

## **Final report: Identification of different ecotypes and centers of adaptive genetic diversity in American chestnut**

### **Principal investigators:**

Oliver Gailing, Professor for Ecological Genetics, Michigan Technological University, 1400 Townsend Drive, Houghton 49931, Michigan, ogailing@mtu.edu.

Brian C. McCarthy, Professor of Forest Ecology & Chair, Dept. Env. & Plant Biology, 416 Porter Hall, Ohio University, Athens, OH 45701-2979 USA, mccarthy@ohio.edu

C. Dana Nelson, Project Leader/Research Geneticist, US Forest Service, Southern Institute Forest Genetics, Southern Research Station, Saucier, MS 39574 USA, dananelson@fs.fed.us, phone: 228-832-2747-201

### **Outputs**

Detailed results and methods are described in the draft article which is attached. A summary of the main methods and results is given below. Additional analyses to associate genetic variation patterns with environmental variable are planned.

### **Plant material and marker analyses**

A total of nine populations (~32 samples per population) that covered the distribution range of the species (Table 1, (Kubisiak and Roberds 2006) have been characterized at genetically mapped EST-SSRs (Kubisiak et al. 2013). We have tested a total of 25 EST-SSRs that have been developed in *Castanea mollissima*, 17 of them with clear and polymorphic amplifications products were selected for the population analysis. We have developed a multiplex PCR touchdown protocol that allowed us to analyze up to four markers in one PCR reaction. The touchdown program in the Biometra Thermocycler Tprofessional (Jena, Germany) was as follows: initial denaturation at 95 °C for 15 min, 10 touchdown cycles of 1 min at 94 °C, 1 min at 60 °C (- 1 °C per cycle) and 1 min at 72 °C, followed by 25 cycles at 94 °C for 1 min, 50 °C for 1 min and 72 °C for 1 min and a final extension at 72 °C for 20 min. The PCR mix consisted of 0.2 µl (5 U/ µl) Hotstar Taq polymerase from Solis Biodyne (Estonia), 1.5 µl 10 X reaction buffer B, 1.5 µl MgCl<sub>2</sub> (25 mM), 0.2 µl (5 pikomol/µl) tailed forward primer, 0.5 µl (5 pikomol/µl) pig-tailed reverse primer (Kubisiak et al. 2013), 1.5 µl M13 primer (5 pikomol/ µl), 1 µl dNTPs (2.5 mM each dNTP), 2µl DNA (ca. 0.6 ng/ µl) and 5 µl H<sub>2</sub>O. PCR products were separated on an ABI 3130xl Genetic analyzer (Applied Biosystems) and alleles were called using Genemapper v. 4.0 (Applied Biosystems).

Based on the review of last year's progress report we tried to identify additional populations in the southwest of the species' distribution range. As suggested by the reviewers we contact Dr. Hill Craddock about further populations from the south/southwest of the distribution range. Unfortunately, these populations were no longer in storage. We decided to use the resources to analyze our populations at chloroplast DNA markers to get additional information on genetic diversity patterns and potential postglacial remigration routes. We tested a total eleven universal chloroplast microsatellite markers (Weising and Gardner 1999; Deguilloux et al. 2003) and selected five polymorphic markers for population analyses. PCR protocols followed (Weising and Gardner 1999; Deguilloux et al. 2003; Gailing et al. 2009).

## **Data analyses**

Genetic variation in populations was assessed as the number of alleles per locus ( $N_a$ ), observed heterozygosity ( $H_o$ ) and expected heterozygosity ( $H_e$ ) (Nei, 1973) in the program GeneALEX v.6.41 (Peakall and Smouse 2006) (Table 2). Genetic differentiation among populations and between population pairs was calculated as  $F_{ST}$ . Additionally Nei's unbiased genetic distance (Nei 1978) between populations was assessed. A Principal Coordinate Analysis was performed to represent genetic distances between populations. Additionally, the program STRUCTURE (Pritchard et al. 2000) was used to determine the number of genetic clusters  $K$  and to assign individual samples to genetic clusters following the approach described in (Lind and Gailing 2013). Significance of pairwise differentiation and linkage disequilibrium was tested in GENEPOP (Raymond and Rousset 1995). Relationships among populations are presented in a dendrogram using POPULATIONS (Langella 1999). We identified loci with genetic differentiation above neutral expectations (outlier loci) in the program LOSITAN (Antao et al. 2008).

## **Results and Discussion**

### *Genetic variation*

Gene diversity or expected heterozygosity ( $H_e$ ), observed heterozygosity ( $H_o$ ) and allelic diversity ( $N_a$ ) were highest in population MNC and lowest in population PCT (Table 2). No latitudinal variation in genetic diversity was observed. However,  $N_a$  showed a significant decline from southwest to northeast ( $R^2 = 0.856$ ). The northernmost population OCA (Canada, Ontario) showed the second highest values for  $H_e$  and  $N_a$ , but comparative low values for  $H_o$  which might reflect moderate levels of inbreeding. The decrease of allelic diversity from southwest to northeast was in accordance with earlier analyses at anonymous RAPDs and allozyme markers and suggest repeated founder events during postglacial migration (Kubisiak and Roberds 2006).

