Final report for funded research: **Assessing the functional genetic diversity of blight resistance in Chinese chestnut (*Castanea mollissima* Blume) by whole-genome resequencing of a diverse germplasm collection.**

Principal Investigators

Nicholas R. LaBonte, M.S.

Purdue University Department of Forestry and Natural Resources

715 W. State Street, West Lafayette, IN 47907, USA

nlabonte@purdue.edu; 262-389-5815

Dr. Keith E. Woeste

USDA Forest Service Northern Research Station Hardwood Tree Improvement and Regeneration Center at Purdue University, 715 W. State Street, West Lafayette, IN 47907, USA

Summary

Several quantitative trait loci associated with interspecific variation in blight resistance have been described in hybrids of American (*Castanea dentata*) and Chinese chestnut (*C. mollissima*), but the genes underlying these QTL remain unknown. Disease resistance genes in plants (and animals) may be subject to balancing or diversifying selection, in which heterozygosity at the individual level, and allelic diversity at the population level, confer a fitness advantage. If high allelic diversity is present at chestnut blight resistance loci, maintaining allelic diversity at blight resistance loci would be essential to developing a blight-resistant chestnut population for restoration in the eastern United States. To determine the likelihood of diversifying selection at blight resistance loci, we sequenced whole genomes of 24 individual chestnuts with varying levels of blight resistance. We initially analyzed the established blight resistance QTL scaffold sequences and found that a few predicted genes contained most of the polymorphisms statistically associated with differences in blight resistance, especially at the *cbr1* locus. These genes showed somewhat higher nucleotide diversity in the most resistant Chinese chestnuts, versus less-resistant Chinese chestnuts and highly susceptible American chestnuts and hybrids. We also assembled and analyzed the full genomes of the same 24 individuals using a pseudochromosome assembly provided by Dr. John Carlson. The whole-genome analysis

identified additional predicted genes with polymorphisms significantly associated with blight resistance. Some of these predicted genes aligned to cDNA contigs from chestnut transcriptomes. It also confirmed that some putative blight resistance loci show evidence of moderate diversifying selection in the most resistant Chinese chestnut, but also demonstrated that other putative blight-resistance loci have very low nucleotide diversity in both Chinese and American chestnut. These results indicate that allelic diversity at some blight-resistance loci in Chinese chestnut contributes to enhanced resistance, while at other loci there is essentially one resistant allele possessed by all Chinese chestnuts.

Completion of short-term goals from grant proposal (2015)

-Sequence the genomes of 20 mature Chinese chestnuts with variable blight resistance

-17 Chinese chestnuts sequenced (some with Japanese chestnut admixture), 3 complex hybrids, a European/American chestnut hybrid, two American chestnuts, and one BC1 chestnut

-Sequence the genomes of 'Clapper' and 'Graves,' resistance donors for the ACF breeding program

-Sequenced genome of 'Clapper;' 'Graves' sample not available, sequenced 'Mahogany'

-Identify polymorphisms associated with resistant Chinese chestnuts and determine whether they occur in protein-coding regions

-Genome scanned for coding and non-coding SNPs statistically associated with blight resistance

-Assess the sequence diversity of QTL regions associated with blight resistance and candidate resistance genes

-Nucleotide diversity statistics calculated for genes potentially involved in blight resistance

-Identify SNP haplotypes associated with elevated blight resistance

-SNPs identified with one allele fixed in Chinese chestnut, and an alternate allele fixed in American chestnut

Introduction

Chinese chestnut (*Castanea mollissima* Blume) has a large native range in temperate eastern and central China and was introduced to the United States in the early 1900s as a nut tree. Chinese chestnut is believed to be a native host of chestnut blight. Because its resistance to blight

damage is consistently higher than other *Castanea* species, Chinese chestnut was selected as the resistance donor in the backcross breeding program devised by Dr. Charles Burnham and colleagues (Burnham et al. 1986). In this plan, after three generations of backcrossing hybrids to American chestnut and only advancing trees with elevated resistance and similar appearance to American chestnut, selected trees are intercrossed to produce a population (BC3F2) in which a small percentage of trees breed true for blight resistance.

This strategy was based on a hypothesis that a small number of major genes controlled blight resistance. Since then, quantitative trait locus (QTL) mapping using neutral markers (Kubisiak et al. 1997, 2013) has indicated that three genomic regions account for three-quarters of the variation in blight resistance among Chinese/American chestnut hybrids with minor loci making up the balance. The major QTL regions make up a small percentage of the genome of Chinese chestnut, but the individual genes that confer blight resistance remain unknown.

The success of the American chestnut restoration breeding effort depends on recovering nearly all of Chinese chestnut's blight resistance in advanced backcross progeny. Since there is considerable variation in blight resistance among individual Chinese chestnuts, choosing the best possible resistance donors would increase the likelihood of meeting the program's goals. The primary goal of our research was to determine if functional genetic diversity at blight resistance loci in Chinese chestnut confers greater resistance to heterozygous trees, or, conversely, if blight-resistance genes are so highly conserved in Chinese chestnut that all individuals of the species carry an identical resistance allele.

