RNA-SEQ DRIVEN ANALYSIS OF SEASONAL GENE EXPRESSION VARIATION IN
MISCANTHUS × GIGANTEUS

BY

ADAM ROBERT BARLING

DISSERTATION
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Crop Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:
Professor Stephen P. Moose, Chair
Associate Professor Matthew Hudson
Assistant Professor Patrick J. Brown
Assistant Professor Carl Bernacchi
ABSTRACT

Miscanthus × giganteus is a C4 grass that has generated a large amount of interest as a potential biofuel crop due to its high level of biomass production, its perenniality, and its sterility. In this study, two separate M. × giganteus RNA-Seq datasets were generated to help explore the characteristics of M. × giganteus at the level of gene expression: a ten-tissue dataset suitable for examining genes with tissue-preferred expression, and a twenty-four sample dataset for examining the changes in gene expression that occur over the growing season. Aided by these datasets, aspects and potential mediators of M. × giganteus’ seasonal developmental cycle and changes in the utilization, storage, and long-distance mobilization and remobilization of the essential nutrient nitrogen have been studied. These RNA-Seq datasets have been verified with RT-qPCR and compared to amino acid and elemental concentration profiles; as a result, many seasonal changes in gene expression corresponding to the growth and development of M. × giganteus have been documented in order to better define the traits that make this crop such an outstanding biofuel candidate.
ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Professor Stephen P. Moose for his years of guidance and dedication to helping me reach this milestone. I would also like to thank my committee members, Professor Matthew Hudson, Professor Patrick Brown, and Professor Carl Bernacchi, for their knowledge, input, and understanding. I am very thankful to work with such great fellow lab-mates, and would like to thank both Dr. Kankshita Swaminathan and Brandon James in particular for their support and the aid they gave me in experimental design, troubleshooting procedures, and problem solving. I would also like to thank Dianne Carson for her help, motivation, and friendly demeanor in working through what can be a somewhat unfriendly, pedantic, and bureaucratic process. I am indebted to Cody Postin and Mike Masters, who helped me complete the amino acid and elemental analysis of the seasonal dataset, respectively, and thank them for their help in doing so. Finally, I would like to thank my parents, Bruce and Beth, as well as my brothers, Phil and Matt, my wife Jing, and the rest of my family for their endless love and support throughout the years. Without you, none of this would have been possible.
# TABLE OF CONTENTS

CHAPTER ONE “BACKGROUND: THE BIOFUEL INDUSTRY” ........................... 1

CHAPTER TWO “A TEN-TISSUE RNA-SEQ STUDY OF *MISCANTHUS × GIGANTEUS*” ........................................................................................................... 17

CHAPTER THREE “CONSTRUCTION AND UTILIZATION OF A SEASONAL *MISCANTHUS × GIGANTEUS* RNA-SEQ DATASET” .................. 31

CHAPTER FOUR “ANALYSIS OF SEASONAL NITROGEN MOBILIZATION AND REMOBILIZATION IN *MISCANTHUS × GIGANTEUS*” ........................................................................................................... 54

CHAPTER FIVE “TABLES AND FIGURES” .................................................. 79

REFERENCES .................................................................................................................. 121
1.1 Biofuels

In 2011, the United States consumed approximately 19,000 barrels of oil every day (USEIA, 2013). This level of consumption has far-reaching social, economic, and environmental impacts for both the United States and the greater world beyond. Perhaps most notably are the detrimental environmental effects that the use of petroleum-based fuel has had on the Earth’s atmosphere, as a large amount of greenhouse gases currently present in the atmosphere are a direct result of the world’s petroleum fuel use.

The proliferation of greenhouse gases within the Earth’s atmosphere has intensified the Earth’s “greenhouse effect.” As a result, a greater amount of the sun’s infrared radiation remains trapped within the atmosphere, which has the potential to slowly alter temperatures across the globe (IPCC, 2007). Altered global temperatures have been hypothesized to be able to cause a number of detrimental effects in nearly every biome present on Earth—polar ice caps could melt, raising sea-level and altering ocean currents and ocean-based ecosystems, and precipitation patterns could change and cause once-temperate environments to become flooded or arid, making plants that once flourished in their native ecosystem instead struggle through their growing season.

A search for energy sources outside of petroleum is occurring worldwide. In the United States, many efforts are taking place to find a means of producing energy that would not only diminish or eliminate the threat of global climate change, but also have the potential to boost the United States’ economy. If a new, environmentally-friendly
energy source is found that can be utilized within the United States, that energy source will not only possess an increased innate level of energy security, but it will also have the potential to improve the United States’ economy by creating new jobs. A prime example of an environmentally-friendly energy source that could be produced locally and would foster the growth of a new “green” industry can be found in biofuels.

Biofuels, or fuel derived from biological material, are often cited as an attractive alternative fuel option that could reduce the rate at which atmospheric carbon dioxide levels are increasing by displacing the use of fossil fuels, which currently dominate the transportation sector and are known to emit large quantities of greenhouse gas upon their combustion. Biofuels are often cited as a “carbon-neutral” source of energy, as the use of plant-material for fuel will not release new carbon into the atmosphere—only previously existing atmospheric carbons that the plant captured, sequestered, and incorporated into its cells during its growth are released upon combustion (Ragauskas, et al., 2006). Due to this, the use of biofuels could have the potential to help alleviate the problems associated with global climate change. In fact, if biofuel production waste and emissions were successfully re-sequestered underground, it could even be possible to remove carbon from the atmosphere in order to further combat global climate change (Batjes, 1998; Lemus & Lal, 2005).

Today, most liquid biofuel in the United States is either bioethanol that is derived from annual food crops with high starch or sugar content (such as the starch in a kernel of maize or the sugar in a sugar beet) or it is biodiesel derived from mostly annual food crops with seeds that possess high oil content, such as canola and soy (Agarwal, 2007; Demirbas, 2007). However, a variety of concerns are raised when food crops are used to make biofuel. The central complaint focuses on the fact that using land to grow food-crops for biofuel both reduces the amount of food crop that is produced and drives up
food prices across the globe, thereby potentially removing food resources from the hands of the hungry, an issue commonly referred to as the “Food vs. Fuel” debate (Cassman & Liska, 2007). Many recognize this as a major fault in biofuels produced from feedstocks derived from annual food crops and have instead moved to promoting the use of a new generation of biomass-for-biofuel crops, which would use their harvestable lignocellulosic biomass as the biofuel input. The crops that have been considered for this type of biomass feedstock application generally possess two common traits: one, they produce very large amounts of biomass within a growing season and, two, they can be grown on lands typically deemed unfit for the production of agricultural food crops (commonly referred to as “marginal lands”). It is the hope of many that research in areas concerning the production of advanced biofuels from lignocellulosic biomass, as well as advances in agricultural technology, can help eliminate the threat that “Food vs. Fuel” potentially poses on the world’s hungry.

Yet, regardless of the feedstock means, it is essential that the biofuel industry becomes a profitable enterprise in order for biofuel to truly become a feasible energy source. A large degree of work and resources are being devoted to maximizing the value of biofuel production from a variety of feedstock sources (Wang, et al., 2012; Jingura, 2011; Martín, Ahmetović, & Grossmann, 2010; Sukumaran, Singhania, Mathew, & Pandey, 2009). Maximizing the efficiency of biofuel production, however, is only one single piece of a much greater puzzle. In order to fully foster the growth of the United States’ burgeoning biofuel industrial system, a combination of supportive policies, research efforts, and studies on biofuel potential and utilization are needed.
1.2 The Billion Ton Vision

The Biomass Research Development Act of 2000 set into motion many events, the ramifications of which can still be seen today. The explicit goal of this act was to foster the growth of the biofuel industry in order for biofuels to be a more readily obtainable energy source for the United States (BRDI, 2000). At the time of the Biomass Research Act’s passing, the United States was seeking a secure source of energy that would reduce reliance on foreign oil. In order to help the biofuel industry grow, the Biomass Research Development Act sought to better understand the use of biomass as a fuel source and develop new techniques and technologies to improve the economic efficiency of large-scale biofuel production without sacrificing the environmental-friendliness or energy security traits inherent to the use of biofuels. In order to accomplish this task, the act created three groups: the Biomass Research and Development Board, the Biomass Research and Development Technical Advisory Committee, and the Biomass Research and Development Initiative.

Although all three groups have essential roles in the growth of the biofuel industry, it is the Biomass Research and Development Technical Advisory Committee that had the most direct role in informing future policies concerning the use of biomass feedstocks for the production of biofuel. The Technical Advisory Committee has many duties, but chief among these duties is the responsibility of advising the United States Secretary of Agriculture and the United States Secretary of Energy on issues related to biofuels as well as facilitating communication between interested parties at the federal, state, and private industry level. In order to help accomplish this task, the Technical Advisory Committee drafted a report entitled “The Billion Ton Study” through a partnership with the Department of Energy and the Department of Agriculture. This
The report sought to make biofuel a more widely accessible and obtainable source of energy that would also build the nation’s economy by supporting the growth of agriculture, forestry, and the rural areas of the United States.

The Billion Ton Study was completed in 2005, when biomass provided for approximately three percent of the United States’ total energy consumption (Perlack, et al., 2005). The explicit goal of this study was to determine if it was possible to replace thirty percent of the United States’ petroleum consumption with biofuels by the year 2030. In order to reach this considerable goal, the study estimated that approximately one-billion tons of biomass would be needed to be produced annually for conversion into biofuel, hence the “Billion Ton” moniker within Billion Ton Study. As a result, this report has set the direction in bioenergy research and development since its publication.

The generation of one-billion tons of biomass would require about a five-fold increase in biomass production (Perlack, et al., 2005). Such a large increase in production carries with it a number of potential pitfalls, so it was very important that the proper research was carried out to ensure that increasing biomass production within the United States would be not be to the detriment of the economy, the environment, or the quality of life for American citizens. For example, since the implementation of this plan was put in place with environmental friendliness in mind, the propagation of biomass used for bioenergy needed to occur in the most sustainable manner possible. Due to this fact, the report had to address not only whether the United States actually had the land available to grow such a large supply of biomass, but also if the United States would be able to grow this biomass on its lands without negatively impacting the environment or the economy.

Upon completion, the Billion Ton Study concluded that would be possible for the United States to supply the annual biomass needed to replace thirty percent of the
nation’s petroleum use. In fact, the report determined that it would be possible to exceed the billion ton requirement under ideal circumstances, and the United States would have the potential to produce approximately 1.3 billion tons of biomass annually by 2030. However, the report outlined that the means to accomplish such a feat would be no easy task, and a great deal of both forestland and agricultural resources would need to be utilized in order to reach such a significant demand.

According to the report, the biggest portion of the approximately 1.3 billion tons of biomass that could be produced annually would be grown via standard agricultural practices on agricultural lands. In total, it was hypothesized that the United States would be able to produce approximately 998 million tons of dry biomass on agricultural lands within the continental United States. According to the model used in this report, a value of 998 million dry tons was the highest level at which the production of biofuel from agricultural biomass would not impact demands on food or feed within the United States or the demands on food and feed export. In order to reach 998 million tons, the Billion Ton Study reported that a total of 448 million acres of cropland would need to be utilized. This figure includes the 342 million acres of cropland that was already active at the time, as well as 39 million acres of currently unutilized (or “idle”) cropland and 67 million acres of non-permanent pasture land.

Even with such great stretches of cropland, many improvements to existing agriculture technologies, feedstocks, and infrastructure components would need to occur in order to generate the enormous amount of agriculturally-derived biomass proposed by the Billion Ton Study. Factors such as increased crop yield and composition, as well as improving the efficiency of management practices, are all essential steppingstones on the path to maximizing the potential of agricultural biomass. For example, technological advances such as increasing the fuel efficiency of agricultural equipment, as well as the
more widespread implementation of agricultural practices such as the use of no-till over conventional or reduced tillage, are pivotal developments that would ensure that the environmental benefits of biofuel utilization are maintained.

The existing biofuel sources—maize seed for bioethanol, soybean oil for biodiesel, and other similar sources—make up the smallest percentage of feedstocks from agricultural lands within the Billion Ton Study. The report proposed that 87 million dry tons could be available annually from grains. Although the use of food crops for biofuel remains controversial, the availability and ease of use inherent to these feedstocks for conversion into biofuel still makes it unlikely that grain-based biofuels will be eliminated by 2030. However, as more sustainable options hopefully become more valuable in terms of economic worth and conversion efficiency, the use of annual seed crops as feedstock input should begin to see their numbers wane, and, in effect, reduce the issues associated with using seed crops as outlined in the “Food vs. Fuel” debate.

The largest proposed source of dry tonnage from agricultural lands would come from the residues of currently grown annual crops. In total, the report concluded that 428 million dry tons of biomass could be generated from annual crop residues in the future. This particular source of biomass is obtained from the “leftover” parts of a plant aside from the part of the plant for which the plant is typically grown. For example, when maize is typically harvested, all that is usually taken away from the field is the seed. The rest of the plant—the leaves, the stalks, and the corn cobs—is left to remain on the field. This harvesting behavior is typically profit driven, as it is much more economically advantageous to only have to transport accumulated weight that is solely the desired grain since that is the commodity to be sold, but it also has environmental benefits due to the nutrients that are released into the soil from the rest of the plant as it is broken down into soil organic matter.
There is a trade-off, therefore, to using annual crop residues as biofuel feedstocks since the removal of crop residues will deny the soil the nutrients that could have acquired if the residues had been left on the field to deteriorate. However, if plant biomass is to be used for the production of biofuel, the use of annual crop residues is a logical choice with a variety of innate benefits. These annual crops are already being grown across the nation, so their supply is already present, and the means to harvest the left-over biomass typically already exists (usually with hay collecting and baling equipment). In addition, as this biomass is a product of the growth of existing commodity crops, the generation of annual crop residue does not impact the amount of farmland available to grow food crops.

However, annual crop residues are not the only option. In the Billion Ton Study, perennial crops used as biomass feedstocks for biofuel input composed the second largest source of agricultural biomass for the generation of biofuel and were proposed by the report to be able to supply 377 million dry tons of biomass annually. Using whole-plant biomass as a biofuel feedstock is a much more demanding proposition than using the seed from annual crops. For example, a corn kernel for biofuel production has high starch content, and starch is composed of 6-carbon sugar unit chains that are easily broken down by fermenting organisms for the production of fuel. In this manner, the fermenting microorganisms produce ethanol as a byproduct. This is a fairly simple, well-known process that has been utilized for centuries to produced alcoholic beverages. In comparison, the use of lignocellulosic material as a biofuel feedstock is a much more complicated process. The plant cell wall’s three primary components—lignin, hemicellulose, and cellulose—are the constituents of lignocellulosic biomass. Cellulose is both very abundant within the cell wall and very similar in composition to that of starch, as it is made up of long chains of 6-carbon sugars that, once the long chains have been
broken down, can be readily utilized by fermenting microorganisms for the production of ethanol. However, the other two primary components of the plant cell wall both pose a much more substantial challenge to the production of biofuel. Hemicellulose, the second-most abundant cell wall component, is composed of both the easy-to-ferment 6-carbon sugar units as well as 5-carbon sugar units. These 5-carbon sugars are not naturally fermented by most known microorganisms, which has typically made hemicellulose a much less attractive fermentation input than the pure 6-carbon sugars found within cellulose (Saha B. C., 2003). However, in order for lignocellulosic feedstocks to be ideal biofuel inputs, the conversion of lignocellulosic biomass needs to be as efficient as possible, which means discovering ways to utilize 5-carbon sugars alongside 6-carbon sugars for the fermentation of bioethanol. Due to this, a large amount of research is being performed in order to discover and test potential methods for the efficient utilization of the 5-carbon sugars that are found in hemicellulose as fermentation inputs for fermenting microorganisms (Du, Li, & Zhao, 2010).

The third plant cell wall component, lignin, poses an even greater challenge than hemicellulose. Lignin cannot be fermented, as it is broken down into a series of phenolic compounds that, when depolymerized, will inhibit fermentation by means of their toxicity (Ragauskas, et al., 2006). Since lignin cannot be fermented and will reduce final ethanol yield, lignin often needs to be removed from biomass in order to make bioethanol. This removal of lignin is an expenditure that holds back the economic viability of ethanol, and, as such, a great deal of research is focused on either reducing lignin content in plant cells or finding cheaper, more efficient techniques with which to remove lignin from biomass after it has been harvested (Weng, Li, Bonawitz, & Chapple, 2008).
Beyond the three main players of agricultural biomass feedstocks—grains, annual crop residues, and perennial crops—the Billion Ton Study proposed that a total of 106 million dry tons of miscellaneous agricultural feedstocks could be obtainable. This category contained residual biomass from industrial food and feed processing as well as manure produced by confined animal feeding operations. All together, the biomass generated from agricultural lands and practices would need to be supplemented by approximately 368 million dry tons from American forestlands in order to reach the billion ton goal. The use of these additional sources would be absolutely necessary, as every possible source of usable biomass is needed in order to obtain the Billion Ton Vision’s goal.

1.3 A Post-Billion Ton Study America

In 2005, the Energy Policy Act was passed with the goal of creating a more secure energy climate for the United States (US DOE, 2010). In doing so, the Energy Policy Act funded studies and technology development in many energy-related industries. Both traditional petroleum-based industries as well as alternative energy source industries received funding in order to both promote American business and the economy while simultaneously reducing the United States’ dependence on foreign energy. As one of the primary duties of the committee that authored the Billion Ton Study was to promote the development of the biofuel industry for reasons very similar to the Energy Policy Act, these two documents possess nearly analogous goals. The most direct tie between the Billion Ton Study and the Energy Policy Act, however, was the implementation of the Renewable Fuel Standard by the Energy Policy Act (US EPA, 2013). In the Renewable Fuel Standard, the United States instituted its first blending mandate for renewable fuels
by requiring that, by the year 2012, 7.5 billion gallons of renewable fuel would need to be blended into gasoline. The implementation of this blending mandate created a huge boon for the biofuel industry. While the Billion Ton Study proved that a “source” of biofuel was feasible, this Renewable Fuel Standard mandated a “sink” for biofuel to help make biofuel production not just something that could be done, but something that was required to be done. This push by the Renewable Fuel Standard was more than a mere incentive, it was a requirement that ensured that large-scale biofuel production would become something beyond a theoretical possibility.

