The dissertation of Joshua R. Herr was reviewed and approved* by the following:

John E. Carlson  
Professor of Molecular Genetics  
Dissertation Advisor  
Chair of Committee

Claude dePamphilis  
Professor of Biology

Seogchan Kang  
Professor of Plant Pathology

Istvan Albert  
Associate Professor of Bioinformatics

Teh-hui Kao  
Professor of Biochemistry and Molecular Biology  
Chair, Intercollege Graduate Degree Program in Plant Biology

*Signatures are on file in the Graduate School
ABSTRACT

Macro- and microbial organisms naturally inundate plants and these organisms may be beneficial or pathogenic. Forest trees are particularly interesting for the wide array of microorganisms they come into contact with along with the complexity of stresses that are subjected upon them year after year. Additionally, not only being environmentally important, trees are the focus of a shifting economy based on cellulosic bioenergy. To meet global expectations and current benchmarks, we, as a society, will have to figure out how to grow trees faster, with more yield, with less inputs, and with a greater incidence of climate induced stress, which comes in many forms. The advancement of a new “green” revolution will have to take into account microorganisms that interact with these trees, but our understanding of these interactions is lacking in many ways. This dissertation is a small contribution to fill a large gap in our knowledge of forest tree stress responses, microbial diversity associated with native forests and also *Populus* hybrids planted as bioenergy crops. Additionally, there are chapters included here on the identification of fungal species from nucleotide data and methods to improve short rotation woody crops (SRWCs) for increased yield and sustainability.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................... v

ACKNOWLEDGEMENTS .................................................................................................. vi

Chapter 1 Phylogenomics of the CYP74 cytochrome P450 gene family in *Populus* and other plants .................................................................................................................. 1

1.1 Abstract .................................................................................................................... 1
1.2 Introduction .............................................................................................................. 3
1.3 Materials & Methods ............................................................................................ 6
  1.3.1 CYP74 member library development and phylogeny construction .......... 6
  1.3.2 Promoter Motif Characterization for CYP74 genes .............................. 7
  1.3.3 Induced Expression of CYP74 Genes ...................................................... 7
1.4 Results & Discussion ............................................................................................ 9
  1.4.1 CYP74 Gene Family Phylogeny ................................................................. 9
  1.4.2 The Evolution of the Octadecanoid Pathway in Plants ......................... 11
  1.4.3 CYP74 Conserved Motif Regions ............................................................. 13
  1.4.4 Induced Gene Expression ......................................................................... 13
1.5 Conclusions .......................................................................................................... 14
1.5 Acknowledgements ............................................................................................... 15
1.7 References ............................................................................................................. 15

Chapter 2 The challenge of identifying operational taxonomic units from environmentally-derived fungal amplicon data ........................................................................... 32

2.1 Abstract ................................................................................................................. 32
2.2 Introduction ............................................................................................................. 33
2.3 OTU based analysis of fungi from environmental samples ........................... 35
2.4 Computational methods for the identification of OTUs ................................. 39
  2.4.1 Pre-clustering ............................................................................................. 40
  2.4.2 Quality Assessment and Denoising ........................................................... 41
  2.4.3 Chimera Detection ..................................................................................... 41
2.5 Local Sequence Similarity .................................................................................... 43
2.6 Database Based Clustering Methods .................................................................. 44
  2.6.1 Sequence Similarity Search Methods ....................................................... 44
  2.6.2 Database methods for sequence comparison ........................................... 45
  2.6.3 Phylogenetic Methods for Clustering ....................................................... 45
2.7 Evaluation of Clustering Methods ...................................................................... 47
2.8 Association networks and Environmental Factors ......................................... 47
  2.8.1 Diversity Measurements .......................................................................... 47
  2.8.2 Network Analysis ...................................................................................... 48
2.9 Conclusions .......................................................................................................... 49
Chapter 5 Comparative Genomics of Forest Trees: A Fast-Track For Understanding Diverse Stress Responses ................................................................. 126

5.1 Introduction.................................................................................................................. 126
5.2 Integrating Comparative Genomics and Systems Biology ........................................ 128
5.3 Hormone Response ...................................................................................................... 130
5.4 Crosstalk Between Responses .................................................................................... 131
5.5 Understanding Stress Response Through Co-Expression ......................................... 133
5.6 Conclusions.................................................................................................................. 134
5.7 Acknowledgements..................................................................................................... 136
5.8 References................................................................................................................... 136

Chapter 6 Bioenergy From Trees ...................................................................................... 146

6.1 Introduction.................................................................................................................. 146
6.2 Using Genomics to optimize tree biomass ................................................................ 148
6.3 Understanding cell walls to take them apart ............................................................. 149
6.4 From field to fuel – saccharification and life-cycle analysis ........................................ 151
6.5 How do get there from here? ...................................................................................... 153
6.5 Acknowledgements..................................................................................................... 155
6.5 References................................................................................................................... 156

Chapter 7 Summary and Broader Impacts ..................................................................... 161

7.1 Introduction.................................................................................................................. 161
7.2 Characterization of CYP74 Proteins ......................................................................... 163
7.3 Soil Microbial Diversity & Sequence Identification .................................................... 165
7.4 Short-Rotation Woody Crop Biomass Breeding & Plantation Experiments ................ 166
7.5 Future Directions ........................................................................................................ 167

Appendix The Correlation With Leaf Area, Leaf Damage, and an Ethylene-Blocking Agent on the Growth of a hybrid Populus Grown for Biomass........... 172

A.1 Abstract....................................................................................................................... 172
A.2 Introduction.................................................................................................................. 173
A.3 Materials and Methods.............................................................................................. 175
  A.3.1 Experimental Plot Design .................................................................................... 175
  A.3.2 Treatment with 1-MCP ........................................................................................ 176
  A.3.3 Field Data Collection and Analysis .................................................................... 176
  A.3.4 The Characterization of Biomass Composition and Conversion to Ethanol .......... 177
A.4 Results & Discussion................................................................................................. 178
A.5 References................................................................................................................... 180
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>CYP74 Enzymatic pathway</td>
<td>21</td>
</tr>
<tr>
<td>1.2</td>
<td>Model of Jasmonic Acid pathway in plants</td>
<td>22</td>
</tr>
<tr>
<td>1.3</td>
<td>Gene Expression Data for AOS in Poplar tissues</td>
<td>23</td>
</tr>
<tr>
<td>1.4</td>
<td>Phylogeny of CYP74 gene family (Part One)</td>
<td>24</td>
</tr>
<tr>
<td>1.5</td>
<td>Phylogeny of CYP74 gene family (Part Two)</td>
<td>25</td>
</tr>
<tr>
<td>1.6</td>
<td>Amino Acid Alignment for CYP74 genes in 17 plants with sequenced genomes</td>
<td>26</td>
</tr>
<tr>
<td>1.7</td>
<td>Predicted promoter motifs for AOS in Populus</td>
<td>27</td>
</tr>
<tr>
<td>1.8</td>
<td>Predicted promoter motifs for HPL in Populus</td>
<td>28</td>
</tr>
<tr>
<td>1.9</td>
<td>GUS reporter assay for AOS in Arabidopsis</td>
<td>29</td>
</tr>
<tr>
<td>3.1</td>
<td>Map showing South Korean soil sampling locations</td>
<td>79</td>
</tr>
<tr>
<td>3.2</td>
<td>Rarefaction Curves for soil sample reads</td>
<td>80</td>
</tr>
<tr>
<td>3.3</td>
<td>Phylogenetic Tree of Sequence reads</td>
<td>81</td>
</tr>
<tr>
<td>3.4</td>
<td>Sequencing Read Cluster Diagram</td>
<td>82</td>
</tr>
<tr>
<td>3.5</td>
<td>Table of Micronutrients from Korean Forest Soils</td>
<td>83</td>
</tr>
<tr>
<td>3.6</td>
<td>Principle Component Analysis of Forest Soil Metagenomes</td>
<td>85</td>
</tr>
<tr>
<td>3.7</td>
<td>Phylogenetic Tree of Fungal Hits to Sequence Databases</td>
<td>86</td>
</tr>
<tr>
<td>3.8</td>
<td>Pearson Heat-Map of Microbial Co-Occurence</td>
<td>87</td>
</tr>
<tr>
<td>3.9</td>
<td>Spearman Heat-Map of Microbial Co-Occurence</td>
<td>88</td>
</tr>
<tr>
<td>3.10</td>
<td>Kendall Heat-Map of Microbial Co-Occurence</td>
<td>89</td>
</tr>
<tr>
<td>3.11</td>
<td>Overall Microbial Family Hits For Each Forest Soil</td>
<td>90</td>
</tr>
</tbody>
</table>
Figure 3.12: Relative Presence of Microbial Families With and Without Normalization ................................................................. 91

Figure 3.13: Cluster Analysis of Sequence Reads Using GLIMMER .................. 92

Figure 4.1: Graph Showing Growth in SRWC Publications ............................. 124

Figure 4.2: Graphical Representation of Genomics in SRWC Breeding ........... 125

Figure 5.1: A Systems Approach for Studying Forest Tree Environmental Responses .................................................................................. 145

Figure 7.1: Synopsis of Possible Role of Soil Fungi in Mediating a Induced Systemic Response in Plants ................................................................. 170

Figure 7.2: Poplar Root With Multiple Mycorrhizal Fungi ............................... 171

Figure A.1: Poplar Biomass – Total Leaf Area X 1st Year Height ......................... 182

Figure A.2: Poplar Biomass – 1st Year Height X 2nd Year Height ...................... 183

Figure A.3: Poplar Biomass – 1st Year Height X Leaf Area Removed ................. 184

Figure A.4: Poplar Biomass – 2nd Year Height X Leaf Area Removed .............. 185
ACKNOWLEDGEMENTS

There’s an African proverb that states “it takes a village to raise a child” and I would argue that it also “takes a village to raise a PhD student”. I’ve had innumerous help from many people; people who were there for me and also people who let me grow on my own by not being there for me. It was all a help to me during this strenuous and chaotic growing experience.

I’d first like to thank the members of my dissertation committee for taking time to both serve on that committee and give me constructive criticism along the way. Dr. Claude dePamphilis, as well as his laboratory members, was instrumental in trouble-shooting different stages of my many projects and allowed me to access lab equipment to check the quality of my pre-sequencing nucleotides. Dr. Seogchan Kang was there for me through interesting talks about plant-fungal interactions and also through support via the Penn State Microbial Genomics Training Fellowship that supported me for two years and allowed me to initiate so much rewarding collaborative science. Lastly, Dr. Istvan Albert was very helpful in pointing me in the right direction when it came to data analysis, and helped to introduce me to new levels of computer proficiency. He always welcomed me to use the computers in the Bioinformatics Consulting and Core Facility.

I’d like to thank Dr. Teh-hui Kao for his role as head of the Plant Biology Graduate Degree Program. Dr. Kao always was there for me with his support and was always there for me when I needed someone to talk with. He was supportive – but never too soft – and
was instrumental in making sure that I have upheld the high standards of the program. I could have gone to a number of highly ranked schools in plant biology, but Dr. Kao reminded me that Penn State’s Plant Biology program was the number one ranked program in the country, not exclusive to his hard work to make it so. My most fond memory of Dr. Kao comes from the first time we spoke: When I asked him why I should attend Penn State’s Plant Biology program over any other program, he paused for a moment and said “Well, I throw a pretty great party in my backyard every year for the program…”

I’ve been generously funded by many sources, most notably through an Excellence in Graduate Recruitment Fellowship via the Penn State Graduate School, support through the Dorothy & Lloyd Huck Biosciences Fellowship, teaching support from the Department of Biology, a fellowship from the Penn State Functional Microbial Genomics Training Grant, a PUF Fellowship jointly supported through the US French Embassy and a grant from the office of collaborative science from the European Union, and lastly through support to John Carlson on the USDA SunGrant for development of plant based bioenergy.

There’s been many friends who I have shared my time with at Penn State (most of all who have gone before me) and I thank them for knowing when to distract me and knowing when to leave me alone.
Everyone in the Carlson Lab has been incredibly supportive and has made coming into work every day very pleasurable. They all have been great role models to me in their work ethic. Most of all, Tyler Wagner, our former lab technician, was instrumental in helping me maintain my sanity and has provided a ton of support during this research.

I would like to greatly thank my dissertation advisor, John E. Carlson, for all his support during my time in his laboratory. I have greatly appreciated his guidance and support. He’s gone above and beyond the call of duty in support of my growth. John always added insight into whatever I was working on, in a magical, obvious (but never condescending), and subtle way that always made me feel at ease. He has given me a lot of freedom to explore and make my own mistakes along the way, which was a really very insightful stance to take during my growth process. I appreciate his support, mentorship, and friendship.

Quite most of all, my family both immediate and extended have made this whole process happen. The Baker Family including most importantly Allie Baker, as well as my mother, father, as well as my sisters, Julie McCauley and Jennifer Maloney and their respective families, have been extremely supportive in this whole process.

I’d like to dedicate this dissertation to my uncle Michael W. Farmer, from Williamsport PA, who recently passed away in October of 2012. Uncle Mike was instrumental in
getting me to “come back” to Pennsylvania and he was always an amazing role model in his work ethic and drive. He will be remembered and always missed.
Chapter 1

Phylogenomics of the CYP74 cytochrome P450 gene family in Populus and other plants

Joshua R Herr\textsuperscript{1}, Haiying Liang\textsuperscript{2}, John E. Carlson\textsuperscript{1,3}

\textsuperscript{1}The Schatz Center for Tree Genetics, Department Of Ecosystem Science & Management, Pennsylvania State University, University Park, PA 16801 USA
\textsuperscript{2}Department of Genetics & Biochemistry, Clemson University, Clemson, SC 29634 USA
\textsuperscript{3}Department of Bioenergy Science & Technology, Chonnam National University, Gwangju, 500-757 South Korea

Note: Haiying Liang conducted the transformation of Arabidopsis and the GUS reporter analysis.

Chapter formatted for planned submission to BMC Plant Biology

1.1 Abstract

Jasmonic acid (JA) is a plant-signaling hormone critical for regulating plant carbon allocation, growth, and response to biotic and abiotic stresses. JA is synthesized from membrane-derived fatty acids via the octadecanoid (or lipoxygenase) pathway. While many of the key genes involved in the JA pathway have been identified, their evolutionary phylogeny has not been well explored due to the lack of reference genomes.
In particular, the first dedicated step in JA synthesis is presumably regulated by allene oxide synthase (AOS), a member of the large CYP74 cytochrome P450 family, which converts lipoxygenase-derived fatty acid hydroperoxide to allene epoxide. We have characterized the CYP74 cytochrome P450 gene family in *Populus* as a model species. The gene expression of AOS, as expected, is highly upregulated following herbivore damage, and also varies among different plant tissues. More importantly, we present a phylogeny of the CYP74 cytochrome P450 gene family across the plant kingdom. The phylogeny shows three distinct clades in the CYP74 cytochrome P450 gene family corresponding to genes for divinyl ester synthase (DES), fatty acid hydroperoxide (HPL), and a large AOS clade with Monocot and Dicot specific sub-clades. Remarkably, despite the critical role of JA as a phytohormone, we found just one AOS gene in the *Populus* reference genome.

*Keywords:* Jasmonic Acid, Oxylipins, Octadecanoic Pathway, Plant Defense, Stress Response, *Populus*

reductase, PUFAs: poly-unsaturated fatty acids, SAMT: salicylic acid methyl transferase, TE: thioesterase,

1.2 Introduction

The plant hormone jasmonic acid (JA) regulates plant growth and response to wounding and stress. Although JA possesses a role in the induction of flowering, senescence, leaf abscission, and tuber formation, it is typically associated with responses to abiotic stresses, such as drought or air pollution, or biotic stresses, such as herbivore attack (Wasternack 2007, Hughes et al 2009). Jasmonic acid is the product of the octadecanoid arm of the oxylipin pathway, which also produces other notable octadecanoid compounds such as prostaglandins (Gretchkin 2002). Plant oxylipins are typically associated with signal transduction and defense reactions and are analogous to animal prostaglandins in structure and function (Gretchkin 2002). Oxylipins are derived from the oxygenation of polyunsaturated fatty acids (PUFAs) (Blee 2002, Pohnert 2005) and are enzymatically modified or spontaneously catalyzed to form a diverse array of biochemicals that are involved in a suite of functions within the plant. Green leaf volatiles (GLVs), are volatile oxylipins that are produced in response to biotic and abiotic stresses, such as herbivore attack, ozone exposure, or pathogen infection (Matsui 2006).

The CYP74 cytochrome P450 enzymes function as the first dedicated step of the oxylipin pathway that modifies lipoxygenase-derived hydroperoxides to oxylipins (Figure 1.1). CYP74 enzymes are unique in that they do not require oxygen or NADH-dependent reductase to catalyze hydroperoxides to JA precursors, which is atypical for cytochrome
P450 enzymes (Stumpe & Feussner 2006, Wasternack 2007). It has been hypothesized that three different enzymatic activities are regulated by CYP74 family proteins (Stumpe & Feussner 2006, Hughes et al 2009). Allene oxide synthase (AOS) forms allene oxide, which eventually become converted via the octadecanoid pathway to JA and other oxylipins (Brash 2009). Hydroperoxide lyase (HPL) ultimately produces various aldehydes, esters, and alcohols termed “green leaf volatiles” (GLVs) due to their odor likeness to freshly mown green grass. GLVs can act as volatile or nonvolatile compounds involved in plant defense and growth regulation (Froehlich et al 2001, Gretchkin 2002, Stumpe & Feussner 2006). Divinyl ester synthase (DES) forms divinyl esters and, while their exact role is not entirely clear, appear to be involved in plant defense (Stumpe & Feussner 2006, Brash 2009). Products of DES undergo variable acidic hydrolysis easily, which contributes to their chemical diversity. Volatile scents, such as etherolic and etherolenic acids, are products of DES and constitute portions of the smell of garlic and onion (Gretchkin 2002, Stumpe & Feussner 2006, Brash 2009).

Literature on CYP74 gene family isoforms has yielded conflicting reports on the location of enzyme activity and the specificity for 9- or 13-hydroperoxy-linolenic acid, the product of lipoxygenase (LOX) on both Linoleic and Linolenic acid (Figure 1; Stumpe & Feussner 2006, Hughes et al 2009).

Important questions exist regarding the evolutionary history and expression patterns of the CYP74 genes as well as the resulting activity of their enzymes. Previous assessments of the CYP74 gene family have hypothesized that either DES (Gretchkin 2002) or AOS (Nelson 2009, Brash 2009) are the most basal member based on similarity of sequence
motifs between Arabidopsis CYP74s and those of Physcomitrella. However, these hypotheses remain controversial because until recently there was not sufficient sequence data to construct a robust, rooted phylogeny. A chemically active AOS has been identified in a different form in animals (both in the corals and Cephalochordates) and not found in any bacteria, algae or fungi (Lee et al 2008). As a result, a robust phylogenetic study within the plants in needed to understand the evolution and diversification of CYP74-mediated lipid biochemistry during plant wounding. The CYP74 gene family appears to be actively expressed in different sub-cellular locations and in different tissues, but there is no comprehensive understanding of spatial and temporal patterns of CYP74 expression, which hinders our full understanding of enzyme substrate specificity. For example, CYP74 enzymes can theoretically use either 9- or 13-hydroperoxy-linolenic acid in different sub-cellular locations (chloroplast vs. peroxisome vs. cytoplasm) (Stumpe & Feussner 2006, Hughes et al 2009). Part of the confusion may be alleviated by understanding general expression patterns of the CYP74 genes in different tissues or under different conditions.

Although members of the CYP74 gene family have been identified in numerous plant species (Song et al 1993; Laudert & Weiler 1998; Park et al 2000; Sivasankar et al 2000; Itoh et al 2002), we address the relatedness of these genes by characterizing them here in Poplar (Populus deltoides X Populus trichocarpa hybrid) and Arabidopsis thaliana. The JA pathway is vital for the understanding of stress response in plants, so understanding the regulatory pathways responsible for its synthesis is of critical importance. Little is known about CYP74 biology, their phylogenetic diversity, reaction specificity, or their
spatial distribution in plant cells, and surprisingly, there have been no phylogenetic studies of the CYP74 gene family to date. We chose *Arabidopsis* for studying the CYP74 gene family because it has the best-characterized plant genome and its ease of study in the laboratory, but its streamlined genome may not provide a characterization that would be comparable to the entire lineage of plants. We also chose *Populus* for characterization of the CYP74 gene family to represent an economically important species that is a model for woody plants (Tuskan et al 2006) and that has last undergone a genome duplication event at approximately more recent than 80 million years ago (Jiao et al 2011) and estimated presently at 60 to 65 million years (Tuskan et al 2012). *Arabidopsis*, on the other hand, has undergone a duplication event at approximately 60 million years ago and has experienced subsequent genome streamlining (Jiao et al 2012). As a genus, *Populus* is distributed widely across the northern hemisphere and subsists in a diverse array of environmental conditions, and a thorough characterization of the JA stress response pathway in Poplar should be of great interest. Phylogenetic studies within the plants will contribute to our understanding of the important CYP74-mediated lipid biochemistry (Lee et al 2008).