Materials and Methods

Plant Material

Chestnuts (n=20) were selected from the germplasm collection of the Empire Chestnut Company, along with two American chestnuts from Purdue University and 'Clapper' and 'Mahogany' samples provided by Dr. Laura Georgi of TACF (Table 1). Chinese chestnuts from Empire Chestnut Company were selected based on blight symptoms: the most resistant and most susceptible Chinese chestnut trees available were sampled. Controlled inoculations were not performed; phenotypes were based on the health of trees at the time of sampling and long-term observations collected by Dr. Greg Miller of Empire Chestnut Company. DNA was sampled

from leaves and dormant twigs. Tissue was ground in liquid nitrogen, then extracted using CTAB buffer and a phenol:chloroform extraction protocol. OneStep PCR Inhibitor Removal spin columns (Zymo Industries) were used to purify samples prior to submission to the Purdue Genomics Core facility for sequencing. Sequencing was carried out on an Illumina HiSeq 2500 platform and paired-end 100 bp reads were produced.

Assembly and SNP calling

The cbr QTL sequences (Kubisiak et al. 1997, 2013; Staton et al. 2015) consisted of several hundred individual scaffold sequences. Assembling short reads to each short scaffold and identifying polymorphisms proved to be computationally taxing, so we concatenated the individual scaffolds with a 300-bp sequence of missing (N) nucleotides between each individual scaffold reference sequence. Since we didn't know the actual linkage order of the scaffolds within each QTL, they were concatenated in random order to obtain a single reference sequence for each QTL. These concatenated sequences were the references we used for our initial assemblies. Once the whole genome was obtained from Dr. John Carlson, pseudochromosome sequences were used as references. A slightly modified version of the Genome Analysis Toolkit (GATK) best practices workflow was used to assemble short reads to reference and call SNPs (Depristo et al. 2011). Chloroplast sequences were also assembled for each individual using the same procedure, using the Chinese chestnut complete chloroplast sequence (Jansen et al. 2011) as a reference.

Gene Prediction and Annotation

Genes were predicted for the cbr QTL scaffold sequences and for the whole-genome sequence using AUGUSTUS gene prediction software (Stanke et al. 2006). Protein sequences of predicted genes were aligned to the UniProt-SwissProt curated protein database (Boutet et al. 2016) using the DIAMOND sequence aligner (Buchfink et al. 2015). The best alignment from this database was used to assign a hypothetical function for each predicted gene.

Predicted proteins from AUGUSTUS were used to generate a protein database in DIAMOND, and cDNA contigs from chestnut transcriptomes (Barakat et al. 2009, 2012; Serrazina et al. 2015) were aligned to this database using DIAMOND's blastx algorithm. A

predicted protein was counted as having transcriptomic support if it was the best protein alignment for at least one transcriptome contig.

Association Analysis and Calculation of Statistics

After filtering the whole-genome SNP dataset for quality (sum of quality scores/site > 1000) and read depth per individual (minimum: 10, maximum: 45), association analysis was performed using Plink software (Chang et al. 2015). Permutation tests were used to determine statistical significance because the number of tests performed (all SNP loci in the genome) was very large. Regions with unusually large numbers of blight-associated SNPs were discovered by dividing the genome variant file into 5000-SNP bins and tallying the number of SNPs statistically associated below a given P-value cutoff (0.005 or 0.001) for each bin. VCFtools (Danecek et al. 2011) was used for SNP filtering and calculation of F_{ST}, Tajima's D, π, and heterozygosity.

Results and discussion

Chloroplast sequence analysis

Analysis of the sample's chloroplasts revealed expected (and unexpected) hybrid origin for some the 24 trees we sequenced (Figure 1). Based on chloroplast markers, the mother of the hybrid tree "Paragon" was a European sweet chestnut, *C. sativa*. Some Chinese chestnuts that had been collected in the United States from unknown provenances, like 'Schmucki,' actually had a *C. dentata* chloroplast. This is somewhat surprising considering the near-immunity of that particular tree to blight damage. The "northern Chinese" chestnuts we sequenced were actually derived from Korea and possessed Japanese chestnut chloroplasts (Table 2). 'Clapper' had a different (*Cm*) chloroplast from all the other *Cm* sequenced; all *Cm* other than 'Clapper' had chloroplasts identical to the reference. Based on sequences of wild Chinese chestnuts we sampled, the 'Clapper' haplotype is most common in orchard trees from northern China and is found in several wild populations in northern and southern China. 'Mahogany,' 'Nanking,' and all the other southern Chinese chestnuts sequenced have the same chloroplast, which we found to be most common in southern Chinese trees. We also identified a single *Cd* chloroplast haplotype among the four *Cd* chloroplasts in our sequences.

Sequence diversity in blight resistance QTL sequences

On average, the three cbr sequences had fairly high Tajima's D values ranging from 1.6-1.7 when only coding sequences were used to calculate the statistic (Table 2), indicating diversifying selection. Across most of cbr1, Tajima's D was lower among susceptible Chinese chestnuts than resistant trees (Figure 2). While Tajima's D was also lower in general across cbr2 and cbr3 (Figure 3, Figure 4) the pattern was less distinct. In cbr1 there were 90 individual 3000-bp genome segments (windows) in which Tajima's D for resistant trees was greater than 1 (near average) and the statistic for susceptible trees was less than -1. Conversely, there were only 6 windows where the opposite was true (resistant < -1 and susceptible > 1). This indicates that there may be relatively low allelic diversity in susceptible Chinese chestnuts at this locus. In cbr2, there were 10 windows with D(resistant) >1 and D(susceptible) <-1 versus only one with D(resistant) < -1 and D(susceptible) > 1. In cbr3, the results were more equivocal. There might be some increased haplotype diversity in resistant trees at cbr2, but it does not appear to be a major factor as it does in cbr1. Nucleotide diversity values averaged 0.00694 -0 .00717 across the cbr sequences.