However, this particular biofuel boon would not end with the Energy Policy Act. In 2007, The Energy Independence and Security Act built upon the existing Renewable Fuel Standard to intensify the mandated biofuel blending requirement and the “sink” that it created even further. Under the Energy Independence and Security Act, the Renewable Fuel Standard’s mandate for the blending of renewable fuels into gasoline would be increased to a requirement of 36 billion gallons by 2022. Furthermore, the Energy Independence and Security Act opened up the mandate to also include biodiesel in addition to gasoline. In addition, the Energy Independence and Security Act required the EPA to analyze renewable fuels and put thresholds on the greenhouse gas emissions of these renewable fuels in order to ensure that they would actually be more environmentally-friendly than the petroleum fuel that they were replacing, as biofuels that emit more greenhouse gases than traditional fuels would be of little value. In doing so, the Energy Independence and Security Act ensured that not only would the United States continue to support the growth and increase the scope of the renewable fuel industry, but it would guarantee that the environmental benefits of renewable fuel that originally made their use so attractive would also be conserved.
More recently, the American Recovery and Reinvestment Act of 2009 was enacted with the goal of attempting to improve the economy of the United States. Although the act had broad impacts on many areas beyond the renewable energy sector, it did promote the creation of new domestic job opportunities, including those in “green” technology fields such as biofuel. Specifically, the American Recovery and Reinvestment Act supported the rapid deployment of “...biofuel projects at the pilot or demonstration scale...[that are] likely to become commercial technologies and will produce transportation fuels that substantially reduce life-cycle greenhouse gas emissions...” (US GPO, 2009), an action that would substantially aid in the growth of the biofuel industry.

The combination of supportive federal policies and the verification that the Billion Ton Vision is obtainable seemed to signal a bright portent for the future of the biofuel industry. However, as time passed, more and more experts examined the results of the Billion Ton Study and subsequently found the report lacking in a variety of areas. The largest criticism focused on the economic variables used in the Billion Ton Study, as it was felt that proper consideration towards advances in technology and differences in feedstock production by region would result in discrepancies in feedstock price that could drastically alter feedstock availability (Downing, et al., 2011). For example, while certain species of perennial grasses could potentially produce the greatest volume of biofuel feedstock at a farm in central Illinois, the farmer could already be making much more money selling maize that the market’s purchasing price-point of perennial grass biomass would need to be prohibitively high in order for the farmer to consider growing it. In this case, it would be more feasible to have the farmer use the corn stover left on the farmer’s field than to attempt to persuade the farmer to switch out his/her corn crop for a perennial grass.
With concerns such as this in mind, it was clear that the Billion Ton Study needed to be updated and revised, and such an act was completed in 2011. Entitled “The Billion Ton Update,” the newly published report was almost three times the length of the Billion Ton Study that preceded it. In order to fully ensure that the mistakes of the past were not repeated, it was decided that a series of workshops would be held where large groups of experts from industry, government, and academia could all express their opinions and concerns about the Billion Ton Vision, with the hope that such a wide and varying number of views and standpoints would address every possible facet of such a momentous task.

The most substantial change from the Billion Ton Study to the Billion Ton Update was the decision to perform biofuel feedstock analysis on a county-by-county basis for each contiguous state in the Billion Ton Update. The model used in the Billion Ton Study estimated feedstock availability at the national level and did not build in any cost analyses, which caused many of the problems centered around the “economic inadequacy” criticisms of the Billion Ton Study. The Billion Ton Update attempted to rectify this misstep by examining each individual county in the contiguous United States in terms of potential reasonable and cost-effective biofuel feedstock sources and performing a corresponding cost analysis. Other factors beyond spatial considerations were also taken into account—for example, as the United States implemented the Renewable Fuel Standard following the publication of the Billion Ton Study, the Billion Ton Update needed to take into account the potential effects of the Renewable Fuel Standard’s requirements in order to accurately assess the altered economics of feedstock availability. As a result of these and other similar actions performed by the Billion Ton Update’s authors, the nation’s final feedstock availability was both more accurate and more analytical than the findings published by the Billion Ton Study.
A notable difference between the Billion Ton Study and the Billion Ton Update is how each categorizes feedstock sources. While the Billion Ton Study breaks its sources down into two groups—those that come from forestlands and those from agricultural lands—the Billion Ton Update created a new, third category of “energy crops.” The formation of this category helps to further differentiate the sources of feedstock and allows for more care and consideration to be taken in the examination of feedstock availability that results from the categories’ unique intrinsic traits. The Billion Ton Update’s new energy crop category is composed of lignocellulosic feedstocks originally found in the Billion Ton Study’s agricultural category as well as farmed woody crops. A handful of crops are specifically mentioned in this category as possessing the highest likelihood of being used as biofuel feedstocks in the United States, and the three most promising of these crops—switchgrass, sugarcane, and *Miscanthus × giganteus*—are each perennial grasses.

1.4 *Miscanthus × giganteus*

*Miscanthus × giganteus* has generated interest as a possible source of lignocellulosic biomass for the American bioenergy industry in the Billion Ton Update and other similar studies. The *Miscanthus* genus is a perennial C4 grass genus that belongs to the Andropogoneae tribe of the Poaceae family, a large family that includes the agriculturally-important crops of sugarcane, sorghum, and maize. As perennial crops require less inputs than annual crops since they do not need to be replanted each year and are able to sequester nutrients for remobilization in the following growing season, perennial crops are often looked upon favorably in terms of environmental sustainability.
The *Miscanthus* genus consists primarily of diploid and tetraploid species (Hodkinson, Chase, & Renvoize, 2002). The genus is an obligate outcrosser with a highly-repetitive 2.5 Gb genome that is distributed among nineteen chromosomes. Hybridization events between *M. sinensis* and *M. sacchariflorus* have previously been observed (Matumura, Hasegawa, & Saijoh, 1987; Dwiyanti, et al., 2012), and evidence strongly suggests that the sterile triploid hybrid *M. × giganteus* (*3n = 57*) is the result of a hybridization event between a diploid *M. sinensis* (*2n = 38*) and a tetraploid *M. sacchariflorus* (*4n = 76*) (Greef & Deuter, 1993). As *M. × giganteus* is sterile, there is little chance of it cross-breeding with weedy relatives, yet it must be asexually propagated through cutting and planting *M. × giganteus* rhizome tissue. As rhizomes do not come in uniform shapes and sizes, the planting of rhizomes is far less standardized than that of seed crops and requires specialized machinery. Despite this drawback, this plant’s ability to quickly produce a large quantity of biomass during the growing season combined with a relatively small amount of required inputs for its growth (Bullard & Nixon, 1999) has made *M. × giganteus* an ideal biofuel candidate for many regions of the United States. In addition, *M. × giganteus* has been shown to be able to generate amounts of harvestable biomass superior to many of its peers: while corn stover is able to produce around 7.4 Mg ha⁻¹ (megagram or metric ton per hectare) of harvestable biomass in a growing season, and the US-native biofuel-candidate switchgrass can produce up to 10.4 Mg ha⁻¹, *M. × giganteus* has been shown to generate 29.6 Mg ha⁻¹ of harvestable biomass in a single growing season (Heaton, Dohlman, & Long, 2008).

As previously mentioned, *Miscanthus × giganteus* is a sterile hybrid plant and it therefore cannot be propagated by seed. Instead, new *Miscanthus × giganteus* plots are established through the cutting of existing, healthy rhizomes and their subsequent replanting. The rhizome itself is a very interesting plant tissue; since the rhizome is
modified underground stem tissue that can be used as a storehouse for nutrients, the rhizome’s functional characteristics support *M. × giganteus*’ status as an outstanding biomass crop since the rhizome is the key to one of *M. × giganteus*’ key traits: its perenniality. In effect, nutrients that have been assimilated by the plant during the growing season are mobilized to the rhizome once fall senescence begins and winter starts to approach. Whereas *M. × giganteus*’ above-ground biomass is effectively dead during the winter, the rhizome remains a healthy living organ throughout this time period as it safely stores nutrients until the winter ends. Once spring has arrived, the rhizome remobilizes the nutrients that were stored from the previous growing season in order to spur forward a new seasonal generation of plant growth.

Despite *Miscanthus* only being introduced to the western world in the 1930s (Lewandowski, Clifton-Brown, Scurlock, & Huisman, 2000), members of this genus can currently be found growing as ornamental crops in many regions of the United States typically due to their characteristically large stature and attractive late-season inflorescence. Although *Miscanthus* has held horticultural interest for some time, it essentially remains a genus of wild species. Conventional breeding efforts have thusly been largely concentrated on traits of interest to the horticultural and landscaping industry; there have been few focused research efforts that have targeted traits to enhance the potential of the *Miscanthus* genus, including *M. × giganteus*, as perennial biomass-for-biofuel feedstocks. As a result, the proper application of any newly acquired genomic knowledge on *M. × giganteus* has the potential to be greatly aid in the goal of making *Miscanthus* a more viable and valuable biofuel feedstock source.
CHAPTER TWO

“A TEN-TISSUE RNA-SEQ STUDY OF Miscanthus × Giganteus”

2.1 Introduction

As mRNA transcripts largely represent the qualities and capabilities of an organism at the level of gene expression, the identification and characterization of an organism’s transcript expression patterns and variable expression responses under different environmental and physiological conditions provides a means with which to interpret the organism’s functional behavior at the molecular level. Specifically, the accumulating of data on gene expression patterns within M. × giganteus gives the ability to investigate genes associated with the plant’s biofuel feedstock related traits. Using a technique commonly identified as RNA-Seq, the reads from modern high-throughput sequencing can be used to create expression profiles by measuring the quantity of reads that align to reference gene models and the differences in expression levels that occur between separately sequenced samples (Wang, Gerstein, & Snyder, 2009).

The expression profiles that can be created through RNA-Seq are useful for a number of downstream studies, as the knowledge of where and under what conditions a gene is being up- or down-regulated can reveal more about the function and purpose of the gene within the studied organism (Bouchez & Höfte, 1998). This type of method has proven to both highly accurate and highly sensitive in its quantification of gene expression levels (Wang, Gerstein, & Snyder, 2009). Once genes that correspond to a certain metabolic pathway or plant trait are identified, the extent of the identified genes’ impacts on the trait of interest can be studied. Even in organisms without a fully
sequenced genome, the relationship of high-throughput sequencing data to a closely-related reference genome can provide useful insights into the organism’s functional characteristics.

Previous to the completion of this study, very little mRNA resources or RNA-Seq studies on \( M. \times giganteus \) existed (Hodkinson, Chase, & Renvoize, 2002; Swaminathan, et al., 2010; Dwiyanti, et al., 2012; Chouvarine, et al., 2012). As such, this project sought to generate a dataset that would be useful for all labs considering investigating \( M. \times giganteus \) transcriptomics by producing data that provided a wide range of high-quality expression information derived from a variety of different \( M. \times giganteus \) tissue types (Barling & Swaminathan; in review). In addition to the RNA-Seq dataset produced by this project, a transcriptome for \( Miscanthus \times giganteus \) was also constructed from the same collection of sequenced high-throughput paired-end reads discussed herein (available online at ftp://ftp.jgi-psf.org/pub/JGI_data/Miscanthus/transcriptome/).

An organism’s transcriptome is simply a documentation of an expressed mRNA transcript collection from that organism. Ideally, a complete transcriptome will identify all possible transcripts produced by an organism within every tissue, under every possible set of environmental conditions, and during every stage of development. In this regard, a transcriptome can be very valuable on its own; however, in an organism that does not yet have its full genome sequenced, this information can additionally aid in a future genome assembly by effectively establishing its gene content. The ability to use an assembled representation of an organism’s transcriptome for the annotation of its genome is a gainful, recognized application of transcriptomic data that can help identify previously unknown genes and exons (Saha, et al., 2002; Shah, et al., 2006). This type of strategy has already been successfully applied to crops such as rice (Tyagi, et al., 2004;
Jiang, Christoffels, Ramamoorthy, & Ramachandran, 2009) and could work just as well for other plants such as *M. × giganteus*.

2.2 Materials and Methods

2.2.1 Tissue Sampling and Processing

Ten *M. × giganteus* tissues were selected for use in this study: Mature Leaf, Vegetative Shoot Apex, Vegetative Sub-Apex, Spring Rhizome, Immature Inflorescence, Emerging Shoot, Pre-Flowering Apex, Rhizome Bud, Root, and Mature Inflorescence (Table 1). Sampling for most tissues took place at a small field plot located at the University of Illinois’ Turf Farm in Urbana, Illinois, with the exception of Root tissue and Rhizome Bud tissue, which were grown in Turface soilless medium in greenhouses at the University of Illinois Champaign/Urbana campus. Tissues were flash-frozen in liquid nitrogen upon sampling and ground by hand via mortar and pestle to a fine powder while submerged in liquid nitrogen in order to preserve the sample as effectively as possible. RNaseZAP was used to clean all labware and utensils in order to minimalize exposure to ambient environmental ribonucleases, which are known to rapidly catalyze the degradation of RNA (D’Alessio & Riordan, 1997). Samples were stored at -80°C until they were ready for RNA extraction.

2.2.2 RNA Extraction

For total RNA extraction, approximately two grams of ground tissue were homogenized with a CTAB-based extraction buffer and 2% beta-mercaptoethanol. After the sample had been homogenized, an equal volume of acidic phenol:chloroform:isoamyl
alcohol (125:24:1) was added and then the solution was centrifuged at high speed for ten minutes. After centrifugation, the upper aqueous layer was removed and transferred to a sterile 50ml tube and an equal amount of chloroform:isoamy alcohol (24:1) was then added. The solution was re-centrifuged, after which the upper layer was once again transferred to a new 50ml tube. Lithium chloride was added to precipitate the RNA and the solution was stored overnight at 4°C. On the following day, the sample’s precipitated RNA was pelleted via high-speed centrifugation at 4°C for thirty minutes. Once completed, the supernatant was carefully removed and the pellet was washed with a 70% ethanol solution and then resuspended in nuclease-free water and transferred to a new nuclease-free 1.5ml tube. Samples were then stored at -20°C for one hour in an ethanol and ammonium acetate solution. Following this, samples were re-pelleted via rapid centrifugation at 4°C. The resultant pellet was washed and resuspended in 70% ethanol, upon which point the solution was re-centrifuged and pelleted once again. Ethanol was then removed from the sample via vacuum concentrator centrifugation and the pellet was suspended in certified nuclease-free water. The amount of total RNA present in the sample was then quantified with a NanoDrop Spectrophotometer.

Following the manufacturer’s protocol, Dynabeads (Invitrogen, 2013) were employed to purify mRNA from the total RNA, as Dynabeads are able to bind to the polyA tail of mRNA to separate it from the rest of the RNA found in the sample (Jakobsen, Breivold, & Hornes, 1990). In order to generate the quantity of mRNA required for sequencing, multiple runs of total RNA extraction followed by mRNA purification were required for each sample. Each mRNA extraction had its sample yield quantitated with a NanoDrop Spectrophotometer, and sample name and mRNA yield was recorded before the mRNA extracts were stored at -80°C until it was time for sequencing.
2.2.3 Sequencing

Prior to sequencing, sample quality was verified with an Agilent 2100 Bioanalyzer; only samples with a 260/280 of 2±0.1 and a minimum RNA integrity number of 8 were used for sequencing. Two runs of Illumina high-throughput paired-end sequencing were completed for this project: the first in November, 2008 and the second in February, 2009. The libraries for both sequencing runs were made and run on an Illumina Genome Analyzer IIx by the W. M. Keck Center at the University of Illinois.

2.2.4 RNA-Seq

Reads were adapter-trimmed and quality filtered via Perl scripts to remove Illumina sequencing adaptors and low quality bases from the ends of reads in order to reduce the amount of erroneous results. Specifically, the freely available “Trim.pl” script (wiki.bioinformatics.ucdavis.edu/index.php/Trim.pl) was used to remove bases below a phred quality score of 10 via windowed adaptive trimming. Reads were then imported into CLC Genomics Workbench Version 3.7 (CLC bio, 2013), which was used to align the reads to the Phytozome 7.0 version of the unmasked Sorghum bicolor genome (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v7.0/Sbicolor/). The following CLC settings were used: require 94.4% identity, extend annotated gen regions for 300 flanking residues upstream and downstream, and only use reads that aligned to a maximum of five locations. In addition, CLC Genomics Workbench’s “exon discovery” feature was enabled during alignments, with the requirements that discovered exons were required to have a relative expression level of 0.2 and a minimum length of ten reads containing 50 or more residues. Unique read counts were exported from CLC
Genomics Workbench and converted into RPKM via Microsoft Excel. RPKM results were compared to one another via Rank Products (RP) analysis (Breitling, Armengaud, Amtmann, & Herzyk, 2004).

2.2.5 RT-qPCR

Nine genes that appeared to have mostly rhizome-specific expression and high ranking RP scores were chosen for potential candidates for use in RT-qPCR. Primers were tested on a Roche LightCycler 480, and, of the nine candidates, five primer pairs had an amplification efficiency of $2 \pm 0.1$ and were of high enough quality for RT-qPCR validation as determined by version 1.5.0.39 of the LightCycler Software Package (Roche, 2013). In addition to these five rhizome-centric primer sets, two primer sets for genes with known preferential leaf expression were added to this verification. Since no housekeeping genes (i.e., genes with stable, constant, and widespread expression) have been tested or verified for use in $M. \times giganteus$, five potential control candidates were deduced from the Rank Product data. These potential control candidates contained gene models with near-equal RPKM values in all five tissues used in this verification as determined by the sequencing project. The two best-performing gene models from these five candidates, in terms of amplification efficiency and closest-to-equal expression as determined by the LightCycler Software Package, were chosen as control genes for this study (Table 2).