### *Genetic differentiation*

The eastern populations MA, NYUC and PCT were differentiated from the western populations as revealed by the Principal Coordinate Analysis based on Nei's unbiased genetic distances (Figure 2). PCA 1 explained 62.62% of the variation and was significantly associated with longitude ( $R^2 = 0.860$ ,  $p < 0.001$ ) (Figure 3). Likewise, the STRUCTURE analysis identified two genetic clusters and separated the eastern populations MA, NYUC, PCT from the western populations, while PYC (Pennsylvania) seemed to occupy an intermediate position (Figure 1, 4).

The results suggest a significant east-west differentiation potentially as result of limited gene flow across the Appalachian range. We also identified significant genetic differentiation between populations from one region which might suggest different local adaptations (e.g. between PCT and MA).

Population OCA from Ontario clustered with southwestern populations in the Bayesian analysis of population structure and in the dendrogram. Also, OCA shared a rare chloroplast haplotype with western populations. The results suggested a divergence of postglacial migration routes to the north and the northeast and potentially more recent human-mediated dispersal west of the Appalachian axis. We suggest to sample and analyze additional populations from Ontario and from the western distribution range of American chestnut. Genetic differentiation above neutral expectations was found at one EST-SSR which showed as steep decline in allele frequencies from east to west. Patterns of genetic variation at gene-based markers should be considered for the collection of reproductive material for breeding purposes.

**Table 1** Sample locations

Name	N	Latitude	Longitude
Massachusetts_a2-a40	32	42.22	72.31
Asheville_br2-br35	32	35.46	82.10
Ontario_ca2-ca34	32	43.08	80.30
Portland_CT1-CT33	32	41.35	72.37
Murphy_NC (fc1-jb18)	32	35.05	84.01
Kentucky_pc1-pc40	32	37.50	83.51
New York_uc1-uc29	32	41.44	74.13
Pennsylvania_yc1-yc35	32	39.48	76.59
Maryland_md15-md58)	32	39.37	79.07