Genes potentially associated with blight resistance

In each cbr sequence, one or two predicted genes stood out as containing the largest number of polymorphisms associated with blight resistance. In cbr1, of 10 SNPs with the highest association scores, six were located in or near a single predicted gene occurring around base 4,280,000. When the predicted protein sequence was submitted to a BLAST search, the most similar proteins in the database were MATE-like proteins from plants. MATE (Multidrug and toxic compound extrusion) proteins are cation-driven efflux pumps that move compounds between compartments in cells. In plants, they seem to mainly transport small organic molecules and compounds produced by other organisms (i.e. pathogens). Several MATE-family proteins have been shown to play a role in disease resistance in *Arabidopsis*. Our earlier analyses identified several copper-containing oxidase genes in cbr1. Genes in this family have a role in lignin biosynthesis, and lignin plays a role in plant defense against fungi. However, when permutation tests were used to calculate the significance of association scores and a depth filter

was used to reduce the number of spurious SNPs, these genes were no longer clearly the best candidates[1]. It is possible that their initial high scores were due to gene duplication resulting in misassembly of duplicated genes and spurious SNPs, or their association scores were inflated by missing data in some individuals. The highest single association score belonged to a SNP in a short gene around base 632,000 that contains a senescence-associated domain.

In cbr2, the result of the association analysis was not as decisive. All of the top 10 most highly-associated SNPs were located near one extremely small predicted gene that did not align to any known protein in the BLAST databases. This small protein, however, was similar to several RNA transcripts from previous chestnut transcriptome studies, and was located near a glycerol-3-phosphate acyltransferase-like gene. Two significantly associated SNPs were located in a predicted NBS-LRR (nucleotide binding site leucine-rich repeat) gene. NBS-LRR genes are also located on cell membranes, and they are thought to act as gatekeepers that recognize compounds from pathogens and trigger defensive reactions in the cell. They are the largest and best-understood group of disease resistance genes in plants.

cbr3 was similar to cbr2 in that several fragmentary predicted genes, retrotransposon parts, and other potentially spurious features made up most of the most highly associated SNPs. It is possible that these features are closely linked to the actual causative gene, or have some biological relevance because they represent disrupted or degraded former gene sequences. If this was the case, however, BLAST alignments should more clearly indicate it. Four associated SNPs were located in a predicted gene with similarity to known epoxide hydrolase-lyase genes. These genes function in lipid metabolism, but their possible roles in disease resistance is not well-documented or understood.

Sequence diversity measures in cbr QTL genes most closely associated with blight resistance

We calculated Tajima's D, heterozygosity, and π for the genes most associated with differences in blight resistance (Table 2). We also calculated it separately for the set of resistant Chinese chestnuts and hybrids, susceptible Chinese chestnuts and hybrids, and trees known to

have 50% or less *C. mollissima* ancestry (the two American chestnuts, "Paragon" and its offspring, and 'Clapper'). Tajima's D values for all three genes, across all three samples of chestnut, were lower than the averages over the entire cbr sequences (~1.5) (Table 3). This indicates that they are probably not subject to strong diversifying selection. When Tajima's D was calculated separately for three groups of trees (resistant Chinese chestnut, susceptible Chinese chestnut, and non-Chinese chestnuts) some interesting patterns emerged. For the most strongly-associated gene in cbr1, Tajima's D was considerably lower in susceptible Chinese chestnuts than in resistant trees. We can't conclude that diversifying selection is at work in cbr1 or at this gene – the Tajima's D value for resistant trees is still fairly low—but it does appear that more haplotypes are present in resistant *C. mollissima* than susceptible *C. mollissima*. The value for resistant Cm is similar to that for non-Cm trees (mostly *C. dentata*). π, the fraction of loci that are polymorphic (Table 4), is essentially the same for the three groups. Heterozygosity is higher in resistant Chinese chestnuts (0.429) across the cbr1 gene than it is in susceptible Chinese chestnuts (0.156) or American chestnuts (0.048). For the cbr2 gene, all three groups (resistant Cm, susceptible Cm, and susceptible species have similar values of heterozygosity. Results for the cbr3 gene were similar to those described for cbr1, but less striking. Based on association results, the resistance genes in cbr3 are weaker candidates for blight resistance than the predicted gene with most associated SNPs in cbr1.

Identification of genes associated with blight resistance from the draft genome assembly of Chinese chestnut

Our analysis identified hundreds of individual SNPs statistically associated with differences in blight resistance on every linkage group, but many of these were concentrated in relatively small portions of the genome. LGA hosted the largest number of regions with concentrations of resistance-associated polymorphisms. Although blight-associated regions in our study were considerably smaller (in terms of base pairs) than the currently understood chestnut blight resistance QTL regions, taken together they still contain several hundred predicted gene sequences. We assessed the strength of each candidate gene in these regions based on several criteria. First, does the predicted gene sequence contain nucleotide variants that are likely to change the protein sequence, and if so, are these variants significantly associated with blight resistance? Second, does the predicted protein sequence align to known proteins that

have a biological function related to the molecular mechanisms of chestnut blight infection response? Third, is there evidence in databases of chestnut gene expression (RNA-seq) data that the predicted gene is actually expressed? Finally, is there any evidence of differential expression in healthy/cankered tissue, or differential expression in American vs. Chinese chestnut?