1.3 Materials & Methods

1.3.1 CYP74 member library development and phylogeny construction

We searched public databases such as NCBI, EMBL, and Phytozome (Goodstein et al 2012) using BLASTX to create a database of CYP74 family genes homologous with
those previously described in *Arabidopsis* (Delker et al 2006). Both genomic and EST data were manually inspected for full-length transcripts and were aligned using the multiple sequence alignment algorithm MUSCLE (Edgar 2004) for all-against-all sequence alignment. When appropriate genomic data was present for a respective species this was selected preferentially over EST data. Duplicate sequences were culled from the data set, excluding some sequences used for statistical tests. A final alignment of 223 sequences was edited manually. A phylogenetic tree was constructed using the computer programs RAxML (Stamatakis et al 2006) for maximum likelihood methods and BEAST (Drummond et al 2012) for Bayesian methods. Maximum likelihood bootstrap values were estimated from 1000 rapid bootstrap replicates and Bayesian posterior probabilities were obtained from the MCMC algorithm in BEAST. Sequences obtained from Physcomitrella were used as outgroups for the analysis since they were estimated to be earliest plant with CYP74 genes expressing homology to other plants.

### 1.3.2 Promoter Motif Characterization for CYP74 genes

Plant promoters for genomic sequences of *Populus trichocarpa* were predicted using the NSITE-PL algorithm and validated in the conjoined PlantProm DB (Abdulazimova et al 2010) along 4Kb prior to the transcription start site to compare how environmental factors or other plant hormones would affect the expression of CYP74 genes.

### 1.3.3 Constitutive and Induced Expression of CYP74 Genes
We measured spatial and temporal variability in CYP74 gene expression in *Populus deltoides* X *Populus nigra* OGY hybrid following herbivore damage. Poplar trees were grown in a greenhouse maintained at 20°C and approximately 50% relative humidity with no supplemental lighting. Clonal *Populus* cuttings were taken after parent material senesced in November of the previous year and cuttings were acclimated at 4°C until March and allowed to grow until July of that year in the greenhouse. These plants were subjected to herbivore feeding treatment using gypsy moth (*Lymantria dispar*) larvae for 48 hours (n=5 per treatment, 16 reps per treatment per variable) (after Osier & Lindroth 2001). Non-sink (5th node and greater) leaves, phloem and xylem were collected along a time course following herbivory (1, 2, 4, 8, 12, 24, 36, and 48 hours), photographed, and flash frozen in liquid nitrogen and immediately placed in freezer and stored at -80°C until processing. RNA was extracted from the plant tissue using a modified CTAB method and cDNA was made using the Quantitect Reverse Transcription Kit for cDNA synthesis (Qiagen, Gaithersburg, Maryland USA). Using qRT-PCR, expression of CYP74 genes was measured relative to a suite of co-expressed house-keeping genes including EF1β (elongation factor 1-beta), UBQ (poly-ubiquitin), TUB (beta-9 tubulin chain), ACT11 (actin 11), and 18S (18S ribosomal RNA) (Brunner et al 2004).

We cloned the promoter sequence region approximately 2Kb upstream from the start site of the Poplar *AOS* gene, which was fused to a GUS reporter gene using the pWGB vector (Earley et al 2005). This was done to track the induction of expression of AOS in leaves after mechanical wounding which was used to mimic wounding by herbivory. *Arabidopsis* leaves were transformed according to (Acosta & Farmer 2010) and
transformed leaves mechanically damaged and subsequently collected at 1, 2, 4, 8, 12, 24, 36, and 48 hour intervals to match the intervals during the greenhouse herbivory experiment. Damaged leaves were stained with X-GLUC to detect induced GUS expression, and the level of stain was measured at the time of sample collection as a surrogate for AOS expression.

1.4 Results & Discussion

1.4.1 CYP74 Gene Family Phylogeny

Three distinct monophyletic clades were evident from our phylogenetic analysis of the gene family coding for the CYP74 cytochrome P450 proteins in plants (Figure 4). Division of the Angiosperm clades into clear monophyletic Monocot and Dicot subclades was evident for divinyl ester synthase (DES), fatty acid hydroperoxide (HPL) and allene oxide synthase (AOS) genes. Gymnosperm and basal plant CYP74s formed distinct clades suggesting that the CYP74 gene family evolved early in land plants underscoring the development of a role for JA in response to both biotic and abiotic stresses in the basal plants. While orthologous members of the CYP74 gene family were found in the basal plants, gymnosperms, and angiosperms, the gene family diversified in the angiosperms with a single ancestral gene found in the basal plants duplicating twice to form the three paralogous clades represented here (DES, AOS, HPL). In agreement with previous reports (Lee et al 2008, Nelson 2009), no homologs for CYP74 genes were found across the algal genome of Chlamydomonas reinhardtii or other algal plant genomes by BLAST searches.
The phylogeny did not show the hypothesized partitioning between predicted orthologs specific for 9- or 13-hydroperoxides (Stumpe & Feussner 2006) although that specificity may be epigenetic or spatially maintained by transit peptides (Froehlich et al 2001). Contrary to all previous reports (Froehlich et al 2001), this comprehensive phylogeny indicates that *HPL* is the most basal clade in the CYP74 gene family within the flowering plants, with *AOS* and *DES* genes being more derived. Previous work on the CYP74 gene family has suggested that either *DES* (Gretchkin 2002) or *AOS* (Nelson 2009, Brash 2009) is the evolutionary oldest member of the CYP74s. Two angiosperm-wide gene duplications are observed in our phylogeny, one giving rise to *HPL* and another gene that duplicated again to form paralogous *AOS* and *DES* clades. The split giving rise to *HPL* and *DES/AOS* may correspond with the *epsilon* duplication event and the *DES* and *AOS* split may correspond to the *delta* duplication event (Jiao et al 2011). *DES* is most likely the last gene member to diverge, after the Monocot/Dicot divergence according to both the phylogeny and the molecular clock analysis. While diversity is noted within the basal and Gymnosperm CYP74 clades, more genomic and biochemical data is needed to determine if there are specific differences within CYP74 family enzymes with regards to chemical specificity in the basal plants.

Unique differences were noticed in both the numbers of CYP74 genes in *Arabidopsis* and *Populus*. *Arabidopsis thaliana* contained just one *AOS* and one *HPL* gene and did not have a DES homolog in the phylogeny (Figure 1.2). This observation is validated by the
fact that colneleic and colnelenic acid exogenously supplied to *Arabidopsis* had no effect on the production of phytochemicals commonly found as derivatives from *DES* (Vellosillo et al 2007). The *Populus trichocarpa* genome, on the other hand, contained a single *AOS* gene, responsible for JA biosynthesis, as well as two copies of *HPL* and three copies of *DES*.

### 1.4.2 The Evolution of the Octadecanoid Pathway in Plants

Important questions exist regarding the evolutionary history and expression patterns of the CYP74 genes, and the resulting activity of their enzymes. The questions remain because until recently there was not sufficient sequence data to construct a robust, rooted phylogeny. *AOS* has been identified in a different form in animals (corals and Cephalochordates) and not found in any bacteria, single-celled algae, or fungi in the same functional form (Lee et al 2008).

HPL ultimately produces various short-chain aldehydes and ω-fatty acids known as green leaf volatiles (GLVs). Little is known about the role HPL plays in the production of volatile compounds, but based on our phylogeny, the presence of HPL was the first of the CYP74 cytochrome P450 enzymes to arise in plants. This suggests that GLVs may have been selected for expression during plant tissue damage or stress early in the evolution of multi-cellular plant structures. GLVs can act as volatile or nonvolatile compounds involved in plant defense and growth regulation (Froehlich et al 2001, Gretchkin 2002, Stumpe & Feussner 2006). Some of the probable leaf compounds made by HPL in damaged leaves are aldehydes, alcohols, and the defensive compound traumatin (Matsui
There may be a benefit to GLV production from an individual fitness perspective, considering that plants are most likely to be related to other plants in their close surroundings. Since plants cannot move in response to stress, these volatile alcohols and aldehydes could be used as a signal to notify other members of the same species of attack.

Allene oxide synthase (*AOS*) produces unstable allene oxides, which are converted via the octadecanoid pathway to JA and other oxylipins (Brash 2009). First recognized as hydroperoxide isomerase and hydroperoxide dehydratase (Gardner 1991), AOS conducts the first committed step in the oxylipin pathway by producing allene oxides. Unstable allene oxides derived from *AOS* have two fates: they may either be hydrolyzed into α- and γ–ketol fatty acids or they may undergo non-enzymatic cyclization (Brash 2009).

Divinyl ether synthase (*DES*) forms a diverse array of divinyl esters that appear to be involved in plant defense although their exact role is not entirely clear (Stumpe & Feussner 2006, Brash 2009). *DES* transcripts are induced during plant-pathogen interactions (Brash 2009). *DES* is responsible for the catalysis of hydroperoxides into various divinyl ether fatty acid compounds and has been reported to produce colneleic and colnelenic acids from HPOT (Fammartino et al 2007). Although these compounds are assumed to have a role in plant defense (Stumpe & Feussner 2006) more research is needed on the specificity and functional role they may play in the oxylipin pathway. Products of *DES* readily undergo variable acidic hydrolysis, which contributes to their chemical diversity (Brash 2009). Volatile scents, such as etherolic and etherolenic acids,
are products of *DES* and constitute portions of the smell of garlic and onion (Gretchkin 2002, Stumpe et al 2008). Fammartino et al (2007) reported that in *Nicotiana DES* is locally responsible for defense against pathogens because its expression was reported only adjacent to infected tissues.

### 1.4.3 CYP74 Conserved Motif Regions

Amino acids in CYP74 proteins were highly conserved (Figure 1.3) among six dicots across all described motifs Itoh *et al* (2002). A total of 179 promoter binding motifs were identified for *AOS* and 253 for *HPL*. The promoter analysis for Poplar *DES* genes has not been completed.

### 1.4.4 AOS “real-time” Gene Expression and Promoter GUS Reporting

Analysis of *AOS* transcript expression by “realtime” qRT-PCR, was conducted in hybrid poplar (OGY: *P. deltoides* x *P. nigra*) (Figure 1.3). Expression of AOS was observed to be induced in leaf tissues following gypsy moth feeding damage. Without herbivore damage, AOS expression was highest in petiole and phloem tissues, suggesting a role independent of herbivory.

The induction of *AOS* by wounding was also demonstrated *in situ* in *Arabidopsis* leaves transformed with poplar a *AOS* promoter::GUS fusion (Figure 1.8). Staining for GUS activity revealed activation of the *AOS* upstream promoter region within seconds after mechanical wounding. Initially, there was a slight diffuse response, but the GUS
expression from the *Populus AOS* promoter is clearly within the wounded region of the leaf. Expression of the promoter region became more diffuse in the leaf after 24 hours. It appears that there was also GUS reporter expression along the leaf midrib, which may suggest that a damage signal is communicated via the vascular tissue.

1.5 Conclusions

The sessile nature of plants adds to the critical nature of responding to both abiotic and biotic stresses. The complex evolution of a multi-faceted “plant immune system” has arisen to deal with these stresses (Jones & Dangl 2006). The first-dedicated step of the JA stress response pathway has evolved to no only generate oxylipins that contribute to the formation of JA, but there is a diversifying step in the generation of oxylipin compounds that may have other functions in plant defense. We determined that the CYP74 gene family diverged into three previously recognized, although uncharacterized, genes which putatively have different function *in planta*. These three gene sub families (*AOS, DES*, and *HPL*), presumably diversified after numerous genome duplication events. The evolution of the CYP74 cytochrome P450 gene family in land plants was presumably a step in the direction of not only taking an environmental affect and translating that signal into a hormone response (through *AOS*), but also signaling other plants or defending against other stresses (through *HPL* and *DES*). The maintenance of a potential bottleneck on such a presumably vital signaling phytohormone is the subject of ongoing work, including the interrogation of other plant reference genomes. Some evolutionary geneticists have argued that the most vital genes, possibly a category in which we could place *AOS*, are kept single copy to prevent interference with the function
of the protein that could arise from divergence of duplicate genes, although there are several explanations for the retention of single copy genes in the flowering plants (Duarte et al 2010). This evolution in plants of the use of fatty acid derivatives in defense, the oxylipins, has accentuated the diversity and roles of lipid compounds that are not at a premium in damaged plant tissues.

By mechanisms that are still not fully understood, JA either directly or indirectly regulates a diverse array of developmental processes in addition to its role as a defense mechanism. These developmental roles include pollen formation, the regulation of carbon allocation, and possibly may include secondary metabolite production. Jasmonates and other oxylipins have been implicated in plant development and growth in numerous ways such as leaf senescence, root development and tuber formation, germination and seedling development, flower and pollen formation, and tendril coiling. Thus, understanding the relationships, evolution, and regulatory pathways hypotheses responsible for JA synthesis is of critical importance.

1.6  Acknowledgements
We would like to thank Claude dePamphilis for critical reading of an early draft of this manuscript. Chris Frost generously provided the gypsy moth caterpillars and assisted with the herbivory treatments. This work was supported by the World Class University Project R31-2009-000-20025-0 from the Ministry of Education, Science and Technology of Korea, with additional support from the Schatz Center for Tree Genetics at
Pennsylvania State University. A portion of this study was funded by a grant from the Pennsylvania State University College of Agriculture Competitive Grants Program to JRH.

1.7 References


Figure 1.1 – Hypothesized enzymatic pathway of derivatives from lipoxygenase (LOX) and the start of the octadecanoid pathway and oxylipin formation by AOS to form jasmonic acid.
Figure 1.2 – Simplified representation of the Jasmonic Acid pathway in plants
Figure 1.3 – Gene expression data (qRT-PCR) for AOS in specific Poplar tissue types. Gene expression studies were carried out in hybrid poplar (OGY: *P. deltoides* x *P. nigra*).
Figure 1.4 – Phylogeny of the CYP74 gene family constructed from highly significant BLAST hits (less than $2e^{-20}$) across plant EST libraries. Phylogeny constructed in BEAST. Consensus tree Bayesian analysis with 473900 burn in states and more than 6 million states. Collapsed polytomies were expanded to show clade resolution.

<table>
<thead>
<tr>
<th>Prior versus alpha</th>
<th>mean</th>
<th>stderr of mean</th>
<th>median</th>
<th>geometric mean</th>
<th>95% HPD lower</th>
<th>95% HPD upper</th>
<th>auto-correlation time (ACT)</th>
<th>effective sample size (ESS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-845.636</td>
<td>1.0147</td>
<td>0.1897</td>
<td>1.3894E-3</td>
<td>-855.0896</td>
<td>0.9708</td>
<td>3018.9897 6455.4225</td>
<td>592.2511 276.9764</td>
</tr>
</tbody>
</table>
Figure 1.5 – Consensus tree from Bayesian analysis with posterior support values.
Figure 1.6 – Amino acid alignment for seventeen plants with sequenced genomes representing all recognized members of the CYP74 family: HPL, DES, and AOS for Arabidopsis, Oryza, Populus, and Vitis. Two Vitis DES genes are shown; one that shows sequence homology to other DES genes and one with a large insert. Phylogenetic trees were constructed with protein alignments and agree with the nucleotide phylogeny. Bootstrap values are shown below branches. Green color corresponds to hydrophillic amino acids, Red color corresponds to hydrophobic.
Figure 1.7 – Promoter motifs found within 4 kb upstream of the allene oxide synthase (AOS) gene in *Populus trichocarpa*. 
**Figure 1.8** - Promoter motifs found within 4 kb upstream both of the hydroperoxide lyase (HPL) genes in *Populus trichocarpa*
**Figure 1.9** – GUS reporter assay showing immediate induction of expression of the AOS promoter from *Populus* transiently transformed *Arabidopsis thaliana*. Assays were conducted and measured immediately after wounding, and stopped by fixation after 1 minute, 1 hour, 2 hours, 6 hours, and 24 hours.
Figure 1.9 - Continued

Unwounded companion

Fixed after 1 hour

Fixed after 2 hours
Figure 1.9 - Continued

Unwounded companion

Fixed after 6 hours

Fixed after 24 hours
Chapter 2

The challenge of identifying operational taxonomic units from environmentally-derived fungal amplicon data

Joshua R. Herr
The Schatz Center for Tree Molecular Genetics, School of Ecosystem Science and Management, Pennsylvania State University, University Park, PA 16802, USA

Key Words: Fungal Ecology, Next-Generation Sequencing, Operational Taxonomic Units (OTUs), Community Analysis, ITS rRNA, Phylogenetics

For submission to the journal Fungal Ecology

Abbreviations: OTUs: operational taxonomic units; rRNA: ribosomal RNA; NMI: normalized mutual information; MSA: multiple sequence alignment; PSA: pairwise sequence alignment; LSU: 26S, Large-subunit ribosomal RNA; ITS: Inter-transcribed spacer region; SSU: 18S, Small-subunit ribosomal RNA

2.1 Abstract

Next-generation sequencing techniques, leading to both amplicon sequencing and metagenomic analysis, can be used to assess fungal diversity in environmental samples
without culturing. These new techniques are revolutionizing the way we observe environmental samples. We can begin to address questions regarding fungal biology and ecology that we could not have addressed in the past, mainly how we understand species interactions and assemblages in different environments, how fungal communities are structured, and how fungi may be influenced by their plant and animals hosts. Many new technologies have required a new array of data analysis and methods to identify this newly recognized diversity. Here, I review some computational methods for the analysis of next-generation amplicon sequence data.

2.2 Introduction

Fungal communities are highly diverse – a single gram of forest soil may contain up to thousands of individuals (Buée et al. 2009) – and are notoriously difficult to identify. Perhaps the majority of fungi are unculturable, many lack conspicuous fruiting bodies, and when morphological characters are easily observable many exhibit cryptic morphology with a paucity of identifying characters (Hibbett et al. 2011). As a result, fungal ecologists, along with microbial ecologists sampling bacteria, archaea, and microscopic Eukaryotes, have been among the first to embrace the use of DNA sequencing for the identification of taxa from environmental samples (Pace 1997).

A reliable taxonomic identification is the foundation to address a wide array of biological, ecological, and pathological questions. When addressing environmental sequencing, assessments such as species biodiversity, the causation of plant or animal diseases, or identifying the roles of fungi in biogeochemical cycles – among other
questions – begin with identifying the fungal taxa present in a sample. Even disregarding environmental samples, if a particular fungus exhibits well-defined macro- and microscopic morphological characters, molecular taxonomy is vital for taxonomic verification or identification of cryptic species. Taxon boundary determination for fungi is particularly vexing, where some taxa are only known from environmental molecular sequences and where cryptic speciation may exist in many others (Schoch et al 2012).

The need for a ubiquitous and appropriately conserved marker gene for determination of fungi from the environment has led researchers to use the ribosomal DNA (rDNA) operon (Schoch et al 2012). The ribosomal small subunit (SSU, 18S), the large ribosomal subunit (LSU, 26S) and the 5.8S, most importantly with two internal transcribed spacer regions, ITS1 and ITS2, among other markers, have all became de facto markers of choice for studies surveying fungi from environmental samples, with the ITS sequence regions becoming the region of choice (Schoch et al. 2011).

One of the benefits of sequence data is that it is well suited for data analysis, hypothesis testing, and model-prediction, including robust methods of sequence differentiation based on phylogenetic inference or fingerprinting. All methods seek to identify operational taxonomic units (OTUs) ideally representing branch tips on a phylogenetic tree (Hamady et al 2010). Ideally, any fungus representing an individual terminal node of a phylogenetic tree – what one could consider a species, sub-species, cryptic species, or population variant – can represent an OTU. While sequence database query has remained the gold-standard for fungal identification, this technique is inherently low-
throughput. Fungal ecologists have made good use of numerous techniques to determine OTUs, such as T-RFLPs (Terminal Restriction Fragment Length Polymorphisms), DGGE (Denaturing Gradient Gel Electrophoresis), and ARISA (Automated Ribosomal Intergenic Spacer Analysis) to increase throughput and identify numerous taxa from mixed environmental samples (Kuczynski et al 2010).

Here I review the rationale for defining OTUs from environmental samples from fungal amplicons, or meta-taxonomic data, generated from next-generation sequencing techniques. I address the computational approaches for classifying amplicons into OTUs and discuss numerous approaches for the analysis of co-occurrence and measuring diversity in fungal ecological studies using amplicon data.

2.3 Putting a name on it? Operational Taxonomic Unit (OTU) Based Analysis of Fungal Individuals from Environmental Samples

Identifying OTUs using sequence markers assumes caveats from both the biological and morphological OTU concepts. When comparing two sequences using a particular homologous marker, a lack of diversity can indicate two options: the sequences represent the same individual or there is not enough resolution to separate two OTUs based on that particular marker. Sequence differences result in opposing conclusions regarding the separation of OTUs. Homologous marker differences are typically the result of distinctive taxa, but may also be attributed to sequencing error, intra-genomic heterogeneity of multi-copy markers, horizontal gene transfer events, species hybridization, or population level differences manifesting themselves as SNPs, insertions
or deletions, or small repeats that reflect the normal within-OTU variation.