Newly collected tissue samples were gathered in triplicate from the University of Illinois Turf Farm’s $M. \times giganteus$ plot in April and May, 2011. In total, five tissue types corresponding to originally sequenced tissues were tested in the RT-qPCR: Mature Leaf, Emerging Shoot, Spring Rhizome, Rhizome Bud, and Root. A Roche LightCycler 480 and
its default software package were once again utilized to run the RT-qPCR tests and determine relative gene expression. In these runs, each individual sample had three biological replicates and four technical replicates. Once completed, RT-qPCR result data was examined and transformed within the Roche LightCycler Software Package before being exported into Microsoft Excel in order to finalize the relative expression analyses via the ΔΔCt method (Schmittgen & Livak, 2008; Arocho, Chen, Ladanyi, & Pan, 2006).

2.3 Results and Discussion

In order to create a dataset that would represent as much of Miscanthus × giganteus’ potential gene expression as possible, ten distinct M. × giganteus tissue types were sampled for this project. Most of these samples were obtained at the University of Illinois’ Turf Farm in a plot that has been maintained since the 1980s. Two separate paired-end Illumina sequencing runs were employed for this sequencing project; in the time between the sequencing of the first run and the sequencing of the second, Illumina technology advanced to the point where a greater number of reads could be sequenced with an increased base-pair length.

Due to this, tissue samples belonging to the earlier, first run contained reads of thirty-six base pair length whereas the tissue samples ran in the newer, second run of sequencing contained reads with a length of seventy-six base pairs. For both runs, all sequenced mRNA was not subjected to any pre-sequencing RNA normalization. Normalization did not occur in order to ensure that the sequenced data would be useful in two parallel projects: the assembly of the M. × giganteus transcriptome and a corresponding RNA-Seq expression analysis. Since genes with high levels of expression will have more transcribed mRNA present, it is logical that these same highly-expressed...
genes will also have a higher amount of sequenced reads in comparison to genes with low expression due to the overall greater abundance of highly-expressed genes’ mRNA at the time of sampling and subsequent sequencing.

*Miscanthus × giganteus* tissue samples found in the early sequencing run, which was performed in November 2008, are as follows:

1. Mature Leaf
2. Emerging Shoot (1)
3. Vegetative Shoot Apex
4. Sub-Apex Shoot
5. Spring Rhizome
6. Immature Inflorescence

In addition, the following *M. × giganteus* tissues were sequenced in the later run, which was performed in February 2009:

1. Pre-Flowering Apex
2. Rhizome Bud
3. Root
4. Mature Inflorescence
5. Emerging Shoot (2)

It should be noted that each sequence run contains an “Emerging Shoots” sample. These samples were sequenced from the same mRNA stock, and served as a quality control comparison-sample for results obtained from the 36-bp run and the 76-bp run. Upon completion, a total of over 142 non-normalized million paired-end reads were sequenced in *M. × giganteus* (Table 1).

In order to perform an RNA-Seq analysis, a reference genome must be used in order to have reference gene models to which the reads generated from the sequencing can be aligned. Ideally, the reference genome would be the genome of the organism...
sampled. Unfortunately, there is currently no *Miscanthus* genome available; therefore, an appropriate reference genome from a closely related organism needed to be utilized instead. When assembled *M. × giganteus* contigs are compared to the gene models for *Sorghum bicolor*, rice, maize, and *Brachypodium*, as well as sugarcane ESTs, it becomes apparent that *M. × giganteus* shares the highest degree of similarity to the sugarcane ESTs and the *Sorghum* gene models (Figure 1). As such, either of these two organisms has the potential to be an adequate reference for *M. × giganteus*. However, *S. bicolor* possesses a number of characteristics that make it better suited as a reference for this study than sugarcane: in addition to *S. bicolor* having been successfully utilized as a reference in previous genomic studies concerning the *Miscanthus* genus (Swaminathan, et al., 2010), it also comes with the added benefit of having a simpler, fully sequenced diploid genome, whereas sugarcane’s considerably more complex octoploid genome currently only possesses an EST collection (Paterson, et al., 2009). With these factors in mind, *S. bicolor* was chosen as a reference for this *M. × giganteus* project.

Once CLC had finished all alignments, approximately 96% of the total 26,230 *S. bicolor* gene models had five or more mapped *M. × giganteus* reads. When reads that map uniquely to *S. bicolor* gene models are compared to those that map non-uniqueley (in this case, non-uniqueley is defined as between two and five mapping locations), a few patterns can be found (Figure 2). For each *M. × giganteus* tissue, approximately sixty percent of the reads will map to *S. bicolor* gene models. This consistency of mapped read percentage for each tissue sample is notable when the much greater number of reads gathered from the second run of paired-end sequencing when compared to those of the first is considered: despite the higher total read count and longer overall length of reads derived from tissues sequenced in the second sequencing run, every tissue—regardless of run, read length, and read count—possesses approximately the same percentage of reads
that will map to *S. bicolor* gene models. This knowledge indicates that any potential “quality gap” of reads derived from tissues sequenced in the first and second runs is minimal, as all sequenced tissues possess the same overall high read quality despite the generation of Illumina technology utilized for sequencing.

Substantial differences in both read length and total read count between samples sequenced in the first sequencing run and those sequenced in the second sequencing run make direct comparison of uniquely mapped read-values across the two separate runs difficult, as the inherent differences in read length and read count necessitate the use of complex normalization and transformation procedures if a direct comparison of samples across the two runs is to be made (Robinson & Oshlack, 2010). While these normalization and transformation practices are not an issue in themselves, the large amount of data lost in performing them on this particular dataset is troublesome. As such, a transformation and normalization method that would minimize the degree of data loss was desired.

It was for this reason that the Rank Products (RP) method was chosen (Breitling, Armengaud, Amtmann, & Herzyk, 2004), as using RP on RPKM-normalized expression data lessened the presence of the previously mentioned concerns. RPKM, or “Reads Per Kilobase of transcript per Million mapped reads,” is a commonly used normalization method in RNA-Seq experiments that can be utilized to analyze high-throughput sequencing data for the basis of creating gene expression profiles. One of the largest benefits of using RPKM is that comparisons between genes of samples with starkly different library sizes are still able to be made due to RPKM correcting for differences in both library size and gene length (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008).
Once RPKM values had all been calculated, the RP method was implemented on the resultant RPKM datasets in order to aid in the ability to make direct comparisons of gene expression in samples across the separate sequencing runs and facilitate the timely manner in doing so. The RP method is able to determine significant RPKM differences through comparisons of fold change, as RP rankings arise from consistencies in fold change differences observed among a collection of samples. Since our \textit{M. × giganteus} tissue samples were largely unreplicated (with the exception of Emerging Shoot, which had a technical replicate in each sequencing run but not a true biological replicate), every sampled tissue other than the tissue being examined was compared to the singular examined tissue. As a result, the final RP ranking results for each tissue identifies genes with preferential expression for that single particular tissue by comparing each individual tissue to all other tissue types. It should be noted, however, that since there were two samples of Emerging Shoot, both samples needed to be treated as a single sample with expression values averaged between the two in order to reduce the bias in the results.

As a result, the Rank Product method was used to evaluate for significant differences in RPKM (and, in effect, significant differences in gene expression) by means of fold change consistencies and inconsistencies between the sequenced tissues (Breitling, 2013). The information obtained by the Rank Products analyses created listings of highly up- and down-regulated genes that have a greater degree of preferential expression unique to one particular tissue in comparison to the other sampled tissues. As a result, these Rank Product listings are helpful in quickly discovering genes either up- or down-regulated in a specific \textit{M. × giganteus} tissue. The listings for highly up-ranked genes within these files provides many biologically-relevant genes specific to each particular sampled tissue. For example, photosynthetic genes like phosphoenolpyruvate
carboxylase (PEPCase) are highly ranked within the Rank Product list generated for the *M. × giganteus* Mature Leaf sample, which is to be expected as PEPCase expression should be far more prominent in the Mature Leaf tissue than that of the other sampled tissues. As a result, the occurrence of expected results such as PEPCase’s presence in Mature Leaf helps lend credibility to genes that occur highly ranked in a sample that are less expected.

Although the high up-rankings of genes that should logically be highly expressed in certain tissue types enforces the validity of this approach, a secondary “true” validation of the application of the RP method on this data was sought. As validation inferences drawn purely from the sequenced dataset are easily questioned, a technique that would be able to confirm specific RPKM results without using the sequencing data would be needed. To this end, RT-qPCR was used to obtain a validation of this manner and verify the accuracy of the RPKM-derived Rank Product results. Five Spring Rhizome genes with high RP rankings and two Mature Leaf genes with high RP rankings were used in this analysis (Table 2). An integral part of the interpretation of these genes’ qRT-CPR assays is to also assay expression for genes that exhibit highly stable expression among the samples, which serve to control for sources of technical experimental variation. To this end, five genes showing near-equal RPKM values in all five tissue types were selected as possible long control genes; of these five candidates, two genes were found to possess sufficient amplification efficiency and stable expression with little variation across all the newly collected tissue sample types. In addition, each of these two particular genes are predicted to encode proteins with housekeeping functions, as annotations of these genes are consistent with housekeeping functions: the first gene, Sb09g019750, belongs to a group of evolutionarily conserved Bax inhibitor-1 family proteins involved in Golgi vesicle transport, whereas the other, Sb02g041180, encodes
the 51-kDa subunit of the mitochondrial NADH-ubiquinone oxidoreductase. As a result, these two genes were chosen for use as control genes in this study.

The resulting relative expression analyses generated via RT-qPCR with the described genes align well to the expression patterns estimated through their corresponding sequencing-derived RPKM value profiles (Figure 3), thereby verifying the accuracy of the RPKM values and the subsequent use of the Rank Product methodology. Occasionally, gene expression for the root tissue in these results appeared higher than predicted from their RPKM values. This result is most likely attributed to the growth environments of the roots. For the sequencing project, *M. × giganteus* root tissue, as well as some supplemental rhizome bud tissue, was sampled from greenhouse plants grown in calcined clay in order to maximize tissue quantity harvest amount during sampling and reduce the presence of soil and soil microbes in the collected samples. Since only a fraction of the mRNA quantity needed for sequencing is required for RT-qPCR, root tissue for this expression analysis was gathered alongside all other tissue types from the University of Illinois Turf Farm’s *M. × giganteus* field plot. Other possible explanations for the discrepancy between root RPKM and RT-qPCR relative expression value patterns include that the sequenced root reads were simply of poor quality, which limited their ability to align properly and/or adequately to *S. bicolor* gene models, or differences in the relative expression of the two RT-qPCR control genes compared to the other tissue samples, which could introduce a bias against the root tissue’s relative expression values. Other than this small but notable discrepancy, the overall resulting profiles of the Ct-based RT-qPCR relative expression tests match well with their corresponding RPKM expression profiles, thereby confirming the accuracy of the sequencing-derived RNA-Seq results via a procedure that more directly assesses the plant’s physical mRNA expression amounts.
2.4 Conclusion

Nearly sixty-five million *M. × giganteus* paired-end reads were sequenced by this project, the product of over 7.2 billion bases. Since ten distinct tissue types were sequenced in this project, it was possible to create a series of expression profiles that each list and rank genes that appear to have a preference for expression in a single tissue in comparison to the other sequenced tissues. These ranked lists contain both examples of known biologically-relevant genes with tissue-preferred expression, such as photosynthesis-related genes in photosynthetic leaf tissue, as well as genes with less well-known annotations that could help widen the scope of knowledge on lesser-studied tissues such as the rhizome. The validity of these expression profiles was confirmed via RT-qPCR, thus indicating this transcriptomic resource will be highly useful for future studies of *M. × giganteus* biology.
CHAPTER THREE

“CONSTRUCTION AND UTILIZATION OF A SEASONAL MISCANTHUS × GIGANTEUS RNA-SEQ DATASET”

3.1 Introduction

The initial *M. × giganteus* dataset derived from the paired-end sequencing of ten different *M. × giganteus* tissue types provided a wealth of information from which many interesting gene expression characteristics could be discovered. One notable example of interesting gene expression that was observed within these tissues after the project’s completion concerned the expression of jasmonate-related genes within the Spring Rhizome tissue sample. Within the Spring Rhizome and, to a lesser degree, the other subterranean tissues (Root and Rhizome Bud), many genes associated with plant hormone pathways showed dramatic changes in gene expression, in particular those for jasmonic acid synthesis and signaling.

Jasmonic acid is derived from linolenic acid and is a primary member of the jasmonate class of plant growth regulators (Creelman & Mullet, 1997). As such, the term “jasmonate” is often colloquially used to refer to jasmonic acid, the biologically-active intermediates in the jasmonic acid biosynthetic pathway, and the biologically-active derivatives of jasmonic acid (Turner, Ellis, & Devoto, 2002). Jasmonic acid itself is a volatile compound that is known to be involved in a large variety of plant functions, ranging from regulating the growth of roots as well as other aspects of overall plant growth and development, to managing the plant’s responses to abiotic and biotic stressors such as fungal infection and mechanical wounding (Creelman & Mullet, 1995).
As the physical act of digging up rhizomes will inevitably damage the rhizome and other underground tissues—such as that which occurred during tissue harvesting for the sequencing project outline in chapter two—it could be possible that the up-regulated expression of jasmonic acid pathway related genes in underground tissues is simply a consequence of wounding during tissue sampling. However, if the observed up-regulation in jasmonate-related expression is not due to a response to plant wounding from the act of tissue sampling or a similar response to pathogen-induced stress, it appears very likely that jasmonic acid could play an important role in the spring-related growth and development of the rhizome tissue. In previous studies, it has been shown that the presence of exogenous jasmonate can induce the growth of underground tubers in potatoes and yams (Koda, et al., 1991; Koda & Kikuta, 1991). In addition, jasmonate has also been observed to promote the formation of bulbs and stems in tissue-cultured garlic (Ravnikar, Žel, Plaper, & Špacapan, 1993). As the rhizome is a modified stem tissue—similar to how a potato tuber is modified stem tissue, a yam tuber is modified root tissue, and a garlic bulb is a modified stem and leaf combination—it is reasonable to believe that jasmonic acid may play a similar role in promoting the growth-related development and gene expression changes that were observed in the M. × giganteus rhizome (Torrey, 1925; Lebot, 2009).

Unfortunately, it is impossible to determine if the jasmonate related expression is due to stress or to rhizome development solely from the M. × giganteus ten-tissue RNA-Seq dataset outlined in chapter two, as there is only one sequenced rhizome sample; therefore, there are no other rhizome samples with which observations about gene expression changes that occur in the rhizome as the growing season progresses can be made. In order to more fully address the role of jasmonate and jasmonic acid related signaling in the rhizome, a dataset that contained rhizome tissue sampled at different
points in the growing season would be required. Therefore, in order to construct such a resource and help assess the role of jasmonic acid as well as other instances of season-based gene expression patterns within *M. × giganteus*, such as those related to other plant growth regulators and seasonal nutrient cycling, a new RNA-Seq dataset that could document seasonal changes in gene expression was constructed.

3.2 Materials and Methods

3.2.1 Tissue Sampling and Preparation

Twenty-four *Miscanthus × giganteus* tissues were sampled between 2009 and 2012 in order to construct a seasonal RNA-Seq dataset (Table 3). Three types of tissues were sampled—rhizome, internode, and leaf—from two locations in Illinois: the same University of Illinois *M. × giganteus* Turf Farm plot from which the previous ten-sample RNA-Seq project’s samples were collected as well as a second location at a *M. × giganteus* test plot located in Pana, Illinois. Each tissue was sampled with three individual biological replicates at the time of sampling and flash frozen in liquid nitrogen and kept on dry ice and in freezers in order to preserve the quality of the tissue. Following this, samples were ground primarily with both mortar and pestle and, when necessary, with Ika A-11 Analytical Mills until a fine powder consistency was achieved (IKA, 2013). Total RNA was then extracted from ground samples in a manner identical to that described in chapter two with the exception that total RNA, and not purified mRNA, was given to the W. M. Keck Center for sequencing.
3.2.2 Sequencing

A total of seventy-two total RNA samples (twenty-four samples with three biological replicates each) were sequenced by the W. M. Keck Center at the University of Illinois in Urbana, Illinois (Biotechnology Center, 2013). RNA-Seq libraries were prepared using Illumina’s TruSeq RNA-Seq Sample Prep Kit. Prior to sequencing, libraries were pooled and quantified by qPCR. Each pool was sequenced in a single lane for 101 cycles from each fragment end on an Illumina HiSeq 2000 system using a TruSeq SBS sequencing kit (version 3).

3.2.3 Alignment

Prior to alignment, reads were trimmed using Trimmomatic (Lohse, et al., 2012). Adapters were first removed using the ILLUMINACLIP option with two maximum mismatches allowed, a palindrome clip threshold of thirty, and a simple clip threshold of ten. After adapter trimming, both leading and trailing bases were required to have a phred33 score of at least three or be trimmed. Any reads that had been reduced below a length of twenty-five nucleotides by adapter or end trimming were removed from the dataset.

TopHat2 was used to align sequences to Phytozome’s newest available version (9.0) of the *Sorghum bicolor* assembly, which was obtained directly from Phytozome (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Sbicolor/). Supercontigs were removed from the FASTA and GFF3 files prior to alignment with grep (GNU, 2013). Bowtie2-build was then used to build an index from the *S. bicolor* assembly files, and bowtie-inspect was used to verify the integrity of the completed index. The highest read-
yielding sample files for each tissue type sequenced (“10/26/09 Rhizome replicate B,” “9/27/10 Leaf replicate C,” and “9/27/10 Internode replicate B”) were concatenated together using “cat” (GNU, 2013) in order to build a junctions file that would help reduce overall alignment time. TopHat2 was then used to align the concatenated file using conditions outlined below to the *S. bicolor* reference; the resulting junctions.bed file was converted into a .juncs file via “bed_to_juncs,” a script that is included in the TopHat2 install, for use in future alignments in order to identify and align to novel splice junctions not within the *S. bicolor* annotation file that instead originated from the three *M. × giganteus* tissue types. As a result, when future alignments were run, novel non-*S. bicolor* junction sites were included in the alignment that were determined by the collection of all three *M. × giganteus* tissue types and not solely limited to novel junction sites identified from the single tissue used for a particular alignment, thus allowing TopHat2 to search the same scope of tissue-spanning novel *M. × giganteus* splice junctions during each sample’s alignment.

