1		cuff
OsNG08-43	chr08	26884705	26884956	251	136	4.1	0.4	11.9	12.0	0.0			cuff
OsNG08-44	chr08	26926760	26927821	1061	239	1.9	1.9	1.0	1.4	0.0	AFW64978.1		cuff
OsNG08-45	chr08	27042943	27043294	351	53	1.1	4.5	4.9	2.1	0.0	YP_003823916.1		cuff
OsNG08-46	chr08	27060206	27061016	810	175	1.6	0.4	0.5	0.3	0.0	JAA57892.1		cuff
OsNG08-47	chr08	27177441	27178152	711	333	3.6	0.7	0.3	0.3	0.0	CBJ49199.1		cuff
OsNG08-48	chr08	27549744	27551431	1687	312	1.4	5.5	0.4	12.5	19.2	EAY84875.1		cuff
OsNG08-49	chr08	27811772	27813417	1645	67	1.2	999.0	999.0	999.0	3.1	BAD10394.1		cuff
OsNG09-01	chr09	2668887	2670985	2098	308	1.1	2.0	1.1	2.5	13.4	NP_001066945.1	osa-MIRf10317-akr	cuff
OsNG09-02	chr09	2922275	2923948	1673	147	1.0	2.2	31.9	10.4	19.4	BAD33904.1		cuff
OsNG09-03	chr09	4048740	4051582	2842	2124	14.6	1.0	0.8	0.7	3.9			cuff
OsNG09-04	chr09	5432410	5433174	764	9437	93.9	0.7	2.2	1.5	0.0			novel
OsNG09-05	chr09	7300166	7300796	630	1494	18.0	0.5	0.5	0.2	0.0	XP_001383927.2		cuff
OsNG09-06	chr09	10395344	10396022	678	156	2.1	0.3	0.2	0.2	21.1	BAC83426.1		cuff
OsNG09-07	chr09	10804037	10807261	3224	560	1.3	1.2	0.2	0.1	2.4	BAD29107.1		cuff

OsNG09-08	chr09	11927402	11929155	1753	213	1.6	2.3	2.4	1.3	10.4		osa-MIRf11893-akr	cuff
OsNG09-09	chr09	13591995	13592332	337	138	3.1	3.1	0.7	0.7	10.1	EEE69639.1		cuff
OsNG09-10	chr09	13903746	13906664	2918	138	1.5	1.1	1.2	1.1	6.5	BAD28585.1		cuff
OsNG09-11	chr09	14549958	14553321	3363	162	1.3	7.1	4.2	4.5	0.0	EEE69688.1		cuff
OsNG09-12	chr09	15710994	15711692	698	92	1.0	1.3	0.1	0.2	0.0	BAD33798.1		cuff
OsNG09-13	chr09	16059740	16070249	10509	3697	22.0	2.0	1.8	1.1	21.8	NP_001176180.1	osa-MIR2907b	cuff
OsNG09-14	chr09	16863672	16864542	870	186	1.6	0.4	0.9	0.3	0.0	BAD38200.1		cuff
OsNG09-15	chr09	17268163	17268956	793	274	2.6	1.8	2.1	8.5	5.5			both
OsNG09-16	chr09	17791538	17793466	1928	522	3.7	0.3	1.6	0.5	2.6			cuff
OsNG09-17	chr09	17959259	17959872	613	55	1.2	0.9	2.8	1.7	0.0			cuff
OsNG09-18	chr09	18422969	18423624	655	170	2.0	0.6	1.1	0.4	0.0	NP_001063484.1		cuff
OsNG09-19	chr09	18585504	18586396	892	362	3.1	0.2	0.3	0.1	3.3	DAA62005.1		both
OsNG09-20	chr09	18878325	18881334	3009	1363	4.3	2.3	0.3	0.2	7.7	BAD82486.1	osa-MIRf10214-akr	cuff
OsNG09-21	chr09	20332566	20335790	3224	378	1.4	2.6	2.9	3.6	13.5	EEC77149.1		cuff
OsNG09-22	chr09	20580043	20581240	1197	243	1.5	0.3	6.7	0.4	22.1	ZP_08605532.1		cuff
OsNG09-23	chr09	21140997	21141622	625	104	1.3	0.2	3.4	1.6	24.2	BAD34231.1		cuff
OsNG09-24	chr09	22017551	22018359	808	191	1.8	0.7	1.0	0.7	0.0	XP_001821267.1		cuff
OsNG09-25	chr09	22017558	22021971	4413	814	3.3	1.4	5.0	4.1	11.4	EEE70198.1		cuff
OsNG09-26	chr09	22521584	22525689	4105	863	3.9	2.2	1.0	1.4	0.0	EEE70249.1		cuff
OsNG10-01	chr10	1909410	1910683	1273	216	1.3	0.4	0.0	0.0	22.7		osa-MIRf10	novel
OsNG10-02	chr10	3067477	3070085	2608	162	1.4	2.3	3.7	3.2	1.9			cuff
OsNG10-03	chr10	3398048	3400068	2020	152	1.1	1.3	1.2	4.0	0.0	EEE53954.1		cuff
OsNG10-04	chr10	3403565	3404676	1111	31	1.4	999.0	999.0	999.0	0.0	EEE50585.1		cuff
OsNG10-05	chr10	4352992	4353304	312	56	1.4	999.0	999.0	999.0	0.0	ZP_18298819.1		cuff
OsNG10-06	chr10	4960885	4961537	652	168	2.0	0.5	0.6	0.4	9.8	AAM08840.1	osa-MIR441c	cuff
OsNG10-07	chr10	5269076	5271561	2485	680	2.3	1.2	1.1	0.8	19.9	ABF97069.1		both
OsNG10-08	chr10	5389049	5393102	4053	942	5.3	0.8	0.9	0.7	14.3	EAZ07035.1		both
OsNG10-09	chr10	6230132	6233465	3333	1063	7.0	3.0	11.7	9.7	8.2	EEE62189.1	osa-MIRf10581-akr	cuff
OsNG10-10	chr10	6760372	6761495	1123	458	3.4	0.8	1.5	1.7	18.5	AAX96757.1		both
OsNG10-11	chr10	8589303	8593036	3733	851	5.0	3.4	3.7	6.3	15.4	AAM00951.1		both
OsNG10-12	chr10	9880750	9880978	228	38	1.3	0.9	3.9	1.3	0.0			cuff
OsNG10-13	chr10	9897999	9898382	383	114	2.3	0.4	3.3	0.9	21.4	P0CAY4.1		cuff
OsNG10-14	chr10	9898744	9898886	142	44	2.4	0.3	4.6	0.9	0.0			cuff
OsNG10-15	chr10	11508228	11509623	1395	496	3.0	0.7	3.2	1.1	0.0	XP_003940119.1		both
OsNG10-16	chr10	11884010	11884943	933	354	2.9	0.1	2.4	0.2	8.4	XP_003223739.1		both
OsNG10-17	chr10	12694001	12696636	2635	838	2.6	1.5	0.9	0.5	9.2	AAK13094.1	sbi-MIR396c	both
OsNG10-18	chr10	12727676	12732766	5090	684	5.5	0.7	0.7	0.4	2.7	AAK13098.1		cuff
OsNG10-19	chr10	12811074	12814516	3442	439	7.6	59.6	42.1	34.1	3.7	EEE60597.1	osa-MIRf11418-akr	cuff
OsNG10-20	chr10	13232456	13232742	286	105	2.8	2.3	1.5	2.3	10.5	XP_003899869.1		cuff
OsNG10-21	chr10	13947141	13947997	856	2465	21.9	1.0	0.1	0.3	7.1	XP_003976126.1		both
OsNG10-22	chr10	14486434	14487155	721	191	2.0	0.9	0.7	0.3	0.0	YP_003003030.1		cuff
OsNG10-23	chr10	14724359	14725639	1280	206	1.4	0.5	0.2	0.1	23.5	EAY78517.1		cuff
OsNG10-24	chr10	14782658	14784966	2308	169	1.1	1.7	2.5	1.7	24.5	ABB47665.1	osa-MIRf10634-akr	cuff
OsNG10-25	chr10	15397118	15400004	2886	257	1.5	2.3	3.3	2.0	6.3	ABF94503.1		cuff
OsNG10-26	chr10	15811523	15812357	834	124	1.1	0.6	2.1	0.7	18.9	NP_001065536.1	osa-MIRf10450-akr	cuff
OsNG10-27	chr10	16328931	16329934	1003	557	4.2	1.3	0.3	0.9	15.9		osa-MIRf11	novel
OsNG10-28	chr10	17614500	17616341	1841	228	0.9	0.5	0.3	0.1	12.6			novel
OsNG10-29	chr10	18459017	18459522	505	223	3.4	2.5	2.7	1.8	6.1	AAL31066.1	ptc-MIRf12172-akr	cuff
OsNG10-30	chr10	18475387	18475571	184	25	1.0	2.1	0.5	0.5	0.0			cuff
OsNG10-31	chr10	18500115	18501638	1523	212	1.1	0.9	0.8	0.3	7.7	AAR87368.1	csi-MIR396c	cuff
OsNG10-32	chr10	18966724	18967349	625	92	1.1	0.7	0.3	0.3	21.9	ABB47486.1		cuff
OsNG10-33	chr10	19024575	19025097	522	784	11.4	0.3	6.0	1.5	0.0			novel
OsNG10-34	chr10	19489322	19493964	4642	571	1.9	1.2	2.1	1.8	16.2	EEC67289.1		both

OsNG10-35	chr10	19521497	19522741	1244	208	1.3	0.5	23.7	2.4	0.0	EEE51253.1		cuff
OsNG10-36	chr10	20214542	20218992	4450	2618	14.6	1.5	0.8	1.7	6.9	AAM92286.1	osa-MIRf11991-akr	cuff
OsNG10-37	chr10	21017447	21017853	406	66	1.2	1.7	4.0	3.5	23.6	AAG46112.1		cuff
OsNG10-38	chr10	21205848	21217000	11152	429	4.2	0.8	1.0	0.9	24.5	CAE04633.3	osa-MIRf11817-akr	cuff
OsNG10-39	chr10	21510653	21512916	2263	698	4.1	1.2	0.7	0.2	4.5	NP_001065316.1		cuff
OsNG10-40	chr10	22249181	22252443	3262	5497	25.7	1.2	1.1	0.8	0.0	AAG60197.1		cuff
OsNG10-41	chr10	22292622	22292833	211	37	1.3	6.6	2.8	24.0	0.0	EEE51415.1		cuff
OsNG10-42	chr10	22625858	22627277	1419	266	1.4	3.1	2.1	1.9	4.3	XP_954120.1		cuff
OsNG10-43	chr10	22865699	22867356	1657	1671	15.2	0.6	1.0	0.7	0.0	NP_001176288.1		cuff
OsNG10-44	chr10	22882674	22883723	1049	200	1.4	0.5	0.5	0.5	0.0	BAD28066.1		both
OsNG10-45	chr10	22952254	22953539	1285	258	1.5	0.3	1.2	0.5	11.6	AAL58187.1		cuff
OsNG11-01	chr11	164958	168318	3360	1379	3.1	3.1	2.1	2.6	22.0	NP_001176299.1		cuff
OsNG11-02	chr11	704928	707921	2993	1031	6.3	1.0	2.4	2.4	19.4	EEC82439.1		both
OsNG11-03	chr11	1944994	1945795	801	1909	18.1	0.4	0.2	0.1	23.0	NP_001176346.1	osa-MIRf11608-akr	cuff
OsNG11-04	chr11	2126423	2127294	871	115	1.1	0.7	3.0	1.3	0.0	EAY79929.1		cuff
OsNG11-05	chr11	2162852	2163692	840	149	1.3	0.1	0.0	0.0	15.4	EGB10776.1		cuff
OsNG11-06	chr11	2453668	2456731	3063	243	2.1	2.7	4.4	3.8	5.6	EEE69766.1		cuff
OsNG11-07	chr11	3140545	3141473	928	230	2.2	0.4	0.1	0.0	14.0	EEC67737.1		both
OsNG11-08	chr11	4588311	4589231	920	593	4.9	0.3	0.1	0.1	20.2	NP_001067413.2		both
OsNG11-09	chr11	5232024	5236435	4411	692	1.2	1.1	1.1	0.5	14.0	XP_003572955.1	osa-MIRf11737-akr	cuff
OsNG11-10	chr11	6076147	6079485	3338	635	1.4	999.0	999.0	999.0	7.3	XP_003946460.1	osa-MIR169o	cuff
OsNG11-11	chr11	6079715	6081286	1571	224	1.1	0.6	39.2	0.9	16.2		osa-MIR169n	cuff
OsNG11-12	chr11	7117289	7118437	1148	168	4.8	0.8	2.8	1.7	17.0	NP_001172842.1		cuff
OsNG11-13	chr11	7483380	7488749	5369	637	4.2	1.5	5.2	4.0	7.9	EAY86160.1	oru-MIR806	cuff
OsNG11-14	chr11	7802126	7809056	6930	507	2.9	0.8	0.5	0.4	8.8	CAE01710.1	osa-MIR2118q	cuff
OsNG11-15	chr11	8356854	8358382	1528	178	1.8	0.7	0.2	0.1	18.5	AAX94805.1		cuff
OsNG11-16	chr11	10926043	10926495	452	90	1.5	0.3	0.6	1.0	15.5	EAY90715.1		cuff
OsNG11-17	chr11	11365112	11366068	956	184	1.5	1.6	1.0	1.3	4.7	AAX94920.1		cuff
OsNG11-18	chr11	11706995	11710611	3616	535	2.2	2.4	5.4	4.0	4.1	NP_001176494.1		cuff
OsNG11-19	chr11	12366249	12368526	2277	339	1.1	0.7	1.1	0.8	14.6	BAD61974.1		cuff
OsNG11-20	chr11	13167666	13168448	782	257	2.5	0.5	1.2	0.9	23.3	AAX96639.1		cuff
OsNG11-21	chr11	13288797	13293797	5000	1329	9.1	0.6	0.9	0.6	7.9	XP_003558751.1		cuff
OsNG11-22	chr11	13663664	13665984	2320	329	2.8	1.3	3.0	1.9	10.7	EAY80802.1		cuff
OsNG11-23	chr11	13857663	13859599	1936	114	1.4	1.1	2.5	1.3	17.8	CAE05785.2	sbi-MIR396c	cuff
OsNG11-24	chr11	14771235	14771979	744	161	1.6	3.8	2.9	6.0	22.6	EEE62972.1		cuff
OsNG11-25	chr11	14880680	14883189	2509	1030	3.4	12.9	9.4	20.2	18.0	AAX96127.1		both
OsNG11-26	chr11	14885594	14886367	773	313	3.1	1.1	1.0	2.7	8.4	XP_678751.1	ptc-MIRf11913-akr	cuff
OsNG11-27	chr11	14991691	14992030	339	60	1.3	1.7	7.2	4.1	0.0	YP_004906066.1		cuff
OsNG11-28	chr11	15735748	15736688	940	752	6.1	4.0	0.8	1.2	16.4			cuff
OsNG11-29	chr11	16456613	16456773	160	22	1.0	0.6	3.4	1.8	0.0			cuff
OsNG11-30	chr11	17689965	17691452	1487	265	2.0	1.1	1.3	1.0	3.9			cuff
OsNG11-31	chr11	20788778	20789454	676	89	1.5	0.8	0.4	0.2	4.6			cuff
OsNG11-32	chr11	21753142	21753593	451	72	1.2	0.2	1.6	1.2	0.0		osa-MIRf11505-akr	cuff
OsNG11-33	chr11	22426676	22428657	1981	91	1.0	1.1	2.6	2.5	2.2	NP_001068218.1		cuff
OsNG11-34	chr11	22721884	22722632	748	256	2.6	0.8	0.5	0.1	4.4	ZP_05102604.1		cuff
OsNG11-35	chr11	22948963	22949389	426	58	1.0	2.1	0.8	0.8	17.6	CCH61489.1		cuff
OsNG11-36	chr11	24136733	24138405	1672	179	2.6	0.8	0.9	0.1	5.9	EEC68506.1	osa-MIRf11955-akr	cuff
OsNG11-37	chr11	24256949	24258262	1313	313	6.1	0.3	3.7	2.8	12.4	EEE52412.1		both
OsNG11-38	chr11	24461283	24461966	683	162	1.8	0.4	0.5	0.3	0.0	XP_001916877.2		cuff
OsNG11-39	chr11	24813642	24814426	784	211	2.0	2.6	0.3	2.0	15.7	EEE52432.1		cuff
OsNG11-40	chr11	25247920	25252603	4683	992	4.0	21.5	1.7	67.9	7.2	AAX96183.1		cuff
OsNG11-41	chr11	26062171	26065575	3404	1236	6.1	1.9	4.4	4.2	15.7	NP_001050054.1		both