Marker regions should have an adequate amount of variability to be able to resolve species (Schoch et al 2012). When identifying OTUs, it is important to have markers that correspond to orthologous rather than paralogous genes, as the goal of OTU determination should be based on phylogenetics of the actual organisms in question, rather than the phylogeny of a specific gene family. To reduce issues associated with paralogous genes, marker gene should optimally be found in minimal copy number with single copy genes being the most desirable. Some debate on whether fungal OTUs can be truly named or not (Hibbett et al 2011).

The computational determination of this sequence data adds layer of complexity. In other words, in addition to numerous environmental caveats, the determination of OTUs from amplicon sequences adds another variable into consideration. The determination of OTUs depends largely on the type of clustering algorithm used and these different methods of OTU determination may yield different results. Due to these difficulties, mycologists have moved towards the identification of individuals as OTUs.

As taxon boundaries cannot directly be inferred from a phylogenetic tree, researchers mostly apply clustering algorithms in combination with predefined thresholds to pair-wise genetic distances in order to assign sequences to molecular operational taxonomic units. However, the chosen thresholds differ in the literature, even if applied to the same organisms and molecular markers, and are often based on subjective criteria.
or just on tradition. Furthermore, the clustering algorithm also may have a significant impact on the outcome, but it is seldom addressed which algorithm is most appropriate for molecular taxonomy. Further potential sources of biases are alignment ambiguity and rate heterogeneity between sites.

The clustering of OTUs is essentially not reproducible because every sample sequenced from the environment, even from the same location and time, will consist of a slightly different composition. Once OTUs are identified, labeled and have a molecular reference sequence, they may be referred to and referenced in other studies. This can be done by identification to databases of known sequences. OTU identification is really only as good as the database used to differentiate individuals. The limitation of databases is that tag sequences typically represent only a small fraction of these organisms are known and well-studied. Thus, studies based on the relationships of known tag sequences may be biased and do not present full understanding of the microbial diversity in communities. On the other hand, OTUs do not depend on the available information about the known tag sequences and thus present an unbiased view of the microbial diversity. The main challenge during amplicon clustering is the accurate identification of biological taxa with the subsequent goal of being able to identify the same taxa across multiple samples. Tag sequences can be grouped into different clusters such that the sequences in each cluster are similar, but sequences in different clusters may have relatively large differences. The sequences in each cluster form an operational taxonomic unit (OTU). It is difficult to quantify the total abundance of individuals based on sequence recounts from NextGen data. There is extreme variation in the number of nuclei in certain groups of fungi such
as *Glomus*, etc. (Hibbett et al 2011).

Documented known and accepted taxa may be sequenced, and the level of variation can be measured for a particular marker, and resequencing of this individual will provide an observational variation rate per marker. Comparison of variation between OTUs as opposed to within-taxon variation and error will provide the accuracy of the OTU determination. This can be beneficial as many marker genes from differing taxonomic levels have variable rates of substitution and evolutionary clock speeds. Additionally, different sized populations (or populations with differing evolutionary history) of environmental OTUs will contribute different levels of within-taxon variation. Multiple markers can be beneficial in this regard.

Adding to the difficulty of choosing and implementing different algorithms, there are numerous artifacts associated with using sequence data for OTU determination. Sequencing errors may contribute to the overestimation of OTUs by allowing a greater estimation of sequence diversity in a sample. Typically, researchers use a cut-off of 95% to 97% similarity to account for on average 3% sequencing error (Quince et al 2009). Large data sets provided by next-generation sequencing techniques are computationally and memory intensive. Computational memory is typically reduced when handling sequences with errors derived from next-generation sequencing platforms (Quince et al 2009) and the optimal methods for dealing with sequencing errors are still being debated in regards to the clustering methods which give the most biologically meaningful results. Sequencing error is a major problem in OTU clustering because of this. When
sequencing OTUs it’s important to know the rate of sequencing error, the types of sequencing error (transitions vs. transversions, etc.). Identifying the actual OTU and not sequencing error, the actual amount of nucleotide variation that occurs in an OTU.

For fungi that uncontaminated DNA can be extracted, such as those that may be cultured or produce sporocarps, the amount of interspecies homology can be determined by sequencing all of the copies of a particular marker using next-generation sequencing techniques or Sanger sequencing of cloned markers. One of the benefits of using a single copy marker is the lack of interspecific variation in the marker itself. The amount of marker variation can be associated with the levels of morphological variation among species complexes and putative cryptic species.

2.4 Computational methods for the identification of OTUs

Sequence clustering is highly computationally intensive and typically requires a large amount of computer memory. If you have a small dataset, which is unlikely, you may be able to use a web server or a laptop or desktop computer; if this is the case you will have to know how much computational memory will be need to be allocated to your analysis. Larger datasets may require the use of a computer cluster for data analysis. With the exception of the few online clustering methods located on the web, clustering programs are generally downloaded, installed, and operated via the command line on LINUX-based computers, so users are recommended to have some basic competency. Some clustering programs may not be regularly maintained after their introduction so it is important to understand the respective algorithm strategy of each program and keep track of errors.
that may occur during the clustering run.

Typically the initial step to dealing with raw data from an amplicon sequencing study is to bioinformatically remove platform-specific sequencing adapters added during library construction. If numerous mixed-template samples were barcoded and multiplexed on a sequencing run, barcodes are removed as the sequences are de-multiplexed and placed into specific sample FASTA files or recorded in the sequence title in a single FASTA file along with the corresponding sequencing read number.

2.4.1 Pre-clustering: Addressing Read Quality, Sequence De-noising, Chimera Checking and Pre-clustering to Reduce Computational Load

Error rates for sequencing amplicons using high-throughput sequencing methods are less than or equal to Sanger sequencing (Quince et al 2009). Despite accurate base calling, the sheer volume of sequence data generated from high-throughput sequencing platforms will naturally produce sequences with errors whose numbers will be correlated to sequencing run size. Sequencing errors may be easily predicted and identified by assessing both the sequence run quality scores and by comparing the amplicon data to other sequences in the data output of that respective sequencing run (Quince et al 2009). Using read quality data, short or poor quality reads are removed from the dataset and may be manually inspected for quality control purposes. This step is typically computationally intensive based on the size of the data set. Errors are attributed to a number of factors, including the presence of homopolymeric regions, the deterioration of read quality towards the end of the sequencing run, or the physical location on the
sequencing plate (Quince et al 2009), none of which are typically pronounced in amplicon sequencing studies. Nucleotide image quenching may be a factor in amplicon studies where a large number of nucleotides will fluoresce in unison causing an increase in base calling errors. To avoid this, some researchers have suggested using staggered primers to avoid in sync nucleotide reads (Hamady et al 2009).

2.4.2 Quality Assessment and Denoising

One of the initial steps in assessing the quality of sequencing reads is observing quality measures of the sequences. Whether reads originate from 454 Pyrosequencing or Illumina sequencing, the quality assessment does require large computational resources to view and cull poor reads from the dataset, not to mention large processing power to store and analyze data as well as high-capacity network capabilities to move sequence data from sequencers to servers and workstations. Programs for correcting FASTA or FASTQ files use read quality assessed by the sequencer. In the case of de-noising programs, an algorithm removes sequencing errors by looking at read abundance levels and then removing poorly called reads (Quince et al 2009). One recognized downside to error correction is that rare species, as well as cryptically recognized species, may be removed from the data set.

2.4.3 Chimera Detection

In addition to sequencing errors, assembly errors that are the result of sequencing assembly can also be an error associated with sequencing. Hybrid sequences that are closely related to sequences in the data set are very challenging. PCR artifact, hybrid
mixture, two unrelated molecules, incomplete PCR extension, this fragment then ask has a primer for other sequences and the two hybridize, discrete junction between two molecules, unrelated. Thus, the use of 16S/18S rRNA genes may not be optimal for community comparison.

One of the most frequently used ways to reduce computational time used by clustering algorithms is to pre-cluster raw sequencing reads. In all pre-clustering steps of the algorithm, the sequences are filtered from the active data set and it is key that abundance information is recorded. This is commonly done by filtering out reads that exhibit 100% similarity to other sequences in the data set. Because OTU clustering algorithms typically use a strategy that compares each sequence to every other sequence the computational time is approximately proportional to the squared number of sequences in the data set. Filters have been developed at this stage to recognize similar conserved sequence motifs and this can dramatically reduce compute time for those wanting to focus on specific groups of fungi in their meta-taxonomic study (Hibbett et al 2011).

Measures of confidence, of sequence assignment confidence, confident scores below a certain threshold, such as a bootstrap level of 70%, can be de-emphasized, BLAST e-value.

### 2.5 Local Sequence Similarity - *de novo* OTU Clustering

In local sequence similarity methods, sequences are subjected to *de novo* OTU assignment where reads are compared to each other and not to a database or previous collection of OTUs. Most algorithms share this initial step for clustering of OTUs,
although some algorithms only use this sequence similarity clustering step. This initial clustering step is typically the computation of a pairwise or distance matrix between sequences. Benefits include not having to have a reference database, uses all sequence reads.

Algorithms used to cluster sequences into OTUs typically share two steps in common. The initial step of OTU clustering programs calculate distances pairwise between any two sequences or use multiple sequence alignment to align all amplicons reads at one time. Usually a distance measurement matrix is calculated on the basis of this alignment (Schloss & Handelsman 2005, Schloss et al 2009). The second step of OTU clustering relies on using the pairwise distances of these sequences to play sequences in similarity groups (Schloss & Handelsman 2005, Schloss et al 2009, Sun et al 2009, Hao et al 2011). Initial algorithms took complete linkage two groups of OTUs would be compared by taking the maximum distance between two pairs of aligned sequences from each group from hierarchical clustering (Schloss & Handelsman 2005, Schloss et al 2009). Subsequent algorithms have taken a more refined approach by taking the average distance between OTU pairs.

A major issue with the clustering step of OTU determination is the computation of pairwise distances among all sequences from metataxonomic studies. This step of the analysis is computationally challenging due to limitations of computer memory. Clustering algorithms have been developed to reduce this issue by limiting pairwise matrix data by only recording distance values from close neighbor taxa instead of all taxa
in a study (Caporaso et al 2010). Other clustering algorithms, such as UCLUST (Edgar 2010), CD-HIT (Li et al 2001) and ESPRIT-Tree (Cai & Sun 2011), only cluster when new sequences are added to the analysis. Both strategies greatly reduce the number of pairwise distances in memory, especially for amplicon studies – as opposed to shotgun metagenomic sequencing – because the use of a marker provides a majority of highly similar sequences. These clustering algorithms each are slightly different in strategy and may provide differing results and clustering time.

2.6 Database Based methods

2.6.1 sequence similarity search methods

Alignment methods typically do not provide a confidence estimate other than a blast e-value, the number of base match/mis-matches and the amount of sequence similarity. These methods take inquiry sequence and alignment to a reference sequence reference database. Programs that use sequence similarity search methods used the blast algorithm or a modification of the blast algorithm, BLASTX, BLAT, several identified the LCA or lowest common ancestor algorithm, (see MEGAN, Huson et al 2010).

Sequence similarity searches, these alignments, are fairly accurate, especially accounting for the short sequence of some Amplicon reads, two disadvantages, particularly with large databases references that are very large you can take a long time to search each query, if databases are incomplete and a representative is not found in the database then assigning the read can be problematic and this is quite common with fungi.
2.6.2 Database methods, sequence composition models

Sequence composition methods, models for sequence comparison, various methods are used for sequence comparison these can be interpolated Markov models native Bayesian classifiers and K means or K nearest neighbor algorithms. And time is needed to establish models so that’s time-consuming, once these models are built the sequence classification generally faster with alignment then alignment-based methods or comparing alignment-based methods. Fungal LSU (Liu et al 2012), accuracy can be an issue as is the case with databases incomplete databases can be an issue in compiling models. Typically these types of composition models are better associated with longer reads but the algorithms are being improved in recent versions and they may be more applicable to shorter reads the longer reads of fungi should be no problem for the native classifiers. For bacteria taxa, there are numerous 16S databases that have been developed for classification, such as SILVA (Pruesse et al 2007), Greengenes (DeSantis et al 2006) or the RDP project database (Cole et al 2010).

2.6.3 Phylogenetic methods for clustering

The first step in phylogenetic clustering is the generation of a phylogenetic tree from sequence data provided by the researcher prior to clustering. Serving the same function (and often the same sequence data) as a training database. Phylogenetic methods are those that take evolutionary relationships of the sequences into consideration in the comparison of communities. First, a phylogenetic tree, including all the sequences in the samples, is generated using a phylogenetic analysis. Each sequence is labeled according
to the community the sequence comes from. Based on this observed tree, the minimum
number of changes needed to explain the labels, termed parsimony score, is calculated.

Numerous clustering algorithms implement phylogenetic methods to cluster and delineate
OTUs. Phylogenetic clustering programs implement algorithms, sometimes more than
one, such as neighbor-joining (NJ), maximum-likelihood (ML), or Bayesian inference,
for the phylogenetic placement of sequence data and the determination of the OTU.

Phylogenetic placement of Amplicon sequences, various methods are used to locate a
sequence on a phylogenetic tree using different evolutionary models. Programs that
compute the branch length of the inserted Amplicon represents the amount of inquiry
sequence has evolved relative to the rest of the tree.

Because these methods are computationally intensive phylogenetic algorithms are only
really viable for the use of marker sequences, such as amplicons suggested here, and not
feasible for the use of total metagenomic reads. Even for amplicon studies, the
phylogenetic breadth covered in most databases, compounded with the diversity assessed
in experiments is computationally intensive. With the advance of computational power,
parallelization of computer clusters, and the development of newer and faster
phylogenetic algorithms, this cluster time will decrease.

One of the benefits of phylogenetic methods over other clustering methods is that
confidence levels of OTU assignments can be assessed through tests such as bootstrap or
jackknife for maximum likelihood or posterior probabilities for Bayesian analysis.

2.7 Evaluation of clustering

One criterion used to evaluate algorithms for defining OTUs is comparing the number of OTUs given by the algorithms with the known number of species in the simulated community. When comparing different methods for defining OTUs, investigators designed some benchmark data by either experimentally sequencing a community with known microbial species (Huse et al 2010, Quince et al 2009) or computationally selecting a set of species and introducing errors in these sequences according to the sequence error models of sequencing technologies (White et al 2010, Schloss & Westcott 2011, Sun et al 2011).

2.8 Putting it all together: Association networks and assessment of environmental factors

2.8.1 Diversity measurement

Some researchers have compared environmental communities in terms of relative taxon abundance. In which they normalized reads per OTU and take into consideration the intragenomic variability (Lindner & Banik 2012). Intragenomic variation can confound methods to differentiate sequencing errors and understand the diversity that exists and environmental data. There needs to be critical evaluation of rare designations for low abundance taxa. Numerous metagenomics analysis pipelines, including QIIME (Caporaso et al 2010), MOTHUR (Schloss et al 2009), and MG-RAST (Meyer et al 2009) have included diversity analyses for rare taxa.
It should be noted that one should not discard rare OTUs. If you're interested in dominant OTUs in all samples, some of the apparent rare OTUs in one sample might be high abundance in another. Numerous measures – including alpha, beta, and gamma diversity – have been developed to measure the biological diversity in a given environment and methods for separating diversity have been reviewed in Whittaker (1972). Alpha diversity is the local level of diversity, typically recorded in the number of species found in a given habitat.

When comparing two or more habitats, beta-diversity is the relative change in diversity, typically recorded in the number of taxa found only in one habitat versus another. There are numerous measures of beta-diversity that are categorized as quantitative or qualitative. Qualitative measures, as the name suggests, address taxon presence or absence within communities, disregarding their abundance (Hamady et al 2009). Some common qualitative measures of beta-diversity include indices such as Jaccard, Sørenson, Lennon, and Dice (Chao et al. 2005). Quantitative measures of beta-diversity take into consideration the abundance of OTUs and include statistics such as Euclidean, Bray-Curtis, Canberra, and Chao (Chao et al. 2005).

2.8.2 Network Analysis

Really hypothesis generation tools regarding the designation of OTUs and in some cases can be directly constructed from raw OTU data. Network representation tools such as the
popular program Cytoscape have originally been designed to map gene network
associations in cells, but can also be used to conduct a network analysis of fungal OTUs
and environmental factors. How does this work? Ease analyses can manifest themselves
as negative or positive associations that would imply a correlation between presence or a
correlation between absence. Preference for varying seasons competition predation
Symbiosis of bacteria and fungi.

2.9 Conclusions

The field of metagenomics and molecular ecology are rapidly advancing. With large
amounts of data, there is a need for the development of efficient and accurate
computational measures of data. By assessing new datasets, researchers can understand
fungal community structure, compositional change, and environmental associations with
changing communities.

As I have provided an overview on the clustering of next-generation sequence data from
fungi here, there is a need to more exploratory studies addressing new methods to analyze
the large amount of sequence divergence that is observed in the fungi. Besides inherent
problems with trying to identify OTUs based on single genes or amplicons, the value of
determining species diversity is great (Gilbert et al 2009).

2.10 References

identification of bacteria using small subunit ribosomal RNA sequences. Biotechniques
17(1), 144-149


Liu WT, Marsh TL, Cheng H, Forney LJ, 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology* 63(11), 4516-4522


Lozupone CA, Hamady M, Kelley ST, Knight R, 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* 73(5), 1576-1585

Martin AP, 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology* 68(8), 3673-3682


Muyzer G, de Waal EC, Uitterlinden AG, 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* 59(3), 695-700


Schloss PD, 2008. Evaluating different approaches that test whether microbial communities have the same structure. *The ISME Journal* 2(3), 265-275


Schloss PD, Handelsman J, 2005. Introducing DOTUR, a computer program for defining
operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* 71(3), 1501-1506


Schloss PD, Westcott SL, 2011. Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Applied and Environmental Microbiology* 77(10), 3219-3226


Chapter 3

Metagenomic sequencing of South Korean forest soils reveal plant host and spatially mediated microbial structure.

Joshua R Herr¹, Donghwan Shim¹², John E. Carlson¹²

¹The Schatz Center for Tree Genetics, Department Of Ecosystem Science & Management, The Pennsylvania State University, University Park, PA 16801 USA
²Department of Bioenergy Science & Technology, Chonnam National University, Gwangju, 500-757 South Korea

Key Words: Metagenomics, Forest Soil, Microbial Diversity, whole shotgun sequencing,

3.1 Abstract

New sequencing technologies have drastically affected our ability to address microbial ecology in soils. In forest ecosystems it is believed that host trees are the predominant driver of microbial species diversity in soils and biogeochemical status of soils that mediate soil microbial diversity. In order to address soil microbial diversity host-tree affects, we assessed microbial diversity in five monoculture forest plots in South Korea. The dramatic history of forests in South Korea provides a unique opportunity to address the role of tree host in mediating microbial diversity. Using metagenomic shotgun sequencing of total DNA from soils in these forest stands, we identified the taxonomic components identifying bacterial, archaeal, and eukaryotic taxa, notably fungal diversity.
We observed a similar general structure of microbial components in each forest soil. However the data showed minute spatial and host associated differences, particularly in the eukaryotic taxa that were identified. While bacterial presence was fairly constant across all the sampled locations, eukaryotic diversity was spatially patchy.

3.2 Introduction

Soil microorganisms constitute a highly diverse and abundant microcosm responsible for contributing to a suite of ecosystem functions responsible for maintaining soil fertility through biogeochemical cycles and interactions with a wide array of plant species (Fierer et al 2012). We are just now beginning to understand species diversity and structure of microbial communities in soil, with forest soils exhibiting extreme diversity, but the role of plant host in regulating the diversity of associated microbes is not completely understood. With newer next-generation sequencing technologies we are now able to survey portions of whole soil genomes to capture the extent of microbial diversity in forest soils, including the majority of microorganisms that are unable to be identified by culture-based taxonomic approaches (Pace 1997).

There has been a long history of the study of community ecology of plants and animals and more recently, new sequencing technologies have made it possible to study microbial diversity in similar ways. We are now able to investigate how plants, particularly trees in this study, may influence the composition and community structure of associated microbes. Plants contribute to the input of organic carbon to the soil in the form of photosynthate and can also influence numerous geological and chemical components to
soil, such as pH (Rousk et al. 2010). Bacteria, archaea, fungi, and protists all have differing biogeographical patterns based on many factors, such as soil pH, macronutrient availability, micronutrient gradients, nitrogen, moisture, temperature, pH, etc. (Fierer et al., 2007; Rousk et al., 2010; Bates et al., 2011).

South Korea possesses an interesting land history that enables us to address the influence of plant host on microbial diversity. In 1910, Japan and Korea entered into the Japan-Korea Annexation Treaty, which gradually gave way to Japanese occupation of Korea during World War II (Bae et al., 2012). During this time, the Japanese – in need of raw materials for the war – harvested virtually all of South Korea’s native forests. Some locations remained clear cut for many years and were re-planted only when South Korea regained its independence in 1965 (Son et al., 2004). When trees were re-planted they were established as monoculture forests of approximately the same age (Bae et al., 2012). We chose five plots on the basis of host tree replanting, the estimated age of the stands, and the ability to collect meta-data associated with each stand.

