DNA Fingerprinting of *Castanea* Species in the USA

by

Xiaowei Li

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
December 12, 2011

Keywords: *Castanea*, species-specific marker, cpDNA,
phylogenetics, 454 sequencing,

Approved by

Fenny Dane, Chair, Professor of Horticulture
Elina Coneva, Associate Professor of Horticulture
Leslie R. Goertzen, Associate Professor of Biological Sciences

Abstract

Two *Castanea* species (*C. dentata*, the American chestnut, and *C. pumila*, var. *pumila*, the Allegheny chinkapin, and var. *ozarkensis*, the Ozark chinkapin) are native to the USA. It has been difficult to differentiate the species based on morphological characters because of intra-specific variability and the incidence of chestnut blight, which has prevented trees from maturing to the point of flower and fruit production. To develop species-specific markers and to infer historical processes associated with the geographical distribution of plant populations, chloroplast DNA, nuclear DNA and 454 sequences were generated, with special emphasis on one *Castanea* population at Ruffner Mountain, Alabama.

The Ruffner Mountain *Castanea* tree population was analyzed based on leaf morphology, and sequencing analysis using several chloroplast DNA regions (*trn*T-L, *trn*L, *ndh*F, *ndh*C, *orf*62, and *rpL*16), and two informative nuclear regions (17, and 126). Comparative analysis with *C. dentata*, *C. pumila* var. *pumila* and *C. pumila* var. *ozarkensis* populations was conducted to infer the biogeographic history of the AL population. A total of 5 cpDNA haplotypes were detected at the Ruffner Mountain population, which can be used to divide the population into two main groups: *C. dentata* type and *C. pumila* var. *pumila* type group. Some mutational sites (two deletions at *trn*T-L region, one indel at *ndh*F region, one deletion at region *ndh*C, one SNP in region *rpl*16, one SNP in nuclear region 17 and one SNP in nuclear region 126) can be considered as species-specific markers to varying degrees. However, species identification had better be based on morphology and combined sequence analyses. Phylogenetic analyses of the cpDNA data provided some evidence of the relationship among samples from different *Castanea* populations in North America, Moreover, the phylogenetic analyses of the nuclear data showed the possible origin of hybrid taxa.

To obtain more species-specific markers, cDNA from leaves of 5 individual *C. pumila* trees was isolated and sequenced using the 454 GS-FLX at the Genomics Core Facilities of Penn State University. A total of 1221540 reads, about 372 Mb of cDNA, was generated. The read length is between 36-603 bp, with an average length of 305 bp. A total of 47565 contigs and 77547 singletons, and 125112 unigenes were obtained from the 454 sequencing analyses. Through alignment of the individual reads against contigs from the assembly, 143792 SNPs were detected in the contigs, with average length of 222 bp per SNP. The proportion of transition nucleotide substitutions (29273, 21%) is much less than the proportion of transversions (109775, 78.9%). In addition, there are 2415 complex SNPs (variations of more than two nucleotides). Upon alignment of *C. dentata* and *C. pumila* contigs using Model SNP of the CLC genomic workbench software, a total 267874 inter-SNPs were detected. Nineteen contigs with possible species-specific markers were analyzed and two were preliminary validated. Multi-alignment of the *C. pumila* and *C. dentata* contigs with 3714 Arabidopsis single copy genes was conducted. Contigs from both species with a good match to a single copy gene were selected and re-aligned. Ten possible species-specific marker sites were examined and two showed species-specificity. More species-specific markers can be obtained this way. Gene ontology analysis of the *C. pumila* assembly showed high similarities to transcriptomes of other *Castanea* species with known genome sequences in the NCBI database.

Acknowledgments

The author would like to thank the members of his research committee Dr. Fenny Dane, Dr. Elina Coneva, and Dr. Leslie R. Goertzen for their constructive advice and complete support during his research. He appreciates his lab members Ms. Zhuoyu Wang, Ms. Delaine Borden who always gave their help during his research. His deepest gratefulness goes to his parents and family members, who always give their support and encouragement. And finally, special thanks are extended to his wife Shuxiu Wang and his son Yihan Li for their unending support, love and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Literature Review

Systematics of *Castanea*

Castanea Miller (Fagaceae), the genus of the chestnuts and chinkapins, which contains 7 species (Johnson, 1988), is widely distributed in the temperate forest of the Northern Hemisphere. In Japan and the Korean peninsula, the Japanese chestnut (*C. crenata* Sieb & Zucc.) has been cultivated as an important food and timber tree for at least 1000 years. Morphologically, the Japanese chestnut appears to be most closely related to the European chestnut (*C. sativa* Mill.) (Jaynes, 1975). The Chinese chestnut (*C. mollissima* Blume) is commercially the most important chestnut species because of its large size and good quality. China currently produces almost half of the world's chestnut. The Chinese chestnut is cultivated in 22 provinces and more than 300 cultivars have been recognized in China. Seguin chestnut (*C. seguinii* Dode), is commercially less important because of its small nut size. Seguin chestnut occurs sympatrically with Chinese chestnut over central and southwest China (Dane et al., 2003). *Castanea henryi*, sometimes called the Chinese chinkapin because it has a single nut per cupule, is indigenous to southeastern and southwestern China (Xu, 2005). This species is also commercially less important because of its small nut size, but is good for timber. Sweet chestnut (*C. sativa*) is the only *Castanea* species native to Europe. It is currently widespread throughout Europe and South-west Asia as one of the multipurpose and most economically important tree species for the Mediterranean region (Casasoli et al., 2001).

There are two species of *Castanea* in North America. The American chestnut (*C. dentata* Marsh.), was one of the leading tree species in the temperate deciduous forests of the

Appalachian Mountains and was regarded by some to be a keystone species. Its range covered forests from Maine to Mississippi, including West Virginia, Kentucky, Tennessee, Indiana, extreme southeastern Michigan, Arkansas, and Missouri (Stilwell et al., 2003). The species was an important food source for wildlife and a valuable timber crop. The American chinkapin (*C. pumila* L.) has two varieties, var. *pumila* and *var. ozarkensis.* *C. pumila var. pumila*, the Allegheny chinkapin, is endemic to the eastern and southeastern United States from Pennsylvania to Florida, eastern Texas, south western Missouri, and west-central Kentucky. It grows on usually disturbed sites from near sea level to about 1400 m in altitude (Johnson, 1988; Dane et al., 1999). *C. pumila* var. *ozarkensis*, the Ozark chinkapin, is restricted to the Ozark Plateau in Arkansas, where it persists mainly as trees or stump sprouts of various sizes (Johnson, 1988). The Allegheny and Ozark chinkapin differ in reproductive and vegetative characters, as well as leaf micromorphology and flavonoid constituents (Johnson, 1988; Dane et al., 1999).

Unfortunately, the American chinkapin as well as American chestnut are considered endangered species due to a devastating disease, chestnut blight, caused by the bark-inhabiting canker fungus (*Cryphonectria parasitica*), which was introduced from Asia in the late 19th Century (Milgroom et al., 1996). The disease has reduced the American chestnut from an important timber and nut producing tree to an understory shrub (Anagnostakis, 1987; Kubisiak et al., 1997; Dane et al., 1999). The blight was first discovered in New York in 1904, where it spread rapidly across the range of chestnut, and within 50 years had converted this stately tree to a rarely flowering understory shrub across 3.6 million ha of chestnut forest (Anagnostakis, 1987). The American chinkapin, which is closely related to the American chestnut, has almost been extirpated from several states in the southern America (Paillet, 1993).

All species in genus *Castanea* are diploid (2n=2x=24) and hybridize freely, but some interspecific F_1’s usually suffer from low seed germination and male sterility (Jaynes, 1975). There are significantly different levels of resistance to chestnut blight among these species. In Asia, Chinese chestnut species exhibit the highest levels of resistance, while the Japanese chestnut species is also resistant, but less than the Chinese species. The European and North American species are susceptible (Burnham, 1988; Huang et al., 1996; Kubisiak et al., 1997). Although the deadly fungus cannot attack the root systems of these species, it destroys the shoots-the new life form underneath the forest floor. Almost always, the fungus kills the shoots before the trees flower. So reforestation can’t occur before the disease is controlled.

From the early 1950s, many methods have been used to eradicate or to control this devastating disease. But none of these approaches have been effective in orchards or forests (Griffin, 2008). The American Chestnut Foundation (TACF), a non-profit organization was founded in 1983 by a group of prominent plant scientists. It aims to recover the American chestnut tree via an intensive backcross breeding program using the Chinese chestnut (*C. mollissima*) species as a source of blight resistance (http://www.acf.org).

Phylogeography

The word “phylogeography” was coined in 1987 (Avise et al., 1987). It is a relatively new discipline that deals with spatial arrangements of genetic lineages, especially within and among closely related species. The ease and high availability of DNA sequence data has made it possible to resolve high levels of phylogenetic relationships and also to examine intraspecific variation, especially as it relates to geography (Avise, 2009). Phylogeographic research of tree species has been significantly based on amplification of specific chloroplast (cp) DNA loci with universal primers.

Chloroplasts are small organelles within the plant cell that contain the entire machinery necessary for the process of photosynthesis, with much of the size variation attributable to the extent of sequence reiteration in a large inverted repeat region. The chloroplast genome is a circular chromosome of 120~220 kb that consists of two inverted repeats (IRa and IRb), a large single-copy region (LSC), and small single copy region (SSC). Chloroplasts are considered to have originated from cyanobacteria through endosymbiosis, because the plastids share a common ancestor with modern cyanobacteria based on comparison of ribosomal RNA sequences from organelles and certain free-living prokaryotes (Ozeki and Umesonok, 1989). Each chloroplast contains one or more molecules of small circular DNA with genes coding for ribosomal and transfer RNAs plus numerous polypeptides involved in protein synthesis and photosynthesis, and components for the maintenance of their own genetic system. However most of its proteins are encoded by genes contained in the host cell nucleus, with the protein products transported to the chloroplast (Sugiura, 1992).

The chloroplast genome is considered to be conserved in its evolution, and varies little in size, structure and gene content among angiosperms (Olmstead and Palmer, 1994). Moreover, there are several advantages to using cpDNA for taxonomy and evolutionary research: (1) small size, high copy number and simple structure; (2) unlike the mitochondrial and most nuclear genomes, cpDNA gene content and arrangement are more conserved, so it is easier to design primers and clone genes; (3) no genetic reassortment that interferes with molecular phylogenetic relationships because cpDNA is maternally inherited in most species (Tian and Li, 2002). The conserved sequences can be used as heterologous hybridization probes and polymerase chain reaction (PCR) primers in species can facilitate cloned cpDNA from each species under study (Olmstead and Palmer, 1994). Typical cpDNA primers have been constructed on the basis of conserved sequences of chloroplast genes and used to amplify the DNA located between the primer binding sites (Taberlet et al., 1991).

Several recent studies have used the strategy of coupling relatively quickly evolving chloroplast DNA sequences with a phylogeographic sampling scheme to illuminate complex evolutionary histories in plants (Shaw and Small 2005; Chiang et al., 2004; Burban and Petit, 2003). Through analysis of data from the chloroplast genome, many relationships can be resolved based on the characters of evolution, from the level of species and genus to family and even higher levels (Provan et al., 2001; Olmstead and Palmer, 1994; Wagner et al., 1987). Lang et al. (2006, 2007) showed that cpDNA sequences can be used successfully to study systematic relationships within the genus *Castanea*.

An understanding of postglacial colonization routes has come from the analysis of chloroplast DNA variation (King and Ferris 1998; Petit et al., 2002), providing a seed-specific marker derived from seed dispersal which is not blurred by pollen flow. The routes of seed dispersal therefore can be inferred from the geographical distribution of cpDNA variation. Some projects used cpDNA markers to study of population history in trees (Walter and Epperson, 2001; Rendell and Ennos, 2003; Cheng et al., 2005). Phylogenetic analysis of *Castanea* using DNA sequence data from six variable regions of the chloroplast genome indicated migration of extant *Castanea* species from Asia westward to Europe and North America (Lang et al., 2007).

Nuclear genome and mitochondrial genome

Phylogenetic and phylogeographic studies have focused primarily on cpDNA rather than nuclear loci in plants because of the large size of the nuclear genome and the large number and diversity of genes that are involved, and recombination of genes. Interpretation of nuclear genome data is less prejudiced by lineage sorting at individual loci because they represent numerous genealogies across the genome. Moreover, differences in the frequencies of nuclear polymorphisms among subdivided populations should accrue more quickly than differences among chloroplast haplotypes, and they can be easily translated into a genetic distance matrix regardless of recombination (Eidesen et al., 2007). Nuclear DNA (nDNA) markers can readily reveal higher levels of variation. This is especially true using fingerprinting techniques that mainly screen nDNA, regardless of possible problems caused by recombination and heterozygosity. Typically, markers revealing higher levels of variation can reflect a more recent history (Eidesen et al., 2007). For most interspecific phylogenetic studies, nuclear protein coding loci (NPCLs) are the better markers of choice (Rowe et al., 2008), because they have a medium level of variation, have relatively simple detection of paralogs and easy alignment across large phylogenetic distances. Highly conserved coding regions (18S, 26S rDNA) have been used primarily at the family level and above, however the two internal transcribed spacers (ITS1 and ITS2) of the nuclear rDNA are often best suited for comparing species and closely related genera (Baldwin et al., 1995). For phylogeographic studies and phylogenetic analysis of rapid migration exon-primed intron-crossing nuclear sequence markers (EPICs) are widely used. However, anonymous nuclear markers (ANMs) may actually be the better choice because they are relative easy to develop and contain greater variability than EPICs, although ANMs can easy fall in non-coding regions of the genome and a large fraction of non-coding regions include repetitive elements (Thomson et al., 2010).

Molecular study of the mitochondrial (mt) genome involving either restriction site analysis or sequencing primarily has been used in phylogenetic studies of animals. Plant mtDNA generally has a low rate of nucleotide substitutions (Mower et al., 2007). Moreover, it is very large and highly variable in size, structure, and gene order. However, mtDNA has been used in some studies because of the frequent intramolecular recombination resulting in rearrangements of intergenic regions (Elansary et al., 2010).

The nuclear and cytoplasmic genomes have quite different histories and their analysis may result in quite different phylogenetic reconstructions. Combination of data obtained for all the three genomes allows one to better resolve phylogenetic relationships.

Morphology of *Castanea*

The genus *Castanea* have seven deciduous species, they are trees and shrubs with simple ovate or lanceolate leaves with sharply-pointed, widely-spaced teeth, and rounded sinuses. The flowers are two types: catkins, a male staminate type which tends to flower earlier, and a mixed type which has female flowers below a staminate catkin. The fruit including one to three nuts is enclosed within a spiny cupule (Jaynes, 1975).

Morphology is a primary tool used to discern taxa (Krishnankutty and Chandrasekaran, 2008). Taxonomic designations for North American *Castanea* have been based on plant type, leaf type, inflorescences, stamen type of male catkin, fruit shape, fruit glossiness and color (at harvest), ripening period and type of stripes, presence of hair on torch in fruit and contrast of hilum to pericarp. The most discriminating trait was stamen type of the male catkin, which allowed classifying the accessions into longistaminate, mesostaminate, brachystaminate, and astaminate (Binkley, 2008). However, inter- and intra- specific morphological variability can lead to considerable confusion in discriminating species. Morphological ambiguity is apparent in populations of southern Appalachians, where the distribution range of *C. pumila* overlaps with that of *C. dentata*. Species distinction of North American *Castanea* taxa is further complicated because of chestnut blight. High susceptibility in these species to blight prevents almost all young trees from maturing to the point of flower and fruit production. Thus only leaf, twig and stipule morphology can be used to differentiate among *Castanea* species on the US continent. But chestnut blight made *C. dentata* from a big canopy tree to a small tree or shrub, similar as the habit of *C. pumila* var. *pumila* and *C. pumila* var. *ozarkensis*. Reproductive barriers *in Castanea* are incomplete and species are known to hybridize naturally (Johnson, 1988). Hybridized trees between *C. dentata* and *C. pumila* are widespread throughout the southern Appalachian Mountains and they can be difficult to separate from species of *Castanea.*

The American Chestnut Foundation (TACF) has studied morphological features of leaves and twigs of American chestnut, Chinese chestnut, their F_1 hybrid and three successive generations of backcrosses between hybrid populations and American chestnut to determine the rate of recovery of the American chestnut morphology after hybridization. The morphological characters included leaf (shape, apex shape, base shape, margin, interveinal surface and veinal surface), stipule (size and shape), twig (color, surface, lenticels and diameter), and bud (color, shape, tip shape, pitch angle and yaw angle). Results showed that the majority of BC_3 trees differed from Chinese chestnut in every individual characteristic (Diskin et al., 2006).

Boundaries between the two species (*C. dentata* and *C. pumila*) are difficult to establish due to intraspecific variation, interspecific similarities, and possible interspecific hybridization (Shaw and Small, 2004). Thus additional data are necessary to better explain the relationships between these closely related taxa. Integration of molecular and morphological data has proven useful for resolving systematic questions in taxonomically uncertain groups.

The next generation sequencing

In the past few years, next-generation sequencing (NGS) technologies have led to a revolution in genomics and genetics and provided cheaper and faster delivery of sequencing information (Sun et al., 2010). Today's commercial DNA sequencing platforms include the Genome Sequencer from Roche 454 Life Sciences (www.454.com), the Solexa Genome Analyzer from Illumina (www.illumina.com), the SOLiD System from Applied Biosystems (www.appliedbiosystems.com), the Heliscope from Helicos (www.helicos.com), and the commercialized Polonator (www.polonator.org). A distinct and common characteristic of these platforms is that they do not rely on Sanger chemistry as did first-generation machines including the Applied Biosystems Prism 3730 and the Molecular Dynamics MegaBACE (Miller et al., 2010). And the length of theses reads were commonly 500 bp to 1000 bp. The second-generation machines are characterized by decreased costs of sequencing and allow rapid and cost-effective sampling of genome, and much lower cost per read. Also read lengths are much shorter with these new methods than with capillary sequencing (averaging 100–230 bp and 300–400 bp for 454FLX and 454Titanium, respectively, and 35 to up to 76 b for Illumina Solexa platforms). Today's machines are commonly referred to as short read sequencers or next-generation sequencers.

454 GS20, the first commercial NGS platform of Roche Company, was released in 2005 and produces about 200,000 reads with an average read length of 100 bases per run. Since then, 454 sequencing technology has made much progress in data volumes, read length, and difference

in errors and quality. The GS FLX Titanium, the latest 454 sequencing platform, can generate one million reads with an average length of 400 bases at 99.5% accuracy per run. To date, the Genome Sequencer FLX System and GS Junior System are the most widely used in diverse fields of biology (www.454.com).

