EST DISCOVERY IN THREATENED *TSUGA CAROLINIANA*

By

ANNE NJERI MWANIKI

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

UNIVERSITY OF FLORIDA

2009

To my daughters, Zuri and Zuwena

ACKNOWLEDGMENTS

First, I would like to thank the Graduate Program in Plant Molecular and Cellular Biology for accepting me, and allowing me to be a part of it. I also would like to thank my committee members Gary Peter, John Davis and Alison Morse, and the members of their respective labs for their guidance and encouragement through my degree. I would also like to express my gratitude to family for their support and encouragement, and finally to my best friends Kariuki Gikang'a and Leandro De Freitas.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF OBJECTS

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Science

EST DISCOVERY IN THREATEND *TSUGA CAROLINIANA*

By

Anne Njeri Mwaniki

August 2009

Chair: Gary Peter
Major: Plant Molecular and Cellular Biology

Hemlock Woolly adelgid (HWA) has largely infested *Tsuga* species across North America. HWA was introduced into the United States from Japan and other parts of Asia *T. chinensis* is highly susceptible to HWA while *T. caroliniana* is resistant. There are currently no biological controls to reduce *Tsuga* devastation by HWA. The goal of this project is to discover genes that are differentially expressed between the two *Tsuga* species that may be associated with resistance or susceptibility.

This project involved collecting infested and non-infested Carolina hemlock samples while HWA were breaking aestivation. *T. caroliniana* RNA was extracted from needle and stem samples and pooled in a 2:1 ratio prior to cDNA synthesis. High throughput sequencing was done on the GS-FLX pyrosequencing platform. A total of 120,850 ESTs were discovered resulting in 35,761 contigs and 93,464 singletons. Annotation was performed using BLASTx and using Agilent's eArray, probes were created from the sense strand. Gene ontologies (GO) for ESTs were also assigned by performing a BLASTx against the TAIR database. Using an e-value of 10^{-4}, 2,255 GO annotations were assigned and classified into three main categories, molecular functions, biological functions and cellular components. I also identified genes encoding CC-

NBS/LRR and NBS/LRR domains with a stringent e-value of 10^{-12}; members of these families are known to encode plant resistance genes to insects that feed using stylets.

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

The Origin of Hemlock Woolly Adelgid

Hemlock Woolly Adelgid (HWA) is an exotic homopteran pest that is native to China and affects *Tsuga canadensis* and *T. caroliniana*. HWA are 1-2 mm long and as they mature produce a wooly wax coat which protects the adult and egg during cold seasons and also against pests and other predators (Montgomery et. al 2004). In North America, HWA were first discovered in British Columbia in the early 1920s (Annad et al. 1924) and in 1951 in Virginia. Today HWA are spread all along the east coast of the United States (Figure 1-1). Adelgids are small piercing insects that feed by inserting their stylet in the needle traces and extracting nutrients from vascular ray parenchymal cell (Young, Sheilds et al. 1995). HWA feeding causes *Tsuga* needles to desiccate and buds to stop growing. The trees begin to look grayish within a few months of infestation and later start losing foliage within a few years. Major limb dieback is visible within 2 to 4 years of infestation. Although the trees may survive more than 4 years, they are more prone to other disease and pests (McClure and Shields 2001). If left uncontrolled, adelgid have been recorded to advance 17 km/year (Cheah Carole 2004).

HWA populations in North America naturally fluctuate as a result of weather changes, (winter cold) and also tend to decrease as *Tsuga* populations decline due to infestation or drought (Montgomery et. al 2004). HWA have established large populations in cold hardiness zone 6 which has a temperature range of -25 to -18 degrees Celsius (Cheah Carole 2004). In a study by (Skinner, Parker et al. 2003) HWA were found to lose tolerance to cold as the winter progressed, perhaps as a result of reduced food quality. In an attempt to identify the speciation and source of HWA in eastern North America, it was found that a single haplotype of HWA was shared among all *Adelges tsugae* species in North America and those found on *T. sieboldii* in southern Japan

(Havill, Montgomery et al. 2006). The lineage of adelgid found in Northern America is also accustomed to low elevations in Southern Japan. Until the 1980's, *Adelges tsugae* was thought to have a monomorphic life cycle, with only one parthenogenic sisten that occurred in hemlock. However, new developmental stages were identified revealing a complex polymorphic life cycle that occurred on spruce trees (McClure 1992). Adelgid has three generations that develop each year, two of which are winged and one that is non-winged. Their development and emergence, varies with weather conditions and elevations. In North America, nymphs develop and mature in June and adult adelgid remain on the hemlock until July when they start producing ovisacs. Most crawlers undergo aestivation a state of dormancy defined by reduced movement and no feeding and then begin to feed in early February (Figure 1-2). In south West Virginia and North Carolina, the second generation breaks aestivation in late September, (M.McClure 2001) therefore, for this project, hemlock material from infested *T. caroliniana* was collected in November when adelgid were actively feeding. This was done to maximize discovery of genes expressed in Carolina Hemlock during feeding by the adults.

Feeding habits of adelgid have been documented in thin sections of branch material using Safranin O which stains salivary sheaths a red color, adelgid were found to feed both intra- and intercellularly (Young, Sheilds et al. 1995). Insertion of the stylet into the vascular bundle is intracellular, although some intercellular paths were also observed. Upon reaching the vascular bundle, the stylet followed an intercellular path along the vascular bundle in the xylem between the tracheids and ray parenchyma cells. In samples where more than one insect was feeding, it was observed that one stylet bundle followed the tracheids closest to the xylem while the other followed distal tracheids. Towards the paths' end, the stylet bundle traveled intracellularly to the ray parenchyma cells which were the final feeding site. A larger number of feeding tracks were

found in previous year growth collected in early or late summer and in the current years' growth collected in early summer. The salivary track within the plants largely reflects a developing population of adelgids, in early summer the current years' growth only had tracks from the first developing generation, whereas the late summer had both the first and second generation tracks.

Although there is no known parasitoid control of HWA (McClure 1987) in the past, parasites like S*asajiscymnus tsugae* which measures a length of 2 mm and is a small coccinellid that specializes in aphids, scales, and adelgids among other small insects has been tested (Cheah Carole 2004). However, most of these biological controls are generalists (McClure 1987). Non-host specific fungi have also been used to control HWA in its adult stage, but no naturally occurring epizootics have been observed. Thus far, biological controls have not been enough to maintain low populations of HWA. Chemical controls for HWA have been used in various regions; however, their use is made difficult by the close proximity of trees and the sites in which most grow. Carolina hemlocks grow on mountain sides, cliffs and river banks (McClure and Shields 2001). When used, horticultural soaps and oils are the preferred treatments for controlling adelgid and are either applied to the soil or directly injected into the hemlock stems. In both treatments, adelgid ingest insecticides through the sap (M.McClure 2001). Like many other plant diseases and pathogens, the adelgid has shown to be spread mainly by birds, animals and wind mainly as eggs or crawlers (McClure 1990).

Ecological Importance of Hemlock

Hemlock is of great ecological importance because it provides shelter and acts as a food source for a variety of flora and fauna ranging from deer to wood sorrels (McClure and Shields 2001). Eastern Hemlock has also been shown to assist in containing headwater ecosystems (Snyder Craig D. 2002). There are nine species of the genus *Tsuga* two species in Western North America, two in Eastern North America and five in Asia (Farjon 1990). The Western

North America and Asian *Tsuga* species are considered to be resistant to HWA, whereas the eastern species are susceptible (McClure 1992). Species include *Tsuga canadensis* ,*Tsuga caroliniana*; *Tsuga chinesnsis*, *Tsuga diversifolia*, *Tsuga heterophylla*, *Tsuga sieboldii*; *Tsuga mertensiana*, *Tsuga dumosa* and *Tsuga forrestii* and all have been documented to be infested with adelgid (McClure 1992; Lagalante and Montgomery 2003),

Plant Defense Mechanisms Against Insects

Independent of the attacking organism, plants have four main stages in defense mechanisms. These begin with the plants attempt to prevent invasion, if this fails, the plant then tries to compartmentalize the invader. The third stage is to seal and repair damaged plant cells and tissue followed by systematic defenses that prevent such future attacks (Franceschi, Krokene et al. 2005). Plants express various defense genes depending on the type and extent of feeding by insects (A. and Georg 2008). In general, chewing insects cause more damage than phloem feeders. Plants have developed various mechanisms to cope with insect infestation. The mechanisms are either direct, where plant defenses like thorns, toxins and secondary metabolites are produced to deter the attack, or indirect in which plants protect themselves by attracting the pathogen's predators to the plant (A. and Georg 2008). Herbivores mainly induce direct defenses since they cause extensive damage to the plants. Plant defense mechanisms can be costly because they reduce the carbon and nitrogen pool available for growth and development and focus it toward defense, attaining a balance of growth and protection is key to a plants survival during pathogen or pest attack (Walling 2000).

Plants have various mechanisms for defense against insect attack. Different plant-insect interactions have demonstrated that mechanical damage to plants by insects results in the emission of volatile compounds. For example, this is evident in hybrid poplar which responds to forest tent caterpillars by producing high levels of (-) gregarine D (Gen-ichiro, Dezene et al.

2004). Some plants produce enzymes that strengthen the cell wall improving the barriers between the plant and insects whereas other plants increase secondary metabolite production (Richter 2001). A large body of evidence shows that plants have genes that are involved in both direct and indirect defense mechanisms in response to insect attack. For example, proteinase inhibitors are induced by wounding and feeding and limit digestion whereas polyphenol oxidases reduce the nutritional value of plant substrates by crosslinking proteins with phenolics and other toxic compounds like terpenoids and alkaloids (Kessler and Baldwin 2002). Proteinase inhibitors (PI) have been shown to be produced in tomato leaves after wounding and insect feeding. PI production is activate by jasmonic acid production (A. and Georg 2008).More insect responses are summarized inTable 1-1

Genes activated are usually correlated to mode and extent of feeding on the plant (Walling 2000). For example, Arabidopsis in response to aphid feeding induces transcription of genes involved in salicylic acid (SA)-production, *PR-1* and *BGL2* (Moran and Thompson 2001). A similar response has also been observed in tomato and squash (Walling 2000). In pine, it has been observed that pathogen related proteins (PR) are elevated after plants are infected with insects. These PR genes include, stilbene synthase (Regina Preisig-müller 1999), chitinase genes (M, Wu et al. 2002) and peroxidases (Fossdal, Gunnar et al. 2001). However, in pine species resin-based defenses are the most studied defense mechanisms (Phillips MA 1994).