We assessed microbial taxonomic diversity in five different forest soils representing trees cultivated at approximately the same age using shotgun metagenomic sequencing. We propose that plant host will affect the microbial taxonomic composition of these forest soils. Additionally, we address the questions: Do monoculture hosts support different arrays of microbial diversity? What soil metadata, such as micronutrient composition and geochemical data, is influenced with host or microbial composition in these soils? We analyzed the taxonomic and phylogenetic composition of the shotgun metagenomic
sequences and also surveyed for the presence of both 16S and 18S rDNA genes for use in well-curated taxonomic databases for archaea, bacteria, fungi, and soil-inhabiting eukaryotes.

3.3 Materials and Methods

3.3.1 Soil collection.

Five sites were selected on the basis of host tree, plot size, and the approximate age (55 years) of the forest plot: Mixed Oak (*Quercus variabilis* Bl. and *Quercus mongolica* Fisch.) Hardwood Plot with no understory vegetation (Plot 1 - 35.1312°N, 127.0037°E), Mixed Oak (*Quercus variabilis* Bl. and *Quercus mongolica* Fisch.) Hardwood Plot with understory vegetation consisting of *Gymnadenia cucullata* var. *variegata* Y. N. Lee, *Miscanthus changii* Y. N. Lee, and *Maianthemum bicolor* (Nakai) Cubey (Plot 2 - 35.1307°N, 126.9905°E); Korean Red Cedar (*Thuja koraiensis*) (Plot 3 - 35.1301°N, 126.9832°E); Dawn Redwood (*Metasequoia glyptostroboides*) plantation (Plot 4 - 35.1733°N, 126.8973°E), and Ginkgo (*Ginkgo biloba*) plantation (Plot 5 - 35.1754°N, 126.9013°E). Apart from the mixed Oak plot with vegetation listed above, none of the plots had understory plants, and no nitrogen fixing plants were documented in any of the plots.

Each forest plot varied in size from 25 m$^2$ to 200 m$^2$ depending on the original planting plot size. Within each forest plot, the upper 10 cm of soil was sampled using a 10 cm corer, included both the organic and mineral layer, and was taken from 9 locations in the geographic center of each plot using a randomly generated starting placement with at
least 3 m² difference in spacing from one sample to the next to form a square of 3 X 3 samples. Individual soil samples were taken back to the laboratory, sieved using a 2 mm screen, and a 10 g portion of the sieved soil was stored at -20°C until DNA could be extracted and another 10 g portion was also stored at -20°C and used for element and nutrient analysis.

3.3.2 DNA extraction.

For each of the soil samples, 10 g of sieved soil was homogenized with liquid nitrogen in a mortar and pestle. Total soil DNA was extracted from 0.5 g of each soil sample using the PowerSoil DNA Isolation kit (MoBio Laboratories, Carlsbad, CA). Each of the nine replicate subsamples of each plot were extracted for nucleotides in triplicate. The DNA samples were shipped from South Korea to Pennsylvania in a liquid nitrogen acclimated dry shipper. DNA quality was assessed prior to shipping and after the samples arrived with a Nanodrop 2000 spectrophotometer (Thermo-Fisher, Wilmington, DE, USA) to assess the amount of sequence degradation. Samples exhibiting any degradation were discarded from the experiment.

3.3.3 Whole genome shotgun sequence data.

Sequencing was conducted using the Illumina MiSeq platform using the library preparation protocol for 2 x 250 bp paired-end sequencing (Osborne et al., 2011) at Pennsylvania State University in the laboratory of Stephan Shuster. Individual samples were barcoded and run on two lanes of the MiSeq. A range of 380971 to 442559 of reads were obtained from each sample. Paired-end reads were quality filtered and processed
(Caporaso et al., 2011) prior to deposition in the MG-RAST (MetaGenome – Rapid Annotation using Subsystems Technology) platform (Meyer et al 2011) under study accession numbers 4503014 through 4503019 (MG-RAST Study Number 2220). Within the MG-RAST platform, sequences were annotated against the M5NR database using BLAT with the criteria of maximum e-value of 1e-5, a minimum sequence identity of 60% and a minimum alignment length of 15 bases (Bates et al., 2011), using the default parameters for metagenome annotation (Meyer et al 2011). Any reads that could not be annotated against the M5NR database in MG-RAST were discarded from further analysis. Other forest tree plots were sampled but did not provide robust initial sequencing results so the sequencing data was not included in this study. Predicted 16S SSU bacterial and archaeal rDNA reads were identified using MG-RAST and analyzed using the QIIME (Caporaso et al., 2010), MOTHUR (Schloss et al. 2009), and MEGAN4 (Huson et al 2007) platforms.

Predicted 16S sequences from the whole genome shotgun sequencing reads were compared to the non-redundant nucleotide database from GenBank using BLAST. Using the QIIME package and the open-reference phylotype picking python script (Caporaso et al., 2010), the predicted 16S reads from MG-RAST were clustered into phylotypes representing ≥ 97% sequence identity (Fierer et al 2012). Only a single identity percentage was used in this study to cluster OTUs as we did not measure sequence similarity between individual OTUs. OTUs matched the slope of the rarefaction curves for the whole genome shotgun sequencing portion of the study (Figure 3). Even with a deep sequencing depth of an average of 407896 reads from each plot, we were not able to
observe an asymptote, suggesting an extremely high level of diversity typically seen in forest soil environments.

Metagenomic 16S rRNA reads from bacteria and archaea were analyzed using QIIME and the Greengenes database in a closed-reference phylotype process to identify reads to 90% identity. The taxonomy of each phylotype was assigned as the closest hit in the Greengenes database. A similar strategy was employed in MEGAN4 using the NCBI GenBank database as a reference.

Principle component analyses using the biogeochemical data from the soils were determined using R scripts provided from the METAGENassist platform through QIIME (Caporaso et al. 2010) using pairwise distances from the metagenomic data (Bray-Curtis distances) and using analysis of similarities (ANOSIM) tests. These tests were conducted on sequences from each plot sample rarefied to 350000 randomly selected sequences from each plot and conducted pairwise t-tests with P values calculated using a Bonferroni correction for comparison of multiple plots.

3.4 Results & Discussion

3.4.1 General Soil Microbial Community Characteristics - soil pH and other mineral components of the soils.

Soils were sampled from five plots consisting trees different species but approximately the same age (35 years) in geographically similar soil chemistry, temperature, etc. Monoculture plantations were selected to compare the effects of tree species on microbial
diversity in these forest soils. By shotgun metagenome sequencing of these monoculture forest soils, we acquired a total of 2,039,482 paired end sequences. When mate pairs were matched, the resulting sequences averaged 281.6 basepairs in length with an approximate average of 40 base pairs of overlapping sequence. Approximately 6% of the sequenced reads did not pass quality control and were discarded from the analysis. Using MG-RAST’s M5NR database only 32% of the sequences could be accurately annotated. This number was in agreement with other studies that have conducted shotgun metagenomic sequencing of forest soils (Fierer, Leff, et al., 2012; Uroz et al., 2010; Delmont et al., 2011; Delmont et al., 2012).

The pH of soils has been a solid predictor of the bacterial diversity across soils. The pH of the forest soils did not differ enough to explain much bacterial diversity in each of the plots. In fact, all of the forest plots, regardless of type of host tree, provided a similar composition of bacteria. One would expect according to previous research that soils with a neutral pH would harbor the most bacterial diversity. Since all of our soils had approximately the same pH ranging from 4.71 to 6.78, we expected that pH would not affect the species composition nor diversity by a large amount in these forest plots.

3.4.2 Microbial Community Composition.

Reads from the shotgun sequencing provided the basis for our taxonomic assessment of species diversity in the forest soils. The reads provided two sources of species identification for microbial taxa. Both the predicted 16S/18S reads and the direct annotation of the metagenomic reads provided a similar assessment of the species
diversity of the forest soils. This was evident from the direct comparison of the relative
abundances of each value and confirmed with a strong correlation between the Bray-
Curtis distance matrices (Figure 3.5).

Not surprisingly, the majority of the shotgun metagenomic sequencing data (ranging from
93.6 to 95.2% of all reads) consisted of reads that were annotated to bacteria. The
shotgun sequence data was verified by both the overall predicted species identification of
the rRNA small subunit (SSU) and large subunit (LSU) prediction (Fierer et al. 2012).

With regards to bacterial and archaeal diversity, all of the forest soils contained a similar
constituency of bacteria (Figure 3.4) that are abundant and ubiquitous in forest soils
(Fierer et al). This included the following bacterial phyla, in order from most common to
least common: Proteobacteria, Actinobacteria, Acidobacteria, Fimicutes, Bacteroidetes,
Planctomycetes, Cyanobacteria, Chloroflexi, and Verrucomicrobia. Archaea were found
relatively rarely in the forest soils (less than 3% of total reads), but were represented by
three of the five Archaeal phyla: Euryarchaeota, Crenarchaeota, and Thumarchaeota.
The most dominant phyla of archaea were the Thaumarchaeota. The amount of archaea
in soils agreed with previous studies that were conducted by whole genome shotgun
sequencing of forest soils (Fierer, Lauber, et al., 2012).

Eukaryotic representation in soils is generally expected to be low (Delmont et al., 2012)
(Delmont et al., 2011) and our assessment agreed with this observation. There have been
numerous hypotheses for this phenomenon including eukaryotes having a low ratio of
rRNA genes when compared to bacteria per unit biomass (Fierer et al., 2009). All phyla of the fungi were represented in the forest soils with the Ascomycota and Basidiomycota being the most common.

3.4.3 Determination of Operational Taxonomic Units

We saw no relative differences in the levels of diversity derived from the rarefaction curves in the soil samples (Figure 3.3). While we did not achieve a total asymptote, we believe we were close to a depth of sequence to capture the total OTU based diversity of our soils, with an average of 691 individual OTUs per plot. It may not be necessary to sequence each sample to the asymptote of the rarefaction curve to attain the majority of diversity. Although rare OTUs can still be contributing to unknown biogeochemical cycles, the sampling accomplished here provides a fairly complete picture into the total microbial diversity of these forest soils (Green et al 2008, Wall & Virginia 1999). Measurement of other assessments of biodiversity, such as Chao I or ACE (Manzoni et al), agreed with our rarefaction curves.

The slopes of the rarefaction curves were lower for archaea and fungi when compared to the bacteria, but this is probably the product of species diversity or the overwhelming presence of bacterial DNA in soils (show diverse rarefaction curves), reflecting a misbalance of nucleotides in the soil environment rather than actual bacterial OTUs or species diversity present in the soils (Fierer et al 2012).
A total of 641 unique bacterial OTUs was estimated for the five plots measured, which agrees with previous studies of agronomic, tropical forest, boreal forest, and desert soils (Fierer et al. 2012). When comparing the bacterial richness to that of the other microorganisms in the forest soils, we obtained estimated approximately 875 fungal OTUs, 37 archaeal OTUs, and 91 double stranded viral OTUs.

Fungal estimates of diversity in OTUs from shotgun metagenomic analyses are substantially higher than bacterial diversity despite a much lower amount of total DNA present in the soil. Our finding of approximately 875 fungal OTUs in the forest soil samples agreed with other studies regarding the total fungal diversity in soils (Buée et al., 2011).

We were limited in the use of shotgun metagenomic technology by not being able to directly sequence RNA from the soil in the form of RNA viruses or single-stranded DNA viruses. Thus we were not able to completely measure the diversity of viruses in the soil. We assume that this means there is greater viral diversity in the soil than we were able to detect.

Of the five plots there was no individual plot with significant influence of tree host or geochemical influence on diversity of microorganisms. No plot had a statistically significantly more OTUs, in regards to bacteria or archaea (Figure 3.4). Bacterial and archaeal richness in regards to OTUs was very similar across all the forest soils. The eukaryote soil microorganisms were more spatially patchy, so some particular OTUs
were not found in one plot versus another, but all the plots contained a similar OTU level, suggesting replacement of one taxon with another in the microbial community structure. Archaea diversity observed was similar to that documented previously by (Bates et al., 2011).

There have been a number of proposed mechanisms that regulate species diversity in forest soils that are extremely rich in diversity. In forest soils there is heterogeneity in soil properties that could account for drastic differences in micro-ecology related to small soil surfaces, or porous surfaces. Soil bacteria, like all bacteria, have rapid rates of speciation and low rates of extinction. The seasonality of forest components allow for rapid recycling of carbon sources which can keep soil microorganisms in constant turnover cycles. High levels of immigration may occur in these forest sites, especially low land sites (close to stream which may flood; Hughes-Martiny et al 2006)

3.4.4 Alpha & Beta Diversity

Alpha diversity, known as the evenness or richness of individual species or OTUs contained with a community, was similar across the five forest soils that we examined. The observed species diversity or richness of the bacterial communities ranged from 622 to 705 OTUs per sample. Archaea (average of 41 OTUs per plot) and fungal (average of 967 OTUs per plot) species also appeared to be as diverse, if not more diverse, than the bacteria, when samples were normalized to identical sequencing depth.
Only slight differences in microbial composition were distinguishable between the five forest soil communities based on the whole genome shotgun sequencing data (Figure 3.4). Principle component analysis of the samples showed location and tree host species being the two most influential factors on microbial composition. Both location and host species could be affecting micronutrient variability and pH (Rousk et al 2010). Moisture conditions were approximately identical, and temperatures did not differ much among locations. No discernible differences in organic carbon inputs or outputs were observed (Fierer 2011).

3.4.5 Whole Genome Shotgun Metagenomic Analysis

While there are limitations to shotgun metagenomic analysis of forest soils, there are many benefits to using this type of analysis. For some experiments shallow sequencing of a few samples maybe just as useful and more cost-effective than deeper sequencing of the same samples (Knight et al 2012). One problem with this approach is that it may not allow us to fully capture the entire genomic diversity that is contained within a forest soil sample. Greater depths of sequencing could possibly capture the rare yet potentially important organisms that could be contributing important genes or biochemical or geochemical processes. Any further studies on these forest plots should consider deeper sequencing of soil DNAs to better determine the relative abundance differences.

Very few reads from shotgun metagenomic sequencing of soil DNAs can be annotated, and this study is no different, as only approximately 32% of the reads could be annotated. This is not so much limitation of the shotgun metagenome sequencing technology as it is
a reflection of the paucity of soil taxa and soil bacterial sequence data that is available in databases. A concerted effort is needed to sequence the genomes of organisms living in forest soils, possibly through a combination of in vitro culturing and in situ metatranscriptomics. We are probably missing or inappropriately annotating genes that may be important for key biogeochemical processes in soils. This is not just a problem for forest soil metagenomics but for any sequencing technology using shotgun metagenomic sequencing that addresses the diversity of life in natural environments (Delmont et al 2011)

These sites were only examined at a single time point, as it was not our goal during this study to determine the temporal variability at each site. Other soil DNA research has shown temporal variability to be far less than spatial variability (Fierer & Jackson 2006). The role of plant host as a mediator of microbial diversity is an understudied area of microbial ecology, along with the spatial distribution of microbial diversity. New technologies can help address these questions. Sequencing technology and data analysis tools will continue to improve our ability to measure microbial diversity in soils, and our ability to understand the importance of this diversity.

3.5 Conclusions

This study investigated the host-association and spatial distribution of soil microbial taxa in forest soils from South Korea. The forest soils show predictably similar taxonomic structure to those previously described for bacterial taxa in forest soils at sites in other countries and continents (Fierer et al. 2012). While the distribution of soil bacteria in
these forest plots was not unexpected, the presence of eukaryotes in the soils was surprisingly patchy and unevenly distributed. While this may be an artifact of soil sampling, these observations raise the possibility that soil bacterial taxa are generalists that are widely and evenly distributed across forest ecosystems, while eukaryotes such as fungal taxa are more site-specific and perhaps associated with plants at the site. There is a paucity of information regarding eukaryotic diversity in soils, however forest soils are known to be diverse in composition, as are eukaryotes in soils (Delmont et al. 2012). Thus there is a compelling need to further characterize the diversity and range of eukaryotes in soils, particularly fungal taxa, and their roles in forest ecosystems.

3.6 Acknowledgements

We would like to thank Claude dePamphilis for critical reading of an early draft of this manuscript. This work was supported by the World Class University Project R31-2009-000-20025-0 from the Ministry of Education, Science and Technology of Korea, with additional support from the Schatz Center for Tree Genetics at Pennsylvania State University.

3.7 References


Bae, J.S., Joo, R.W. and Kim, Y.S. (2012). Forest transition in South Korea: Reality, path
and drivers. Land Use Policy. 29:198–207.


Figure 3.1 – Map showing sampling locations of the five forest plots that were analyzed in this study and their relative elevation and map coordinates.
Figure 3.2 – Rarefaction curves for each soil sample from 5 forested sites.
Figure 3.3 – Phylogenetic tree of phyla of microorganisms encountered in forest plots.

Bacteria were generally found in all the plots. Eukaryotes, on the other hand, were very patchy and very few were found dispersed across all of the forest plots.
Figure 3.4 – Cluster diagram of the five forest plots clustered on their similarity and shown as a heat map. Each phyla frequency was normalized between 1.00 for the most commonly found phyla and 0.00 for the least common phyla and each group was colored and associated on a scale of zero to one. Both host and geography influenced clustering of the taxa in each respective sample.
Figure 3.5 – Table of micronutrients from the five Korean forest soils. All concentrations are shown in parts per billion (except for Fe, shown in parts per trillion). Measurements were taken ten times for each soil sample within each plot and then the subsequent average measurement was used for that soil sample. The standard deviation in parentheses shows the standard deviation of all the soil samples from a single plot.

<table>
<thead>
<tr>
<th>Plot</th>
<th>Li</th>
<th>B</th>
<th>Na</th>
<th>Mg</th>
<th>Si</th>
<th>P</th>
<th>K</th>
<th>Ca</th>
<th>V</th>
<th>Cr</th>
<th>Mn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.90</td>
<td>3.37</td>
<td>92.79</td>
<td>158.63</td>
<td>051.31</td>
<td>04.48</td>
<td>253.24</td>
<td>767.50</td>
<td>1.20</td>
<td>4.56</td>
<td>47.62</td>
</tr>
<tr>
<td></td>
<td>(±3.82)</td>
<td>(±2.37)</td>
<td>(±2081)</td>
<td>(±78950)</td>
<td>(±55436)</td>
<td>(±8794)</td>
<td>(±78079)</td>
<td>(±38223)</td>
<td>(±1774)</td>
<td>(±400)</td>
<td>(±7091)</td>
</tr>
<tr>
<td>2</td>
<td>5.70</td>
<td>1.36</td>
<td>88.96</td>
<td>254.69</td>
<td>41.13</td>
<td>57.74</td>
<td>083.29</td>
<td>909.82</td>
<td>8.00</td>
<td>3.77</td>
<td>01.77</td>
</tr>
<tr>
<td></td>
<td>(±206)</td>
<td>(±234)</td>
<td>(±957)</td>
<td>(±60169)</td>
<td>(±5570)</td>
<td>(±2263)</td>
<td>(±21253)</td>
<td>(±30112)</td>
<td>(±1292)</td>
<td>(±139)</td>
<td>(±1191)</td>
</tr>
<tr>
<td>3</td>
<td>6.25</td>
<td>1.99</td>
<td>95.58</td>
<td>496.83</td>
<td>49.73</td>
<td>71.64</td>
<td>273.75</td>
<td>937.90</td>
<td>4.18</td>
<td>8.88</td>
<td>121.69</td>
</tr>
<tr>
<td></td>
<td>(±218)</td>
<td>(±097)</td>
<td>(±330)</td>
<td>(±80194)</td>
<td>(±2925)</td>
<td>(±6132)</td>
<td>(±23395)</td>
<td>(±36206)</td>
<td>(±658)</td>
<td>(±288)</td>
<td>(±3716)</td>
</tr>
<tr>
<td>4</td>
<td>3.65</td>
<td>5.24</td>
<td>48.48</td>
<td>4965.44</td>
<td>72.81</td>
<td>070.49</td>
<td>401.79</td>
<td>836.31</td>
<td>76.63</td>
<td>5.22</td>
<td>169.96</td>
</tr>
<tr>
<td></td>
<td>(±119)</td>
<td>(±906)</td>
<td>(±1412)</td>
<td>(±232450)</td>
<td>(±20665)</td>
<td>(±38473)</td>
<td>(±83146)</td>
<td>(±306050)</td>
<td>(±6972)</td>
<td>(±4124)</td>
<td>(±5072)</td>
</tr>
<tr>
<td>5</td>
<td>6.23</td>
<td>4.05</td>
<td>53.89</td>
<td>2079.73</td>
<td>241.84</td>
<td>32.74</td>
<td>411.58</td>
<td>329.75</td>
<td>06.48</td>
<td>2.85</td>
<td>19.90</td>
</tr>
<tr>
<td></td>
<td>(±398)</td>
<td>(±126)</td>
<td>(±720)</td>
<td>(±183277)</td>
<td>(±88303)</td>
<td>(±12303)</td>
<td>(±29493)</td>
<td>(±85927)</td>
<td>(±1283)</td>
<td>(±1530)</td>
<td>(±9838)</td>
</tr>
</tbody>
</table>
Figure 3.5 (continued) – Table of micronutrients from the five Korean forest soils. All concentrations are shown in parts per billion (except for Fe, shown in parts per trillion). Measurements were taken ten times for each soil sample within each plot and then the subsequent average measurement was used for that soil sample. The standard deviation in parentheses shows the standard deviation of all the soil samples from a single plot.