Single nucleotide polymorphisms

A single-nucleotide polymorphism (SNP, pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in an individual. There has been a recent trend for single nucleotide polymorphism (SNP)-based markers to replace other marker types in both animal and plant species (Jung et al., 2010). Because of binary or co-dominant status, they are able to efficiently discriminate between homozygous and heterozygous alleles. In general, SNPs are common in the genome, can easily be standardized between laboratories and can be surveyed on a wide variety of platforms from single polymerase chain reaction (PCR) to a million or more SNPs on a microarray (Novaes et al., 2008). Moreover, their power comes from the large number of loci that can be assessed instead of from the number of alleles. Once the rare SNPs are discovered in a low diversity species, the genetic population discrimination power can be equivalent to the same number of loci in a genetically diverse species. Finally, SNPs are amenable to high throughput automation, allowing rapid and efficient genotyping of large numbers of samples (Hyten et al., 2010)

In plants, SNPs are always designed from whole genome sequences or expressed sequence tags (ESTs) obtained from genetically diverse individuals. Because of this, the identified SNPs are within known expressed genes (Arif et al., 2010). To date, the main application of this sequencing technology has focused on re-sequencing, including whole genome re-sequencing for SNP discovery (Imelfort et al., 2009). PyroBayes is a modificattion of the software PolyBayes, designed for pyrosequencing reads from 454 sequencing technology (Quinlan et al., 2008). It permits accurate SNP calling in re-sequencing applications, even in shallow read coverage. EagleView software allows the combined viewing of data of 454 and Solexa from both long- and short-read technologies. The software offers a compact assembly view and annotation for the interpretation of SNPs in a genomic context (Huang and Marth, 2008). The SNP discovery software AutoSNPdb (Duran et al., 2009) can integrate both Sanger and Roche 454 pyrosequencing data, enabling efficient SNP discovery from next-generation sequencing technologies. The basic principle of SNP and detection method includes preparation of sample reactions using template and primer, performing SNaPshot reactions by thermal cycling and conduction of post -extension treatment of the products. Then automated electrophoresis of the samples and finally, analyzing the data (Imelfort et al., 2009).

SNPs are becoming popular genetic markers in evolution and ecology (Moen et al., 2008). SNPs can discover genetic diversity in plants, particularly in species with limited genetic diversity. SNP-based markers can be used to set up very dense genetic maps. Marker-assisted selection (MAS) programs and the specific genotypes required for quantitative genetic studies could be made based on the maps. SNPs can be used for genome-wide linkage disequilibrium and association studies that assign genes to specific functions or traits. SNPs can also be used to develop allele-specific assays for the examination of cis-regulatory variation within a species (Barbazuk et al., 2007).

Project Objectives

This study was conducted to differentiate the two species, *Castanea dentata* and *Castanea pumila*, in North America based on the sequence data of chloroplast, nuclear and 454 sequences. The objectives of the present investigation were (1) to develop species-specific markers for correct species identification, (2) conduct phylogeographic analysis to infer historical processes of geographical distribution of plant populations.

Chapter 2

Identification of *Castanea dentata* and *C. pumila* Based on Chloroplast and Nuclear DNA Sequence Data

Abstract

Two *Castanea* species and two chinkapin varieties are endemic to the North American continent. It has been difficult to distinguish American chestnut (*C. dentata*) from the chinkapin (*C. pumila*) via morphological characters because of chestnut blight. In this study, we analyzed *Castanea* species from Alabama, Arkansas, and North Carolina for leaf morphological characteristics, and sequence variability at six chloroplast DNA regions (*trn*T-L, *trn*L, *ndh*F, *ndh*C, *orf*62, and *rpL*16) and two polymorphic nuclear regions (17 and 126). Comparative analysis with other *Castanea* populations was conducted. The results of morphology and sequence analyses indicated that the Ruffner Mountain (AL) population contains two main groups: *C. dentata* type and *C. pumila* var. *pumila* type, with in each group different cpDNA haplotypes due to variability at the *ndh*F or *orf*62 region. Based on the presence of unique indels in *C. dentata* it can be hypothesized that the *C. dentata* type diverged from the *C. pumila* type. Some mutational sites (two deletions in *trn*T-L region, one deletion at *ndh*F region, one deletion in *ndh*C region, one SNP in region *rpl*16, one SNP in nuclear region 17 and one SNP in nuclear region 126) can be considered as species-specific markers to varying degrees. Species identification had better be based on morphology and combined sequence analyses. Phylogenetic analyses of the nuclear data showed the possible origin of hybrid taxa.

Key words: *Castanea; trn*T-L; *trn*L; *ndh*F; *ndh*C; *orf*62; *rpL*16; Plylogenetics.

Introcuction

The genus *Castanea* (Fagaceae), which includes 7 species, is widely distributed in the Northern Hemisphere. Two species and two varieties can be found on the North American continent. American chestnut (*Castanea dentata*), with its characteristic three nuts per bur, once was the dominant canopy tree species in the Appalachian forest ecosystem. It possessed a remarkable array of desirable traits, grew very rapidly, often to great size, had outstanding form and wood quality, and provided food and revenue for rural communities. It ranged from Maine and southern Ontario to Mississippi, and from the Atlantic coast to the Appalachian Mountains and the Ohio Valley (http://www.fagaceae.org). American chinkapin, with one nut per bur, has two varieties, the Allegheny chinkapin (*C. pumila* var. *pumila*) and Ozark chinkapin (*C. pumila* var. *ozarkensis*). Ozark chinkapin is distributed in the extreme southwest of Missouri, northwest of Arkansas and the extreme eastern portion of Oklahoma (http://www.ozarkchinquapin.com). The Allegheny chinkapin performs best on well-drained soils in full sun or partial shade. Its range of adaptation is from northern Florida, west to Texas and Oklahoma, north to Kentucky, Virginia, Maryland, and along the Atlantic coastal plain to Cape Cod, Massachusetts. The chinkapin is not economically important for nut and timber production because of its small nut and tree size, and only provides a food source and community for wildlife (Payne et al., 1994).

Chloroplast (cp) genomes of land plants are highly conserved in both gene order and gene content. It can provide phylogenetically useful information at various taxonomic levels (Ames et al., 2007). Sequences from noncoding regions of the cp genome are often used in systematic analysis because such regions tend to evolve relatively rapidly and provide higher percentages of variable and informative characters as compared to cpDNA coding sequences (Taberlet et al., 1991).

The *trn*T-L region is located in the large single copy region of the chloroplast genome in close proximity to *rbc*L. It contains one intergenic spacer *trn*T (UGU)-*trn*L (UAA) and the *trn*L (UAA) intron. These variable noncoding regions can easily be amplified using universal primers, which are homologous to the exons of the *trn*T (transfer RNA threonine UGU) and *trn*L (transfer RNA leucine- UAA) gene (Taberlet et al., 1991). Although the *trn*T-*trn*L has not been widely used in phylogenetic studies, it was found to have a more than 95% probability of identifying the correct species of *Sinningia* s.l. (Gesneriaceae) (Cowan et al., 2006), and has been used for species identification of *Aspalathus* L. (Fabaceae) (Edwards et al., 2008). It was also used to trace phylogenetic analysis in *Coleeae* (Bignoniaceae) (Zjhra et al., 2004); *Clerodendrum* (Lamiaceae) (Yuan et al., 2010); *Arenaria* section *Plinthine* (Caryophyllaceae) (Valcárcel et al., 2006); and *Castanea* (Fagaceae) (Lang et al., 2006).

The *trn*L intron of plants has sequence conservation in the regions flanking both *trn*L exons, but is highly variable in the central part. *trn*L (UAA) intron sequences have been used for phylogeny reconstruction in the genus *Gentiana* (Gentianaceae) (Gielly and Taberlet, 1994), subfamily Secamonoideae (Lahaye et al., 2007) ; *Ceropegia* (Apocynaceae, Ceropegieae) (Meve and Schumann, 2007); and the basal tribes of the subfamily Papilionoideae (Leguminosae) (Pennington et al., 2001).

The plastid DNA of most plants contains 11 *ndh* genes encoding components of the thylakoid ndh complex (NDH polypeptides). *ndh*F is located at one end of the small single-copy region and encodes the ND5 protein of chloroplast NADH dehydrogenase (Olmstead and Sweere, 1994). Because of its higher rate of nucleotide substitution, *ndh*F has been used extensively for phylogenetic studies at the generic level and above (Small et al., 2004). The *ndh*F gene is known to vary in rate of evolution among major plant lineages and different gene regions. Non–coding regions exhibit a higher level of sequence variation among closely related species than the coding region (Gielly and Taberlet, 1994) and are more suitable to study correct and precise identification of taxa (Miller et al., 2009).

In the chloroplast genome, *orf*62-*trn*GM, is located at the large single-copy region, and encodes the ycf 9 protein, a photosystem II (PS II) core subunit. *trn*G encodes transfer RNA glycine, which recognizes codon GGC (Shinozaki et al., 1986). Phylogenetic studies based on comparative sequences with region *orf*62-*trn*G are seldom used, but have been informative in *Citrullus* species (Dane and Lang, 2004) because of the presence of large indels.

The chloroplast gene *rpl*16, which encodes the ribosomal protein L16, is interrupted by an intron in most land plants. The intron in *rpl*16 is missing from the flowering plant families Geraniaceae, Goodeniaceae and Plumbaginaceae (Campagna and Downie, 1998). The *rpl*16 intron belongs to a group II located in the chloroplast gene flanked by *rpl*14 and *rps*3 in the large single copy (LSC) region near the internal region (IR) border of streptophyte plastid genomes. In most angiosperms, the *rpl*16 gene contains two exons separated by an intron that varies in length from ~1000 bp to 1500 bp (Jordan et al., 1996). For amplification of the *rpl*16 intron, the F71/R1661 primer combination was first used. Later, Kelchner and Clark (1997) substituted the reverse primer R1661 for R1516, which is nowadays most frequently used. Olsson et al. (2009) designed a new reverse primer between R1661 and R1516 that performs very well in combination with F71. This approach was recommended because it facilitates sequencing and allows recovery of complete intron sequences (Borsch and Quandt, 2009). The *rpl*16 intron has been predominantly used for interspecies (Ohta et al., 2006; Katoch et al., 2010) and intergeneric

(Borg et al., 2008; Hansen et al., 2009) relationships in plants, although it can offer some potential population level phylogeny. The *rpl*16 intron is one of the most phylogenetically useful intron because its mutational hotspots can easily be identified and excluded in phylogenetic analyses. However, since the *rpl*16 intron exhibits clear interspecific variability, it is better used in combination with other markers for population studies in plants (Borsch and Quandt, 2009).

The *trn*V-*ndh*C intergenic spacer lies within the LSC region. The average length is 1146 bp and it ranges from 318-1800 bp. The *trn*V-*ndh*C region has been used as a good species-specific marker to differentiate wild Rose populations (Fedorova et al., 2010) and can be used as a barcode to distinguish species of the genus *Psiguria* (Steele et al., 2010). The region has also been used to gain insights to southeastern *Castanea* populations with intermediate morphologies (Shaw et al., 2007; Binkley, 2008). A total of four different haplotype groups were identified at a 390 bp section of *trn*V-*ndh*C in American *Castanea* accessions. One haplotype in *Castanea* trees with intermediate morphology was found to be shared among *C. dentata* and *C. pumila* populations.

Although the plant nuclear genome has a large size and large number and diversity of genes, conserved regions are often used for phylogenetic studies. Some nuclear markers, in combination with mitochondrial sequences, were used to study animal evolution (Stöck et al., 2008) and species phylogeography (Gaines et al., 2005), and two nuclear loci (the 28S ribosomal gene regions D2 and D3-5) were examined to check the phylogenetic relationships between oak gallwasps (Zoltanacs et al., 2007). In plant studies, nuclear markers, usually combined with cpDNA, are used to check genetic diversity, differentiation and phylogenetic hypotheses of species, genus and even family (Muir et al., 2004). And nuclear DNA was investigated for plant species delineation and inference of evolutionary relationships (Suda, 2007). Moreover, nuclear DNA was widely used in fungal studies of the taxonomy and phylogeny (Rakotoarisoa, 2010)

In this study, we compared and analyzed sequences from different American *Castanea* populations based on the chloroplast DNA regions of *trn*T-L, *trn*L, *ndh*F, *ndh*C, *orf*62, *rpL*16 and two polymorphic nuclear regions (17, and 126), with the intent to obtain species-specific markers and gain a better understanding the phylogeny of the genus *Castanea* in North America. Our special emphasis is on the most southern *Castanea* population in the Appalachian region, known for its morphological diversity.

Materials and Methods

Plant material and morphological analysis

Thirty-one samples (including *Castanea dentata* and *C. pumila*) from four populations were used (Table 1).Wherever possible fresh leaf samples were collected from *Castanea* populations. Leaves collected from *Castanea* trees at Ruffner Mountain, AL (Figure 1), were examined using diagnostic quantitative and qualitative characteristics as described by Jaynes (1975) and Johnson (1988). Samples were sent to Dr. Fred Hebard at TACF Meadowview Research Farm for species identification and have since been deposited at the AU Herbarium. Seven samples were received from Dr. J. James (TACF Carolina chapter).

DNA extraction, PCR, and nucleotide sequencing.

DNA was extracted from nuts of *Castanea pumila* using the DNeasy^{TM} plant Mini kit (Qiagen, Valencia, CA); DNA extractions were made from fresh leaf material using CTAB

(Hexadecyl trimethylammonium bromide) method (Kubisiak et al., 2003). DNA from different populations in central and southern Appalachian region was kindly provided by Dr. T. L. Kubisiak. The *trn*T-L intergenic spacer was amplified with primers “A” and “B” (Table 2). The *trn*L intron region was amplified with primers “C” and “D” (Taberlet et al., 1991). The 3’ flanking region of *ndh*F was amplified using the primer 1955F - 607R (Olmstead and Sweere, 1994). Primers *orf*62 and *trn*GM were used to amplify the *orf*62-*trn*GM region (Heinze, 2002). Primers exon1 (Forward) and exon2 (Reverse) were used to amplify the *rpl*16 region (Shinozaki et al., 1986). Primers $trnV_2$F and *ndh*CR were used to amplify the *ndh*C region (Shaw et al., 2007). Wound or blight fungus infection responsive expressed sequence tags (ESTs) from American and European chestnut available in the GenBank database (www.ncbi.nlm.nih.gov) were used for the design of primers to identify sequence polymorphisms unique to each *Castanea* species. Primer pair 17, designed based on the American chestnut EST (BG835820), and primer pair 126 (Casasoli et al., 2001) were used in this study to amplify sequences at nuclear regions. Double stranded DNA amplifications were performed in a 55-µl volume containing 1×PCR buffer of 20 mmol/L Tris HCl (pH 8.4) and 50mmol /L KCl, 1.5mmol/L $MgCL_2$, 200µmol/L of each dNTP, 0.2 µmol/L of each primer, 2 U of *Taq* polymerase (New England Biolabs), and 2.5 µl template DNA (50ng/L). PCR products were purified using Qiaquick PCR purification kit (Qiagen,Valencia, CA) to remove excess primers and dNTPs. Sequencing of PCR products was conducted by Auburn Genomics and Sequencing Lab with the ABI3100 sequencer (Applied Biosystems Inc. Foster city, CA).

Sequence alignment and data analyses.

Multiple alignments of the sequences were carried out at CLUSTAL W at the default setting, using the AlignX program implemented in the Vector NTI software, and adjusted manually. Gaps were introduced in the alignment in order to optimize positional homology. Single–base indels were cross-checked to the original chromatograms, to verify that they were not sequencing artifacts missed during base calling. Indels that were potentially parsimonious were scored and added to the end of the data sets as present (1) or absent (0) type characters. Gaps with overlaps were considered nested and treated as single multi-state character according to Simmons and Ochotreana (2000). Areas of ambiguous alignment were excluded from all analyses. A maximum parsimony analysis was conducted using PAUP version 4.0 software (Swofford, 2000). Sequences were aligned with cpDNA sequence information from many other *C. dentata* and *C. pumila* populations (Dane and Lang, 2008; Lang et al., 2007; Dane, 2009) (Table 3). Nuclear DNA sequences (17 and 126 region) were aligned with sequences from other known *Castanea* taxa.

Results

Morphological characterization of Ruffner Mountain, AL, tree samples.

Based on morphological characteristics of leaves collected from trees at Ruffner Mountain (Figure 2), 12 trees showed the *C. dentata* type, while 12 samples were identified as chinkapin type. Samples from Dr. J. James showed morphological characters indicative of *C. pumila* var. *pumila* or var. *ozarkensis* (Table 1).

Variability at *ndh*C cpDNA region.

When the Ruffner Mountain and *C. pumila* (JJ) samples were aligned, only five variable sites were detected at the 560 bp length *ndh*C region, with a transition to transversion ratio of 2:3. When more *ndh*C sequences of samples from different populations in North America were aligned, a total of 14 variable sites were detected, which includes three indels, three multi-nucleotide changes and 8 single nucleotide substitutions, 10 of which are parsimony informative, with a transition to transversion ratio of 11:6 (Table 4). One large 59 bp deletion can only be detected in northeastern *C. dentata* (D type, Dane, 2009) samples. Four variable sites are unique to 11samples from AL (Represented by PT20, 7CN, and IZ1in Fig.3), three variable sites are unique to samples from FL, two variable sites are unique to *C. sativa* samples, and two SNPs are unique to the samples from KY.

Molecular variation at combined cpDNA regions.

In the combined data analysis of the following cpDNA regions: *trn*T-L, *trn*L, *ndh*F, *orf*62, and *rpl*16, a total of 16 gaps with 9 indels and 7 single nucleotide substitutions were introduced into the Vector NTI sequence alignment (3 in the *trn*T- L spacer, 2 in the *trn*L intron, 5 in *ndh*F, 2 in the o*rf*62, and 4 in the *rpl*16 region) (Table 5). These gaps ranged from one to 73 bases, with the largest indel of 73 bp in the *trn*T-L intergenic region. Ruffner Mountain samples could be divided into two main groups I and II. Eleven Samples in group I are characterized by two deletions (12 and 73 bp) at the *trn*T-L region, and a unique 31 bp insertion at *ndh*F region and multi nucleotide SNPs at *ndh*C region, and can be considered as *C. dentata* type (Dane, 2009). Included in this group is the sample PT20, which lacks the 31 bp indel at *ndh*F region characteristic of the *C. dentata* type. One study had noted that two deletions (12 and 73 bp) at the *trn*T-L region could differentiate *C. dentata* from all other *Castanea* species (Kubisiak and

Roberds, 1997). However, Dane (2009) found a few chestnut mother trees with a unique deletion (42 bp), many mother trees from AL, GA, TN, NC and KY without the 12 and 73 bp deletions and one Allegheny chinkapin population in northern Georgia which showed the two deletions (12 and 73bp) at *trn*T-L. So this region cannot reliable be used to distinguish *C. dentata* trees from the other C*astanea* species.

The group II contains the other 13 Ruffner Mountain samples which show high sequence homology to *C. pumila* type(s), which lack variability at the *trn*T-L region. The Ruffner Mountain samples in this group have 2 unique deletions at *ndh*F region with the exception of 4CN. The *ndh*F region of 4CN is homologous to other *C. pumila* haplotypes (Dane and Lang, 2008). The haplotype of M34 is different at the *orf*62 region. Three Ozark chinkapin samples (JJ4, JJ5, and JJ6) and JJ7 can be distinguished from Allegheny *C. pumila* samples based on 4 variable cpSNP sites (position 268 at *trn*L region; position 518 at *orf*62 region; position 221 and 485 at *rpl*16 region). One region of *ndh*F is a mutational hotspot since it shows many mutational changes in *Castanea* species. Eight samples (JJ1, JJ2, JJ3, JJ4, JJ5, JJ6, JJ7 and 4CN) can be differentiated from other *C. pumila* samples based on the variability at *ndh*F (one SNP and one indel). This pattern also occurs in Allegheny and Ozark chinkapin samples (Dane and Lang, 2008). Sample JJ4 has a 24 bp insertion at the *trn*L region which also occurs in other Ozark and some Allegheny chinkapin samples (Dane and Lang, 2008) and a unique 34 bp deletion at the *rpl*16 reigon. In the Ruffner Mountain *Castanea* population 5 different haplotypes were detected.

