Progress report: Identification of different ecotypes and centers of adaptive genetic diversity in American chestnut

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

Ouputs

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, 2005) have been characterized at genetically mapped EST-SSRs (Kubiasik 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. Until now 10 markers have been characterized in all samples and a first data analysis has been performed. 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 1min 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 $MgCl_2$ (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_2O. PCR products were separated on an ABI 3130xl Genetic analyzer (Applied Biosytems) and alleles were called using Genemapper v. 4.0 (Applied Biosytems).

Plans for year 2

Additional 10 to 25 markers will be tested for amplification and polymorphism to obtain a total of ~ 30 markers for the outlier screens. Outlier screens will be performed at all markers in population pairs from contrasting environments. 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.

Budget

A total of $2,000 is requested for the second year of the program. We will do additional marker tests, outlier screens and the final data analyses. The original proposal including all references is attached.

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

Marker_name	GenBank_ID	primer1_sequence	primer2_sequence	Motif	BLASTN	LG and position (cM)
CmSI0031	**290474526**	**AGCGCCACTTTTCTTTTTCA**	**GAATCCCAGCCTGACCAATA**	**AAG**	-	**LGH (25.3 cM)**
CmSI0049	**290474533**	**CCGATGCCGATTTCTACAAC**	**GCGGCAGACACATAGTTTCA**	**TCTTCA**	-	**LGB (34.1 cM)**
CmSI0051	290474536	CGATCATATCCCATACCCACA	GCGGAGACCACTAAGAGACG	CT		
CmSI0327	**290474585**	**CTCCTCCGCCTCCATTCA**	**AGTCCTTGCGGATCACTTTG**	**AG**	60S ribosomal protein L27-like, 8e-43	LGK (5.8 cM)
CmSI0383	290474601	CCTCCTCCACCACCAGTTTA	TGGAGGTGGGGACTTGTACT	CTC		
CmSI0391	**290474604**	**TTTGGAGTGCTGAGGAGACT**	**GACTCAAAGTCCTCGGCAAA**	**TAA**	-	
CmSI0392	290474605	CCATCGGGAAATGTGGTATT	GCTCATCTGGGAACCACTGT	GTG		LGC (26.4 cM)
CmSI0396	**290474606**	**AACTCCCACCACTCACATCC**	**TTTCGGACCATCCAGAACTC**	**CACACC**	uncharacterized protein, $1e^{-4}$	
CmSI0437	290474621	GGGCTTCTTGGAAACTAGCA	CCATACGAACACCGAGGACT	TG		LGH (49.4cM)
CmSI0495	**290474633**	**GAAACAACAGGCTCTGCCTC**	**CTGGGAAAATCCGAACTCAA**	**GA**	probable amino-acid acetyltransferase NAGS2, $9e^{-78}$	
CmSI0527	**290474668**	**TACGACCTAAACGACTCGCC**	**AGGAGGAGAAGTCAGCCACA**	**GTT**	-	LGF (14.8 cM)
CmSI0537	**290474676**	**AGAGATGGGTTGGGAAGGTT**	**GGCCTCTCTGGTTTGTGTGT**	**AG**	-	**LGA (41.4 cM)**
CmSI0541	290474681	CCAATCCCAAAATCCACTTG	GGACATTTTGGAGCCTGAAA	GA		
CmSI0551	**290474692**	**TAACCAATCACGTTCACCGA**	**CGCCACATCTAAAACCCCTA**	**CGC**	-	**LGH (26.3 cM)**
CmSI0559	**290474700**	**AGGTGGGAGTGAAGGTGTTG**	**TATCTCGCTCGCTCCATCTT**	**AGG**	-	**LGC (33.3cM)**
CmSI0561	**290474702**	**CGTATAGGGTGGAAACGGAA**	**GGACAAGCAAATCACGGAAT**	**TCG**	-	
CmSI0594	290476538	GCCCCCAAAGAAAGAGAAAG	GCATGCCCATACCCATTAAC	GGT		LGH (17.2 cM)
CmSI0600	**290474736**	**TCACACTCACACACGCAAAA**	**TGTTCACGAAATACGCAACG**	**CT**	Uncharacterized locus, $1e^{-74}$	
CmSI0603	290474740	ACTCCATGGGAATGATGAGC	TGTGTGTGTGTGTTTCTCTGTGA	TC		
CmSI0608	**290474745**	**TTTGATTGGCCCTCCTAGAC**	**CTGAATCGCCGAAACTCATT**	**AGG**	transducin/WD40 repeat-like superfamily protein, $9e^{-32}$	LGL (17.8 cM)
CmSI0611	**290474746**	**GCTGACCCTGTCAACAACAA**	**CAGAACTAGACAAGGATCACAAGA**	**TC**	-	**LGK (20.1 cM)**
CmSI0632	290474764	TCGGAGTAGTTGGAGCAGTG	TGAGAAAGGAAAGTGCGTCA	GTT		
CmSI0678	**290474795**	**GGTCAGACCCGCTAGCTCT**	**ACCCAAAACCAAAACCAAAA**	**TCT**	-	**LGK (19.3 cM)**
CmSI0683	**290474802**	**CACCAGCACCTCACTTCTCC**	**CCGGAAGATTTAGGTTTAGGG**	**AGA**	-	
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.