A variety of built-in options and settings of TopHat2 were utilized in order to maximize the quality of this project’s alignment while still keeping total run-time at a reasonable length. To this end, reads were discarded if they were discovered to possess more than ten mismatches during alignment, if they contained gaps over six nucleotides in length, or if they possessed over ten nucleotides in total edit distance. Mate inner distance was set to fifty nucleotides and mate standard deviation was set to sixty in accordance with the used sequencing-run parameters and results as outlined by the W. M. Keck Center. Read-pairs closer than twenty-five base pairs were excluded from alignment, and only insertions and deletions under fifteen base pairs and ten maximum hits in multiple locations were allowed. The “microexon search” option was enabled, allowing TopHat2 to search for microexons during alignment. Sensitivity was set to “b2-
very-sensitive,” which enacts a number of calibrations, including having the alignment advance if twenty consecutive seed extension attempts fail, setting the maximum number of re-seeding reads with repetitive seeds to three, allows no mismatches permitted per seed, and sets the length of seed substrings to twenty (Langmead & Salzberg, 2012).

After sequencing, the number of mapped reads for each replication in all samples was determined (Figure 4). The flagstat function of SAMtools (Li, et al., 2009) was used on the “unmapped.bam” file produced by TopHat2 for each sample to identify the number of unmapped reads. Unmapped read information was obtained in lieu of mapped read information due to obfuscations introduced by allowing multiple hits; with unmapped read information properly quantified, the number of mapped reads was then identified by subtracting the number of unmapped reads from the total number of reads input into TopHat2.

3.2.4 Testing for Differential Gene Expression

The Cufflinks package (version 2.1.1) was used to test for differential gene expression within our RNA-Seq samples (Cufflinks, 2013; Trapnell, et al., 2010; Roberts, Trapnell, Donaghey, Rinn, & Pachter, 2011; Roberts, Pimentel, Trapnell, & Pachter, 2011; Trapnell, et al., 2013). Cufflinks was initially run on each individual “accepted_hits.bam” file generated for each TopHat2 alignment using the S. bicolor reference annotation and MultiFASTA file. Two notable settings were enabled for these runs: an option that allowed Cufflinks to utilize an initial estimation procedure in order to more accurately score reads that mapped to more than one genomic location as well as an option that enabled Cufflinks to normalize by the upper quartile of fragment count that mapped to individual loci instead of simply the total number of sequenced fragments, effectively
improving the results for any subsequent differential expression tests concerning the less abundant genes and transcripts. Following the use of Cufflinks, “cuffmerge” (a package included with the default Cufflinks install) was used on all “transcripts.gtf” files produced to produce a merged GTF file that was used for downstream Cuffdiff differential gene expression analysis.

Using all “accepted_hits.bam” files generated by TopHat2 as well as the GTF file created from cuffmerge, a Cuffdiff (version 2.1.1) run was started that tested for differential gene expression using all samples and all replications. In this run, multi-read correction was enabled and the geometric-FPKM normalization method was enabled. Once finished, CummeRbund (CummeRbund, 2013) was installed and used in RStudio (RStudio, Inc., 2013) to create a database of expression information and aid in the analysis of this otherwise large and unwieldy dataset.

3.2.5 PAGE Analysis of RNA-Seq Data

Custom-written R scripts were used in conjunction with built-in cummeRbund commands to perform selected pairwise tests between samples. *S. bicolor* genes found to be significantly different between the selected paired samples were imported into agriGO (version 1.2), along with their corresponding fold-change numbers, for use in analysis via the Parametric Analysis of Gene set Enrichment (PAGE) tool (Du, Zhou, Ling, Zhang, & Su, 2010). Any gene where both samples had an FPKM value of less than five were not included in the import file in order to cut down the number of genes with possible high statistical significance but low biological significance. PAGE analysis was run using *Sorghum bicolor* as the reference species, the Hochberg (FDR) multi-test adjustment method, a significance level requirement of 0.05, and a minimum of ten mapping entries.
for a term to be evaluated. Both agriGO PAGE input information and the resulting agriGO graphical outputs were downloaded and saved locally for future use.

3.3 Results and Discussion

3.3.1 RNA-Seq

Multiple samplings of rhizome, internode, and leaf were taken during the growing seasons of 2009 through 2012 in order to create a seasonal dataset with which to assess season-based changes in *M. × giganteus* gene expression (Table 3). A total of seventy-two samples were taken for this project (twenty-four individual samples, each with three biological replicates) from two different field plots in order to get the widest representation of gene expression throughout the growing season as possible. Analysis of these newly sequenced tissue samples can be used to determine seasonal changes in gene expression within *M. × giganteus*. As the previous RNA-Seq dataset was obtained from plant tissues sampled primarily at a single point in the growing season, the replication and repetition of sampling the same tissue during different parts of the season over numerous seasons and two separate field plots by this project provides a comprehensive array of information related to seasonal changes in gene expression.

Similarly to how high-throughput sequencing technology had evolved in the time between the first and second run of sequencing in the previous ten-sample sequencing project, the top-of-the-line Illumina HiSeq technology currently used for most high-throughput sequencing projects had once again advanced. The technology utilized at the time of this sequencing represents the latest step forward in sequencing technology, offering up vastly greater quantities of sequenced reads that were additionally longer than those of the previous sequencings (Figure 5). In total, this sequencing project
produced over 972 gigabytes of uncompressed read data, representing almost four billion total sequenced mRNA paired-end reads.

In the previous ten-tissue RNA-Seq project, CLC Genomics Workbench was used for read alignment. While CLC Genomics Workbench has some notable features, namely its speed and ease-of-use, it was decided early on to explore other alignment options since, despite its speed and user-friendliness, CLC Genomics Workbench is proprietary software that carries a hefty licensing fee. As much of what is being done “under the hood” in CLC Genomics Workbench is proprietary, a great deal of information about what and how some of the functions of CLC Genomics Workbench are being performed is withheld from the everyday user. In contrast, there are many alignment tools available that are not just free, but also have complete transparency when it comes to their functionality and, in some cases, a larger degree in flexibility in the customization of the alignment tool's functionality. To this end, a variety of powerful, freely available alignment programs were investigated, including Novoalign (Novocraft, 2013), Bowtie 1 & 2 (Langmead, Trapnell, Pop, & Salzberg, 2009; Langmead & Salzberg, 2012), Stampy (Lunter & Goodson, 2011), BWA (Li & Durbin, 2009), and TopHat version 2.0.9 (Kim, et al., 2011). Fortunately, a wealth of information is available to aid in the choice of an alignment tool that will fit a particular project’s needs from a variety of sources (Nookaew, et al., 2012; Medina-Medina, et al., 2012; Grant, et al., 2011; Garber, Grabherr, Guttmann, & Trapnell, 2011; Vijay, Poelstra, Kunstner, & Wolf, 2013). Although many of these alignment tools are far more complicated and less user-friendly than CLC Genomics Workbench, requiring extensive use of a command-line interface more often than not, the level of versatility, flexibility, and transparency over proprietary software such as CLC Genomics Workbench coupled with their low, low cost of nothing can be extremely attractive.
Three primary criteria were considered when different alignment tools were examined for use with this particular project. Firstly, the program would need to be fairly quick in its alignment process. As this project contains seventy-two individual samples that would need to be aligned, runtimes longer than a few days would cause an undesirable delay in data acquisition. Features such as supporting the use of CPU multithreading and the use of efficient data-management algorithms, such as the Burrows-Wheeler algorithm (Burrows & Wheeler, 1994), can aid in the speed at which alignment occurs. Secondly, it was preferred that the alignment tool used would be “splice aware,” which simply means that the aligner would be able to align to the reference even if there are gaps in the alignment. Unspliced read aligners are limited to aligning to previously identified exons and junctions and cannot manage alignments involving unidentified splicing sites (Vijay, Poelstra, Kunstner, & Wolf, 2013; Garber, Grabherr, Guttman, & Trapnell, 2011). Splice aware alignment tools are able to align reads by either first mapping reads “unspliced” and then breaking down the reads into shorter segments and realigning independently, or by breaking reads first into short seeds and aligning them to the genome, at which point regions are then individually examined to determine an exact spliced alignment location for a read. As our reference genome is a different organism than that from which our reads are derived—an occurrence that creates the perfect opportunity to have splice junctions in genes that are not identified in the reference genome—a splice aware read aligner would be of great value to this project. The final criteria in alignment tool selection required that since the reads in this project are all paired-end, the chosen alignment tool would need to be able to capably and accurately handle the alignment of paired-end reads (as opposed to the simpler single end reads).
After considering each alignment tool option in regards to the aforementioned criteria, TopHat version 2.0.9 (or simply “TopHat2”) was chosen as it conforms to the desired standards and possesses the largest scope of features useful for this project from the available researched alignment options (Kim, et al., 2011). TopHat2 is built off of BowTie2, and, as such, is a fairly quick alignment tool that supports gapped and paired-end alignment. Unlike BowTie2, TopHat2 has the added benefit of an ability to find splice junctions without a reference annotation. TopHat2 is able to accomplish this by initially mapping reads to the genome and identifying possible exons where many reads continuously align. This possible exon information is compiled into a database that documents possible splice junctions, at which point TopHat2 then remaps the reads again to confirm the presence of the splice junctions. TopHat2 is able to do this by both remapping whole reads and splitting long reads into smaller segments and mapping the segments independently.

In order to examine the quality of reads alignments to S. bicolor gene models and novel sites identified during alignment, a pair of TopHat2 alignments utilizing sample 5Rc_Sep10 (rhizome tissue sampled on 9/27/2010, biological replicate “c”) were run. One run only allowed uniquely mapping reads to be aligned, while the second run allowed up to ten mapping locations before a read would be discarded. As a result, the run that only allowed uniquely mapping reads to be aligned ended up with 68.26% of M. × giganteus reads aligning to S. bicolor gene models or the novel sites determined by TopHat2, while the second run allowing up to ten maximum alignments had a total of 69.16% accepted mapping hits (Figure 6). As the difference between the two is a negligible 0.9%, most reads in this dataset map uniquely and it was decided that allowing ten maximum multi-hits would be beneficial to this project as it would allow for the collection of more accurate gene expression information without overwhelming the
data with repetitively mapped reads. Therefore, the small overall percentage of multi-hits combined with the quality of TopHat2’s multi-read correction ensured that the multiple alignment data would be beneficial and not detract from this project’s goal of quantifying seasonal changes in gene expression.

After trimming, more than 3.9 billion reads were aligned by this project (Figure 4). Approximately 2.8 billion of these reads aligned to the reference, giving an overall mapping percentage of 72.61%. This figure represents a substantial improvement over the previous two sequencings, as the 36 bp sequencing run possessed a mapping percentage of approximately 60.70% and the 76 bp sequencing run possessed a mapping percentage of approximately 59.04%. Whether this improvement can be attributed to higher quality reads, better alignments from longer reads, or better functionality of the alignment tool is unknown, though it is likely that each of these aspects played a role in the observed increase in mapping percentage. Individually, all of the mapping percentages of this projects alignments ranged from a low of 64.84% in the “10/29/12 Rhizome replicate C” sample to 89.01% in “9/27/10 Internode replicate C” sampling (Figure 7). The median mapping percentage for the entire dataset was 70.76%, while the mean mapping percentage was 72.42% with a standard deviation of 5.18%. For all of this project’s rhizome samples and replications, the median mapping percentage was determined to be 69.47% and the mean mapping percentage was 69.37% with a standard deviation of 1.71%. Across all internode tissue replications, the median mapping percentage was 71.80% and the mean mapping percentage was 72.99% with a standard deviation of 4.43%. Finally, for all samples and replications of leaf samples in this project, the median mapping percentage was 74.72% and the mean mapping percentage was 75.36% with a standard deviation of 6.37%. With these values under consideration, it appears that *M. × giganteus* leaf mRNA shares the most similarity with *S. bicolor,*
while the internode and rhizome share respectively decreasing amount of similarity. This could be from a high conservation of photosynthetic genes—which would logically dominate expression in most leaf tissue—between *M. × giganteus* and *S. bicolor* and/or it could be due to the uniqueness of the rhizome in *M. × giganteus* and the divergent expression patterns within *M. × giganteus* related to its presence.

After alignment with TopHat2, Cufflinks—specifically, the Cuffdiff tool included with the Cufflinks package—was used to test for differential expression. In the previous ten-sample *M. × giganteus* RNA-Seq project, CLC Genomics Workbench was used to determine unique gene counts and Microsoft Excel was used to normalize count information into RPKM (“Reads Per Kilobase of transcript per Million mapped reads”) values for the expression analysis. For this project, however, Cufflinks was used to analyze data and, as such, the normalization method employed was FPKM (“Fragments Per Kilobase of transcript per Million mapped reads”). According to the authors of Cufflinks, the key difference between RPKM and FPKM is that whole fragments instead of reads are measured, meaning that both reads produced from paired-end sequencing in a pair make up a fragment, and this fragment can be mapped as a whole even if the one of the two reads that make up the fragment map poorly (Cufflinks, 2013). Another important difference is that the length value in the FPKM normalization equation performed by Cufflinks is not simply the length of the transcript but is instead an “effective length” (Trapnell, et al., 2010). Calculating effective length is a very complex process in which the expected uniquely mappable area and the statistical weights of areas common to all isoforms of a gene and those specific to a single isoform are determined (Soohyun, et al., 2010).

Whether using an effective length in its calculation or not, it should be noted that both the RPKM and FPKM methods have be recently denigrated (perhaps rightfully),
especially when compared to newer options for normalizing RNA-Seq data (Dillies, et al., 2012; Soneson & Delorenzi, 2013). Two of the most common criticisms focus on an inherent unacceptable false-positive rate as well as a tendency to introduce bias in per-gene variances that occur from correcting for gene length, an issue that is compounded for genes with low levels of expression (Oshlack & Wakefield, 2009). This noted problem can be assuaged, yet not wholly eliminated, by using effective transcript length in place of “real” transcript length (Dillies, et al., 2012).

Fortunately, the version 2.1.0 of Cufflinks has a non-standard FPKM normalization method available in addition to the “classic” FPKM normalization that addresses these issues with RPKM and FPKM. According to the Cufflinks package authors, this method, titled “geometric-FPKM”, scales its read counts in a manner that is “identical to the one used by DESeq” (Cufflinks Manual, 2013) which is another popular and more widely lauded normalization method than that of RPKM/FPKM. The “geometric” of geometric-FPKM is due to this option’s method of scaling expression counts by the geometric mean’s median value across all libraries as illustrated by the DESeq paper (Anders & Huber, 2010). In comparative studies of different RNA-Seq normalization approaches, the DESeq method has a very small coefficient of variation and a minor rate of false positive calls while still maintaining a strong degree of detection power (Dillies, et al., 2012).

As the geometric-FPKM uses an identical method to DESeq to normalize the data, the expression values generated by the Cuffdiff tool should be highly accurate. As such, it is suspected that use of the classic-FPKM option in Cuffdiff would result in false positives not present in the geometric-FPKM data. To test the degree of potential false positive calls and the overall resulting differences between these two normalization methods, the number of gene models that had at least one call of significant gene
expression in all pairwise tests with a significance value of 0.01 for both the classic-FPKM normalization method and geometric-FPKM Cuffdiff results was determined (Figure 8). In this test, almost twice the number of normalization-method-specific gene models are present in the classic-fpkm data; however, as the amount of normalization-method-specific gene models in classic-FPKM is comparatively small when considering the whole dataset (i.e., a total of 763 gene models out of the total 33,815 models), it is quite possible that many of those arise from false positives. As such, it was decided to utilize the geometric-FPKM normalization method of Cuffdiff in order to take advantage of its greater degree of accuracy.

After FPKM values were determined, the novel splice junctions were examined. In total, TopHat2 discovered and utilized 7,104 splice junctions during alignment (Figure 9). The majority of these locations were between 200 and 1,000 bp in length. Unfortunately, more than half of these sites had low expression values (less than ten FPKM) for all samples, leading their presence to not account for a high degree of gene expression. Approximately one-thousand sites were found to have either a length over 1,000 bp or a value of over 50 FPKM in at least one sample, while approximately one-hundred sites fell into both of these categories. After a blastn megablast search against the nt database for these one-hundred sites, most appeared to either be originated within ribosomal, chloroplastic, or mitochondrial RNA (Table 4).

As the raw RNA-Seq expression data produced by Cuffdiff can be fairly cumbersome to navigate, a combination of cummeRbund tools (CummeRbund, 2013) and custom R scripts were utilized within RStudio (RStudio, Inc., 2013) to help efficiently analyze and visualize this project. In order to assess the degree of expression profile similarities between each pair of samples, a dendogram based on Jensen-Shannon distances utilizing all gene models in the dataset was built (Figure 10) as well as
a distance matrix heatmap to visualize the dendogram results at a more granular level (Figure 11). Values for Jensen-Shannon distance are built from determining Jensen-Shannon Divergence, which measures difference through the directed divergence between multiple weight-able probability distributions based on Jensen’s inequality and Shannon entropy (Lin, 1991). In performing this analysis, it becomes apparent that the three tissue types group into three categories, as expected, with only a few peculiarities. Perhaps most prominently of these peculiarities is the close grouping of the 10/26/09 Internode sample with the 10/26/09 Leaf sample, as this is the only case where two different tissue type samples are closer to one another than they are to tissues of their same type. However, when one considers that late stage of the growing season in which October 26th occurs, the similarity of these two tissue samplings to one another seems like a likely result of expression focus—that is to say, the leaf has moved away from a primary concentration on photosynthetic activities and is now devoted to pathways involving senescence and nutrient remobilization, a focus that would logically be very similar to that of the internode. In addition to this observation, it should be noted that the two samples of 10/26/09 Rhizome and 10/29/12 Rhizome are far more similar to one another than they are to the rest of the rhizome—as well as leaf and internode, for that matter—samples. As this pair of rhizome samples were harvested at very close to the same time point in the growing seasons of 2009 and 2012, they both likely have a similar comparable focus on nutrient storage that is not reflected in the expression profiles produced from the rhizome sampled during early times of the growing season.