During infestation by some sucking insect, plants respond by producing resistance or R proteins which match the avirulence gene of an insect. These genes provide resistance for insects, nematodes, fungi, viruses and bacteria. In response to aphid infestation and feeding, tomatoes produce Mi-1 proteins. Mi-1 is a NBS –LRR protein with a highly conserved sites for ATP binding and hydrolysis (Tameling, Elzinga et al. 2002).. Significantly more is known about

phloem feeders as opposed to xylem feeders. Phloem feeders insert their probes intercellularly through the epidermal and mesophyll cell layers with their stylet like mouth parts. In response to phloem feeders like aphids, Arabidopsis plants begin transcription of genes associated with salicylic acid (SA) pathway and PR-1 and BLG2 genes. Aphid feeding also results in the increase of stress-related monosaccharide symporter gene, *STP4* (Moran and Thompson 2001). In a plant more closely related to hemlock, pine responds to insect boring by both physical and chemical defenses. Both mechanisms can be constitutive or inducible. Physical barriers are by lignifying cells and calcium oxalate (Franceschi, Krokene et al. 2005). Pine also produces tepenoids that act as attractants for insect predators. Terpenoid synthesis is regulated by terpene synthase (TPS) genes which are expressed prior, during and after insect attacks. In Douglas fir, monoterpenes are produced in response to spruce budworm defoliation (Chen, Kolb et al. 2002). In different experiments, it has been shown that levels of terpene fluctuate among different species of *Tsuga* depending on their susceptibility or resistance to HWA. Volatiles have shown to be of great importance in plant/insect interactions, acting as signals for reproduction and defense mechanisms against herbivores. Conifers have evolved terpene-based defense mechanisms to deter bark beetles and their associated fungal infestations (Phillips and Croteau 1999). Various attempts have been made to determine what factors are responsible for the susceptibility or resistance of different *Tsuga* species to HWA. Analysis of terpenoids in *Tsuga* species using solid-phase microextraction, gas chromatography/ion trap mass spectrometry revealed that there are elevated levels of isobornyl acetate in *T. caroliniana* and significantly lower levels in *T. chinensis*. It was proposed that isobornyl may function as a HWA attractant (Lagalante and Montgomery 2003).

In the past, *Tsuga* species were found in 20% of US forest ecosystems; however, harvesting has reduced these populations to less than 6%. Hemlock is widely used for pulpwood and, with 274 cultivars of *Tsuga* cultivated in the United States, hemlock is widely used for landscaping purposes (McClure and Shields 2001). *Tsuga* hybrids were made and tested for resistance to adelgids. Five hemlock species from eastern North America and Asia were used in controlled pollinations. Putative hybrids were analyzed 44 with Amplified Fragment Length Polymorphism (AFLP). Fifty-nine *T. caroliniana* x *T. chinesis* hybrids were confirmed and that appeared to have intermediate resistance to HWA (Bentz, Riedel et al. 2001). SNP analysis could be used to further understand HWA resistance. Classical breeding has led to improvements in crop cultivars and disease resistance (Johnson Mar 2000.). Some examples of plants which have attained disease resistance through breeding are barley to powdery mildew (Jørgense 1992) and resistance to yellow and brown rust in wheat (Johnson R. 1983). Genetic resistance is acquired through selective breeding and in the past has been largely used in animal populations. Genetic resistance involves identifying and characterizing candidate genes and increasing their expression in desired plants. A good example of how genetic resistance has been conferred is evident in maize where disease resistance genes are incorporated into commercial lines by selecting resistant progeny derived from crosses of resistant and susceptible plant (Tsuji, Fischer et al. 2003). In the past, genetic resistance has been selected for in pine (Wilcox, Amerson et al. 1996), and rice (Lin, Anuratha et al. 1995).

The development of protein and DNA markers offers new approaches to conferring plant resistance. Markers provide a connection between phenotypes and specific set(s) of genes. Improvements in statistical analyses have increased the utility of markers and microsatellites in extracting information from genes (Sunnucks 2000). With sequences discovered, DNA sequence

variation can be used to introduce resistance traits in hemlock. Molecular markers have been used succesfully in breeding rye and other cereals (Roder, Korzun et al. 1998). 454 sequencing has previously been used to identify markers for resistance in plants where ESTs are analyzed for simple sequence repeats (SSRs) that can be usual as molecular markers (Varshney, Graner et al. 2005). These markers can be used to confirm resistance. Genetic engineering permits the introduction of desirable genes using a single event (Hilder and Boulter 1999). Plant resistance to disease and insects is more efficient through markers than classical breeding and phenotypic selection due to the long maturation periods for trees. However, a combination of classical breeding and creation of marker could be an effective means to improve Hemlock resistance to adelgid.

Objectives

1) Develop an EST resource for *Tsuga* to identify genes that are highly regulated during the trees response to adelgid infestation. Both needle and stem samples were used during gene expression discovery. With the identification of these genes, different approaches can be taken to improve adelgid resistance, one of which includes, creating transgenic trees that have been bred for adelgid tolerance.

2) Characterize the EST resource. Sequencing data was analyzed to determine efficiency of normalization, ESTs that were deposited into Genbank and create Gene Ontologies for the various genes discovered.

3) Design long oligonucleotide arrays to quantify changes in mRNA level from infested and noninfested tissues. For future experiments, sequences from pyrosequencing were also used to design probes for differential gene expression using Agilent micro arrays with the 4x44K platform. Microarray between Carolina infested and non- infested would provide information regarding genes that are expressed during adelgid feeding.

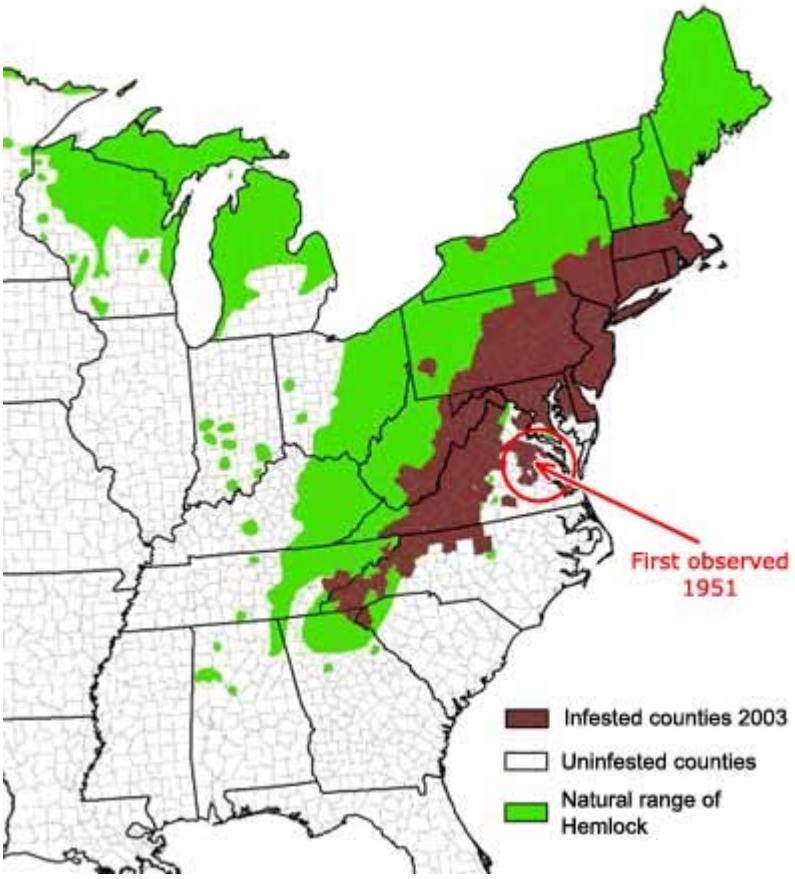

Figure 1-1 Distribution of Hemlock Wooly Adelgid in the United States. The green regions denote native populations of hemlock whereas the brown shown adelgid spread within the population (Cheah Carole 2004).

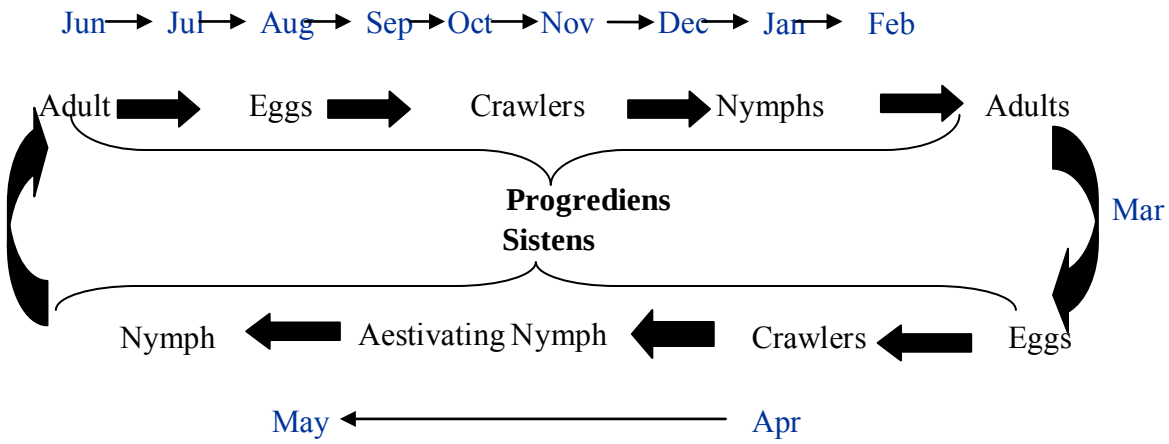

Figure 1-2: Life cycle of Adelgid in Eastern North America. There are two generations a year, wingless Sisters which hatch in the spring and progeny which produce both wingless and winged offspring.

Table 1-1. Various plant responses to insect feeding

Insect	Plant	Response	Reference
Aphid	Arabidopsis	NBS/LRR proteins	(Tameling et al., 2002)
Western Spruce Budworm	Douglas Fir	Monoterpnes	(Chen et al., 2002)
Small white butterfly (Pieris rapae)	Arabidopsis	Terpenes	(Reymond, Bodenhausen et al. 2004)
Yellow-striped flea beetle	Sorghum	cyanogenic glycoside dhurrin	(Howe and Jander 2008)
ForestTent Caterpillars	Hybrid poplar	Terpenoids	(Gen-ichiro et al., 2004)

CHAPTER 2
EST DISCOVERY WITH 454 PYROSEQUENCING

Introduction

For Eukaryotic organisms one of the most efficient methods for gene discovery is the random sequencing of mRNA isolated from tissues of interest (Neil 2007). These expressed sequence tags (ESTs) give insight into the number of genes present in a genome and clues about their biological functions. Until recently, EST discovery was performed through traditional Sanger sequencing of cDNAs that were cloned into plasmid vectors and randomly picked prior to template preparation. Sanger sequencing, is based on electrophoretic separation of DNA fragments that have incorporated the chain terminating dideoxy nucleotides (Sanger, Nicklen et al. 1977). With recent improvements in automation using capillary electrophoresis Sanger reads have increased from 450 to 850 bp and have accuracies of 99% (Neil 2007). However, limitations to the cost and throughput of this method drove the development of alternative DNA sequencing methods.