Molecular Variation at nuclear DNA region

A total of 8 variable parsimony informative sites were detected at the alignment of the two nuclear regions of Ruffner Mountain and JJ samples, with a transition to transversion ratio of

7:1. Two of the variable sites (position 234 at 126 region; position 332 at 17 region) can be used as marker to divide the 31 samples into two groups (Table 6). This result is consistent with that from the five regions of cpDNA except the sample AL_4CN. Group I contains eleven samples (*C. dentata* type) plus the sample AL_4CN. Variability at position 184 at the region 126 was observed in seven samples (AL_5, AL_6, AL_M18, AL_M33, AL_IZ1, AL_M61, and AL_M38). At position 158 of region 126, one SNP was detected in three samples (AL_7CN, AL_AL2 and AL_M60). More variability was detected at the nuclear regions in group II which includes the other 19 samples. One SNP was found at position 119 at the 126 region in three samples (AL_M34, MO_JJ6 and NC_JJ2). One SNP was found in position 172 at the 126 region in three other samples (JJ1, JJ3 and JJ7). At the 17 region, only two samples (AL_M31 and AL_M67) have a SNP at position 402, and more samples (NC_JJ1, NC_JJ3, AL_M65, AL_M40, AR_JJ4 and AL_M37) show variability at position 451 ('G' change to 'A') or heterozygosity.

Phylogenetic analysis

*ndh*C region

The analysis of the data set with indels coded as extra binary characters yielded two equally most-parsimonious trees. Each tree required 17 steps and had a consistency index (CI) of 0.8235, and a retention index (RI) of 0.9302. Using *C. sativa* as outgroup, the 50% majority-rule consensus tree from 170 trees based on bootstrap analysis of the data set is presented in Fig.3. Analysis of the 34 samples with equal character weighing generated one most parsimonious tree. The American *Castanea* species are supported as a monophyletic clade, and 32 taxa were divided into four groups. One group is composed of *C. dentata*, collected from north-eastern

United States; the second group is considered as hybrid group, including most of the samples from central and southern Appalachian mountain area and several *C. pumila* samples from FL. The third group contains the Ozark chinquapin samples from Arkansas, Missouri, and one Georgia chinkapin sample. The fourth group only contains samples (PT20, 7CN, and IZ1) from Ruffner Mountain in Alabama. They are sister to the *C. dentata* hybrid type and Ozark type.

NLRO (*ndh*F+*trn*L+*rpl*16+*orf*62) regions

To increase resolution, phylogenetic analyses with combined data sets were conducted. Only those accessions for which we had sequences from all the studied regions were used in the analysis. Aligned sequences from *ndh*F, *trn*L, *rpl*16, and orf62-*trn*GM were concatenated into a single data set, which contained 45 taxa, 2350 characters with 6 indels coded as binary characters. The 50% majority-rule consensus tree from 106 trees based on bootstrap analysis of the data set with the indels coded as extra binary data is presented in Fig.4. American *Castanea* species are supported as a monophyletic clade with all the 45 taxa divided into 3 types. The first group with seven samples, including 2 samples from Ruffner Mountain (7CN and PT20), can be considered as *C. dentata* type. The second group can be considered as hybrid or *C. pumila* type, although the resolution appears satisfactory, the topology is somewhat confused. These samples originated in the south and central Appalachian mountain region and include samples of *C. dentata*, *C. pumila*, and hybrids between the two species. The third group is Ozark chinkapin type, and contains samples from Arkansas, Missouri, and one hybrid sample (JJ4) from Georgia.

NLRO+NDHC regions

In order to better understand the phylogenetic relationships of *Castanea* species in North America, we concatenated the *ndh*C region of 24 samples into the NLRO region, resulting in 46 parsimony-informative characters, with 15 indels coded as binary characters. The most parsimonious tree has 96 steps, and has a consistency index (CI) of 0.6000 and a retention index (RI) of 0.7821. Using *C. sativa* as outgroup, the 50% majority-rule consensus tree from 180 trees based on bootstrap analysis of the data set with the indels coded as extra binary data is presented in Fig.5. The topology derived from the combined data set was congruent to that generated from the NLRO data analyses. Species from North America are supported as a monophyletic clade. Three major clades are evident in the cpDNA phylogeny: one which contains two Alabama samples (7CN and PT20) and two samples (C_DENT1, and GA_BTB), which can be considered as *C. dentate* type. The second clade includes samples from the central and southern Appalachian mountain area, and can be considered as *C. pumila* type. And a third is the chinkapin clade, which includes samples from around the Ozark mountain area.

Nuclear regions

To check the sequences for informative mutations at the nuclear regions (17 region and 126 region), sequences from 40 samples were concatenated and aligned together. A total of 22 parsimony-informative characters were observed. The most parsimonious tree has 38 steps, with a consistency index (CI) of 0.6667 and a retention index (RI) of 0.9160. The 50% majority-rule consensus tree from 333 trees based on bootstrap analysis of the data set with the indels coded as extra binary data is presented in Fig.6. The results show three clades. The first clade includes group II Ruffner Mountain samples, six JJ samples and four other *C. pumila* samples from GA, FL, and MO, can be considered as *C. pumila* var *pumila* or hybrid type. The second clade contains group I Ruffner Mountain samples (IE1, PT20 and M38) and other samples which from north-eastern America, which belongs to the *C. dentata* type. The two clades are congruent with those observed using cpDNA sequence data. Samples 4CN and 7CN from Ruffner Mountain do not group with the other samples, while JJ7 groups with Asian species. At the nuclear region, especially at region 126, the Asian species show several unique SNPs. Sample JJ7 can thus be considered as a hybrid between Allegheny chinkapin (mother tree) and Chinese chestnut (father tree). The hybrid nature of KY110, which is known to be a F1 between *C. mollissima* × *C. dentata* show the *C. mollissima* cpDNA type and combination of Chinese and American sequences at the nuclear regions. Moreover, sample 4CN can be considered as hybrid tree between *C. pumila* and *C. dentata* which is congruent with results of cpDNA region.

Discussion

Comparison of morphological characteristics and cpDNA haplotypes

During July 2009 fresh leaf samples (24 in total), collected from "chestnut" sprouts at Ruffner mountain, were positively identified by Shawn Yeager at the TACF Meadowview Research Farm as either American chestnut (*C. dentata*: M60, 7CN, 5, 6, AL2, IZ1, PT20, MS33, M18, 4CN, M68, and MS38) or chinkapin (*C. pumila*, other 12 samples). Moreover, the following 10 samples had been identified previously by Drs. Hill Craddock and/or Fred Hebard as *C. dentata* (M 18, M36, 4CN, and 7CN), *C. pumila* (6, M30, M37, and M40) or as hybrids (PT20, and AL2). The morphological classification is not completely congruent with the results from cpDNA analysis and the identification by Shawn Yeager. It is clear that difficult to discern *Castanea* species on basis of morphological characteristics alone, since the criteria used are not distinct and leaf characteristics are influenced by environmental and climatic factors. For example, the presence of hairs on leaves is an important character in plant taxonomy. Simple hairs did appear on midribs of most chinquapin leaves, but also on some American chestnut leaves. So even with the observations on morphological characters, it is difficult to draw a correct conclusion with 100% accuracy.

The taxonomic status of C. *pumila* var. *ozarkensis* is unclear, although the Allegheny and Ozark chinkapins have been considered as two varieties of one species (Johnson, 1988). When we analysed Ruffner Mountain samples with other known *Castanea* taxa (for a total of 43 taxa) using a combined cpDNA data set (*ndh*F+*trn*L+*rpl*16+*orf*62) (see Fig 4), the phylogenetic tree indicated that the North American species are weakly supported as a clade with bootstrap value of 60% with *C. pumila* var. *ozarkensis* as the basal lineage, sister to the group of *C. pumila* var. *pumila* and *C. dentata*. Similar results were obtained from the analysis of another combined cpDNA data set (24 taxa and five regions: *ndh*F +*trn*L+ *rpl*16+*orf*62+*ndh*C) (Fig.5). Moreover, the divergence between the chinkapin varieties is larger than the divergence between the American chestnut (*C. dentata*) and Allegheny chinkapin (*C. pumila* var. *pumila*). Even though Johnson (1988) considered the Ozark chinkapin as ancestral and less highly evolved than the Allegheny chinkapin, the combined data showed separation of the Ozark chinkapin and American chestnut and retention of ancestral characters.

Based on combined data analyses of 4 or 5 cpDNA regions (Figs 4, and 5), it is clear that the relationship among the samples from the southern Appalachian area is very complicated. In this region, the distribution range of *C. pumila* overlaps with that of *C. dentata.* Moreover, since the domestication of economically important chestnut crops, European, Chinese and Japanese chestnuts have been planted all along the Appalachian region. *Castanea* species are wind pollinated and can hybridize freely, and research has recently confirmed that hybridization between the different *Castanea* species did occur over time (Dane, 2009,). Thus there is a possibility for gene flow via introgression between sympatric species in southern Appalachian area, and strong directional selection pressure and human activities can aggravate the genomic divergence.

Comparisons of cpDNA regions

The *trn*T-L spacer region has been widely utilized to resolve phylogenetic relationships at both the generic and species level (Small et al., 1998), but it is rarely used in systematic studies. In this study, the *trn*T-L region has three variable sites, a 73bp deletion in the *trn*T-L intergenic region, the largest mutation in all of the examined regions, which can provide informative markers for tracing the phylogeny among species of the genus *Castanea*. For several samples, this region can be used as a species-specific marker to distinguish the *C. dentata* with all other *Castanea* species (Kubisiak and Roberds, 1997). However, a few chestnut mother trees were reported to have a unique deletion (42bp) instead of the large deletion, and some Allegheny chinquapin populations in northern Georgia have the two deletions (12 and 73bp) (Dane, 2009). An earlier study showed that the *trn*T-L region is variable both with respect to indel number and indel length (Lang et al, 2006), and it can evolve faster than the *trn*L region. In contrast, *trn*L region has only one SNP and is conserved. So the *trn*T-L should be useful for phylogenetic studies at lower taxonomic levels (Fukuda et al., 2001).

When the Ruffner mountain and JJ samples were aligned together, only five mutations at the *ndh*C region were detected. When more samples from different populations were added to the alignment, a total of 13 mutations were detected, some of which are unique to the Ruffner

Mountain population, one SNP (A-G) is unique to the KY samples, and 3 mutations are unique to FL samples (Table 4). Shaw et al. (2005, 2007) had reported that the *ndh*C region is one of the cpDNA regions which showed the greatest variation in comparisons of three groups of angiosperms: rosids, asterids, and monocots. And the *ndh*C region is noted as highly variable by Steele et al. (2010) and Timme et al. (2007). Because this spacer is evolving faster than the other regions, maybe it can be used in phylogenetic studies at higher taxonomic levels.

Nuclear marker SNPs

Since chloroplast DNA haplotypes are maternally inherited, species-specific nuclear DNA markers are needed for proper species and hybrid identification in evolutionary lineages of intermediate morphology. Primers were designed using wound specific *Castanea* ESTs (Connors et al., 2001; Casasoli et al., 2001) and used to screen genomic DNA from different *Castanea* species to identify sequence polymorphisms specific for each species (Dane, 2009). Only the regions amplified with primer pairs 17 (EST BG838520) and 126 (Casasoli et al., 2006) were used for sequence analysis because of intra and interspecific polymorphism (Dane, unpublished results). The presence of conserved SNPs at both regions (Table 6) could divide the Ruffner Mountain and JJ samples into a *C. dentata* specific, *C. pumila* specific or interspecific hybrid type. The only exception was 4CN (*C. dentata* cpDNA haplotype), which was grouped with the *C. pumila* nuclear type. Nuclear SNP analysis of JJ7 showed that the tree is a recent hybrid between *C. pumila* var. ozarkensis (its cpDNA maternal haplotype) and *C. mollissima* based on sequence specific *C. mollissima* SNPs detectable at both regions. The sample KY110, at the same group as sample JJ7, is a known hybrid tree (*C. mollisima*× *C. dentata*). Thus recent interspecific hybrids can be identified using a combination of cp and nuclear SNPs. However, more regions are needed to be able to unambiguously identify samples with intermediate morphology commonly found in southern Appalachian regions.

Identification of Ruffner Mountain, AL samples

Based on the morphology and cpDNA and nuclear DNA information, the Ruffner Mountain samples consist of two main groups. One is the *C. dentata* group of 11 samples (IZ1, M61, M18, M33, 6, 5, M38, 7CN, AL2, M60, PT20), all of which have deletions at *trn*T-L region, and 31bp insertion at *ndh*F region similar to samples from the northeastern America. This group contains 2 cpDNA haplotypes. The 50% Majority-rule consensus trees (Fig.3, 4, 5, 6) support the conclusion. The geographic distribution of cpDNA lineages allows for present population patterns to be connected to post-glacial migration routes from separate thermophilic forest refugia. Study of Davis (1983) showed that the migration route of the genus *Castanea* was from south to north America, and indicated the existence of *C. dentata* in the southern Appalachian region 15,000 ya, while in the northern Appalachian region 5000 ya, and in Connecticut only 2000 ya (Delcourt and Harris.,1980). In comparisons of cpDNA regions of different *Castanea* species, we can deduce that this *C. dentata* haplotype is evolutionary young as compared to other *Castanea* haplotypes. At the *ndh*C region, for example, only the Northern samples (NC_C2, NC_C8, KY_LW23, KY_LW28, and CT9) have a 19bp deletion (Table 3). The Ruffner Mountain samples with multi nucleotide SNP at *ndh*C region are a young *C. dentata* population similar to those in the north-eastern US. The SNP at position 200 of the *orf*62 region does not occur in MS34, but is unique to other samples of this group. Sample 4CN has a hybrid character based on DNA sequences and leaf morphology. Group II includes the other 12 samples, Dr. Fred Hebard had indicated that11 are chinquapin (except sample M68) (Table 1).

However, based on the leaf size and color, the samples resemble chestnut more than chinkapin. The 50% majority-rule consensus trees do not provide conclusive information. Alabama is known to be contact zone for closely related species that survived glacial periods in different refugia (Soltis et al. 2006), moreover, the reproductive barriers are not absolute between *Castanea* taxa (Johnson, 1988), so hybridization could have occurred in the overlapping region. The species have complicated taxonomy and a complex evolutionary and biogeographic history. Thus some of the *C. dentata* trees might be evolutionary older *C. dentata* population remnants or the result of hybridization between *C. dentata* and *C. pumila*. Binkley in her survey of a GA *Castanea* population (The pocket in Floyd County) of intermediate morphology, similarly detected intermediate haplotypes indicative of hybridization between the species. Sample JJ7 was clearly of hybrid origin based on morphology and DNA sequences variety. Lang et al (2007) found *C. ozarkensis* to be sister to a clade with *C. dentata* and *C. pumila* accessions, and our results from the 50% majority-rule consensus trees (Figs 3, 4, 5, and 6) supported this conclusion. However, more research needs to be done before the conclusions can be drawn about the evolutionary relationships of the *Castanea* species in North America.

Table 1. *Castanea* sample description, origin, and species identification

Samples	Compiled observations	Identification	Origin
GA_JJ7	Simple hairs on midrib, minor veins, interveinal, leaf margins, Stalked glandular hairs, low density of glandular hairs off midrib; purplish stem.	*C. pumila*× *C. mollissima*	GA
NC_JJ3	Many simple hairs on midrib, minor veins, interveinal, leaf margins, stalked glandular hairs on minor veins, purplish stem	Chinkapin	Varnamtown, NC
NC_JJ1	Simple hairs on minor veins, midrib, interveinal, leaf margins stalked glandular hairs on midrib; purplish stem.	Chinkapin	Varnamtown, NC
AL_M35	Simple hairs on midrib, minor veins, leaf margins stalked glandular hairs on midrib; purplish stem.	Chinkapin	N33.56889 W86.69630
AL_MS30	Simple hairs on minor veins, midrib, interveinal, leaf margins stalked glandular hairs on midrib; purplish stem.	Chinkapin	N33.56969 W86.69490
AL_XL1	Lots of simple hairs on midrib, leaf margins ,and minor veins stalked glandular hairs off midrib	Chinkapin	N33.58404 W86.69578
AL_MS31	Simple hairs on minor veins, midrib, interveinal, leaf margins stalked glandular hairs on midrib and minor veins; purplish stem.	Chinkapin	N33.56985 W86.69502
MO_JJ5	Lots of simple hairs on midrib, minor veins, interveinal, leaf margins, stalked glandular hairs on minor veins, purplish stem	Ozark Chinkapin	MO
AL_MS65	Simple hairs on midrib, minor veins, interveinal, leaf margins stalked glandular hairs; purplish stem.	Chinkapin	N33.55571 W86.70269
AL_MS40	Simple hairs on midrib, minor veins leaf margin stalked glandular hairs on midrib; purplish stem.	Chinkapin	N33.56887 W86.69483
AL_M67	Simple hairs on minor veins, midrib, interveinal, leaf margins stalked glandular hairs on midrib; purplish stem.	Chinkapin	N33.55699 W86.68112
AR_JJ4	Simple hairs on midrib, minor veins leaf margin stalked glandular hairs on midrib; purplish stem.	Ozark Chinkapin	AR
AL_MS36	Simple hairs on midrib, leaf margins, and minor veins stalked glandular hairs on midrib; low density of glandular hairs off midrib; purplish stem	Chinkapin	N33.56983 W86.69596
AL_M001	Simple hairs on midrib, minor veins, leaf margins stalked glandular hairs on midrib; purplish stem.	Chinkapin	N33.56124 W86.71202
AL_M34	Simple hairs on minor veins, midrib,interveinal stalked glandular hairs on midrib, purplish stem.	Chinkapin	N33.55782 W86.70214
MO_JJ6	Simple hairs on midrib, minor veins leaf margin, long stalked glandular hairs on midrib; purplish stem.	Ozark Chinkapin	MO
NC_JJ2	Simple hairs on midrib, leaf margins, and minor veins stalked glandular hairs on midrib; low density of glandular hairs off midrib; purplish stem	Chinkapin	Varnamtown, NC
AL_M68	Long simple hairs on midrib, rare off midrib few interveinal stellate hairs unstalked glandular hairs.	American chestnut	N33.55861 W86.69405
AL_MS37	Simple hairs on midrib, minor veins, leaf margins stalked	Chinkapin	N33.56944

	glandular hairs on midrib		W86.69516
AL_4CN	Simple hairs on minor veins, midrib, interveinal, leaf margins stalked glandular hairs on midrib; purplish stem.	American chestnut	N33.55565 W86.70406
AL_6	Simple hairs on midrib, occasional minor veins and interveinal unstalked glandular hairs, low glandular hairs off midrib.	American chestnut	N33.55545 W86.70392
AL_M61	Simple hairs on minor veins, midrib, interveinal, leaf margins stalked glandular hairs on midrib; purplish stem.	Chinquapin	N33.56900 W86.69000
AL_IZ1	Simple hairs on minor veins, midrib, interveinal, leaf margins stalked glandular hairs on midrib; purplish stem.	American chestnut	N33.55561 W86.70233
AL_M18	Simple hairs on midrib, occasional minor veins and interveinal unstalked glandular hairs, low glandular hairs off midrib.	American chestnut	N33.55598 W86.70139
AL_M33	Long simple hairs on midrib, rare off midrib few interveinal stellate hairs unstalked glandular hairs.	American chestnut	N33.56957 W86.69512
AL_5	Long simple hairs on midrib unstalked glandular hairs on midrib hot cross buns on minor veins.	American chestnut	N33.55539 W86.70401
AL_MS38	Long simple hairs on midrib, minor veins, interveinal. occasional interveinal stellate hairs; unstalked glandular hairs off midrib; purplish stem	American chestnut	N33.56956 W86.69500
AL_7CN	Long simple hairs on midrib, rare off midrib few interveinal stellate hairs unstalked glandular hairs.	American chestnut	N33.55592 W86.70354
AL_AL2	Simple hairs on midrib, sparse medium length abaxial veins unstalked glandular hairs off midrib.	American chestnut	N33.55543 W86.70219
AL_M60	Long simple hairs on midrib, rare off midrib few interveinal stellate hairs unstalked glandular hairs.	American chestnut	N33.56900 W86.69000
AL_PT20	Long simple hairs on midrib, minor veins, interveinal. Occasional stellate. Stalked and unstalked glandular hairs on midrib.	American chestnut	N33.55562 W86.70242
AL_FD1	Very hairs simple hairs all over; large stipules; hairy stem; stalked glandular hairs on midrib; Stalked mushroom-shaped glandular hairs on minor veins lots	Chinese chestnut	AU campus control