<table>
<thead>
<tr>
<th>Plot</th>
<th>Fe (ppt)</th>
<th>Co</th>
<th>Ni</th>
<th>Cu</th>
<th>Zn</th>
<th>As</th>
<th>Se</th>
<th>Mo</th>
<th>Cd</th>
<th>Cs</th>
<th>Pb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.13</td>
<td>.51</td>
<td>0.14</td>
<td>.46</td>
<td>3.51</td>
<td>5.62</td>
<td>.50</td>
<td>.32</td>
<td>.13</td>
<td>.17</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td>(+5.00)</td>
<td>(+1.45)</td>
<td>(+2.47)</td>
<td>(+1.67)</td>
<td>(+13.05)</td>
<td>(+15.37)</td>
<td>(+0.19)</td>
<td>(+0.07)</td>
<td>(+0.04)</td>
<td>(+0.33)</td>
<td>(+5.02)</td>
</tr>
<tr>
<td>2</td>
<td>1.36</td>
<td>.83</td>
<td>1.18</td>
<td>.48</td>
<td>5.77</td>
<td>8.50</td>
<td>.48</td>
<td>.47</td>
<td>.11</td>
<td>.17</td>
<td>7.50</td>
</tr>
<tr>
<td></td>
<td>(+4.34)</td>
<td>(+0.87)</td>
<td>(+1.02)</td>
<td>(+0.87)</td>
<td>(+6.09)</td>
<td>(+10.05)</td>
<td>(+0.13)</td>
<td>(+0.05)</td>
<td>(+0.01)</td>
<td>(+0.14)</td>
<td>(+1.92)</td>
</tr>
<tr>
<td>3</td>
<td>4.44</td>
<td>2.07</td>
<td>6.38</td>
<td>0.47</td>
<td>0.82</td>
<td>1.39</td>
<td>.48</td>
<td>.43</td>
<td>.09</td>
<td>.67</td>
<td>4.25</td>
</tr>
<tr>
<td></td>
<td>(+3.51)</td>
<td>(+4.22)</td>
<td>(+2.68)</td>
<td>(+1.52)</td>
<td>(+9.96)</td>
<td>(+16.65)</td>
<td>(+0.18)</td>
<td>(+0.14)</td>
<td>(+0.03)</td>
<td>(+0.21)</td>
<td>(+14.88)</td>
</tr>
<tr>
<td>4</td>
<td>3.78</td>
<td>8.95</td>
<td>0.78</td>
<td>1.77</td>
<td>62.25</td>
<td>15.98</td>
<td>.24</td>
<td>.91</td>
<td>.17</td>
<td>.02</td>
<td>4.75</td>
</tr>
<tr>
<td></td>
<td>(+3.48)</td>
<td>(+13.63)</td>
<td>(+18.03)</td>
<td>(+13.67)</td>
<td>(+70.25)</td>
<td>(+76.21)</td>
<td>(+0.81)</td>
<td>(+0.36)</td>
<td>(+0.09)</td>
<td>(+0.39)</td>
<td>(+48.36)</td>
</tr>
<tr>
<td>5</td>
<td>8.22</td>
<td>6.29</td>
<td>9.34</td>
<td>5.07</td>
<td>6.66</td>
<td>3.13</td>
<td>.77</td>
<td>.43</td>
<td>.21</td>
<td>.91</td>
<td>7.36</td>
</tr>
<tr>
<td></td>
<td>(+6.77)</td>
<td>(+2.88)</td>
<td>(+3.76)</td>
<td>(+3.00)</td>
<td>(+15.76)</td>
<td>(+23.68)</td>
<td>(+0.38)</td>
<td>(+0.07)</td>
<td>(+0.05)</td>
<td>(+0.17)</td>
<td>(+3.34)</td>
</tr>
</tbody>
</table>
Figure 3.6 – Principle component analysis of the three most influential factors on sample partitioning. Plant host explains the majority (98.7%) of partitioning of individual samples from Korean forest soils. Host may explain many other factors such as soil chemistry and nutrients.
Figure 3.7 – Phylogenetic tree showing the major components of the fungal biota in the soil plots based on BLAST hits to sequence databases.
Figure 3.8 – Pearson rank heat map of clustering of microbial co-occurrence in each of the forest plots.
Figure 3.9 – Spearman rank heat map of clustering of microbial co-occurrence in each of the forest plots.
Figure 3.10 – Kendall rank heat map of clustering of microbial co-occurrence in each of the forest plots.
Figure 3.11 – Overall read hits for each forest plot shown in raw sequence reads (top plot) and log scale (bottom plot).
Figure 3.12 – Relative presence of the most frequent microbial groups before (left) and after (right) normalization.
Figure 3.13 – Cluster analysis of sequence reads identified with GLIMMER. The overall trend of the sequence identification is in agreement of both the BLAST and BLAT sequence identification.
Chapter 4

Traditional Breeding, Genomics-Assisted Breeding, and Biotechnological Modification of Short Rotation Woody Crops

Joshua R. Herr & John E. Carlson, Penn State

Keywords: Short Rotation Woody Crops (SRWCs), Bioenergy, Biofuels, Biomass, Lignin modification, Populus, Salix


4.1 Introduction

Short Rotation Woody Crops (SRWCs) are a valuable renewable resource with the unique ability to provide both combustion heat and liquid biofuels. With a long history as a feedstock for the pulp and paper industry, SRWCs are now being touted as a component of future plant-based bioenergy allocations by world governments. Fast-growing trees with high productivity, such as Poplar (Populus species and hybrids) and Willow (Salix species and hybrids) have been deemed the most productive SRWCs in temperate climates, while Pines (Pinus species) and Eucalyptus (both species and hybrids) are perhaps most advantageous in seasonally warmer areas (Neale, 2007; Neale and Kremer, 2011). There are many benefits of SRWCs as opposed to other bioenergy crops: flexible year-round harvesting, low-input multiple-growing season growth, reduced storage costs, and reduced storage degradation losses are a few of these benefits (Sims and Venturi, 2004). Both native and exotic tree plantation species can contribute to energy feedstock, as managed, unmanaged, or native forests. A production-based economy has been developed in plantation management of trees, as opposed to natural forests. Although natural forests are still harvested, the use of plantations are gaining
more acceptance through the automation of production, and they may help to maintain biodiversity by reducing the amount of natural forests harvested, by mitigating climate change, and by providing a wider array of commercial value-added chemicals and products (Harfouche et al., 2011). Additionally, forest plantations can help to subvert climate change by replacing fossil fuels with wood bioenergy crops and preventing changes in land-use, such as that seen in the conversion of forests to lands not capable of sequestering carbon (Neale and Kremer, 2011).

In the future, increasing demand for both high quality sources of bioenergy and forest products will create large global productivity targets for renewable fiber production (Herr, 2011). Current targets for SRWCs, such as Poplar and Willow, are 8 to 10 dry tons per acre per year (Cobb et al., 2008). Eastern Cottonwood (Populus deltoides) produces an annual average of 5 to 8 dry tons per acre (Simmons et al., 2008). Early research in short rotation Willow produced site-specific yields in the range of 3 to 7 dry tons per acre per year (Mead, 2005). A 20-year rotation cycle of Loblolly Pine (Pinus taeda) produced a yearly harvest of 4 dry tons per acre (Simmons et al., 2008). Current dry weight production levels for SRWCs are not sufficient to meet future demand for plant based bioenergy – as a result, improvements to SRWC feedstocks are needed.

In order to avoid competition with food crops, SRWCs are typically grown on marginal lands that may not support high productivity crop growth. As a result, it may be most imperative that SRWCs are bred for efficient water and nutrient acquisition to be productive on marginal sites. Additionally, marginal sites are typically associated with increased plant stress, in the form of both abiotic and biotic stresses. Breeding resources
can be developed to address susceptibility to frequently encountered stresses. With the added benefit of optimization for additional end uses (such as chemical co-production, paper and pulp production, climate change mitigation efforts, strip mine reclamation, etc.), genetic improvement of SRWCs have the potential to increase growth and biomass yields, adaptation to marginal lands and environments, and resistance to pests and diseases (Mizrachi et al., 2012). Reduction of inputs such as irrigation, fertilizers and pesticides can be accomplished through breeding as well as, in concert with improved downstream resources for the bioenergy industry, such as enzymes used to digest plant material, catalysts for improved saccharification and improved thermal and chemical pretreatments (Mansfield, 2009).

Applied plant breeding has been successful with forest trees, including SRWCs mainly for the pulp and paper industry, to improve growth, tree volume, and various wood quality traits (Harfouche et al., 2012). There are numerous traits recognized for maximized growth from plantation grown trees. Phenotypes with the ability to grow quickly at close spacing, those showing resistance to various pests, and those able to be vegetatively propagated are the cultivars with the most value (Cobb et al., 2008). Possible breeding pairs are typically selected on projected yields, wood quality, and traits associated with environmental stress tolerance, such as drought or herbivore-tolerant hybrids and cultivars (Cobb et al., 2008).

### 4.2 Breeding Strategies for end use specific energy conversion for SRWCs

Yield may be increased by breeding in a variety of ways. Breeding can be used to
develop SRWCs that both optimize growth in specific locations and are optimized for the maximization of energy gained from the thermal, chemical, or biological (enzymatic) breakdown of woody biomass.

Productivity may be improved by developing SRWCs that optimally grow in a wide array of growing conditions, most probably in marginal lands not suitable for higher commodity crops. These breeding initiatives will have to address both the biological and genetic considerations of the type of land, which includes soil conditions such as microbial and nutrient levels, as well as the overall climate and rainfall of the planting geography.

While breeders are addressing geographic considerations to SRWC growth, improvements in woody quality and energy should be a major priority. Breeding focusing on improving the lignocellulosic feedstock or cell wall density will be the most effective. General recalcitrance of woody tissues to thermal, chemical and biological breakdown is a breeding priority.

**4.2.1 Breeding for maximization of Cellulosic Ethanol**

In recent years, emphasis has been placed on the development of SRWCs for the production of cellulosic ethanol. Woody tissue, predominantly cellulose, is dense with polysaccharides that may be fermented into ethanol. While woody tissue can potentially produce a large amount of ethanol relative to other bioenergy crop plants, the largest hurdle is the general recalcitrance of woody tissue.
Cell wall polysaccharides, the largest component of insoluble compounds in plant cells, contribute the most to this recalcitrance. Secondary cell walls of both angiosperm and conifer wood consist of approximately 45% cellulose, 20% hemicellulose, and 15% lignins. The composition of hemicellulose varies the widest in hardwood SRWCs (typified by Poplar) which consist of 80% xylans, 10% mannans, with the rest accounted for by galactans and arabinogalactans (Mansfield, 2009). Softwood SRWCs (such as Pines) consist of approximately 50% mannans, 30% xylans, with the remaining 20% made up of galactans and arabinogalactans (Harfouche et al., 2012). The compositions of cell walls is variable in species and varieties and represent a possible breeding target to attain the optimal composition for a desired energy output.

There have been numerous strategies for the modification of woody tissue to reduce recalcitrance and to increase cellulosic ethanol yields, and of which, two have emerged as the most common. First, SRWC plant breeders have tried to make polysaccharides more amenable to breakdown (either industrially or through microbial enzyme or a combination of both). A second strategy has been to reduce the presence of recalcitrant compounds that inhibit the fermentation of sugars after the plant tissue has been broken down (Mansfield, 2009).

4.2.2 Modifying the Lignin Content of SRWC Biomass

Manipulating the lignin content in woody biomass has long been a research interest by plant breeders. Lignin interferes with the chemical breakdown of woody tissue for the pulp and paper industry and reduces the digestion of herbaceous forage plants by farm
animals. Now, researchers interested in the development of SRWCs for bioenergy have also been motivated to reduce the effect of lignin in the breakdown of cellulosic materials. As a result, many of the genes coding for enzymes in the lignin pathway have been elucidated and have been down-regulated in experiments with forest trees as well as herbaceous plants such as *Arabidopsis* (Anterola and Lewis, 2002; Barakat et al., 2011; Boerjan et al., 2003; X., Li et al., 2008; Vanholme et al., 2012; Weng et al., 2008).

Two approaches to make SRWCs more efficient in biomass conversion to liquid fuels has been to reduce the overall lignin content of the woody tissue or to modify the existing lignin content so that it is easier to chemically remove or break down the cellulosic material. Syringyl-(S-)lignin, which is higher in oxygen content than guaiacyl-(G-) lignin, is generally believed to be easier to be chemically removed (Mansfield, 2009), although some have reported that lignin composition has little effect on the ease of cell-wall breakdown as opposed to lignin content (Wegrzyn et al., 2010). When genes early in the pathway have been manipulated in expression, differences in the composition of S-to G-lignin have been detected in cell walls. In the case of down-regulation of COMT or of ferulate 5-hydroxylase (F5H), the amount of S-lignin was greatly reduced, while over-expression of F5H led to increased S-lignin (Vanholme et al., 2012).

In *Populus*, numerous pleiotropic effects have been recognized during the down-regulation of lignin pathway genes. These effects include the reduced removal of sugars from leaves and the subsequent reduction in the overall photosynthetic rate when the gene encoding *p*-coumaroyl shikimate 3’-hydroxylase was down-regulated (Coleman et
A reduction in the amount of hemicellulose and an increase in vessel collapse and cavitation have been observed when phenylpropanoid acid pathway genes have been down-regulated (Leplé et al., 2007). The reduction of pleiotropic effects associated with lignin reduction in plants can probably be alleviated through genomic breeding. This will probably lead to optimized regulation of promoters and their target genes (Mizrachi et al., 2012) or by targeting pathway regulation hubs that can affect suites of genes involved in lignin biosynthesis through the phenylpropanoid pathway or pathways that provide precursors compounds to lignin biosynthesis (Mansfield, 2009).

4.2.3 Breeding for maximization of BTUs

Although producing cellulosic ethanol is the main thrust of research for SRWC breeding, the use of woody crops as a combustible fuel is a secondary market for growers and already implemented in numerous markets in North America (Ellis, 2012). When cellulose is burned it liberates a calorific value of approximately 8,000 BTUs per pound of wood, while lignin liberates approximately 11,000 BTUs per pound (Demirbas, 2001). High lignin wood can be burned in co-gasification furnaces with coal and other fuels, or it can be directly burned on its own, for the production of heat, steam, or electricity generation.

Probably the best strategy for increasing the thermal energy produced by wood is to increase both the overall lignin content and increase the cellular density of wood tissue (Demirbas, 2001). Plant breeding for increased lignin content and increased cellular density can probably be accomplished by directly applying the opposite of breeding for
maximized cellulosic ethanol. While there are perhaps limits to cell wall density, SRWCs can be bred to increase both cell wall density and to reduce the per cell lumen volume, which typically results in the thickening of the xylem. It appears that lignin synthesis in plant cell walls and cell wall thickening through increased xylem formation may be genetically associated. Numerous studies (Patzlaff et al., 2003; Goicoechea et al., 2005) have reported that several MYB transcription factors may be responsible for regulating lignin biosynthesis by binding to promoters for lignin biosynthesis genes. These same MYB transcription factors have been implicated in thickening of the xylem through increased deposition in the S2 layer of woody plant cell walls.

4.3 Genomics in SRWC Improvement

Unlike many agricultural crops with short generation times, traditional breeding methods for the genetic improvement of trees is hindered by long generation times, possibly spanning multiple decades (Harfouche et al., 2012).

The use of genomics has revolutionized biology and plant breeding has not been immune to these improvements. New technologies for genome sequencing and rapidly identifying genes in breeding populations allow researchers to capitalize on naturally occurring genetic variation. To take full advantage of these technologies, SRWC breeders must be able to model associations of genotype-phenotype relationships. This should be done with particular focus on phenotype traits associated with abiotic and biotic stress resistance, as the incidence of stress is expected to be high on marginal lands and as a consequence of future climate predictions (Ellis, 2012).
4.3.1 Marker-assisted selection

Marker-Assisted Selection (MAS), the use of a genetic marker assigned to the region of a genome for the indirect selection of a breeding trait of interest, is a valuable resource for SRWC breeding (Grattapaglia and M., D., V., Resende, 2011). Markers used in MAS may be morphological, biochemical, genetic, or based on many different types of genetic polymorphisms. Relying on quantitative trait loci (QTL) analysis, typically MAS uses genetic polymorphisms that have been previously established. The major benefit of MAS is an overall decrease the time involved in plant breeding. Some complaints of MAS are that when a single species or variants are being investigated, QTL analysis provides a small allele range or that there is weak linkage disequilibrium (LD) to work with within a breeding population (M., F., R., Resende et al., 2012). Moving beyond QTL analysis, more plant breeders have relied on association genetics or, when available, means investigating whole-genome data to detect genetic polymorphisms (Neale and Kremer, 2011). Forest trees have been notorious for having high heterozygosity (Neale and Ingvarsson, 2008), weak LD (Lexer and Stölting, 2012), and a large amount of rare alleles found at low frequencies (Eckert et al., 2009), which have all precluded breeders from identifying phenotype associated polymorphisms or large effect (Quesada et al., 2010).

4.3.2 The future is here - selection at the whole genome level

Perhaps the most valuable asset to tree breeders from genomics has been genome-wide association studies (GWAS) (Grattapaglia and M., D., V., Resende, 2011). Unlike MAS,
GWAS addresses large populations and take both frequent and rare alleles into consideration (M., F., R., Resende et al., 2012). The overarching goal of studies incorporating GWAS into breeding programs is to provide polymorphic information of value to plant breeders. For GWAS studies to be valuable, data must be available for both rare and commonly found alleles, and must represent population sizes large enough to determine these alleles. Using GWAS, all allele variants in a breeding population can be determined, in theory. This is more optimally accomplished for forest trees with moderate genome sizes such as *Populus*, at 450 Mb, as opposed to large genomes such as conifer, at over 10 Gb, for which a genome sequence has yet to be completed (M., F., R., Resende et al., 2012).

An alternative to MAS has been the development of Genomic Selection (GS) that incorporates a predictive model using both genotypic and phenotypic data (Grattapaglia and M., D., V., Resende, 2011). GS models are optimized to use all available genetic markers to predict the genomic landscape of future progeny. One model, the genomic breeding value, is a calculation developed by using the number of alleles at any given locus and estimating their breeding effect through linear regression (M., F., R., Resende et al., 2012). Optimal for SRWCs, GS is arguably the best method for traits expressed later in organismal life cycles, with low heritability, or those not easily attainable through lower-throughput methods. The impact of GS in agronomic annual crop plants will be potentially smaller due to already existing advanced breeding strategies implemented over months; it is expected that GS will have the greatest impact for long-lived plants with delayed breeding cycles, typical of forest trees which can take decades for breeding cycles to come to fruition (M., F., R., Resende et al., 2012). The impact of GS will
remain strong as long as there are loci that have yet to be fully characterized (Grattapaglia and M., D., V., Resende, 2011).

Recent breeding experiments using GS for forest trees have shown that predictive models can be developed, even with a low number of markers, for small effective breeding size populations that are comparable to traditional sized populations (Grattapaglia et al., 2009). With larger sizes of breeding populations and an increase in the LD within breeding populations, a higher marker density is required, which is provided by GS (M., F., R., Resende et al., 2012). The role of GS in breeding SRWCs, especially those with long breeding cycles, are just now seeing gains in the reduction of breeding time (Harfouche et al., 2012).

The largest impact of GS in the breeding of SRWCs may be in the reduction of breeding costs through the acceleration of superior phenotypes. It is typical for SRWC breeders to generate many crosses from numerous parents during multi-year experiments and GS provides ideal allele combinations that may reduce reliance on large breeding populations (Harfouche et al., 2012). A secondary benefit of the use of GS will be the reduction in breeding times associated with determining the best germplasm to use in coordinated crosses. Through both the reduction of breeding cost and reduced time associated with crosses, GS is poised to be the most widely used breeding strategy.

4.3.3 Next Generation Sequencing in tree breeding

The development of next-generation sequencing (NGS) methods (Shendure and Ji, 2008), resulting in greater throughput sequence production at lower cost, has drastically
improved our ability to assess genomic variation for breeding efforts. Most importantly, NGS has provided an ability to generate entire genome sequences for organisms much faster than previous sequencing technologies (Hill, 2012). With new sequencing techniques, allele variations can be directly identified in large breeding populations or, in the extreme case, even entire species (Tuskan et al., 2006). NGS techniques are used most frequently for two types of resequencing: for entire genomes of individuals in a species that has a reference genome, such as a *Populus* breeding population (Cohen et al., 2010), or to resequence a suite of genes or regions of interest from large breeding populations to detect variants. There are numerous types of variants that may be determined from resequencing, including single nucleotide polymorphisms (SNPs) or insertion and deletion events, to larger variants such as transposable elements (Harfouche et al., 2012) or rearrangements and structural variants (SVs; Villar et al., 2011). Trait variation, as identified by GWAS, can only typically identify a small fraction of the actual variation in individuals. The use of resequencing approaches are is beneficial in that rare functional variants can also be identified (Elshire et al., 2011).