Pairwise comparisons between interesting pairs of samples were made to find genes that were differentially expressed at an alpha significance level of 0.05. Lists of these genes, along with their fold-change values, underwent analysis via the PAGE (“Parametric Analysis of Gene Set Enrichment”) tool found at agriGO (Du, Zhou, Ling,
Zhang, & Su, 2010). The agriGO website hosts a variety of tools useful for automating the identification of enriched GO (“Gene Ontology”) terms within a given set of gene identifiers for a variety of agricultural genomes. As this particular project’s reference, *S. bicolor*, is present in the agriGO database and each *S. bicolor* gene model has its own GO terms associated with it, differential gene expression data built upon the *S. bicolor* gene models can be analyzed to determine significantly represented GO terms using this valuable freely-available online resource.

GO terms derive from the Gene Ontology project, which was originally instated in order to standardize terms associated with gene products (Ashburner, et al., 2000; The Gene Ontology Consortium, 2013). This standardization has helped further the ability to analyze data from different sources, as even though the sources may be different, having compatible standardized terms between the datasets facilitates quicker and simpler comparisons. For example, the same gene in two different organisms could be annotated as being involved in “transcription” in one organism but “RNA synthesis” in another, and while it’s easy to tell that these two annotations are the same when next to each other, when they are lost within a list of thousands of genes it becomes near-impossible for a human or computer to efficiently pair them up. As such, a standardized system to introduce consistency into gene annotations is a great boon.

With this in mind, the usefulness of agriGO’s PAGE tool becomes apparent. The gene ontology information gained from this type of analysis greatly aids in interpreting the functional activities most prominently behind the significant gene expression differences in a chosen pairwise comparison. Having a large set of genes that are known to be significantly up- or down-regulated between two samples of interest is certainly useful; however, there still remains the issue of analyzing the data to determine the overall “big picture” purpose(s) of the differentially expressed genes. Therefore, an
analysis of enriched GO terms, such as that afforded by agriGO’s PAGE tool, presents a convenient and reliable method to identify groupings of gene expression related to particular functions or traits, thus better clarifying the role of differentially expressed genes in that particular comparison. The PAGE method was originally designed for microarray analysis that improved upon the previously used Gene Set Enrichment Analysis (GSEA) method by reducing bias introduced when working with larger datasets (KIm & Volsky, 2005). As PAGE is based on the central limit theorem, the PAGE method assumes that as the number of random samples in a dataset increases, a normal distribution should begin to occur even when the entire dataset’s population is not normal. As a result, PAGE is a more powerful method that benefits from being able to identify more gene sets that are significantly changed than the GSEA method.

3.3.2 Jasmonic Acid, Revisited

One application of the agriGO PAGE tool on this dataset is the further investigation of the role of jasmonate in *M. × giganteus* as it relates to seasonal development, particularly that of the *M. × giganteus* rhizome. As previously stated at the beginning of this chapter, it was noticed in the previous ten-tissue RNA-Seq study that the plant growth regulator jasmonic acid appeared to play an important role in the activity of underground *M. × giganteus* tissues, particularly in the Spring Rhizome sample, as there was a noticeable amount of preferential gene expression in this rhizome sample for jasmonate-related transcripts. However, as there was only a single unreplicated “Spring Rhizome” sample in the previous ten-tissue RNA-Seq project, it was impossible to determine the possible purpose and the extent of jasmonic acid’s role.
The completion of this study’s replicated seasonal \textit{M. \times \textit{giganteus}} dataset presents an opportunity to better explore the role of jasmonate-related gene expression in \textit{M. \times \textit{giganteus}}. To this end, pairwise comparisons of tissues sampled in the spring versus those sampled in the fall for each of the three tissue types sampled (leaf, internode, and rhizome) were completed by means of custom R scripts and cummeRbund using a significance value of 0.05. The results of these comparisons were imported into agriGO’s PAGE tool for GO term enrichment analysis. As the role of jasmonic acid appeared to play the largest role in the rhizome from the information gleamed in the ten-sample project, the samples for rhizome tissue gathered between the earliest-in-season sample (May 5$^{th}$, 2011) and the latest-in-season sample (October 28$^{th}$, 2012) could be compared to one another to see if GO terms associated with jasmonic acid had significant changes between these two samples (Figure 12). When examining this particular GO term enrichment analysis, it becomes apparent that there is a far greater focus on the biosynthesis of jasmonic acid in the spring 5/5/11 Rhizome sample than that of the fall 10/28/12 sample. In addition to jasmionate-related activities, there are a number of other plant hormone activities—including responses to auxin, ethylene, and abscisic acid stimuli—that are far more represented in the earlier springtime sampling of rhizome than the late fall sampling. However, there is a notable presence of a defense response to bacterium present in the spring rhizome sample, which means that this single pairwise comparison is not enough to eliminate the possibility of an increase in jasmonic acid biosynthesis that is occurring as a part of a defense response.

Therefore, in order to look at jasmonate in a more complete manner, pairwise test information for many spring-fall pairs of different tissue types was compiled and examined as a whole (Table 5). Once completed, this information appears to point to the most prominent period of jasmonic acid biosynthesis being that in the earliest possible
rhizome sample. If jasmonic acid does indeed play a role in the mobilization of nutrients from the rhizome to the growing plant, this is a logical result as most mobilization would be occurring only in the earliest part of spring. Indeed, it may even be possible that the highest point of rhizome-based jasmonic acid biosynthesis occurs immediately following the plant’s release from dormancy in spring. Unfortunately, this study lacks the necessary dormancy-breaking time period rhizome sample to either confirm or deny this notion.

In addition to the high-point of GO term related jasmonic acid biosynthesis, the 10/26/09 Leaf sample has the compelling result of having a high degree of enrichment of GO terms related to the response of jasmonic acid stimulus without having an enrichment of GO terms related to defense responses or responses to wounding while also having a concurrent high degree of GO term enrichment focused on amino acid transport. As amino acids are known transporters of nitrogen, an essential plant nutrient, there may be a correlation between the mobilization/remobilization of nitrogen in response to jasmonic acid, as a similar increase in amino acid transport can be seen when the spring rhizome samples also have an enrichment of GO terms related to amino acid transport at the same time that the response to jasmonic acid stimulus occurs.

In the previous ten-tissue RNA-Seq study, three *S. bicolor* jasmonate-related gene models in particular stood out as possessing rhizome or underground-tissue preferred expression: Sb01g033020, Sb01g045190, and Sb01g045180. All three of these gene models are annotated as “jasmonate ZIM-domain protein.” In *Arabidopsis*, jasmonate ZIM-domain transcripts are found to become highly expressed in response to events which also increase systemic jasmonate levels, and it is thought that these jasmonate ZIM-domain proteins play a role in regulating the extent and manner in which an internal jasmonate cause-and-effect response occurs (Chung, et al., 2008). The
overall FPKM profiles for these three gene models were examined using the new dataset in order to determine their seasonal profiles and periods in which their expression is significantly altered (Figure 13). In these profiles, there appears to be two patterns of gene expression: one in which the only prominently high level of gene expression occurs solely in the earliest rhizome sampling (5/5/11), and another where expression starts high in the spring rhizome and then tapers off to obscurity for the rest of the season while simultaneously starting low in the leaf and internode samples, but then dramatically rising at the end of the season in both leaf and internode. Whereas the first pattern could be interpreted as a result of a multitude of activities, ranging from a seasonally-localized defense response or a jasmonate-regulated period of rhizome growth and development, the second pattern is supremely interesting as it coincides with the pattern of nutrient mobilization expected to be observed: a high amount of mobilization from the rhizome to the rest of the plant in the spring which disappears during the summer and is subsequently replaced in the fall by a high amount of nutrient remobilization from the aboveground tissues back into the rhizome for storage over winter.

To examine this further, the jasmonic acid precursor OPC-8:0 was studied within the new seasonal dataset. OPC-8:0—also known as 3-oxo-2(2'-pentenyl)-cyclopentane-1-octanic acid—is produced in the cytoplasm by genes such as OPR3 (12-oxophytodienoate reductase 3) before it is imported into the peroxisome to undergo three rounds of β-oxidation and finally becomes jasmonic acid (Leon & Sanchez-Serrano, 1999). In the seasonal dataset, two 12-oxophytodienoate reductase genes were found that match the dual patterns observed in the jasmonate ZIM-domain protein profiles (Figure 14). For these genes, Sb07g022500 matched the “solely in spring rhizome” pattern while Sb10g007320 aligns with the “high in spring rhizome, fall leaf, and fall internode”
pattern with the exception of a sharp decline between the second-to-last (10/8) and the very last (10/29) fall samplings of leaf and internode tissues, which creates a sudden drop off in expression from the steadily rising peak observed in the preceding fall-sampled data points (8/17 to 10/8). A possible explanation of this occurrence is that, at this particular time in the growing season, there is an adequate level of endogenous jasmonic acid, which no longer necessitates the production of more jasmonic acid precursor. Yet, whatever the cause, it has quickly become apparent that the role of jasmonic acid in *M. × giganteus* is exceedingly complex. The seasonal changes in the expression of jasmonate-related genes in *M. × giganteus* appear to have some relationship to the developmental and growth-related priority changes of the plant, such as the mobilization of nutrients away from the rhizome in the spring and the subsequent remobilization of nutrients back to the rhizome in the fall, which occur as the growing season progresses.

3.4 Conclusion

When the entire wealth of information supplied by this RNA-Seq project is considered, examining the seasonal role of jasmonic acid in *M. × giganteus* is only a slight morsel of what could possibly be learned about *M. × giganteus* through the proper study of this newly acquired dataset. For this project, approximately four billion reads of three different tissue types from a pool of twenty-four samples were sequenced and aligned to *S. bicolor* gene models and analyzed using modern, freely-available RNA-Seq tools and custom written scripts. Of course, *M. × giganteus* isn’t alone on this RNA-Seq frontier, as the combination of decreasing sequencing costs and advancing analysis technology has recently paved the way for more and more expression analysis studies on
non-model plant organisms such as tomato (Van der Hoeven, Ronning, Giovannoni, Martin, & Tanksley, 2002), soybean (Shoemaker, et al., 2002), potato (Ronning, et al., 2003), sugarcane (Vettore, et al., 2003; Ma, et al., 2004), apple (Newcomb, et al., 2006), lodgepole pine (Parchman, Geist, Grahnen, Benkman, & Buerkle, 2010), and grape (Venturini, et al., 2013). The newly formed dataset created by this project expands this list further, allowing for the analysis of seasonal changes in gene expression within M. × giganteus.

Many expression-related focuses concerning the development of M. × giganteus and its physiological traits throughout the growing season could be deciphered through examining gene expression patterns present in this RNA-Seq project. As a result, the expression profiles created through an RNA-Seq analysis of this dataset are useful for a number of downstream studies, as the knowledge of where and when a gene is being up- or down-regulated can reveal more about the function and purpose of the gene within M. × giganteus. The accumulation of information on gene expression profiles as they relate to time-of-season allows for better the identification and characterization of gene expression patterns related to M. × giganteus’ seasonal development, resulting in an increased capacity to study traits such as biomass accumulation, nutrient use, and responses to stress in M. × giganteus. Overall, the creation of this M. × giganteus dataset provides for a deeper understanding of the growth and development of this important biomass-for-biofuel crop.
CHAPTER FOUR

“ANALYSIS OF SEASONAL NITROGEN MOBILIZATION AND REMOBILIZATION IN MISCANTHUS × GIGANTEUS”

4.1 Introduction

The completion of the seasonal \textit{M. × giganteus} dataset described previously in chapter three opened the door to an enormous landscape of possible studies on gene expression patterns in this noteworthy potential biofuel feedstock source. To this end, a focused analysis on nitrogen use in \textit{M. × giganteus} via the newly constructed seasonal RNA-Seq dataset, as well as separate analyses of both elemental and amino acid content changes throughout the growing season, has been completed to better understand the seasonal mobilization and remobilization of nitrogen within \textit{M. × giganteus}.

As the key component of amino acids, nitrogen is an essential plant nutrient. The availability of nitrogen coupled with a plant’s efficiency at nitrogen uptake and internal nitrogen use (i.e., the plant’s nitrogen use efficiency, or “NUE”) are crucial drivers of plant growth, as nitrogen is known to promote plant growth to a large extent and, similarly, deficiency in nitrogen can place a severe burden on the health of a plant that can drastically limit plant biomass production (Zhao, Reddy, Kakani, & Reddy, 2005; Lawlor, Lemaire, & Gastal, 2001). For plants that require nitrogen to be obtained through soluble compounds in the soil (i.e., those without symbiotic nitrogen-fixing organisms), crop growers are often required to provide nitrogen to the crop via nitrogen fertilizer in order to maximize yield.
As a result, large quantities of nitrogen fertilizers are being applied to farmlands. In 2011 alone, almost thirteen million nutrient short tons of nitrogen were applied to fields in the United States, a figure almost twice the amount used in 1970 (USDA ERS, 2013). Since \( \textit{M. \times giganteus} \) is a proposed biomass-for-biofuel crop, it is vital that the plant produces as much biomass as possible in a growing season while still remaining cost effective. If a farmer is required to spend more money on nitrogen fertilizer than he/she will receive upon selling the crop after harvest, then there is little incentive for the farmer to grow the crop. However, as \( \textit{M. \times giganteus} \) is a perennial crop, its ability to remobilize and store nutrients over the winter reduces the requirement of exogenous nitrogen application by the farmer for the next season’s growth.

As a result, a plant’s ability to efficiently use and re-use nitrogen plays a significant role in a plant’s overall fitness strategy. In annual plants, a substantial change is undertaken in nutrient-use priorities and physiology at the end of the growing season, as the plant typically reprioritizes away from its own growth and instead focuses on increasing reproductive fitness by producing the most viable seed possible (Christensen, Below, & Hageman, 1981). During this period, proteins are broken down in vegetative tissues in order to recover nutrients that are then mobilized to the seed (Schiltz, et al., 2004). While this process is pivotal for the survival of an annual plant’s progeny, perennial crops can utilize a simultaneous alternative mechanism to ensure another healthy season of growth in the next year. By remobilizing nutrients into a storage tissue such as a rhizome for sequestration during the winter, a perennial plant ensures that there will be nutrients available for its own growth in the future, thereby increasing its own fitness.

An example of this behavior can be seen in sugarcane, a close relative of \( \textit{M. \times giganteus} \) that stores carbohydrates as sucrose in its stalk for later utilization (Hatch &
Glasziou, 1963; Pan, Luo, & Li, 2009). *M. × giganteus* behaves somewhat similarly to sugarcane in this regard, but instead appears to remobilize nutrients to the rhizome at the end of the growing season so that it can survive through the winter. However, the *Miscanthus* rhizome is not as well-studied as the sugarcane stalk and much about the rhizome remains unknown. The differences between the functional behavior of the sugarcane stalk and the *Miscanthus* rhizome, as well as the intricacies of nitrogen mobilization, storage, and remobilization in the rhizome, could very well be a vital component of *M. × giganteus*’ ultra-productiveness. Additionally, it should be noted that as a sterile plant *M. × giganteus* is unable to produce viable seed. Due to this, the perennial habit of nutrient recycling is of the utmost importance in *M. × giganteus*: since *M. × giganteus* does not produce usable seed for the next generation of growth, the perennial growth of existing *M. × giganteus* plants is the only fitness strategy available to the crop.

Many of the nitrogen cycle’s common characteristics in higher plants are very well studied. Generally speaking, nitrogen assimilation in plants mostly begins when nitrate (NO$_3^-$) is absorbed through the roots and then is either stored in the vacuole, transported through the xylem to leaf mesophyll cells, and/or reduced into ammonia or ammonium (NH$_3$ or NH$_4^+$) within root cortical and leaf mesophyll cells. This ammonia/ammonium can then be fixed into amino acids, which are the major long-distance transporters of nitrogen through the xylem and phloem (Herschbach, Gessler, & Rennenberg, 2012; Peuke, 2010). For example, ammonia/ammonium can be fixed onto glutamate to produce glutamine via glutamine synthetase, which can then be transported to wherever it is needed within the plant through the xylem or phloem (Abrol, Chatterjee, Kumar, & Jain, 1999). Of all possible amino acid fates for nitrogen, the most commonly found forms of amino acids transported in herbaceous plants are glutamine,
glutamic acid (i.e., glutamate), aspartic acid (i.e., aspartate), and serine (Herschbach, Gessler, & Rennenberg, 2012; Amiard, et al., 2004; Arlt, Brandt, & Kehr, 2001; Lohaus & Moellers, 2000), with glutamine being the most efficient carrier of nitrogen among the four, as it carries two nitrogen atoms as opposed to the single nitrogen atom carried by glutamate, serine, and aspartate. In addition, asparagine synthetase can convert glutamine into asparagine, another dual-nitrogen containing amino acid that has been shown to be involved in long distance transportation of nitrogen (Pate, 1980).

Although both asparagine and glutamine are more efficient carriers of nitrogen, there is a key difference separating their roles in nitrogen transportation: glutamine possesses a very active role in metabolism and donates nitrogen for the biosynthesis of the vast majority of nitrogen-requiring compounds (Kung & Wu, 1992), whereas asparagine is a comparatively inert transporter for nitrogen that possesses only a small net charge under physiological conditions and has a soluble form that is only used as a substrate in a minor amount of enzymatic reactions (Lea, Sodek, Parry, Shewry, & Halford, 2007). These characteristics lend asparagine the capability of being an effective transporter of nitrogen suitable for long-distance movement in addition to being able to be stored in the vacuole, and, as such, makes asparagine a supremely effective molecule for both nitrogen movement and nitrogen storage (Kung & Wu, 1992).

There is also a substantial difference in the ease with which the nitrogen that asparagine and glutamine carry can be utilized, as glutamate synthase can be used to transfer glutamine’s nitrogen to α-ketoglutarate to form glutamate—which can then be used in a variety of downstream transamination reactions—or glutamine could be hydrolyzed by glutaminase, whereas asparagine needs to be hydrolyzed by asparaginase in order to free its ammonium (Skokut, Varner, Schaefer, Stejskal, & McKay, 1982). This further reinforces the role of glutamine as a transporter of nitrogen more suited for post-
transportation utilization, whereas asparagine is a transporter that appears to generally be more suited for having a post-transportation storage role. Since these two amino acids are the predominant carriers of dual nitrogen atoms in herbaceous plants, following their biosynthesis and movement through different tissues of *M. × giganteus* during different points in the growing season will aid in the clarification of seasonal nitrogen mobilization and remobilization patterns and behaviors within *M. × giganteus*.