Two such methods are now commercialized, pyrosequencing with emulsion PCR (Schuster 2008) and sequencing by synthesis (Mardis 2008). Sequencing by synthesis with the Illumina/Solexa or ABI SOLiD platforms, gives short 30-40 bp read lengths that are most useful for organisms with a genome sequence, due to the problems associated with *de novo* assembly of short reads into longer sequences. In contrast, pyrosequencing methods such as 454 give 200-500 base read lengths, depending on the chemistry and equipment, and thus are useful not only for organisms with a genome sequence, but more importantly ones that lack a genome sequence.

In Roche/Life Sciences 454 pyrosequencing platform, DNA is randomly fragmented and the fragments are captured onto beads for clonal amplification prior to sequencing. During sequencing, the pyrophosphate released during chain elongation is converted to ATP by

sulfurylase. The ATP is converted by luciferase and the emitted photons are quantified by digital cameras (Ronaghi, Karamohamed et al. 1996). The library to be sequenced is integrated with a solution of agarose beads which carry complementary oligonucleotides to 454-adapters that are located on the fragmented library to be sequenced (Mardis 2008). 454 sequencing has dramatically increased sequencing depth and coverage while reducing labor, cost and time. Sequence errors are as low as <1% (Margulies, Egholm et al. 2005). However, base calling when using 454 sequencing is limited as the equipment is only calibrated to interpret less than 6 identical nucleotides in a row. In the past, most pyrosequencing has been limited to model organisms but recently, non model animals and plants (Garcia-Reyero, Griffitt et al. 2008; J. Cristobal Vera 2008; Vera, F et al. 2008) have been successfully sequenced and analyzed.

This new technology enables sequencing of large genome organisms by reducing cost and throughput limitations. These high throughput DNA sequencing methods have been applied in various organisms including the sequencing of genomic DNA from remains of Neanderthal (Noonan, Coop et al. 2006) using a metagenomic approach and mammoth (Poinar, Schwarz et al. 2006). 454 sequencing has also been used to sequence more mainstream genomes like (Cheung, Haas et al. 2006), Zea mays (Emrich, Barbazuk et al. 2007), Arabidopsis (Jones-Rhoades, Borevitz et al. 2007), Medicago (Cheung, Haas et al. 2006) and to discover ESTs from normalized cDNA from *Eucalyptus grandis* (Novaes, Drost et al. 2008).

The previously mentioned studies were performed on 454 Roche/Life Science GS-20, however a new addition to the pyrosequencing technology is the GS-FLX which can sequence between 80 -100 million base pairs with reads ranging from 200 – 300bp. Sequencing has been used to discover ESTs and SNPs in various non model insect species. 454 has been used to characterize the transcriptome of non-model organisms like fly *Sarcophaga crassipalpis,*

hornworm *Manduca sexta* and *Melitaea cinxia*; the Glanville fritillary butterfly (Hahn, Ragland et al. 2009) (Zou, Najar et al. 2008)and(Vera, F et al. 2008)). In human and plants, it has also been used to identify methylated regions on DNA (Poinar, Schwarz et al. 2006). Pyrosequencing has also facilitated the discovery of SNPs (Barbazuk, Emrich et al. 2007) by sequencing multiple genomes with parallel tagged sequencing (PTS) which is a type of barcoding that also facilitates sequencing of double stranded nucleic acids (Meyer, Stenzel et al. 2007). For this project, 454 sequencing was preferred to maximize read length, creating as much transcriptome coverage as possible.

Tsuga chinensis is known to have resistance to HWA. Crosses between Chinese and Carolina hemlocks produce F1 hybrids that are more resistant to HWA than Carolina alone (Bentz, Riedel et al. 2001). Thus, generating resistance to HWA in Carolina hemlock appears possible with hybrid breeding and backcrossing. However, this process is expected to take a very long time since no hemlock breeding programs exist and the reproductive life cycles are rather long. One way to accelerate the development of HWA resistance in Carolina hemlock is to identify the gene(s) that mediate this resistance. However, there are no known EST and genome sequences from *Tsuga* species. Therefore, EST sequencing will provide a stepping stone to future genetic breeding by possible discovery of genetic markers. DNA analysis for insect resistance can also be performed using additional biological replicates of RNA with microarrays designed from this sequence resource. 454 pyrosequencing was chosen over Solexa and Sanger because a large number of genes can be discovered with reasonable length to annotate their functions with genes present in the databases. We performed a single 454 Life Science GS-FLX run from pooled infested and non-infested *T. caroliniana* stems and needles. The purpose of this sequencing run was to discover genes from *T. caroliniana* to use to design arrays for studying

differential gene expression between infested and non-infested susceptible *T. carolinina* and between susceptible and resistant hemlock hybrids and species.

Materials and Methods

Plant Material

Needle and branch samples were collected for *T. caroliniana.* Tissue samples were collected from three *T.caroliniana* trees/genotypes. Samples of tissue infested with HWA were located in the state forest in Asheville, NC. Two of the three genotypes had non-infested regions from which needles and branches were also collected. Samples were quick frozen in liquid nitrogen to maintain RNA integrity and transported in dry ice. Additionally, *T. chinensis* samples were collected from the USDA South farm, Beltsville, MD and were stored and transported in the same manner as the *T. Caroliniana*.

Samples were ground in liquid nitrogen using a mortar and pestle and RNA was extracted from 2 grams of stem and 1 gram of needle tissues (Chang, Puryear et al. 1993). CTAB extraction buffer containing β-mercaptoethanol was heated to 65° C and 17 ml was added to the ground tissue and homogenized for 30 seconds. Following this, 20 ml of chloroform was added to the extraction buffer and the conical tubes were thoroughly mixed by shaking for 2 minutes. The samples were spun for 10 minutes at 10,000 RPM in Beckman JA-20 rotor. The aqueous phase was transferred into an Oakridge tube and the chloroform extraction was repeated three more times. After the third chloroform extraction, equal amounts of aqueous layer 4M $LiCl_2$: 50mM EDTA were combined and incubated overnight at 4° C. The precipitated RNA was harvested by centrifuging at 10, 000 RPM for 20 minutes and the supernatant was carefully removed. The RNA pellet was dissolved in 500 µl of SSTE incubated at 65° C for 2 - 10 minutes. One more chloroform extraction was performed with 500 µl, and the RNA was precipitated with 2 volumes of 95% ethanol. The RNA pellet was washed in 200 µl of 70%

ethanol and air dried. RNA was quantified using a ND-1000 Spectrophotometer (NanoDrop USA, Wilmington DE) and its quality assessed after electrophoresis on a 1.2% agarose gel stained with ethidium bromide.

cDNA Synthesis

Prior to cDNA synthesis, total RNA was pooled from infested and non-infested samples from one *T. caroliniana* tree in a 2:1 ratio. cDNA was synthesized from 2µg of total RNA that had been DNase treated and cleaned using RNeasy Plant Mini Kit (Qiagen USA, Valencia, CA) using the Clontech SMART cDNA Library Construction Kit (Clontech, USA, Mountain View, CA) following manufacturer's protocol with the exception that the Clontech CDSIII/ 3' Primer was replaced with the Evrogen CDS-3M adaptor (Evrogen, Moscow). cDNA was amplified using PCR Advantage II polymerase (Clontech USA, Mountain View, CA) for 17 cycles (7s at 95° C, 20S at 66 ° C and 4 mins at 72° C) and purified with the QIAquick PCR purification Kit (Qiagen USA, Valencia, CA).

The cDNA was normalized to reduce transcript over representation using the Evrogen Trimmer-Direct Kit (Evrogen Moscow) which uses a duplex-specific nuclease (DSN) isolated from Kamchatka crab that shows a preference for double stranded DNA and RNA –DNA hybrids compared to RNA and single stranded DNA (Zhulidov, Bogdanova et al. 2004). Following normalization cDNA was amplified for 13 cycles at 95° C, 20S at 66° C and 4 min at 72° C). A SfiI digest was performed on the library prior sequencing, followed by purification using QIAquick PCR purification Kit (Qiagen USA, Valencia, CA) to remove adaptors incorporated.

454 sequencing and assembly

Seventeen µg of normalized cDNA were used for 454 sequencing at the Interdisciplinary Center for Biotechnology Research (ICBR) at the University of Florida following procedures

previously described by Margulies et al. (2005). Initially, a titration run (1/4 of a plate) produced 8,687 reads with 2,094,333 base pairs, demonstrating that the cDNA was of sufficient quality for a full run. For this pooled *T. caroliniana* run, there was a 6.5 fold increase in the number of bases during the first production run. However, due to lower than expected yield, two more production runs were performed each having ~ 15X increase in the total number of base pairs (Table 2-1). One full sequencing run was processed on the GS-FLX platform. Initial sequence assembly was performed using Newbler version 1.1.02.15 which used pairwise comparisons. Newbler assembly considers normalization intensity of each nucleotide flux as opposed to base calling. However, it doesn't mask sequence repeats and generates 454 contigs and 454 reads (Novaes, Drost et al. 2008) Newbler assembly is followed by Paracel Transcript Assembler version 3.0.0 which assembles the 454 reads and contigs again after masking repeats and adaptor containing reads and also performs a low base call quality

Results

DNA Sequencing

From the pooled *T. caroliniana* samples, a total 79.3 megabases of sequence were obtained from 1 titration and 3 production runs (Table 2-1 and Table 2-2). Compared with the titration run, the high throughput production runs gave 45 fold more reads (Table 2-2). Overall, the average read length was 240.37 bp, with 85% being between 100 and 500 bp and ~ 62% having lengths between 100-250 bp (Table 2-3). Newbler and Paracel Transcript Assembler generated a total of 128,819 ESTs with 35,761 contigs and 93,464 singletons (78.7%) comprising 77.9 Mbp of sequence (Table 2-2). The maximum contig length was 2,561 bp.