Next generation sequencing technologies are expected to change rapidly in the future such that they will become routine in breeding applications with forest trees. This has already becoming the case in *Populus*, where microarrays and SNP genotyping chips have been replaced by methods for genotyping-by-sequencing (GBS; Elshire et al., 2011; Davey et al., 2011). New resequencing technologies are most applicable in species with sequenced genomes, but one advantage of GBS technologies is that the initial investment of genome sequencing or SNP chip development are not mandatory pre-
requisites for genotyping in forest trees. Forest trees are especially prone to weak LD and the identification of a large number of low cost markers from GBS are a resolution to a lack of genetic markers (Elshire et al., 2011).

4.3.4 Integration of molecular data in breeding technologies

Information derived from new sequencing technologies in the form of genomic and transcriptomic data will progressively be integrated with new technologies in the areas of proteomics and metabolomics in SRWC breeding initiatives (Cohen et al., 2010; Hamanishi et al., 2010). Integrating data collected from breeding populations developed by GS or allele variants from GWAS studies, holistic information from the integration of many forms information from the genetic to cell chemistry will increase productivity and quality of cultivars developed by SRWC plant breeders. The fields of proteomics and metabolomics are still in the developmental stage for forest trees (Nilsson et al., 2010), but there is much promise for SRWC breeding initiatives.

Transcriptomics, which include methods such as microarrays and RNA-Seq, are comprehensive ways to measure genome-wide expression coverage and change (Baginsky et al., 2010; Qiu et al., 2011). RNA-Seq in particular is contributing to the understanding of transcript profiling at more highly resolved levels than microarrays (Wang and Brummer, 2012), through the increasing availability of NGS methods. Additionally, transcriptomics can be used to understand epigenetic regulation of gene expression in forest trees and how yield may be affected by specific environments. Forest breeders can implement transcriptional markers to predict heterosis in mating pairs
as has been accomplished in *Arabidopsis* (Lister *et al.*, 2008).

Metabolic markers have been more highly correlated with phenotype than transcript levels (Boerjan, 2005). As a result, monitoring metabolites may be a more rapid way of advancing phenotype selection for SRWC breeders than following gene expression. Metabolic markers have been developed in *Arabidopsis* where correlations between metabolomic markers and biomass accumulation were observed (Meyer *et al.*, 2012). Proteomics is a key technological development in understanding the role of proteins in the accumulation of biomass or environmental stress resistance. Proteomic methods can yield information regarding structural conformation and variation from post-translational modification or quantitative traits of proteins such as stability and tissue specific protein accumulation (Gstaiger and Aebersold, 2009). Proteomics benefits SRWC breeding because it can transcend only measurements of mRNA levels to predictions of gene products and post-translational modification events (Baginsky *et al.*, 2010).

Integrating transcriptomics, metabolomics, and proteomics into SRWC breeding programs will complete the already vibrant functional genomics toolbox for the development of superbly performing cultivars. This will enable a systems biology approach in plant improvement that will be crucial to the understanding of regulatory mechanisms for complex traits such as environmental stress resistance and characterization of hormone pathways (Neale and Ingvarsson, 2008). Integration of these genomic tools will enhance our knowledge of the complexity of the genes, proteins, and metabolites involved in networks governing phenotypic traits (Baginsky *et al.*, 2010).
The end result of the integration of transcriptomics, metabolomics, and proteomics with GBS and GWAS is the goal of exploiting natural biological and genetic variation to expand the tree improvement toolbox available to SRWC breeders.

4.4 Biotechnology in SRWC Improvement – The use of Genetic Engineering

4.4.1 What is genetic engineering?

Genetic engineering (GE) can also be used to complement conventional breeding programs for SRWC species. Incorporating GE with existing traditional and genomics-based breeding programs allows breeders to take advantage of natural genetic variation and introduce genes for traits which may not be available within the species, or be difficult to introduce via other breeding methods (Busov et al., 2005). The important SRWC targets of high yields digestible woody biomass are increasingly feasible using GE methods (Grattapaglia et al., 2009; Mizrachi et al., 2012).

Numerous strategies can be implemented using GE. One approach involves directly adding or modifying genes responsible for increased plant growth, improved stress tolerance or environmental adaptability, or increased secondary metabolites or byproducts (Grattapaglia et al., 2009). Another approach can be to express transcription factors to up-regulate expression of target genes to accomplish a specific trait or suite of traits (Mizrachi et al., 2012). The combination of these strategies can be used in tandem with many different genes, known as trait stacking.
Genetic engineering can support plant breeding in numerous ways. It is particularly useful in woody plants that have long juvenile periods before breeding can be conducted (Harfouche et al., 2011). Unwanted phenotypic variation is often observed in breeding populations resulting from segregation and recombination, and traits have been known to breakdown in elite breeding genotypes after years of cultivation. These issues could be addressed through GE by the use of dominant and highly expressed genes (Harfouche et al., 2011). Complex traits, such as disease resistance or nutrient use efficiency, which have eluded traditional breeders of SRWCs, may soon be addressed with multi-genic GE techniques.

4.4.2 Implementation of Genetic Engineering in SRWC breeding

Genetic engineering will feature prominently when improvement of SRWCs is needed for traits that are not observed in a breeding population or the species as a whole, or when a lack of natural variation in a trait makes it particularly difficult to use GS for selection. One example of when GE will be most useful is the addition of a single gene – for which there is no analogue in the breeding plant’s respective genome – such as genes to confer resistance to a new pest or disease (Quesada et al., 2010). If there is no genetic variation for a trait of interest, GE may be used to introduce variation from other species. Lastly, GE methods may be able to produce early flowering in some SRWC plants allowing for faster breeding cycles (Flachowsky et al., 2009). In contrast, it is more likely that GE methods will be used to reduce, delay, or completely knock-out flowering in cases where containment of gene flow from GE plants to other populations of SRWCs would be an issue.
4.4.2.1 Cell Wall Modification through Genetic Engineering Means

Perhaps the most excitement in the use of GE for SRWCs is the modification of cell wall composition. Modifying the components of the cell wall may increase the amount of energy attained from SRWC material or may reduce the amount of recalcitrance associated with conversion of plant tissues into energy or materials (Campbell et al., 2003). Studies incorporating GE have influenced cell wall development by modification of a β-1,4-endoglucanase (CEL1) responsible for enhanced growth in hybrid *Populus* (Shani et al., 2006). Additional studies have shown that using GE to add and express a cytosolic glutamine synthetase (GS1) in hybrid *Populus* derived from *Pinus* increased cell wall biomass accumulation, most notably in leaf tissue (Bernard and Habash, 2009). The expression of the transgenic GS1 was verified with increased enzyme activity along with an overall increase in amino acid concentration (Man et al., 2005) and drought tolerance (El-Khatib et al., 2004). Numerous studies have used GE methods to modify the structure and composition (G versus S)-lignin in SRWC plant cell walls (Fu et al., 2003; Boerjan, 2005).

4.4.2.2 Stress Resistance Traits Enhanced Through Genetic Engineering Means

Genes involved in leaf size and structure, stem development, timing of bud flush and leaf senescence all may influence biomass accumulation, rotation time and growth rate. It is also possible that genes that confer stress tolerance or delay senescence could also improve growth under environmental stress conditions or extend the growing season, and it has been shown that suppression of a gene, deoxyhypusine synthase (DHS) which is
part of the stress-response pathway, increases vegetative and reproductive growth in the model plant *Arabidopsis* (Duguay et al 2007).

For example, transgenesis has led to the development of cold-tolerant hybrid eucalyptus trees (Eucalyptus grandis X Eucalyptus urophylla) which are awaiting field trials [44]. Introduced genes being tested in *Populus* include the *Populus tremula* and *Arabidopsis* stable protein 1 (SP-1) gene, as well as genes involved in metabolic processes responsive to drought, redox proteins, transporter proteins, signal transduction proteins, and transcription factors (Polle et al. 2006).

Another area for improvement is the engineering of belowground traits. Roots are the key to a second ‘Green Revolution’ in agriculture [48], but the same holds true for establishing SRWC trees with higher productivity. Although efforts to boost root growth by conventional breeding are underway in some species, GE may prove to be more efficient. For example, discovery of the WUSCHEL-related homeobox gene (WOX11), which is involved in the activation of crown root emergence and growth [49], and UPBEAT1, which is a key transcription factor that controls the transition between cell proliferation and differentiation in Arabidopsis roots [50], suggests that it may be possible to design root systems tailored to certain environments. Thus, it may be possible to reduce the time needed for root systems to become established on recalcitrant forest tree species, therefore reducing the time to harvest. By simultaneously engineering several genes involved in root growth, GE can support research to obtain a more rapid understanding of complex, multigenic traits, such as adventitious rooting.
4.4.2.3 Development of Mutant lines through Genetic Engineering

As a complement to transgenesis, forward-genetic approaches such as cisgenesis and intragenesis, implemented as activation tagging, are used to develop mutant lines and identify genes controlling specific traits in SRWCs (Ahuja, 2011). Several biotechnology programs have the goal of developing thousands of activation-tagged mutant lines for *Populus* in this fashion.

Activation-tagged genes have potential use for the further development of transgenic and cisgenic GE SRWC plants. Additionally, activation tagging may elucidate novel targets for improved SRWC breeding. The majority of GE work on activation tagging has been conducted in maize, Arabidopsis and more recently poplar, with numerous genes showing promise for breeding using reverse-genetic approaches. Most notably, the gene LATERAL ORGAN BOUNDARIES DOMAIN (*LBD*) showed positive regulation of phloem formation during secondary growth (Yordanov *et al.*, 2010). In another study, a gene responsible for coding gibberellin catabolism (GA 2-oxidase) was found to regulate tree height (Busov *et al.*, 2003).

Another method for the development of mutant lines that has potential use with SRWCs is the implementation of Zinc finger nucleases (ZFNs). This strategy uses ZFN enzymes to target a specific locus by creating a double stranded DNA break that can include both targeted insertions or point mutations (Durai *et al.*, 2005). The major benefit of using ZFNs is avoiding the randomness found with traditional transgene insertion methods.
The use of ZFNs has yet to be fully realized in SRWCs, but the potential is there for the enhancement of breeding and genomic selection.

4.4.2.4 The use of small RNA interference for gene silencing

Perhaps the most widely used method for targeted gene silencing is the use of small non-transcribed RNA fragments identified as RNA interference (RNAi). These methods have been implemented with both direct introduction of designed RNA fragments or microRNAs (miRNAs); both methods take advantage of a plants innate small interfering RNA (siRNA) processing pathway (B., Li et al., 2011). The most notable use of RNAi with SRWCs has been the suppression of genes in *Populus* including the STYISH1 (STY1) and SHORT INTER-NODES (SHI) genes, which showed increased shoot and root growth and increased the amount of xylem tissue (Zawaski et al., 2011). The insertion of miR156 class miRNA has been effective in down-regulating the SQUAMOSA PROMOTER BINDING PROTEIN (SBP) modulating the growth of lateral branches and meristems (Chuck et al., 2007). Further studies with *Populus euphratica* have yielded a suite of miRNAs associated with long-term drought stress that could be implemented in plant breeding initiatives (B., Li et al., 2011).

4.4.3 Regulatory Requirements and Associated Risk Assessment for Biotech Trees

The use of GE strategies may prove to be essential for the development of economically feasible implementation of biofuels from SRWCs and may provide environmental, social, as well as economic benefits (Ahuja, 2011). The use of GE in SRWC breeding will allow
novel traits to be introduced in a timely fashion not capable with traditional breeding techniques.

Numerous concerns must be addressed before GE modified SRWCs may be implemented, however. Perhaps the most important concern is the dispersal of GE pollen, seeds, or other vegetative propagules (Strauss et al., 2004). SRWCs are generally wind pollinated and it is possible that GE pollen can travel long distances and pollinate receptive native species. Although traits being bred into bioenergy SRWCs – such as improved yield, wood composition, or abiotic and biotic stress resistance – should be considered low risk, the overall goal or requirement for crop release could be to eliminate all possibility of escape of GE modified genes. Many SRWCs are clonally propagated and do not reach flowering before they are harvested, but as a precaution, GE plants for of SRWCs may be produced in genetic backgrounds that are pollen sterile, such a triploids (Busov et al., 2005). This can been implemented through the addition of sterility knock-out genes which prevent the formation of pollen without affecting the general vigor of the plant (Strauss et al., 2004; Ahuja, 2011).

The use of GE in breeding of SRWCs for bioenergy purposes will require regulatory guidance and approval prior to commercialization. While field tests of GE modified trees have been implemented in numerous countries, the United States has no GE trees in field production (Harfouche et al., 2012). The only GE SRWC trees grown industrially are insect resistant poplar genotypes in China (Ahuja, 2011), while freeze-tolerant
Eucalyptus lines await (Harfouche et al., 2011), and currently the only non-regulated GE trees in the United States are papaya and plum.

Regulations for SRWC plants bred using GE technology should meet clearly defined simple standards agreed upon by global contributors. Studies considering the safety and implementation of GE plants, as well as a thorough cost-benefit analysis should be a priority for SRWCs (Strauss et al., 2004). There is a growing consensus among the tree improvement community that guidelines for all plants developed by genomic or biotechnological means should address traits and not the method by which superbly performing cultivars were developed (Harfouche et al., 2011).

4.5 Conclusion

It is clear that in order to meet future renewable energy standards, we will have to have clearly designated plant-based bioenergy production benchmarks consisting of a variety of bioenergy crops, of which SRWCs must play a significant role. Poised to make minor contributions at current levels of production, in order to substantially provide future energy benchmarks, SRWCs will have to be bred for higher productivity and greater resistance to abiotic and biotic stresses found associated with marginal lands.

The use of SRWCs is diverse as they may provide energy in a wide range of uses such as cellulosic ethanol as a liquid fuel or through combustion to generate electric power. Additionally, unlike other bioenergy plants, SRWCs may be used as raw materials in the pulp and paper industry, as well as building materials, and in carbon sequestration for
climate change abatement. Purpose-bred SRWCs grown in high-rotation plantations will alleviate pressure to harvest native forests and preserve biodiversity. In order to meet all these demands, SRWCs may be bred to optimally meet all of these criteria.

The integration of genomics, biotechnology, and SRWC breeding has a great potential to improve the yield and sustainability of bioenergy production. Dedicated SRWC breeding initiatives can make large contributions to climate change abatement and economies, both at local and global scales. For the ultimate economic and environmental impacts to be achieved, the development of SRWC breeding must be supported by government and private funding, meet regulatory acceptance and be welcomed by the public.

4.6 References


Martin, F. et al. (2008). The genome of Laccaria bicolor provides insights into


Figure 4.1 – Graph showing the number of published papers over the last 20 years that include both a SRWC species name (Eucalyptus, Populus, Pinus, and Salix) and the words “plant breeding” in the title, abstract, or keywords.
Figure 4.2. Graphical representation of the use of genomics implemented in both genomic-based and biotechnology-based SRWC breeding approaches.
Chapter 5

Comparative genomics of forest trees: A fast-track for understanding diverse stress responses

Joshua R Herr\textsuperscript{1}, John E. Carlson\textsuperscript{1,2}

\textsuperscript{1}The Schatz Center for Tree Genetics, Department of Ecosystem Science & Management, The Pennsylvania State University, University Park, PA 16801 USA

\textsuperscript{2}Department of Bioenergy Science & Technology, Chonnam National University, Gwangju, 500-757 South Korea


5.1 Introduction

Natural forests and timberlands form extremely complex biological systems, covering almost 40 million km\textsuperscript{2} of a diverse array of physical sites and ecosystems. Forests provide many economic benefits, some of which include long-term carbon sequestration, stabilization of riparian areas and watersheds, habitat and sustenance for wildlife, and fuels and building materials for human consumption, as well as cultural significance. The increasing incidence of plant stress, disease, and mortality in forest trees – in the form of introduced exotic pests, pathogenic diseases, and invasive plants, and compounded with climate change and forest fragmentation – threaten the sustainability of many forest ecosystems.
Understanding forest tree responses to stress is not a trivial task. Forest trees possess large and complex genomes with typically high levels of heterozygosity (Tuskan et al., 2006). Despite this complexity, forest trees are fundamentally no different from other perennial plants – they require light, water, nutrients from the soil, and sense abiotic and biotic stimuli and stresses from the environment – but opposed to herbaceous plants they must regulate central processes, pathways, and signals across much larger structures. At a local scale, trees respond to cumulative and synergistic stresses by intricately moderating gene expression and adjusting these responses seasonally through epigenetic means (Raj et al., 2011). Fluctuations in environmental conditions, such as drought or heat-stress, not only reduce growth and yield for trees, but also reduce the ability of trees to respond to other stresses. The difficulty of studying forest trees in the laboratory or glasshouse has hindered our ability to understand these stress responses.

Our primary knowledge of stress responses at the molecular level is based on controlled laboratory studies of single stresses on small model plant species, such as Arabidopsis or Medicago, and molecular responses to stress are extremely complex. Trees are subjected to a diverse and complex combination of stresses and consist of a greater physical area to respond to environmental signals, this requires tight coordination across a large physical area between multiple genomes: nuclear, plastid, and mitochondrial. Trees, a synthetic polyphyletic categorization, naturally show a wide range of tissue diversification – from pliable fine roots and leaves or needles to complex cell walls responsible for rigidity and height – which respond to stresses in differing ways. Adding to this complexity, genomic
responses to abiotic and biotic stresses are dynamic: plant hormone responses can be on the order of seconds, and may be reversible, while epigenomic changes may occur over many years and may largely be irreversible (Raj et al., 2011).

Recent advances in genome sequencing technology have considerably enhanced our abilities to understand the genomic framework and gene expression responses to environmental stresses. Network analysis and systems biology approaches in gene expression studies provide a holistic picture of complex physiological and cellular responses. Nevertheless, understanding the specific functions and roles of genes that respond to stresses remains a daunting task. Comparative genomics, in addition to gene network analysis of responses to stress, can advance our knowledge of stress responses in forest trees. Through these comparative approaches, we can develop a better understanding of the genotype-phenotype relationship within locally adapted susceptible and resistant stress response candidates (Neale and Kremer, 2011). Common garden studies of clonal trees (such as Poplar, Willow, and Eucalyptus) and well-characterized progeny from exclusion breeding from genotyped and phenotyped parents can also contribute valuable insights to knowledge gained from genomics.

5.2 Integrating a comparative genomics and systems biology approach to understanding forest health

Understanding stress responses in phylogenetically disparate forest trees can be accelerated by the development of both a comparative genomics and systems biology approach incorporating both genome sequencing efforts and – through the measurement
of transcriptomes, proteins, and metabolites – molecular surveys of stress responses. Sequencing of the *Populus* genome (Tuskan et al., 2006) provided a framework to understand gene responses in both an ecological and bioenergy context for woody plants, and the recent release of numerous draft tree genomes and ongoing genome sequencing projects will contribute greatly to comparative studies (Neale and Kremer, 2011).

Gene network responses to the environment involve interactions with many different regulatory pathways often centered around hormones (Fujita et al., 2006). Stress responses in forest trees – like other plants – begin at the local site affected by the stress and may vary on the basis of location. More specifically, tissue specific responses to insect herbivory on leaves vary when compared to nutrient stress in roots, but the overall product of both such interactions may ultimately be the closure of leaf stomata. The severity and length of different stress events may drastically affect the intricacy of each subsequent or concomitant reaction and this intricacy challenges our current knowledge of plant stress network responses (Fujita et al., 2006). Stress signaling responses are temporal and are dependent on cell, tissue, or organ histories (Dinneny et al., 2008), but these sensing events may have lasting repercussions. Disease and adverse environmental conditions may reduce growth and yield for the trees for many years. One example of the synergistic effects of stresses on forest trees is drought stress. For instance, drought stress can reduce a tree’s the response to insect herbivory or ozone toxicity (Raj et al., 2011). Water stress reduces conductance and the flow of signaling compounds into different plant tissues. Chronic water stress results in xylem embolisms that do not allow for movement.
Perhaps the fastest way to understand stress response and stress induced signal transduction in forest trees is to use comparative genomics methods. There are a number of large gene family expansions in forest trees, notably in *Populus* (Tuskan et al., 2006), and understanding how these gene expansions affect stress response is a daunting task.

Transcriptomics, proteomics, and metabolomics are at the forefront of the discovery of understanding environmental stresses in forest trees. These new techniques can facilitate useful comparisons between the effects of drought stress, saline soils, ozone fumigation, light stress, heat, or cold stress in trees. Such comparisons can lead to the identification of shared versus unique stress responses specific to each type of stress and how gene expression, protein modification and metabolites are correlated to each stress.

Comparative genomics studies are increasingly including observations of natural populations and how the genetic basis of diversity may effect forest health on different temporal and spatial levels (Whitham et al., 2012). Systems biology studies, genomic data analysis, and the development of new bioinformatics tools for large data sets will help forest tree breeder overcome delays resulting from long-life cycles (Neale and Kremer, 2011), more so than with of annual crop plants and model plants such as *Arabidopsis* and *Medicago*.