4.2 Materials and Methods
4.2.1 Elemental Analysis

Perhaps the most appropriate way to begin a study into the long-distance transportation of nitrogen within *M. × giganteus* is to examine the changes in raw elemental nitrogen content throughout the growing season. To this end, the exact same tissue stock from the twenty-four sequenced tissues and their biological replicates in the seasonal RNA-Seq dataset (Table 3) was subsetted for elemental quantification via combustion analysis. Each sample (seventy-two in total: twenty-four samplings, each with three biological replicates) was first lyophilized and then individually wrapped in tin before being weighed. Samples were then given to the Genomic Ecology of Global Change Research theme of the Institute for Genomic Biology at the University of Illinois, where they were ran on an Costech CHCS-O (ECS 4010) elemental combustion system according to the manufacturer’s instructions (Costech, 2002). In this process, each sample was flash combusted and analyzed via gas chromatography for carbon and nitrogen content.
4.2.2 Amino Acid Analysis

Another subset of the exact same tissue stock used for both sequencing and elemental analysis was also used for amino acid analysis. For each of the seventy-two samples in this study, lyophilized tissue for each sample was first weighed and aliquoted into individual tubes. Five-percent trichloroacetic acid was added to each sample at a rate of 400 µl for every 100 mg of tissue while keeping both the trichloroacetic acid and the samples on ice. Tubes were then vortexed every ten minutes for one hour. Once the hour had passed, all samples were centrifuged for ten minutes at maximum rpm and at 4°C. Following this, samples were moved to a 96 well filter plate and filtered using a vacuum filtering apparatus into a fresh 0.5 ml 96 well plate. Samples were then stored at -20°C until they could be analyzed via HPLC.

4.2.3 Expression Analysis

The RNA-Seq dataset generated by Cuffdiff in chapter three was examined using a combination of cummeRbund tools, custom-written R scripts, and Microsoft Excel for this analysis of nitrogen transportation. Genes related to the many aspects of nitrogen metabolism and transportation were divided into groups based on role in the regulation, use, and/or biosynthesis of specific amino acids. When these lists became too large, gene models were subsetted to allow for more efficient analysis by removing any gene model that did not have a score of at least 50 FPKM in at least one tissue sample. Resulting gene lists were then analyzed for expression that appeared to have season-dependent changes in gene expression. When one particular gene required more detailed investigation, the “findSimilar” tool of cummeRbund was used to search for other genes.
with similar expression profiles. When a group of genes required further analysis, cummeRbund’s “csCluster” tool was used to group genes into clusters based on similarities in their expression profiles via K-means clustering using Jensen-Shannon distance between each sample.

Guide trees to cluster sequences by their relatedness were produced for gene families of interest by first retrieving protein sequence for *S. bicolor* and *Z. mays* from Phytozome (www.phytozome.net) and *M. × giganteus* contigs and their Interproscan annotations from their online repository at JGI (ftp://ftp.jgi-psf.org/pub/JGI_data/Miscanthus/transcriptome/). All annotations are determined by *Arabidopsis* best hits. *M. × giganteus* contigs were translated using the translation tool available on ExPASy (http://web.expasy.org/translate/) and all sequences were collected and imported into Jalview and aligned using Clustal with default settings. Unaligned ends and gaps of all imported sequences were trimmed so that an equal size fragment for all inputs that would represent an identical region in each were generated; *M. × giganteus* contigs that aligned poorly and resulted in a sum fragment size of less than 150 residues were removed from the alignment. Once completed, the file was exported from Jalview into the Clustal Omega tool available at EMBL-EBI (http://www.ebi.ac.uk/Tools/msa/clustalo/) and run with the following settings: use Clustal with numbers, do not realign input sequences, and use mBed-like clustering guide tree and iteration.

4.3 Results and Discussion

By analyzing changes in *M. × giganteus’* elemental nitrogen content throughout the growing season, information concerning resource availability and environmental
factors for the growing plant is revealed. For example, the concentration of nitrogen in a plant is known to vary due to changes in organ growth and nutrient accumulation, as cooler temperatures negatively impact the rate of organ growth more than it does nitrogen uptake and assimilation, which leads to an overall increased ratio of nitrogen to carbon (Lawlor, 2002). Similarly, by studying changes in amino acid content over the growing season, patterns in amino acid production and transportation that are relevant to the use and storage of nitrogen within *M. × giganteus* over the growing season can be identified. On their own, both of these individual resources are valuable. However, when information generated from the elemental analysis is combined with that of both the amino acid analysis and the RNA-Seq study, a more complete understanding of nitrogen mobilization and remobilization can be acquired.

By utilizing the same stock of collected seasonal leaf, stem, and rhizome samples in both the elemental and amino acid analysis that was also used for the high-throughput sequencing project, it is possible to obtain a better representation of long-distance nitrogen transportation in *M. × giganteus* through the comparison of each sample's results and the seasonal patterns that occur within these analyses. Furthermore, when the data generated by these analyses is compared to nitrogen-transportation related gene expression, the extent to which gene expression fluctuations physiologically manifest can be observed.

Firstly, when examining the elemental changes in nitrogen throughout the growing season within our collected *M. × giganteus* samples, a few patterns are identifiable that clarify how the plant is partitioning its nitrogen as the growing season progresses (Figure 15). Overall nitrogen concentration appears the highest in the earliest spring sampling of leaf, which then steadies out through the summer and slowly declines as fall progresses. As the earliest spring sample represents a period of focus centered on
growth, it is reasonable to assume that this high amount of nitrogen is being utilized by the leaf sample to promote its rapid growth. Then, once summer has fully arrived, the plant can essentially reach the maximum quantity of leaf biomass that it can sustain. When this occurs, the nitrogen content in the leaf stabilizes until fall senescence begins, at which point the plant then begins reallocating nitrogen to the rhizome to be stored for the next growing season.

For the internode sample, nitrogen concentration is stable through summer but a small increase is observed at the end of the season. A likely explanation for this observation is simply that at the end of the season the above-ground plant is remobilizing nitrogen to the rhizome for storage, and in doing so, nitrogen is passing through the internode in quantities higher than those that occurred during the rest of the growing season. As a result, it can be observed that nitrogen concentration in the fall begins to drop in the leaf at approximately the same time it begins to increase in the internode. This change in nitrogen concentration for both the leaf and internode at the very end of the year (i.e., between the 10/8/12 and 10/26/12 samples) is a statistically significant difference (t-test, p<0.01).

When the nitrogen concentration of the rhizome is studied, it appears that internal nitrogen levels are fairly steady all season long, with a possible small decrease at the end of the season. What needs to be considered when observing this is the fact that, unlike the senescing above-ground tissues, the rhizome is still a growing tissue as the end of the season approaches. Although overall nitrogen concentration is stable, the biomass of the underground rhizome network is increasing; as such, it appears that remobilized nitrogen is not used to “load up” existing rhizomatic tissue with nitrogen but is instead being used to promote the growth of new rhizomes and/or is being stored in newly formed rhizomatic mass. This observation could account for a portion of the
plant’s stand density increases that occur in *M. × giganteus* plots as time passes (Lewandowski, Clifton-Brown, Scurlock, & Huisman, 2000), as nitrogen is being partitioned not only for the next season’s growth from existing rhizomes, but also to new rhizomes in the same area of the plot that will sprout the next generation of the stand’s *M. × giganteus* plants.

The elemental analysis of nitrogen is very useful for observing changes in overall nitrogen content as it is mobilized and remobilized throughout the growing season, but it does little in the way of revealing what types of long-distance transportation vessels that the nitrogen is utilizing as it moves through the plant. As previously mentioned, the five most commonly found amino acid long-distance carriers of nitrogen in herbaceous plants are glutamate (i.e., glutamic acid), glutamine, asparagine, aspartate (i.e., aspartic acid), and serine, with both asparagine and glutamine being the more efficient carriers of nitrogen as they possess two units of nitrogen whereas the other three carriers possess one unit of nitrogen. The presence of the more efficient dual-nitrogen carrying amino acids with respect to their required single-nitrogen carrying amino acid precursor substrates was assessed in the internode tissue (Figure 16). For the leaf and rhizome tissues, there does not appear to be a specific trend for the ratio of dual-nitrogen carrying amino acids in relation to their single-nitrogen carrying precursor substrates, with the possible exception of a dip in glutamine-to-glutamic acid, as well as asparagine-to-aspartic acid ratio, in the leaf tissue at the near the end of the growing season (10/8), which is followed by a subsequent return to heightened levels of the more efficient long-distance amino acid transporters of nitrogen in the final sampling (10/26). A possible explanation of this could reflect an expected feature of nitrogen remobilization, where first free amino acids are exported, which is then followed by the exportation of protein degradation products. The 10/8 sampling could reflect the depletion of stored free amino
acids, and the spike in free amino acids in the last sampling occurring as a result of the sudden shock of chilling temperatures and final proteolysis of remaining proteins that are not transported to the rhizome for storage.

Perhaps most noticeable from this comparison, however, is the increasingly heightened level of both asparagine and glutamine to their precursors (aspartic acid and glutamic acid, respectively). These changes may reflect an increased focus as the growing season draws closer to its end on long-distance transport nitrogen through the internode as the more efficient amino acid carriers of nitrogen instead of their less efficient substrate amino acids. The pattern of amino acid accumulation is consistent with senescing above-ground tissues favoring the transport of nutrients for overwinter storage and the next season’s growth. In fact, at the last point of sampling for these two tissues (10/26/12), the ratio of asparagine-to-aspartic acid and glutamine-to-glutamic acid has increased to over eleven-fold of the more efficient amino acid carrier of nitrogen compared to its less efficient nitrogen carrying substrate. This end-of-season increase in both the asparagine-to-aspartic acid and glutamine-to-glutamate ratio represents a substantial, statistically significant (t-test, p<0.01) increase in the presence of asparagine and glutamine over their single nitrogen-unit substrates from the late-spring/early-summer internode sample (6/13/09).

In order to better understand the potential transportation roles of the five commonly known amino acid nitrogen carriers in herbaceous plants, bias for one or more particular amino acid out of the five was studied (Figure 17). From this data, it appears that both leaf and internode have large amounts of glutamine in comparison to the other four transporters at every sampling, while the rhizome typically contains asparagine at the most prominent levels. In the internode samplings, a steady decrease in the amount of aspartate/aspartic acid and glutamate/glutamic acid can be observed
across the late fall samples of 9/17/12, 9/27/10, 10/8/12, and 10/26/09. As aspartate is the precursor for asparagine and glutamate is the precursor for glutamine, the decreasing levels of these two amino acids appears to indicate that their role is shifting away from building proteins for growth and development and instead being transported themselves or incorporated into glutamine and/or asparagine for remobilization. Additionally, despite serine being documented as an important long-distance transporter for nitrogen, this dataset seems to indicate that the role serine plays in *M. × giganteus* seasonal mobilization/remobilization is not of the same magnitude as that of glutamine and asparagine, as comparative levels of serine are small and relatively stable for all samplings and there seems to be no increase in serine levels to correspond with seasonal mobilization/remobilization.

Documenting elemental nitrogen content and amino acid presence builds a good framework for understanding seasonal mobilization and remobilization of nitrogen in *M. × giganteus*, but examining gene expression related to nitrogen utilization and the biosynthesis of nitrogen’s known long-distance amino acid carriers allows for examining season-based nitrogen transportation at a more fundamental level. As previously stated, the most common first step of nitrogen uptake and assimilation is importing nitrate into root cells. Once nitrate is in the root cell, it is either directly stored as nitrate in the vacuole, incorporated into amides that are transported to leaf mesophyll cells through the phloem, or it simply remains nitrate and is transported to leaf mesophyll cells through the xylem. To study the transportation of nitrate, genes with nitrate transportation annotations were study and examined to determine if there was any expression changes that appeared to occur as a result of time-of-season (Figure 18). Three nitrate transporter genes were found that possessed a change in their expression that appeared to be related to season: Sbo1g029470, Sbo3g025300, and Sb10g022800.
The Sb01g029470 gene has a dramatic level of expression solely in the earliest rhizome sampling, making this gene appear to play a role in remobilizing stored nitrate or new nitrate away from the rhizome to the above-ground growing plant. Once nitrate arrives in a leaf mesophyll cell, it can either be shuttled to the nucleus to make mRNA and nitrogen-containing enzymes, stored in the vacuole until it is needed, or incorporated into amides. One possible explanation for the Sb03g025300 gene showing preferential expression for late season expression that increases as the season progresses in both the leaf and internode is that this particular gene is aiding in the remobilizing of nitrate that had been stored in leaf vacuoles back to the rhizome through the xylem. The Sb10g022800 gene, on the other hand, has high rhizome expression in early spring that quickly fades away in summer yet possesses peaks of expression in late fall for both leaf and internode, indicating that this gene is seasonally regulated for all three tissue types and is utilized to mobilize nitrate away from the rhizome in the spring and then back to the rhizome in the fall.

As previously stated, if nitrogen isn’t being transported as nitrate, then it is likely being moved as an amino acid. Before nitrate can be incorporated into amino acids, however, it must be reduced first into nitrite by nitrate reductase and then once more by nitrite reductase into ammonia. To this end, a few nitrate reductase and nitrite reductase genes were found that appeared to have changes in expression due to a tissue’s current time-period in the growing season (Figure 19). The Sb04g024300 nitrate reductase gene has its most notable expression solely in the earliest spring rhizome, making it a potential spring-only entry point for nitrate assimilation into amino acids within the rhizome for either transportation of nitrogen to the above-ground tissues or for growth and development of the spring rhizome. Conversely to the expression of Sb04g024300, the Sb04g034470 and Sb07g022750 nitrate reductase genes see their most significant
expression late in the season, possibly as a part of producing amino acids for the transportation of nitrogen back to the rhizome for over-winter storage. When proteins generated by these two genes are compared with translated \( M. \times giganteus \) contigs assembled from the ten-tissue sequencing project described in chapter two, the translation of the \( M. \times giganteus \) contig “Mxg_TContig47204” appears to be very similar to the Sbo4g034470.1 “Nitrate Reductase 1” protein. As such, this particular contig possesses a high probability of corresponding with the seasonal expression profile observed for Sbo4g034470 and would be a suitable starting point for a future study on seasonal regulation of nitrate reductase in \( M. \times giganteus \) (Figure 20).

Only a single nitrite reductase gene, Sbo4g034160, which appeared to have expression changes that align with seasonal changes was found in this dataset. This nitrite reductase gene is very weakly expressed in the rhizome all season long, but there are significant increases in its expression at the end of the season in the leaf and internode tissues, making it similar to the Sbo4g034470 and Sbo7g022750 nitrate reductase genes in that this gene may also be involved in the production of late-season amino acid nitrogen transporters for remobilization of nitrogen back to the rhizome.

After nitrate is fully reduced to ammonia, the next step for its assimilation into the predominant long-distance amino acid carriers of nitrogen is its incorporation with glutamate in order to produce glutamine. The interaction between glutamate and glutamine in carrying nitrogen is a complex balance that requires a number of different enzymes (Figure 21). Therefore, the next logical step to chart the production of long-distance nitrogen carrying amino acids for seasonal mobilization and remobilization is to examine the biosynthesis and utilization of glutamate (Figure 22). Glutamate synthase catalyzes the reversible reaction of generating two glutamates from glutamine and 2-oxoglutarate (Suzuki & Knaff, 2005). As such, it plays an important role in the cycling
between glutamine and glutamate. In this dataset, only one glutamate synthase gene appeared to have season-based changes in gene expression: Sb02g041740, a gene which shows the strongest relationship with M. × giganteus contig “Mxg_TContig38056” and Z. mays' GRMZM2G036609_T01 “Glutamate Synthase 1” (Figure 23).

Of particular note in this guide tree developed for glutamate synthase is the close clustering of the translated M. × giganteus “Mxg_TContig35815” and “Mxg_TContig49333” contigs. When contigs were assembled to build the transcriptome (see chapter two), it was known that most paralogous genes (i.e., homologous genes that are present due to a past genome duplication event) would likely collapse during assembly if not enough read depth and/or sequence differences were present/detectable, resulting in the construction of single contigs that would represent the assembly of reads originating from both paralogs. If it were the case that contig assembly and annotation worked perfectly, there would be a cluster on a finished guide tree that possesses two translated M. × giganteus contigs—one for each paralog—grouped tightly together, followed by a branch of a closely related Z. mays and/or S. bicolor proteins. This pattern is observed with “Mxg_TContig35815” and “Mxg_TContig49333,” which were both annotated via InterProScan as being related to glutamate synthase, as well as the closely related NADH-dependent glutamate synthase 1 protein “GRMZM2G085078_T01” from Z. mays. The likely reasons for this pattern not occurring more often in the guide trees produced for this project are twofold: firstly, it is likely that the two paralogs collapsed during assembly in many cases, which produced only a single corresponding contig as a result, and, secondly, in order for a translated M. × giganteus contig to be considered for use in a guide tree for this project, it needed to possess an annotation related to the family of interest (in this case, glutamate synthase). Specifically, this means that the protein domain from which annotation information originates would need to be
conserved in both paralogs. As many paralogous genes do not retain identical functionality to one another, it is also possible to have cases where individual contigs could have been generated for both paralogous genes, yet differences in protein domain presence and preservation didn’t identify both contigs as having similar annotation to one another, leading to only the contig with the proper family annotation being used in the generation of the guide trees.

The FPKM profile for the Sb02g041740 glutamate synthase gene found to have season-based changes in its expression shows fairly stable expression for each tissue until the end of the season approaches, at which point expression levels begin to rise in leaf, internode, and the rhizome (Figure 22). In particular, the leaf tissue shows a sharp increase in expression at the beginning of fall that slowly declines near the end of the season. The increases noted for the tissues as fall begins could signal a focus of this gene helping begin the synthesis of amino acids for nitrogen transport.