BLAST Annotation

BLASTx searches with the 128,819 ESTs, identified 50,036 ESTs (38.8%) which have significant sequence similarity to genes in the NR database at an e-value of 10^{-4} or below. For the

contigs, 58% had significant sequence similarity to genes in the database (hits). As expected, the average length of contigs with hits was longer (> 2 fold) than contigs with no hit in the database (Table 2-4 and Figure 2-1). A sensitivity analysis with decreased e-values at 10^{-9} and 10^{-12} showed the expected larger percent decrease for singletons hits compared to contigs (Table 2-5 and Figure 2-2). For contigs, increasing the stringency of the e-value threshold from 10^{-4} to 10^{-9} reduced the annotated sequences by 16.7% and from 10^{-9} to 10^{-12} and by additional 8.6%. By comparison, the annotated ESTs for the singletons was reduced to a greater extent, dropping by approximately 47% from 10^{-4} and 10^{-12}, with the most significant reduction being between 10^{-4} to 10^{-9} at 33.5%.

Genes with the most sequence similarity to Drosophila genes were also annotated during the BLASTx (Table 2-6). These genes were discovered at a cutoff e-value of e 10^{-12} suggesting that they may be actual adelgid genes and not from hemlock. This maybe due to the insertion of the stylet into the vascular bundle during feeding on the ray parenchyma cells (Young, Sheilds et al. 1995). However, overall this was a very small fraction of the genes discovered appear to be adelgid genes.

To learn more about the contigs, we counted the number of first hits ones with the highest sequence similarity from various plants (Table 2-7). Of the 30,840 sequences, 21.5 % of these hits were annotated with a *Picea* gene. In contrast, only 1.2% had the strongest similarity with *Pinus* genes even though both organisms have similar numbers of ESTs in Genbank and all three species are closely related. Interestingly, although the entire Poplar genome has been sequenced, there were no annotations of *T. caroliniana* that had the highest Hit assigned to Poplar. Based on the same stringent e-value of 10^{-12}, we screened for genes that were related to disease resistance

proteins or pathogen related proteins(Table 2-8). We found 9 genes that are related to pathogenesis related proteins (Table 2-9).

GO Annotation

To obtain GO annotations for queries acquired from sequencing, BLASTx was performed with the 128,819 hemlock ESTs against the TAIR database at an e-value of 10^{-4}. 26,131 sequences annotated to 9,888 unique Arabidopsis locus identifiers. However, these 9,888 identifiers only had 2,255 GO annotations. GO annotations were then classified into three main categories: molecular function, cellular component, and biological processes. The majority of the hemlock genes discovered were categorized as biological processes and molecular functions representing 48.7% and 40.3%, respectively (Figure 2-3). In each of the three categories the genes were further classified based on specific cellular location, molecular function or biological functions (Tables 2-10 – 2-12).

A total of 1095 genes had biological process GO annotations, where 40% represented cellular processes which include membrane transporters. Under biological processes, sex determination, mitotic cell cycle and life circadian rhythm and response to nutrients were all collapsed into other cellular processes (Table A-1). We discovered 102 genes categorized as responsive to stress, as well as biotic and abiotic stimuli (Table 2-13). For the genes with cellular component GO annotations, the majority (247) were located to the nucleus (Table 2-11). Prior to a summarization (Figure 2-3), all genes from various plastids had separate annotations for a compressed summary of all plastid genes. Mitochondrial inner membrane and envelope were all collapsed into mitochondria cellular components. Nuclear pore, outer membrane and envelope were all classified under nucleus located genes (Object A-2). We also noticed that only 1 gene was associated with the chloroplast. A total of 905 genes had molecular function GO annotations

(Table 2-12 and Object A-3). These include kinases, transcriptional regulators and DNA or RNA binding proteins. A total of 29.6% represent genes involved in enzyme activity such as aldehyde dehydrogenase, desaturase and isomerase activity. A broad spectrum of coverage would be expected assuming the normalization procedure was successful.

Pathogen and Stress Responsive Genes

Fifty-seven genes were annotated as responding to stress including but not limited to environmental factors like temperature, humidity, ionizing radiation. These also included genes that respond to fungal and viral infection. Jasmonic acid and ethylene dependent pathways have been shown to be induced during insect infestation. Among these, are genes associated with response to wounding, a gene assigned a GO ID of 0009611, was described as being involved in the jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signaling pathway, defense response to virus and virus induced signaling among others (Table A-1).

I discovered a total of 130 genes with NBS/LRR domains, 50% of which were annotated from contigs only (Table 2-8). 11 of these NBS /LRR proteins also had N-terminal coiled-coil (CC) domains. At an e-value threshold of 10^{-12}, 21 putative TIR/NBS/LRR disease resistance proteins were identified. Despite having the overall highest number of Hit annotations from *Picea*, 76.1% of these disease resistance proteins were annotated from *Pinus*. Of the 21 TIR/NBS/LRR genes, 13 are contigs (Table 2-8). Although, this class of proteins are also involved in other biological processes, the material sequenced was from a pool of infested and non infested hemlock tissues, thus it is possible to these genes could mediate some level of resistance in hemlock. Other genes that may be useful in future attempts to identify genes that confer HWA resistance are pathogen related proteins (Table 2-9).

Discussion

EST Discovery

My goal for this experiment was to identify ESTs in *T. caroliniana* to provide the first large-scale EST resource for hemlock. Prior to this, there has been no attempt to sequence the transcriptome or genome of hemlock. From our analysis, 128,819 ESTs were discovered; however, only 38.8% had significant hits identified after a BLASTx against the non-redundant database. The average length of the hits was 309.5 bp which was higher than the average read length for each sequence read acquired through 454 (Margulies, Egholm et al. 2005) due to the improved FLX platform. Previous sequencing runs on the GS_FLX attained similar results with the read lengths averaging 209.89 bp. Short sequence reads limit the ability to find sequence similarity during annotation (Novaes, Drost et al. 2008).

The various genes with annotations to BLAST at an e value of 10^{-12} from Drosophila could be due to contamination from feeding adelgid. In previous experiments, using safranin O (Young, Sheilds et al. 1995), aldgegid feeding stylets have been observed in the ray parenchyma cells. The cDNA library consisting of infested stem needles may have contributed to a gene pool from adelgid.

Tissue type representation in GO

BLAST searches with the sequences acquired from 454 sequencing against the TAIR database yielded 9,888 genes with an e-value 0.5 10^{-4}. These hits were analyzed for GO categories using Arabidopsis as an intermediate since there are no annotations for pine or any other tree with a genome similar to that of hemlock. Genes were categorized based on molecular or biological functions or cellular components. However, some genes fell under more than one category. These genes will be more informative after the differential expression using microarrays is complete (Table 2-7).

Microarray design

Probes were designed from 454 sequence data using Agilent eArray program. Prior to probe design an annotation was performed by doing a BLASTx against NCBI sequences to determine gene ontology and the sense strand of each 454 contig that had a e-value stronger than 0.0001. All probes were printed on Agilent microarray. For probe design, Base Composition Methodology platform was used. Probes were created without 3' Bias and also using best probe methodology where probe design is weighted to creating the best probe resulting in one probe/ transcript target. A total of 49,549 probes were designed and printed out on a 4X44 K array.

During the microarray experiment, contamination from genes from HWA must be considered since during sample preparation for sequencing, plant material was pooled from some infested samples. This may also explain why there were some sequences annotated by BLASTx at e-value 10^{-12} as Drosophila genes. We also discovered various genes that annotated as disease resistance proteins with TIR/NBS/LRR domains. More disease resistance analyses can also be performed with a focus on CC-NBS/LRR domains since they too have been shown to have association with disease response with two models proposed for their activation. CC-NBS/LRR proteins interact with viral coat proteins which induces conformational changes in the nucleotide binding of the NBS. Majority of these proteins were annotated from contigs. These genes have been shown to be highly conserved, therefore in future differential gene expression analysis comparing expression of uninfested and infested tissues may help determine if one of these genes is involved in resistance.

Table 2-1. A summary of 454 sequences prior to assembly.

Run statistics:	Titration run	Production run (low throughput)	Production run *	Production run*
# of reads	8,687	86,470	148,076	150,902
# of clean reads	8,674 (99.8%)	86,044 (99.5%)	147,500 (99.6%)	150,352 (99.6%)
# of bases	2,094,333	14,228,599	30,296,823	31,368,822
# of clean bases	1,994,562 (95.2%)	13,281,468 (93.3%)	28,493,400 (94.0%)	29,476,029 (94.0%)
Avg. length	241.1	164.6	204.6	207.9
Min.length	44	35	30	25
Max. length	382	327	376	411

*The Production runs denoted by an asterisk are re-runs of a production run that are performed when the total number of reads doesn't meet desired average. This is why the production runs have higher number of bases and increased sequence lengths.

Table 2-2. Summary of base pairs and total reads. A comparison between titration run and fill production run of GS-20 454 sequencing.

	Titration Run	Complete Run
Total Bases	2,094,333	77,245,459
Total Reads	8,687	392,570
Minimum Length	44 bp	35 bp
Maximum Length	241bp	411bp
Number of Contigs	906	35,761
Number of Singletons	5,884	93,464
Average Contig length	217.7bp	391.5bp

Table 2-3. Summary and distribution of assembled sequences. Length distribution and characteristics of contigs

Length	Number of sequence	% of total
<100 bp	12265	9.52
101-250 bp	80026	62.12
251-500 bp	29940	23.24
501-750 bp	3313	2.57
751-1000 bp	1767	1.37
1000 bp	1508	1.17

Table 2-4. Total number of Hits and no Hits, from BLASTx with and e-value of 10^{-4}

	Total #	Max. Length (bp)	Min. length (bp)	Mean length (bp)
All Singleton	96,739	323	60	190.2
All contigs	32,080	2561	73	391.5
Hit Singletons	31,426	323	79.4	216.48
Hit Contigs	18,610	2561	115	466.83
No Hit Singletons	56,833	321	896	192.8
No Hit Contigs	13,676	1849	66	286.05