### 5.3 Hormone response

Plant hormones function as the regulators of abiotic and biotic stresses responses. There is significant cross talk between plant hormone pathways and this interaction can be both
gradual and instant. Biotic stresses, such as pathogen defense, are associated with 
saliclyc acid (SA), which modulates pathogen resistance to biotrophic pathogens and is 
associated with systemic acquired resistance (SAR, (Jones and Dangl, 2006). Jasmonic acid (JA), more typically associated with defense to necrotrophic pathogens, is 
responsible for induced systemic response (ISR).

Abiotic stresses, such as those regulated by abscisic acid (ABA) and ethylene (ET) 
(Wilkinson and Davies, 2010), may have a diverse array of initiating stressors. ABA is a 
principal controller of environmental stress responses in relation to water (Lee and Luan, 
2012). ABA, like JA and SA, is responsible for quick responses to abiotic stresses, such 
as the regulation of ion and water transport processes as well as slower regulation of 
general down-stream stress responses such as effects on growth and development (Lee 
and Luan, 2012). Abiotic stresses are known to initiate the rapid down regulation of 
energy metabolism and sugar movement. There is also typically a reduction in protein 
synthesis, which may represent a protective strategy rather than the investment in growth 
and reproduction.

5.4 Crosstalk between responses

Early responses to abiotic and biotic stresses may differ, but can result in similar stress 
responses in the plant. Early signals in stress response include reactive oxygen species 
(ROS), reactive nitrogen species (RNS) and oxylipin signatures derived from broken 
plant membranes. All of these constituents (ROS, RNS, and oxylipins), as well as others, 
may modify gene expression and protein synthesis and activity (Mittler et al., 2011).
Energy metabolism is drastically affected by oxidative stresses, and the resultant biochemical changes to sugar transport, lipid biosynthesis, photosynthesis and electron transport (Smeekens et al., 2010).

Carbon flow in the form of sugar is either directly or indirectly affected by stress pathways and modifications may occur in the sugar synthesis, metabolism, transport, or storage of sugars. Additionally, many tree species use photosynthate as currency with which to acquire phosphorus, nitrogen, and water from symbiotic organism, such as mycorrhizal fungi or nitrogen fixing bacteria. Plant growth and development is directly affected by stress-induced changes in sugar signaling. Sugars, such as the simple sugars glucose and fructose, are additionally osmoprotectant and have been recognized as mediating drought and temperature stress (Smeekens et al., 2010).

Although recognized as important in stress response pathways, there are few studies addressing the role of stress in the induction of phosphorylation cascades, such as mitogen-activated protein kinases (MAPKs), and other forms of post-translational modification (Yoo et al., 2009) in trees. There is known to be significant regulation of MAPKs via plant hormone and light-regulation pathways. In forest trees, there is a great need for holistic experiments that not only measure transcriptomes, but measure proteomes and metabolic data as well, to obtain a full picture of these post-translational modifications in response to stress. For example, in regards to ABA signaling, protein phosphorylation and dephosphorylation are important stress regulating factors, but we do not fully understand the intricate details of how this process works in small plants, such
as Arabidopsis, let alone large plants, such as forest trees, which have to regulate water balances across large living structures (Lee and Luan, 2012). We have a great deal to learn about the general principles in abiotic stress signaling across the structures of woody plants.

5.5 Understanding stress response through co-expression

Arabidopsis has been the model for understanding gene co-expression in plants, and recent studies have investigated the comparative genomics of stress response with Glycine max and, the first model tree, Populus trichocarpa (Weston et al., 2011). The Populus trichocarpa genome has more than 42,000 predicted gene models, some of which are known to exhibit environmentally regulated post-translational modifications (Tuskan et al., 2006). Existing co-expression analyses have revealed stress regulation networks, that are genetically-based, and that probably reflect phenotypic associations (Usadel et al., 2009).

Functional genomics Studies have addressed whole plant transcriptome changes in forest trees with various stresses, such as dehydration, cold, heat, high-salinity and ABA. Transcript abundance has been observed to be affected by a variety of known, as well as unknown, stress response genes. These stress responses can be exacerbated by drought, flooding, or other water stress conditions (Skirycz and Inzé, 2010).

Transcriptomics data, which is easier to collect using next-generation nucleotide sequencing techniques, has eclipsed proteomic data in sheer numbers. Proteomic data,
which is collected using techniques such as ELISA and mass spectrometric methods such as MALDI-TOF, is more expensive and more difficult to produce. By using proteomic methods, the phosphorylation or ubiquitination state of a protein may be determined, and this has implications in cell signaling and downstream modification of stress responses (Kaufmann et al., 2011).

There is low correlation between profiles of plant transcriptomes and the “downstream” proteomic and metabolomic signatures (Hey et al., 2010). Post-translational modification is a common way for transcripts to result in differing protein or metabolic signatures. Epigenetic regulation and methylation change gene expression as the result of environmental factors. For example, three different *Populus trichocarpa* clones grown in common garden studies from differing drought conditions, showed epigenetic changes in the form of DNA methylation across different time frames (Raj et al., 2011).

### 5.6 Conclusions

Plant biologists studying trees have a wealth of data from long-term provenance trials and progeny tests from many years of tree breeding to model phenotype-genotype and genotype by environment interactions. The production of sequence databases for expressed genes under different environmental scenarios, genetic markers and genetic linkage maps, and resources for reference populations and geographically diverse common garden studies for comparative purposes that are currently being developed will provide data for generating hypotheses of how the genetic networks are in forest trees are integrated. This will enable the forest tree research community to better characterize
genes for growth, adaptation, and environmental abiotic and biotic stress responses to drought, heat, insect herbivory, and fungal and bacterial pathogens.

There is a daunting amount of work needed to fully understand responses of forest trees to abiotic and biotic stresses. Hormone pathways and their interconnectedness are extremely complex and species specific, and even studying clonal trees grown in plantations for bioenergy or biomass, such as *Populus* or *Salix*, stress responses vary on a large number of factors, including the organ, tissue, cell type, developmental stage, as well as the type of stress or stresses affecting the plant and the intensity and duration of the stress. We need more information on how plants sense stress in their environments, and the downstream gene and protein action in regards to stress response, what are the alternative splicing events, and how this affects metabolites which may act as defense mechanisms.

Comparative genomics and systems biology (the “multiple omics”) approaches will help us understand molecular responses to stress. Time series experiments are required along with common garden experiments with holistic and detailed metadata for each location. These resources will be vital for understanding condition-dependent responses. Genome mapping of stress response at cellular, tissue, and organ levels are needed, especially for tree species, as well as the level of transcriptome and proteome. Additionally, the modeling of genetic traits to phenotypic traits and mapping genetic loci to loci explaining phenotypic diversity will be a key accomplishment. When QTL on genetic maps can be linked to new genomes, greater understanding of basis of stress responses will be gained.
There are many challenges to working with forest trees but the stakes are high and the potential rewards are great.

5.7 Acknowledgements

Research on the genomics of stress-response in the Carlson lab was supported by grants to JEC from the Ministry of Education, Science and Technology of Korea (World Class University Project R31-2009-000-20025-0) and from The National Science Foundation Plant Genome Research Program (Award #1025974). JRH is supported by The Schatz Center for Tree Genetics at Pennsylvania State University and a grant to JEC from the USDA NIFA NE Sun Grant Project NE 11-48.

5.8 References


Martin, F. et al. (2008). The genome of Laccaria bicolor provides insights into


Figure 5.1 – A systems approach to the study of forest tree responses to environmental stresses will require integrating comparative genomics data from transcriptomics, proteomics, and metabolomics, with molecular and landscape level modeling.
Chapter 6

Bioenergy From Trees

Joshua R. Herr*

The Schatz Center for Tree Genetics, The Interdepartmental Program in Plant Biology, and The Center for Lignocellulose Structure and Function, The Pennsylvania State University, University Park, Pennsylvania, 16802, USA

(*Author for correspondence: tel +01 814 865 4440; email: jherr@psu.edu)

Key words: biofuels, ethanol, cellulose, lignin, trees, Populus, climate change, life cycle analysis.

A modified version of a commentary that appeared in New Phytologist, Volume 192, Number 2, pages 313-315, 2011

6.1 Introduction

Any cursory look at current newspaper headlines reveal escalating food prices, increased demands for energy consumption, and – often ignoring the unrealized economic costs associated with elevated atmospheric carbon – the rising costs of fossil fuels. Recently, the crude oil spill in the Gulf of Mexico caused economic damage to surrounding ecosystems, and severe weather events – flooding, storms, high temperatures, and drought associated fires – have been linked to anthropomorphic climate change and infectious diseases (Shuman 2010). Fears derived from the meltdown of Japan’s Fukushima Power Plant have persuaded some governments to move away from nuclear
power. Acknowledging that current energy strategies are unsustainable and detrimental to global health, it is not surprising that the UK Parliament, European Union, U.S. Department of Energy, and Chinese Government have all set lofty benchmarks for greater reliance on plant based bioenergy by the year 2020 (Upham et al. 2011).

The desire to maximize bioethanol production from tree-based cellulosic materials was the central theme of the meeting. Some participants pointed out early in the discussion that solid biomass, such as wood pellets, would fare best as a replacement for fossil fuels currently used for thermodynamic heat and electricity production (Somerville et al. 2010). Other participants stressed that specific transportation industries – namely air travel and heavy-weight shipping – need the force and distance of combustion engines powered by volumetric and gravimetric energy dense fuels. These industries are also not amenable to carrying the heavy weight of electric batteries or waiting for long battery charging times (Ohlrogge et al. 2009). It is also important to note that, unlike electric recharging, the infrastructure of liquid refueling for small vehicles is well established. As a liquid fuel substitute for petroleum is needed in the immediate future, bioethanol from woody biomass, in particular, has potential to satisfy this need.

When considering both liquid and solid fuel production, lignocellulosic feedstocks from fast growing short rotation tree species arguably show the most promise. Trees provide high productivity on a per hectare basis, just behind the C₄ grasses Switchgrass
(Panicum) and Miscanthus (Schmer et al. 2008, Jørgensen 2011), and carefully managed, trees offer an opportunity for sustainable, and – unlike other bioenergy crops – non seasonally dependent harvesting. First generation bioethanol production from either sugar (sugarcane, sugar beet) or starch (corn, sorghum) based materials elevates the price of agricultural commodities subsequently driving up the cost of food (Karp & Richter 2011). Algal biodiesel shows promise, but photosynthetic bioreactors are costly and still years away from general production (Wijffels & Barbosa 2010, Stephens et al 2010). Despite hindrances to their utilization for bioethanol, lignocellulosic feedstocks have the potential to make a considerable contribution toward future energy benchmarks.

6.2 Using genomics to optimize tree biomass

The greatest emphasis on tree biomass has been placed on Poplar (Populus sp. hybrids), Willow (Salix sp.), Eucalyptus, and numerous temperate Pinus species. The promise of a model tree for plant genomics, population genetics, and cell biology came to the fruition with the release of the Poplar (Populus trichocarpa) genome (Tuskan et al. 2006). More recently, with the ongoing characterization of the Poplar pan-genome, efforts have focused on understanding genetic diversity across natural populations, recording detailed phenotypic databases, and mapping genotypic data associated with desirable bioenergy traits (Neale & Kremer 2011). By extending our basic knowledge of woody plant gene expression – from methylation to transcription factors – we hope to understand the genomic basis influencing plant growth and adaptation (Neale & Ingvarsson 2008). Beyond elucidating plant growth at the genomic level, meeting participants stressed that
the two main goals should be to understand the basic mechanisms behind cell wall formation and to reduce the level of chemical recalcitrance of lignocellulosic material in cell walls.

Advanced tree breeding approaches, such as genomic selection, attempt to develop individual hybrids optimized for biomass production while maintaining inherent genetic diversity for adaptation to a diverse array of abiotic and biotic stresses present in the field (Karp et al. 2011). Genome-wide markers are typically used to accomplish direct selection through traditional breeding techniques. Breeding efforts have focused on maximizing both biomass yield and plant health by selecting for traits such as increased carbon partitioning to woody tissues, optimized growth forms, improved hydraulic conductivity, increased pulp yield, and stress adaptability and disease resistance (Grattapaglia et al. 2009, Wegrzyn et al. 2010). Target genes for maximized bioethanol yield are being identified using various methods, such as quantitative trait locus mapping and genetical genomics (Grattapaglia & Krist 2008). Genetic modification techniques are used to promote or alter the expression of already existing genes linked to desirable traits or add genes via transgenic modification (Strauss et al. 2010).

6.3 Understanding cell walls to take them apart

A large portion of the meeting discussion was devoted to addressing how data provided by researchers working in the realm of genomics and molecular biology can provide tools for the elucidation and modification of plant cell walls. Lignocellulosic materials, the components of secondary cell walls, vary wildly in trees but typically consist of cellulose
(30 to 60%), hemicellulose (20 to 40%), and lignin (15 to 35%) (Weng & Chapple 2010). With an end goal of modifying this composition of plant cell walls, researchers have been identifying transcription factors underpinning the phenotypic and physicochemical attributes of cell walls, including wood chemistry and ultrastructure, optimum growth parameters, and tree physiology (Legay et al. 2010). Recent research on cell signaling (Nieminin et al. 2008) and sugar partitioning (Payyavula et al. 2011) during procambial development in cell wall formation has been conducted with the goal of elucidating cell wall crystallinity, wood porosity and density, and the size and complexity of cellulose fibers. Considerable efforts have been placed on understanding carbon allocation to growing cell walls, with the goal of maximizing traits associated with source and sink components, including cellulose content (Joshi & Mansfield 2007). There is an interest to understand key plant enzymes in sugar metabolism, such as fructokinase and sucrose synthase, which provide wood polymer precursors and energy for cell wall biosynthesis. It was noted that increased cell wall density is important if lignocellulose is broken down via thermochemical or gasification techniques. Understanding the accessibility and interactions of cell wall polysaccharides to microorganisms and their enzymes – particularly understanding the breakdown of both C5 and C6 sugars – was emphasized in some presentations (Rubin 2008). A discussion centered on what cell wall traits can be modified to reduce bioethanol conversion cost and time and increase yields.

The recalcitrance of lignin is perhaps the key scientific challenge for establishing highly efficient biofuels from woody biomass and numerous presenters focused on this topic. Goals for plant biologists and wood chemists include producing plants with less lignin
and making the remaining lignin more accessible to degrading enzymes (Vanholme et al. 2008). Recent studies down-regulating genes in the phenylpropanoid pathway have been important in understanding lignin biosynthesis (Studer et al. 2011). Moving to a systems biology approach, researchers hope to understand how molecular perturbations in the synthesis of monolignols, such as modifying the ratio of guaiacyl to syringyl ratio, might incur wide-ranging consequences on various metabolic processes (Voelker et al. 2011, Zhou et al. 2011). It is important to note that the combustion of lignin is highly energetic, and as a result, increased amounts of lignin could be a desirable trait for increased thermal energy from wood.

6.4 From field to fuel - saccharification and life-cycle analysis

A well-recognized benefit of woody biomass from trees is the ability to utilize marginal lands not suitable for food or other agricultural crops, however, sustainable cultivation presents a number of potential problems (Taylor 2008). Details of tree biomass accumulation for bioethanol yields in field settings are minimal and edaphic conditions found frequently on marginal sites, such as poor soil quality, can affect both abiotic and biotic plant stress, greater susceptibility to insect herbivory and disease, and increased lignin content, all of which translate into reduced biomass (Karp & Shield 2008). Trees may be the least nutrient intensive of the bioenergy crops, and symbiotic associations, such as mycorrhizae, may increase water use efficiency and improve phosphorus and nitrogen uptake (Luo & Polle 2009). Willow is conceivably leading the way in field studies (Karp et al 2011), evidenced by many oral and poster presentations underscoring
its value in bioenergy experiments. Field trials for tree species of interest are needed, but especially in relation to overall bioenergy yield.

The production of ethanol fuel from woody biomass is accomplished by the process of saccharification, or hydrolysis of starch and cellulose polymers to sugars, and fermentation of those sugars to ethanol. The process involves the harvesting, cleaning, and physical breaking of biomass material, pretreatment, and subsequent separation to liquid and solid components (Richard 2010). To improve saccharification, pretreatments are utilized and these handling steps most often involve high pressure steam with the addition of lactic or sulfuric acid or sulphur dioxide (Monavari et al 2011). While lignocellulose pre-treatment technologies are still being developed and optimized at the fuel production phase, chemical treatments to improve the yield of biomass feedstocks prior to harvest were suggested during the meeting discussion (Kumar et al. 2009). The fermentation phase is possibly optimized, but the majority of work needs to occur in the pretreatment phase of process.

With possibly the most to offer, life cycle analysis (LCA) has exhibited the most scrutiny over the efficiency of biofuel production. The meeting consensus agreed that LCA should critically address the feasibility of energetic, economic, and environmental sustainability of cellulosic ethanol. Potential benefits and tradeoffs of cellulosic bioethanol can be investigated comprehensively using process modeling, techno-economic analysis and attributional LCA (Cherubini & Strømman 2011). To make
lignocellulosic bioethanol economically feasible, at least in the immediate future, all components of the input material should be utilized: ethanol can first be produced from carbohydrates, and the remaining soluble organic compounds and solids can be used to produce electricity, biogas, secondary products and agricultural fertilizers (Murphy et al 2011).

6.5 How do we get there from here?

Perhaps a sign of a very productive meeting, many symposium participants left the discussion regarding the use of trees in future energy schemes with more questions than answers. Participants called for more collaborative science across multiple disciplines and agreed that all stakeholders could benefit from the development of more specific research questions.

The meeting consensus recognized gaps in our current knowledge and identified technology bottlenecks existing in three research areas. The first bottleneck is our knowledge in planta as we could benefit from biotechnological advances yielding improved genomic selection and rapid trait identification of bona fide growth characters. We need to stress further characterization of pathways for the biosynthesis of cellulose, hemicelluloses, and lignins. A second bottleneck is in our knowledge of how plants perform in the environment and how this translates into increased growth. Numerous researchers stressed our need for demonstration trials or common garden studies to determine natural population variability, develop surveys for novel trait breeding, and
make in situ measurements of plant performance. We do not have much data from bioenergy field trials regarding the long term consequences of coppice growth on marginal lands where nutrient depletion may occur over time in soil reserves. We can also benefit from understanding the yield reducing effects of abiotic and biotic stresses on trees. Numerous presenters emphasized the evaluation of overall feedstock quality and noted that even in short rotation Poplar plantations various stresses (such as wind and ice) can reduce yield, particularly trees with reduced lignin. The last bottleneck identified is in our knowledge and ability to move from field to fuel. To produce even a portion of the bioenergy earmarked in the next decade, a large increase in the scale of cellulosic material processing and bioreactor size is needed. Uniformity of feedstocks is a concern in downstream processes, as not all feedstocks are equal in quality or 100% uniform. Continued refinement of LCA models and assessments, by a diverse array of collaborating researchers, can make inroads to the most efficient production of bioenergy fuels. Just as the petroleum industry has developed secondary chemical markets, products based on woody biomass, such as those made from lignin, can be an economic supplement for bioenergy industries.

The greatest obstacles toward the utilization of woody tree biomass, and bioenergy in general, are policy hindrances, both at local and global scales. The consensus at the meeting was that researchers need to directly address policy changes, even have an ethical obligation to do so, to ensure that science is addressing societal needs (Buyx & Tait 2011). Noting the difficulty in initiating, changing or enhancing policy, some
suggested that we, as scientists, should develop synergistic interactions with government policy makers, at many different levels, from the start. Additional open public discourse on the science of tree transgenics is needed (Strauss et al. 2009). Field to fuel demonstration sites – many of which are underway – will show current technology and collaborative science in effect. Some researchers called for increased collaborations with industry, to work with engineers designing flexible-fuel vehicles or to collaborate with the paper production industry looking to offset increased recycling and reduced production from the electronic document revolution.

The production of renewable and environmentally sustainable energy is one of the principle goals, if not the most pressing goal, currently facing scientists. Despite numerous and considerable hurdles to accomplishing this goal, the research on lignocellulosic biofuels presented at this meeting suggests that tree derived woody biomass, in conjunction with other forms of renewable energy, can satisfy a portion of our energy needs. The New Phytologist Symposium - Bioenergy Trees was a large success because it opened up scientific discussion and dialogue, communicated cutting edge knowledge from advancing and veteran researchers in the realm of plant-based bioenergy, and outlined research directions and future goals so that we can move from tree pulp to gas pump.

6.6 Acknowledgements
I wish to thank the organizing committee of Francis Martin, Michele Morgante, Andrea Polle, Steve Strauss, Gail Taylor, and Jerry Tuskan for their efforts toward an interesting and productive meeting and Brian Ellis for his thoughtful post meeting discussion on the state of lignocellulosic bioenergy.

References


Zhou G, Taylor G, Polle A. 2011. FTIR-ATR-based prediction and modeling of lignin and energy contents reveals independent intra-specific variation of these traits in bioenergy poplars. Plant Methods 7:
Chapter 7

Summary and Broader Impacts

7.1 Introduction

The research conducted and presented in this dissertation has contributed knowledge to several aspects of the soil microbiome in forest ecosystems and interactions of Poplar, an ecologically and economically important woody biomass-producing tree, with its environment. Although the topics presented here may seem to be somewhat disparately related, I believe they have succeeded in providing novel information that contributes to addressing a very wide gap in our knowledge, and a platform for future research.