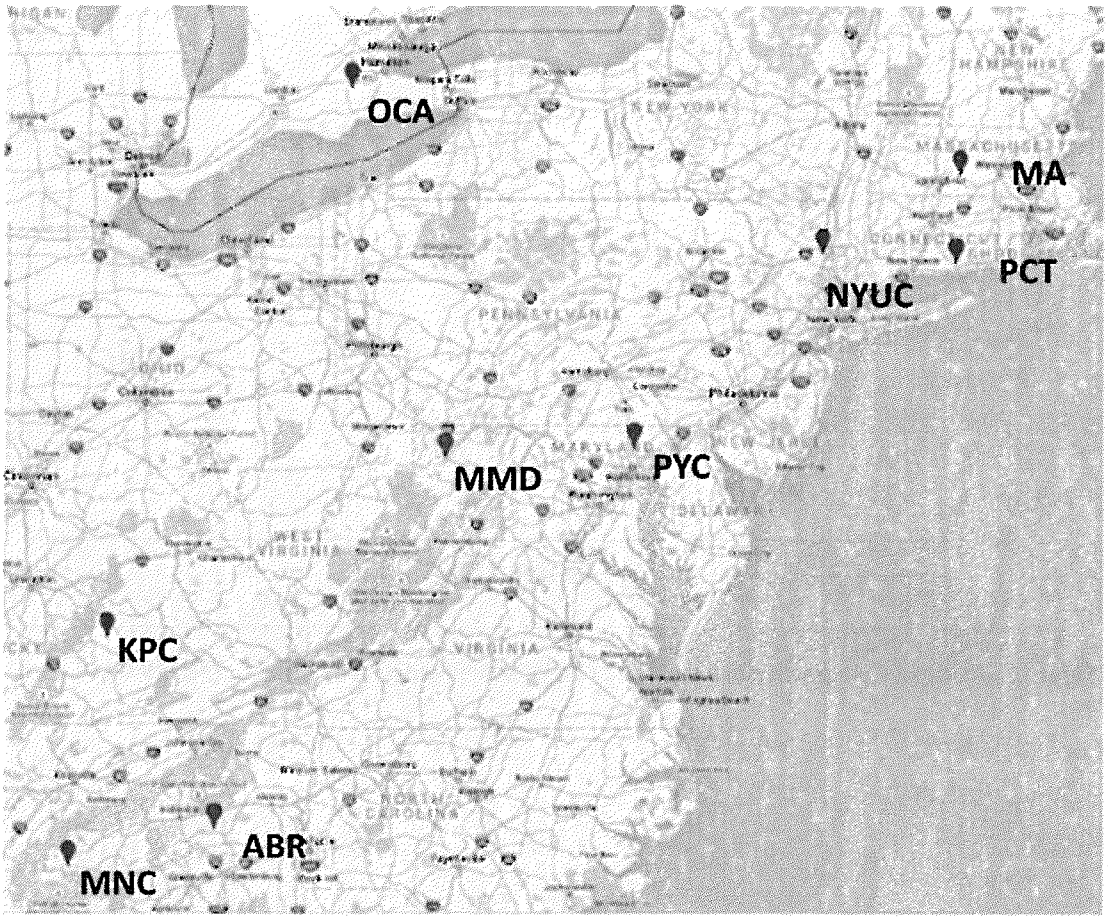

Figure 1. Sample locations

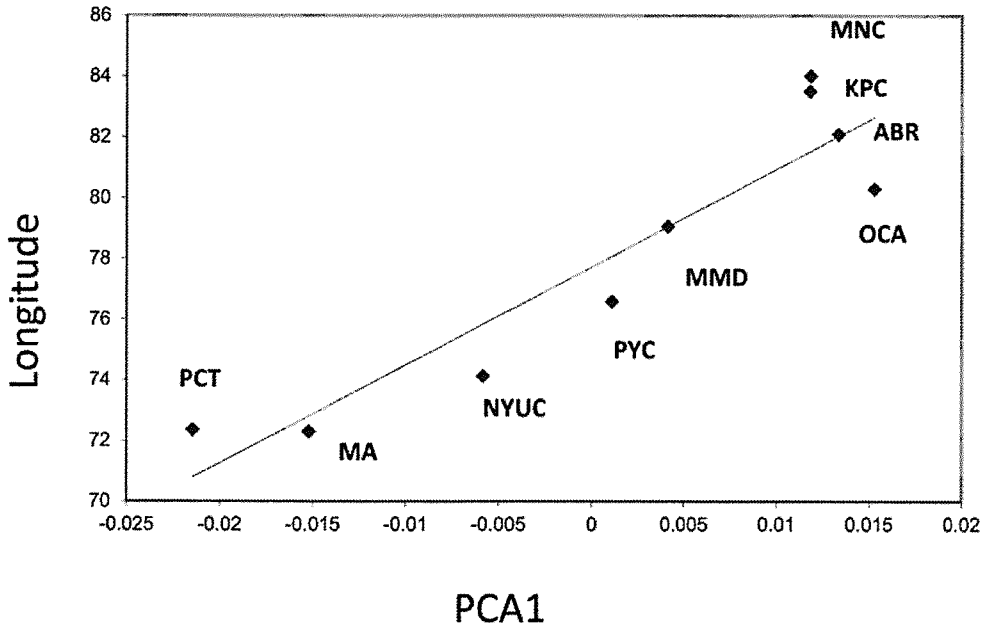

Figure 3. Association between longitude and PCA1.

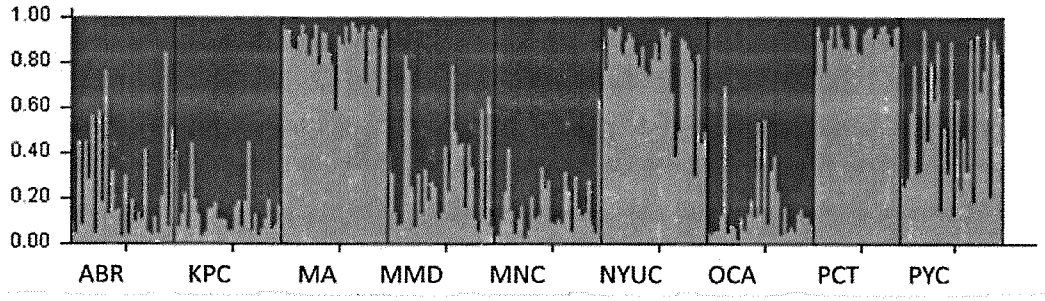

Figure 4. STRUCTURE plots for K = 2.

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. *Phytophtora 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

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 GeneAlEx v.6.41 (Peakall and Smouse, 2006).

Association of allele frequencies with environmental variables

h. Timeline:

	Year 1				Year 2			
Activity	**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.

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.