**Table 2** *Castanea mollissima* primers tested for amplification and polymorphism in *C. dentata*

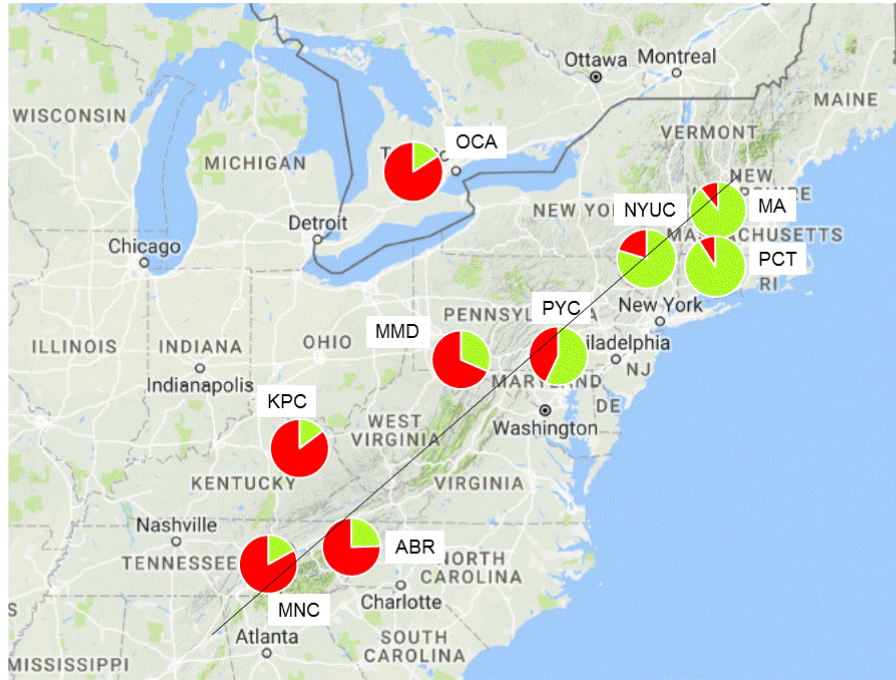
Marker name	GenBank ID	primer1 sequence	primer2 sequence	Motif	BLASTN	LG and position (cM)
<b>CmSI0031</b>	<b>290474526</b>	<b>AGCGCCACTTTTCTTTTCA</b>	<b>GAATCCCAGCCTGACCAATA</b>	<b>AAG</b>	-	<b>LGH (25.3 cM)</b>
<b>CmSI0049</b>	<b>290474533</b>	<b>CCGATGCCGATTTCTACAAC</b>	<b>GCGGCAGACACATAGTTTCA</b>	<b>TCTTCA</b>	-	<b>LGB (34.1 cM)</b>
CmSI0051	290474536	CGATCATATCCCATACCACA	GCGGAGACCACTAAGAGACG	CT		
<b>CmSI0327</b>	<b>290474585</b>	<b>CTCCTCCGCCTCCATTCA</b>	<b>AGTCCTTGCGGATCACTTTG</b>	<b>AG</b>	60S ribosomal protein L27-like, 8e-43	<b>LGK (5.8 cM)</b>
CmSI0383	290474601	CCTCCTCCACCACCAGTTTA	TGGAGGTGGGGACTTGTACT	CTC		
<b>CmSI0391</b>	<b>290474604</b>	<b>TTTGGAGTGCTGAGGAGACT</b>	<b>GACTCAAAGTCCTCGGCAAA</b>	<b>TAA</b>	-	
CmSI0392	290474605	CCATCGGGAAATGTGGTATT	GCTCATCTGGGAACCACTGT	GTG		LGC (26.4 cM)
<b>CmSI0396</b>	<b>290474606</b>	<b>AACTCCCACCACTCACATCC</b>	<b>TTTCGGACCATCCAGAATC</b>	<b>CACACC</b>	uncharacterized protein, 1e <sup>-4</sup>	
CmSI0437	290474621	GGGCTTCTGGAAACTAGCA	CCATACGAACACCGAGGACT	TG		LGH (49.4cM)
<b>CmSI0495</b>	<b>290474633</b>	<b>GAAACAACAGGCTCTGCCTC</b>	<b>CTGGGAAAATCCGAACTCAA</b>	<b>GA</b>	probable amino-acid acetyltransferase NAGS2, 9e <sup>-78</sup>	
<b>CmSI0527</b>	<b>290474668</b>	<b>TACGACCTAACGACTCGCC</b>	<b>AGGAGGAGAAGTCAGCCACA</b>	<b>GTT</b>	-	<b>LGF (14.8 cM)</b>
<b>CmSI0537</b>	<b>290474676</b>	<b>AGAGATGGGTTGGGAAGGTT</b>	<b>GGCCTCTCTGGTTTGTGTGT</b>	<b>AG</b>	-	<b>LGA (41.4 cM)</b>
CmSI0541	290474681	CCAATCCCAAAATCCACTTG	GGACATTTTGAGCCTGAAA	GA		
<b>CmSI0551</b>	<b>290474692</b>	<b>TAACCAATCACGTTCAACGA</b>	<b>CGCCACATCTAAAACCCCTA</b>	<b>CGC</b>	-	<b>LGH (26.3 cM)</b>
<b>CmSI0559</b>	<b>290474700</b>	<b>AGGTGGGAGTGAAGGTGTTG</b>	<b>TATCTCGCTCGCTCCATCTT</b>	<b>AGG</b>	-	<b>LGC (33.3cM)</b>
<b>CmSI0561</b>	<b>290474702</b>	<b>CGTATAGGTTGGAAACGGAA</b>	<b>GGACAAGCAAATCACGGAAT</b>	<b>TCG</b>	-	
CmSI0594	290476538	GCCCCAAAGAAAGAGAAAG	GCATGCCCATACCCATTAAC	GGT		LGH (17.2 cM)
<b>CmSI0600</b>	<b>290474736</b>	<b>TCACACTCACACACGCAAAA</b>	<b>TGTTACGAAAATACGCAACG</b>	<b>CT</b>	Uncharacterized locus, 1e <sup>-74</sup>	
CmSI0603	290474740	ACTCCATGGGAATGATGAGC	TGTGTGTGTGTTTCTCTGTGA	TC		
<b>CmSI0608</b>	<b>290474745</b>	<b>TTTGATTGGCCCTCCTAGAC</b>	<b>CTGAATCGCCGAAACTCATT</b>	<b>AGG</b>	transducin/WD40 repeat-like superfamily protein, 9e <sup>-32</sup>	<b>LGL (17.8 cM)</b>
<b>CmSI0611</b>	<b>290474746</b>	<b>GCTGACCCTGTCAACAACAA</b>	<b>CAGAACTAGACAAGGATCACAAGA</b>	<b>TC</b>	-	<b>LGK (20.1 cM)</b>
CmSI0632	290474764	TCGGAGTAGTTGGAGCAGTG	TGAGAAAAGAAAGTGCCTCA	GTT		
<b>CmSI0678</b>	<b>290474795</b>	<b>GGTCAGACCCGCTAGCTCT</b>	<b>ACCCAAAACCAAAACCAAAA</b>	<b>TCT</b>	-	<b>LGK (19.3 cM)</b>
<b>CmSI0683</b>	<b>290474802</b>	<b>CACCAGCACCTCACTTCTCC</b>	<b>CCGGAAGATTTAGGTTTAGGG</b>	<b>AGA</b>	-	
CmSI0689	290474809	TCCCAAATGAAATGAAATGAAA	TGAAAATCCCTCCATCATCA	TGAGA		LGD (55.8 cM)

Markers labeled in bold face amplified easy-to-interpret and polymorphic loci and were characterized in all nine populations. BlastN 2.231 results (Zheng et al., 2000) and sequence descriptions are included. Linkage group (LG) position (cM) on the *C. mollissima* linkage map (Kubisiak et al., 2013) is indicated.