Analyzing genes based on these criteria (Table 3), one or several "best" candidate genes were selected for each of the associated-SNP regions (referred to from now on as "loci") on the twelve linkage groups. For some loci, the best candidate gene contained SNPs that were strongly associated with differences in blight resistance and were predicted to cause amino acid changes to the predicted protein (LGA.a, LGA.d-e, LGB.a-d, LGD.a, LGJ.a, LGL.a), although some of these predicted genes, most notably on LGB and LGL, did not have support from available transcriptome data. g3006 (LGB.b locus) corresponds exactly to the "best" gene from the standalone analysis of cbr1. By contrast, the 12 genes that showed differential expression in cankers vs. healthy stems in either Chinese or American chestnut (Barakat et al. 2012) did not often have highly-associated nonsynonymous polymorphisms within the predicted exons of the predicted gene. This was not particularly surprising, because differences in mRNA expression are generally not due to differences in the actual coding sequences, but rather to differences in promoters and enhancers near the gene, or by the methylation of the gene's DNA. For genes at the loci LGA.a, LGA.c, LGA.d, LGC.a, LGD.b, LGG.c, LGJ.a, and LGK.a there were SNPs immediately upstream (< 500 bp) of the predicted transcription start site of a candidate gene that had statistically significant associations with blight resistance. Several of these predicted genes (LGA.d, LGK.a, LGG.c) were differentially expressed in cankers vs. healthy stem tissue of American and Chinese chestnuts. Others (LGC.a, LGD.b) showed evidence of transcription in Chinese chestnut, but not in American chestnut.

Since resistant Chinese chestnuts, relatively susceptible Chinese chestnuts, and highly susceptible American chestnuts and hybrids were included in the study sample and analyzed together, the most statistically significant associations at SNP loci were for those loci where resistant Chinese chestnuts share an allele that was found in neither susceptible Chinese chestnuts nor American chestnut. The strongest statistical association would be for a hypothetical locus with genotype 1/1 in all resistant trees and 0/0 in all susceptible trees, so that allele frequencies for the 0 allele would be 1/0 in susceptible and 0/0 in resistant trees. For the

strongest statistical associations that were actually observed, allele frequencies for the 0 allele were 1.0 in susceptible trees (i.e., susceptible trees would have all susceptible alleles), and ~ 0.5 in resistant trees; the most resistant Chinese chestnuts tended to be 0/1 heterozygotes for these SNPs. In these cases, a rare alternate allele present in the most resistant Chinese chestnut was associated with resistance. This was the pattern observed in genes at the LGA.d and LGA.e locus, for example, which had the strongest signal of statistical association with blight resistance. These loci most likely affect the marginal differences in resistance within Chinese chestnut and differences in resistance between American and Chinese chestnut. Next, we considered SNPs that contributed only to the (large) difference in resistance between American chestnut and all Chinese chestnuts, including the most susceptible. This type of SNP would have one allele (0) fixed in all Chinese chestnuts and another (1) fixed in American chestnut. Because a large part of the most susceptible trees we sampled were American chestnuts and 'Paragon' offspring, SNPs with this pattern had a statistically significant association with blight resistance, but the strength of the association was lower than those where resistant Chinese chestnuts had a unique allele. This pattern was observed at loci on LGF and LGG in particular, which likely correspond to *cbr2* and *cbr3*. Predicted genes with associations to blight resistance in the standalone cbr2 and cbr3 analysis, which were relatively poor candidates, were not re-discovered in the whole-genome analysis.

Nucleotide diversity at the predicted genes deemed the most likely blight-resistance candidates did not follow a consistent pattern; at some genes, Tajima's D, nucleotide diversity, and heterozygosity were lowest in the most resistant Chinese chestnuts, particularly in blight-associated regions on LGF and LGG. At others, statistics indicated evidence for diversifying selection in Chinese chestnut, in particular, across blight-associated genes on LGA and LGL. (Table 4).

Hints to the molecular basis of blight resistance and susceptibility

Most of the known disease resistance genes in annual crop plants (R genes) are involved in detecting a pathogen by binding to some molecule the pathogen produces and initiating a defensive response. Often, these resistance genes encode a protein that spans the cell membrane, with a receptor for fungal molecules (pattern-recognition receptor or PRR) protruding outside the cell and a protein kinase or other domain to transmit a message inside the cell. The PRR

portions of genes like this tend to be involved in the "arms race" between plant and pathogen, and often show a great deal of allelic diversity. Often, membrane bound PRR genes detect biotrophic pathogens (like rusts), which depend on living plant cells. The resistance gene initiates a defensive response that kills local cells—sort of a "scorched-earth" strategy to starve out the biotrophic fungus or kill it outright with reactive oxygen species and other damaging chemicals (Agrios 2005). This is often referred to as the hypersensitive response (HR) and the intentional death of cells as programmed cell death (PCD). Chestnut blight is not a biotroph; it is a necrotroph, which means that it kills plant cells prior to digesting the contents. There are documented cases in crop plants of necrotrophic pathogens that stimulate the HR and PCD in order to kill plant tissue more easily – effectively tricking the plant into killing its own tissue. Is this a potential mechanism for American chestnut's ineffective response to chestnut blight? Loci LGA.d and LGL.c both contain clusters of genes associated with resistance to biotrophs. Both LGA.d and LGL. c contained predicted genes that aligned to transcripts that were differentially expressed in inoculated stems of American chestnut, but not Chinese chestnut (Barakat et al. 2012). Part of Chinese chestnut's resistance to chestnut blight might be its ability to suppress genes associated with the HR. Another gene expressed more in American, but not Chinese, chestnut cankers is the predicted kinase at LGB.d, which is involved in regulating PCD and disease resistance responses to a variety of pathogens in *Arabidopsis* (Christiansen et al. 2011).