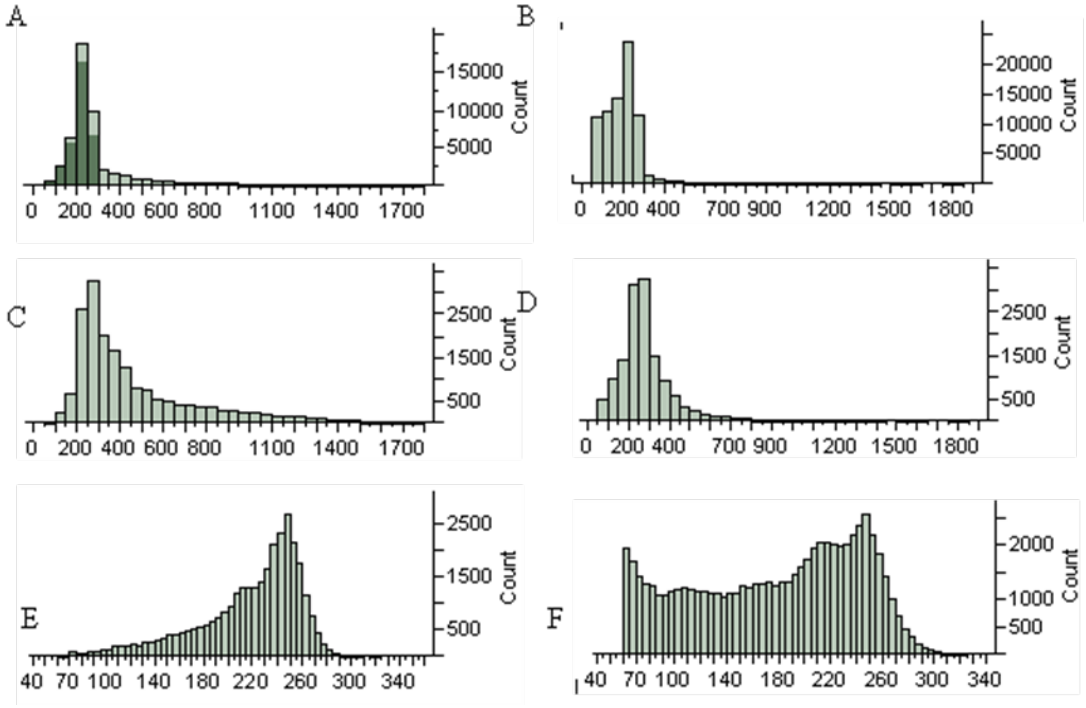

Figure 2-1. These are distribution plots of data acquired from 454-pyrosequencing and assembly. A represents the total number of Hits , C and E are contigs and singletons with Hits. Panel B is a representation of sequences with no blastx hits, D and F are contigs and singletons respectively. E-value was set to $10^{-4.}$

Table 2-5. Number of contigs and singletons with annotations as the e-value threshold gets more stringent.

E value	Contigs	Singletons	Total
10^{-4}	18610	31426	50036
10^{-9}	15486	20914	36401
10^{-10}	15078	19550	34628
10^{-11}	14597	18103	32700
10^{-12}	14150	16690	30840

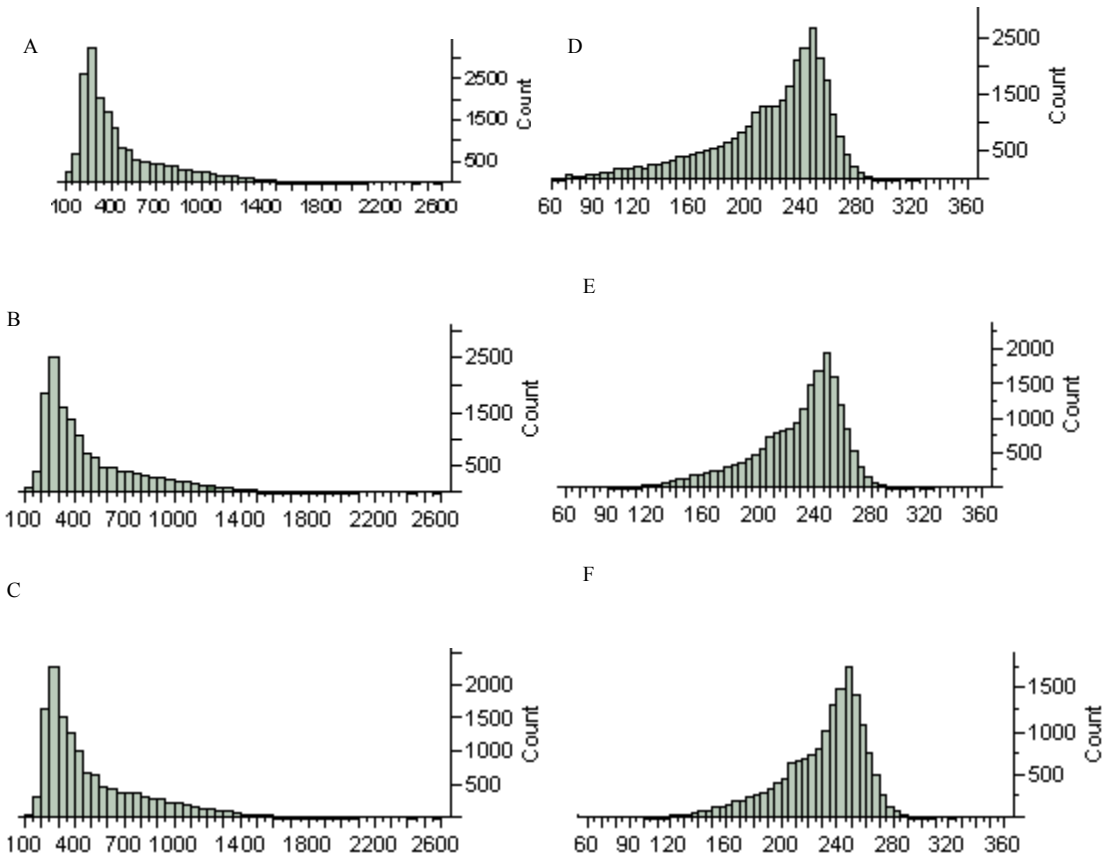

Figure 2-2. Frequency distributions of the length of contigs (A, B, C) singletons (D, E, F) and that had BLASTx Hits at e-values thresholds of 10^{-4}(A,D), 10-^{10}(B,E) and 10^{-12}.(C,F).

Table 2-6. Potential HWA genes from BLASTx using an e-value of 10-12.

Description	Query Name	E-value	Length(bp)
GJ15760 [Drosophila virilis]	CH.17028.C1	1.00E-25	256
GL23535 [Drosophila persimilis] gb|EDW24159.1| GL23535 [Drosophila persimilis]	CH.22362.C1	4.00E-40	262
TMS membrane family protein / tumour differentially expressed (TDE)	CH.23088.C1	5.00E-23	233
vermilion [Drosophila ananassae] gb|EDV44673.1| vermilion [Drosophila ananassae]	CH.24677.C1	1.00E-14	255
GF20391 [Drosophila ananassae] gb|EDV44937.1| GF20391 [Drosophila ananassae]	CH.4909.C1	3.00E-38	929
GL21511 [Drosophila persimilis] gb|EDW34566.1| GL21511 [Drosophila persimilis]	CH.5367.C1	7.00E-22	226
GJ18505 [Drosophila virilis] gb|EDW57418.1| GJ18505 [Drosophila virilis]	E8IB5KT06DUJ5P	7.00E-13	183
GA22494 [Drosophila pseudoobscura pseudoobscura] gb|EDY71215.1|	E9M19J301AJIH4	5.00E-14	245
GJ13683 [Drosophila virilis] ref|XP_002060434.1| GJ14590	E9M19J301AQC3C	1.00E-30	218
GH17878 [Drosophila grimshawi] gb|EDV95315.1| GH17878 [Drosophila grimshawi]	E9M19J301CFGV5	4.00E-24	211
GF24147 [Drosophila ananassae] gb|EDV40058.1| GF24147 [Drosophila ananassae]	E9M19J301CNJL6	2.00E-14	184
GH01866p [Drosophila melanogaster]	E9M19J301CWRRT	2.00E-15	215
GJ22948 [Drosophila virilis] gb|EDW67674.1| GJ22948 [Drosophila virilis]	E9M19J301DZYQ1	3.00E-15	254
GK19626 [Drosophila willistoni] gb|EDW73744.1| GK19626 [Drosophila willistoni]	E9M19J301EVL35	3.00E-27	269
Contains similarity to CG10338 gene product from Drosophila melanogaster gb	FAFEF5V01A5HJH	7.00E-16	240
GF17263 [Drosophila ananassae] gb|EDV41904.1| GF17263 [Drosophila ananassae]	FAFEF5V01B0NI0	4.00E-18	274
GK22207 [Drosophila willistoni] gb|EDW77139.1| GK22207 [Drosophila willistoni]	FAFEF5V01BBTV6	1.00E-19	195
GK15001 [Drosophila willistoni] gb|EDW75799.1| GK15001 [Drosophila willistoni]	FAFEF5V01BNG18	2.00E-18	288
GI13073 [Drosophila mojavensis] gb|EDW18150.1| GI13073 [Drosophila mojavensis]	FAFEF5V01BYI6R	8.00E-18	193
GH10535 [Drosophila grimshawi] gb|EDW03384.1| GH10535 [Drosophila grimshawi]	FAFEF5V01C6SMF	2.00E-35	263
GD11318 [Drosophila simulans] gb|EDX07576.1| GD11318 [Drosophila simulans]	FAFEF5V01CCVQD	3.00E-19	262
GH10818 [Drosophila grimshawi] gb|EDW02844.1| GH10818 [Drosophila grimshawi]	FAFEF5V01CI3ND	4.00E-28	260
GL15240 [Drosophila persimilis] gb|EDW31045.1| GL15240 [Drosophila persimilis]	FAFEF5V01CNDXK	9.00E-16	227
GA20772 [Drosophila pseudoobscura pseudoobscura] gb	FAFEF5V01CTM65	3.00E-15	231
GE17244 [Drosophila yakuba] gb|EDX01871.1| GE17244 [Drosophila yakuba]	FAFEF5V01CW66F	2.00E-22	295
GG16295 [Drosophila erecta] gb|EDV52817.1| GG16295 [Drosophila erecta]	FAFEF5V01CX4ZZ	1.00E-26	260
GL13671 [Drosophila persimilis] gb|EDW38942.1| GL13671 [Drosophila persimilis]	FAFEF5V01D61UB	3.00E-13	243
histone 4 [Drosophila takahashii]	FAFEF5V01D6J14	7.00E-14	238
GM12569 [Drosophila sechellia] gb|EDW52931.1| GM12569 [Drosophila sechellia]	FAFEF5V01DAUI1	3.00E-24	261
GE11745 [Drosophila yakuba] gb|EDW92002.1| GE11745 [Drosophila yakuba]	FAFEF5V01DGZ14	5.00E-19	236
GJ24151 [Drosophila virilis] gb|EDW67446.1| GJ24151 [Drosophila virilis]	FAFEF5V01DJCH0	3.00E-28	245
GG16295 [Drosophila erecta] gb|EDV52817.1| GG16295 [Drosophila erecta]	FAFEF5V01DPJ8Z	1.00E-26	260