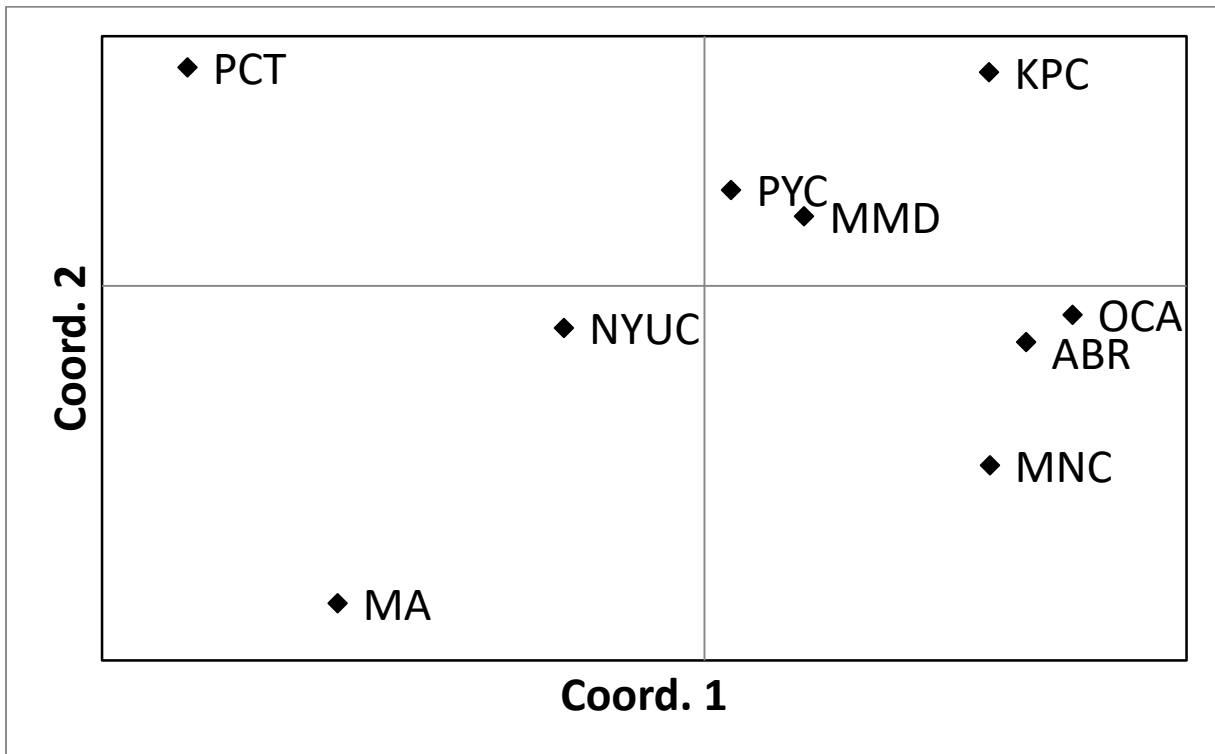
**Table 3** Genetic variation within populations

<b>Pop</b>		<b>N<sub>a</sub></b>	<b>H<sub>o</sub></b>	<b>H<sub>e</sub></b>	<b>F</b>
<b>ABR</b>	<b>Mean</b>	4.706	0.451	0.468	0.084
	<b>SE</b>	0.527	0.064	0.060	0.067
<b>KPC</b>	<b>Mean</b>	4.706	0.481	0.498	0.050
	<b>SE</b>	0.460	0.061	0.057	0.028
<b>MA</b>	<b>Mean</b>	3.882	0.499	0.489	-0.034
	<b>SE</b>	0.410	0.045	0.044	0.034
<b>MMD</b>	<b>Mean</b>	4.471	0.493	0.475	-0.047
	<b>SE</b>	0.576	0.059	0.057	0.024
<b>MNC</b>	<b>Mean</b>	5.824	0.517	0.531	0.033
	<b>SE</b>	0.637	0.064	0.062	0.043
<b>NYUC</b>	<b>Mean</b>	4.059	0.513	0.489	-0.058
	<b>SE</b>	0.424	0.059	0.057	0.018
<b>OCA</b>	<b>Mean</b>	5.294	0.477	0.517	0.084
	<b>SE</b>	0.435	0.050	0.049	0.042
<b>PCT</b>	<b>Mean</b>	3.706	0.439	0.396	-0.101
	<b>SE</b>	0.400	0.052	0.045	0.029
<b>PYC</b>	<b>Mean</b>	4.471	0.490	0.475	-0.033
	<b>SE</b>	0.394	0.055	0.052	0.030
<b>Total</b>	<b>Mean</b>	4.569	0.484	0.482	-0.002
	<b>SE</b>	0.164	0.019	0.018	0.013

SE: Standard error, N<sub>a</sub>: number of alleles per locus, H<sub>o</sub>: observed heterozygosity, H<sub>e</sub>: expected heterozygosity, F: inbreeding coefficient,  $F = (H_e - H_o)/H_e$