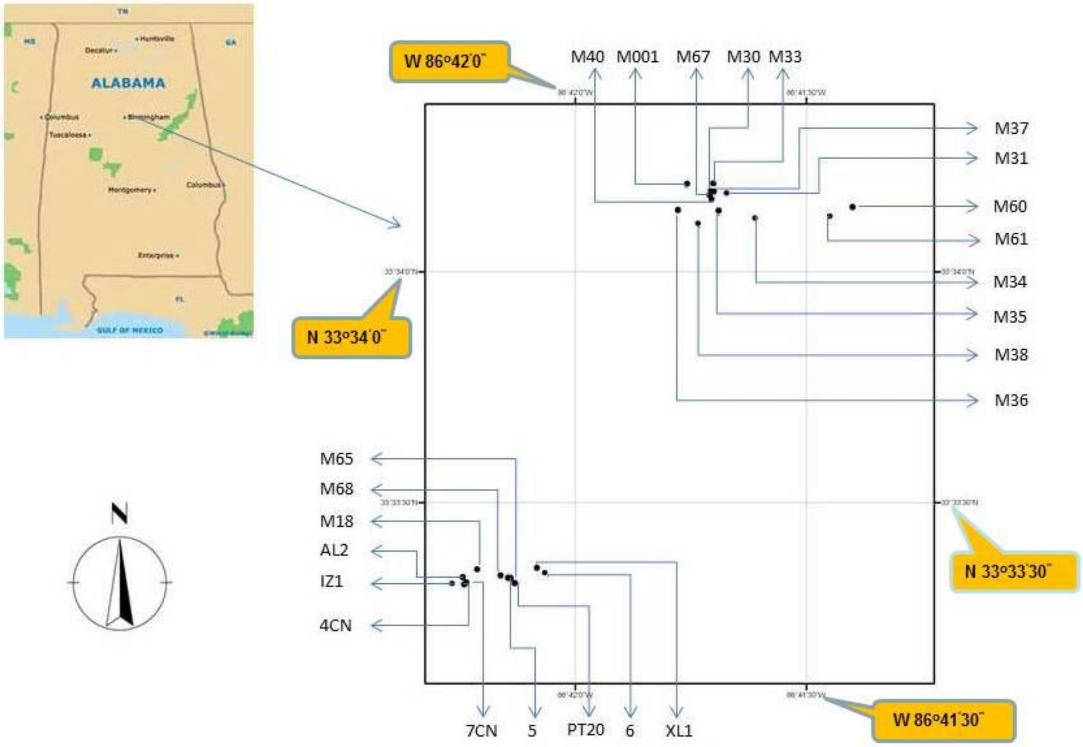

Fig.1. Samples distribution in Ruffner Mountain in AL.

Table 2. Primers used for PCR reaction and sequencing (from 5’ to 3’).

Region	Primers	Sequence	PCR	Reference
*trn*T-L	A	CATTACAAATGCGATGCTCT	55/65°C	Taberlet et al., 1991
	B	TCTACCGATTTCGCCATATC		
*trn*L	C	CGAAATCGGTAGACGCTACG	55/65°C	Taberlet et al., 1991
	D	GGGGATAGAGGGACTTGAAC		
*rpl*16	exon1	AATAATCGCTATGCTTAGTG	54/65°C	Shinozaki et al., 1986
	exon2	TCTTCCTCTATGTTGTTTACG		
*ndh*F	1955F	TATATGATTGGTCATATAATCG	54/65°C	Olmstead and Sweere,1994
	607R	ACCAAGTTCAATGTTAGCSAGATTAGTC		
*orf*62	trnGM	ACCCCGCATCTTCTCCTCGG	52/65°C	Heinze,2002
	orf62P	CTTGCTTTCCAATTGGCTGT		
ndhc	trnV2F	TATTATTAGAAATAAATATCATATTC	54/65°C	Shaw et al., 2007
	ndhCR	GTCTACGGTTCGARTCCGTA		
17	17F	ATTTCATGGGGTGCCTTAAT	55/72°C	Kubisiak, 2003
	17R	GGAGGTTTTGAAAGGGATGG		
126	126F	ACCCTTACCCTGCGACTTCT	53/72°C	Casasoli et al., 2006
	126R	TGCTCAAGAGGCTGTGAAGA		

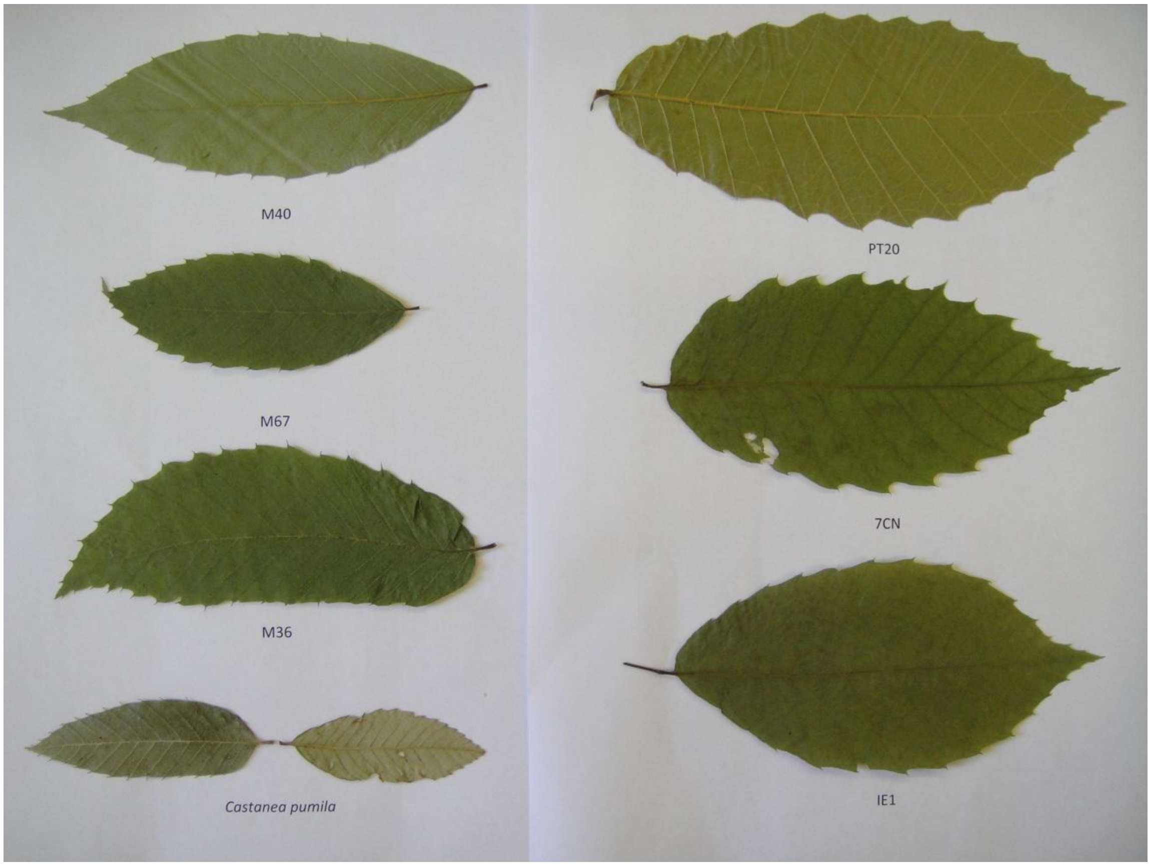

Fig.2. Comparison of leaf characteristics of *Castanea* trees at Ruffner Mountain in AL, showing abaxial side of leaves from M40 and PT20, adaxial side of leaves from M67, M36, 7CN and IE1, and *Castanea* pumila adaxial and abaxial side of leaves from MS.

Table 3. Collection site information of *Castanea* samples used for comparative analysis

Species	Samples ID	County, State or County
C. dentata	AL_LAC	Lacon Mountain Grove, Morgan County, AL
C. pumila	AL_FOR, AL_M1	Mobile, AL
C. dentata	AL_TAL	Talladeega, AL
C. pumila	AR_P7	Russellville , AR
C. pumila	AR_M20	AR
C. pumila	AR_B43, AR_S41, AR_E13, AR_S5	Sylamore Ranger, District, AR
C. dentata	C_DENT1	CT
C. mollissiama	C_MOL2	SLR1T15, New Haven, CT
C. sativa	C_SAT1,C_SAT3, C_SAT4	R2T21, R2T41, R1T2, New Haven, CT
C. crenata	C_CREN1,C_CREN5	NL R34T6, New Haven, CT
C. pumila	FL_P6	HH R4T1, New Haven, CT
C. pumila	FL_E3, FL_P4, FL_D22, FL_I1, FL_B144	Eglin Air Force Base, Okaloosa County, FL
C. pumila	FL_L5, FL_E9, FL_G11	Eglin Air Force Base, Okaloosa County, FL
C. dentata	GA_BTB	Brass Town Bald, GA
C. dentata	GA_TP5	Floyd County, GA
C. dentata	GA_LL38, GA-LL43, GA_LL4	Lula Lake, GA
C. dentata	GA_LL36, GA_LL7, GA_LL2, GA_LL7	Lula Lake, GA
C. pumila	GA_JUN	Juno, GA
C. dentata	GA_RAB	Rabun County, GA
C. dentata	GA_OAK	Oak Mountain, GA
C. dentata	GA_JM, GA_JMR	John Mountain, GA
C. dentata	GA_RSF6	Rattle Snake Falls, GA
C. dentata	KY_LW21,KY_LW32,KY_LW18,KY_LW11	Laurel and Whitney County, KY
C. pumila	MS_314, MS_39, MS_314, MS_39, MS_2	Saucier, MS
C. pumila	MO_1	Oregon County, MO
C. dentata	NC_C2, NC_C8, NC_C9, NC_14, NC_C10	Coweeta County, NC
C. dentata	NC_C60, NC_C45, NC_C6, NC_C58, NC_37	Coweeta County, NC
C. sativa	C_SAT7	Romania
C. pumila	VA_C16, VA_C18, VA_C15, VA_1510	Snakeden Mountain, VA
C. dentata	MUSCK, VA_BA22, VA_BA33	VA
C. dentata	WB385	VA
C. pumila	VA_A3, VA_B13, VA_B5	Iron Mountain, VA
C. dentata	WV_52, WV_23	McDowell, WV
HYBRID	KY_110	*C. mollissima* × *C. dentata*

Table 4. Substitutions of variable sites at *Castanea* samples at *ndh*C (cpDNA) region.

Samples	126	134	142	157	174	179	181	210	265	321	325	465	482
NC_C8	A	--	-	--------	-----	T	A	T	G	G	TA	C	-
NC_C2	A	--	-	--------	-----	T	A	T	G	G	TA	C	-
KY_LW23	A	--	-	--------	-----	T	A	T	G	G	TA	C	-
KY_LW28	A	--	-	--------	-----	T	A	T	G	G	TA	C	-
CT_9	A	--	-	--------	-----	T	A	T	G	G	TA	C	-
AL_M40	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
AL_M65	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
MS_2	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	A	G	TA	C	-
KY_LW21	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	A	G	TA	C	-
KY_LW18	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	A	G	TA	C	-
GA_LL38	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
GA_LL43	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
KY_LW32	A	AA	A	ACAAAACAA	ATTTAA	A	A	C	G	G	TA	C	-
NC_C10	A	AA	A	ACAAAACAA	ATTTAA	A	A	C	G	G	TA	C	-
NC_C37	A	AA	A	ACAAAACAA	ATTTAA	A	A	C	G	G	TA	C	-
GA_RSF6	A	AA	A	ACAAAACAA	ATTTAA	A	A	C	G	G	TA	C	-
NC_C58	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
FL_L5	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	C	GC	A	-
FL_G11	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	C	GC	A	-
FL_E9	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	C	GC	A	-
NC_C60	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
GA_LL4	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
AL_PT20	A	TT	C	TTGTTTTGT	ATTTAA	T	A	T	G	G	TA	C	-
AL_7CN	A	TT	C	TTGTTTTGT	ATTTAA	T	A	T	G	G	TA	C	-
AL_IZ1	A	TT	C	TTGTTTTGT	ATTTAA	T	A	T	G	G	TA	C	G
NC_JJ2	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	G
NC_JJ3	A	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-
SATIVA_3	A	AA	A	ACAAAACAA	-----------	T	T	C	G	G	TA	C	G
SATIVA_4	A	AA	A	ACAAAACAA	-----------	T	T	C	G	G	TA	C	G
AR_JJ4	T	AA	A	ACAAAACAA	ATTTAA	T	A	T	G	G	TA	C	G
GA_JJ7	T	AA	A	ACAAAACAA	ATTTAA	T	A	T	G	G	TA	C	G
MR_JJ6	T	AA	A	ACAAAACAA	ATTTAA	T	A	T	G	G	TA	C	-
AR_M20	T	AA	A	ACAAAACAA	ATTTAA	T	A	T	G	G	TA	C	-
AR_S5	T	AA	A	ACAAAACAA	ATTTAA	T	A	T	G	G	TA	C	-
AL_4CN	T	AA	A	ACAAAACAA	ATTTAA	T	A	C	G	G	TA	C	-

Note: Samples AL_PT20, AL_7CN, AL_IZ1 represent the *C. dentata* type 11 Ruffner Mountain samples.
Samples AL_40, and AL_65 represent the 13 *C. pumila* type Ruffner Mountain samples.

Table 5. Substitutions of variable sites in *Castanea* samples at combined cpDNA regions

	region A B			region C D		region *ndh*f					*orf62*		region *rpl*16		
Samples	295	486	490	**268**	402	240	260	280	288	380	200	518	220-221	236	485
GA_JJ7	++	A	++	G	--	C	++	--	T	++	C	I-20	GA	A	++
NC_JJ3	++	A	++	T	--	C	++	--	T	++	C	--	T --	C	++
NC_JJ1	++	A	++	T	--	C	++	--	T	++	C	--	T --	C	++
AL_M35	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_M30	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_XL1	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_M31	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
MO_JJ5	++	A	++	G	--	C	++	--	T	++	C	I-20	GA	A	++
AL_M65	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_M40	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_M67	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AR_JJ4	++	A	++	G	I-24	C	++	--	T	++	C	I-20	GA	A	D-34
AL_M36	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_M001	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_M34	++	A	++	T	--	A	D-8	--	T	D-6	C	--	T --	C	++
MO_JJ6	++	A	++	G	--	C	++	--	T	++	C	I-20	GA	A	++
NC_JJ2	++	A	++	T	--	C	++	--	T	++	C	--	T --	C	++
AL_M68	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_M37	++	A	++	T	--	A	D-8	--	T	D-6	A	--	T --	C	++
AL_4CN	++	A	++	T	--	C	++	--	T	++	C	--	T --	C	++
AL_6	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_M61	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_IZ1	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_M18	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_M33	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_5	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_M38	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_7CN	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_AL2	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_M60	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++
AL_PT20	D-12	C	D-73	G	--	C	++	--	C	++	C	--	G--	C	++
CT1	D-12	C	D-73	G	--	A	++	I-31	C	++	C	--	G--	C	++

I: insertion; D: deletion

Table 6. Substitutions at variable sites in *Castanea* samples at nuclear regions

Region	126					17		
	Position					Position		
Samples	119	158	172	184	234	332	402	451
GA_JJ7	T	A	T	G	T	A	T	A
NC_JJ3	T	A	T	G	T	A	T	G
NC_JJ1	T	A	T	G	T	A	T	G
AL_M35	T	A	C	G	T	A	T	A
AL_M30	T	A	C	G	T	A	T	A
AL_XL1	T	A	C	G	T	A	C	A
AL_M31	T	A	C	G	T	A	C	A
MO_JJ5	T	A	C	G	T	A	T	A
AL_M65	T	A	C	G	T	A	T	G
AL_M40	T	A	C	G	T	A	T	G
AL_M67	T	A	C	G	T	A	C	A
AR_JJ4	T	A	C	G	T	A	T	G
AL_M36	T	A	C	G	T	A	T	A
AL_M001	C	A	C	G	T	A	T	A
AL_M34	C	A	C	G	T	A	T	A
MO_JJ6	C	A	C	G	T	A	T	A
NC_JJ2	C	A	C	G	T	A	T	A
AL_M68	T	A	C	G	T	A	T	A
AL_M37	T	A	C	G	T	A	T	G
AL_4CN	T	A	T	G	C	T	T	A
AL_6	T	A	T	A	C	T	T	A
AL_M61	T	A	T	R	C	T	T	A
AL_IZ1	T	A	T	R	C	T	T	A
AL_M18	T	A	T	R	C	T	T	A
AL_M33	T	A	T	A	C	T	T	A
AL_5	T	A	T	A	C	T	T	A
AL_M38	T	A	T	A	C	T	T	A
AL_7CN	T	G	T	G	C	T	T	A
AL_AL2	T	G	T	G	C	T	T	A
AL_M60	T	G	T	G	C	T	T	A
AL_PT20	T	A	T	G	C	T	T	A

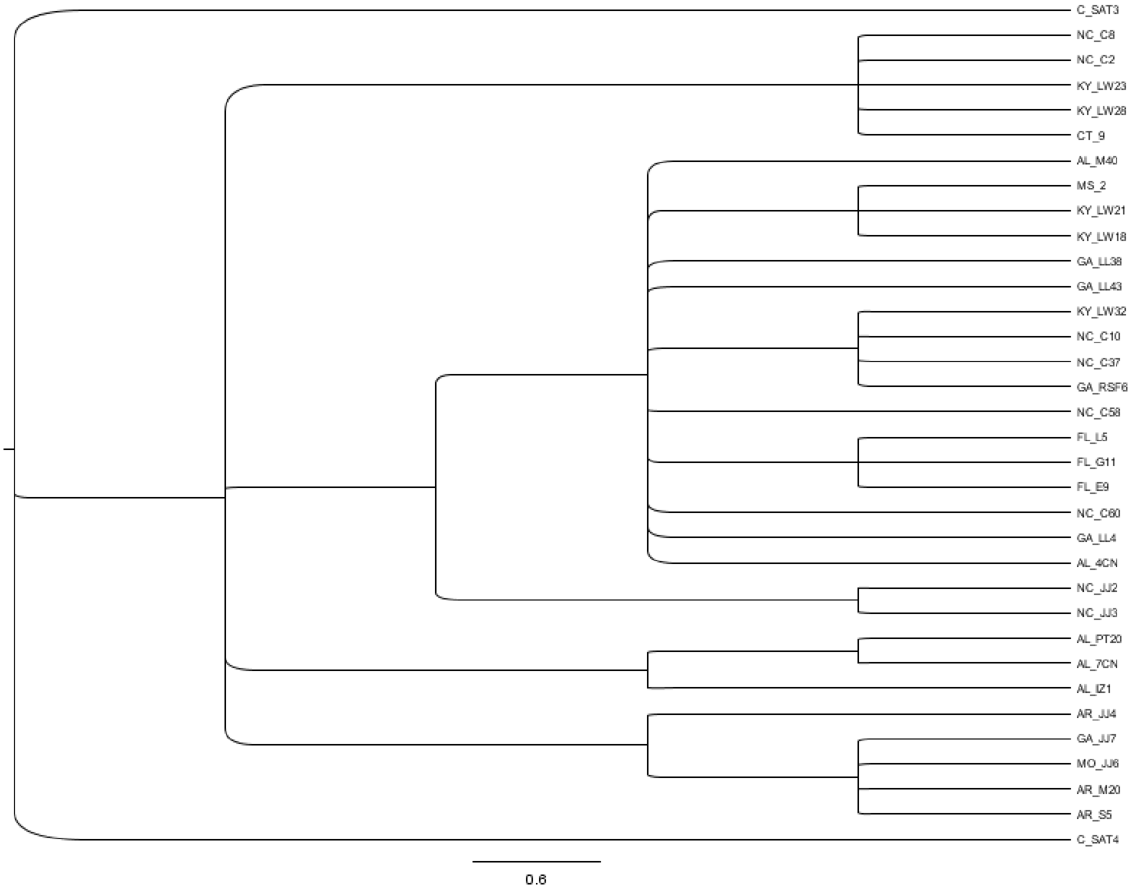

Fig.3. 50% Majority-rule consensus of 170 trees inferred from comparative analysis of *Castanea ndh*C sequences (CI=0.8235 RI=0.9302) using *C. sativa* as outgroup.