Table 2-6. continued

GI14567 [Drosophila mojavensis] gb|EDW11330.1| GI14567 [Drosophila mojavensis]	FAFEF5V01DQYCU	8.00E-20	259
GJ22871 [Drosophila virilis] gb|EDW67831.1| GJ22871 [Drosophila virilis]	FAFEF5V01EJ4QM	2.00E-13	266
GG16295 [Drosophila erecta] gb|EDV52817.1| GG16295 [Drosophila erecta]	FAFEF5V01EL6NO	1.00E-26	260
GD25889 [Drosophila simulans] gb|EDX06653.1| GD25889 [Drosophila simulans]	FAFEF5V01EN581	9.00E-19	236
GJ13732 [Drosophila virilis] gb|EDW70344.1| GJ13732 [Drosophila virilis]	FAFEF5V01EP70V	2.00E-15	190
GA14320 [Drosophila pseudoobscura pseudoobscura] gb|EAL27434.2|	FAFEF5V02F3S2H	5.00E-20	292
GK15001 [Drosophila willistoni] gb|EDW75799.1| GK15001 [Drosophila willistoni]	FAFEF5V02FJSQQ	8.00E-15	271
GK13532 [Drosophila willistoni] gb|EDW83714.1| GK13532 [Drosophila willistoni]	FAFEF5V02FT7HS	7.00E-19	243
GF16734 [Drosophila ananassae] gb|EDV42994.1| GF16734 [Drosophila ananassae]	FAFEF5V02G4UPC	5.00E-18	227
GJ15784 [Drosophila virilis] gb|EDW66034.1| GJ15784 [Drosophila virilis]	FAFEF5V02GOFCE	1.00E-22	265
GA11612 [Drosophila pseudoobscura pseudoobscura] ref|XP_002027292.1|	FAFEF5V02GOZPG	3.00E-26	227
GJ19383 [Drosophila virilis] gb|EDW65659.1| GJ19383 [Drosophila virilis]	FAFEF5V02GSAZP	1.00E-16	202
GH16892 [Drosophila grimshawi] gb|EDV97470.1| GH16892 [Drosophila grimshawi]	FAFEF5V02GUMTS	8.00E-15	252
GA28953 [Drosophila pseudoobscura pseudoobscura] gb|EDY70417.1|	FAFEF5V02GW488	3.00E-36	251
GD24357 [Drosophila simulans] gb|EDX15244.1| GD24357 [Drosophila simulans]	FAFEF5V02HB01B	2.00E-16	217
GJ22386 [Drosophila virilis] gb|EDW62059.1| GJ22386 [Drosophila virilis]	FAFEF5V02HEI1U	7.00E-32	250
GL10126 [Drosophila persimilis] gb|EDW32857.1| GL10126 [Drosophila persimilis]	FAFEF5V02HWLTK	2.00E-25	246
GH20881 [Drosophila grimshawi] gb|EDW00675.1| GH20881 [Drosophila grimshawi]	FAFEF5V02IED6Z	5.00E-15	249
GF17183 [Drosophila ananassae] gb|EDV42086.1| GF17183 [Drosophila ananassae]	FAFEF5V02IHEE9	4.00E-15	208
GJ22386 [Drosophila virilis] gb|EDW62059.1| GJ22386 [Drosophila virilis]	FAFEF5V02IJ898	7.00E-32	250
GK13546 [Drosophila willistoni] gb|EDW83686.1| GK13546 [Drosophila willistoni]	FAFEF5V02IZ31E	2.00E-34	249
GF11363 [Drosophila ananassae] gb|EDV37521.1| GF11363 [Drosophila ananassae]	FAFEF5V02JAW0C	5.00E-35	256

Table 2-7. Number of top hits from each organism.

Species	Total Hits	Contigs	Singletons
Pinus	388	218	170
Picea	6646	4537	2109
Zea mays	365	176	189
Oryza sativa	1849	956	893
Arabidopsis	1054	601	453
Populus	0	0	0

Table 2-8. Disease resistance proteins from BLASTx using an e-value of 10^{-12}. Genes are annotated both from contigs and singletons

Query Name	Hit	Description	Length	E-value
CH.11008.C1	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	324	2.00E-26
CH.11030.C1	gb|AAM28908.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	369	6.00E-18
CH.13863.C1	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	260	5.00E-17
CH.14400.C1	gb|AAR36911.1|	disease resistance gene [Pinus sylvestris]	858	8.00E-42
CH.18379.C1	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	255	5.00E-19
CH.20987.C1	gb|AAM28907.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	296	9.00E-17
CH.21925.C1	gb|AAM28908.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	838	3.00E-28
CH.22894.C1	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	227	4.00E-14
CH.28879.C1	gb|AAM28908.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	438	1.00E-13
CH.3023.C1	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	369	3.00E-22
CH.4518.C5	gb|AAV34188.1|	disease resistance associated protein [Picea abies]	1250	1.00E-84
CH.4518.C6	gb|AAV34188.1|	disease resistance associated protein [Picea abies]	1181	1.00E-111
FAFEF5V01BOHD9	gb|AAV34188.1|	disease resistance associated protein [Picea abies]	273	1.00E-39
FAFEF5V01BT73A	gb|AAV34188.1|	disease resistance associated protein [Picea abies]	248	7.00E-30
FAFEF5V01CN0L9	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	271	9.00E-14
FAFEF5V01DK1L2	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	252	2.00E-16
FAFEF5V01E0GRP	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	267	8.00E-28
FAFEF5V02F5JCE	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	245	4.00E-14
FAFEF5V02F8VD2	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	239	2.00E-13
FAFEF5V02HSU6E	gb|AAV34188.1|	disease resistance associated protein [Picea abies]	248	8.00E-18
FAFEF5V02IY1WA	gb|AAM28917.1|	putative TIR/NBS/LRR disease resistance protein [Pinus taeda]	264	2.00E-15

Table 2-9. Pathogenesis related Proteins from BLASTx using an e-value of 10^{-12}. Genes are annotated both from contigs and singletons.

Description	Query Name	Length(bp)	e-value
putative intracellular pathogenesis-related protein [Picea glauca]	CH.13601.C1	507	1.00E-19
pathogenesis-related protein 10-2.1 [Pinus monticola] gb	CH.14254.C1	760	3.00E-75
putative intracellular pathogenesis-related protein [Picea mariana]	CH.26919.C1	244	3.00E-34
nonexpresser of pathogenesis-related 1 [Gossypium hirsutum]	CH.28746.C1	543	3.00E-52
pathogenesis-related protein 10-3.3-like [Picea glauca]	contig31123	481	3.00E-14
pathogenesis associated protein Cap20	E9M19J301DFXU9	242	
Os07g0128800 [Oryza sativa (japonica cultivar-group)] dbj|	FAFEF5V01DDAL7	185	9.00E-14
nonexpresser of pathogenesis-related 1 [Gossypium hirsutum]	FAFEF5V01E1UHG	226	6.00E-13
pathogenesis associated protein Cap20	FAFEF5V02GFVLY	236	

Figure 2-3. A pie chart representation of various GO categories. Represented as molecular function (F) cellular component (C) and biological processes.(P)

Table 2-10. A summary of all the ESTs annotated under biological processes

Biological process	Genes represented	% of total
Electron transport or energy pathways	3	0.27
Transcription	19	1.74
Other metabolic processes	25	2.28
DNA or RNA metabolism	29	2.65
Signal transduction	34	3.11
Cell organization and biogenesis	45	4.12
Response to abiotic or biotic stimulus	46	4.20
Response to stress	56	5.12
Other biological processes	79	7.25
Protein metabolism	80	7.31
Developmental processes	132	12.05
Other cellular processes	441	40.27
Transport	106	9.68

Table 2-1. Cellular components represented as a percent of the total. All unknown cellular components were collapsed into the other cellular category.

Cellular Component	Genes represented	% of total
Chloroplast	1	0.40
Plasma membrane	1	0.40
Cell wall	2	0.81
Extracellular	4	1.62
Ribosome	4	1.62
ER	7	2.83
Cytosol	9	3.64
Mitochondria	12	4.86
Other cellular components	13	5.26
Golgi apparatus	13	5.262
Plastid	30	12.16
Other membranes	32	12.99
Other cytoplasmic Components	37	14.98
Other intracellular Components	40	16.19
Nucleus	42	17.06

Table 2-12. Molecular functions were collapsed into 14 categories which including but limited to RNA binding among other binding activities and kinase activities.

Molecular function	Genes represented	% of Total
Transcription factor activity	1	0.11
Structural molecule activity	4	0.44
Receptor binding or Activity	10	1.10
Nucleic acid binding	14	1.55
DNA or RNA binding	22	2.43
Protein binding	33	3.64
Other molecular functions	40	4.42
Other binding	57	6.29
Kinase activity	58	6.40
Transporter activity	101	11.15
Hydrolase activity	130	14.35
Transferase activity	168	18.54
Other enzyme activity	268	29.58

Table 2-13. Stress response genes based on GO annotation from TAIR

GO ID	Description	Molecular Function
GO:0000302	response to reactive oxygen species	response to stress
GO:0000303	response to superoxide	response to stress
GO:0000304	response to singlet oxygen	response to stress
GO:0001666	response to hypoxia	response to stress
GO:0002213	defense response to insect	response to stress|response to abiotic or biotic stimulus
GO:0006950	response to stress	response to stress
GO:0006952	defense response	response to stress
GO:0006970	response to osmotic stress	response to stress|response to abiotic or biotic stimulus
GO:0006974	response to DNA damage stimulus	response to stress
GO:0006979	response to oxidative stress	response to stress
GO:0006987	activation of signaling protein activity involved in unfolded protein response	response to stress|signal transduction
GO:0006995	cellular response to nitrogen starvation	response to stress|other cellular processes
GO:0009267	cellular response to starvation	response to stress|other cellular processes
GO:0009269	response to desiccation	response to stress|response to abiotic or biotic stimulus
GO:0009408	response to heat	response to stress|response to abiotic or biotic stimulus
GO:0009409	response to cold	response to stress|response to abiotic or biotic stimulus
GO:0009413	response to flooding	response to stress|response to abiotic or biotic stimulus
GO:0009414	response to water deprivation	response to stress|response to abiotic or biotic stimulus
GO:0009432	SOS response	response to stress|other cellular processes
GO:0009611	response to wounding	response to stress
GO:0009616	virus induced gene silencing	response to stress|response to abiotic or biotic stimulus
GO:0009626	plant-type hypersensitive response	response to stress|other cellular processes
GO:0009627	systemic acquired resistance	response to stress|response to abiotic or biotic stimulus
GO:0009631	cold acclimation	response to stress|response to abiotic or biotic stimulus
GO:0009651	response to salt stress	response to stress|response to abiotic or biotic stimulus
GO:0009695	jasmonic acid biosynthetic process	response to stress|other cellular processes|
GO:0009814	defense response, incompatible interaction	response to stress|response to abiotic or biotic stimulus