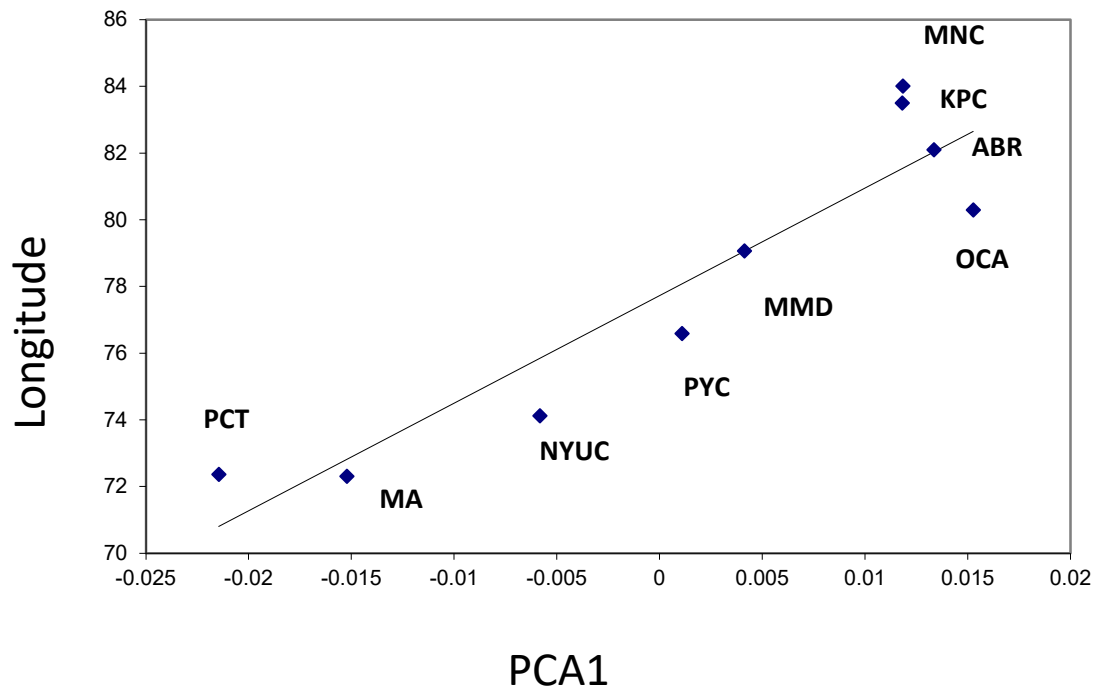


**Figure 1.** Sample location and results of genetic assignment analysis in STRUCTURE.

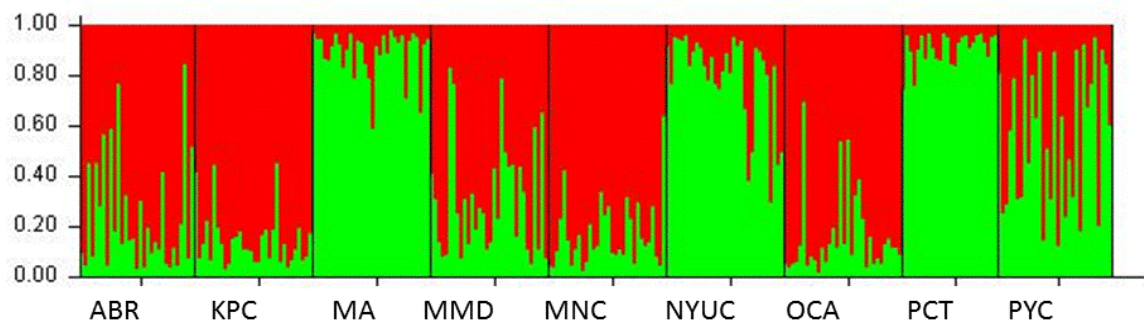


**Figure 2.** Principal Coordinate Analysis based on Nei's unbiased genetic distance (Nei 1978).





**Figure 3.** Association between longitude and PCA1.



**Figure 4.** STRUCTURE plots for  $K = 2$ .

## Proposal

### **a. Title: Identification of different ecotypes and centers of adaptive genetic diversity in American chestnut**

#### **b. Summary:**

We will develop a set of new gene-based DNA markers with potential role in local adaptation to (1) identify centers of genetic diversity in American chestnut, (2) detect gene-based markers under divergent selection in contrasting environments and to (3) associate allele frequency distributions at gene markers with environmental variables. We will map the genes back to the *Castanea mollissima* linkage map (*in silico* mapping) to analyze the distribution of these genes on the linkage map and their potential association with QTL for phenotypic and resistance traits.

#### **c. Principal investigators:**

Oliver Gailing, Associate Professor for Ecological Genetics, Michigan Technological University, 1400 Townsend Drive, Houghton 49931, Michigan, ogailing@mtu.edu.

Brian C. McCarthy, Professor of Forest Ecology & Chair, Dept. Env. & Plant Biology, 416 Porter Hall, Ohio University, Athens, OH 45701-2979 USA, mccarthy@ohio.edu

C. Dana Nelson, Project Leader/Research Geneticist, US Forest Service, Southern Institute Forest Genetics, Southern Research Station, Saucier, MS 39574 USA, dananelson@fs.fed.us, phone: 228-832-2747-201

#### **d. Duration of project: 2 years**

**e. Total amount requested:** We requested a total of \$ 6,000. The estimated costs for the marker analyses are \$ 9,500 excluding labor costs. The additional costs will be covered with internal funds.

**f. Short and long-term goals of the project:** Short-term goals of the project are the identification of (1) center of genetic diversity and (2) of genes under divergent selection in contrasting environments.

The long-term goals are to (1) provide information for the selection of breeding material and reintroduction of hybrid American chestnut and (2) to use a combined outlier screening and QTL mapping approach to identify genes related to biotic and abiotic stress resistance for use in marker-assisted breeding.

## **g. Narrative**

### ***Background***

Considerable progress has been made to develop blight resistant chestnuts for restoration purposes using mainly genetic marker-assisted back-cross breeding to incorporate blight resistance from *C. mollissima* into a *C. dentata* genetic background. Restoration programs are more likely to be successful if hybrid American chestnut populations are genetically diverse (James et al., TACF project 2010 - 2013) and locally adapted. For this purpose genetic variation of the American chestnut parents should be captured from many individuals originating from different geographic regions and climatic zones. The conservation of genetic variation in fitness-related traits (adaptive genetic variation) is crucial for the successful restoration of American chestnut since the species' reintroduction is threatened by other biotic (e.g. *Phytophthora cinnamomi*) and abiotic stressors. While growing genomic resources and gene-based markers are becoming available for American chestnut and related species (Barakat et al., 2012; Bodénès et al., 2012; Kubisiak et al., 2013; Nishio et al., 2011), genetic variation at these markers with annotated function (e.g. gene-based microsatellites) has not yet been analyzed in natural populations of American chestnut. In a preliminary study we have already identified and characterized a set of gene-based markers in American chestnut (Table 1).