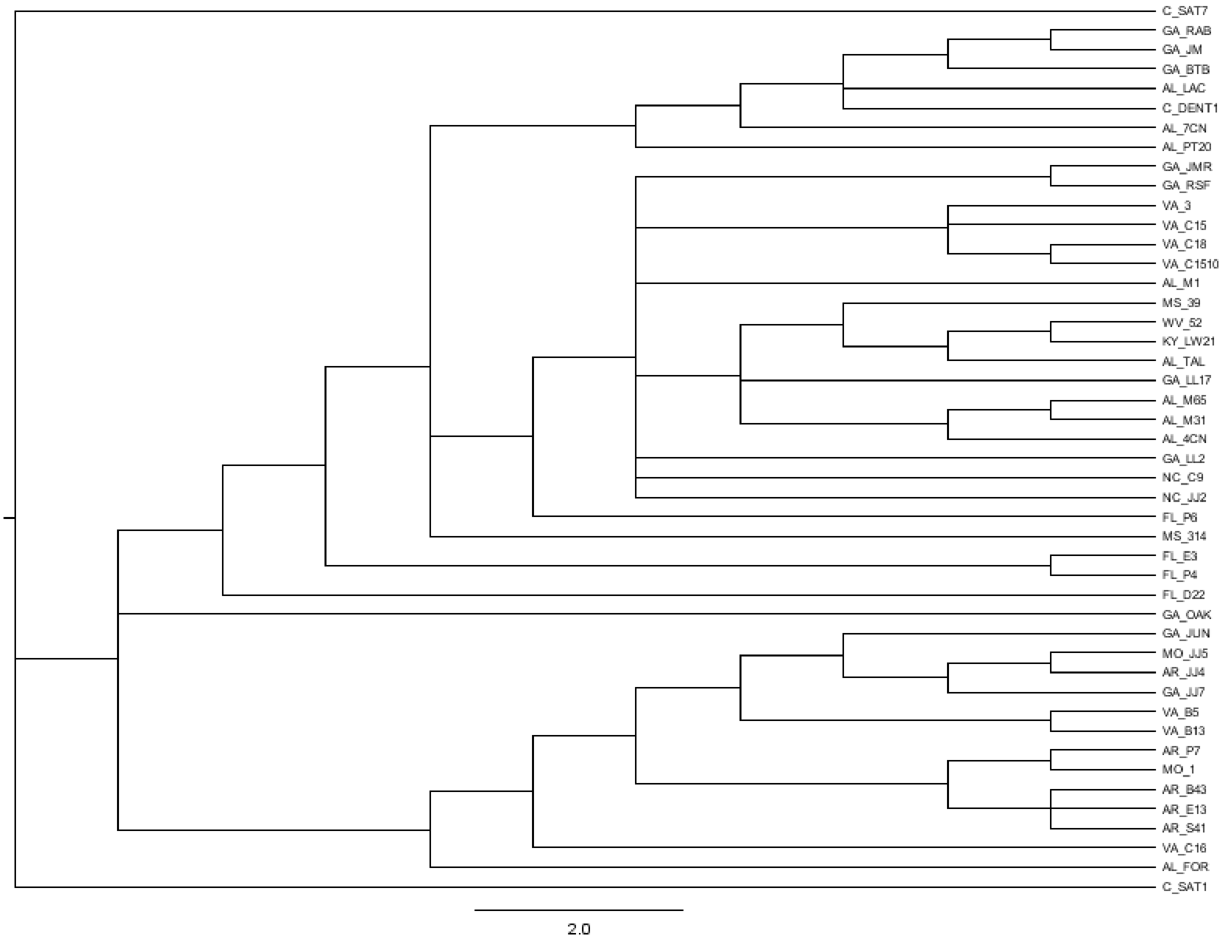

Fig.4. 50% Majority-rule consensus of 620 trees inferred from comparative analysis of *Castanea* NLRO combined sequences (CI=0.4717 RI=0.7804) using *C. sativa* as outgroup.

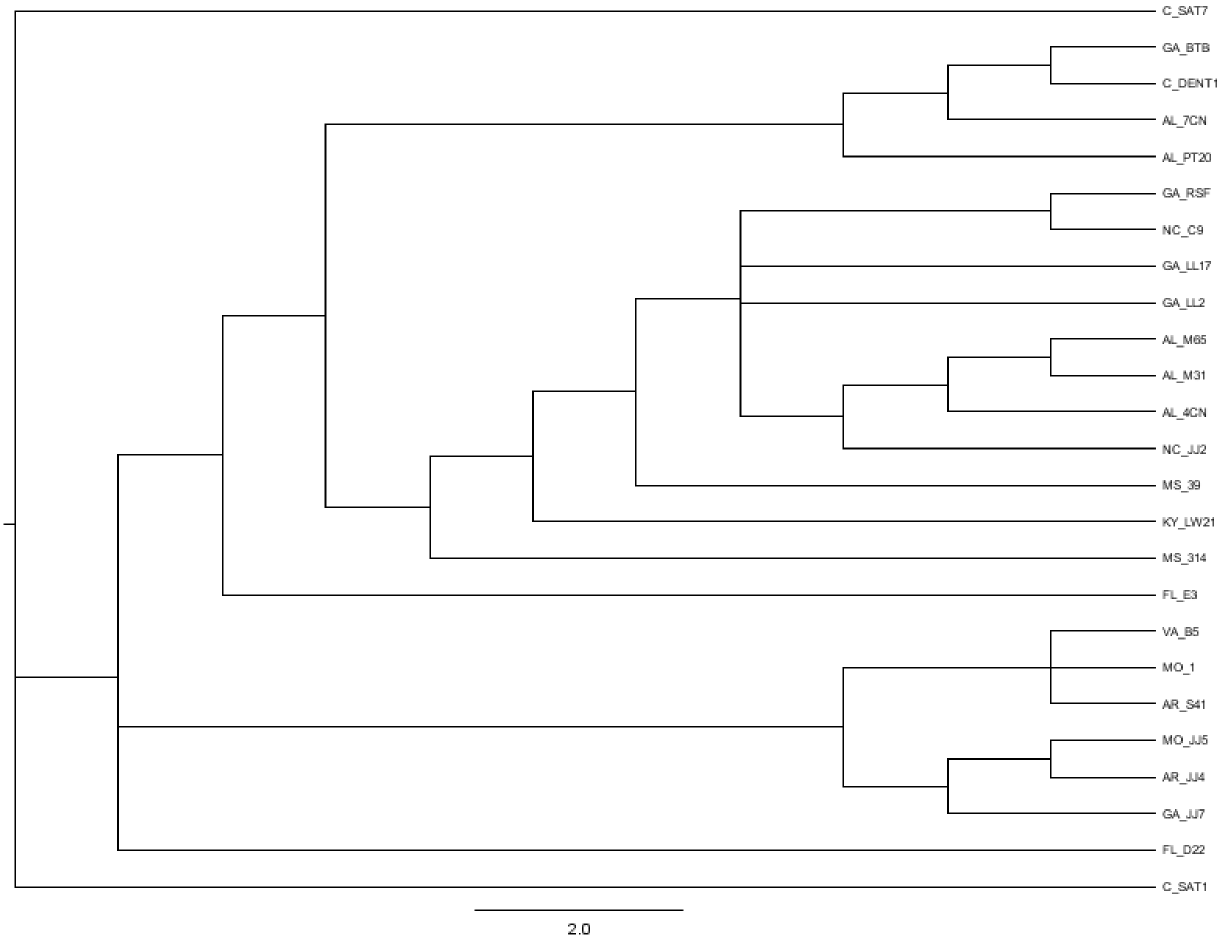

Fig.5. 50% Majority-rule consensus of 180 trees inferred from comparative analysis of *Castanea ndh*C +NLRO sequences (CI=0.6000 RI=0.7821) using *C. sativa* as outgroup.

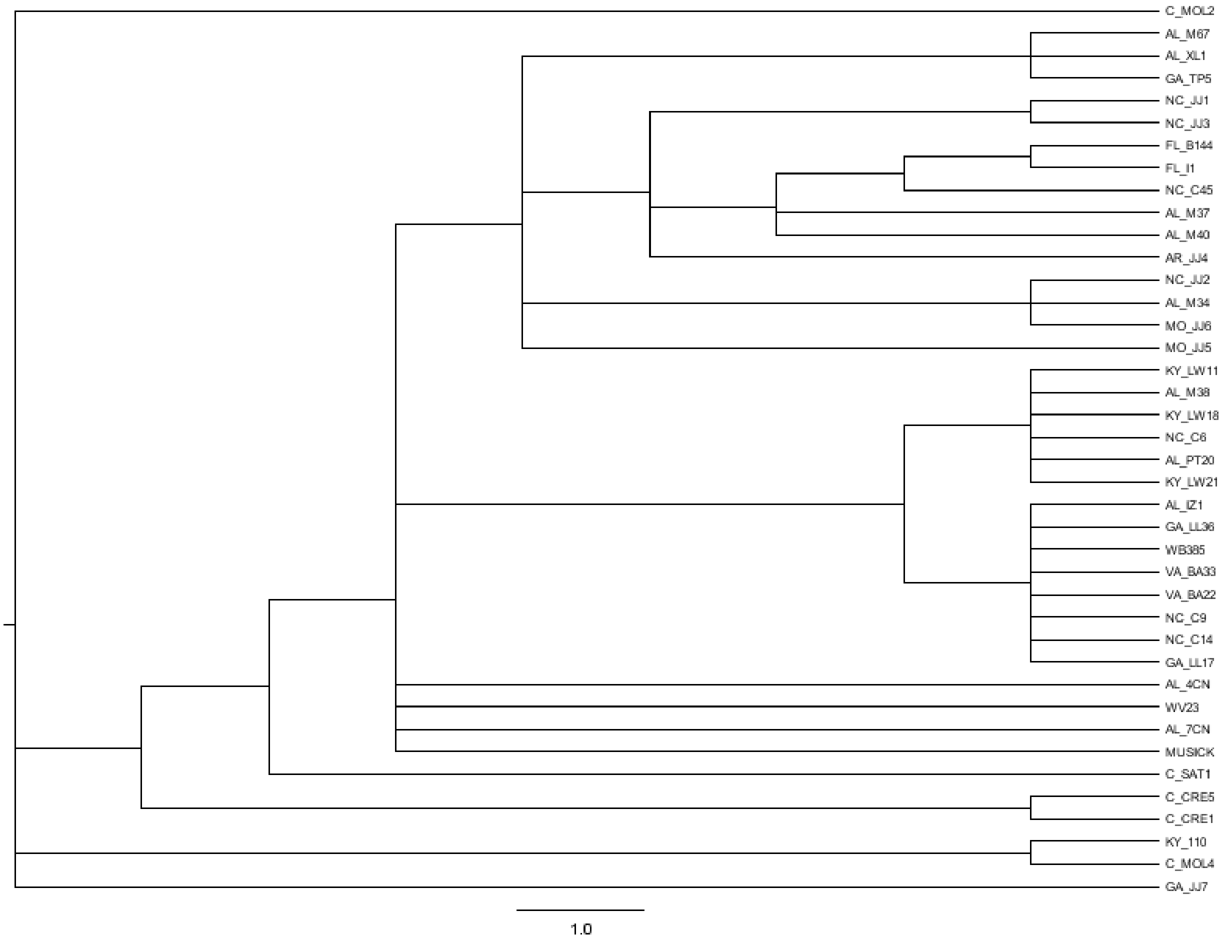

Fig.6. 50% Majority-rule consensus of 333 trees inferred from comparative analysis of *Castanea* nuclear sequences (CI=0.6667 RI=0.9160) using *C. sativa*, *C. mollissima*, and *C. crenata* as outgroup.

Chapter 3

GS-FLX 454 Pyrosequencing and Analysis of *C. pumila* Transcriptome Dataset

Abstract

cDNA samples from leaves of five individual *C. pumila* trees was isolated and sequenced used the 454 GS-FLX at the Genomics Core Facility of PennState University. A total of 1221540 reads, about 372 Mb of cDNA, with an average length of 305 bp, were generated. A total of 125112 unigenes were obtained from the 454 sequencing analyses. Through alignment of the individual reads against contigs from the assembly, 143792 single nucleotide polymorphisms (SNPs) and 2415 complex SNPs (variations of more than two nucleotides) were detected in 47565 contigs, with average length of 222 bp per SNP. Upon alignment of *C. dentata* and *C. pumila* contigs using Model SNP of the CLC genomic workbench software, a total of 267874 inter-SNPs were detected, and 19 *C. pumila* putative SNP sites were selected for primer design of which two were preliminary validated. Meanwhile, *C. pumila* and *C. dentata* contigs were used for multi-alignment with 3714 *Arabidopsis* single copy genes. Contigs of both species with a good match to single copy gene(s) were selected and re-aligned using the CLC genomic workbench software. Ten possible species-specific marker sites were amplified and three were preliminary validated. More species-specific markers should be obtained when more candidate contigs are checked. Gene ontology analysis of the *C. pumila* assembly showed high similarities to transcriptomes of the other *Castanea* species in the NCBI GenBank.

Key words: *Castanea*; species-specific marker; SNPs; Gene ontology

Introduction

The genus *Castanea*, members of the family Fagaceae, is represented in Asia, North America and Europe by three sections and seven species. Section *Eucastanon* contains five species, Chinese chestnut (*C. mollissima*) and Seguin chestnut (*C. seguinii*) in China, Japanese chestnut (*C. crenata*) in Japan and the Korean peninsula, American chestnut (*C. dentata*), and European chestnut (*C. sativa*). Section *Hypocastanon* consists of only the Chinese chinkapin (*C. henryi*) found in a restricted area in southeast China. Section *Balanocastanon* has just one species, *C. pumila*, with two varieties, Allegheny chinkapin (var. *pumila*) and Ozark chinkapin (var. *ozarkensis*) (Johnson, 1988). American chestnuts were the dominant canopy tree species in the eastern hardwood forests with a 200 million acre range, extending from Maine south along the Appalachian Mountains to Alabama and westward to the Mississippi river (Barakat et al., 2009). The American chestnut possessed a remarkable array of desirable traits. It grew very rapidly, and had outstanding form and wood quality. Tannins were extracted from bark and wood chips, and the chips were subsequently pulped for the production of paper. The tree grew well on dry uplands, a trait that today would make it a valuable biofuel species in these regions (Http://www.acf.org). Historically, its seeds provided food and revenue for rural communities, and a wide range of animals were dependent on the mast. Because of its utility, rapid growth, ability to quickly colonize burned or clearcut areas, and edible nuts, the American chestnut has been described as the perfect tree (Tudge, 2007). Chinkapins are less used by humans, mainly because it only grows to shrub size and has small nuts. The Allegheny chinkapin has a wide distribution from northern Florida, west to Texas and Oklahoma, north to Kentucky, Virginia, Maryland, and along the Atlantic coastal plain to Massachusetts (Johnson, 1988). Ozark chinkapin has a limited and fragmented distribution in the Ozark Mountains of eastern Oklahoma and Arkansas (Johnson, 1988). American *Castanea* species, especially *C. dentata* virtually ceased to exist as an economically and ecologically relevant forest tree by the mid-1900s, having fallen victim to chestnut blight (*Cryphonectria parasitica*), an introduced fungal pathogen. The blight killed some four billion trees, one of the greatest ecological disasters in American history (Diskin et al., 2006).The American chestnut persists only as rare escapes and stump sprouts not yet infected with the disease, but just a few trees can live long enough to mature and produce flowers. Meanwhile, the American chinkapin is also susceptible to chestnut blight. For young trees, inter-and intra- specific morphological variability can lead to considerable confusion in discriminating species, Morphological ambiguity is apparent in populations of southern Appalachia, where the distribution range of *C. pumila* overlaps with that of *C. dentata* (Johnson, 1988; Small et al., 2004). Thus it is difficult to discern the two American *Castanea* species using morphological traits. Also European, Chinese and Japanese chestnuts have been planted all along the Appalachian region. Research has recently indicated that natural hybridization between the different *Castanea* species did occur over time and evidence for a separate evolutionary lineage with intermediate morphology is becoming more apparent (Dane, 2009; Binkley, 2008). Accurate and unambiguous identification of *Castanea* species is still needed, especially in southern regions of the Appalachian Mountains.

Novel sequencing technologies have been introduced in the past few years. The current commercially available platforms are marketed by Roche (454), Illumina (Solexa/genome Analyzer), and Applied Biosystems (SOLID). They offer greatly reduced per-base sequencing costs, time savings and provide powerful tools for the detection of mutations (Mardis, 2008). 454-sequencing, which was developed by 454 Life Sciences, has opened new possibilities for high-throughput genome analysis. The approach allows parallel sequencing of millions of individual templates immobilized on microbeads, so it can produce megabases of sequence data per single run (Macas et al., 2007). Till now, it has been successfully applied to the sequencing of many species. To study the molecular mechanism behind the co-adaptation in plant-insect interactions, Zagrobelny et al. (2009) used a genomics strategy founded on 454 pyrosequencing of the *Z. filipendulae* transcriptome to identify enzymes in plants. Ovaskainen et al. (2010) developed a modeling approach that utilizes the self-consistency of the reference database to transfer sequence similarity to the probability of correct identification to a given taxonomic level based on the 454 sequencing data of dead wood-inhabiting fungi. The results showed that it is possible if a high-quality reference database with broad coverage is available. Comparative analysis of 454 pyrosequencing transcriptome of fungal infected and healthy stem tissues collected from blight-sensitive American chestnut and blight-resistant Chinese chestnut was conducted by Bakarat et al. (2009). A large number of genes were identified involved in resistance to *Cryphonectria parasitica*. Meanwhile, many studies have used 454 sequencing for the discovery of single nucleotide polymorphism (SNP) (Oliver et al., 2011; Hyten et al., 2010; Wu et al., 2010) and the identification of genetic modifications (Bundock et al., 2009).

In this study, 454 sequencing analysis was conducted on the leaf transcriptome of *C. pumila* var. *pumila*. Comparative analysis of the transcriptome of this species with that of other species in the Fagaceae database will be used for the detection of informative SNPs.

Materials and Methods

Plant material:

Castanea pumila var. *pumila* leaves were collected by Dr. T.L. Kubisiak at Harrison County, MS and from five trees growing at the Auburn University Paterson Greenhouse complex. Nuts had been collected from Allegheny chinkapin tree populations from the Eglin Air Force Base in FL. Leaf samples from each tree were immediately immersed in liquid nitrogen and frozen at -80°C until use.