OsNG11-42	chr11	27259888	27262587	2699	408	1.1	2.0	2.1	1.5	17.6	EAY87140.1	osa-MIRf10942-akr	cuff
OsNG11-43	chr11	28306228	28312417	6189	263	1.6	1.8	2.3	2.0	10.6	NP_001176737.1		cuff
OsNG12-01	chr12	2035188	2039632	4444	154	1.1	1.6	2.2	2.1	9.8	AAL78105.1	osa-MIR5160	cuff
OsNG12-02	chr12	2283813	2285332	1519	117	1.3	1.8	1.5	1.2	9.5		oru-MIR806	cuff
OsNG12-03	chr12	3638265	3638401	136	35	2.0	0.0	7.9	7.4	0.0			cuff
OsNG12-04	chr12	3781048	3782094	1046	415	3.0	0.5	0.5	0.3	10.6	EGG17833.1		cuff
OsNG12-05	chr12	4657303	4657643	340	53	1.2	0.4	0.4	0.3	16.2	EEE52896.1		cuff
OsNG12-06	chr12	5101590	5102508	918	255	2.1	0.7	0.8	0.4	19.0	EEE52959.1		cuff
OsNG12-07	chr12	5164440	5165651	1211	160	1.0	2.6	1.3	0.5	5.1	CAH66018.1	osa-MIRf10710-akr	both
OsNG12-08	chr12	5164616	5167635	3019	377	3.8	1.6	1.3	0.7	3.8	CAH66018.1	osa-MIRf10710-akr	cuff
OsNG12-09	chr12	5166458	5167649	1191	219	2.4	1.0	1.1	0.8	7.1	XP_003581613.1		cuff
OsNG12-10	chr12	6167509	6168456	947	168	1.6	0.9	10.0	5.4	13.0	EIW67319.1		cuff
OsNG12-11	chr12	6488284	6489504	1220	223	1.4	0.3	0.1	0.1	17.8	NP_001066413.1		cuff
OsNG12-12	chr12	6491569	6492428	859	51345	454.2	0.4	0.1	0.1	16.2	EAY82618.1		both
OsNG12-13	chr12	7116488	7117553	1065	188	1.3	1.2	0.6	1.1	22.8	EEE63434.1		cuff
OsNG12-14	chr12	7554325	7555047	722	371	3.9	0.8	0.2	0.2	17.6		osa-MIRf12045-akr	both
OsNG12-15	chr12	8142747	8144051	1304	162	1.2	1.5	0.3	0.5	22.5	AAK84447.1	tae-MIR160	cuff
OsNG12-16	chr12	8540499	8541223	724	151	1.6	0.7	3.7	5.4	23.2	XP_003588296.1		cuff
OsNG12-17	chr12	8692697	8696771	4074	910	7.7	1.0	1.4	0.9	17.4	EEE68158.1		both
OsNG12-18	chr12	10801685	10803437	1752	750	3.4	14.2	1.1	17.1	7.5	NP_001057587.1		both
OsNG12-19	chr12	12119477	12122068	2591	156	2.0	0.8	5.1	1.7	13.1	NP_001050205.1	osa-MIRf11955-akr	cuff
OsNG12-20	chr12	12459744	12463082	3338	380	1.9	0.8	1.0	0.6	13.1	AAU90278.1		both
OsNG12-21	chr12	13049017	13049193	176	900	38.9	0.6	1.1	2.4	20.5	XP_003588355.1		cuff
OsNG12-22	chr12	13349872	13353258	3386	1051	2.4	0.8	1.5	2.5	8.4			novel
OsNG12-23	chr12	14905236	14907394	2158	131	1.1	1.4	2.1	2.4	0.0	NP_001176935.1		cuff
OsNG12-24	chr12	15843452	15845775	2323	150	2.1	2.2	3.5	3.4	0.0			cuff
OsNG12-25	chr12	18536740	18537708	968	300	2.4	0.5	1.2	0.7	14.4	NP_001173345.1	osa-MIRf11389-akr	cuff
OsNG12-26	chr12	20810472	20813379	2907	1848	20.9	1.1	0.5	0.7	18.4	AAT44289.1		cuff
OsNG12-27	chr12	20869358	20870637	1279	2735	16.2	25.1	0.7	31.3	7.2	EAY83357.1		both
OsNG12-28	chr12	21560495	21563550	3055	412	3.3	1.0	2.4	1.9	20.6	BAD03716.1		cuff
OsNG12-29	chr12	21988550	21997608	9058	490	2.5	1.4	1.4	1.1	18.9	CAE76056.1	mtr-MIR156e	cuff
OsNG12-30	chr12	22206065	22207400	1335	755	4.3	2.2	1.0	1.0	11.3	EEC69357.1	ptc-MIRf12172-akr	both
OsNG12-31	chr12	23231765	23231990	225	31	1.0	1.9	5.3	1.5	0.0			cuff
OsNG12-32	chr12	24473703	24475279	1576	210	1.0	1.8	2.9	2.9	3.3	XP_001882482.1		cuff

Supplemental Table S3

Locus	**Name**	**Control**	**ABA Treated**	**ABA Fold Change**	**Log$_2$ Fold Change**
LOC_Os11g26750	rab16d	6	740	123.34	6.95
LOC_Os11g26760	rab16c	40	4,309	107.73	6.75
LOC_Os11g26780	rab16b	176	35,966	204.35	7.67
LOC_Os11g26790	rab16a	634	40,447	63.80	6.00
LOC_Os11g30500	TB2/DPI HVA22	0	79	Infinite	Infinite
LOC_Os08g36440	Similar to HVA22	37	5,959	161.05	7.33
LOC_Os05g46480	LEA	911	97,294	106.80	6.74

LOC_Os01g50910	WS118	269	27,557	102.44	6.68
LOC_Os05g28210	EMP1	100	8,192	81.92	6.36

Supplemental Table S4

Locus	**Name**	**Control**	**GA Treated**	**GA Fold Change**	**Log$_2$ Fold Change**
LOC_Os02g52710	RAmy1A	39,173	129,526	3.31	1.73
LOC_Os01g25510	RAmy1B	2,158	8,262	3.83	1.94
LOC_Os02g52700	RAmy1C	2,626	43,959	16.74	4.07
LOC_Os06g49970	RAmy2A	8	102	12.74	3.67
LOC_Os09g28400	RAmy3A	8	7	0.94	-0.10
LOC_Os09g28420	RAmy3B	53,385	114,020	2.14	1.09
LOC_Os08g36910	RAmy3D	369	757	2.05	1.04
LOC_Os08g36900	RAmy3E	787	14,009	17.8	4.15

Supplemental Table S5

Novel Gene	**Coordinates**	**Forward Primer**	**Reverse Primer**	**sequenced size (bp)**	**# of introns**
OsNG01-22	chr01:14990747-14991289	cctggaaaaacccagcaagattgctac	gcaggcgaagcaagacgccgtac	456	0
OsNG03-04	chr03:727244-728844	ctcctcagattaggataaaaaacatcagttcctctc	gagtctgttcaagacgaagatgaagtgc	259	1
OsNG03-42	chr03:22339382-22344060	cctaccgtgtagcaaccatatatcagcaactc	gagctctgatacggtgtgcccatg	727	6
OsNG04-40	chr04:29221922-29223124	ggtcgacatccaagtcactccaaacc	gatccgtcatatcatgatccagcaattc	637	0
OsNG06-24	chr06:19729629-19734412	ccaccttcttatacggctgttggag	gtctcagaacatttggagcaatggaataagac	367	3
OsNG09-04	chr09:5432410-5433174	gctcttgatgtaaggtgtttcattggtgcatatcagtgtag	gctctgcattctgaagcatcaagagaggagag	372	1
OsNG09-05	chr09:7300166-7300796	cacgttgcaaataagagatggag	gatagtataacagcacaaactgctactacatatag	546	0
OsNG10-40	chr10:22249181-22252443	ccatgtattggcttgtacaaggttgccag	gcaacaatacatgcaacatgttagctctatttatgaaatg	808	3
OsNG10-43	chr10:22865699-22867356	ctgagctttaagcctagctgaacatgctg	gcatcgcgtgattccgatgag	537	2
OsNG12-12	chr12:6491569-6492428	ctccactgttcacttcacaatataagaacgag	gaccgtaggccaagattaacctcatcc	621	0

Supplemental Table S6

Locus_id	Chromosome	Start Position	End Position	RPKE	% Similarity to Another Genomic Region	% Overlap with MSU Gene	Footprint Excluding Introns	# of exons	TA/Cuff
TAOs01g00028	chr01	185093	185840	795.4545	15.24064	0	748	1	TA
TAOs01g00035	chr01	224894	225664	1044.099	7.522698	0	771	1	TA
TAOs01g00101	chr01	718306	719199	140.9836	9.060403	0	305	2	TA
TAOs01g00300	chr01	2046117	2046744	154.4586	21.49682	0	628	1	TA
TAOs01g00352	chr01	2415958	2418971	160	15.8925	0	1000	2	TA
TAOs01g00356	chr01	2459475	2462454	765.6702	16.04027	0	1037	2	TA
TAOs01g00360	chr01	2481179	2483190	159.6244	11.0835	0	1278	2	TA
TAOs01g00378	chr01	2599362	2599727	915.3005	0	0	366	1	TA
TAOs01g00480	chr01	3238049	3238878	215.6627	24.6988	0	830	1	TA
TAOs01g00578	chr01	3856425	3857472	110.687	9.732824	0	1048	1	TA
TAOs01g00604	chr01	4071299	4071635	121.6617	0	0	337	1	TA
TAOs01g00651	chr01	4438152	4439091	115.9574	10	0	940	1	TA
TAOs01g00722	chr01	5015280	5015539	138.4615	0	0	260	1	TA
TAOs01g00804	chr01	5643861	5645834	418.2825	7.294833	0	722	3	TA
TAOs01g00920	chr01	6655919	6656782	280.0926	0	0	864	1	TA
TAOs01g00937	chr01	6770412	6770690	100.3584	0	0	279	1	TA
TAOs01g00974	chr01	7098295	7098676	138.7435	0	0	382	1	TA
TAOs01g01031	chr01	7472815	7476977	1039.047	3.074706	0	3022	4	TA
TAOs01g01033	chr01	7485930	7486632	123.9936	0	0	621	2	TA
TAOs01g01037	chr01	7491375	7492264	119.3112	3.483146	0	813	2	TA
TAOs01g01053	chr01	7593514	7595372	146.7577	16.78322	0	586	3	TA
TAOs01g01081	chr01	7930359	7933251	105.5456	10.64639	0	2795	2	TA
TAOs01g01170	chr01	8722663	8724800	1161.833	13.51731	0	2138	1	TA
TAOs01g01218	chr01	9047135	9049990	123.4076	23.87955	0	1256	4	TA
TAOs01g01267	chr01	9464875	9475174	844.3278	0.495146	0	4002	12	TA
TAOs01g01268	chr01	9478270	9479155	406.3205	6.884876	0	886	1	TA
TAOs01g01284	chr01	9644134	9649493	232.6964	0	0	2355	7	TA
TAOs01g01285	chr01	9649516	9652397	174.5098	0	0	1530	4	TA
TAOs01g01298	chr01	9732721	9733630	227.4725	24.17582	0	910	1	TA
TAOs01g01356	chr01	10270974	10274417	1537.589	1.88734	0	2820	7	TA
TAOs01g01375	chr01	10398573	10399220	152.7778	12.96296	0	648	1	TA
TAOs01g01442	chr01	11191710	11192439	789.0411	8.356164	0	730	1	TA
TAOs01g01472	chr01	11495925	11496381	175.0547	14.22319	0	457	1	TA
TAOs01g01478	chr01	11546899	11547092	103.0928	0	0	194	1	TA
TAOs01g01522	chr01	11987900	11989033	201.94	23.45679	0	1134	1	TA
TAOs01g01534	chr01	12080126	12081310	124.0506	10.21097	0	1185	1	TA
TAOs01g01636	chr01	13198379	13199696	255.3366	13.20182	0	1218	2	TA
TAOs01g01782	chr01	14802782	14804514	294.8718	8.020773	0	1248	2	TA
TAOs01g01836	chr01	15547413	15548658	161.3162	22.47191	0	1246	1	TA
TAOs01g02012	chr01	17275537	17278327	284.4858	15.04837	0	2791	1	TA
TAOs01g02017	chr01	17306292	17307707	106.087	6.567797	0	575	2	TA
TAOs01g02082	chr01	17959741	17960394	239.6396	17.737	0	555	2	TA
TAOs01g02152	chr01	18668026	18668927	435.6984	22.949	0	902	1	TA
TAOs01g02163	chr01	18839068	18839623	487.4101	9.892086	0	556	1	TA
TAOs01g02177	chr01	19001347	19002405	125.1682	13.50331	0	743	2	TA
TAOs01g02187	chr01	19106556	19107601	469.2388	10.61185	0	959	2	TA
TAOs01g02297	chr01	20316710	20317555	117.0213	3.309693	0	658	2	TA
TAOs01g02325	chr01	20502204	20502771	542.2535	0	0	568	1	TA
TAOs01g02487	chr01	22291730	22292346	145.8671	0	0	617	1	TA
TAOs01g02691	chr01	23934267	23937218	754.8161	7.757453	0	1142	6	TA
TAOs01g02702	chr01	24125098	24125645	129.562	24.63504	0	548	1	TA
TAOs01g02713	chr01	24187243	24190017	10842.87	5.981982	0	1241	8	TA
TAOs01g02733	chr01	24390678	24391812	6724.286	15.3304	0	700	2	TA
TAOs01g02798	chr01	25004052	25004744	6770.563	14.43001	0	693	1	TA
TAOs01g02800	chr01	25019183	25021624	155.2007	18.01802	0	2442	1	TA
TAOs01g02856	chr01	25391397	25412122	296.2227	6.735501	0	503	5	TA
TAOs01g02919	chr01	26051760	26053742	265.2757	20.87746	0	671	3	TA

TAOs01g02979	chr01	26777274	26777445	162.7907	0	0	172	1	TA
TAOs01g03111	chr01	27933641	27935127	1401.481	0	0	1350	2	TA
TAOs01g03113	chr01	27950292	27951455	837.6289	20.189	0	1164	1	TA
TAOs01g03238	chr01	28823357	28824179	141.844	12.02916	0	705	2	TA
TAOs01g03309	chr01	29368731	29369496	100.5222	10.31332	0	766	1	TA
TAOs01g03315	chr01	29380749	29381400	113.4969	0	0	652	1	TA
TAOs01g03364	chr01	29755860	29756122	281.3688	0	0	263	1	TA
TAOs01g03598	chr01	31365897	31366704	111.3861	10.27228	0	808	1	TA
TAOs01g03610	chr01	31415137	31417039	461.3768	13.34735	0	1903	1	TA
TAOs01g03618	chr01	31491844	31493035	868.2886	6.795302	0	1192	1	TA
TAOs01g03755	chr01	32512491	32513069	240.0691	0	0	579	1	TA
TAOs01g03788	chr01	32696776	32698254	113.6045	0	0	713	2	TA
TAOs01g03798	chr01	32810953	32812792	244.6886	12.22826	0	1365	4	TA
TAOs01g03863	chr01	33194781	33195066	101.3986	16.08392	0	286	1	TA
TAOs01g03908	chr01	33521362	33522488	326.5306	14.46318	0	1127	1	TA
TAOs01g03948	chr01	33707382	33707984	126.0365	18.07629	0	603	1	TA
TAOs01g03955	chr01	33764549	33765308	309.2105	0	0	760	1	TA
TAOs01g04039	chr01	34256595	34257664	167.2897	0	0	1070	1	TA
TAOs01g04042	chr01	34262103	34262731	114.4674	10.96979	0	629	1	TA
TAOs01g04043	chr01	34262821	34266181	222.2553	24.45701	0	3361	1	TA
TAOs01g04066	chr01	34477841	34478281	104.3084	0	0	441	1	TA
TAOs01g04069	chr01	34495709	34498529	3587.847	17.68876	0	757	2	TA
TAOs01g04070	chr01	34501834	34503578	564.4699	18.45272	0	1745	1	TA
TAOs01g04100	chr01	34656890	34658024	257.2687	0	0	1135	1	TA
TAOs01g04108	chr01	34701906	34702766	148.6643	4.065041	0	861	1	TA
TAOs01g04137	chr01	35003824	35005656	480.6259	20.2946	0	1342	2	TA
TAOs01g04219	chr01	35745327	35749596	173.9023	7.236534	0	4054	3	TA
TAOs01g04226	chr01	35781693	35782296	197.0199	7.450331	0	604	1	TA
TAOs01g04247	chr01	35956603	35959771	766.1483	10.79205	0	1672	3	TA
TAOs01g04260	chr01	36015189	36016031	107.9478	16.37011	0	843	1	TA
TAOs01g04277	chr01	36093828	36094071	147.541	0	0	244	1	TA
TAOs01g04283	chr01	36133744	36134456	130.4348	5.329593	0	713	1	TA
TAOs01g04459	chr01	37543878	37551305	549.1439	3.284868	0	1577	10	TA
TAOs01g04529	chr01	38025199	38026185	127.6596	0	0	987	1	TA
TAOs01g04588	chr01	38450620	38450971	102.2727	0	0	352	1	TA
TAOs01g04631	chr01	38840244	38842424	813.8651	20.58689	0	1053	4	TA
TAOs01g04670	chr01	39037355	39039880	274.91	24.03009	0	833	2	TA
TAOs01g04733	chr01	39516707	39518032	377.0739	23.9819	0	1326	1	TA
TAOs01g04742	chr01	39592443	39594211	111.0057	2.148106	0	1054	8	TA
TAOs01g04855	chr01	40489813	40490035	139.0135	18.83408	0	223	1	TA
TAOs01g04899	chr01	40912474	40914190	434.9882	20.9668	0	846	3	TA
TAOs01g04922	chr01	41071287	41073402	1697.624	1.795841	0	926	7	TA
TAOs01g04932	chr01	41110392	41111128	289.0095	0	0	737	1	TA
TAOs01g05111	chr01	42361293	42363671	2907.921	0	0	1010	4	TA
TAOs01g05216	chr01	43044783	43046259	146.1794	10.56195	0	602	4	TA
TAOs02g00055	chr02	420688	421908	105.6511	0	0	1221	1	TA
TAOs02g00422	chr02	2648642	2649070	132.8671	0	0	429	1	TA
TAOs02g00507	chr02	3091712	3092543	238.7792	3.365385	0	557	2	TA
TAOs02g00546	chr02	3271374	3272036	974.359	0	0	663	1	TA
TAOs02g00666	chr02	4208582	4208975	162.4365	15.22843	0	394	1	TA
TAOs02g00696	chr02	4514486	4514657	139.5349	0	0	172	1	TA
TAOs02g00697	chr02	4520037	4526949	265.3509	3.847823	0	912	9	TA
TAOs02g00784	chr02	5231489	5232359	477.6119	23.42135	0	871	1	TA
TAOs02g00808	chr02	5451147	5451830	111.1111	0	1.754386	684	1	TA
TAOs02g00880	chr02	5938321	5939467	671.3165	16.04185	0	1147	1	TA
TAOs02g00924	chr02	6241842	6242800	129.562	16.99687	0	548	2	TA
TAOs02g00963	chr02	6534481	6538151	330.4277	14.30128	0	3671	1	TA
TAOs02g00980	chr02	6710968	6712030	126.3048	0	0	958	2	TA
TAOs02g00981	chr02	6713337	6713598	122.1374	0	0	262	1	TA
TAOs02g00987	chr02	6778996	6779349	115.8192	0	0	354	1	TA
TAOs02g00993	chr02	6806805	6809992	363.9847	0	0	261	2	TA
TAOs02g01009	chr02	6994116	6995905	553.4665	16.03352	0	851	2	TA
TAOs02g01029	chr02	7236935	7238266	303.3708	0	0	623	2	TA
TAOs02g01092	chr02	7934659	7935464	380.8933	24.31762	0	806	1	TA
TAOs02g01141	chr02	8383191	8384333	460.1925	11.54856	0	1143	1	TA
TAOs02g01154	chr02	8486071	8489550	336.1345	17.35632	0	595	5	TA