Another important gene for glutamate biosynthesis is glutamate dehydrogenase, which catalyzes the reversible reaction of the conversion of glutamate and water into 2-oxoglutarate and free ammonium. Along with glutamate synthase, glutamate dehydrogenase has been shown in past studies to play an important role in nitrogen metabolism (Robinson, et al., 1991). As 2-oxoglutarate is needed to synthesize glutamate, and it plays an important role in the potential future freeing of nitrogen from glutamine-derived glutamate as well, glutamate dehydrogenase and glutamate synthase can be seem as biochemical partners in the process of making long-distance transport of nitrogen via glutamate and/or glutamine possible. In this dataset, a glutamate dehydrogenase gene, Sb06g024150, was found that closely matches the seasonal pattern of mobilization/remobilization that this project seeks to study (i.e., high in rhizome only in the spring, as well as high in leaf and internode only in the fall) (Figure 22). As such, it
is likely that this specific glutamate dehydrogenase plays an important role in season-based nitrogen mobilization and remobilization. The protein produced by this gene matches closely with the protein translated by the *M. × giganteus* contig “Mxg_TContig46072”, and both of these two are grouped fairly closely to the maize “GRMZM2G427097_T01” glutamate dehydrogenase 2 protein (Figure 24), making this cluster of genes and contigs a potential candidate for a future cross-genus comparison of glutamate dehydrogenase in the Poaceae family.

Although there is only a single glutamate synthase and glutamate dehydrogenase gene found with changes in expression that correspond with time-of-season, there are a few additional glutamate-related genes that appear to have interesting preferential expression for certain tissues during certain periods of the growing season. Glutamate decarboxylase is used to catalyze the decarboxylation of glutamate into gamma-aminobutyric acid (GABA), a product that is typically associated with stress responses but has been shown to play a role in nitrogen metabolism and the balance of carbon-to-nitrogen ratios in *Arabidopsis* (Bouche & Fromm, 2004). The glutamate decarboxylase gene Sb07g022670 shows an expression increase in all tissues at the end of the growing season, while the glutamate decarboxylase gene Sb01g041700 shows high expression in spring rhizome tissue that disappears in the summer and never rebounds as well as low expression in leaf and internode until the end of the growing season, at which point its expression rapidly rises, possibly indicating a use of GABA in *M. × giganteus* nitrogen metabolism.

Lastly, a glutamate:glyoxlate aminotransferase gene (Sb02g000780) was discovered in this dataset that appears to be up-regulated in the fall for all tissues. This gene presents an alternative method of generating glutamate and 2-oxoglutarate by catalyzing the reversible production of 2-oxoglutarate and glycine from glutamate and
glyoxylate. Overexpression of this gene has been shown in past studies to produce a major increase in the free levels of the known long-distance nitrogen transporter serine as well as citrulline and glycine (Igarashi, Tsuchida, Mitsue, & Ohsumi, 2006). As this gene is showing across-the-board expression increases in the fall, its up-regulation could be due to a focus on generating free nitrogen compounds.

As previously stated, glutamate is required for the biosynthesis of glutamine via glutamine synthetase. In order to better understand any potential season-related regulation of glutamine, and its subsequent possible role as a long-distance transporter of nitrogen in the spring and fall during nutrient mobilization/remobilization, gene expression related to the production and utilization of glutamine was examined next (Figure 25). In this search, a glutamine synthetase gene was discovered (Sb01g010270) that has a very large increase in the late fall for the leaf and internode tissues, but low rhizome expression throughout the year with a small concurrent late fall rise. This gene could reflect an increase in the production of glutamine with a focus on transportation to the rhizome for over-winter storage. When translated M. × giganteus contigs that possess glutamine synthetase related annotations are compared to proteins translated from glutamine synthetase genes in S. bicolor and Z. mays, the translation of M. × giganteus contig “Mxg_TContig29199” matches most closely with the Sb01g010270.1 S. bicolor glutamine synthetase protein found to have seasonal changes in its expression (Figure 26).

Along with the Sb01g010270 gene for glutamine synthetase, the glutamine dumper gene Sb10g024750 was found to possess high levels of expression in the rhizome only during the early spring and high internode expression only during the late fall; at all other data points, the expression levels of this gene are near insignificant. The aptly named glutamine dumper proteins are cellular membrane proteins that induce the
export of glutamine into the apoplast (Pratelli, Voll, Horst, Frommer, & Pilot, 2009); therefore, this increase in Sb10g024750 corresponds to more glutamine being “dumped” into the apoplast for transport in the spring rhizome and fall internode, which directly corresponds to expected nutrient mobilization and remobilization patterns.

While the leaf and internode tissues showed a huge percentage of glutamine in comparison to the other four known long-distance amino acid carriers of nitrogen in the amino acid analysis (Figure 17), the rhizome tissue generally showed a much larger focus on asparagine instead of glutamine. Although no genes related to aspartate biosynthesis (such as aspartate aminotransferase) were found to have changes in their expression that aligned with seasonal nutrient mobilization/remobilization, two genes for asparagine synthesis were found that did have this pattern of expression (Figure 27). The first glutamine-dependent asparagine synthase gene, Sbo1g038460, has a moderate level of expression in the spring rhizome that slowly tapers off into insignificance by the fall. Leaf expression remains fairly low except for a single data point in mid-August, but the internode has a sharp increase in expression late in the season. This particular asparagine synthase gene could represent a focus in the fall internode to produce more asparagine for transportation, as noted by the larger percentage of asparagine compared to the other studied nitrogen-transporting amino acids in the 10/26/09 internode sample than previous internode samplings (Figure 17) as well as the overtaking of asparagine over aspartic acid late in the fall for the internode (Figure 16).

The other asparagine synthase gene, Sbo5g000440, also appears to have season-based expression, but in a manner far different than most other genes observed in this study (Figure 27). This particular asparagine synthase gene appears to have close to insignificant expression in all three tissues up until late August, at which point its expression rapidly rises in the rhizome tissue. The increase observed in late rhizome is
immense: the August 16th sample has a meager 19 FPKM expression, which rose to 336 FPKM on the October 8th sample and ended at 1,289 FPKM in the October 29th sample, representing an almost 70x increase in gene expression by the end of growing season. As the rhizome has higher overall asparaginase percentage compared to the other four nitrogen-transporting amino acids at most samplings, this extremely large increase could reflect a focus of the rhizome to store nitrogen internally as asparagine or as a asparagine-requiring protein, and/or could reflect a focus of producing asparagine to transport nitrogen to freshly produced rhizomes that had grown during the current growing season. As this gene shows such a remarkable increase and both season and tissue-preferred expression, it becomes a strong candidate for future studies that seek to further investigate the rhizome’s role in seasonal nitrogen mobilization, remobilization, and storage. Fortunately, a translated M. × giganteus contig annotated as an asparagine synthase (Mxg_TContig45475) clusters closely with Sb05g000440’s protein, making a focus on this particular M. × giganteus contig a good place to start such a future study (Figure 28).

The biosynthesis and use of glutamate, glutamine, aspartate, and asparagine for nitrogen transport is all interconnected, but, as previously mentioned, there is one other amino acid that has been observed being an important amino acid nitrogen-transporter: serine. Serine hydroxymethyltransferase is used to catalyze a reversible reaction that generates glycine and water from serine; in this dataset, two of these genes were observed that had apparent season-based changes in their expression (Figure 29). Both Sb03g041410 and Sb08g019520 have increases in leaf and internode expression in the late fall, while Sb03g041410 additionally has increased expression in the spring rhizome that decreases in summer and never recovers. Despite the lower overall fraction of serine in relation to the other four amino acid nitrogen-transporters previously observed, this
information suggests that serine could still play some role in the long-distance transportation of nitrogen during seasonal nutrient mobilization/remobilization, albeit at a lesser magnitude than glutamine and/or asparagine. The effects of these genes could be better analyzed in a future study of *M. × giganteus* through the examination of the “Mxg_TContig41021” and “Mxg_TContig44275” contigs or the “GRMZM2G452630” serine hydroxymethyltransferase 7 gene in *Z. mays* (Figure 30).

When gene expression profiles for nitrogen metabolism and transportation genes found to have season-based expression changes are directly compared to one another, the strongest focuses on seasonal expression alterations can be identified (Table 6). The largest “swings” in expression can be found in the rhizome’s focus on transporting nitrate in the spring and producing asparagine in the fall, whereas both the leaf and internode have their most notable changes in expression in the fall production of glutamine. This emphasis on the rhizome producing asparagine in the fall and the leaf and internode producing glutamine in the fall was previously observed in the amino acid analysis, where it was found that leaf and internode generally have a larger concentration of glutamine in comparison to asparagine and the rhizome tissue generally has a larger concentration of asparagine in comparison to glutamine (Figure 17). As such, it appears that glutamine is a very important amino acid transport vessel for the long-distance remobilization of nitrogen during the fall remobilization of nutrients from the aboveground tissues to the rhizome for storage, and that the rhizome has a large focus on producing asparagine during the late season for either storage, further mobilization to other rhizomes, and/or as a precursor to a larger molecule for nitrogen storage. This possible method of utilizing glutamine for long-distance transport while asparagine is also used for transport as well in addition to a role of storage follows established logic, as
glutamine is the more metabolically-active amino acid of the pair while asparagine is more inert and more suited to vacuole-bound storage (Kung & Wu, 1992).

During this process of studying seasonal nitrogen transport in *M. × giganteus*, many peculiar seasonal gene expression facets make themselves known. How meaningful these facets actually are sometimes difficult—if not impossible—to determine, yet some have the potential to be extremely interesting. An example of one such facet is the expression of two genes whose best-hit *O. sativa* annotations are related to aspartic proteinase nepenthesin (Figure 31). One of these genes, Sbo4g029400, shows a notable increase in expression in late season internode as well as a lesser increase in late season leaf tissue, while the other—Sbo2g035320—has fairly erratic expression through the growing season yet a markedly higher level of expression in its 5/5/11 spring rhizome sample. What makes these two genes interesting is the function of aspartic proteinase nepenthesin (a.k.a. nepenthesin). This enzyme has been best characterized and studied in two carnivorous plant genera: the *Nepenthes* genus (whose members are commonly referred to as “pitcher plants”), which is the namesake of this enzyme, and the *Drosera* genus (Athauda, et al., 2004; Takahashi, 2013). In these plants, nepenthesins are secreted externally (in the case of pitcher plants, they are secreted directly into the pitcher fluid) in order to cleave peptide bonds to free up amino acids for absorption by the plant—essentially, nepenthesins are used by carnivorous plants to digest their victims and absorb their nitrogen. As this study focused on nitrogen use and transport, it was interesting that these genes seemed to have season-based changes in their expression, if only because of the allure of their use in carnivorous plants. In the case of *M. × giganteus*, the expression patterns of these two genes appears to have Sbo2g035320 involved in spring nitrogen mobilization away from the rhizome and
Sb04g029400 involved in freeing nitrogen in the fall above-ground tissues for remobilization back to the rhizome.

Unfortunately, as these two genes are only putative pre-cursors to nepenthesin, and other types of aspartic proteinases are common in all plants, it is a challenge to determine their exact purpose. The most likely explanation, however, is that they are simply proteinases that cleave peptide bonds to free nitrogen for transportation and use. In fact, both of these genes best-hit Arabidopsis annotations are simply as “eukaryotic aspartyl protease family protein.” When the structure of the two S. bicolor proteins translated from these genes is compared to Nepenthes genus nepenthesin proteins and translated M. × giganteus contigs that are annotated as being related to aspartic proteinases, there are no close similarities between the Nepenthes nepenthesins and the S. bicolor proteins or the translated M. × giganteus contigs. Due to this, it seems unlikely that these genes are from an M. × giganteus aspartic proteinase that is more closely related to a nepenthesin than a standard aspartic proteinase. Although this search leads to a dead-end of sorts, it is helpful in illustrating the interesting wealth of information that exists within this dataset outside of what is being specifically sought, as well as the risk one runs of becoming too enamored with possibly inconsequential results that nonetheless require a great deal of time and focus to investigate.

4.4 Conclusion

The use of plant-derived biomass as a biofuel feedstock holds interest worldwide, and, as such, many studies in the United States and elsewhere have focused on examining the often-lauded biomass crop M. × giganteus as a potential biofuel feedstock (Heaton, Dohlman, & Long, 2008; Dohleman, 2009; Clifton-Brown, Neilson,
Lewandowski, & Jones, 2000; Lewandowski, Clifton-Brown, Scurlock, & Huisman, 2000; Price, Bullard, Lyons, Anthony, & Nixon, 2004). These studies have better defined the potential role that \( M. \times giganteus \) could play in the growing bioenergy industry through the documentation of the plant’s developmental and growth-related capabilities and characteristics. With the new seasonal dataset employed by this project, one of the key aspects of \( M. \times giganteus \) has been more acutely studied: its perennial growth cycle and, specifically, the movement and storage of nitrogen as it relates to \( M. \times giganteus \)’ perenniality. Perenniality is a great boon for any crop, as perennial crops are able to store nutrients over the winter for use during the next season’s growth, cutting down on yearly nutrient application requirements needed to yield maximum growth. As a biofuel feedstock should be as efficient in its growth as possible in order to generate a lot of biomass for as low of a cost as possible, the benefits of perennial crops not requiring replanting every year and storing nutrients at the end of the season to offset fertilizer application requirements often makes perennial crops very appealing for biofuel feedstock use.

In this project, the perenniality of \( M. \times giganteus \) was examined through the study of season-based mobilization and remobilization of nitrogen, one of the most important nutrients for plant growth. In order to better document the seasonal movement and accumulation of nitrogen and nitrogen-carrying amino acids, a diverse and replicated \( M. \times giganteus \) tissue collection sampled over a number of growing seasons was utilized. As little information about \( M. \times giganteus \) exists at the genomic and transcriptomic level—mostly due to the relatively newfound American interest in this particular plant in addition to the difficulty of performing genomic-type studies on a triploid organism with no reference of its own—generating information that investigates \( M. \times giganteus \)’ gene expression driven growth-related characteristics is a worthy yet
challenging endeavor. To this end, this project sought to examine important aspects related to perennial nutrient cycling at the level of gene expression through following the production and movements of nitrogen and its long-distance amino acid carriers in *M. × giganteus* in order to better understand its seasonal patterns of long-distance nitrogen mobilization and remobilization. In effect, this project has identified a number of genes related to a variety of different aspects of nitrogen-transporting amino acid production and mobilization that follow a few different season-preferred times of expression, most notably of which are those found to have high expression in the spring rhizome that declines as the season progresses while simultaneously having low springtime expression in both leaf and internode tissues that increases around the time of senescence in the fall. Genes that follow this pattern are very likely to be key components of seasonal nitrogen mobilization and remobilization, and the documentation of their presence and their seasonal expression profiles adds to the resources available to better understand this promising biofuel feedstock.
## CHAPTER FIVE

“TABLES AND FIGURES”

### 5.1 Tables

Table 1. Sample information for tissues sequenced in the ten-tissue RNA-Seq project.

<table>
<thead>
<tr>
<th>Read Type</th>
<th>Tissue</th>
<th>Collection Location</th>
<th>Collection Time</th>
<th>Total Reads</th>
<th>Total Bases Sequenced (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>36 bp Paired End</td>
<td>Mature Leaf</td>
<td>U of I Turf Farm</td>
<td>June, 2008</td>
<td>7,498,344</td>
<td>269,940,384</td>
</tr>
<tr>
<td>36 bp Paired End</td>
<td>Vegetative Shoot Apex</td>
<td>U of I Turf Farm</td>
<td>June, 2008</td>
<td>8,062,179</td>
<td>290,238,444</td>
</tr>
<tr>
<td>36 bp Paired End</td>
<td>Vegetative Sub-Apex</td>
<td>U of I Turf Farm</td>
<td>June, 2008</td>
<td>8,350,477</td>
<td>300,617,172</td>
</tr>
<tr>
<td>36 bp Paired End</td>
<td>Spring Rhizome</td>
<td>U of I Turf Farm</td>
<td>Mar/Apr, 2008</td>
<td>7,736,703</td>
<td>278,521,308</td>
</tr>
<tr>
<td>36 bp Paired End</td>
<td>Immature Inflorescence</td>
<td>U of I Turf Farm</td>
<td>Sept/Oct, 2008</td>
<td>5,679,747</td>
<td>204,470,892</td>
</tr>
<tr>
<td>36 bp Paired End</td>
<td>Emerging Shoot (1)</td>
<td>U of I Turf Farm</td>
<td>Mar/Apr, 2008</td>
<td>8,683,357</td>
<td>312,600,852</td>
</tr>
<tr>
<td>76 bp Paired End</td>
<td>Emerging Shoot (2)</td>
<td>U of I Turf Farm</td>
<td>Mar/Apr, 2008</td>
<td>12,386,166</td>
<td>941,348,616</td>
</tr>
<tr>
<td>76 bp Paired End</td>
<td>Pre-flowering Apex</td>
<td>U of I Turf Farm</td>
<td>Sept/Oct, 2008</td>
<td>12,238,323</td>
<td>930,112,548</td>
</tr>
<tr>
<td>76 bp Paired End</td>
<td>Rhizome Bud</td>
<td>U of I Turf Farm + Greenhouse</td>
<td>Mar/Apr, 2008; 1-2 Mo. Old</td>
<td>12,568,236</td>
<td>955,185,936</td>
</tr>
<tr>
<td>76 bp Paired End</td>
<td>Root</td>
<td>Greenhouse</td>
<td>1-2 Months Old</td>
<td>13,681,100</td>
<td>1,039,763,600</td>
</tr>
<tr>
<td>76 bp Paired End</td>
<td>Mature Inflorescence</td>
<td>U of I Turf Farm</td>
<td>Oct, 2008</td>
<td>13,676,869</td>
<td>1,039,442,044</td>
</tr>
</tbody>
</table>
Table 2. Identification of *S. bicolor* gene models used in RT-qPCR.