RNA preparation and cDNA library synthesis

Total RNA was extracted using the hot-borate method of Wan and Wilkins (1994). About five grams of frozen tissue were weighed, ground to a fine powder under liquid nitrogen, and dispersed into15 ml preheated borate extraction buffer, followed by incubation for 1.5 hours at 42°C with proteinase K, and potassium chloride (KCl) extraction. $LiCl_2$ was used to precipitate and wash the RNA pellet three times. The RNA pellet was dissolved in Tris-HCl (pH 7.5), followed by resuspension in 100µl of DEPC-treated water, and incubated for 5 min at room temperature. 2 µl RNA was used to assess the quality by agarose gel electrophoresis. Poly (A) RNA was separated from total RNA using the MicroPoly (A) Purist kit (AM1919, Ambion).

The RNA samples were shipped to the Genomics Core Facility at PennState University where a NanoDrop spectrophotometer was used to measure the RNA concentration, cDNA library was prepared and sequencing was conducted using a Roche/454 GS FLX sequencer (Roche Diagnostics).

454 sequencing data trimming, assembly and annotation

SFF-formatted sequences were obtained from PennState University for analysis and converted using the SFF converter of the Galaxy software (http://main.g2.bx.pse.edu/) to fasta format. Adaptor sequences (Forward primer A: CCATCTCATCCCTGCGTGTCTCCGACTC AG. Reverse primer B: CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) were removed, as well as poly-A tails and ambiguous sections of the ends of ESTs using the trimest and trimseq model at the default setting of galaxy software. The 454 read sequence data were assembled into transcript contigs using CLC genomics workbench Assembler software (http://www.clcbio.com/index.php?id=1240). The Gene Ontology (GO) (Consortium, 2008) system was used to summarize possible functional classifications of the unigenes via alignment with non-redundant protein database of NCBI. BLASTX parameters were set to e –value =10e-25 and maximum number of descriptions and alignments to report = 250. GO annotation was performed on those HSPs using the e-value selection criteria and supporting sequences described for Blast2GO. Gene identifiers with the strongest BLASTx alignments to the corresponding *C. pumila* 454 reads were used. Comparison of the distribution of biological processes or molecular function obtained using GO annotation was done using the GOstat program.

SNP discovery

To detect SNPs in the cDNA pool, the consensus assembly generated from all sequencing runs was used as reference sequence to which individual reads were aligned using the CLC bio workbench. Each read was aligned to only a single best homologous site in the reference sequence. Reads aligning in more than one location to the reference genome were discarded. The CLC bio only scores polymorphisms when two or more reads contain the variant allele. If more than 50 reads are assembled together, at least 10% of the reads contain the variant allele. In this study, all indels and variants involving more than one nucleotide were discarded, only single nucleotide polymorphisms (SNPs) were reported.

Species-specific marker discovery

Comparison with *C. dentata* sequences. To discover the species-specific markers for *C. pumila* and *C. dentata*, the *C. dentata* sequences 'AC454 contigs V3.fasta' was downloaded from the Fagaceae website (www.fagaceae.org). Alignment of the *C. pumila* and *C. dentata* sequences was conducted using the CLC genomics software workbench. Only differences between the two species, lacking intra-SNP at the same site on either sequence of the two species sequences were reported. To validate the possible species-specificity of the identified SNP markers, *C. pumila* contigs with high target hit number were selected. Primers were designed around the SNP sites and sequencing analysis of PCR fragments was conducted.

Basic Local Alignment Search Tool (BLAST) of *C. dentata* and *C. pumila* 454 fasta contigs and single copy gene fasta file of Arabidopsis.

The single copy gene fasta file of Arabidopsis was downloaded from NCBI. BLAST analysis of the single copy Arabidopsis fasta file with *C. dentata* and *C. pumila* contigs was conducted using the CLC workbench software. Only sequences with E-value of zero were selected. The contigs of both species were blasted and checked manually. Only sequences with rare SNPs at conserved regions were selected and analyzed. To validate their species-specificity, primers were designed around the SNP regions. Regions were amplified and fragments from different species were aligned.

Results

Sequencing analysis and assembly

The *C. pumila* cDNA library was constructed from a pool of RNA isolated from leaves of a population of trees in MS and from a small population in FL. Because of a higher level of rRNA purified from leaves of MS trees, we decided to conduct sequencing analysis on RNA extracted from Eglin Air Force (FL) Base trees. One half plate of sequence analysis was conducted using the 454 GS FLX system. Samples were redone using with 2×1/4 plate. A total of 1221540 reads, about 372 Mb of cDNA, was generated. The read length is between 36-603 bp, with an average length of 305 bp (Figure 1). A total of 47565 contigs from the 1143993 matched reads, with an average contig length of 670 bp was generated. A total of 77547 reads didn’t overlap other sequences and were considered as singletons (coverage depth =1), and 125112 unigenes were obtained from the 454 sequencing analysis.

SNP discovery (47565 contigs with 143791 SNPs)

The CLC genomic workbench software was used to identify SNPs among ESTs by aligning individual reads against contigs from the assembly. To make sure a sequence difference is a true polymorphism, at least four individual reads with alignments to the consensus sequences must have the variant allele and at least 4 others must have the allele of the consensus. By following this criterion, 143792 SNPs were detected in a total of 47565 contigs, with average length of 222 bp per SNP. The proportion of transition nucleotide substitutions (29273, 21%) is much less than the proportion of transversions (109775, 78.9%), moreover, there are 2415 complex SNPs (variations of more than two nucleotides).

Species-specific marker detection

Comparison with *C. dentata* sequences

Upon alignment of *C. dentata* and *C. pumila* contigs using Model SNP of the CLC genomic workbench software, a total of 267874 inter-SNPs were detected. Based on the coverage number of the two sequences from each SNP, *C. pumila* contigs with a high number count both on number 1 and number 2 were selected. These contigs have the highest probability to contain species-specific markers. To validate SNPs detection by CLC genomic workbench, 19 primer pairs were designed around the SNP location on the *C. pumila* contigs, and each primer pair covered one or two SNP sites. DNA from *C. dentata* and *C. pumila* samples was used for PCR amplification and sequencing using ABI sequencer in Auburn University. Sequences were aligned using vectorNTI software to check for putative species-specific markers. Two sequences (10.5%) and two species-specific marker sites were validated (Fig 2).

BLAST of *C. dentata* and *C. pumila* 454 fasta contigs and single copy gene fasta file of Arabidopsis.

A total of 3714 Arabidopsis single copy genes were downloaded from the NCBI website. We blasted *C. pumila*, *C. dentata* and Arabidopsis single copy genes together. Of the multi blast result, there are 2968 *C. pumila* contigs which blast to Arabidopsis single copy genes with a “0” E-value, 8608 contigs did not match any single copy genes, and other contigs showed high E-values from 1.15E-180 to 9.99. We checked the *C. pumila* multi blast results of the contigs with the “0” E-value manually and selected pairs of contigs of *C. dentata* and *C. pumila* which matched and shared with single copy genes with Arabidopsis single copy gene database. Each contig pair was aligned with the 'Create alignment' Model of CLC bio again and many possible species-specific marker sites were obtained. To validate the putative species-specific marker sites, 6 primer pairs were designed based on the SNP containing region between the *C. dentata* and *C. pumila* contigs. We PCR amplified three samples of both *C. dentata* and *C. pumila*. Sequencing of the samples and alignment were conducted to check the SNPs (putative species-specific markers). Three species-specific marker sites were preliminary detected (Fig 7).

Functional annotation and gene ontology analyses

Transcripts of the *C. pumila* assembly contigs were annotated via BLASTx search against the NCBI non-redundant protein database using the Blast2GO algorithm. Blast result accessions were used to retrieve associated gene names and gene ontology (GO) terms. The annotated sequences were classified into three general categories associated with cellular, molecular and biological functionalities. The biological processes constituted the most abundant component of the GO assignment of the transcripts (4916 counts, 42.7%), followed by the cellular components (3240counts, 28.1%) and the molecular function component (3361, 29.2%). The largest proportion of GO assigned sequences fell into broad categories for all three major GO functional domains as presented in Figure 8. Among the biological process categories, 28.1% genes are associated with metabolic processes and 27.2% are related to cellular processes. Of the molecular functional category, 52.4% are related to catalytic activity, followed by 38.3% associated with protein binding. Within the cellular component, 46.7% of the genes are related to the cell and 41.5% to the membrane-bounded organelles (level 2). The BLASTx top-hit species distribution of gene annotations showed the highest homology to grape (*Vitis vinifera*), followed by the castor bean (*Ricinus communis*) and poplar (*Populus trichocarpa*) (Fig 9). Moreover, the *C. pumila* sequences showed significant homologies to the following three species of genus *Castanea*: Chinese chestnut (*C. mollissima*), European chestnut (*C. sativa*), and Japanese chestnut (*C. crenata*).These results indicate that the *C. pumila* genes have a high level of phylogenetic conservation compared to these species. But on the other hand, the closely related *C. dentata*, showed very low homology to the *C. pumila* proteins, which maybe related to the limited number of *C. dentata* proteins which are currently deposited in the NCBI database.

Detection of candidate disease resistance related genes in *C. pumila*

When the 47565 *C. pumila* contigs were blasted into the NCBI database using the CLC genomic workbench software, a total of 27567 contigs hit proteins with an E-value less than 1E-10. Following comparisons with the candidate genes involved in chestnut response to *Cryphonectria parasitica* infection (Barakat et al., 2009), 428 contigs showed significant homologies to several disease resistance genes. Some genes related to hypersensitive cell death, such as ABC transporter, C2-domain-containing gene, elongation factor-1 alpha, and peroxidase were expressed. These genes are involved in controlling the extent of the cell death in the defense response. Several genes involved in plant resistance are pathogen encode proteins involved in lignin biosynthesis, such as cinnamyl alcohol dehydrogenase (CAD), cinnamoyl-CoA reudctase (CCR), o-methyltransferase 1, cytochrome P450, 4-counmarate-CoA ligase, succinyl-CoA ligase, and S-adenosyl-methionine synthase 3. Polyphenol oxidases (PPO) catalyzing the oxygen-dependent oxidation of phenols to quinines are known to increase plant resistance against some pathogens. ATP-binding cassette transport proteins and omega-3 fatty acid desaturase, which are required for systemic resistance, were identified. ATPase is required for the attenuation of the hypersensitive response. Several genes involved in the regulation of resistance gene expression such as SNF, Zinc finger, and Myb were also identified. Moreover, other disease resistance related genes such as beta-glucanase, catalase, chitinase, disease resistance protein, were detected in the *C. pumila* transcriptome (Fig 11). Most of these genes play an important role in plant response to pathogen infection, and they are very useful for future research, especially related to breeding for chestnut blight resistance.

Discussion

Advances in DNA sequencing technology, especially bead-based pyrosequencing have dramatically impacted genome sequencing and transcriptome analyses. Unlike other techniques such as microarrays and SAGE, 454 pyrosequencing has been successfully used to analyze the transcriptome of non-model plant species (Hyten et al., 2010; Parchman et al., 2010). The large number of reads generated per run together with the low sequencing error rate of the contigs makes it a good tool to deeply sequence the transcriptome of plants. It has been used successfully for analyzing the transcriptomes of maize, Arabidopsis, and non-model tree species as *C. dentata* and *C. mollissima* (Barakat et al., 2009). *C. pumila* is closely related to the American chestnut. Only a few hundred sequences from *C. pumila* have been deposited in the EST database at NCBI. The study generated a large number of cDNA resources and analyzes the transcriptome of *C. pumila* for the first time and can be used to relate research about *C. pumila*, and even the genus *Castanea* in America on its ecology, evolution and phylogeography.

Because of the large amount of data obtained, low cost per run, deep and redundant coverage produced over many genes, Next Generation sequencing is considered ideal for SNP discovery and analysis (Wall et al., 2009). In this study, although our sampling was limited to leaves of five different individuals, and only half plate cDNA sample was conducted, more than 140 thousand high quality SNPs were detected in our contigs. Moreover, about one third of these SNPs reside in annotated genes, and some hit to shared single copy genes, which will allow for the identification of open reading frames and facilitate more detailed analyses on the significance of molecular variation. The SNP frequency in the *C. pumila* transcriptome is 0.45/100 bp, similar as that reported for other studies using 454 pyrosequencing of cDNA pooled from multiple individuals, such as 0.6/100 bp in *Pinus taeda* (Parchman et al., 2010) , 0.33/100 bp in maize (Barbazuk et al., 2007), 0.72/100 bp in *Sarcophaga crassipalpis* (Hahn et al., 2009). For SNP discovery using the transcriptome fewer SNPs will be obtained since genes are more conserved than non-coding DNA, in additional, without a genomic reference sequence, the proportion of successful SNP assays will also decrease because of the present of introns interfering with oligo hybridization (Hyten et al., 2010).The large numbers of SNPs should facilitate population genomic and gene-based association studies in *C. pumila*.

Many nuclear genes in angiosperms are members of gene families and may exhibit copy number variation. This complicates the identification of potentially orthologous nuclear genes that could be used for applications such as molecular systematics and mapping of markers. Meanwhile, other conserved single copy genes, which are truly shared in single copy throughout seed plant are ideal nuclear phylogenetic markers. Because of pervasive gene duplication in the angiosperms, it is possible to obtain single copy genes shared among Arabidopsis, *C. pumila*, and *C. dentata*. Multi--blast results showed 223 single copy gene hits among contigs of both *C. pumila* and *C. dentata* and confirm this hypothesis. One of the main aims of this study is detection of species-specific markers between *C. pumila* and *C. dentata*. Because of the slow rate of molecular evolution in single copy genes, the conserved sequence will also lead to primers or probes hybridizing to both the gene sequence that contains the SNP as well as any conserved paralogous sequences, thus decreasing the success rate of SNP. The mutations between the two species which appear in single copy gene regions, especially SNPs, have the highest possibility to be species-specific markers. Moreover, shared single copy nuclear genes have many valuable applications as mapping markers and phylogenetic markers (Duarte et al., 2010). So the large scale transcriptome datasets of *C. pumila*, combined with the chloroplast data, can be used for phylogenetic studies of the genus *Castanea* in North America.

Although the *C. pumila* sequences had significant homologies to the species of genus *Castanea*, there are 35617 contigs (account for 74.9%) that did not show significant similarity to any protein in the databases and could not be annotated. Studies have shown that shorter sequences are less likely to align with a significant E-value (Blanca et al., 2011). However, in this study, the average length of the contigs is 607 bp, with 50% of the contigs longer than 550 bp. For homology searches against known genes, unigenes longer than 200 bp could be assigned effectively for functional annotations (Li et al., 2010). Previous studies using closely related species, Chinese chestnut and American chestnut, showed that there are more than 50% of 454 reads which could not be annotated using either the *Arabidopsis* proteome or the *Populus* proteome. Only a small fraction of 454 reads could be queried against the fungi database at NCBI (Barakat et al., 2009). For those contigs which did hit databases of NCBI, most of them probably lacked conserved functional domains. Another possible reason is that some of these unigenes might be non-coding RNAs, and maybe some sequences might contain potential chestnut-specific or chinkapin-specific genes, but there are no *C. pumila* sequences, and a limited number of *Castanea* and even Fagaceae sequences in the EST database at NCBI (Barakat et al., 2009).

These contigs were further classified into different functional categories using plant-specific GO slims and can provide a broad overview of the ontology content (Riggins et al., 2010). It can be deduced from the figure of the functional classification of *C. pumila* virtual unigenes into plant specific GO slims within the biological process category, cellular processes and metabolic processes in the most highly represented group. This indicated that the *C. pumila* leaves were undergoing rapid growth and strong metabolic activities. Moreover, genes involved in other important biological processes such as response to stimulus, biological regulation, immune system, and developmental process were also identified through GO annotations. A lot of these genes are known to be involved in response to biotic or abiotic stimuli and stress in general, so they are an important resource to research chestnut blight resistance in the genus *Castanea*.

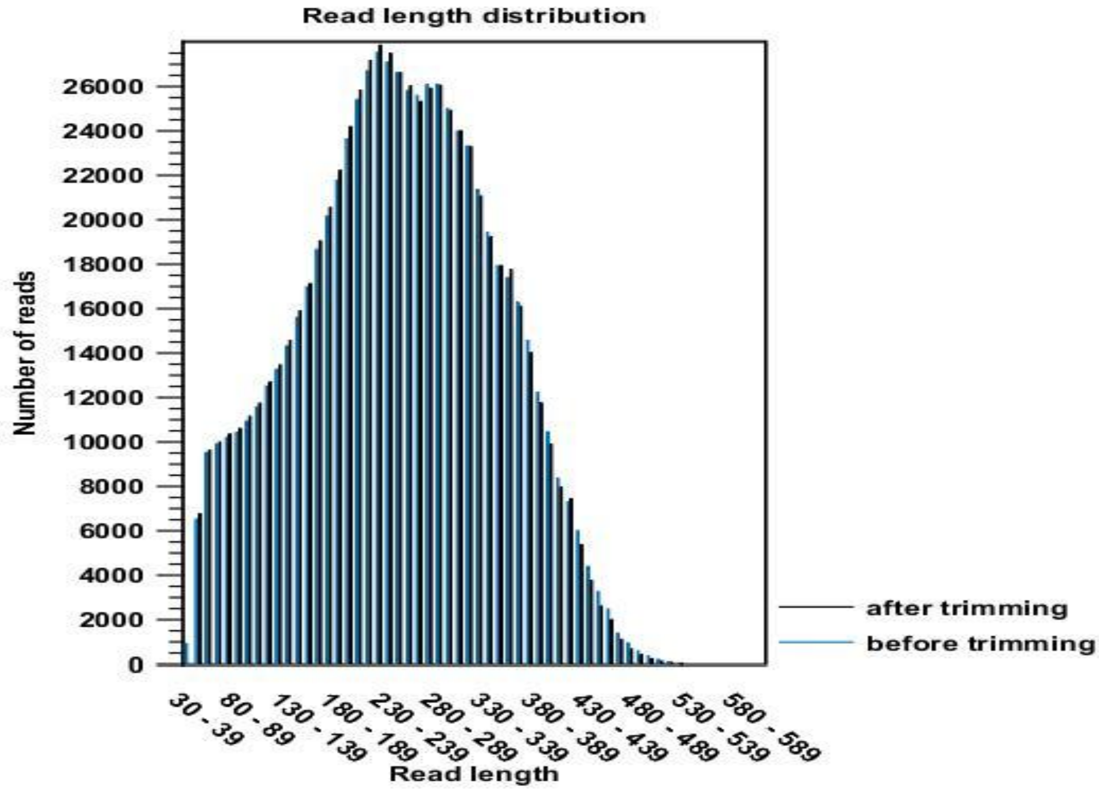

Fig.7 Average length distribution of pyrosequencing reads of *the C. pumila* var. *pumila* leaf transcriptome.