TAOs02g01189	chr02	8814641	8815360	106.383	0	0	517	3	TA
TAOs02g01506	chr02	12188622	12189521	176.6667	8.777778	0	900	1	TA
TAOs02g01520	chr02	12409626	12410687	258.9454	11.11111	0	1062	1	TA
TAOs02g01577	chr02	13041526	13042706	419.9831	3.979678	0	1181	1	TA
TAOs02g01595	chr02	13252592	13253783	125	16.94631	0	1192	1	TA
TAOs02g01600	chr02	13306682	13316088	840.6902	13.69193	0	2492	12	TA
TAOs02g01638	chr02	13855737	13859371	3586.689	22.47593	0	1217	4	TA
TAOs02g01641	chr02	13878566	13880469	175.1497	16.17647	0	1336	4	TA
TAOs02g01650	chr02	13969898	13970262	136.9863	22.19178	0	365	1	TA
TAOs02g01714	chr02	14977302	14977978	1331.07	15.65731	0	589	2	TA
TAOs02g01731	chr02	15167268	15172453	523.9787	3.278056	0	563	5	TA
TAOs02g01755	chr02	15489339	15490050	121.9963	19.38202	0	541	2	TA
TAOs02g01818	chr02	16104596	16110797	1131.8	2.370203	0	2261	9	TA
TAOs02g01832	chr02	16253239	16254040	118.4539	0	0	802	1	TA
TAOs02g01833	chr02	16257392	16258146	263.5762	0	0	755	1	TA
TAOs02g01864	chr02	16634981	16636645	273.8739	13.15315	0	1665	1	TA
TAOs02g01942	chr02	17548196	17548574	189.9736	0	0	379	1	TA
TAOs02g02020	chr02	18379444	18380300	920.6534	10.50175	0	857	1	TA
TAOs02g02073	chr02	18948688	18951429	383.908	14.3326	0	2610	2	TA
TAOs02g02129	chr02	19346763	19351521	305.2699	10.46438	0	4668	2	TA
TAOs02g02133	chr02	19377444	19378109	141.1411	11.71171	0	666	1	TA
TAOs02g02137	chr02	19385178	19385874	129.6296	6.312769	0	540	3	TA
TAOs02g02224	chr02	20134787	20135311	175.2381	15.04762	0	525	1	TA
TAOs02g02237	chr02	20368948	20391617	635.1001	2.474636	0	5396	38	TA
TAOs02g02469	chr02	22416039	22417182	114.2191	0	0	429	3	TA
TAOs02g02470	chr02	22417307	22419063	160.1423	8.537279	0	281	4	TA
TAOs02g02475	chr02	22435611	22435779	124.2604	0	0	169	1	TA
TAOs02g02483	chr02	22526066	22526905	266.6667	15.71429	0	840	1	TA
TAOs02g02512	chr02	22888425	22888618	103.0928	0	0	194	1	TA
TAOs02g02514	chr02	22898277	22898994	1408.078	0	0	718	1	TA
TAOs02g02529	chr02	23032341	23032880	214.8148	15.55556	0	540	1	TA
TAOs02g02550	chr02	23232099	23232958	110.4651	0	0	860	1	TA
TAOs02g02726	chr02	24564058	24565142	171.4286	7.926267	0	1085	1	TA
TAOs02g02739	chr02	24706678	24707447	206.4935	19.87013	0	770	1	TA
TAOs02g02812	chr02	25410319	25410946	170.7317	0	0	533	2	TA
TAOs02g02891	chr02	26130106	26132398	289.577	0	0	2293	1	TA
TAOs02g02967	chr02	26774333	26774990	140	0	0	500	2	TA
TAOs02g03035	chr02	27238074	27238529	104.1667	10.52632	0	192	2	TA
TAOs02g03042	chr02	27273427	27279573	350.1712	4.001952	0	4672	2	TA
TAOs02g03048	chr02	27302075	27302873	223.301	0	0	721	2	TA
TAOs02g03080	chr02	27594133	27598584	134.0782	13.05031	0	358	3	TA
TAOs02g03111	chr02	27771108	27773148	120.5292	1.910828	0	2041	1	TA
TAOs02g03160	chr02	28174331	28180809	324.3338	2.577558	0	2852	2	TA
TAOs02g03197	chr02	28428930	28430543	568.1537	12.82528	0	1614	1	TA
TAOs02g03221	chr02	28556825	28557640	275.7353	17.76961	0	816	1	TA
TAOs02g03330	chr02	29483363	29495763	3231.822	1.636965	0	3947	24	TA
TAOs02g03410	chr02	30101989	30102826	538.1862	10.62053	0	838	1	TA
TAOs02g03427	chr02	30238299	30238752	689.4273	0	0	454	1	TA
TAOs02g03450	chr02	30358416	30359167	115.2174	0	0	460	2	TA
TAOs02g03452	chr02	30385497	30386291	1465.409	18.23899	0	795	1	TA
TAOs02g03575	chr02	31456660	31457461	142.1446	0	0	802	1	TA
TAOs02g03797	chr02	33067444	33069473	102.4631	8.275862	0	2030	1	TA
TAOs02g03805	chr02	33117861	33124099	396.408	3.04536	0	3118	17	TA
TAOs02g03830	chr02	33417764	33421385	153.3269	0	0	1037	3	TA
TAOs02g03891	chr02	33746846	33748578	230.303	0	2.769763	330	2	TA
TAOs02g03922	chr02	33966121	33966819	132.2034	21.74535	0	590	2	TA
TAOs02g03923	chr02	33967385	33969353	161.2903	14.42357	0	1271	6	TA
TAOs02g03956	chr02	34233191	34234599	101.5453	20.79489	0	453	3	TA
TAOs02g04024	chr02	34706721	34707861	115.688	2.716915	0	1141	1	TA
TAOs02g04052	chr02	34904613	34905039	129.3706	0	0	286	2	TA
TAOs02g04144	chr02	35475922	35477678	708.2658	10.52931	0	617	2	TA
TAOs02g04181	chr02	35679605	35681462	2714.637	2.314316	0	813	2	TA
TAOs03g00076	chr03	528996	530017	104.7904	4.990215	0	334	3	TA
TAOs03g00077	chr03	530247	536177	284.153	4.434328	0	366	3	TA
TAOs03g00107	chr03	728243	729851	2884.956	8.328154	0	339	2	TA
TAOs03g00122	chr03	837545	841763	307.934	7.466224	0	1273	3	TA

TAOs03g00156	chr03	1007579	1008073	179.798	0	0	495	1	TA
TAOs03g00160	chr03	1048989	1075683	3575.366	1.445964	0	6422	50	TA
TAOs03g00171	chr03	1117820	1118121	102.649	12.25166	0	302	1	TA
TAOs03g00472	chr03	3414111	3416780	554.9133	5.580524	0	1211	2	TA
TAOs03g00560	chr03	4020974	4022317	195.6845	16.22024	0	1344	1	TA
TAOs03g00587	chr03	4164342	4165585	823.0703	24.51768	0	1153	2	TA
TAOs03g00619	chr03	4347051	4348290	2056.863	3.467742	0	1020	3	TA
TAOs03g00667	chr03	4675257	4680061	680.5195	24.20395	0	1155	4	TA
TAOs03g00728	chr03	5159058	5160249	204.698	0	0	1192	1	TA
TAOs03g00782	chr03	5541701	5542853	216.8257	6.244579	0	1153	1	TA
TAOs03g00812	chr03	5694580	5697322	330.9115	9.223478	0	1514	3	TA
TAOs03g00856	chr03	6004437	6005022	172.3549	7.849829	0	586	1	TA
TAOs03g00871	chr03	6175653	6176985	152.4752	4.726182	0	505	2	TA
TAOs03g00875	chr03	6179669	6181789	346.1538	18.90618	0	1170	5	TA
TAOs03g00898	chr03	6317892	6318361	442.5532	24.04255	0	470	1	TA
TAOs03g00905	chr03	6342372	6352694	216.7217	5.376344	3.167684	2727	12	TA
TAOs03g00971	chr03	6857897	6871226	568.0991	1.807952	0	1373	6	TA
TAOs03g01029	chr03	7176890	7177551	125.3776	0	0	662	1	TA
TAOs03g01035	chr03	7235102	7235492	143.2225	0	0	391	1	TA
TAOs03g01133	chr03	8067447	8071919	5716.312	6.438632	0	2115	10	TA
TAOs03g01145	chr03	8206772	8207138	147.139	0	0	367	1	TA
TAOs03g01572	chr03	11085483	11087675	278.6138	19.92704	0	2193	1	TA
TAOs03g01659	chr03	11836591	11838251	556.7452	1.866346	0	467	2	TA
TAOs03g01662	chr03	11841701	11841898	196.9697	0	0	198	1	TA
TAOs03g01678	chr03	11993786	11995290	197.2556	0	0	583	3	TA
TAOs03g01761	chr03	12681982	12682626	320.9302	0	0	645	1	TA
TAOs03g01793	chr03	12920531	12921772	203.7037	5.152979	0	1242	1	TA
TAOs03g01797	chr03	12937583	12938288	263.4561	14.02266	0	706	1	TA
TAOs03g01809	chr03	13081895	13082925	215.3249	16.97381	0	1031	1	TA
TAOs03g01810	chr03	13083081	13084316	518.6084	19.09385	0	1236	1	TA
TAOs03g01820	chr03	13131441	13132169	283.9506	0	0	729	1	TA
TAOs03g01879	chr03	13707109	13709058	270.5167	0	0	1316	2	TA
TAOs03g01889	chr03	13807781	13808603	608.7485	3.888214	0	823	1	TA
TAOs03g01963	chr03	14562252	14562780	423.4405	0	0	529	1	TA
TAOs03g02053	chr03	15735410	15739160	1326.496	6.664889	0	585	2	TA
TAOs03g02094	chr03	16197821	16202117	119.3856	5.56202	0	4297	1	TA
TAOs03g02216	chr03	17690823	17692106	454.0498	23.52025	0	1284	1	TA
TAOs03g02378	chr03	19715778	19717716	451.3166	18.05054	0	1633	2	TA
TAOs03g02456	chr03	20696076	20697973	584.5666	23.70917	1.580611	946	2	TA
TAOs03g02487	chr03	21016800	21017404	148.7603	11.07438	0	605	1	TA
TAOs03g02502	chr03	21164810	21165444	2424.307	0	0	469	2	TA
TAOs03g02523	chr03	21338306	21340307	468.1648	16.53347	0	534	2	TA
TAOs03g02596	chr03	22107374	22107863	183.6735	0	0	490	1	TA
TAOs03g02622	chr03	22340894	22345572	3422.025	12.88737	0	1353	8	TA
TAOs03g02643	chr03	22544318	22547288	125.7046	5.45271	0	1774	2	TA
TAOs03g02648	chr03	22612488	22614479	983.7398	0	0	615	2	TA
TAOs03g02660	chr03	22715320	22718489	169.9577	8.391167	0	1418	3	TA
TAOs03g02667	chr03	22748966	22749733	558.5938	0	0	768	1	TA
TAOs03g02685	chr03	22873834	22876995	1108.818	0	0	1066	2	TA
TAOs03g02702	chr03	23126895	23127110	134.2593	0	0	216	1	TA
TAOs03g02733	chr03	23392192	23392389	398.9899	14.14141	0	198	1	TA
TAOs03g02762	chr03	23631586	23631952	321.5259	11.71662	0	367	1	TA
TAOs03g02799	chr03	24004460	24006186	258.4148	10.53851	0	921	3	TA
TAOs03g02800	chr03	24008505	24009127	261.6372	0	0	623	1	TA
TAOs03g02837	chr03	24584043	24584809	158.1028	5.215124	0	506	2	TA
TAOs03g02864	chr03	24848340	24848843	103.1746	9.722222	0	504	1	TA
TAOs03g03066	chr03	26972795	26975408	207.0039	14.61362	0	1285	2	TA
TAOs03g03259	chr03	28802649	28803622	117.0431	23.40862	0	974	1	TA
TAOs03g03285	chr03	29013322	29014269	292.1941	24.36709	0	948	1	TA
TAOs03g03293	chr03	29086674	29088257	15152.21	16.98232	0	611	2	TA
TAOs03g03294	chr03	29095680	29096672	186.3041	0	0	993	1	TA
TAOs03g03350	chr03	29480309	29484950	702.765	2.434296	0	1736	10	TA
TAOs03g03466	chr03	30399222	30400737	261.9543	9.102902	0	481	2	TA
TAOs03g03560	chr03	30985029	30986025	147.139	22.4674	0	367	3	TA
TAOs03g03578	chr03	31103960	31107635	287.8577	20.2938	0	3599	2	TA
TAOs03g03698	chr03	31806877	31807701	301.8182	0	0	825	1	TA