If American chestnut's ineffective response to chestnut blight is caused by disease resistance genes engaging in a runaway PCD reaction, how does Chinese chestnut manage that response better? One intriguing predicted protein at LGE (g8469) encodes a protein with LISH and HEAT domains and is up-regulated in Chinese chestnut infected stems. The predicted structure of this protein includes two extracellular domains and a short cytoplasmic region. The HEAT repeat portion of the protein may be involved in protein-protein interactions, so this protein could interact with other disease response-related membrane proteins to modulate the plant's immune response. Other predicted genes in our putative blight-resistance loci that showed evidence of up-regulation in Chinese chestnut cankers included two (LGB.b PIN-LIKES 2, LGG.b nicotinamidase) involved in auxin and abscisic acid hormone signalling, and one (MIEL1 at LGK.a) directly involved in the negative regulation of PCD in *Arabidopsis.* Abscisic acid and auxin both repress PCD in disease-resistance responses (Mauch-Mani and Mauch 2005, Danquah et al. 2014, Eshragi et al. 2014). One auxin-ABA-related gene (ELF3) was up-

regulated in American chestnut infected stems. A predicted serine carboxypeptidase and a predicted carboxylesterase were also up-regulated in infected Chinese chestnut stems, but since the most similar genes in *Arabidopsis* are not thoroughly annotated it is difficult to infer their molecular function.

Are there links between *Cryphonectria parasitica* and *Phytophthora cinammomi* resistance?

Another pathogen important to the chestnut restoration effort is *Phytophthora cinammomi,* an oomycete and the causative agent of a root rot disease that eliminate American chestnut from many low-elevation habitats. *Phytophthora* spp are hemibiotrophs; they begin infection by infiltrating living tissue and transition to necrotrophy shortly after they've entered the plant, so their lifestyle is somewhat different from chestnut blight. We aligned transcriptome sequences from a study of root rot disease in European (susceptible) and Japanese (resistant) chestnut. Some cDNA sequences that were upregulated in response to root rot infection in that study aligned to predicted proteins from our study: an ALF4-like protein at LGA.a (distinct from the one presented in tables here), an ethylene-responsive transcription factor at LGG.D, and one disease resistance-like gene at LGL.c (the one up-regulated in American chestnut cankered stems) were differentially expressed in infected vs. non-infected roots of Japanese chestnut. In European chestnut, the only differentially expressed genes in infected vs. non-infected roots that coincided with blight resistance loci were three separate NBS-LRR genes at LGL.c, including the one that was also upregulated in blight-infected American chestnut stems. This indicates that the resistance gene cluster at LGL.c might include some NBS-LRR genes that are involved in general disease resistance responses, and others that relate to a specific pathogen.

Interspecific genetic variation at disease resistance genes

Because American and Chinese chestnut have been reproductively isolated for > 10 million years, numerous genetic differences have accumulated between the species. For many predicted gene sequences, extremely high (0.9 and greater) values of F_{ST} were observed. High interspecific F_{ST} indicates genes where a different allele is fixed in each species; heterozygosity is low within each species, but very high in an F1 hybrid. Low interspecific F_{ST} indicates either a gene that is so highly conserved that Chinese and American chestnut share the same allele (low heterozygosity in species and hybrids) or a gene that evolves so rapidly that most heterozygosity

is found within, rather than among, the species (high heterozygosity in species and hybrids). We isolated the predicted genes with the highest (>0.9) and lowest (<0.1) F_{ST} calculated among species and submitted them to gene ontology (GO) enrichment analysis. Gene ontology enrichment analysis identifies biological functions that are over-represented (enriched) in a sample of genes. We found that both the high- and low-interspecific F_{ST} gene sets were enriched for GO terms related to disease resistance. Manually examining the genes in the low-F_{ST} set, we found that most had uniformly high heterozygosity in hybrids and pure species; i.e., they are rapidly-evolving, highly diverse sequences. These results are interesting because they indicate that disease resistance genes in the chestnut genome include some that are more highly conserved within species (low diversity) than the rest of the genome, and others that are less conserved within species than the rest of the genome. This is probably due to the fact that plant disease resistance includes many proteins involved in pattern recognition (rapidly evolving) and many that are involved in conserved signalling pathways (slowly evolving).

Conclusions

We discovered evidence that the most blight-resistant Chinese chestnuts possess higher heterozygosity at some putative blight-resistance genes than relatively susceptible Chinese chestnuts and highly susceptible species. Some putative blight-resistance loci, on the other hand, seem to have a single allele fixed in all Chinese chestnuts. 'Clapper' does not possess a Chinese allele at the cbr1 locus, but probably does at cbr2 and cbr3. Our results indicate that the liability involved in depending on one or two resistance donors depends on exactly which genes cause differences in blight resistance among species. In general, our results support the TACF initiative to include more resistance donors. It is possible that high nucleotide diversity and heterozygosity only confer a marginal advantage in resistant Chinese chestnuts, and do not play a role for genes involved in the difference in resistance between American and Chinese chestnuts. Even if this is the case, the American chestnut blight resistance breeding program is most likely to succeed on a large scale if highly resistant planting stock is developed. This requires incorporating all the blight resistance of Chinese chestnut, which is likely to require a diverse pool of blight resistance donors. We are currently developing and screening SNP markers from the blight resistance candidate genes described here and screening them in the Indiana ACF

breeding population of BC3F1 trees to validate the genes as candidates and develop targeted marker-assisted selection for chestnut blight resistance in hybrid breeding populations.