Table 7. Preliminarily validated species-specific markers

Primer NO.	*C. pumila* contig	Position	Mutation	Primer Sequences	Covered Gene in NCBI	Technology
1	11672	2300	T/C	F: TCCATGCTGAGAGAGGAGGT R: CTCCTTCCTCATACCCCACA	*Vitis vinifera* hypothetical protein (XM 002284207.1)	Tech 1*
2	44421	274	G/A	F: TCCGAGCTGGATCTTATGCT R: GCTTCCTTCACAGCCAATTC	Chloroplast latex aldolase-like protein (AY 818399.1)	Tech 1
3	51738	415	C/A	F: AATGGCACAAAATGGGAGAG R: CTGGTTTGGTTGGAGCAAAT	*Vitis vinifera* hypothetical protein (XM 002277155.1)	Tech 2*
4	11569	900	T/C	F: GGTATGCCAGAACGCAAAAT R: CCCTCGCATCCTGATCTTAG	dihydrorotate dehydrogenase (XM 002533805.1)	Tech 2

Tech 1: Comparison *C. pumila* sequences from Ruffner Mountain with *C. dentata* sequences.
Tech 2: BLAST of *C. pumila* 454 fasta contigs and *C. dentata* and single copy gene fasta file of Arabidopsis.

```
                              (Primer 1)
   AL2-P1   (106) CATAAACGTGTTGTGGGTCAAGATCCTGCAGTGAAATCAGTAGCTGAGGC
    M5-P1   (106) CATAAACGTGTTGTGGGTCAAGATCCTGCAGTGAAATCAGTAGCTGAGGC
   CT1-P1   (102) CATAAACGTGTTGTGGGTCAAGATCCTGCAGTGAAATCAGTAGCTGAGGC
   M65-P1   (107) CATAAACGTGTTGTGGGTCAAGACCCTGCAGTGAAATCAGTAGCTGAGGC
   M40-P1   (106) CATAAACGTGTTGTGGGTCAAGACCCTGCAGTGAAATCAGTAGCTGAGGC
    MI-P1   (104) CATAAACGTGTTGTGGGTCAAGACCCTGCAGTGAAATCAGTAGCTGAGGC
 Pu-11672   (756) CATAAACGTGTTGTGGGTCAAGACCCTGCAGTGAAATCAGTAGCTGAGGC
Consensus   (108) CATAAACGTGTTGTGGGTCAAGATCCTGCAGTGAAATCAGTAGCTGAGGC


                              (Primer 2)
   AL2-P2   (373) GTATAGAAAATTGTTGCATCACCAGGGCGTGGAATTTTGGCCATGGATG
   CT1-P2   (374) GTATAGAAAATTGTTGCATCACCAGGGCGTGGAATTTTGGCCATGGATG
   M65-P2   (373) GTATAGAAAATTGTTGCATCGCCAGGGCGTGGAATTTTGGCCATGGATG
    MI-P2   (429) GTATAGAAAATTGTTGCATCGCCAGGGCGTGGAATTTTGGCCATGGATG
  PU-4421   (827) GTATAGAAAATTGTTGCATCGCCAGGGCGTGGAATTTTGGCCATGGATG
Consensus   (432) GTATAGAAAATTGTTGCATCGCCAGGGCGTGGAATTTTGGCCATGGATG


                              (Primer 3)
  AC_45170 (117) ATGGGAGAGCCTCAGGTTCCTTGGATGACAACGCGGCTGTTCCTAATCCA
    IE1-P3 (165) ATGGGAGAGCCTCAGGTTCCTTGGATGACAACGCGGCTGTTCCTAATCCA
    AL2-P3 (163) ATGGGAGAGCCTCAGGTTCCTTGGATGACAACGCGGCTGTTCCTAATCCA
     MS-P3 (166) ATGGGAGAGCCTCAGGTTCCTTGGCTGACAACGCGGCTGTTCCTAATCCA
    M36-P3 (183) ATGGGAGAGCCTCAGGTTCCTTGGCTGACAACGCGGCTGTTCCTAATCCA
  pu_51738 (387) ATGGGAGAGCCTCAGGTTCCTTGGCTGACAACGCGGCTGTTCCTAATCCA
 Consensus (392) ATGGGAGAGCCTCAGGTTCCTTGGCTGACAACGCGGCTGTTCCTAATCCA


                              (Primer 4)
  AC-32044 (104) GTAAATGCCAAAGCTACGGTTCCTGTTTGGGCCAAGATGACTCCTAACAT
    AL2-P4  (71) GTAAATGCCAAAGCTACGGTTCCTGTTTGGGCCAAGATGACTCCTAACAT
     MI-P4  (71) GTAAATGCCAAAGCTACGGTTCCCGTTTGGGCCAAGATGACTCCTAACAT
  pu_11569 (879) GTAAATGCCAAAGCTACGGTTCCCGTTTGGGCCAAGATGACTCCTAACAT
 Consensus (883) GTAAATGCCAAAGCTACGGTTCCTGTTTGGGCCAAGATGACTCCTAACAT
```

Fig.8 Aligned sequences of *Castanea* samples to preliminarily validate the putative species-specific markers

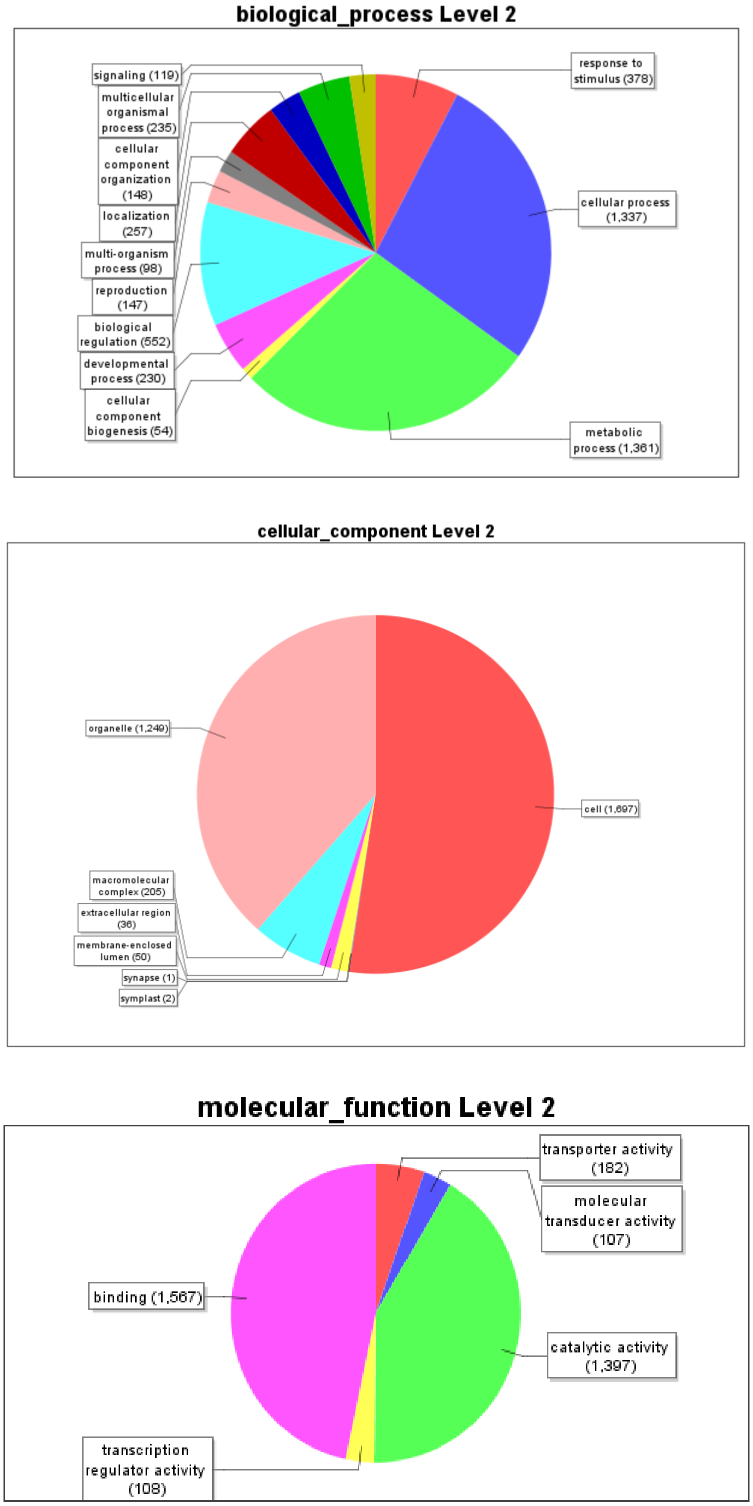

Fig.9 Gene Ontology (GO) assignment (2nd level GO terms) of the *C. pumila* 454-pyrosequencing assembly.

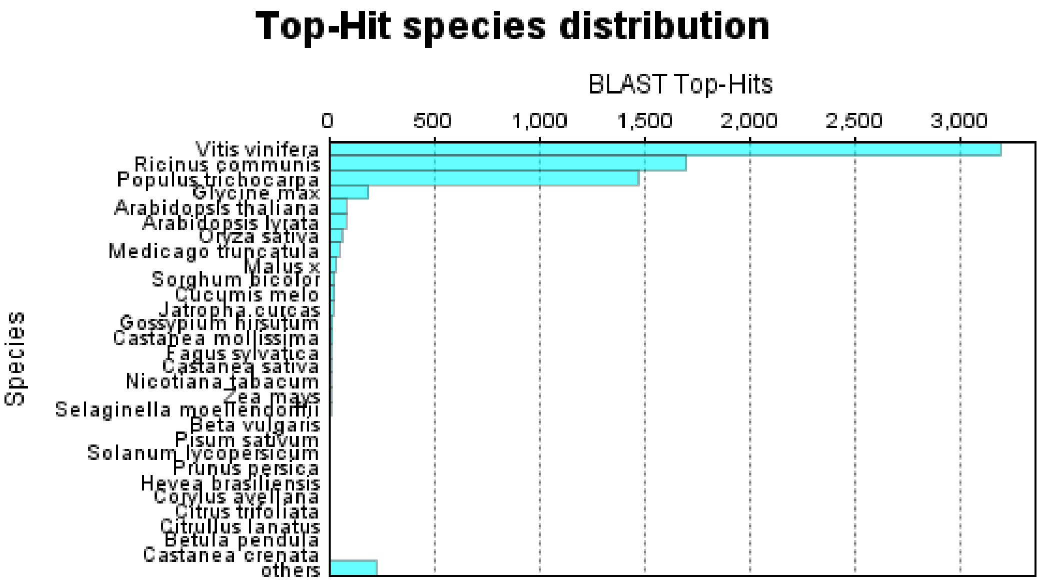

Fig.10 BLASTx top-hit species distribution of gene annotations showing high homology to known genome sequences

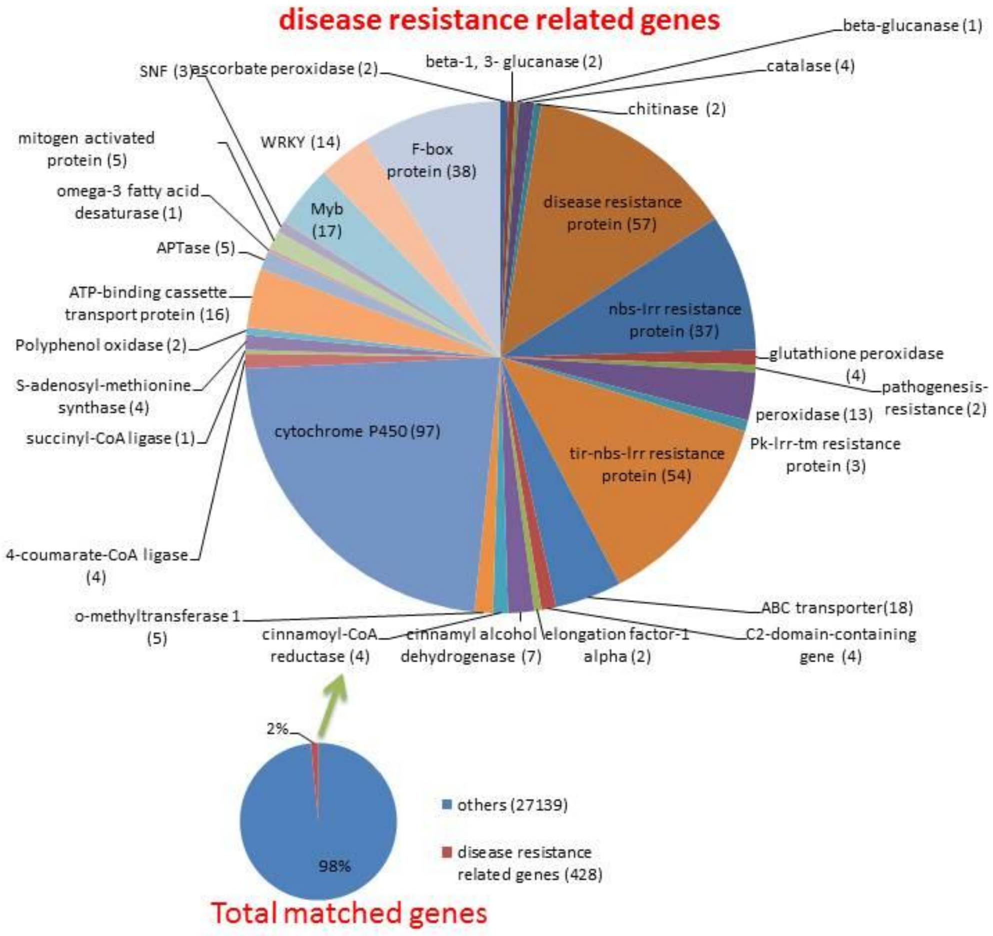

Fig. 11 Similarity match of *Castanea pumila* contig to genes related to disease-resistance

References

AMES, M., A. SALAS, AND D. M. SPOONER. 2007. The discovery and phylogenetic implications of a novel 41 bp plastid DNA deletion in wild potatoes. *Plant Systematics and Evolution* 268: 159-17.

ANAGNOSTAKIS, S. A. 1987. Chestnut blight: the classical problem of an introduced pathogen. *Mycologia* 79: 23–37.

ARIF, I. A., M. A. BAKIR, H. A. KHAN, A. H. ALRARHAN, A. A. HOMAIDAN, A. H. BAHKALI, M. A. SADOON, AND M. SHOBRAK. 2010. A Brief review of molecular techniques to assess plant diversity. *International Journal of Molecular Sciences* 11: 2079-2096.

AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, J. E. NEIGEL, C. A. REEB, AND N. C. SAUNDERS. 1987. Intraspecific phylogeography- the mitochondrial-DNA bridge between population genetics and systematics. *Journal of Biogeography* 18: 489-522.

AVISE, J. C. 2009. Phylogeography: retrospect and prospect. *Journal of biogeography* 36: 3-15.

BALDWIN, B. G., M. J. SANDERSON, J.M.PORTER, M. F. WOJCIECHOWSKI, C. S. CAMPBELL, AND M. J. DONOGHUE. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82: 247-727.

BARAKAT, A., D. S. DILORETO, Y. ZHANG, C. SMITH, K. BAIER, W. A. POWELL, N. WHEELER, R. SEDEROFF, AND J. E. CARLSON. 2009. Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biology* 9: 51.

BARBAZUK, W. B., J. EMRICH S, H. D. CHEN, L. LI, AND P. S. SCHNABLE. 2007. SNP discovery via 454 transcriptome sequencing. *Plant Journal* 5: 910-918.

BINKLEY, M. 2008. Phylogeography of Northeast America *Castanea*. Ms Thesis of University of Tennessee at chattanooga.

BORG, A. J., L. A. MACDADE, AND J. SCHONENBERGER. 2008. Molecular phylogenetics and morphological evolution of Thunbergioideae (Acanthaceae). *Taxon* 57: 811-822.

BORSCH, T., AND D. QUANDT. 2009. Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Systematics and Evolution* 282: 169-199.

BUNDOCK, P. C., F. G. ELIOTT, G. ABLETT, A. D. BENSON, R. E. CASU, K. S. AITKEN AND J. HENRY. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal* 7: 347-

354.

BURBAN, C., AND R. J. PETIT. 2003. Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. *Molecular Ecology* 12:1487-1495.

BURNHAM, C. R. 1988. The restoration of the American chestnut. *American Scientist* 76: 478-487.

CASASOLI, M., C. MATTIONI, M. CHERUBINI, AND F. VILLANI. 2001. A genetic linkage map of European chestnut (Castanea sativa Mill.) based on RAPD, ISSR and isozyme markers. *Theoretical and Applied Genetics* 102: 1190-1199.

CAMPAGNA, M. L., AND S. R. DOWNIE. 1998. The intron in chloroplast gene *rpl*16 is missing from to flowering plant families Geraniaceae, Goodeniaceae , and Plumbaginaceae, *Transactions of the Illinois State Academy of Science* 91: 1-11.

CHENG, Y. P., HWANG, S. Y, AND T. P. LIN. 2005. Potential refugia in Taiwan revealed by the phylogeographical study of *Castanopsis carlesii* Hayata (Fagaceae). *Molecular Ecology* 14: 2075-2085.

CHIANG, T. Y., K. H. HUNG, T. W. HSU, AND W. L. WU. 2004. Lineage sorting and phylogeography in *Lithocarpus formosanus* and *L. dodonaeifolius* (Fagaceae) fromTaiwan. *Annals of the Missouri Botanical Garden* 91: 207-222.

CONNORS, B. J., C. A. MAYNARD, AND W. A. POWELL. 2001. Expressed sequence tags from stem tissue of the American chestnut, *Castanea dentata*. *Biotechnology Letters* 23:1407-1411.

COWAN, R. S., M. W. CHASE AND W.J. KRESS. 2006. 300000 Species to identify: problems, progress, and prospects in DNA barcoding of land plants. *Taxon* 55: 611-616.

DANE, F., P. LANG, H. HUANG, AND Y. FU. 2003. Intercontinental genetic divergence of *Castanea* species in eastern Asia and eastern North America. *Heredity* 91: 314-321.

DANE, F., L. K. HAWKINS, AND H. HUANG. 1999. Genetic variation and population structure of *Castanea pumila* var. *ozarkensis*. *Journal of the American Society for Horticultural Science* 124: 666-670.

DANE, F., AND P. LANG. 2004. Sequence variation at cpDNA regions watermelon and related wild species: implications for the evolution of *Citrullus* haplotypes. *American Journal of Botany* 91:1922-1929.

DANE, F. 2009. Comparative phylogeography of *Castanea* species. Proceeding of the Fourth International Chestnut Symposium pp211-222.

DANE, F., AND P. LANG. 2008. Biodiversity and evolution of *Castanea*. In: Plant Genome

Biodiversity and Evolution. Phanerograms (Gymnosperm) and (Angiosperm-Monocotyledons) Eds A.K. Sharma and A. Sharma. Chapter 4. Science Publishers, Enfield. NH pp 79-100.

DAVIS, M. B. 1983. Quaternary history of deciduous forests of eastern North American and Europe. *Annals of the Missouri Botanical Garden* 70: 550-563.

DELCOURT, H. R., AND W. F. HARRIS. 1980. Carbon budget of the southeastern U.S. biota: Analysis of historical change in trend from source to sink. *Science* 210: 321-323.

DISKIN, M., K. C. STEINER, AND F.V. HEBARD. 2006. Recovery of American chestnut characteristics following hybridization and backcross breeding to restore blight-ravaged *Castanea dentata*. *Forest Ecology and Management* 223: 439-447.

DUARTE, J. M., P. K. WALL, P. P. EDGER, L. L. LANDHERR, H. MA, J. C. PIRES, J. L. MACK, AND C. W. DEPAMPLILIS. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10: 61.

DURAN, C., N. APPLEBY, AND T. CLARK. 2009. AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Research* 39: D951-D954.

EIDESEN, P. B., I. G. ALSOS. M. POPP, J. SUDA, C. BROCHMANN. 2007. Nuclear vs. plastid data: complex Pleistocene history of a circumpolar key species. *Molecular Ecology* 16: 3902-3925.

EDWARDS, D., A. HORN, AND D. TAYLOR. 2008. DNAbarcoding of a large genus, *Aspalathus* L. (Fabaceae). *Taxon* 57: 1317-1327.

ELANSARY, H. O., K. MULLER, M. S. OLSON, AND H. STORCHOVA. 2010. Transcription profiles of mitochondrial genes correlate with mitochondrial DNA haplotypes in a natural population of *Silene vulgaris*. *BMC Plant Biology*10.