TAOs03g03766	chr03	32257639	32257961	260.0619	11.14551	0	323	1	TA
TAOs03g03786	chr03	32398221	32399926	28460.06	5.275498	0	676	2	TA
TAOs03g03961	chr03	33462251	33463217	126.4916	0	0	419	3	TA
TAOs03g04044	chr03	34010884	34024744	962.2698	1.774764	0	3313	27	TA
TAOs03g04069	chr03	34215594	34217999	512.1359	1.330008	0	824	3	TA
TAOs03g04084	chr03	34301228	34303639	202.5518	12.23051	0	627	3	TA
TAOs03g04111	chr03	34469290	34473218	358.6957	9.111733	0	828	5	TA
TAOs03g04169	chr03	34789283	34789663	215.2231	18.63517	0	381	1	TA
TAOs03g04254	chr03	35409135	35409845	112.5176	0	0	711	1	TA
TAOs03g04291	chr03	35565690	35568282	878.4965	13.03509	0	1144	4	TA
TAOs04g00018	chr04	244205	244854	976.9231	7.384615	0	650	1	TA
TAOs04g00056	chr04	659122	659924	118.3064	13.5741	0	803	1	TA
TAOs04g00135	chr04	1744765	1746478	198.1221	21.99533	0	1065	6	TA
TAOs04g00259	chr04	4314832	4316074	163.0435	6.11424	0	644	3	TA
TAOs04g00265	chr04	4382139	4385581	428.5714	13.30235	0	2219	5	TA
TAOs04g00266	chr04	4405992	4407520	161.7934	1.896664	0	513	5	TA
TAOs04g00267	chr04	4407628	4408907	198.0198	0	0	606	5	TA
TAOs04g00268	chr04	4409033	4418898	1314.554	2.990067	0	1917	17	TA
TAOs04g00292	chr04	4644239	4645013	11948.39	0	0	775	1	TA
TAOs04g00293	chr04	4645275	4648910	1094.698	9.29593	0	3527	2	TA
TAOs04g00309	chr04	4915217	4915612	126.2626	19.94949	0	396	1	TA
TAOs04g00403	chr04	5799512	5800084	164.0489	0	0	573	1	TA
TAOs04g00580	chr04	7939727	7943871	212.7822	7.961399	0	2613	6	TA
TAOs04g00656	chr04	8764573	8767226	189.5288	4.107008	0	955	3	TA
TAOs04g00772	chr04	9677403	9678634	808.4416	16.96429	0	1232	1	TA
TAOs04g00860	chr04	10470569	10472444	619.1336	9.701493	0	554	2	TA
TAOs04g00932	chr04	11244705	11245171	353.3191	14.3469	0	467	1	TA
TAOs04g01126	chr04	13593622	13594178	105.9246	0	0	557	1	TA
TAOs04g01147	chr04	13868777	13869244	264.2643	0	0	333	2	TA
TAOs04g01215	chr04	14725178	14728136	173.8693	17.67489	0	995	2	TA
TAOs04g01281	chr04	16021488	16023796	462.1711	6.409701	0	1216	7	TA
TAOs04g01333	chr04	16799655	16800471	1866.585	17.62546	0	817	1	TA
TAOs04g01339	chr04	16840420	16842448	160.1774	10.49778	0	2029	1	TA
TAOs04g01348	chr04	17087729	17088715	350.2377	11.95542	0	631	3	TA
TAOs04g01419	chr04	18285065	18286437	140.5681	9.104151	0	1373	1	TA
TAOs04g01423	chr04	18369036	18371145	15667.03	7.440758	0	919	4	TA
TAOs04g01437	chr04	18478648	18483237	461.8096	16.88453	0	2553	6	TA
TAOs04g01571	chr04	19956098	19956978	349.6027	16.34506	0	881	1	TA
TAOs04g01572	chr04	19962938	19963794	2154.026	8.4014	0	857	1	TA
TAOs04g01598	chr04	20131675	20132883	120.761	9.42928	0	1209	1	TA
TAOs04g01633	chr04	20431620	20435772	9722.494	3.29882	0	1636	6	TA
TAOs04g01688	chr04	20838524	20839294	114.6067	22.43839	0	445	3	TA
TAOs04g01690	chr04	20847195	20851045	184.0178	7.037133	0	3141	5	TA
TAOs04g01795	chr04	21643961	21658218	1013.225	4.285314	0	983	6	TA
TAOs04g01872	chr04	22256267	22257072	307.6923	0	0	806	1	TA
TAOs04g01880	chr04	22304655	22305064	280.4878	0	0	410	1	TA
TAOs04g01904	chr04	22461311	22464114	6788.618	5.777461	0	984	4	TA
TAOs04g01966	chr04	22980583	22981411	989.1435	0	0	829	1	TA
TAOs04g01969	chr04	23004679	23005361	1837.482	11.71303	0	683	1	TA
TAOs04g01984	chr04	23053527	23057018	264.1243	0.91638	0	708	2	TA
TAOs04g02131	chr04	24213721	24217640	2323.507	0	0	1289	4	TA
TAOs04g02143	chr04	24288274	24289565	139.3189	14.70588	0	1292	1	TA
TAOs04g02240	chr04	24906483	24909187	20538.83	18.66913	0	631	4	TA
TAOs04g02291	chr04	25240745	25241104	136.1111	0	0	360	1	TA
TAOs04g02358	chr04	25751279	25753416	21173.91	5.659495	0	851	5	TA
TAOs04g02372	chr04	25917859	25918631	129.3661	22.5097	0	773	1	TA
TAOs04g02416	chr04	26314014	26317276	241.335	5.669629	0	779	3	TA
TAOs04g02509	chr04	27106573	27107837	168.3794	0	0	1265	1	TA
TAOs04g02521	chr04	27182238	27182651	101.4493	0	0	414	1	TA
TAOs04g02531	chr04	27238721	27240654	760.0827	6.101344	0	1934	1	TA
TAOs04g02550	chr04	27349991	27351229	105.4994	20.33898	0	891	2	TA
TAOs04g02570	chr04	27536867	27537499	363.3491	4.897314	0	633	1	TA
TAOs04g02647	chr04	28174887	28175635	154.8732	18.42457	0	749	1	TA
TAOs04g02752	chr04	28889417	28891861	653.9877	7.893661	0	2445	1	TA
TAOs04g02759	chr04	29054017	29055785	1365.178	23.57264	0	1769	1	TA
TAOs04g02772	chr04	29149223	29150810	136.6255	3.526448	0	1215	4	TA

TAOs04g02774	chr04	29163336	29165163	971.7742	0	0	744	4	TA
TAOs04g02800	chr04	29407232	29408258	4357.352	4.186952	0	1027	1	TA
TAOs04g02808	chr04	29453600	29454886	149.4024	0	0	502	2	TA
TAOs04g02861	chr04	29716202	29716583	159.6859	10.4712	0	382	1	TA
TAOs04g02984	chr04	30445864	30446127	185.6061	0	0	264	1	TA
TAOs04g02999	chr04	30515043	30516341	141.6579	2.232487	0	953	3	TA
TAOs04g03117	chr04	31360616	31361805	1942.857	9.579832	0	1190	1	TA
TAOs04g03177	chr04	31712145	31713955	162.3412	8.006626	0	1811	1	TA
TAOs04g03178	chr04	31715040	31715352	121.4058	0	0	313	1	TA
TAOs04g03204	chr04	31864982	31867348	167.3182	9.04098	0	1279	5	TA
TAOs04g03218	chr04	31922302	31925319	261.5385	0	0	455	2	TA
TAOs04g03336	chr04	32735740	32736133	279.1878	21.06599	0	394	1	TA
TAOs04g03352	chr04	32831857	32832435	208.981	21.58895	0	579	1	TA
TAOs04g03383	chr04	33051739	33052324	126.2799	24.74403	0	586	1	TA
TAOs04g03458	chr04	33750891	33751323	120.0924	0	0	433	1	TA
TAOs04g03561	chr04	34468712	34488106	622.0934	0.763083	0	14665	13	TA
TAOs04g03621	chr04	34826917	34827946	125.2427	22.13592	0	1030	1	TA
TAOs04g03686	chr04	35393449	35393677	104.8035	0	0	229	1	TA
TAOs04g03699	chr04	35473445	35478041	11975.7	8.331521	0	1934	8	TA
TAOs05g00027	chr05	153679	154610	282.1888	22.85408	0	932	1	TA
TAOs05g00069	chr05	458849	459822	684.3931	17.96715	0	865	2	TA
TAOs05g00071	chr05	483359	484595	4065.551	17.54244	0	717	3	TA
TAOs05g00083	chr05	568650	572534	1513.41	2.496782	0	1566	8	TA
TAOs05g00128	chr05	887631	889381	261.244	6.110794	0	1045	3	TA
TAOs05g00151	chr05	1008828	1009268	111.1111	0	0	441	1	TA
TAOs05g00240	chr05	1647796	1648328	292.6829	0	0	533	1	TA
TAOs05g00283	chr05	2124173	2127836	264.2225	15.63865	0	2373	5	TA
TAOs05g00293	chr05	2205408	2207051	182.6568	22.20195	2.798054	542	2	TA
TAOs05g00327	chr05	2393596	2399263	379.085	3.440367	0	918	4	TA
TAOs05g00433	chr05	2995243	2997751	122.0826	15.26505	0	557	2	TA
TAOs05g00560	chr05	4165096	4167375	146.4912	0	0	2280	1	TA
TAOs05g00602	chr05	4365604	4370550	315.5316	13.34142	0	2041	4	TA
TAOs05g00652	chr05	4801910	4803361	519.2837	18.04408	0	1452	1	TA
TAOs05g00660	chr05	4870723	4871314	4052.365	8.783784	0	592	1	TA
TAOs05g00697	chr05	5284679	5286795	123.8806	2.031176	3.448276	670	2	TA
TAOs05g00743	chr05	5818064	5821396	985.5538	4.680468	0	623	3	TA
TAOs05g00808	chr05	6758356	6759066	101.2658	24.89451	0	711	1	TA
TAOs05g00916	chr05	8264733	8290967	1158.418	11.38063	0	8143	58	TA
TAOs05g01003	chr05	9194975	9198490	7118.721	2.275313	0	876	4	TA
TAOs05g01015	chr05	9445703	9446971	182.3362	21.5918	0	702	2	TA
TAOs05g01146	chr05	10955267	10955627	257.6177	0	0	361	1	TA
TAOs05g01165	chr05	11257260	11259646	296.748	13.11269	0	492	3	TA
TAOs05g01358	chr05	13554144	13555516	386.7444	22.28696	0	1373	1	TA
TAOs05g01415	chr05	14181391	14186416	6128.901	22.06526	0	2211	12	TA
TAOs05g01421	chr05	14309596	14310242	168.4699	12.36476	0	647	1	TA
TAOs05g01645	chr05	16752468	16774654	1043.536	1.162843	0	5995	37	TA
TAOs05g01716	chr05	17432987	17433691	125.817	14.1844	0	612	2	TA
TAOs05g01742	chr05	17653933	17654970	819.8459	13.39114	0	1038	1	TA
TAOs05g01770	chr05	17852721	17853464	131.7204	8.602151	0	744	1	TA
TAOs05g01837	chr05	18383587	18385718	274.6331	5.300188	0	954	3	TA
TAOs05g01892	chr05	18877686	18878360	103.7037	13.18519	0	675	1	TA
TAOs05g02014	chr05	20123240	20123776	171.3222	0	0	537	1	TA
TAOs05g02025	chr05	20223458	20224364	112.7013	14.66373	0	559	2	TA
TAOs05g02079	chr05	20715385	20716413	167.1526	0	0	1029	1	TA
TAOs05g02090	chr05	20844493	20845536	152.1984	0	0	887	2	TA
TAOs05g02129	chr05	21254396	21255455	250	10.18868	0	1060	1	TA
TAOs05g02235	chr05	22033995	22035711	279.9071	17.99651	0	861	5	TA
TAOs05g02251	chr05	22192652	22197674	126.5734	9.396775	0	4290	2	TA
TAOs05g02337	chr05	22785299	22785522	330.3571	16.07143	0	224	1	TA
TAOs05g02438	chr05	23628448	23631982	227.6215	1.103253	0	1173	3	TA
TAOs05g02439	chr05	23633953	23635584	153.799	10.78431	1.409314	1632	1	TA
TAOs05g02449	chr05	23724671	23725250	334.4828	22.06897	0	580	1	TA
TAOs05g02554	chr05	24494817	24495368	132.2464	11.23188	0	552	1	TA
TAOs05g02560	chr05	24536105	24537367	239.1132	21.93191	0	1263	1	TA
TAOs05g02603	chr05	24912958	24914443	474.428	11.37281	0	1486	1	TA
TAOs05g02636	chr05	25195272	25197851	359.3023	18.10078	0	2580	1	TA

TAOs05g02784	chr05	26274824	26276353	954.902	3.529412	0	1530	1	TA
TAOs05g02834	chr05	26602002	26602599	112.0401	0	0	598	1	TA
TAOs05g02838	chr05	26634850	26640882	3864.715	0.828775	0	1981	15	TA
TAOs05g02908	chr05	27074053	27074728	176.0355	0	0	676	1	TA
TAOs06g00193	chr06	1661179	1662850	814.9123	18.54067	0	1140	2	TA
TAOs06g00318	chr06	2455940	2457055	175.6272	20.2509	0	1116	1	TA
TAOs06g00332	chr06	2527046	2528306	101.9284	7.295797	0	726	2	TA
TAOs06g00396	chr06	2884955	2885249	118.6441	0	0	295	1	TA
TAOs06g00429	chr06	3152486	3153733	207.0031	0	0	971	2	TA
TAOs06g00704	chr06	5190504	5191962	140.5072	3.906785	0	1459	1	TA
TAOs06g00883	chr06	6785429	6787398	106.0914	6.497462	0	1970	1	TA
TAOs06g00935	chr06	7270443	7274678	1115.824	20.65628	0	613	3	TA
TAOs06g00936	chr06	7287794	7290623	447.4808	18.62191	0	1171	2	TA
TAOs06g00981	chr06	7729949	7730662	310.9244	0	0	714	1	TA
TAOs06g00985	chr06	7765012	7766269	154.8443	0	0	1156	2	TA
TAOs06g01014	chr06	8082660	8082932	106.2271	0	0	273	1	TA
TAOs06g01066	chr06	8875716	8875987	117.6471	0	0	272	1	TA
TAOs06g01068	chr06	8880081	8884813	528.3117	0	0	1925	6	TA
TAOs06g01072	chr06	8939399	8942000	130.4348	2.152191	0	437	3	TA
TAOs06g01073	chr06	8943652	8944641	204.0404	0	0	990	1	TA
TAOs06g01095	chr06	9198545	9199484	162.766	16.48936	0	940	1	TA
TAOs06g01358	chr06	12351539	12351990	148.2301	0	0	452	1	TA
TAOs06g01360	chr06	12359454	12360202	67089.45	0	0	749	1	TA
TAOs06g01457	chr06	13280867	13284576	176.2861	16.92722	0	3188	2	TA
TAOs06g01583	chr06	14614196	14615691	139.0071	6.350267	0	705	2	TA
TAOs06g01653	chr06	15253808	15254514	1417.256	0	0	707	1	TA
TAOs06g01749	chr06	16292538	16295349	179.2511	0	0	1629	2	TA
TAOs06g01752	chr06	16307541	16309617	421.7578	15.11796	0	1866	3	TA
TAOs06g01772	chr06	16456047	16461032	199.5588	17.44886	0	4986	1	TA
TAOs06g01776	chr06	16464382	16466591	100	23.93665	0	2210	1	TA
TAOs06g01814	chr06	16788570	16789339	263.6364	11.55844	0	770	1	TA
TAOs06g01816	chr06	16815845	16817906	544.6169	16.44035	0	2062	1	TA
TAOs06g01821	chr06	16887172	16887732	117.6471	9.803922	0	561	1	TA
TAOs06g01828	chr06	16962393	16966574	2252.788	24.36633	0	807	3	TA
TAOs06g01883	chr06	17573048	17573595	135.0365	0	0	548	1	TA
TAOs06g01905	chr06	17902032	17902724	431.4574	17.17172	0	693	1	TA
TAOs06g01911	chr06	17981845	17982381	1467.412	13.40782	0	537	1	TA
TAOs06g01963	chr06	18913205	18915779	179.1639	15.84466	0	1507	4	TA
TAOs06g02020	chr06	19730627	19735314	6218.845	5.268771	0	987	4	TA
TAOs06g02046	chr06	20090455	20091067	145.1876	6.362153	0	613	1	TA
TAOs06g02084	chr06	20716133	20717211	123.2623	10.47266	0	1079	1	TA
TAOs06g02109	chr06	20977671	20982402	410.1633	5.156382	0	551	3	TA
TAOs06g02227	chr06	22090357	22091292	139.1753	19.12393	0	194	2	TA
TAOs06g02256	chr06	22387794	22388346	220.6148	15.55154	0	553	1	TA
TAOs06g02259	chr06	22542333	22546589	162.8664	6.295513	0	307	2	TA
TAOs06g02263	chr06	22556205	22556871	154.4228	19.04048	0	667	1	TA
TAOs06g02401	chr06	23728897	23731661	1964.545	5.858951	0	1100	3	TA
TAOs06g02450	chr06	24196806	24197533	130.4945	7.82967	0	728	1	TA
TAOs06g02469	chr06	24303849	24306445	173.1449	0	0	283	2	TA
TAOs06g02480	chr06	24411881	24412123	205.7613	0	0	243	1	TA
TAOs06g02489	chr06	24509636	24509963	128.0488	0	0	328	1	TA
TAOs06g02529	chr06	24946204	24946620	119.9041	0	0	417	1	TA
TAOs06g02530	chr06	24946752	24950722	201.7773	0	0	3826	2	TA
TAOs06g02531	chr06	24951977	24952898	107.4965	4.446855	0	707	3	TA
TAOs06g02582	chr06	25448424	25449786	162.8895	7.776963	0	706	4	TA
TAOs06g02627	chr06	25904783	25906491	240.4915	18.49035	0	1709	1	TA
TAOs06g02676	chr06	26465179	26469652	1035.986	6.951274	0	4474	1	TA
TAOs06g02680	chr06	26485482	26486343	107.8886	12.8	0	862	1	TA
TAOs06g02684	chr06	26493569	26493990	170.6161	8.056872	0	422	1	TA
TAOs06g02822	chr06	27643501	27646676	142.9471	1.22796	0	3176	1	TA
TAOs06g02823	chr06	27652089	27656715	460.1254	5.273395	0	4627	1	TA
TAOs06g02860	chr06	27958452	27960160	979.1873	22.9959	0	1009	2	TA
TAOs06g02927	chr06	28446539	28447974	1394.15	9.610028	0	1436	1	TA
TAOs06g03014	chr06	28992502	28993012	105.6751	0	0	511	1	TA
TAOs06g03189	chr06	30145673	30149398	190.8714	6.17284	0	241	2	TA
TAOs06g03339	chr06	31090178	31093307	232.9045	7.028754	0	1477	8	TA