We hypothesize (1) significant differences in the level of genetic variation for populations from different geographic regions and (2) significant differentiation among regions at some gene-based markers that reflect different local adaptations of the populations (ecotypes) across their distribution range.

These markers that are identified as under divergent selection between populations (outlier loci) from contrasting environments and/or associated with environmental variables across populations will be mapped back to the *Castanea mollissima* linkage map (Kubisiak et al. 2013) to test for a possible co-location with QTL regions. The expected results will be important to identify centers of genetic diversity and to select appropriate breeding material to produce locally adapted material for the reintroduction of American chestnut.

In the future, a combined outlier screening and QTL mapping approach based on nextgen sequencing markers will allow us to test for a co-location of genome-wide outliers with genomic regions that underlie QTL for traits related to biotic and abiotic stress resistance.

### ***Work plan***

We propose to characterize genetic variation within and among *C. dentata* populations covering the distribution range of the species using 16 gene-based microsatellite markers with annotated function (Expressed Sequence Tag- Simple Sequence Repeats, EST-SSRs). In this preliminary study we will focus on 10 populations that represent the five US climatic zones within the species' native range. Leaf material has been sampled from about 30 trees per population recording the GPS position of each tree (Kubisiak and Roberds, 2005). In order to select the 16 gene-based microsatellite markers for the range-wide study, we will screen about 30 markers with annotated function in abiotic and biotic stress tolerance

and known location on genetic linkage maps (Kubisiak et al., 2013) in two populations from different climatic zones. We will include orthologous candidate genes that had been identified as being under divergent selection in interfertile red oak species with different adaptations to drought (Lind and Gailing, 2013; Sullivan et al., 2013; Lind-Riehl et al. 2014). A total of 12 of these markers have already been tested in a panel of 8 American and Chinese chestnut individuals and 7 of them amplified polymorphic products in the expected size range (Table 1, Figure 1, unpublished results). Two of the loci, FIR013 and FIR039, were identified as putative outlier loci under strong divergent selection in the North American red oak species and have putative functions in drought stress response and the control of flowering time (Sullivan et al. 2013; Lind-Riehl et al. 2014). EST-SSR markers that display signatures of selection relative to potentially neutral genomic SSRs that had been analyzed in the same sample set (Kubisiak and Roberds, 2005) will be selected for the population analyses. Calculation of genetic variation within and among populations will follow standard procedures. Associations of genetic variation patterns with geographic distance and environmental parameters will also be tested. A detailed description of the methods is given below.

## ***Methods***

### *Marker selection*

We have already identified 7 gene-based markers for American chestnut that were originally developed for oak (Durand et al. 2010, Lind-Riehl et al. 2014). Additional EST-SSRs that are genetically mapped in *Castanea mollissima* (Kubisiak et al., 2013) will be selected based on their annotated function in vegetative bud burst and biotic and abiotic stress tolerance, and genomic co-location with QTL for chestnut blight resistance. Our aim is to obtain at least 30 markers for the outlier analysis. Selection of the markers will be based on their reproducibility, absence of null alleles and potential importance in adaptation and disease resistance. Functional annotation will be assigned to microsatellite-containing ESTs using the BLASTx algorithm (Altschul et al., 1997) by comparison to homologous sequences in the NCBI database. Based on the initial outlier screen of the 30 polymorphic markers in two populations from different climate zones (outlier analysis), 16 EST-SSRs will be selected for the population analyses.

### *Marker analyses*

DNA and leaf samples will be provided by Dana Nelson and Tom Kubisiak who performed a range-wide sampling for the assessment of neutral genetic variation patterns at genomic SSR and RAPD markers (Kubisiak and Roberds, 2005). We will be able to combine data sets and compare patterns of genetic variation at candidate gene associated EST-SSR markers with neutral variation (Kubisiak and Roberds, 2005). Samples will be amplified with fluorescent labeled forward primers using a Peltier Thermal Cycler (Geneamp® PCR system 2700, Applied Biosystems) and PCR products will be separated on an ABI PRISM® 3730 Genetic Analyzer (Applied Biosystems). PCR reactions will be adapted from Kubisiak et al. (2013). Even though species identity was tested with a chloroplast marker

that differentiates between American chestnut and the native cogener species chinkapin (*Castanea pumila*) (Kubisiak and Roberds, 2005), the occurrence of interspecific hybrids cannot be excluded. We therefore use the generated marker information to assign individual samples to species and interspecific hybrids in the program STRUCTURE 2.3.4 (Pritchard et al., 2000). For this purpose we will include 20 *C. pumila* reference samples that were identified based on morphology and chloroplast marker information.

#### *Outlier screens*

A total of 30 EST-SSRs will be selected for the outlier screen and amplified in two populations from different adjacent climate zones to identify gene loci (outlier loci) that show a higher or lower differentiation between populations than expected under selective neutrality. We will use the program LOSITAN that implements the  $F_{ST}$ -based algorithms of FDIST to identify outliers that deviate significantly from a simulated neutral confidence envelope (Antao et al., 2008; Beaumont and Nichols, 1996). Loci with higher differentiation between populations than expected under neutrality are identified as potential outliers under divergent selection. Those falling below the lower bound of the neutral envelope might be under balancing selection. Since the confidence interval converges at extreme values for expected heterozygosity ( $H_e$ ), candidate genes under balancing selection were not consistently identified in different simulations (Sullivan et al., 2013) while loci under divergent selection were highly reproducible (Sullivan et al., 2013). We will therefore run the simulations at least three times for each pairwise comparison. To identify signatures of selective sweeps we will also run the LnRH test statistic that estimates variability between populations at individual loci instead of population divergence to identify selection (Schlötterer, 2002).