<table>
<thead>
<tr>
<th><em>S. bicolor</em> Gene Model</th>
<th>Annotation</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sb07g004190</td>
<td>&quot;Basic helix-loop-helix (bHLH) family protein&quot;</td>
<td>RP Rank #5 in Spring Rhizome</td>
</tr>
<tr>
<td>Sb01g005150</td>
<td>&quot;Similar to Indole-3-glycerol phosphate lyase&quot;</td>
<td>RP Rank #98 in Spring Rhizome</td>
</tr>
<tr>
<td>Sb04g025430</td>
<td>&quot;ATEP3; chitinase&quot;</td>
<td>RP Rank #248 in Spring Rhizome</td>
</tr>
<tr>
<td>Sb03g043280</td>
<td>&quot;Similar to Proline transport protein-like&quot;</td>
<td>RP Rank #371 in Spring Rhizome</td>
</tr>
<tr>
<td>Sb10g022200</td>
<td>&quot;Pathogenesis-related protein BET v I family&quot;</td>
<td>RP Rank #473 in Spring Rhizome</td>
</tr>
<tr>
<td>Sb10g028120</td>
<td>&quot;Oxygen-evolving enhancer protein 1, chloroplast precursor&quot;</td>
<td>RP Rank #12 in Mature Leaf</td>
</tr>
<tr>
<td>Sb09g028720</td>
<td>Similar to Chlorophyll A-B binding protein</td>
<td>RP Rank #13 in Mature Leaf</td>
</tr>
<tr>
<td>Sb09g019750</td>
<td>&quot;Bax inhibitor-1 family protein&quot;</td>
<td>Control 1</td>
</tr>
<tr>
<td>Sb02g041180</td>
<td>&quot;NADH-ubiquinone oxireductase 51 kDa subunit, mitochondrial precursor, putative, expressed&quot;</td>
<td>Control 2</td>
</tr>
</tbody>
</table>
Table 3. List of samples sequenced in seasonal RNA-Seq project.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Tissue</th>
<th>Collection Location</th>
<th>Collection Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1R_29Oct12</td>
<td>Rhizome</td>
<td>University of Illinois Turf Farm</td>
<td>10/29/2012</td>
</tr>
<tr>
<td>2L_8Oct12</td>
<td>Leaf</td>
<td>University of Illinois Turf Farm</td>
<td>10/8/2012</td>
</tr>
<tr>
<td>2N_8Oct12</td>
<td>Internode</td>
<td>University of Illinois Turf Farm</td>
<td>10/8/2012</td>
</tr>
<tr>
<td>2R_8Oct12</td>
<td>Rhizome</td>
<td>University of Illinois Turf Farm</td>
<td>10/8/2012</td>
</tr>
<tr>
<td>3L_Sep12</td>
<td>Leaf</td>
<td>University of Illinois Turf Farm</td>
<td>9/17/2012</td>
</tr>
<tr>
<td>3N_Sep12</td>
<td>Internode</td>
<td>University of Illinois Turf Farm</td>
<td>9/17/2012</td>
</tr>
<tr>
<td>3R_Sep12</td>
<td>Rhizome</td>
<td>University of Illinois Turf Farm</td>
<td>9/17/2012</td>
</tr>
<tr>
<td>4L_May11</td>
<td>Leaf</td>
<td>University of Illinois Turf Farm</td>
<td>5/5/2011</td>
</tr>
<tr>
<td>4R_May11</td>
<td>Rhizome</td>
<td>University of Illinois Turf Farm</td>
<td>5/5/2011</td>
</tr>
<tr>
<td>5L_Sep10</td>
<td>Leaf</td>
<td>Pana, IL</td>
<td>9/27/2010</td>
</tr>
<tr>
<td>5N_Sep10</td>
<td>Internode</td>
<td>Pana, IL</td>
<td>9/27/2010</td>
</tr>
<tr>
<td>5R_Sep10</td>
<td>Rhizome</td>
<td>Pana, IL</td>
<td>9/27/2010</td>
</tr>
<tr>
<td>6L_Aug10</td>
<td>Leaf</td>
<td>Pana, IL</td>
<td>8/17/2010</td>
</tr>
<tr>
<td>6N_Aug10</td>
<td>Internode</td>
<td>Pana, IL</td>
<td>8/17/2010</td>
</tr>
<tr>
<td>6R_Aug10</td>
<td>Rhizome</td>
<td>Pana, IL</td>
<td>8/17/2010</td>
</tr>
<tr>
<td>7L_Oct09</td>
<td>Leaf</td>
<td>Pana, IL</td>
<td>10/26/2009</td>
</tr>
<tr>
<td>7N_Oct09</td>
<td>Internode</td>
<td>Pana, IL</td>
<td>10/26/2009</td>
</tr>
<tr>
<td>7R_Oct09</td>
<td>Rhizome</td>
<td>Pana, IL</td>
<td>10/26/2009</td>
</tr>
<tr>
<td>8L_Aug09</td>
<td>Leaf</td>
<td>Pana, IL</td>
<td>8/18/2009</td>
</tr>
<tr>
<td>8N_Aug09</td>
<td>Internode</td>
<td>Pana, IL</td>
<td>8/18/2009</td>
</tr>
<tr>
<td>8R_Aug09</td>
<td>Rhizome</td>
<td>Pana, IL</td>
<td>8/18/2009</td>
</tr>
<tr>
<td>9L_Jun09</td>
<td>Leaf</td>
<td>Pana, IL</td>
<td>6/13/2009</td>
</tr>
<tr>
<td>9N_Jun09</td>
<td>Internode</td>
<td>Pana, IL</td>
<td>6/13/2009</td>
</tr>
<tr>
<td>9R_Jun09</td>
<td>Rhizome</td>
<td>Pana, IL</td>
<td>6/13/2009</td>
</tr>
</tbody>
</table>
Table 4. Annotated novel splice junction sites identified with a length greater than 1,000 bp, at least one sample with over 50 FPKM, and at least 25% coverage.

<table>
<thead>
<tr>
<th>CuffDiff ID</th>
<th>Locus</th>
<th>Coverage</th>
<th>Identity</th>
<th>e value</th>
<th>Subject ID</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_003507</td>
<td>chromosome_1:32911808-32913330</td>
<td>98.36%</td>
<td>98.26%</td>
<td>0</td>
<td>gi</td>
<td>512124087</td>
</tr>
<tr>
<td>XLOC_005288</td>
<td>chromosome_1:51006578-51007990</td>
<td>94.69%</td>
<td>93.12%</td>
<td>0</td>
<td>gi</td>
<td>118201104</td>
</tr>
<tr>
<td>XLOC_001337</td>
<td>chromosome_1:52349707-52352872</td>
<td>26.48%</td>
<td>90.93%</td>
<td>0</td>
<td>gi</td>
<td>226503758</td>
</tr>
<tr>
<td>XLOC_004809</td>
<td>chromosome_1:73835653-73838089</td>
<td>52.71%</td>
<td>98.60%</td>
<td>0</td>
<td>gi</td>
<td>209361311</td>
</tr>
<tr>
<td>XLOC_011861</td>
<td>chromosome_2:71449169-71451222</td>
<td>32.39%</td>
<td>85.26%</td>
<td>5E-148</td>
<td>gi</td>
<td>195607443</td>
</tr>
<tr>
<td>XLOC_013644</td>
<td>chromosome_3:28855847-28859169</td>
<td>63.49%</td>
<td>99.86%</td>
<td>0</td>
<td>gi</td>
<td>118201104</td>
</tr>
<tr>
<td>XLOC_017065</td>
<td>chromosome_3:28860673-28862289</td>
<td>70.54%</td>
<td>100%</td>
<td>0</td>
<td>gi</td>
<td>118201104</td>
</tr>
<tr>
<td>XLOC_013655</td>
<td>chromosome_3:38360056-38367693</td>
<td>32.94%</td>
<td>99.96%</td>
<td>0</td>
<td>gi</td>
<td>114309646</td>
</tr>
<tr>
<td>XLOC_025062</td>
<td>chromosome_6:33729505-33734963</td>
<td>26.88%</td>
<td>86.84%</td>
<td>0</td>
<td>gi</td>
<td>359359138</td>
</tr>
</tbody>
</table>
Table 5. Collection of selected enriched GO Terms for pairwise comparisons of spring versus fall tissue samples focused on jasmonic acid and the overall activity of the tissue.

<table>
<thead>
<tr>
<th>1 2 3 4 5 6 7 8 9</th>
<th>1 2 3 4 5 6 7 8 9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spring Sample</strong></td>
<td><strong>Fall Sample</strong></td>
</tr>
<tr>
<td><strong>Go Term</strong></td>
<td><strong>Value</strong></td>
</tr>
<tr>
<td>5/5/11 Rhizome</td>
<td>9/27/10 Rhizome</td>
</tr>
<tr>
<td>5/5/11 Leaf</td>
<td>9/27/10 Leaf</td>
</tr>
<tr>
<td>5/5/11 Rhizome</td>
<td>10/26/09 Rhizome</td>
</tr>
<tr>
<td>5/5/11 Leaf</td>
<td>10/26/09 Leaf</td>
</tr>
<tr>
<td>6/13/09 Rhizome</td>
<td>9/27/10 Rhizome</td>
</tr>
<tr>
<td>6/13/09 Leaf</td>
<td>9/27/10 Leaf</td>
</tr>
<tr>
<td>6/13/09 Internode</td>
<td>9/27/10 Internode</td>
</tr>
<tr>
<td>6/13/09 Rhizome</td>
<td>10/26/09 Rhizome</td>
</tr>
<tr>
<td>6/13/09 Leaf</td>
<td>10/26/09 Leaf</td>
</tr>
<tr>
<td>6/13/09 Internode</td>
<td>10/26/09 Internode</td>
</tr>
</tbody>
</table>

- Significance Level of Up-Regulation. Z-score of result is found in each cell.
Table 6. FPKM profile comparison for nitrogen-related genes found to possess some degree of season-related differential gene expression.

<table>
<thead>
<tr>
<th>Rhizome</th>
<th>Leaf</th>
<th>Internode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sbo1g2034170: Nitrate Transporter 1.1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Sbo1g0055688: Nitrate Transporter 2.2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Sbo1g0233015: Nitrate Transporter 1.2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Sbo1g0244600: Nitrate Reductase 1a</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Sbo1g0344470: Nitrate Reductase 1b</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Sbo1g0237750: Nitrate Reductase 2a</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Sbo1g0341100: Nitrate Reductase 1c</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Sbo1g0417160: Glutamine Synthetase</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Sbo1g0311570: Glutamine Synthetase Close 1111</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Sbo1g0247500: Glutamine Dumper 1</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Sbo1g0004480: Glutamine-dependent Asparagine Synthetase 1</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Sbo1g0305900: Glutamine-dependent Asparagine Synthetase 2</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Sbo1g0411410: Gluconolactone Oxidase</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Sbo1g0106590: 6-Phosphogluconate Dehydrogenase</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>FPKM Scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhizome</td>
<td>Leaf</td>
<td>Internode</td>
</tr>
<tr>
<td>11</td>
<td>33</td>
<td>100</td>
</tr>
<tr>
<td>300</td>
<td>600</td>
<td>900</td>
</tr>
</tbody>
</table>

(Rhizome, Leaf, Internode)
5.2 Figures

Figure 1. Similarity of *M. × giganteus* contiguous sequences to gene models or ESTs of other grasses.
Figure 2. Reads mapping to *S. bicolor* gene models in the ten-tissue RNA-Seq project.
Figure 3. RPKM values obtained from sequencing versus relative expression verification values obtained via RT-qPCR for the ten-tissue RNA-Seq project.
Figure 4. Number of reads mapped to reference models by Cuffdiff in the seasonal RNA-Seq project for each sample’s three biological replicates.
Figure 5. Comparison of raw sequencing results for each of the three sequencing runs.
Figure 6. Alignment information difference when using “only allow unique alignments” versus “allow 10 max multihits” for biological replicate C of rhizome tissue sampled on 9/27/2010.
Figure 7. Percent of mapped reads to reference models in each sample’s replicates in the seasonal dataset.
Figure 8. Number of significant gene models (alpha = 0.01) determined from all pairwise between the two tissue samplings with the highest distance between one another (10/26/09 Leaf and 8/18/09 Rhizome) for both classic and geometric FPKM normalization options.

Classic FPKM  Geometric FPKM

763  24,804  402

33,815 Total Possible Models
Figure 9. Comparison of novel splice junction sites identified by TopHat2 and used in alignment.
Figure 10. Dendrogram illustrating relatedness of samples in seasonal dataset based on similarity of expression profiles via Jensen-Shannon Distance.
Figure 11. Distance matrix for seasonal dataset showing pairwise similarities between samples based on Jensen-Shannon Distance.
Figure 12. Subset of enriched GO terms in the comparison of 5/5/11 Rhizome versus 10/29/12 Rhizome via PAGE.
Figure 13. Seasonal FPKM profiles for three jasmonate ZIM-domain proteins.
Figure 13 (cont.)

![Graph showing FPKM for Rhizome, Leaf, and Internode](image)

<table>
<thead>
<tr>
<th></th>
<th>Rhizome</th>
<th>Leaf</th>
<th>Internode</th>
<th>FPKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>19</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>85</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>133</td>
<td>38</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>57</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>5/5/11</th>
<th>6/13/11</th>
<th>7/15/11</th>
<th>8/17/11</th>
<th>9/27/11</th>
<th>10/12/12</th>
<th>10/26/12</th>
<th>11/26/12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>26</td>
<td>120</td>
<td>501</td>
<td>2000</td>
<td>10000</td>
<td>50000</td>
<td>100000</td>
<td>200000</td>
</tr>
</tbody>
</table>

Scale: 2 - 2000
Figure 14. Seasonal FPKM profiles for two OPR genes, which generate the jasmonic acid precursor OPC-8:0.
Figure 15. Nitrogen content of seasonal *M. × giganteus* dataset’s tissues as determined by combustion analysis.

- Leaf: $y = 0.009x^4 - 0.3449x^3 + 4.8195x^2 - 29.14x + 65.773$
  $R^2 = 0.9502$
- Internode: $y = 0.0314x^3 - 0.731x^2 + 5.5278x - 13.213$
  $R^2 = 0.9389$
- Rhizome: $y = -0.0355x + 0.6173$
  $R^2 = 0.4333$

Legend:
- Leaf
- Internode
- Rhizome
- Leaf Trendline
- Internode Trendline
- Rhizome Trendline
Figure 16. Content comparison of dual-nitrogen carrying asparagine and glutamine to their single-nitrogen containing amino acid precursors (i.e., asparagine-to-aspartic acid and glutamine-to-glutamic acid) in the seasonal M. × giganteus dataset.
Figure 17. Percent distributions among five amino acids known to be long-distance transporters of nitrogen in seasonal dataset.
Figure 18. Nitrate transporter genes with apparent season-based expression changes.
Figure 18 (cont.)

![Graph showing nitrate transporter expression in rhizome, leaf, and internode regions over months 5 to 11.](image-url)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhizome</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Leaf</td>
<td>23</td>
<td>17</td>
<td>24</td>
<td>19</td>
<td>21</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td>Internode</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>13</td>
<td>37</td>
</tr>
</tbody>
</table>

FPKM Scale: 3 10 20

Month
Figure 19. Nitrate reductase and nitrite reductase genes with apparent season-based changes in expression.
Figure 19 (cont.)
Figure 20. Guide tree for nitrate reductase using translated *M. × giganteus* contigs and protein sequences from *S. bicolor* and *Z. mays* clustered by relatedness.
Figure 21. Overview of nitrogen movement through glutamate and glutamine biosynthesis.
Figure 22. Select glutamate-related genes that have apparent season-based changes in gene expression.
Figure 22 (cont.)
Figure 22 (cont.)

![Graph showing expression levels of Glutamate:Glyoxylate Aminotransferase in different parts of the plant: Rhizome, Leaf, and Internode.](image)

<table>
<thead>
<tr>
<th></th>
<th>5/5/11</th>
<th>6/13/09</th>
<th>8/16/09</th>
<th>9/17/10</th>
<th>9/17/12</th>
<th>9/27/12</th>
<th>10/8/12</th>
<th>10/26/12</th>
<th>10/29/12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rhizome</strong></td>
<td>109</td>
<td>120</td>
<td>119</td>
<td>152</td>
<td>154</td>
<td>121</td>
<td>154</td>
<td>237</td>
<td>357</td>
</tr>
<tr>
<td><strong>Leaf</strong></td>
<td>256</td>
<td>242</td>
<td>168</td>
<td>212</td>
<td>293</td>
<td>277</td>
<td>222</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td><strong>Internode</strong></td>
<td>76</td>
<td>87</td>
<td>125</td>
<td>93</td>
<td>123</td>
<td>105</td>
<td>522</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FPKM Scale: 75-150-350
Figure 23. Guide tree for glutamate synthase using translated *M. × giganteus* contigs and protein sequences from *S. bicolor* and *Z. mays* clustered by relatedness.
Figure 24. Guide tree for glutamate dehydrogenase using translated *M. × giganteus* contigs and protein sequences from *S. bicolor* and *Z. mays* clustered by relatedness.
Figure 25. Glutamine-related genes that show changes in expression characteristic of time-of-season-based nutrient mobilization and remobilization.
Figure 26. Guide tree for glutamine synthetase using translated \( M. \times giganteus \) contigs and protein sequences from \( S. bicolor \) and \( Z. mays \) clustered by relatedness.
Figure 27. Asparagine synthase genes found to have changes in expression based on time-of-season.
Figure 28. Guide tree for asparagine synthase using translated *M. × giganteus* contigs and protein sequences from *S. bicolor* and *Z. mays* clustered by relatedness.
Figure 29. Genes involved in serine biosynthesis with apparent season-based changes in expression.
Figure 30. Guide tree for serine hydroxymethyltransferase using *M. × giganteus* contigs and gene models from *S. bicolor* and *Z. mays* clustered by relatedness.
Figure 31. Seasonal expression profiles of two potential aspartic proteinase nepenthesin related gene.
REFERENCES


http://cufflinks.cbcb.umd.edu/manual.html#library_norm_meth

CummeRbund. (2013). *CummeRbund.* Retrieved from

http://compbio.mit.edu/cummeRbund/


http://bib.oxfordjournals.org/content/early/2012/09/15/bib.bbs046.full


US EPA. (2013, August 1). *Renewable Fuel Standard (RFS).* Retrieved from United States Environmental Protection Agency:
http://www.epa.gov/otaq/fuels/renewablefuels/index.htm

http://www.gpo.gov/fdsys/pkg/BILLS-111hr1enr/pdf/BILLS-111hr1enr.pdf


http://www.eia.gov/countries/country-data.cfm?fips=US

Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *The Plant Cell Online, 14*(7), 1441-1456.