Works Cited

Agrios GN (2005) Plant Pathology: 5th Edition. Elsevier Academic Press, Burlington, MA, USA ISBN-13 978-0-12-044565-3.

Barakat, A., DiLoreto, D.S., Zhang, Yi, Smith, C., Baier, K., Powell, W.A., Wheeler, N., Sederoff, R., Carlson, J.E. 2009. Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (C. mollissima) in response to the chestnut blight infection. BMC Plant Biology 9:51.

Barakat A, Staton M, Cheng C-H et al. (2012) Chestnut resistance to the blight disease: insights from transcriptome analaysis. BMC Plant Biology 12:38.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. Methods Mol Biol 1374:23-54.

Buchfink B, Xie C, Huson D (2015) Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59-60.

Burnham, C.R., Rutter, P.A., French, D.W. 1986. Breeding Blight-Resistant Chestnuts. Plant Breeding Reviews 4: 347-397.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

Christiansen KM, Gu Y, Rodibaugh N, Innes RW (2011) Negative regulation of defence signaling pathways by the EDR1 protein kinase. Mol Plant Pathol 12:746-758.

Danecek P, Auton A, Abecasis G et al. (2011) The Variant Call Format and VCFtools. Bioinformatics27(15):2156-2158.

Danquah A, de Zelicourt A, Colcombet J, Hirt H (2014) The role of ABA and MAPK signaling pathways in plant abiotic stress responses. Biotechnology Advances 32:40-52.

DePristo M, Banks E, Garimella K et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43:491-498.

Eshraghi L, Anderson JP, Aryamanesh N, McComb JA, Shearer B, St J E Hardy G (2014) Suppression of the auxin response pathway enhances susceptibility to *Phytophthora cinammomi* while phosphate-mesiated resistance stimulates the auxin signaling pathway. BMC Plant Biology 14:68 doi:10.1186/1471-2229-14-68.

Jansen RK, Saski C, Lee S-B, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three rosids (*Castanea, Prunus, Theobroma*): Evidence for at least two independent transfers of rpl22 to the nucleus. Mol Biol Evol 28(1):835-847.

Kubisiak, T.L., Hebard, F.V., Nelson, C.D., Zhang, J., Bernatzky, R., Huang, J., Anagnostakis, S.L., Doudrick, R.L. 1997. Molecular mapping of resistance to blight in an interspecific cross in the genus Castanea. Phytopathology 87:751-759.

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 and Genomes 9:557-571.

Mauch-Mani B, Mauch F (2005) The role of abscisic acid in plant-pathogen interactions. Current Opinion in Plant Biology 8:409-414.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J (2007) g:Profiler- a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res 35:W193-W200.

Serrazina SMT, Santos C, Machado H, Pesquita C, Vicentini R, Pais M, Sebastiana M, Costa RL (2015) *Castanea* root transcriptome in response to *Phytophthora cinnamomi* challenge. Tree Genet Genomes 11(1) DOI: 10.1007/s11295-014-0829-7

Stanke M, Schoeffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62.

Staton M, Zhebentyayeva T, Olukolu B, Fang GC, Nelson CD, Carlson JE, Abbott AG (2015) Substantial genome synteny preservation among woody angiosperm species: comparative genomics of Chinese chestnut (*Castanea mollissima*) and plant reference genomes."BMC Genomics 16(1):744.

Table 1. Summary of plant material and provenances used in the study. R: resistant; S: susceptible; HS: highly susceptible.

Tree	Phenotype	Origin
72-132	S	ECC[1]: Southern Chinese (old introduction)
72-139	R	ECC: Southern Chinese (old introduction)
72-41.5	S	ECC: Southern Chinese (old introduction)
72-49.5	S	ECC: Southern Chinese (old introduction)
SC1 aka B66	R	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
SC2	S	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
SC3	S	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
SC4	R	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
'Paragon'	HS	ECC: *C. sativa* × *C. dentata*
Paragon-1	HS	ECC: (*C. sativa* × *C. dentata*) × *C. mollissima*
B21	R	ECC: (*C. sativa* × C. *dentata*) × C. *mollissima*
B32	HS	ECC: (*C. sativa* × C. dentata) × C. mollissima
'Schmucki'	R	ECC: (*C. dentata* × *C. mollissima*) × *C. mollissima* (?) [2]
NC1	S	ECC: B16 (Korean origin) × *C. mollissima*
NC2	S	ECC: B16 (Korean origin) × *C. mollissima*
NC3	S	ECC: *C. dentata* × *C. mollissima* (?) *
NC4	R	ECC: B16 (Korean origin) × *C. mollissima*
NC5	R	ECC: B16 (Korean origin) × *C. mollissima*
NC6	R	ECC: B16 (Korean origin)×*C. mollissima*
'Clapper'	S	TACF: (*C. mollissima* × *C. dentata*) × *C. dentata* "BC1"
'Nanking'	R	ECC: Southern Chinese (old introduction)
'Mahogany'	R	TACF: Southern Chinese (old introduction)
Roselawn-1	HS	Northern Indiana, *C. dentata* from outside accepted native range
'Johnson'	HS	Southern Indiana, *C. dentata* from within native range