#### *Identification of genetic diversity centers*

Genetic variation within and among populations and climatic regions will be calculated for all markers and separately for potentially adaptive (outlier markers) and neutral markers. Specifically the following genetic variation parameters will be calculated: number of alleles per locus ( $N_a$ ), observed heterozygosity ( $H_o$ ) and Nei's unbiased gene diversity ( $H_e$ ) (Nei, 1973). Pairwise genetic differentiation between populations and corresponding significances will be calculated in GenePop4.1 (Raymond and Rousset, 1995). To visualize genetic distances among populations an unweighted pair-group method with arithmetic means (UPGMA) dendrogram (Sneath and Sokal, 1973) will be calculated in Populations 2.0 (Langella, 1999) using 1,000 bootstrap replicates. An Analysis of Molecular Variance (AMOVA, (Excoffier et al., 1992) will be performed in Arlequin 3.5 (Excoffier and Lischer, 2010) in order to assess genetic variation within and among populations and climatic regions. To test for associations between geographic and genetic distances we will perform a Mantel test as implemented in GeneAIEx v.6.41 (Peakall and Smouse, 2006).

#### *Association of allele frequencies with environmental variables*

We will search for associations between geographic (longitude, latitude) and environmental variables (e.g. minimum and maximum temperature, precipitation) on the one hand and allele frequencies and genetic variation parameters on the other hand using stepwise regression analysis (Kubisiak and Roberds, 2005) and association mapping approaches. Associations between SSR alleles and environmental variables will be performed using a standard association mapping approach by substituting the phenotype with environmental variables according to Eckert et al. (2010b). Specifically, following Du et al. (2013) associations between SSR marker locus and climatic variables will be analyzed using a mixed linear model accounting for population structure and kinship (Yu et al., 2006) and P-values will be generated for each test using 10,000 permutations of genotypes with respect to environmental values. All analyses will be conducted using TASSEL v. 3.0 (Bradbury et al., 2007). Multiple testing corrections will be performed using the false discovery rate (FDR) method (Storey, 2002; Storey and Tibshirani, 2003). Data on six nuclear microsatellite loci and 19 RAPD loci available for the same populations (Kubisiak and Roberds, 2005) will be used to estimate pair-wise kinship coefficients in the software SPAGEDi (Hardy and Vekemans, 2002) and population structure coefficients in STRUCTURE 2.3.4 (Pritchard et al., 2000).

Climate data (annual minimum and maximum temperatures, monthly temperature, precipitation) will be obtained from the PRISM website (Oregon State University, <http://www.prismclimate.org>) and from an ESRI 30 arc-second grid file available at the WorldClim website (<http://www.worldclim.org/current>) according to Eckert et al. (2010a). Accumulated growing degree-days above 5 °C will be calculated and an aridity index will be calculated according to (Eckert et al., 2010a; Eckert et al., 2010b; Sork et al., 2010).

The location of all markers relative to QTL for phenotypic and disease-resistance traits will be assessed on the *C. mollissima* linkage map to analyze the genomic location of genes with signatures of selection.

**h. Timeline:**

Activity	Year 1				Year 2			
	1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12
Outlier screening	x	x						
Range-wide marker analyses			x	x	x	x		
Data analysis and publication		x	x	x	x	x	x	x

**j. Budget:** Total costs are estimated as \$9,500. A total of \$ 6,000 is requested from the American Chestnut Foundation. Based on our experience with these analyses we estimate \$ 2,000 for the marker development and \$ 7,500 for the population genetic analyses (\$25 per sample x 300 samples, including DNA isolation, PCR, labeled primers and genotyping services). These estimates do not include labor costs. **Requested funds:** Year 1: \$ 4,000. Year 2: \$2,000.