FEDOROVA, A. V., I. A. SCHANZER, A. A. KAGALO. 2010. Local differentiation and hybridization in wild rose populations in western Ukraine. *Wulfenia* 17: 99-115.

FUKUDA, T., J. TOLOYAMA, AND H. OHASHI. 2001. Phylogeny and biogeography of the genus *Lycium* (Solanaceae): inferences from chloroplast DNA sequences. *Molecular Phylogenetics and Evolution* 19: 246-258.

GAINES, C. A., M. P. HARE, S. E. BECK, AND H. C. ROSENBAUM. 2005. Nuclear markers confirm taxonomic status and relationships among highly endangered and closely related right whale species. *Proceedings of the Royal Society B-Biologyical Sciences* 272: 533-542.

GIELLY, L., AND P. TABERLET. 1994. The use of chloroplast DNA to resolve plant phylogenies-noncoding versus *rbc*L sequences. *Molecular Biology and Evolution* 11: 769-777.

GRIFFIN, G. 2008. Recent advances in research and management of chestnut blight on American chestnut. *Phytopathology* 98: S7-S7.

HAHN, D. A., G. J. RAGLAND, D. D. SHOEMAKER, AND D. L. DENLIGER. 2009. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics* 10: 234.

HANSEN, D. R., G. S. SPICER, AND R. PATTERSON. 2009. Phylogenetic relationships between and within *Phacelia* sections *Whitlavia* and *Gymnobythus* (Boraginaceae). *Systematic Botany* 34:737-746.

HEINZE, B. 2002. Chloroplast DNA primers. *http://fbva.forvie.ac.at/200/1982.html.*

HUANG, W.C., AND G. MARTH. 2008. Eagle view: a genome assembly viewer for next-generation sequencing technologies. *Genome Research* 18: 1538-1543.

HUANG, H., W. A. CAREY, F. DANE, AND J. D. NORTON. 1996. Evaluation of Chinese chestnut cultivars for resistance to *Cryphonectria parasitica*. *Plant Disease* 80: 45-47.

HYTEN, D. L., J. SONG Q, W. FICKUS E, C. V. QUIQLEY, J. S. LIM, I.Y. CHOI, E. Y. HWANG, M. PASTOR-CORRALES, P.B. CREQAN. 2010. High-throughput SNP discovery and assay development in common bean. *BMC Genomics* 11: 475.

IMELFORT, M., C. DURAN, J. BATLEY, AND D. EDWARDS. 2009. Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnology Journal* 7: 312-317.

JAYNES, R. 1975. Chestnut. *In* J. Moore [ed.], Advances in Fruit Breeding, 490-503. *Purdue University Press*, West Lafayette.

JOHNSON, G. P. 1988. Revision of *Castanea* sect. Balanocastanon (Fagaceae). *Journal of the Arnold Arboretum* 69: 25-49

JORDAN, W. C., M. W. COURTNEY, AND J. E. NEIGEL. 1996. Low levels of intraspecific genetic variation at a rapidly evolving chloroplast DNA locus in North American duckweeds (Lemnaceae). *American Journal of Botany* 83: 430-439.

JUNG, J. K., S. W. PARK, W. Y. LIU, AND B. C. KANG. 2010. Discovery of single nucleotide polymorphism in *Capsicum* and SNP markers for cultivar identification. *Euphytica* 175: 91-107.

KATOCH, M., R. KUMAR, S. RAL, AND A. AHUJA. 2010. Identification of *Chlorophytum* species (*C. borivilianum, C. arundinaceum, C. laxum, C. capense* and *C. comosum*) using molecular markers. *Industrial Crops and Products* 32: 389-393.

KELCHNER, S. A., AND L. G. CLARK. 1997. Molecular evolution and phylogenetic utility of the chloroplast *rpl*16 intron in *Chusquea* and the Bambusoideae (Poaceae). *Molecular Phylogenetics and Evolution* 8: 385-397.

KING, R. A., AND C. FERRIS. 1998. Chloroplast DNA phylogenography of *Alnus glutinosa* (L.) Gaertn. *Molecular Ecology* 7:1151-1161.

KRISHNANKUTTY, N., AND S. CHANDRASEKARAN. 2008. Linnaeus 300: Tips for tinkering morphological taxonomy. *Current Science* 94: 565-567.

KUBISIAK, T. L., F. V. HEBARD, C. D. NELSON, C. DANA, J. S. ZHANG, R. BERNATZKY, H. HUANG, S. L. ANAGNOSTAKIS, AND R. L. DOUDRICK. 1997. Molecular mapping of resistance to blight in an interspecific cross in the genus *Castanea*. *Phytopathology* 87: 751-759.

KUBISIAK, T. L., AND J. H. ROBERDS. 2003. Genetic variation in natural populations of American chestnut. Science and natural history XVI: 43-48.

LAHAYE, R., J. KLACKENBERG, M. KALLERSJO, E. V. CAMPO, AND L. CIVEYREL. 2007. Phylogenetic relationships between derived *Apocynaceae* S. L. and within Secamonoideae based on chloroplast sequences. *Missouri Botanical Garden* 94: 376-391.

LANG. P., F. DANE, AND T. L. KUBISIAK. 2006. Phylogeny of *Castanea* (Fagaceae) based on chloroplast *trn*T-L sequence data. *Tree Genetics and Genomes* 2: 132-139.

LANG. P., F. DANE, T. L KUBISIAK, AND H. W. HUANG. 2007. Molecular evidence for an Asian origin and a unique westward migration of species in the genus *Castanea* via Europe to North America. *Molecular Phylogenetics and Evolution* 43: 49-59.

LI, Y., H. M. LUO, C. SUN, J. Y. SONG, Y. Z. SUN, Q. WU, N. WANG, H. YAO, A. STEINMETZ, AND S. L. CHEN. 2010. EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. *BMC Genomics* 11: 268.

MACAS, J., P. NEUMANN, AND A. NAVRATILOVA. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Gemomics* 8: 427.

MARDIS, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 24: 3.

MEVE, U., AND S. LIEDE-SCHUMANN. 2007. Ceropegia (Apocynaceae, Ceropegieae, Stapeliinae): Paraphyletic but still taxonomically sound. *Annals of the Missouri Botanical Garden* 94: 392-406.

MILLER, J.S., A. KAMATH, AND R. A. LEVIN. 2009. Do multiple tortoises equal a hare? The utility of nine noncoding plastid regions for species-level phylogenetics in tribe Lycieae (Solanaceae). *Systematic Botany* 34: 796-804.

MILGROOM, M. G., K.WANG, Y. ZhOU, S. E. LAPARI, AND S. KANKO.1996. Intercontinental population structure of the chestnut blight fungus, *Cryphonectria parasitica*. *Mycologia* 88: 179–190.

MILLER, J. R., S. KOREN, AND G. SUTTON. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95: 315-327.

MOEN, T., B. HAYES, F. NILSEN, M. DELGHANDI, K. T. FJALESTAD, S. E. FEVOLDEN, P. R. PAUL, AND S. LIEN. 2008. Identification and characterization of novel SNP markers in atlantic cod: Evidence for directional selection. *BMC Genetics* 9.

MOWER. J.P., P. TOUZET, J.S. GUMMOW, L. F. GELPH, AND J. D. PALMER. 2007. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. *BMC Evolutionary Biology* 7:135.

MUIR, G., A. J. LOWE, C. C. FLEMING, AND C. VOGL. 2004. High nuclear genetic diversity, high levels of outcrossing and low differentiation among remnant populations of *Quercus petraea* at the margin of its range in Ireland. *Annals of Botany* 93:691-697.

NOVAES, E., D. R. DROST, W. G. FARMERIE, G. J. PAPPAS, D. GRATTAPAGLIA, R. SEDEROFF, AND M. KIRST. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.

OHTA, S., S. OSUMI, T. KATSUKI, I. NAKAMURA, T. YAMAMOTO, AND Y. I. SATO. 2006. Genetic characterization of flowering cherries (*Prunus* subgenus *cerasus*) using *rpl*16-*rpl*14 spacer sequences of chloroplast DNA. *Journal of the Japanese Society of Horticultural Sciences* 75: 72-78.

OLIVER, R. E., G. R. LAZO, J.D. LUTZ, M. F.RUBENFIELD, N. A. TINKER, J. M. ANDERSON, N. H. WISNIEWSKI, D. ADHIKARY, E. N. JELLEN, P. F. MAUGHAN, G. L. GUEDIRA, S. CHAO, A. D. BEATTIE, M. L. CARSON, H. W. RINES, D. E. OBERT, J. M. BONMAN, AND E. W. JACKSON. 2011. Model SNP development for complex genomes based on hexaploid oat using high-thoughput 454 sequencing technology. BMC Genomics 12: 27.

OLMSTEAD, R. G., AND J. A. SWEERE. 1994. Combining data in phylogenetic systematics- an empirical- approach using 3 molecular-data sets in the Solanaceae. *Systematic Biology* 43: 467-481.

OLMSTEAD, R. G., AND J. D. PALMER.1994. Chloroplast DNA systematics –a review of methods and data-analysis. *American Journal of Botany* 81:1205-1224.

OLSSON, S., V. BUCHBENDER, AND J. ENROTH. 2009. Evolution of the Neckeraceae (Bryophyta): resolving the backbone phylogeny. *Systematics and Biodiversity* 7: 419-432.

OVASKAINEN, O., J. NOKSO- KOIVISTO, J. HOTTOLA, T. RAJALA, T. PENNANEN, H. ALI-KOVERO, O. MIETTINEN, P, OINONEN, P. AUVINEN, L. PAULIN, K. H. LARSSON, AND R. MAKIPAA. 2009. Identifying wood-inhabiting fungi with 454 Sequencing – what is the probability that BLAST gives the correct species? *Fungal Ecology* 3: 274-283.

OZEKI, H., K. UMESONO, H. INOKUCHI, T. KOHCHI, AND K. OHYAMA. 1989. The chloroplast genome of plants –a unique origin. *Genome* 31:169-174.

PAILLET, F. L. 1993.Growth form and life-histories of American chestnut and Allegheny and Ozark chinquapin at various North-American sites. *Bulletin of the Torrey Botanical Club* 120: 257-268.

PARCHMAN, T. L., K. S. GEIST, J. A. GRAHNEN, G. W. BENKMAN, AND C. A. BUERKLE. 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* 11: 180.

PAYNE, J., G. MILLER, G. P. JOHNSON, S. D. SENTER. 1994. *Castanea-pumila* (L) Mill- an underused native nut tree. Horticulture Science 29: 62-& .

PENNINGTON, R. T., M. LAVIN, H. IRELAND, B. KLITGAARD, J. PRESTON, AND J. M. HU. 2001.Phylogenetic relationships of basal Papilionoid legumes based upon sequences of the chloroplast *trn*L intron. *Systematic Botany* 26: 537-556.

PETIT, R. J., U. M. CSAIKL, S. BORDACS, K. BURG, E. COART, J. COTTRELL, B. V. DAM, J. D. DEANS, S. D. LAPEGUE, AND A. KREMER. 2002. Chloroplast DNA variation in European white oaks- Phylogeography and patterns of diversity based on data from over 2600 populations. *Forest Ecology and Management* 156: 5-26.

RIGGINS, C. W., Y. H. PENG, C. N. STEWART, AND P. J. TRANEL. 2010. Characterization of *de novo* transcriptome for waterhemp (Amaranthus tuberculatus) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes. *Pest Management* 66: 1042-1052.

PROVAN, J., W. POWELL, AND P. M. HOLLINGSWORTH. 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology and Evolution* 16: 142 -147.

QUINLAN, A. R., D. A. STEWART, AND M. P. STROMBERG.2008. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods* 5: 179-181.

RAKOTOARISOA, J. E., M. RAHERIARISENA, AND S. M. GOODMAN. 2010. Phylogeny and species boundaries of the endemic species complex, *Eliurus anstsingy* and *E. carletoni*

(Rodentia: Muroidea: Nesomyidae), in Madagascar using mitochondrial and nuclear DNA sequence data. *Molecular Phylogenetics and Evolution* 57: 11-22.

RENDELL, S., AND R.A. ENNOS. 2003. Chloroplast DNA diversity of the dioecious European tree *llex aquifolium* L. (English holly). *Molecular Ecology* 12: 2681-2688.

ROBERT. C., I. J THOMSON, WANG, AND J. R. JOHNSON. 2010. Genome- enabled development of DNA markers for ecology, evolution and conservation. *Molecular Ecology* 19: 2184-2195.

ROWE, T., T. H. RICH, P.VICHERS-RICH, M. SPRINGER, AND M. O. WOODBURNE. 2008. The oldest platypus and its bearing on divergence timing of the platypus and echidna clades. *Proceedings of the National Academy of Sciences of the United States of America* 105: 1238-1242.

SHAW, J., AND. R. L. SMALL. 2005. Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). *American Journal of Botany* 92: 2011-2030.

SHAW, J., AND R. L. SMALL. 2004. Addressing the “hardest puzzle in American pomology”: phylogeny of *Prunus* sect. *prunocerasus* (Arosaceae) based on seven noncoding chloroplast DNA regions. *American Journal of Botany* 91: 985-996.

SHAW, J., E. B. LICKEY, E. E. SCHILLING, AND R. L. SMALL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* 94: 275-288.

SHINOZAKI, K., M. OHME, M. TANAKA, T. WAKASUGI, N. HAYASHIDA, T. MATSUBAYASHI, N. ZAITA, J. CHUNWONGSE, J. OBOKATA, K. YAMAGUCHI-SHINOZAKI, C. OHTO, K. TORAZAWA, B. Y. MENG, M. SUGITA, H. DENO, T. KAMOGASHIRA, K.YAMADA, J. KUSUDA, F. TAKAIWA, A. KATO, N. TOHDOH, H. SHIMADA, AND M. SUGIURA. 1986. The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression. *European Molecular Biology Organization Journal* 5: 2043-2049.

SIMMONS, M. P., AND H. OCHOTREANA. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* 49: 369-381.

SMALL, R. L., J. A. RYBURN, R. C. CRONN, T. SEELANAN, AND J. F. WENDEL. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstructin in a recently diverged plant group. *American Journal of Botany* 85: 1301-1315.

SMALL, R. L. 2004. Phylogeny of *Hibiscus* sect. *Muenchhusia* (Malvaceae) based on chloroplast *rpL*16 and *ndh*F, and nuclear ITS and GBSSI sequences. *Systematic Botany* 29: 385-392.

SSOLTIS, D. E., A. B. MORRIS, J. S. MCLACHLAN, P. S. MANOS, P. S. SOLTIS. 2006. Comparative phylogeography of unglaciated eastern North American. *Molecular Ecology* 15: 4261-4293.

STEELE, P. R., L. M. FRIAR, L. E. GILBERT, AND R. K. JANSEN. 2010. Molecular systematics of the neotropical genus *Psiguria* (Cucurbitaceae): implications for phylogeny and species identification. *American Journal of Botany* 97:156-173.

STILWELL, K, L., H. M. WILBUR, C. R. WERTH, AND D. R.TAYLOR. 2003. Heterozygote advantage in the American chestnut, *Castanea dentata* (Fagaceae). *American Journal of Botany* 90: 207-213.

STOCK, M., A. SICILIA, N. M. BELFIORE, D. BUCKLEY, S. L. BRUTTO, M. L. VALVO, AND M. ARCULEO. 2008. Post-messinian evolutionary relationships across the Sicilian channel: mitochondrial and nuclear markers link a new green toad from Sicily to African relatives. *BMC Evolutionary Biology* 8:56.

SUGIURA, M. 1992. The chloroplast genome. *Plant Molecular Biology* 19: 149-168.

SUDA, J., A. KRAHULCOVA, P. TRAVNICEK, R. ROSENBAUMOVA, T. PECKERT AND F. KRAHULEC. 2007. Genome size variation and species relationships in *Hieracium* Sub-genus *Pilosella* (Asteraceae) as inferred by flow cytometry. *Annals of Botany* 100: 1323-1335.

SUN, Y. J., Y. P. CAI, V. MAI,W. FARMERIE, F.YU, J. LI, AND S. GOODISON. 2010. Advanced computational algorithms for microbial community analysis using massive 16S rRNA sequence data. *Nucleic Acids Research* 38

SWOFFORD, D. L. 2000. PAUP. Phylogenetic Analysis Using Parsimony. Version 4. Sinauer Associates, Sunderland, Massachusetts.

TABERLET, P., L. GIELLY, G. PAUTOU, AND J. BOUVET. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* 17: 1105-1109.

THOMSON, R.C., I. J. WANG, J. R. JOHNSON. 2010. Genome-enabled development of DNA markers for ecology and conservation. *Molecular Ecology* 19: 2184-2195.

TIAN, X., AND D. Z. LI. 2002. Application of DNA sequences in plant phylogenetic study**.** *Acta Botanica Yumnan* 24: 170-184.

TIMME, R. E., J. V. KUEHL, J. L. BOORE, AND R. K. JANSEN. 2007. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *American Journal of Botany* 94: 302-312.

TUDGE, C. 2007. American chestnut: The life, death, and rebirth of a perfect tree. *Nature* 450:169-169.

VALCARCEL, V., P. VARGAS, AND G. N. FELINER. 2006. Phylogenetic and phylogeographic analysis of the western Mediterranean *Arenaria* section *Plinthine* (Caryophyllaceae) based on nuclear, plastid, and morphological markers. *Taxon* 55: 297-312.

WAGNER, D. B., G.R. FURNIER, M. A. SAGHAI-MAROOF, S. M. WILLIAMS, B. P. DANCIK, AND R.W. ALLARD. 1987. Chloroplast DNA polymorphisms in lodgepole and jack pines and their hybrids. *Proceedings of the National Academy of Sciences of the United States of America* 84: 2097-2100.

WALL, P. K., J. LEEBENS-MACK, A. S. CHANDERBALI, A. BARAKAT, E. WOLCOTT, H. LIANG, L. LANDHERR, L. P. TOMSHO, Y. HU, J. E. CARLSON, H. MA, S. C. SCHUSTER, D. E. SOLTIS, P. S. SOLTIS, N. ALTMAN AND C. W. DEPAMPHILIS. 2009. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* 10: 347.

WAN, C. Y., AND T. A. WILKINS.1994. A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum* L.). *Analytical Biochemistry* 223: 7-12.

WALTER. R., AND B.K. EPPERSON. 2001. Geographic pattern of genetic variation in *Pinus resinosa*: area of greatest diversity is not the origin of postglacial populations. *Molecular Ecology* 10: 103-111.

WU, X. L., C. W. REN, T. JOSHI, T. VUONG, D. XU AND H. NGUYEN. 2010. SNP discovery by high-throughput sequencing in soybean. BMC Genomics 11: 469.

XU, J. S. 2005. The effect of low-temperature storage on the activity of polyphenol oxidease in *Castanea* henryi chestnuts. *Postharvest Biology and Technology* 38: 91-98.

YUAN, Y.W., D. J. MABBERLEY, D. A. STEANE AND R. G. OLMSTEAD. 2010. Further disintegration and redefinition of *Clerodendrum* (Lamiaceae): implications for the understanding of the evolution of an intriguing breeding strategy. *Taxon* 59: 125-133.

ZAGROBELNY, M., S. A. KARSTEN, N. B. JENSEN, B. L. MOLLER, J. GORODKIN, AND S. BAK. 2009. 454 pyrosequencing based transcriptome analysis of Zygaena filipendulae with focus on genes involved in biosynthesis of cyanogenic glucosides. *BMC Genomics* 10: 574.

ZJHRA, M. L., K. J. SYTSMA, AND R. G. OLMSTEAD. 2004. Delimitation of Malagasy tribe *Coleeae* and implications for fruit evolution in Bignoniaceae inferred from a chloroplast DNA phylogeny. *Plant Systematics and Evolution* 245: 55-67.