TAOs07g00071	chr07	596774	598213	225	13.19444	0	1440	1	TA
TAOs07g00140	chr07	1164350	1165079	2169.863	0	0	730	1	TA
TAOs07g00187	chr07	1598802	1599557	109.7884	0	0	756	1	TA
TAOs07g00239	chr07	2029131	2031915	248.5066	9.658887	0	1674	2	TA
TAOs07g00243	chr07	2092634	2093739	10636.53	6.600362	0	553	2	TA
TAOs07g00259	chr07	2262168	2266809	258.6491	3.856097	0	607	3	TA
TAOs07g00272	chr07	2400242	2401107	4117.91	12.47113	0	670	3	TA
TAOs07g00438	chr07	3783162	3785040	134.5196	0	0	1405	4	TA
TAOs07g00466	chr07	3922847	3926342	326.8116	12.78604	0	1380	3	TA
TAOs07g00500	chr07	4213339	4214189	109.2832	10.34078	0	851	1	TA
TAOs07g00553	chr07	4567730	4568390	10538.58	12.55673	0	661	1	TA
TAOs07g00620	chr07	5212462	5215364	693.3614	9.920772	0	949	4	TA
TAOs07g00632	chr07	5365943	5366634	453.7572	15.60694	0	692	1	TA
TAOs07g00748	chr07	6572726	6573341	204.5455	8.116883	0	616	1	TA
TAOs07g00755	chr07	6661054	6663953	462.6168	11.24138	0	2568	3	TA
TAOs07g00772	chr07	6931468	6932646	1591.179	9.584394	0	1179	1	TA
TAOs07g00804	chr07	7237175	7237890	303.1496	0	0	508	3	TA
TAOs07g00810	chr07	7255089	7255571	217.3913	14.49275	0	483	1	TA
TAOs07g00825	chr07	7356117	7356853	153.4296	14.51832	0	554	2	TA
TAOs07g01006	chr07	10112207	10114161	857.4194	7.365729	0	1550	2	TA
TAOs07g01040	chr07	10279498	10280179	196.4809	21.84751	0	682	1	TA
TAOs07g01146	chr07	11843960	11844604	254.2636	19.53488	0	645	1	TA
TAOs07g01148	chr07	11861194	11861725	148.4962	21.99248	0	532	1	TA
TAOs07g01281	chr07	13193435	13194943	288.2704	21.2061	0	1509	1	TA
TAOs07g01398	chr07	14485306	14489280	184.1105	20.12579	0	2895	2	TA
TAOs07g01417	chr07	14716784	14717827	100.5747	19.25287	0	1044	1	TA
TAOs07g01456	chr07	15187611	15191338	312.2855	19.68884	0	1457	3	TA
TAOs07g01506	chr07	15951955	15952642	127.907	9.738372	0	688	1	TA
TAOs07g01513	chr07	16035046	16038627	267.4484	17.75544	0	1791	4	TA
TAOs07g01533	chr07	16497885	16498403	119.4605	18.30443	0	519	1	TA
TAOs07g01585	chr07	17174120	17174658	135.436	0	0	539	1	TA
TAOs07g01720	chr07	18639089	18639293	170.7317	16.09756	0	205	1	TA
TAOs07g01780	chr07	19187268	19187562	118.6441	0	0	295	1	TA
TAOs07g01819	chr07	19539547	19541056	123.3974	0	0	624	2	TA
TAOs07g02040	chr07	21428889	21431514	109.152	21.47753	0	1191	3	TA
TAOs07g02070	chr07	21672489	21674178	35567.49	11.2426	0	689	3	TA
TAOs07g02113	chr07	21935188	21935649	365.8009	8.441558	0	462	1	TA
TAOs07g02165	chr07	22385526	22386833	122.3242	2.522936	0	1308	1	TA
TAOs07g02208	chr07	22720527	22721420	270.6935	0	0	894	1	TA
TAOs07g02279	chr07	23204476	23207843	132.4935	11.34204	0	2302	3	TA
TAOs07g02314	chr07	23450500	23451129	271.4286	17.61905	0	630	1	TA
TAOs07g02342	chr07	23682457	23683011	648.6486	0	0	555	1	TA
TAOs07g02515	chr07	25018410	25019225	112.7451	14.70588	0	816	1	TA
TAOs07g02546	chr07	25260150	25261001	1190.141	5.868545	0	852	1	TA
TAOs07g02565	chr07	25427619	25429475	373.4104	12.27787	0	865	2	TA
TAOs07g02584	chr07	25540456	25542735	7001.159	1.973684	0	863	6	TA
TAOs07g02604	chr07	25649926	25652209	3834.291	6.654991	0	1044	2	TA
TAOs07g02960	chr07	28239051	28239546	197.5806	24.19355	0	496	1	TA
TAOs07g02990	chr07	28460255	28462421	181.6881	9.413936	0	699	3	TA
TAOs07g03005	chr07	28523186	28523802	179.9028	0	0	617	1	TA
TAOs07g03020	chr07	28704194	28704812	142.8571	0	0	462	2	TA
TAOs07g03037	chr07	28839941	28844837	151.7826	0.735144	0	2833	7	TA
TAOs07g03105	chr07	29285792	29286114	142.8571	0	0	238	2	TA
TAOs07g03110	chr07	29328874	29329993	331.7647	19.46429	0	850	2	TA
TAOs07g03125	chr07	29412624	29413724	729.337	22.2525	0	1101	1	TA
TAOs08g00120	chr08	1014893	1015510	122.9773	22.16828	0	618	1	TA
TAOs08g00284	chr08	2317001	2317405	113.5802	9.382716	0	405	1	TA
TAOs08g00309	chr08	2571983	2573294	132.3877	22.86585	0	423	3	TA
TAOs08g00331	chr08	2816384	2819061	535.8904	19.34279	0	1825	5	TA
TAOs08g00361	chr08	3030919	3031216	100.6711	0	0	298	1	TA
TAOs08g00375	chr08	3117834	3118326	322.5152	15.01014	0	493	1	TA
TAOs08g00378	chr08	3146486	3150122	7635.684	5.334067	0	936	4	TA
TAOs08g00452	chr08	3832620	3835080	1426.494	7.436002	0	619	2	TA
TAOs08g00488	chr08	4308612	4309807	193.75	13.04348	0	1120	2	TA
TAOs08g00500	chr08	4380421	4383037	775.1856	19.33512	0	943	2	TA
TAOs08g00880	chr08	8172793	8177412	242.1959	11.94805	0	3716	2	TA

TAOs08g00882	chr08	8182347	8182831	222.6804	9.278351	0	485	1	TA
TAOs08g00902	chr08	8335685	8339925	313.2137	8.394247	0	1839	6	TA
TAOs08g00996	chr08	9487714	9491818	196.1987	15.46894	0	1631	3	TA
TAOs08g01075	chr08	10578335	10579290	1444.561	19.87448	0	956	1	TA
TAOs08g01147	chr08	11374330	11376203	160.6715	13.18036	0	1251	2	TA
TAOs08g01243	chr08	12369950	12373650	3799.178	4.134018	0	2918	2	TA
TAOs08g01392	chr08	13909971	13910925	223.0366	23.4555	0	955	1	TA
TAOs08g01399	chr08	13959153	13960012	133.7209	23.72093	0	860	1	TA
TAOs08g01425	chr08	14283832	14287296	333.5	17.4026	0	2000	2	TA
TAOs08g01501	chr08	15093664	15096737	156.4626	0	0	882	5	TA
TAOs08g01557	chr08	16074288	16077123	151.7241	11.21298	0	870	3	TA
TAOs08g01752	chr08	18691479	18692522	610.1533	8.237548	0	1044	1	TA
TAOs08g01773	chr08	19059603	19065892	2866.747	19.82512	0	1666	9	TA
TAOs08g01822	chr08	19657700	19659296	148.5149	12.52348	0	404	2	TA
TAOs08g01840	chr08	19808905	19809818	396.0613	23.19475	0	914	1	TA
TAOs08g01852	chr08	19908152	19909510	188.5522	20.67697	0	297	2	TA
TAOs08g01896	chr08	20254914	20260027	267.2673	22.9957	0	1998	7	TA
TAOs08g02074	chr08	21880374	21881416	178.3317	8.533078	0	1043	1	TA
TAOs08g02088	chr08	21970314	21973214	187.8663	7.790417	0	2901	1	TA
TAOs08g02143	chr08	22624412	22627708	791.762	10.4034	0	2185	5	TA
TAOs08g02214	chr08	23234051	23234302	559.5238	19.04762	0	252	1	TA
TAOs08g02262	chr08	23602097	23602916	106.0976	4.634146	0	820	1	TA
TAOs08g02269	chr08	23691223	23692138	188.8646	22.81659	0	916	1	TA
TAOs08g02386	chr08	24795754	24796389	223.2704	5.503145	0	636	1	TA
TAOs08g02395	chr08	24892823	24893394	108.3916	22.72727	0	572	1	TA
TAOs08g02410	chr08	25053516	25056293	213.9831	5.795536	0	472	2	TA
TAOs08g02451	chr08	25393638	25394396	588.9328	0	0	759	1	TA
TAOs08g02537	chr08	26017741	26019051	255.5301	16.93364	0	1311	1	TA
TAOs08g02559	chr08	26189595	26192048	689.4942	0	0	1285	6	TA
TAOs08g02629	chr08	26756342	26759102	268.75	0	0	480	2	TA
TAOs08g02641	chr08	26887420	26887675	519.5313	0	0	256	1	TA
TAOs08g02729	chr08	27551988	27554146	256.2634	15.00695	0	1397	2	TA
TAOs08g02795	chr08	27814487	27815782	288.4615	3.935185	0	156	2	TA
TAOs08g02796	chr08	27815808	27816132	106.5574	0	0	244	2	TA
TAOs09g00099	chr09	1168867	1172536	185.0789	8.991826	0	697	3	TA
TAOs09g00306	chr09	2923276	2924949	108.8918	19.41458	0	1552	2	TA
TAOs09g00394	chr09	4049741	4052583	2709.147	3.869152	0	973	3	TA
TAOs09g00632	chr09	7301167	7301797	2366.086	0	0	631	1	TA
TAOs09g00856	chr09	10396314	10397031	240.9471	19.91643	0	718	1	TA
TAOs09g00890	chr09	10805039	10809454	167.3088	2.626812	0	3634	2	TA
TAOs09g00940	chr09	11779747	11785462	530.1174	3.411477	0	3918	16	TA
TAOs09g00958	chr09	11928404	11930336	250	17.53751	0	1180	2	TA
TAOs09g01047	chr09	12840203	12844258	267.9739	18.31854	0	306	2	TA
TAOs09g01078	chr09	13130795	13131418	171.4744	24.67949	0	624	1	TA
TAOs09g01094	chr09	13266468	13267018	105.2632	0	0	551	1	TA
TAOs09g01117	chr09	13592957	13593334	367.7249	8.994709	0	378	1	TA
TAOs09g01149	chr09	13904748	13907678	198.9101	6.516547	0	734	2	TA
TAOs09g01185	chr09	14351455	14351877	113.4752	13.94799	0	423	1	TA
TAOs09g01239	chr09	14919702	14920506	119.2547	3.478261	0	805	1	TA
TAOs09g01252	chr09	15065376	15072107	527.1163	3.966132	0	6509	3	TA
TAOs09g01309	chr09	15563072	15569739	793.667	3.659268	0	979	4	TA
TAOs09g01323	chr09	15712200	15712694	163.6364	0	0	495	1	TA
TAOs09g01399	chr09	16460351	16461518	446.9178	12.92808	0	1168	1	TA
TAOs09g01442	chr09	16864674	16865487	227.2727	0	0	814	1	TA
TAOs09g01472	chr09	17121350	17122104	108.6093	11.92053	0	755	1	TA
TAOs09g01490	chr09	17269165	17269958	348.8665	5.541562	0	794	1	TA
TAOs09g01557	chr09	17792865	17794468	721.4854	2.680798	0	754	2	TA
TAOs09g01562	chr09	17828974	17830012	109.7209	0	0	1039	1	TA
TAOs09g01565	chr09	17840443	17841762	103.0303	17.27273	0	1320	1	TA
TAOs09g01651	chr09	18423971	18424498	329.5455	0	0	528	1	TA
TAOs09g01681	chr09	18586663	18587398	493.2065	3.940217	0	736	1	TA
TAOs09g01850	chr09	19799368	19802657	2234.375	0	0	1280	6	TA
TAOs09g01853	chr09	19858382	19859902	107.8238	17.81723	0	1521	1	TA
TAOs09g01874	chr09	19986529	19988435	121.2914	0	0	1146	4	TA
TAOs09g01875	chr09	19988568	19988997	134.8837	12.7907	0	430	1	TA
TAOs09g01971	chr09	20580601	20581722	216.5775	23.61854	0	1122	1	TA

TAOs09g02019	chr09	20922260	20925077	100.1541	4.43577	0	1947	2	TA
TAOs09g02162	chr09	21915886	21916909	140.8451	0	0	639	3	TA
TAOs09g02206	chr09	22228534	22230069	160.9756	3.320313	0	410	2	TA
TAOs09g02243	chr09	22519455	22535536	811.3499	1.361771	0	6467	36	TA
TAOs10g00075	chr10	1118274	1118847	120.2091	24.04181	0	574	1	TA
TAOs10g00232	chr10	3068501	3071046	204.5455	0	0	836	3	TA
TAOs10g00246	chr10	3399068	3401092	155.122	0	0	1025	2	TA
TAOs10g00307	chr10	3983853	3986727	1802.474	9.321739	0	2668	2	TA
TAOs10g00328	chr10	4336455	4336767	172.524	0	0	313	1	TA
TAOs10g00343	chr10	4581872	4582429	107.5269	11.29032	0	558	1	TA
TAOs10g00367	chr10	4944348	4945001	256.8807	9.785933	0	654	1	TA
TAOs10g00381	chr10	5252540	5255029	302.4096	19.87952	0	2490	1	TA
TAOs10g00386	chr10	5372435	5376566	722.1429	14.39981	0	1400	3	TA
TAOs10g00427	chr10	5836744	5837439	130.7471	0	0	696	1	TA
TAOs10g00457	chr10	6214430	6217695	445.0019	8.236375	0	2591	3	TA
TAOs10g00507	chr10	6744750	6745911	453.9535	17.90017	0	1075	2	TA
TAOs10g00536	chr10	6944850	6945588	212.4493	24.89851	0	739	1	TA
TAOs10g00683	chr10	8660442	8664175	429.1805	15.42582	0	2111	3	TA
TAOs10g00731	chr10	9177098	9177700	122.7197	24.04643	0	603	1	TA
TAOs10g00792	chr10	9951698	9952158	104.1215	0	0	461	1	TA
TAOs10g00793	chr10	9969177	9970066	316.9811	9.213483	0	530	2	TA
TAOs10g00810	chr10	10118344	10119955	497.1382	0	0	1223	2	TA
TAOs10g00860	chr10	10649831	10650430	108.3333	0	0	600	1	TA
TAOs10g00956	chr10	11579414	11580809	409.5536	0	0	1277	2	TA
TAOs10g00978	chr10	11784364	11786740	246.1538	24.86327	0	520	4	TA
TAOs10g00979	chr10	11786813	11787604	107.3232	8.459596	0	792	1	TA
TAOs10g00995	chr10	11955317	11956129	431.7343	9.594096	0	813	1	TA
TAOs10g01041	chr10	12765158	12767822	351.9644	9.080675	0	2469	3	TA
TAOs10g01042	chr10	12798936	12803952	880.4598	2.710783	0	870	3	TA
TAOs10g01053	chr10	12882259	12885748	130.9456	3.667622	0	3490	1	TA
TAOs10g01104	chr10	13303688	13303975	361.1111	10.41667	0	288	1	TA
TAOs10g01151	chr10	14018441	14019232	3132.576	6.944444	0	792	1	TA
TAOs10g01257	chr10	14853867	14856961	137.2549	18.28756	0	1938	2	TA
TAOs10g01321	chr10	15468351	15471237	226.6667	6.33876	0	1350	4	TA
TAOs10g01364	chr10	15882753	15883588	184.2105	18.89952	0	836	1	TA
TAOs10g01638	chr10	18530461	18530966	440.7115	6.126482	0	506	1	TA
TAOs10g01641	chr10	18546797	18547042	105.6911	0	0	246	1	TA
TAOs10g01744	chr10	19095894	19096542	1268.105	0	0	649	1	TA
TAOs10g01810	chr10	19592996	19594240	167.0683	0	0	1245	1	TA
TAOs10g01844	chr10	19877026	19878044	133.4642	19.13641	0	1019	1	TA
TAOs10g01855	chr10	19919910	19920294	106.4935	0	0	385	1	TA
TAOs10g01858	chr10	19921753	19922006	125.9843	0	0	254	1	TA
TAOs10g01859	chr10	19922130	19923327	118.2222	0	0	1125	2	TA
TAOs10g01860	chr10	19923473	19927268	388.9816	4.373024	0	599	2	TA
TAOs10g01906	chr10	20286057	20290531	3126.165	6.815642	0	1395	11	TA
TAOs10g02030	chr10	21277371	21288523	671.8547	24.45082	0	771	2	TA
TAOs10g02060	chr10	21445886	21453596	1216.099	1.11529	0	4137	21	TA
TAOs10g02085	chr10	21580887	21584440	1141.145	6.443444	0	751	4	TA
TAOs10g02185	chr10	22320799	22323970	4406.25	0	0	1504	5	TA
TAOs10g02219	chr10	22526080	22527262	107.8717	15.8918	0	686	2	TA
TAOs10g02237	chr10	22697385	22698764	339.7683	4.42029	0	777	2	TA
TAOs10g02261	chr10	22936830	22938832	1711.636	0	0	1186	4	TA
TAOs10g02264	chr10	22954203	22955280	184.6011	0	0	1078	1	TA
TAOs10g02277	chr10	23023840	23025012	217.3913	12.70247	0	1173	1	TA
TAOs11g00023	chr11	166095	169328	498.1447	18.95485	0	3234	1	TA
TAOs11g00116	chr11	705926	708919	991.9872	19.57248	0	1248	3	TA
TAOs11g00280	chr11	1764850	1765274	185.8824	19.76471	0	425	1	TA
TAOs11g00349	chr11	2130566	2131438	150.1926	0	0	779	2	TA
TAOs11g00353	chr11	2166995	2167740	198.3914	17.29223	0	746	1	TA
TAOs11g00379	chr11	2333850	2334889	100	0	0	1040	1	TA
TAOs11g00388	chr11	2363858	2364491	323.3438	0	0.473186	634	1	TA
TAOs11g00403	chr11	2457765	2460828	345.9821	5.548303	0	896	4	TA
TAOs11g00426	chr11	2627399	2629862	567.2043	13.87987	0	744	3	TA
TAOs11g00471	chr11	3144620	3145571	258.4034	13.7605	0	952	1	TA
TAOs11g00603	chr11	4504987	4505853	7561.707	14.41753	0	867	1	TA
TAOs11g00605	chr11	4592410	4593330	647.1227	20.19544	0	921	1	TA