[1]Empire Chestnut Company; [2]Pedigree uncertain, inferred from nuclear and chloroplast genotypes

Table 2. Summary of Tajima's D statistic in assemblies of the cbr blight resistance QTL regions (Kubisiak et al. 2013) and genes chosen for concentrations of associated SNPs for each region.

	cbr1: MATE-like gene (LGB.b.3)	cbr1 average	cbr2: NBS-LRR gene	cbr2 average	cbr3: Epoxide hydrolase gene	CBR3 average
Resistant Cm	1.061	1.643	1.052	1.677	1.275	1.714
Susceptible Cm	-0.935	0.901	1.091	1.457	-0.190	1.680
Non-Cm	0.644	1.234	0.602	1.341	0.791	1.641

Table 3. List of predicted genes in regions where most blight-associated polymorphisms were found showing predicted function and evidence for association with blight resistance based on statistical association and publicly available cDNA data (Barakat et al. 2013).

Locus	Gene[a]	Exon[b]	NSyn[c]	Inferred function[d]	Transcript[e]	Diff[f]	Clust[g]
LGA.a	g1418	3	2	Aberrant root formation protein 4	CC 2, AC 1	na	1
LGA.b	g2361	0	0	Lysophospholipase	CC 2, AC 3	na	0
LGA.c	g3528	1	0	NIP5-like aquaporin	CC 1, AC 2	CC	0
LGA.d	g4191	0	0	LRK10-like rust resistance	AC 1	na	3
LGA.d	g4193	0	0	LRK10-like rust resistance	CC 4, AC 3	AC	3
LGA.d	g4196	4	3	LRK10-like rust resistance	CC 1	na	3
LGA.e	g8459	19	7	LISH/HEAT-domain protein	CC 7, AC 1	CC	0
LGA.e	g8465			Serine carboxypeptidase	CC 1	CC	0
LGB.a	g2160	16	9	TAO1-like TMV resistance protein	na	na	0
LGB.b	g2214	0	0	Protein PIN-LIKES 5	CC 2, AC 1	CC	0
LGB.b	g2245	1	1	Cytochrome P450 90B1	CC 1, AC 1	na	0
LGB.c	g3006	2	1	DETOXIFICATION 27 MATE-like	na	na	2
LGB.d	g5043	1	1	F-box protein	CC 1	na	6
LGB.d	g5048	2	1	Protein kinase EDR1	AC 1	AC	0
LGC.a	g3384	0	0	MLP-like protein 328	CC1	na	2
LGC.a	g3419	0	0	LRK10-like rust resistance	CC1	na	0
LGD.a	g1162	0	0	EARLY FLOWERING 3 -like	CC 4, AC 1	AC	0
LGD.a	g1179	2	1	Cationic peroxidase	CC 1, AC 1	na	0
LGD.b	g2262	0	0	Pectinesterase inhibitor	CC 1	na	0
LGD.b	g2282	3	1	Cysteine-rich RLK	CC 1, AC 1	na	1
LGE.a	g7940			GDSL esterase-lipase 2-like	CC1, AC 1	na	0
LGF.a	g1803	1	0	Periodic tryptophan protein	AC 1	na	1
LGF.a	g1804	1	0	Periodic tryptophan protein	CC 1	na	1
LGF.b	g2785			ERDL sugar transporter	CC 1	na	
LGG.a	g2311	1	0	Senescence/dehydration-associated	AC 2, CC 2	na	0
LGG.b	g3657	0	0	Nicotinamidase 1	AC 1, CC 2	CC	0
LGG.c	g4295	0	0	Probable carboxylesterase 5	AC 3, CC 1	CC	5
LGG.c	g4298	2	0	2-hydroxyisoflavanone dehydratase	AC 1, CC 1	na	2
LGJ.a	g238	0	0	Pathogenesis-related protein	AC 1, CC 2	na	2
LGJ.a	g240	4	3	Cytosolic carboxypeptidase	na	na	0
LGJ.b	g1363	0	0	FLX-like protein	AC 1, CC 2	na	0
LGK.a	g2007			MIEL1 ubiquitin-protein ligase	CC 1	CC	0
LGL.a	g4222	9	5	MAIN-like protein phosphatase	na	na	2
LGL.c	g6971	0	0	Probable disease resistance protein	AC 3, CC 2	AC	10
LGL.c	g6992	2	0	Probable disease resistance protein	CC 2	na	10
LGL.c	g8953			Probable disease resistance protein			
LGL.c	g8955	0	0	Retrovirus-related POL polyprotein	AC 1	AC	0

[a] Number assigned to predicted gene by AUGUSTUS gene prediction software; [b] Number of polymorphisms in predicted exons with Plink association p-value < 0.01; [c] SNPs predicted to

cause an amino acid change with Plink association p-value <0.01; [d] Function inferred from alignment to SwissProt/UniProt database; [e] Number of cDNA contigs from Barakat et al. (2013) matching predicted protein (>75% ID) in American (AC) and Chinese (CC) chestnut; [f] Differential expression in cankers vs. healthy stem tissue in American (AC) and Chinese (CC) chestnut (Barakat et al. 2013); [g] Size of gene cluster, i.e. number of genes with same or similar predicted function adjacent to the named gene.