For many years I have been interested in the roles microorganisms – most notably mycorrhizal fungi – play in the interactions with forest trees. One of the areas I investigated in the past, during my undergraduate honor’s thesis and Master’s degree was the specificity of fungi and how they provided both nutrients and water to plants in exchange for tree photosynthate in the form of carbohydrates. Carbon and nutrients are transferred from fungus to plant, and vice versa, via fungal cellular penetration into apoplastic plant tissues, an infection process similar to plant-pathogenic fungal relationships (Harrison 2005). Despite an infection-like appearance, this relationship is beneficial to the plant, and appears to have strong evolutionary selection pressure because approximately 92% of all land plants are obligate mycorrhizal formers (Wang & Qui 2006). There have been questions for some time if this process provided other benefits to the plant, in addition to the added ability to acquire nutrients and water. Perhaps more importantly, these microorganisms may influence plant defense responses (Figure 1),
including "priming" and its role in Induced Systemic Resistance (ISR) (Goellner & Conrath 2008). While soil microflora can potentially have strong impacts on plant biology, their species and functional diversity is relatively unknown. Moreover, very little is known about how soil microbial diversity will be affected by environmental change, or how such diversity may affect plant responses to environmental change.

When I first entered the doctoral degree program in Plant Biology at Penn State in 2006, I was expecting to continue to investigate these interactions focusing on understanding how trees interacted with their microbes. At the time I planned to use “di-deoxy” Sanger sequencing – and my research did begin there, but a drastic change in technology over the last six years changed the route of my studies. Joining the Carlson Lab, I found that I could work with others who were interested in tree genetics and forest health. We benefited greatly by the advance of technology over the past six years, from the movement to microarrays designed on the sequenced genome of Poplar (Tuskan et al 2006), to the measure of gene expression through RNA-Seq methods, all within a few years’ time. The ability to sequence the genome of a tree, and now get a “fairly” complete picture of all the organisms – many of which we can’t even name yet – in soils is changing the way we observe plant-microbe interactions. My contributions here, in the dissertation, and elsewhere not represented in this dissertation, have been in this area.

### 7.2 Characterization of CYP74 proteins

When I first joined the Carlson lab I found a common ground with other researchers in the lab already investigating the role that stress play in forest trees. We typically used
poplar in laboratory experiments because of its vast array of genomics resources. There was already some research investigating the fact that soil fungi may be able to enhance plant growth and survival by not only providing nutrients and water, but by enhancing the oxylipin pathway to allow interactions of one (or multiple) compatible fungi, but exclude others. When I first became interested in this area it was hypothesized that this regulation step took place early in the Jasmonic Acid (JA) pathway, so we began to dissect the pathway and hypothesized that the first step, regulated by CYP74 cytochrome P450 oxidases would be the modifying hub that would be hijacked by mutualistic and pathogenic fungi. I was actually shocked that such an important hub in this plant hormone pathway was not studied in great detail, nor even a complete phylogeny addressed in the literature. That was our rationale for focusing on the phylogenetics and evolution of this group. Just last year, two groups, one in Francis Martin’s group and the other in Natalie Requena’s group, published that fungi hijack a step later in the JA pathway (Plett et al 2011, Kloppholz et al 2011).

It was during this time that I devised an experiment to study the effects of soil fungi, ozone, and insect herbivory on the gene expression of hybrid *Populus* trees in the greenhouse and laboratory. We used clonal trees from the same parent material to reduce any genotypic effects. We had a sequenced genome for *Populus* (Tuskan et al., 2006) and we acquired the mycorrhizal fungus Laccaria bicolor from Francis Martin’s lab, which in conjunction with the *Populus* sequencing project, also recently had its genome sequenced (Martin et al., 2008). We established plants in the greenhouse with the mycorrhizal inoculum and fumigated these plants with ozone and placed gypsy moth
(Lymantria dispar) caterpillars on them to induce insect damage responses. When I measured differences in the effect of fungal inoculum on the Populus roots, I found more than one group of fungi in our sterilized soil (Figure 7.2). This was not expected and we soon determined that the fungi were most likely emerging from within the surface sterilized plant tissue as endophytic fungi. I realized it would be quite impossible to understand the “priming” effect of Laccaria bicolor in the context of the presence of other fungi that may have been initiating a induced systemic response. This observation, along with an opportunity to study for a few months in the lab of Francis Martin, along with the development of new sequencing technologies, made me interested in observing the diversity of microorganisms in and around a plant.

### 7.3 Soil Microbial Diversity & Sequence Identification

The next two chapters of my dissertation address the identification of soil microbes using new next-generation sequencing techniques. I was able to conduct two studies in the lab of Francis Martin using 454 pyrosequencing that are not presented here, but these two studies allowed me to prepare a whole shotgun soil genome sequencing experiment conducted with forest soil samples collected in South Korea whilst visiting with John Carlson.

In this study we identify the microbial components of various forest soils and attempt to observe the role that host tree plays in mediating microbial community structure. The most novel observation from this study is that for bacteria and archaea, we found that all the groups were represented across all forest soil samples, suggesting that there was little
effect on the prokaryotic community from host plant. This agreed with other studies that recently addressed this phenomenon (Fierer et al. 2012a; Fierer et al., 2012b).

Interestingly we observed partitioning, or “patchiness”, with the microbial eukaryotes, particularly the fungi. Some microbial eukaryotes were found in some specific forest soils, others were found only elsewhere. This suggests some kind of plant mediation in microbial community structure for the eukaryotes. I have also collected soil samples from a number of additional plantations and forest sites populated by monocultures of other forest tree species for future metagenome analyses.

A separate review paper here was invited for the inclusion in a special issue on fungal diversity in environments, so I prepared a chapter on some of the techniques to cluster and identify sequences from the environment, both whole genome shotgun data and marker based amplicons. This clustering can be done using phylogenetics or nearest neighbor clustering, and for the accurate identification of fungal taxa, the use of a curated database is the most helpful.

7.4 **Short-Rotation Woody Crop Biomass Breeding & Plantation Experiment**

The last few chapters are focused on the role of *Populus* in the production of biomass for either cellulosic ethanol or combustible heat. These chapters go with the other studies and provide a more field-oriented side of the experiments related to improving yield from bioenergy plantations.
The first of these chapters is an invited book chapter which addresses the role that breeding plays in the improvement of yield and stress responses in woody biomass crops, notably *Populus*. The next chapter is an invited commentary which was published in the journal *New Phytologist* in slightly modified version (Herr, 2011). The next chapter is another short invited commentary/mini-review on the development of a holistic strategy to understand forest health through the union of genomics, transcriptomics, proteomics, and metabolomics. These tools will provide comparative mechanisms to make strides in understanding forest health and comparative tree biology.

The last chapter, presented as an appendix, is somewhat of a progress report in an ongoing large-scale project to provide an environmental framework for woody biomass on marginal lands. We have collected thousands of data points on numerous clonal hybrid *Populus* trees in order to describe the effects of environmental conditions, such as soil microorganism diversity, nutrient availability, stress responses, and the effects of tree spacing and an ethylene-blocking agent, 1-methyl-cyclopropene, on the production of lignin in woody tissues and drought response. The effect of the soil microbial community above ground phenotypes observed in these field trials will also be addressed through metagenomic analysis of soil samples that were collected during the study.

### 7.5 Future Directions

I see my research moving in directions that I have addressed here in this dissertation. I plan to continue to address plant associated microbial diversity and how this variability affects plant health. This is a very exciting time, as new plant and microbial genomes are
being completed every day. I have tried to develop my bioinformatics skills in the past few years to begin to address the large amounts of data that are now present and that which will continue to grow.

7.6 References


Figure 7.1 – Synopsis of the possible role of soil fungi in mediating a induced systemic response (ISR) in plants.
Figure 7.2 – Image of a root emerged from surface sterilized plant cutting and autoclaved soil that shows multiple fungi growing on a single root. This observation confounded the experiment and causing me to rethink my research direction that led me to trying to address soil microbial diversity using next-generation sequencing techniques.
Appendix 1

The correlation with leaf area, leaf damage, and an ethylene-blocking agent on the growth of a hybrid Populus grown for biomass

Joshua R Herr\textsuperscript{1}, Tyler K. Wagner\textsuperscript{1}, John E. Carlson\textsuperscript{1,2}

\textsuperscript{1}The Schatz Center for Tree Genetics, Department of Ecosystem Science & Management, The Pennsylvania State University, University Park, PA 16801 USA
\textsuperscript{2}Department of Plant Bioenergy, Chonnam National University, Gwangju, South Korea

Keywords: Populus, biomass, leaf area, insect herbivory, folial disease, cellulosic ethanol

A.1 Abstract

A number of woody crops grown at an agricultural scale have the possibility to provide large biomass yields, which can then be used for fuel, including conversion to ethanol. In order to not compete with high commodity crops, woody biomass is typically grown on edaphic marginal crop lands. These edaphic conditions reduce yield and plants grown on these lands typically show a higher rate of stresses associated with these conditions which can manifest themselves not just in lower yields but a higher incidence of insect herbivory, fungal disease infections, and air pollution, such as ozone, toxicity. Stress responses, such as an increase in plant lignin, increases the recalcitrance associated with saccharification of woody plant tissues. In order to address these concerns in an experimental fashion, we established a Populus biomass plantation with various spacing and treatment sub-plots. One treatment was tree spacing, planted at one, three, and five foot spaced rows, and another treatment was application of a ethylene blocking agent reported to reduce lignin accumulation and enhance drought response. In this study we
observed the first and second year growth characteristics with the incidence of insect herbivory damage in these treated sub-plots.

### A.2 Introduction

One benefit of sustainable biomass production from woody cellulosic materials (trees and grasses) is the ability to use agriculturally unproductive marginal lands unsuitable for food or other high commodity crops. The use of such lands presents the problem of reduced plant health and biomass yield due to poor edaphic soil conditions. These poor growing conditions may reduce both plant yield and the ability to survive multiple years of growth – one of the benefits of short rotation woody crops (SRWCs). Plants that are under stress from edaphic conditions may be at an increased susceptibility to other stresses, namely insect herbivory and fungal infections affecting plant leaves. This reduction in leaf tissue and photosynthetic efficiency in turn reduces plant yield which then can cause a downward spiral of stress that may be detrimental to SRWCs biomass production and bioenergy yield.

Some of the most efficient SRWCs are fast growing tree species, such as Poplar (*Populus* species and hybrids), which may be coppiced, allowing a continuous regime of low-input harvesting, regeneration, and subsequent re-harvesting. Although these are great benefits to the use of *Populus* in a short rotation crop cycle, there are some disadvantages. One such disadvantage is the recalcitrance of woody tissue from SRWCs. This recalcitrance can be beneficial, as when biomass is required to be stored until needed, or it may be detrimental, such as when the woody tissue needs to be broken down in the
saccharification process to make various fuel products, such as cellulosic ethanol. The recalcitrance of woody tissue is largely regulated by the amount of lignin present in the plant cell walls; high lignin content is known to reduce herbivory and pathogen infection, thus allowing a sustained biomass yield, but high lignin content may also reduce the amount of energy gained from these woody tissues. Therefore, there is a trade-off between lignin content and energy yields.

Although numerous lignocellulose pre-treatment technologies are being developed for liquid fuel production (Kumar et al 2009), opportunities to improve the suitability and yield of biomass feedstocks prior to harvest exist and should be investigated. Pre-harvest treatments with the ethylene blocking agent 1-methylcyclopropene (1-MCP), may sustain yields in the presence of environmental stresses while lowering total lignin content, which can interfere with conversion of biomass tissues to fuels. The 1-MCP treatment is applied to leaf tissue and is absorbed through the leaf stomata, but it is not known if 1-MCP could preferentially reduce lignin in leaf tissue before woody tissue in the stems and branches. Reduced lignin in leaf tissue may allow for the synergistic susceptibility of insect herbivory and fungal infections on leaves already stressed from growth on edaphic marginal lands. Higher disease incidence was observed in bamboo following 1-MCP treatments (Luo et al. 2007), so insect and disease damage estimates will be monitored closely in our study. In addition, ozone damage, visible as leaf stippling, is a common environmental stress in the northeastern United States which may affect biomass yields, and which is known to be compounded by other stresses.
We address here the effects of the application of 1-MCP on the growth of a fast growing hybrid Poplar developed for bioenergy production and the effects of this application on herbivore damage in the first growing year. We also address how leaf area/photosynthetic area and reduction from herbivory may reduce future growth of these woody crop trees. The overall goal of this study is to understand productivity of *Populus* plantations for woody biomass at the system level and to provide recommendations for farmers to reduce the effects of marginal lands while sustainably maximize woody biomass yields for ethanol production.

### A.3 Materials & Methods

#### A.3.1 Experimental Plot Design.

A complete randomized block design plantation (40.700897°N, -77.962107°E) consisting of a *Populus trichocarpa* X *P. deltoides* hybrid was planted in replicated blocks with rows, spaced at 1 meter apart, of various plant spacing of one, three, and five feet separating the planted trees, across 15 sub-plots. Each subplot consisted of six 60-foot rows with the two outside rows serving as buffer rows to reduce treatment effect from other sub-plots. The source of cuttings were clonally propagated trees from identical parent material acquired from Greenwood Resources (Portland, Oregon, USA). After a one-month storage period at 4°C after plant shipment, the plantlet cuttings – which were all approximately 20 cm long and 1 cm in diameter – were planted by hand with no hormonal treatments for root induction directly into the research plots following a spring rain in May 2009. After 1 month, trees that failed to produce viable shoots, overall less than 1% of the plantlets, were replaced with new cuttings. Plants were surveyed every
week for mortality over the first three months of plant growth and were replaced when mortality was observed. The plots were located at the Penn State University experimental farms at Rock Springs, PA, in an unmaintained fallow, hill-side farm field area that been cleared approximately 10 years earlier from the adjacent forest had grown a cover of both native and invasive plants. Prior to planting, the overall plot areas were mowed with a commercial mower and a were treated with Roundup® herbicide two weeks prior to planting. Competing vegetation was then controlled throughout the test by mowing between rows as needed.

### A.3.2 Treatment with 1-MCP.

1-methyl-cyclopropene (1-MCP) is an ethylene action inhibitor used in agriculture to enhance post-harvest shelf life of produce and protect crop yields from the effects of heat and drought stress (Choi & Huber 2009). It is applied as an aerial spray and is known to be absorbed through plant stomata. In addition to these known uses, 1-MCP has been previously shown to inhibit plant respiration and ethylene production, impede lignin accumulation through the repressing of key enzymes associated with lignification, and increase total sugar content in bamboo (Luo et al 2007). A volatile liquid emulsion of 1-MCP was applied to half of the research plots with a commercial sprayer on June 19th and August 4th of the first years growing season (2009); spraying dates coincided with periods predicted to not be affected by rainfall and in anticipation of water stress conditions.

### A.3.3 Field Data Collection and Analysis.
Growth characteristics were gathered for trees planted in inner rows in all plots over the course of three growing seasons. Outer rows were maintained as buffer rows to reduce the effect of plot blocking against other plot treatments. Characteristics that were measured in the field included: plant survival, plant height, shoot diameter, number of leaves per plant (sub-sampled), total leaf area per plant (sub-sampled), Maximum leaf area per plant, minimum leaf area per plant, estimated amount of Leaf Area Removed (LAR) from insect herbivore damage and fungal infection (sub-sampled), and total harvestable biomass accumulation (sub-sampled). Physiological measurements such as photosynthesis and conductance rates, and leaf gas exchange measurements were taken using a LiCor 6400 open-gas exchange photosynthesis system to assess the health of plants during environmental stress conditions, as well as + and - 1-MCP treatments. All of these measurements were repeated during and at the end of the 2010 and 2011 growing seasons.

A.3.4 The characterization of biomass composition and saccharification tests.

At the end of the third growing season, in October 2011, all of the stems and branches were harvested, sectioned and each tree individually packaged, after drying. Biomass composition analysis was conducted by colleagues in the Bioenergy Center at Chonnam National University in South Korea, for stem samples from each plot. Total lignin content was determined by the acetyl bromide method and lignin monomer composition by thioacidolysis, as well as fiber content in the cell wall (by neutral detergent fiber (NDF) and acid detergent fiber assays (ADF)) (Fukushima & Hatfield, 2004). Quantitative “real-time” PCR (RT-PCR) was used to confirm stochiometric relationships
between 1-MCP treatments, physiological responses, and biomass yields. Cellulose was quantified (by subtracting lignin content from ADF), hemicellulose (by subtracting ADF from NDF), and total phenolics contents in the cell walls (Duguid et al., 2007). The cell wall polysaccharides released by this method were hydrolyzed to their constituent monomers and hydrolyzed monosaccharides were derivatized to their alditol acetates and quantified by GS-MS. As a surrogate for ethanol conversion efficiencies, saccharification tests were conducted from 30 samples of *Populus* stems, two selected from each of the control and 1-MCP treated plots using a reduced severity pretreatment method followed by saccharification (Duguid et al., 2007).

### A.4 Results & Discussion

We measured a large number of characteristics across 15 sub-plots according to treatment of tree spacing and 1-MCP application. With regards to leaf characteristics, we measured leaf area and assessed insect herbivory and fungal damage for 93 trees chosen from all sub-plot treatment types for a total of 6494 leaf measurements. Not surprisingly, we found a strong correlation between the total leaf area per tree and the overall height of each tree during the first year of growth (Figure A.1).

There was also a correlation in growth from the first year to the second year, which was also not surprising, because one would expect that trees which grew well in the first year of growth after planting would continue that trend in future years. The effect of tree spacing was not profound, from first year’s growth to the second, but there were differences in the trend of growth between the spacing of trees in each sub-plot. One
might expect that trees planted close to each other might be in competition for both above-ground resources, such as light, and below-ground resources, such as nutrients and water. There might also be benefits to competition for nearest neighbors such that trees in close proximity would compete with each other to gain advantage in growth. We saw a slight effect in both first year and second year growth for the more densely planted sub-plots (Figure A.2). However, unexpectedly, the trees with the greatest space between them (five feet) showed the least amount of growth in both year one and year two.

There was a slight effect in the first year of growth on the amount of leaf area removed by insect herbivory and fungal damage (Figure A.3). It is interesting to note that trees with the largest gap between them showed the most leaf damage. This could be explained in numerous ways: chewing insects, such as caterpillars, usually do not travel long distances and once in place would not move to other trees. In plots where branches or leaves would never move, chewing insects would not move far and would concentrate their activity on one tree which may have skewed the data in the plot that was spaced at five feet distance, versus the plot with trees spaced a one foot apart, which showed a more consistent level of insect damage. Additionally, the trees with more space in between them might be more susceptible to wind to allow fungal spores to penetrate leaf tissue or ozone to damage leaves. The closer packing of trees might allow less damage due to a reduction in accessibility.

The effect of insect herbivory or fungal and/or ozone leaf damage in the first year had little lasting effect in the subsequent years growth. The growth differences in the
treatments were not as evident as in the first year (Figure A.4) and each spacing treatment showed less effect from leaf area removed than the prior first year of planting.

A.5 References


Choi ST & DJ Huber (2009) Differential sorption of 1-methylcyclopropene (1-MCP) to fruit and vegetable tissues, storage and cell wall polysaccharides, oils, and lignins. *Postharvest Biology and Technology* 52: 62-70


Figure A.1 – For all *Populus* trees in the experiment, regardless of treatment, there was a correlation with the number of leaves per tree and the overall tree height in the first year of growth.
Figure A.2 – For the sub-plot treatments of tree row spacing (one, three, and five foot spacing regimes), there was a slight correlation with first year growth and the growth in the subsequent growing season.
Figure A.3 – For the sub-plot treatments of tree row spacing (one, three, and five foot spacing regimes), there was an effect in the first year growth (height) and the amount of leaf area that was removed via insect damage or fungal infection. Trees planted in the widest spacing exhibited the widest range of leaf damage.
Figure A.4 – For the sub-plot treatments of tree row spacing (one, three, and five foot spacing regimes), there was a reduced effect in the second year of growth when compared to the first year growth (height) and the amount of leaf area that was removed via insect damage or fungal infection.
# VITA

Joshua R. Herr

## Education

<table>
<thead>
<tr>
<th>Institution</th>
<th>Degree</th>
<th>Location</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Pennsylvania State University</td>
<td>Ph. D. in Plant Biology</td>
<td>University Park, PA</td>
<td>2006-2013</td>
</tr>
<tr>
<td></td>
<td>Advisor: JE Carlson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Wyoming</td>
<td>M.S. in Botany</td>
<td>Laramie, WY</td>
<td>May 2002</td>
</tr>
<tr>
<td>Virginia Tech (VPI &amp; SU)</td>
<td>B.S. in Biochemistry</td>
<td>Blacksburg, VA</td>
<td>May 1997</td>
</tr>
</tbody>
</table>

## Publications (conducted during, but not included in dissertation)