TAOs11g00680	chr11	5236274	5240685	736.2086	14.00725	0	997	3	TA
TAOs11g00738	chr11	5630674	5635002	211.7647	14.27581	0	1700	4	TA
TAOs11g00768	chr11	5858235	5868175	169.7652	7.212554	0	2044	11	TA
TAOs11g00785	chr11	6080396	6083734	191.9736	7.307577	0	3339	1	TA
TAOs11g00786	chr11	6083964	6085478	147.8548	16.76568	0	1515	1	TA
TAOs11g00905	chr11	7200176	7214073	199.095	4.640956	0	663	5	TA
TAOs11g00912	chr11	7327007	7327521	110.6796	16.50485	0	515	1	TA
TAOs11g00963	chr11	7803010	7803150	134.7518	0	0	141	1	TA
TAOs11g00964	chr11	7806382	7813242	416.1765	8.905407	0	1360	3	TA
TAOs11g01019	chr11	8362519	8363773	340.6814	20.79681	0	499	2	TA
TAOs11g01128	chr11	10090976	10091470	121.2121	0	0	495	1	TA
TAOs11g01166	chr11	10931727	10932179	196.468	15.45254	0	453	1	TA
TAOs11g01186	chr11	11370796	11371768	189.1059	4.624872	0	973	1	TA
TAOs11g01203	chr11	11712679	11716295	381.0445	4.064142	0	1551	5	TA
TAOs11g01233	chr11	12077030	12078594	1494.718	17.0607	0	568	3	TA
TAOs11g01237	chr11	12143135	12148329	3176.402	4.79307	0	856	7	TA
TAOs11g01291	chr11	12829993	12832206	153.5682	15.04065	0	2214	1	TA
TAOs11g01300	chr11	12911030	12919649	315.738	14.06032	0	2046	3	TA
TAOs11g01307	chr11	13088533	13089284	131.6489	24.33511	0	752	1	TA
TAOs11g01341	chr11	13631345	13632072	453.2967	25	0	728	1	TA
TAOs11g01361	chr11	13752477	13757447	1484.058	7.583987	0	1035	5	TA
TAOs11g01385	chr11	14127399	14129762	406.3158	10.53299	0	950	4	TA
TAOs11g01405	chr11	14321338	14323284	217.7294	17.66821	0	643	4	TA
TAOs11g01452	chr11	15234922	15235666	214.7651	22.55034	0	745	1	TA
TAOs11g01459	chr11	15344367	15346418	529.0488	13.49903	0	1945	2	TA
TAOs11g01461	chr11	15349281	15350054	403.1008	8.397933	0	774	1	TA
TAOs11g01470	chr11	15455380	15455939	119.6429	9.464286	0	560	1	TA
TAOs11g01532	chr11	16200879	16201819	802.3379	16.36557	0	941	1	TA
TAOs11g01542	chr11	16366304	16367550	2086.608	8.901363	0	1247	1	TA
TAOs11g01609	chr11	17289449	17289922	154.0541	0	0	370	2	TA
TAOs11g01671	chr11	17983639	17984306	321.8563	10.32934	0	668	1	TA
TAOs11g01684	chr11	18155008	18157595	418.5137	8.230294	0	767	2	TA
TAOs11g01839	chr11	19722867	19734456	4436.508	7.299396	0	2646	16	TA
TAOs11g01880	chr11	20189986	20191844	4359.333	5.433029	0.645508	1859	1	TA
TAOs11g01904	chr11	20386378	20389200	839.6065	23.20227	0	2338	2	TA
TAOs11g02037	chr11	21700130	21705713	115.1155	6.840974	4.190544	2858	3	TA
TAOs11g02172	chr11	22892806	22894787	151.4706	2.21998	0	680	4	TA
TAOs11g02344	chr11	24723110	24723876	410.691	13.82008	0	767	1	TA
TAOs11g02358	chr11	24927410	24928057	246.9136	0	0	648	1	TA
TAOs11g02378	chr11	25136359	25137536	481.9413	7.979626	0	886	2	TA
TAOs11g02391	chr11	25279803	25280587	272.6115	15.66879	0	785	1	TA
TAOs11g02423	chr11	25714054	25718322	770.2889	7.89412	0	1454	4	TA
TAOs11g02508	chr11	26533757	26537161	580.59	15.71219	0	2339	2	TA
TAOs11g02705	chr11	28700690	28702732	802.0086	0	0	697	3	TA
TAOs11g02727	chr11	28828591	28835652	248.0758	9.303314	0	1689	8	TA
TAOs12g00337	chr12	2031775	2033882	258.3774	10.05693	0	1134	5	TA
TAOs12g00338	chr12	2033988	2035709	110.3368	14.518	0	1722	1	TA
TAOs12g00339	chr12	2036042	2040492	192.8072	9.795552	0	1001	4	TA
TAOs12g00534	chr12	3733593	3734248	103.6585	17.37805	0	656	1	TA
TAOs12g00540	chr12	3782050	3783096	394.4604	10.60172	0	1047	1	TA
TAOs12g00624	chr12	4455473	4455963	107.943	0	0	491	1	TA
TAOs12g00637	chr12	4658316	4658665	151.4286	15.71429	0	350	1	TA
TAOs12g00656	chr12	4821312	4822211	118.8889	12.77778	0	900	1	TA
TAOs12g00677	chr12	4990301	4990800	110	24.2	0	500	1	TA
TAOs12g00697	chr12	5102552	5103436	288.1356	19.66102	0	885	1	TA
TAOs12g00702	chr12	5165455	5168664	396.7391	3.613707	0	1104	4	TA
TAOs12g00810	chr12	6168717	6169569	251.4205	14.4197	0	704	2	TA
TAOs12g00831	chr12	6489395	6490615	182.6372	17.77232	0	1221	1	TA
TAOs12g00832	chr12	6492680	6493479	64326.25	17.375	0	800	1	TA
TAOs12g00906	chr12	7242076	7242627	106.8841	11.05072	0	552	1	TA
TAOs12g00939	chr12	7556978	7557684	523.338	16.69024	0	707	1	TA
TAOs12g00997	chr12	8145447	8146646	184.7345	24.41667	0	904	2	TA
TAOs12g01020	chr12	8543093	8543817	234.4828	23.17241	0	725	1	TA
TAOs12g01038	chr12	8695281	8699365	991.778	17.38066	0	973	4	TA
TAOs12g01139	chr12	10138199	10139448	156.0403	10.08	0	596	2	TA
TAOs12g01192	chr12	10804320	10806072	473.747	7.529949	0	1676	2	TA

TAOs12g01211	chr12	11083997	11084310	111.465	0	0	314	1	TA
TAOs12g01303	chr12	12122142	12124733	324.5902	24.09756	0	610	4	TA
TAOs12g01327	chr12	12240385	12242807	10507.78	6.273215	0	707	2	TA
TAOs12g01354	chr12	12460460	12465425	393.2656	9.323399	0	1574	4	TA
TAOs12g01379	chr12	12833832	12839306	913.6894	10.10046	0	5318	3	TA
TAOs12g01392	chr12	13051683	13051855	5138.728	20.80925	0	173	1	TA
TAOs12g01507	chr12	14907846	14910075	152.0875	0	0	1006	6	TA
TAOs12g01508	chr12	14910184	14910813	149.4253	0	0	174	2	TA
TAOs12g01659	chr12	17400758	17401254	122.7364	0	0	497	1	TA
TAOs12g01731	chr12	18565305	18566273	317.8535	14.34469	0	969	1	TA
TAOs12g01732	chr12	18569264	18569785	526.8199	0	0	522	1	TA
TAOs12g01918	chr12	20510835	20511768	127.409	14.98929	0	934	1	TA
TAOs12g01977	chr12	20813606	20813965	144.4043	0	0	277	2	TA
TAOs12g01982	chr12	20843924	20846832	3298.22	18.3912	0	674	2	TA
TAOs12g01990	chr12	20902922	20903631	3829.577	12.95775	0	710	1	TA
TAOs12g02061	chr12	21593949	21597004	478.4644	20.58246	0	1068	4	TA
TAOs12g02116	chr12	22239519	22240854	583.8323	11.3024	0	1336	1	TA
TAOs12g02118	chr12	22287150	22287596	304.2506	9.619687	0	447	1	TA
TAOs12g02241	chr12	23439619	23440599	131.4985	17.22732	0	981	1	TA
TAOs12g02357	chr12	24507244	24508836	134.3377	3.264281	0	1593	1	TA
TAOs12g02518	chr12	25879533	25880874	911.3924	16.61699	0	395	2	TA
TAOs12g02575	chr12	26204366	26205394	109.8154	16.71526	0	1029	1	TA
TAOs12g02670	chr12	27062026	27063257	111.7318	4.301948	0	358	2	TA
TAOs12g02747	chr12	27493653	27493934	109.9291	0	0	282	1	TA
CUFF.1367.1	chr01	12621994	12623945	107.0697	19.41598	0	1952	1	Cuff
CUFF.2896.1	chr01	30206232	30208707	101.2048	0	0	415	2	Cuff
CUFF.4348.1	chr01	41217629	41219195	1042.017	6.06254	0	476	2	Cuff
CUFF.4364.1	chr01	41303825	41305109	148.7179	0	0	195	2	Cuff
CUFF.5658.1	chr02	8200463	8201477	131.2169	0	0	945	2	Cuff
CUFF.5741.1	chr02	9228099	9232659	131.9886	16.22451	0	4561	1	Cuff
CUFF.6506.3	chr02	19849018	19858052	205.1064	10.04981	2.390703	2350	5	Cuff
CUFF.7920.1	chr02	32896235	32901636	927.627	21.97334	0	1437	9	Cuff
CUFF.8353.1	chr02	35830821	35831264	137.3874	0	0	444	1	Cuff
CUFF.11059.1	chr03	26248946	26250571	537.2881	6.642066	0	590	2	Cuff
CUFF.11380.1	chr03	29167133	29168010	186.8557	3.758542	0	776	2	Cuff
CUFF.12618.1	chr04	5296178	5306496	1544.311	0	0	1670	18	Cuff
CUFF.16458.1	chr05	17854830	17855745	688.958	0	0	643	2	Cuff
CUFF.17925.1	chr06	802493	807035	212.8549	5.238829	0	4543	1	Cuff
CUFF.17926.1	chr06	807635	808000	250	0	0	216	2	Cuff
CUFF.17927.1	chr06	808196	808735	159.0909	0	0	352	2	Cuff
CUFF.17928.1	chr06	808811	810540	196.6019	0	0	412	2	Cuff
CUFF.20500.1	chr06	30885815	30886439	102.4	18.08	0	625	1	Cuff
CUFF.21866.2	chr07	19193352	19196976	346.1872	9.324138	0	2308	3	Cuff
CUFF.22526.1	chr07	25607417	25611497	1281.297	1.347709	0	2652	7	Cuff
CUFF.23643.1	chr08	7553224	7553476	106.7194	0	0	253	1	Cuff
CUFF.24767.1	chr08	23146579	23146895	1233.438	10.41009	0	317	1	Cuff
CUFF.25053.1	chr08	27045546	27046008	123.1102	0	0	463	1	Cuff
CUFF.25226.2	chr08	28312425	28315175	372.9097	12.28644	0	1196	6	Cuff
CUFF.26142.1	chr09	10921011	10921452	110.8597	7.918552	0	442	1	Cuff
CUFF.30588.1	chr11	24113459	24116607	222.6104	2.508733	0	3149	1	Cuff

Supplemental Table S7. RNA-seq read counts on the individual samples

Sample	**Reads**	**Not Mapped**	**Mapped 1x**	**Mapped >1x**	**% Mapped**
Control1	31,868,774	3,408,575	22,735,052	5,725,147	89.30
Control2	37,019,058	3,975,839	26,130,920	6,911,299	89.26
Control3	51,713,330	5,195,555	36,686,978	9,830,797	89.95
ABA1	43,934,073	4,523,048	31,507,355	7,903,670	89.70
ABA2	45,986,449	4,953,384	31,673,143	9,359,922	89.23

ABA3	40,236,999	4,751,220	16,227,440	19,258,339	88.19
Total	250,758,683	26,807,621	164,960,888	58,989,174	89.31

Supplemental Table S8. Nine previously reported ABA-inducible rice genes

Locus	Name	Control	ABA Treated	Fold	Log2 Fold	p value
LOC_Os11g26750	rab16d	0.103	1.846	17.892	4.161	4.05E-06
LOC_Os11g26760	rab16c	1.662	88.843	53.447	5.740	0
LOC_Os11g26780	rab16b	8.883	492.626	55.459	5.793	0
LOC_Os11g26790	rab16a	22.815	566.651	24.837	4.634	0
LOC_Os11g30500	TB2/DPI HVA22	0.244	4.136	16.972	4.085	7.68E-08
LOC_Os08g36440	similar to HVA22	5.243	117.292	22.372	4.484	0
LOC_Os05g46480	LEA	64.602	1762.550	27.283	4.770	2.22E-16
LOC_Os01g50910	WS118	38.049	745.502	19.593	4.292	0
LOC_Os05g28210	EMP1	8.739	323.988	37.072	5.212	0

Supplemental Table S9. Sugar sensitive genes αAmy3, αAmy8, OsMST3, OsMST4 and OsMST6 responded to the α-amylase treatment as expected. The ABA synthesis genes SDR1, AAO3, NCED2 and NCED3 showed low expression and no significant response to α-amylase treatment. ABA3 had low expression and showed modest ABA sensitivity. Highlighted rows are significantly differentially expressed. AAT – α-amylase treatment, NAAT – non-α-amylase treatment.

Name	Locus ID	AAT	NAAT	log2 fold	significant	Notes
αAmy3	LOC_Os08g36910	1.04931	228.25	7.76503	yes	Sugar sensitive
αAmy8	LOC_Os08g36900	18.9524	23.7028	0.322682	no	Not sugar sensitive
OsMST3	LOC_Os07g01560	66.5595	270.905	2.0250705	yes	Sugar sensitive
OsMST4	LOC_Os03g11900	138.255	248.364	0.8451244	yes	Sugar sensitive
OsMST6	LOC_Os07g37320	0.892603	169.69	7.5706672	yes	Sugar sensitive
ABA3	LOC_Os06g45860	3.44157	10.6575	1.63074	yes	Sugar sensitive
SDR1	LOC_Os03g59610	8.37165	8.41551	0.00754	no	Not sugar sensitive
AAO3	LOC_Os03g57680	0.227807	0.129184	-0.81839	no	Not sugar sensitive
NCED2	LOC_Os12g24800	0	0.009893	Inf	no	Not sugar sensitive
NCED3	LOC_Os03g44380	0	0.072043	inf	no	Not sugar sensitive

Supplemental Table S10. Number of occurrences of the ABREN in the upstream regions of ABA-inducible genes in Arabidopsis thaliana

A. thaliana **DNA sequence**	base pairs	expected	matches	matches/expected
all 5' UTR TAIR	3,699,306	107.33	226	2.106
promoter TAIR	27,378,224	794.38	672	0.846
ABA ind genes -promoter (Seki et al.)	102,000	2.96	0	0

ABA ind genes - 5utr (Seki et al.)	25,119	0.75	1	1.342
ABA ind genes - promoter - (Li et al.)	141,000	4.09	6	1.467
ABA ind genes - 5utr - (Li et al.)	17,608	0.52	1	1.914
ABA ind gene – promoter (Seki + Li)	243,000	4.6134	6	1.301

Supplemental Table S11. Methylation sensitive restriction enzymes that can cut the ABREN

Restriction Enzyme	**Cut site**	**Notes**	
BfuCI	N|GATCN NCTAG|N		Blocked by CpG methylation
DpnI	NGA^{m}|TCN NCT |A^{m}GN	Cuts only if A is methylated	Blocked by CpG methylation
DpnII	N|GATCN NCTAG|N	Blocked by dam methylation	
MboI	N|GATCN NCTAG|N	Blocked by dam methylation	Blocked by CpG methylation
Sau3AI	N|GATCN NCTAG|N		Blocked by CpG methylation

Supplemental Table S12

Description	**Primer**	**Oligo**
Oleosin Promoter	Forward	5’-CGCGAATTCCATGTCGGCGTCCACGCAGCAAC-3’
	Reverse	5’-GCGTCGACGGGTACTACTGATCGACCTCAAGAAAATG-3’
mutCE3-like	NA	5’-GCAGAGCTCACCCCTCGCCCACGGTGGATCCACCCCAGCCACACTCC-3’
mutABRE-like	NA	5’-GCTGCAGAGGCTGGCGCCTGCATGCCAAGCTTCCAAGCATCGCCTATC-3’
mutABREN in promoter	NA	5’-CTCGAATCATACTCGACCCATCCATGTCGACCCATCAGGATCTCGATC-3’
mutABREN in 5’ UTR	NA	5’ ATCTCCCTCCCCTGCATCCATCCATCCATGGCGCGCCCACAACATTTTCTTGAGGTCGA 3’

REFERENCES

Alexandrov, N.N., Brover, V.V., Freidin, S., Troukhan, M.E., Tatarinova, T.V., Zhang, H., Swaller, T.J., Lu, Y.P., Bouck, J., Flavell, R.B., et al. (2009). Insights into corn genes derived from large-scale cDNA sequencing. Plant Mol Biol 69:179-194.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol 215:403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Asano, T., Miyao, A., Hirochika, H., Kikuchi, S., and Kadowakia, K. (2013). A pentatricopeptide repeat gene of rice is required for splicing of chloroplast transcripts and RNA editing of ndhA. Plant Biotechnology 30:57-64.

Baginsky, S., Hennig, L., Zimmermann, P., and Gruissem, W. (2010). Gene expression analysis, proteomics, and network discovery. Plant Physiol 152:402-410.

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28-36.