Table 2-13 continued

GO:0009816	defense response to bacterium, incompatible interaction	response to stress|response to abiotic or biotic stimulus
GO:0009817	defense response to fungus, incompatible interaction	response to stress|response to abiotic or biotic stimulus
GO:0009861	jasmonic acid and ethylene-dependent systemic resistance	response to stress|response to abiotic or biotic stimulus
GO:0009862	systemic acquired resistance, salicylic acid mediated signaling pathway	
GO:0009864	induced systemic resistance, jasmonic acid mediated signaling pathway	
GO:0009870	defense response signaling pathway, resistance gene-dependent	response to stress|signal transduction
GO:0009871	jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signaling pathway	
GO:0009970	cellular response to sulfate starvation	response to stress|other cellular processes
GO:0010048	vernalization response	response to stress|response to abiotic or biotic stimulus
GO:0010186	positive regulation of cellular defense response	response to stress
GO:0010204	defense response signaling pathway, resistance gene-independent	response to stress|signal transduction
GO:0010286	heat acclimation	response to stress|response to abiotic or biotic stimulus
GO:0016036	cellular response to phosphate starvation	response to stress|other cellular processes
GO:0030968	endoplasmic reticulum unfolded protein response	
GO:0031347	regulation of defense response	response to stress
GO:0031348	negative regulation of defense response	response to stress
GO:0042149	cellular response to glucose starvation	response to stress|other cellular processes
GO:0042538	hyperosmotic salinity response	response to stress|response to abiotic or biotic stimulus
GO:0042542	response to hydrogen peroxide	response to stress
GO:0042594	response to starvation	response to stress
GO:0042631	cellular response to water deprivation	
GO:0042742	defense response to bacterium	response to stress|response to abiotic or biotic stimulus
GO:0042744	hydrogen peroxide catabolic process	
GO:0043619	regulation of transcription from RNA polymerase II promoter in response to oxidative stress	
GO:0045087	innate immune response	response to stress

Table 2-13 continued

GO:0050826	response to freezing	response to stress|response to abiotic or biotic stimulus
GO:0050832	defense response to fungus	response to stress|response to abiotic or biotic stimulus
GO:0051607	defense response to virus	response to stress|response to abiotic or biotic stimulus
GO:0051788	response to misfolded protein	response to stress|response to abiotic or biotic stimulus
GO:0052542	callose deposition during defense response	response to stress

CHAPTER 3
CONCLUSION

This project involved using EST discovery through highthrough put sequencing. ESTs can be used to generate genetic markers to enhance breeding through faster selection means or also for transgenic hemlock trees. From pooled infested and non infested Carolina leaves and branches, a CDNA library was synthesized and normalized using clontech SMART and Evrogen protocol. A complete sequencing run, consisting of 3 production and a titration run generated 128,819 ESTs covering 77.9Mbp. These sequences assembled into 93,454 singletons and 35,761 contigs. BLASTx was used to annotate these EST sequences. At an e value of 10^{-4}, 50,036 EST's were annotated against the nonredundant database. The average length of the hits was 309.5 bp which was higher than the average read length for each sequence read acquired through 454 (Margulies, Egholm et al. 2005) Using a stringent e value of 10^{-12}, ESTs was analyzed for contigs and singleton length. The annotated ESTs for the singletons were greatly reduced, when the e value was changed from $10^{-4}$to 10^{-12}. At this e value, BLASTx results were also analyzed for genes that were annotated as disease resistance proteins and or pathogen related protein. Some of these disease resistant proteins contained TIR/NBS/LLR domains, which have been shown to be highly involved in responses to pathogen responses and therefore may provide a good guide in the analysis of hemlock in response to adelgid feeding (Tameling, Elzinga et al. 2002) Other disease and stress responsive genes were discovered and in future experiments will be used to assist in determining the genetic groundwork for resistance in hemlock.

BLAST against the TAIR database was performed to identify Gene Ontologies, (GO). Of all the 2, 255 genes with assigned GO annotations most genes were annotated under biological processes. We discovered some genes responsible for stress response both biotic and abiotic. Another interesting finding was identification of 20 disease resistant proteins containing

The ESTs that had hits were used to design 49,549 probes for gene expression analysis using microarray for future quantitative analyses of gene expression.

APPENDIX
GO ANNOTATION TABLES

Table A-1 At e-value e-4 these are the genes that were assigned roles in stress response.

GO ID	GO Description	Biological Processs
GO:0000302	response to reactive oxygen species	response to stress
GO:0000303	response to superoxide	response to stress
GO:0000304	response to singlet oxygen	response to stress
GO:0001666	response to hypoxia	response to stress
GO:0002213	defense response to insect	response to stress|response to abiotic or biotic stimulus
GO:0006950	response to stress	response to stress
GO:0006952	defense response	response to stress
GO:0006970	response to osmotic stress	response to stress|response to abiotic or biotic stimulus
GO:0006974	response to DNA damage stimulus	response to stress
GO:0006979	response to oxidative stress	response to stress
GO:0006987	activation of signaling protein activity involved in unfolded protein response	response to stress|signal transduction|other cellular processes
GO:0006995	cellular response to nitrogen starvation	response to stress|other cellular processes
GO:0009267	cellular response to starvation	response to stress|other cellular processes
GO:0009269	response to desiccation	response to stress|response to abiotic or biotic stimulus
GO:0009408	response to heat	response to stress|response to abiotic or biotic stimulus
GO:0009409	response to cold	response to stress|response to abiotic or biotic stimulus
GO:0009413	response to flooding	response to stress|response to abiotic or biotic stimulus
GO:0009414	response to water deprivation	response to stress|response to abiotic or biotic stimulus
GO:0009432	SOS response	response to stress|other cellular processes
GO:0009611	response to wounding	response to stress
GO:0009616	virus induced gene silencing	response to stress|response to abiotic or biotic stimulus
GO:0009626	plant-type hypersensitive response	response to stress|other cellular processes
GO:0009627	systemic acquired resistance	response to stress|response to abiotic or biotic stimulus
GO:0009631	cold acclimation	response to stress|response to abiotic or biotic stimulus
GO:0009651	response to salt stress	response to stress|response to abiotic or biotic stimulus
GO:0009695	jasmonic acid biosynthetic process	response to stress|other cellular processes|other metabolic processes|
GO:0009814	defense response, incompatible interaction	response to stress|response to abiotic or biotic stimulus
GO:0009816	defense response to bacterium, incompatible interaction	response to stress|response to abiotic or biotic stimulus
GO:0009817	defense response to fungus, incompatible interaction	response to stress|response to abiotic or biotic stimulus
GO:0009861	jasmonic acid and ethylene-dependent systemic resistance	response to stress|response to abiotic or biotic stimulus
GO:0009862	systemic acquired resistance, salicylic acid mediated signaling pathway	response to stress|signal transduction|response to abiotic or biotic stimulus

Table A-1. Continued

GO:0009864	induced systemic resistance, jasmonic acid mediated signaling pathway	response to stress|signal transduction|response to abiotic or biotic stimulus
GO:0009870	defense response signaling pathway, resistance gene-dependent	response to stress|signal transduction
GO:0009871	jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signaling pathway	response to stress|signal transduction|response to abiotic or biotic stimulus
GO:0009970	cellular response to sulfate starvation	response to stress|other cellular processes
GO:0010048	vernalization response	response to stress|response to abiotic or biotic stimulus
GO:0010186	positive regulation of cellular defense response	response to stress
GO:0010204	defense response signaling pathway, resistance gene-independent	response to stress|signal transduction
GO:0010286	heat acclimation	response to stress|response to abiotic or biotic stimulus
GO:0016036	cellular response to phosphate starvation	response to stress|other cellular processes
GO:0030968	endoplasmic reticulum unfolded protein response	response to stress|signal transduction|other cellular processes|response to abiotic or biotic stimulus
GO:0031347	regulation of defense response	response to stress
GO:0031348	negative regulation of defense response	response to stress
GO:0042149	cellular response to glucose starvation	response to stress|other cellular processes
GO:0042538	hyperosmotic salinity response	response to stress|response to abiotic or biotic stimulus
GO:0042542	response to hydrogen peroxide	response to stress
GO:0042594	response to starvation	response to stress
GO:0042631	cellular response to water deprivation	response to stress|other cellular processes|response to abiotic or biotic stimulus
GO:0042742	defense response to bacterium	response to stress|response to abiotic or biotic stimulus
GO:0042744	hydrogen peroxide catabolic process	response to stress|other cellular processes|other metabolic processes
GO:0043619	regulation of transcription from RNA polymerase II promoter in response to oxidative stress	transcription|response to stress|other cellular processes|other metabolic processes
GO:0045087	innate immune response	response to stress
GO:0050826	response to freezing	response to stress|response to abiotic or biotic stimulus
GO:0050832	defense response to fungus	response to stress|response to abiotic or biotic stimulus
GO:0051607	defense response to virus	response to stress|response to abiotic or biotic stimulus
GO:0051788	response to misfolded protein	response to stress|response to abiotic or biotic stimulus
GO:0052542	callose deposition during defense response	response to stress

Object A-1. Summary of genes collectively annotated under biological processes

Object A-2. Summary of genes collectively annotated under cellular components

Object A-3. List of genes annotated under molecular functions

LIST OF REFERENCES

A., H.G., and Georg, J. (2008). Plant Immunity to Insect Herbivores. Annual Review of Plant Biology **59,** 41-66.

Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L., and Schnable, P.S. (2007). SNP discovery via 454 transcriptome sequencing. Plant J **51,** 910 - 918.

Bentz, S.E., Riedel, L.G.H., Pooler, M., and Townsend, A.M. (2001). Hybridization and self-compatibility in controlled Pollination of Eastern North American and Asian Hemlock (Tsuga)Species. Journal of Arboriculture **28,** 200-204.

Chang, S., Puryear, J., and Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. Plant Molecular Biology Reporter **11,** 113-116.

Cheah Carole , M.M., Salem Scott, Bruce Parke, Skinner Margaret and Costa Scott (2004). Biological Control Of Hemlock Woolly Adelgid.

Chen, Z., Kolb, T.E., and Clancy, K.M. (2002). The Role of Monoterpenes in Resistance of Douglas Fir to Western Spruce Budworm Defoliation. Journal of Chemical Ecology **28,** 897-920.