## References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389-3402.
- Antao, T., Lopes, A., Lopes, R.J., Beja-Pereira, A., Luikart, G. (2008) LOSITAN: A workbench to detect molecular adaptation based on a F(st)-outlier method. *BMC Bioinformatics*, **9**.
- Barakat, A., Staton, M., Cheng, C.H., Park, J., Yassin, N.B.M., Ficklin, S., Yeh, C.C., Hebard, F., Baier, K., Powell, W., Schuster, S.C., Wheeler, N., Abbott, A., Carlson, J.E., Sederoff, R. (2012) Chestnut resistance to the blight disease: insights from transcriptome analysis. *BMC Plant Biology*, **12**.
- Beaumont, M.A. and Nichols, R.A. (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **263**, 1619-1626.
- Bodénès, C., Chancerel, E., Gailing, O., Vendramin, G.G., Bagnoli, F., Durand, J., Goicoechea, P.G., Soliani, C., Villani, F., Mattioni, C., Koelewijn, H.P., Murat, F., Salse, J., Roussel, G., Boury, C., Alberto, F., Kremer, A., Plomion, C. (2012) Comparative mapping in the Fagaceae and beyond with EST-SSRs. *BMC Plant Biology*, **12**, 153.
- Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633-2635.
- Du, Q.Z., Pan, W., Xu, B.H., Li, B.L., Zhang, D.Q. (2013) Polymorphic simple sequence repeat (SSR) loci within cellulose synthase (PtoCesA) genes are associated with growth and wood properties in *Populus tomentosa*. *New Phytologist*, **197**(3), 763-776.
- Eckert, A.J., Bower, A.D., Gonzalez-Martinez, S.C., Wegrzyn, J.L., Coop, G., Neale, D.B. (2010a) Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology*, **19**, 3789-3805.
- Eckert, A.J., van Heerwaarden, J., Wegrzyn, J.L., Nelson, C.D., Ross-Ibarra, J., Gonzalez-Martinez, S.C., Neale, D.B. (2010b) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, **185**, 969-982.
- Excoffier, L. and Lischer, H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564-567.
- Excoffier, L., Smouse, P.E., Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479-491.
- Hardy, O.J. and Vekemans, X. (2002) SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618-620.
- Kubisiak, T. and Roberds, J.H. (2005) Genetic structure of American Chestnut populations based on neutral DNA markers. In *Restoration of American Chestnut to forest lands* (K.C. Steiner & J.E. Carlson, eds): pp 1-13.
- Kubisiak, T.L., Nelson, C.D., Staton, M.E., Zhebentyayeva, T., Smith, C., Olukolu, B.A., Fang, G.C., Hebard, F.V., Anagnostakis, S., Wheeler, N., Sisco, P.H., Abbott, A.G., Sederoff, R.R. (2013) A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*). *Tree Genetics & Genomes*, **9**, 557-571.
- Langella, O. (1999) Populations, 1.2.30, CNRS UPR9034.
- Lind, J. and Gailing, O. (2013) Genetic structure of *Quercus rubra* L. and *Q. ellipsoidalis* E. J. Hill populations at gene-based EST-SSR and nuclear SSR markers. *Tree Genetics & Genomes*, **31**, 231-239.
- Lind-Riehl J., Sullivan A., Gailing O. (2014) Evidence for selection on a CONSTANS-like gene between two red oak species. *Annals of Botany*, **113**, 967-975.
- Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA*, **70**, 3321-3323.

- Nei, M. (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583-590.
- Nishio, S., Yamamoto, T., Terakami, S., Sawamura, Y., Takada, N., Nishitani, C., Saito, T. (2011) Novel genomic and EST-derived SSR markers in Japanese chestnuts. *Scientia Horticulturae*, **130**, 838-846.
- Peakall, R. and Smouse, P.E. (2006) GENEALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288-295.
- Pritchard, J.K., Stephens, M., Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.
- Raymond, M. and Rousset, F. (1995) GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.
- Schlötterer, C. (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, **160**, 753-763.
- Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy* W.H. Freeman, San Francisco.
- Sork, V.L., Davis, F.W., Westfall, R., Flint, A., Ikegami, M., Wang, H.F., Grivet, D. (2010) Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Nee) in the face of climate change. *Molecular Ecology*, **19**, 3806-3823.
- Storey, J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **64**, 479-498.
- Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440-9445.
- Sullivan, A., Lind, J., McCleary, T.S., Romero-Severson, J., Gailing, O. (2013) Development and characterization of genomic and gene-based microsatellite markers in North American red oak species. *Plant Molecular Biology Reporter*, **31**, 231-239.
- Yu, J.M., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203-208.
- Zheng, Z., Schwartz, S., Wagner, L., Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000, **7**, 203-14.
- Antao, T., Lopes, A., Lopes, R.J., Beja-Pereira, A., and Luikart, G. 2008. LOSITAN: A workbench to detect molecular adaptation based on a  $F_{ST}$  outlier method. *BMC Bioinformatics* **9**: 323.
- Deguilloux, M.-F., Dumolin-Lapegue, S., Gielly, L., Grivet, D., and Petit, R.J. 2003. A set of primers for the amplification of chloroplast microsatellites in *Quercus*. *Molecular Ecology Notes* **3**: 24-27.
- Gailing, O., Wachter, H., Heyder, J., Rogge, M., and Finkeldey, R. 2009. Chloroplast DNA analyses of very old, presumably autochthonous *Quercus robur* L. stands in North Rhine-Westphalia. *Allgemeine Forst Und Jagdzeitung* **180**: 221-227.
- Kubisiak, T., and Roberds, J.H. 2006. Genetic structure of American Chestnut populations based on neutral DNA markers. Pages 109-122 *in* Restoration of American Chestnut to forest lands. National Park Service, Washington, DC.
- Kubisiak, T.L., Nelson, C.D., Staton, M.E., Zhebentyayeva, T., Smith, C., Olukolu, B.A., Fang, G.C., Hebard, F.V., Anagnostakis, S., Wheeler, N., Sisco, P.H., Abbott, A.G., and Sederoff, R.R. 2013. A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*). *Tree Genet. Genomes* **9**: 557-571.
- Langella, O. 1999. Populations, 1.2.30. CNRS UPR9034.

- Lind, J., and Gailing, O. 2013. Genetic structure of *Quercus rubra* L. and *Q. ellipsoidalis* E. J. Hill populations at gene-based EST-SSR and nuclear SSR markers. *Tree Genet. Genomes* **9**: 707-722.
- Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **83**: 583-590.
- Peakall, R., and Smouse, P.E. 2006. GENEALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**: 288-295.
- Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959.
- Raymond, M., and Rousset, F. 1995. GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.* **86**: 248-249.
- Weising, K., and Gardner, R.C. 1999. A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. *Genome* **42**: 9-19.