Bartels, D., Singh, M., and Salamini, F. (1988). Onset of desiccation tolerance during development of the barley embryo. Planta 175:485-492.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science 306:2242-2246.

Bethke, P.C., Hwang, Y.S., Zhu, T., and Jones, R.L. (2006). Global patterns of gene expression in the aleurone of wild-type and dwarf1 mutant rice. Plant Physiol 140:484-498.

Bethke, P.C., Schuurink, R., and Jones, R.L. (1997). Hormonal signalling in cereal aleurone. J. Exp. Bot. 48:1137-1356.

Bewley, J.D. (1997). Seed Germination and Dormancy. Plant Cell 9:1055-1066.

Blythe, M.J., Kao, D., Malla, S., Rowsell, J., Wilson, R., Evans, D., Jowett, J., Hall, A., Lemay, V., Lam, S., et al. (2010). A dual platform approach to transcript discovery for the planarian Schmidtea mediterranea to establish RNAseq for stem cell and regeneration biology. PLoS One 5:e15617.

Bonizzoni, M., Afrane, Y., Dunn, W.A., Atieli, F.K., Zhou, G., Zhong, D., Li, J., Githeko, A., and Yan, G. (2012). Comparative transcriptome analyses of deltamethrin-resistant and -susceptible Anopheles gambiae mosquitoes from Kenya by RNA-Seq. PLoS One 7:e44607.

Bright, J., Desikan, R., Hancock, J.T., Weir, I.S., and Neill, S.J. (2006). ABA-induced NO generation and stomatal closure in Arabidopsis are dependent on H2O2 synthesis. Plant J 45:113-122.

Brown, T.A. (2002). Genomes, 2nd edition: Oxford: Wiley-Liss.

Bruce, W.B., Christensen, A.H., Klein, T., Fromm, M., and Quail, P.H. (1989). Photoregulation of a phytochrome gene promoter from oat transferred into rice by particle bombardment. Proc Natl Acad Sci U S A 86:9692-9696.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78-94.

Campbell, M.A., Zhu, W., Jiang, N., Lin, H., Ouyang, S., Childs, K.L., Haas, B.J., Hamilton, J.P., and Buell, C.R. (2007). Identification and characterization of lineage-specific genes within the Poaceae. Plant Physiol 145:1311-1322.

Cao, P., Jung, K.H., Choi, D., Hwang, D., Zhu, J., and Ronald, P.C. (2012). The Rice Oligonucleotide Array Database: an atlas of rice gene expression. Rice (N Y) 5:17.

Chen, P.W., Chiang, C.M., Tseng, T.H., and Yu, S.M. (2006). Interaction between rice MYBGA and the gibberellin response element controls tissue-specific sugar sensitivity of alpha-amylase genes. Plant Cell 18:2326-2340.

Cheng, W.H., Endo, A., Zhou, L., Penney, J., Chen, H.C., Arroyo, A., Leon, P., Nambara, E., Asami, T., Seo, M., et al. (2002). A unique short-chain dehydrogenase/reductase in Arabidopsis glucose signaling and abscisic acid biosynthesis and functions. Plant Cell 14:2723-2743.

Chomczynski, P., and Sacchi, N. (2006). The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. Nat Protoc 1:581-585.

Chung, H.J., Fu, H.Y., and Thomas, T.L. (2005). Abscisic acid-inducible nuclear proteins bind to bipartite promoter elements required for ABA response and embryo-regulated expression of the carrot Dc3 gene. Planta 220:424-433.

Dai, X., and Zhao, P.X. (2011). psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 39:W155-159.

Dekkers, B.J., Schuurmans, J.A., and Smeekens, S.C. (2008). Interaction between sugar and abscisic acid signalling during early seedling development in Arabidopsis. Plant Mol Biol 67:151-167.

Deveau, H., Garneau, J.E., and Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64:475-493.

Endo, A., Okamoto, M., and Koshiba, T. (2012). ABA biosynthetic and catabolic pathways. In: Abscisic Acid: Metabolism, Transport and Signaling--Zhang, D., ed.: Springer Netherlands. 21-45.

Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C., Dwight, S.S., Hitz, B.C., Karra, K., Nash, R.S., et al. (2014). The reference genome sequence of *Saccharomyces cerevisiae*: then and now. G3: Genes, Genomes, Genetics 4:389-398.

Eulgem, T., Rushton, P.J., Robatzek, S., and Somssich, I.E. (2000). The WRKY superfamily of plant transcription factors. Trends Plant Sci 5:199-206.

Eversole, K., Feuillet, C., Mayer, K.F., and Rogers, J. (2014). Slicing the wheat genome. Introduction. Science 345:285-287.

Fang, J., and Chu, C. (2008). Abscisic acid and the pre-harvest sprouting in cereals. Plant Signal Behav 3:1046-1048.

Fawal, N., Savelli, B., Dunand, C., and Mathe, C. (2012). GECA: a fast tool for gene evolution and conservation analysis in eukaryotic protein families. Bioinformatics 28:1398-1399.

Finkelstein, R. (2013). Abscisic Acid synthesis and response. Arabidopsis Book 11:e0166.

Finkelstein, R.R., Gampala, S.S., and Rock, C.D. (2002). Abscisic acid signaling in seeds and seedlings. Plant Cell 14 Suppl:S15-45.

Finkelstein, R.R., and Rock, C.D. (2002). Abscisic acid biosynthesis and response. In: The Arabidopsis Book. e0058.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res 42:D222-230.

Frech, C., Choo, C., and Chen, N. (2012). FeatureStack: Perl module for comparative visualization of gene features. Bioinformatics 28:3137-3138.

Gelvin, S.B. (2000). Agrobacterium and Plant Genes Involved in T-DNA Transfer and Integration. Annu Rev Plant Physiol Plant Mol Biol 51:223-256.

Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. Proc Natl Acad Sci U S A 99:7821-7826.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296:92-100.

Gomez-Cadenas, A., Zentella, R., Walker-Simmons, M.K., and Ho, T.H. (2001). Gibberellin/abscisic acid antagonism in barley aleurone cells: site of action of the protein kinase PKABA1 in relation to gibberellin signaling molecules. Plant Cell 13:667-679.

Gomez-Porras, J.L., Riano-Pachon, D.M., Dreyer, I., Mayer, J.E., and Mueller-Roeber, B. (2007). Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in Arabidopsis and rice. BMC Genomics 8:260.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., et al. (2011). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 40:D1178-D1186.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644-652.

Guan, S., and Lu, Y. (2013). Dissecting organ-specific transcriptomes through RNA-sequencing. Plant Methods 9:42.

Gulledge, A.A., Roberts, A.D., Vora, H., Patel, K., and Loraine, A.E. (2012). Mining Arabidopsis thaliana RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. Am J Bot 99:219-231.

Gutierrez, F.R.S., Güiza, M.L.T., and Magally del Carmen, M.E. (2013). Prevalence of *Trypanosoma cruzi* infection among people aged 15 to 89 years inhabiting the Department of Casanare (Colombia). PLOS Neglected Tropical Diseases 7:e2113.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature 431:99-104.

Harris, T.W., Baran, J., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., Done, J., Grove, C., Howe, K., et al. (2013). WormBase 2014: new views of curated biology. Nucleic Acids Research 42:D789-D793.

Hattori, T., Totsuka, M., Hobo, T., Kagaya, Y., and Yamamoto-Toyoda, A. (2002). Experimentally determined sequence requirement of ACGT-containing abscisic acid response element. Plant Cell Physiol 43:136-140.

Hedden, P., and Thomas, S.G. (2012). Gibberellin biosynthesis and its regulation. Biochem J 444:11-25.

Higo, K., and Ugawa Y, I.M., Korenaga T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 27:297-300.

Hilbricht, T., Salamini, F., and Bartels, D. (2002). CpR18, a novel SAP-domain plant transcription factor, binds to a promoter region necessary for ABA mediated expression of the CDeT27-45 gene from the resurrection plant Craterostigma plantagineum Hochst. Plant J 31:293-303.

Hobo, T., Asada, M., Kowyama, Y., and Hattori, T. (1999). ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. Plant J 19:679-689.

Hoth, S., Morgante, M., Sanchez, J.P., Hanafey, M.K., Tingey, S.V., and Chua, N.H. (2002). Genome-wide gene expression profiling in Arabidopsis thaliana reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant. J Cell Sci 115:4891-4900.

Hu, C.D., Chinenov, Y., and Kerppola, T.K. (2002). Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. Mol Cell 9:789-798.

Ingram, J., and Bartels, D. (1996). The Molecular Basis of Dehydration Tolerance in Plants. Annu Rev Plant Physiol Plant Mol Biol 47:377-403.

Iwasaki, T., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1995). Identification of a cis-regulatory region of a gene in Arabidopsis thaliana whose induction by dehydration is mediated by abscisic acid and requires protein synthesis. Mol Gen Genet 247:391-398.

Izawa, T., Foster, R., and Chua, N.H. (1993). Plant bZIP protein DNA binding specificity. J Mol Biol 230:1131-1144.

Joshee, N., Kisaka, H., and Kitagawa, Y. (1998). Isolation and characterization of a water stress-specific genomic gene, pwsi 18, from rice. Plant Cell Physiol 39:64-72.

Kao, C.Y., Cocciolone, S.M., Vasil, I.K., and McCarty, D.R. (1996). Localization and interaction of the cis-acting elements for abscisic acid, VIVIPAROUS1, and light activation of the C1 gene of maize. Plant Cell 8:1171-1179.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) 6:4.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. 12:996-1006.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11:345-355.

Khan, A.A., and Downing, R.D. (1968). Cytokinin reversal of abscisic acid inhibition of growth and alpha-amylase synthesis in barley seed. Physiologia Plantarum 21:1301-1307.

Kim, H., Hwang, H., Hong, J.W., Lee, Y.N., Ahn, I.P., Yoon, I.S., Yoo, S.D., Lee, S., Lee, S.C., and Kim, B.G. (2012). A rice orthologue of the ABA receptor, OsPYL/RCAR5, is a positive regulator of the ABA signal transduction pathway in seed germination and early seedling growth. J Exp Bot 63:1013-1024.

Kizis, D., and Pages, M. (2002). Maize DRE-binding proteins DBF1 and DBF2 are involved in rab17 regulation through the drought-responsive element in an ABA-dependent pathway. Plant J 30:679-689.

Kobayashi, Y., Yamamoto, S., Minami, H., Kagaya, Y., and Hattori, T. (2004). Differential activation of the rice sucrose nonfermenting1-related protein kinase2 family by hyperosmotic stress and abscisic acid. Plant Cell 16:1163-1177.

Koornneef, M., and Reuling G, K.C. (1984). The isolation and characterization of abscisic-acid insensitive mutants of Arabidopsis thaliana. Physiol. Plant. 61:377-383.

Kumar, R., Moseley, B., Vassilvitskii, S., and Vattani, A. (2015). Fast Greedy Algorithms in MapReduce and Streaming. ACM Transactions on Parallel Computing - Special Issue for SPAA 2013 2:14:11-22.

Kunkel, T.A. (1985). Rapid and efficient site-specific mutagenesis without phenotypic selection. Proc Natl Acad Sci U S A 82:488-492.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. Nucleic Acids Research 39:D19-21.

Leprince, O., Hendry, G.A.F., and McKersie, B.D. (1993). The mechanisms of desiccation tolerance in developing seeds. Seed Science Research 3:231-246.

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P., and Rombauts, S. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30:325-327.

Li, Y., Lee, K.K., Walsh, S., Smith, C., Hadingham, S., Sorefan, K., Cawley, G., and Bevan, M.W. (2006). Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine. Genome Res 16:414-427.

Li, Z., and Trick, H.N. (2005). Rapid method for high-quality RNA isolation from seed endospern containing high levels of starch. Biotechniques 38:872-876.

Lionikas, A., Meharg, C., Derry, J.M., Ratkevicius, A., Carroll, A.M., Vandenbergh, D.J., and Blizard, D.A. (2012). Resolving candidate genes of mouse skeletal muscle QTL via RNA-Seq and expression network analyses. BMC Genomics 13:592.

Liu, X., Brutlag, D., and Liu, J. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput. 6:127-138.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., and Pallen, M.J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30:434-439.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res 33:6494-6506.

Loraine, A.E., McCormick, S., Estrada, A., Patel, K., and Qin, P. (2013). RNA-Seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. Plant Physiology 162:1092-1109.

Lu, B., Zeng, Z., and Shi, T. (2013a). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Science China Life Sciences 56:143-155.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., Feng, Q., Zhao, Y., Guo, Y., Li, W., et al. (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res 20:1238-1249.

Lu, X., Chen, D., Shu, D., Zhang, Z., Wang, W., Klukas, C., Chen, L.L., Fan, Y., Chen, M., and Zhang, C. (2013b). The differential transcription network between embryo and endosperm in the early developing maize seed. Plant Physiol 162:440-455.

Ma, Y., Szostkiewicz, I., Korte, A., Moes, D., Yang, Y., Christmann, A., and Grill, E. (2009). Regulators of PP2C phosphatase activity function as abscisic acid sensors. Science 324:1064-1068.

Makashir, S.B., Kottyan, L.C., and Weirauch, M.T. (2015). Meta-analysis of differential gene co-expression: application to lupus. Pac Symp Biocomput:443-454.

Mao, S., Souza, A.L., Goodrich, R.J., and Krawetz, S.A. (2012). Identification of artifactual microarray probe signals constantly present in multiple sample types. Biotechniques 53:91-98.

Marcotte, W.R., Jr., Russell, S.H., and Quatrano, R.S. (1989). Abscisic acid-responsive sequences from the em gene of wheat. Plant Cell 1:969-976.

Marth, P.C., Audia, W.V., and Mitchell, J.W. (1956). Effects of Gibberellic Acid on Growth and Development of Plants of Various Genera and Species Botanical Gazette 118:106-111.

Mauch-Mani, B., and Mauch, F. (2005). The role of abscisic acid in plant-pathogen interactions. Current Opinion in Plant Biology 2005:4.

McWha, J.A., and Langer, H.J. (1979). The effects of exogenously applied abscisic acid on bud burst in Salix spp. Ann. Bot. 44:47-55.

Miyawaki, K.N., and Yang, Z. (2014). Extracellular signals and receptor-like kinases regulating ROP GTPases in plants. Front Plant Sci 5:449.

Mizuno, H., Kawahara, Y., Sakai, H., Kanamori, H., Wakimoto, H., Yamagata, H., Oono, Y., Wu, J., Ikawa, H., Itoh, T., et al. (2010). Massive parallel sequencing of mRNA in identification of unannotated salinity stress-inducible transcripts in rice (Oryza sativa L.). BMC Genomics 11:683.

Mohanty, S., Wassmann, R., Nelson, A., Moya, P., and Jagadish, S.V.K. (2013). Rice and climate change: significance for food security and vulnerability IRRI Discussion Paper Series 49:14.

Moons, A., De Keyser, A., and Van Montagu, M. (1997). A group 3 LEA cDNA of rice, responsive to abscisic acid, but not to jasmonic acid, shows variety-specific differences in salt stress response. Gene 191:197-204.

Morris, P.C., Kumar, A., Bowles, D.J., and Cuming, A.C. (1990). Osmotic stress and abscisic acid induce expression of the wheat Em genes. Eur J Biochem 190:625-630.

Mundy, J., and Chua, N.H. (1988). Abscisic acid and water-stress induce the expression of a novel rice gene. EMBO J 7:2279-2286.

Murase, K., Hirano, Y., Sun, T.P., and Hakoshima, T. (2008). Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. Nature 456:459-463.

Nagalakshmi, U., Waern, K., and Snyder, M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol Chapter 4:Unit 4 11 11-13.

Nakashima, K., and Yasunari Fujita, K.K., Kyonoshin Maruyama, Yoshihiro Narusaka, Motoaki Seki, Kauzo Shinozaki, and Kazuko Yamaguchi-Shinozaki. (2006). Transcriptional regulation of ABI3- and ABA-responsive genes including RD29B and RD29A in seeds, germinating embryos, and seedlings of Arabidopsis. Plant Molecular Biology 60:51-68.

Narusaka, Y., Nakashima, K., Shinwari, Z.K., Sakuma, Y., Furihata, T., Abe, H., Narusaka, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. Plant J 34:137-148.

Nijhawan, A., Jain, M., Tyagi, A.K., and Khurana, J.P. (2008). Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. Plant Physiol 146:333-350.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997). Evolution of genetic redundancy. Nature 388:167-171.

Obata, T. (1975). Gibberellic acid-induced secretion of hydrolases in barley aleurone layers. Plant and Cell Physiology 17:63-71.

Ohkawa, H., Kamada, H., Sudo, H., and Harada, H. (1989). Effects of Gibberellic Acid on Hairy Root Growth in Datura innoxia. Journal of Plant Physiology 134:633-636.

Olszewski, N., Sun, T.P., and Gubler, F. (2002). Gibberellin signaling: biosynthesis, catabolism, and response pathways. Plant Cell 14 Suppl:S61-80.

Oshlack, A., Robinson, M.D., and Young, M.D. (2010). From RNA-seq reads to differential expression results. Genome Biol 11:220.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35:D883-887.

Pages, M., Vilardell, J., Jensen, A.B., Albà, M.M., Torrent, M., and Goday, A. (1993). Molecular biological responses to drought in maize. In: Interacting stresses on plants in a changing climate--Jackson, M.B., and Black, C.R., eds.: Berlin/Heidelberg: Springer-Verlag. 583–591.

Palmieri, N., Nolte, V., Suvorov, A., Kosiol, C., and Schlotterer, C. (2012). Evaluation of different reference based annotation strategies using RNA-Seq - a case study in Drososphila pseudoobscura. PLoS One 7:e46415.