Cheung, F., Haas, B., Goldberg, S., May, G., Xiao, Y., and Town, C. (2006a). Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. BMC Genomics **7,** 272.

Cheung, F., Haas, B.J., Goldberg, S.M., May, G.D., Xiao, Y., and Town, C.D. (2006b). Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. BMC Genomics **7,** 272.

Emrich, S.J., Barbazuk, W.B., Li, L., and Schnable, P.S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res **17,** 69 - 73.

Farjon, A. (1990). Pinaceae. Drawings and Descriptions of the Genera Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix, and Picea. Koeltz Scientific Books **121,** 330.

Fossdal, Gunnar, C., Praveen, S., and Anders, L. (2001). Isolation of the first putative peroxidase cDNA from a conifer and the local and systemic accumulation of related proteins upon pathogen infection Plant Molecular Biology **47,** 423-435

Franceschi, V.R., Krokene, P., Christiansen, E., and Krekling, T. (2005). Anatomical and Chemical Defenses of Conifer Bark against Bark Beetles and Other Pests. New Phytologist **167,** 353-375.

Garcia-Reyero, N., Griffitt, R.J., K., L.L., W., J.K., Farmerie, G., Barber, D.S., and Denslow, N.D. (2008). Construction of a robust microarray from a non-model species

largemouth bass, *Micropterus salmoides* (Lacèpede), using pyrosequencing technology. Journal of Fish Biology **72,** 2354-2376.

Gen-ichiro, A., Dezene, P.W.H., and Jörg, B. (2004). Forest tent caterpillars (*Malacosoma disstria*) induce local and systemic diurnal emissions of terpenoid volatiles in hybrid poplar (*Populus trichocarpa*): cDNA cloning, functional characterization, and patterns of gene expression of -germacrene D synthase. The Plant Journal **37,** 603-616.

Hahn, D., Ragland, G., Shoemaker, D., and Denlinger, D. (2009). Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly Sarcophaga crassipalpis. BMC Genomics **10,** 234.

Havill, N.P., Montgomery, M.E., Yu, G., Shiyake, S., and Caccone, A. (2006). Mitochondrial DNA from Hemlock Woolly Adelgid (Hemiptera: Adelgidae) Suggests Cryptic Speciation and Pinpoints the Source of the Introduction to Eastern North America. Annals of the Entomological Society of America **99,** 195-203.

Hilder, V.A., and Boulter, D. (1999). Genetic engineering of crop plants for insect resistance - a critical review. Crop Protection **18,** 177-191.

J. Cristobal Vera, C.W.W.H.W.F.M.J.F.D.L.C.I.H.J.H.M. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology **17,** 1636-1647.

Johnson, R. (Mar 2000.). Classical plant breeding for durable resistance to diseases

Johnson, R. Journal of Plant Pathology **82,** 3-7.

Johnson R. (1983). Genetic background of durable resistance. In NATO Advanced Study Institute Series. Series A, Life Science . 5-26.

Jones-Rhoades, M.W., Borevitz, J.O., and Preuss, D. (2007). Genome-Wide Expression Profiling of the Arabidopsis Female Gametophyte Identifies Families of Small, Secreted Proteins. PLoS Genet **3,** e171.

Jørgense, I.H. (1992). Discovery, characterization and exploitation of Mlo powdery mildew resistance in barley. Euphytica **63,** 141-152.

Kessler, A., and Baldwin, I.T. (2002). Plant Response to Insect Herbivory: The Emerging Molecular Analysis. Annual Review of Plant Biology **53,** 299-328.

Lagalante, A.F., and Montgomery, M.E. (2003). Analysis of Terpenoids from Hemlock (Tsuga) Species by Solid-Phase Microextraction/Gas Chromatography/Ion-Trap Mass Spectrometry. Journal of Agricultural and Food Chemistry **51,** 2115-2120.

Lin, W., Anuratha, C.S., Datta, K., Potrykus, I., Muthukrishnan, S., and Datta, S.K. (1995). Genetic Engineering of Rice for Resistance to Sheath Blight. Nat Biotech **13,** 686-691.

M, D.j., Wu, H., Janice, C., E. K.; Reed, M.Luce, J., Scott, K., and Michler, C.H. (2002). Pathogen challenge, salicylic acid, and jasmonic acid regulate expression of chitinase gene homologs in pine. Molecular Plant-Microbe Interactions **14,** 380-387.

M.McClure. (2001). Biological control of hemlock woolly adelgid in the Eastern

United States. U.S. Department of Agriculture Forest Service, Morgantown, WV,

FHTET-2000-08**,** 10.

Mardis, E.R. (2008a). Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics **9,** 387-402.

Mardis, E.R. (2008b). The impact of next-generation sequencing technology on genetics. Trends in Genetics **24,** 133-141.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature **437,** 376 - 380.

McClure, M. (1987). Biology and control of hemlock woolly adelgid. Bulletin, Connecticut

Agricultural Experiment Station**,** 9.

McClure, M., and Shields, S.M. (2001). Hemlock woolly adelgid. U.S.

Department of Agriculture Forest Service, Morgantown, WV, FHTET-2001-03**,** 14p.

McClure, M.S. (1990). Role of wind, birds, deer, and humans in the dispersal of hemlock woolly adelgid (Homoptera: Adelgidae). Pests of plants;Forestry - General aspects **19,** 36-43.

McClure, M.S. (1992). Hemlock woolly adelgid. Am. Nurseryman**,** 82-86.

Meyer, M., Stenzel, U., Myles, S., Prufer, K., and Hofreiter, M. (2007). Targeted high-throughput sequencing of tagged nucleic acid samples. Nucl. Acids Res. **35,** e97-.

Moran, P.J., and Thompson, G.A. (2001). Molecular Responses to Aphid Feeding in Arabidopsis in Relation to Plant Defense Pathways. Plant Physiol. **125,** 1074-1085.

Neil, H. (2007). Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol **210,** 1518-1525.

Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Paabo, S., Pritchard, J.K., and Rubin, E.M. (2006). Sequencing and Analysis of Neanderthal Genomic DNA. Science **314,** 1113-1118.

Novaes, E., Drost, D., Farmerie, W., Pappas, G., Grattapaglia, D., Sederoff, R., and Kirst, M. (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics **9,** 312.

Phillips, M.A., and Croteau, R.B. (1999). Resin-based defenses in conifers. Trends in Plant Science **4,** 184-190.

Phillips MA, C.R. (1994). Resin-based defenses in conifers. Trends Plant Science**,** 184-190.

Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R.D.E., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., Rampp, M., Miller, W., and Schuster, S.C. (2006). Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. Science **311,** 392-394.

Regina Preisig-müller, A.S.I.B., Hans-Jörg Reif and Helmut Kindl. (1999). Characterization of a pine multigene family containing elicitor- responsive stilbene synthase genes Plant Molecular Biology **39**.

Reymond, P., Bodenhausen, N., Van Poecke, R.M.P., Krishnamurthy, V., Dicke, M., and Farmer, E.E. (2004). A Conserved Transcript Pattern in Response to a Specialist and a Generalist Herbivore. Plant Cell **16,** 3132-3147.

Richter, T.E. (2001). The evolution of disease resistance genes. Plant Molecular Biology **42,** 195-204.

Roder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.-H., Leroy, P., and Ganal, M.W. (1998). A Microsatellite Map of Wheat. Genetics **149,** 2007-2023.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Analytical Biochemistry **242,** 84-89.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America **74,** 5463-5467.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat Meth **5,** 16-18.

Skinner, M., Parker, B.L., Gouli, S., and Ashikaga, T. (2003). Regional Responses of Hemlock Woolly Adelgid (Homoptera: Adelgidae) to Low Temperatures. Environmental Entomology **32,** 523-528.

Snyder Craig D. , Y.J.A., . Lemarié David P, and Smith David R. (2002). Influence of eastern hemlock (Tsuga canadensis)forests on aquatic invertebrate assemblages in headwater streams. Canadian Journal of Fisheries and Aquatic Sciences **59,** 262–275.

Sunnucks, P. (2000). Efficient genetic markers for population biology. Trends in Ecology & Evolution **15,** 199-203.

Tameling, W.I.L., Elzinga, S.D.J., Darmin, P.S., Vossen, J.H., Takken, F.L.W., Haring, M.A., and Cornelissen, B.J.C. (2002). The Tomato R Gene Products I-2 and Mi-1 Are Functional ATP Binding Proteins with ATPase Activity. Plant Cell **14,** 2929-2939.

Tsuji, R., Fischer, A.J., Yoshino, M., Roel, A., Hill, J.E., and Yamasue, Y. (2003). Herbicide-Resistant Late Watergrass (Echinochloa phyllopogon): Similarity in Morphological and Amplified Fragment Length Polymorphism Traits. Weed Science **51,** 740-747.

Varshney, R.K., Graner, A., and Sorrells, M.E. (2005). Genic microsatellite markers in plants: features and applications. Trends in Biotechnology **23,** 48-55.

Walling, L.L. (2000). The Myriad Plant Responses to Herbivores. Journal of Plant Growth Regulation **19,** 195-216.

Wilcox, P.L., Amerson, H.V., Kuhlman, E.G., Liu, B.H., O'Malley, D.M., and Sederoff, R.R. (1996). Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proceedings of the National Academy of Sciences of the United States of America **93,** 3859-3864.

Young, R.F., Sheilds, K.S., and Berlyn, G.P. (1995). Hemlock Woolly Adelgid (homoptera:Adelgidae):Stylet Bundle Insertion and Feeding Sites. Annals of the Entomological Society of America **88,** 827-835.

Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Vagner, L.L., Khaspekov, G.L., Kozhemyako, V.B., Matz, M.V., Meleshkevitch, E., Moroz, L.L., Lukyanov, S.A., and Shagin, D.A. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucl. Acids Res. **32,** e37-.

Zou, Z., Najar, F., Wang, Y., Roe, B., and Jiang, H. (2008). Pyrosequence analysis of expressed sequence tags for Manduca sexta hemolymph proteins involved in immune responses. Insect Biochemistry and Molecular Biology **38,** 677-682.

BIOGRAPHICAL SKETCH

Anne Mwaniki was born in Nairobi, Kenya. She moved to the United States of America in January 2001, where she completed her high school education at Chapel Hill High School, Chapel Hill, NC. She then attended North Carolina State University where she majored in biochemistry. During her undergraduate studies she worked in various USDA laboratories analysis soybeans for phytic acid and a pilot project for breeding wheat resistant to Fusarium rust. Upon graduation, she joined the University of Florida to acquire her Master of Science in plant molecular and cellular biology while working on differential gene expression in *Tsuga* in response to woolly adelgid infestation under the supervision of Dr. Gary Peter.