Pandey, S., Nelson, D.C., and Assmann, S.M. (2009). Two novel GPCR-type G proteins are abscisic acid receptors in Arabidopsis. Cell 136:136-148.

Park, S.Y., Fung, P., Nishimura, N., Jensen, D.R., Fujii, H., Zhao, Y., Lumba, S., Santiago, J., Rodrigues, A., Chow, T.F., et al. (2009). Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. Science 324:1068-1071.

Pons, P., and Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In: Computer and Information Sciences - ISCIS 2005--Yolum, P., Güngör, T., Gürgen, F., Özturan, C., ed.: Springer. 284-293.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341.

Quinn, E.M., Cormican, P., Kenny, E.M., Hill, M., Anney, R., Gill, M., Corvin, A.P., and Morris, D.W. (2013). Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PLoS One 8:e58815.

Raghavendra, A.S., Gonugunta, V.K., Christmann, A., and Grill, E. (2010). ABA perception and signalling. Trends Plant Sci 15:395-401.

Rambaldi, D., and Ciccarelli, F.D. (2009). FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. Bioinformatics 25:2281-2282.

Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13:278-289.

Rice Full-Length c, D.N.A.C., National Institute of Agrobiological Sciences Rice Full-Length c, D.N.A.P.T., Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., et al. (2003). Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science 301:376-379.

Richa, K., Tiwari, I.M., Kumari, M., Devanna, B.N., Sonah, H., Kumari, A., Nagar, R., Sharma, V., Botella, J.R., and Sharma, T.R. (2016). Functional Characterization of Novel Chitinase Genes Present in the Sheath Blight Resistance QTL: qSBR11-1 in Rice Line Tetep. Front Plant Sci 7:244.

Richard, G.F., Kerrest, A., and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686-727.

Richard, P., and Manley, J.L. (2009). Transcription termination by nuclear RNA polymerases. Genes & development 23:1247-1269.

Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27:2325-2329.

Ross, C., and Shen, Q.J. (2006). Computational prediction and experimental verification of HVA1-like abscisic acid responsive promoters in rice (Oryza sativa). Plant Mol Biol 62:233-246.

Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516-522.

Salipante, S.J., Kawashima, T., Rosenthal, C., Hoogestraat, D.R., Cummings, L.A., Sengupta, D.J., Harkins, T.T., Cookson, B.T., and Hoffman, N.G. (2014). Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol 80:7583-7591.

Sander, J.D., and Joung, J.K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 32:347-355.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:e37.

Schroeder, J.I., Kwak, J.M., and Allen, G.J. (2001). Guard cell abscisic acid signalling and engineering drought hardiness in plants. Nature 410:327-330.

Seki, M., Ishida, J., Narusaka, M., Fujita, M., Nanjo, T., Umezawa, T., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., et al. (2002). Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray. Funct Integr Genomics 2:282-291.

Seo, M., and Koshiba, T. (2002). Complex regulation of ABA biosynthesis in plants. Trends Plant Sci 7:41-48.

Sharp, R.E., and LeNoble, M.E. (2002). ABA, ethylene and the control of shoot and root growth under water stress. J Exp Bot 53:33-37.

Sheard, L.B., and Zheng, N. (2009). Plant biology: Signal advance for abscisic acid. Nature 462:575-576.

Shen, Q., and Ho, T.H. (1995). Functional dissection of an abscisic acid (ABA)-inducible gene reveals two independent ABA-responsive complexes each containing a G-box and a novel cis-acting element. Plant Cell 7:295-307.

Shen, Q., Uknes, S.J., and Ho, T.H. (1993). Hormone response complex in a novel abscisic acid and cycloheximide-inducible barley gene. J Biol Chem 268:23652-23660.

Shen, Q., Zhang, P., and Ho, T.H.D. (1996). Modular nature of abscisic acid (ABA) response complexes: composite promoter units that are necessary and sufficient for ABA induction of gene expression in barley. Plant Cell 8:1107-1119.

Shen, Q.J., Casaretto, J.A., Zhang, P., and Ho, T.H. (2004). Functional definition of ABA-response complexes: the promoter units necessary and sufficient for ABA induction of gene expression in barley (Hordeum vulgare L.). Plant Mol Biol 54:111-124.

Singh, H., LeBowitz, J.H., Baldwin, A.S., Jr., and Sharp, P.A. (1988). Molecular cloning of an enhancer binding protein: isolation by screening of an expression library with a recognition site DNA. Cell 52:415-423.

Slamet-Loedin, I.H., Chadha-Mohanty, P., and Torrizo, L. (2014). Agrobacterium-mediated transformation: rice transformation. Methods Mol Biol 1099:261-271.

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Smith, H. (1977). The Molecular biology of plant cells: Berkeley : University of California Press, 1977.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13:613-626.

Srivastava, L.M. (2002). Abscisic Acid Signal Perception and Transduction. In: Plant Growth and Development: Hormones and Environment CA: Academic Press. 569-588.

St. Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P., and Consortium, F. (2014). FlyBase 102--advanced approaches to interrogating FlyBase. Nucleic Acids Research 42:D780-D788.

Steijger, T., Abril, J.F., Engstrom, P.G., Kokocinski, F., Hubbard, T.J., Guigo, R., Harrow, J., and Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. Nature Methods 10:1177-1184.

Stern, D.B., Goldschmidt-Clermont, M., and Hanson, M.R. (2010). Chloroplast RNA metabolism. Annu Rev Plant Biol 61:125-155.

Stripecke, R., Oliveira, C.C., McCarthy, J.E., and Hentze, M.W. (1994). Proteins binding to 5' untranslated region sites: a general mechanism for translational regulation of mRNAs in human and yeast cells. Mol Cell Biol 14:5898-5909.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science 302:249-255.

Sun, Y., Zhang, X., Wu, C., He, Y., Ma, Y., Hou, H., Guo, X., Du, W., Zhao, Y., and Xia, L. (2016). Engineering Herbicide-Resistant Rice Plants through CRISPR/Cas9-Mediated Homologous Recombination of Acetolactate Synthase. Mol Plant.

Syed, N.H., Kalyna, M., Marquez, Y., Barta, A., and Brown, J.W. (2012). Alternative splicing in plants--coming of age. Trends Plant Sci 17:616-623.

Takaiwa, F., Oono, K., Iida, Y., and Sugiura, M. (1985). The complete nucleotide sequence of a rice 25S.rRNA gene. Gene 37:255-259.

Takaiwa, F., Oono, K., and Sugiura, M. (1984). The complete nucleotide sequence of a rice 17S rRNA gene. Nucleic Acids Res 12:5441-5448.

Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., and Li, J.B. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Res 23:201-216.

Tanaka, Y., Sano, T., Tamaoki, M., Nakajima, N., Kondo, N., and Hasezawa, S. (2006). Cytokinin and auxin inhibit abscisic acid-induced stomatal closure by enhancing ethylene production in Arabidopsis. Journal of Experimental Botany 57:2259-2266.

Thiel, T., Graner, A., Waugh, R., Grosse, I., Close, T.J., and Stein, N. (2009). Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. BMC Evol Biol 9:209.

Thomas, S.G., and Sun, T.P. (2004). Update on gibberellin signaling. A tale of the tall and the short. Plant Physiol 135:668-676.

Toung, J.M., Morley, M., Li, M., and Cheung, V.G. (2011). RNA-sequence analysis of human B-cells. Genome Res 21:991-998.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511-515.

Tzen, J.T., and Huang, A.H. (1992). Surface structure and properties of plant seed oil bodies. J. Cell Biol. 117:327-335.

Ueguchi-Tanaka, M., Ashikari, M., Nakajima, M., Itoh, H., Katoh, E., Kobayashi, M., Chow, T.Y., Hsing, Y.I., Kitano, H., Yamaguchi, I., et al. (2005). GIBBERELLIN INSENSITIVE DWARF1 encodes a soluble receptor for gibberellin. Nature 437:693-698.

Ulitsky, I., and Bartel, D.P. (2013). lincRNAs: genomics, evolution, and mechanisms. Cell 154:26-46.

Umezawa, T., Sakurai, T., Totoki, Y., Toyoda, A., Seki, M., Ishiwata, A., Akiyama, K., Kurotani, A., Yoshida, T., Mochida, K., et al. (2008). Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. DNA Res 15:333-346.

Van Verk, M.C., Hickman, R., Pieterse, C.M., and Van Wees, S.C. (2013). RNA-Seq: revelation of the messengers. Trends Plant Sci 18:175-179.

Wang, H.J., Wan, A.R., Hsu, C.M., Lee, K.W., Yu, S.M., and Jauh, G.Y. (2007). Transcriptomic adaptations in rice suspension cells under sucrose starvation. Plant Mol Biol 63:441-463.

Wang, M., Oppedijk, B.J., Lu, X., Van Duijn, B., and Schilperoort, R.A. (1996). Apoptosis in barley aleurone during germination and its inhibition by abscisic acid. Plant Mol Biol 32:1125-1134.

Wang, X., Yu, Z., Yang, X., Deng, X.W., and Li, L. (2009a). Transcriptionally active gene fragments derived from potentially fast-evolving donor genes in the rice genome. Bioinformatics 25:1215-1218.

Wang, Y., You, F.M., Lazo, G.R., Luo, M.C., Thilmony, R., Gordon, S., Kianian, S.F., and Gu, Y.Q. (2013). PIECE: a database for plant gene structure comparison and evolution. Nucleic Acids Res 41:D1159-1166.

Wang, Z., Gerstein, M., and Snyder, M. (2009b). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63.

Watanabe, K.A., Homayouni, A., Tufano, T., Lopez, J., Ringler, P., Rushton, P., and Shen, Q.J. (2015). Tiling Assembly: a new tool for reference annotation-independent transcript assembly and novel gene identification by RNA-sequencing. DNA Res 22:319-329.

Watanabe, K.A., Ringler, P., Gu, L., and Shen, Q.J. (2014). RNA-sequencing reveals previously unannotated protein- and microRNA-coding genes expressed in aleurone cells of rice seeds. Genomics 103:122-134.

Williams, M.E., Foster, R., and Chua, N.H. (1992). Sequences flanking the hexameric G-box core CACGTG affect the specificity of protein binding. Plant Cell 4:485-496.

Wu, J., Anczukow, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic Acids Res. 41:5149-5163.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434:338-345.

Xue, T., Wang, D., Zhang, S., Ehlting, J., Ni, F., Jakab, S., Zheng, C., and Zhong, Y. (2008). Genome-wide and expression analysis of protein phosphatase 2C in rice and Arabidopsis. BMC Genomics 9:550.

Yamaguchi-Shinozaki, K., Mundy, J., and Chua, N.H. (1990). Four tightly linked rab genes are differentially expressed in rice. Plant Mol Biol 14:29-39.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13:329-342.

Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G., and Chen, L.L. (2011a). Genomewide characterization of non-polyadenylated RNAs. Genome Biol 12:R16.

Yang, X., Tschaplinski, T.J., Hurst, G.B., Jawdy, S., Abraham, P.E., Lankford, P.K., Adams, R.M., Shah, M.B., Hettich, R.L., Lindquist, E., et al. (2011c). Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res 21:634-641.

Ye, N., Jia, L., and Zhang, J. (2012). ABA signal in rice under stress conditions. Rice (N Y) 5:1.

Yoshida, T., Mogami, J., and Yamaguchi-Shinozaki, K. (2015). Omics approaches toward defining the comprehensive abscisic acid signaling network in plants. Plant Cell Physiol 56:1043-1052.

Young, K.H. (1998). Yeast two-hybrid: so many interactions, (in) so little time. Biol Reprod 58:302-311.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296:79-92.

Zeeman, S. (2002). Carbohydrate Metabolism--Buchanan, B.B., Gruissem, W., and Jones, R.L., eds. Biochemistry and Molecular Biology of Plants: Wiley Blackwell.

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., et al. (2010a). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 20:646-654.

Zhang, J., Jia, W., Yang, J., and Ismail, A.M. (2006). Role of ABA in integrating plant responses to drought and salt stresses. Field Crops Res. 97:111-119.

Zhang, X., and Cai, X. (2011). Climate change impacts on global agricultural land availability.

Environmental Research Letters 6.

Zhang, Y., and Wang, L. (2005). The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. BMC Evol Biol 5:1.

Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., Wang, T., Ling, Y., and Su, Z. (2010c). PMRD: plant microRNA database. Nucleic Acids Res 38:D806-813.

Zhang, Z.L., Shin, M., Zou, X., Huang, J., Ho, T.H., and Shen, Q.J. (2009). A negative regulator encoded by a rice WRKY gene represses both abscisic acid and gibberellins signaling in aleurone cells. Plant Mol Biol 70:139-151.

Zhang, Z.L., Xie, Z., Zou, X., Casaretto, J., Ho, T.H., and Shen, Q.J. (2004). A rice WRKY gene encodes a transcriptional repressor of the gibberellin signaling pathway in aleurone cells. Plant Physiol 134:1500-1513.

Zou, J., Abrams, G.D., Barton, D.L., Taylor, D.C., Pomeroy, M.K., and Abrams, S.R. (1995). Induction of lipid and oleosin biosynthesis by (+)-abscisic acid and Its metabolites in microspore-derived embryos of Brassica napus L.cv Reston (biological responses in the presence of 8[prime]-hydroxyabscisic acid). Plant Physiol. 108:563-571.

CURRICULUM VITAE

Kenneth A. Watanabe

2115 Pilar Ave.
North Las Vegas, NV 89032
702-234-6034
watana43@unlv.nevada.edu

Goal:

To pursue a career in academia that involves research, bioinformatics and education of students.

Education:

Bachelor's degree, Chemical Engineering from Cornell University, Ithaca NY
Master's Degree, Computer Science from Cornell University, Ithaca NY
PhD, Bioinformatics/Molecular and Cell Biology from the University of Nevada Las Vegas, NV

Teaching Experience:

Teaching Assistant for Intro to Biology (Bio196) (2012-2014).
Duties include short lecture explaining the purpose of the lab; preparation, administration and grading of quizzes; instruction of use of lab equipment, lab techniques and experiments.

Research Experience

- January 2010 to present: Dr. Shen's lab at the University of Nevada.
- January 2009 to December 2009: Dr. Abel Santo's lab at the University of Nevada
- January 2008 to December 2008: Dr. Fiscus at the Nevada Cancer Institute

Programming Languages

My knowledge of various programming languages, software and operating systems is quite extensive. Below is a *limited* list and extent of my knowledge with programming languages and software.

- Very strong programming knowledge of PERL and PERL/Tk. PERL was the primary language used for development of bioinformatics tools for transcriptome analysis.
- Strong knowledge of the LINUX operating system. All the servers at the UNLV Shen Lab were running on the Fedora LINUX operating system.
- Strong programming knowledge in Python and to a lesser extent C+ and R.
- Strong knowledge of SQL database programming and relational database concepts.
- Strong knowledge of HTML, CSS , DHTML, Javascript and PHP. I was the primary developer of the online Transcript Structure Display webpage and helped develop our online WRKY Gene database.
- Strong knowledge of RNA-sequencing concepts and data analysis including the Tuxedo Suite data analysis software: Bowtie, Tophat, Cufflinks, Cuffdiff as well as custom

written software for transcript assembly. I have knowledge of the Trinity software for de novo transcript assembly.

Laboratory procedures: This includes but is not limited to the following:

- DNA/RNA extraction.
- Polymerase Chain Reaction (PCR) for DNA amplification.
- Primer design for PCR or site-directed mutagenesis.
- Single-stranded DNA (ssDNA) extraction via VCS m13-K07 helper phage.
- ssDNA mutagenesis via oligonucleotide base-paring.
- Bacterial transformation via heat shock and electroporation.
- Plasmid preparation such as Qiagen Miniprep, Midiprep and Maxiprep
- Restriction digest of DNA via NEB restriction enzymes
- Ligation of PCR-amplified DNA fragments into plasmids.
- Reverse Transcription PCR (RT-PCR)
- Agarose gel electrophoresis
- Particle bombardment of plasmids into plant cell tissues via a gene gun
- Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE) followed by Western Blotting

Laboratory equipment: This includes but is not limited to the following:

Tecan fluorometer, gene gun, PCR machine, optical density, luminometer, flurometer, sonicators, autoclaves, high speed large and small volume centrifuges

Publications:

Watanabe, **K.A.**, Ma K., Homayouni, A., Shen, Q. J.: Transcript Structure and Domain Display: A Customizable Transcript Visualization Tool. Submitted to Bioinformatics on Oct 1, 2015, in press

Watanabe, K.A., Homayouni, A., Tufano, T., Lopez, J., Ringler, P., Rushton, P. and **Shen, Q.J.:** Tiling Assembly: a new tool for reference annotation independent transcript assembly and novel gene identification by RNA-sequencing, DNA Research, 2015, in press

Watanabe, K.A., Ringler P., Gu L., Shen, Q.J.: RNA-sequencing reveals previously unannotated protein- and microRNA-coding genes expressed in aleurone cells of rice seeds Volume 103, Issue 1, January 2014, Pages 122–134

Cheng, J., Anastasi, J., **Watanabe, K.A.**, **Shen, Q.J.**, and Vardiman, J.: A novel epigenetic mechanism involving hyper-trimethylation of histone H3 lysine 27 underlying the Myelodysplastic Syndrome. Leukemia, 2013, 27: 1291-1300.

Submitted publication under review:

Watanabe, K.A., Homayouni, A., Gu, L., Ringler P., Huang KY., Hong YF., Ho THD., and Shen, Q.J.*In Silico* Identification and Experimental Verification of a Novel Abscisic Acid Response Element in Rice