Table 4. Nucleotide diversity, interspecific *F*st, heterozygosity, and Tajima's D statistic for groups of chestnuts sampled at selected blight resistance candidate genes. Genes where "Clapper" most likely possesses a Chinese chestnut allele are highlighted in green.

Locus	Gene[a]	π_{CM}	π_{NCM}	F_{ST}	H_{CMR}	H_{CMS}	H_{NCM}	H_{HYB}	H_{Clap}	D_{CMR}	D_{CMS}	D_{NCM}
LGA.a	g1418	0.00166	0.00180	0.461	0.32	0.25	0.32	0.21	0.00	1.985	1.567	-0.138
LGA.b	g2361	0.00099	0.0011	0.876	0.06	0.10	0.12	0.46	0.10	0.543	0.904	-0.723
LGA.c	g3528	0.00228	0.00279	0.643	0.11	0.16	0.19	0.37	0.00	nan	nan	nan
LGA.d	g4191	0.00208	0.00265	0.718	0.13	0.16	0.13	0.36	0.00	0.854	0.994	1.801
LGA.d	g4193	0.00415	0.00385	0.499	0.30	0.22	0.24	0.49	0.29	2.163	0.285	1.298
LGA.d	g4196	0.01619	0.01200	0.016	0.078	0.095	0.074	0.427	nc	2.121	0.554	0.317
LGA.e	g8459	0.00256	0.00183	0.662	0.07	0.13	0.12	0.26	0.08	-0.306	1.349	0.059
LGA.e	g8465	0.00344	0.00314	nc	0.314	0.279	0.266	0.426	nc	0.536	1.282	-0.018
LGB.a	g2160	0.01205	0.0064	0.508	0.42	0.35	0.19	0.50	0.45	2.081	2.108	**-0.011**
LGB.b	g2214	0.00146	0.0013	0.530	0.25	0.19	0.13	0.42	0.00	1.366	1.206	-0.483
LGB.b	g2245	0.00058	0.0003	0.888	0.08	0.11	0.04	0.58	0.00	0.506	-1.28	-0.186
LGB.c	g3006	0.00196	0.001	0.626	0.13	0.20	0.14	0.48	0.00	0.459	-0.21	-1.145
LGB.e	g5043	0.00676	0.0054	0.369	0.43	0.49	0.34	0.46	0.30	0.724	1.411	1.624
LGB.e	g5048	0.00313	0.0023	0.534	0.27	0.19	0.18	0.32	0.08	0.927	0.843	-0.173
LGC.a	g3384	0.00559	0.0085	0.305	0.48	0.42	0.50	0.59	0.65	1.475	0.281	-0.089
LGD.a	g1162	0.00399	0.00081	0.772	0.27	0.32	0.03	0.50	0.05	1.236	1.720	nan
LGD.a	g1179	0.00159	0.00087	0.497	0.32	0.20	0.04	0.43	0.00	0.015	1.696	nan
LGD.b	g2282	0.00149	0.00093	0.588	0.19	0.24	0.06	0.53	0.54	0.641	-0.055	nan
LGE.a	g7940	0.00179	0.00122	0.049	0.49	0.57	0.28	0.36	0.22	nan	nan	nan
LGF.a	g1803	0.00225	0.00169	0.471	0.39	0.24	0.14	0.52	0.56	1.311	1.197	-0.403
LGF.a	g1804	0.00294	0.00197	0.601	0.33	0.20	0.18	0.47	0.55	1.671	2.002	-0.286
LGG.a	g2307	0.00055	0.00056	0.939	0.10	0.03	0.08	0.53	0.10	-1.373	-0.480	-1.110
LGG.b	g3657	0.00297	0.00103	0.556	0.24	0.18	0.08	0.44	0.49	0.835	2.183	-0.492
LGG.c	g4295	0.00286	0.00337	0.429	0.35	0.05	0.36	0.39	0.57	1.559	0.874	-0.255
LGG.c	g4298	0.00313	0.00267	0.505	0.30	0.10	0.22	0.34	0.20	1.288	0.710	0.848
LGJ.a	g238	0.00093	0.01437	0.515	0.31	0.26	0.17	0.46	0.43	1.791	-0.771	-1.220
LGJ.a	g240	0.01166	0.00867	0.800	0.800	0.344	0.272	0.195	0.340	1.786	1.993	-0.250
LGK.a	g2007	0.00057	0.00108	0.847	0.12	0.18	0.33	0.54	0.26	1.383	nan	nan
LGL.a	g4222	0.00157	0.00053	0.304	0.48	0.11	0.00	0.32	0.29	2.514	-0.638	nan
LGL.c	g6971	0.00158	0.00447	0.704	0.11	0.14	0.53	0.63	0.39	-0.231	-1.245	0.636
LGL.c	g6992	0.00238	0.00271	0.703	0.16	0.17	0.22	0.51	0.12	1.357	-0.693	nan
LGL.d	g8955	0.00317	0.00429	0.795	0.18	0.13	0.23	0.70	0.72	nan	nan	1.024

Table 5. Gene ontology enrichment analysis for predicted chestnut genes with low (<0.1) interspecific F_{ST} values, based on analysis of the best *Arabidopsis* alignment for each predicted gene, using g:profiler software (Reimand et al. 2007).

GO term	p-value	Genes (of 312 total)
defense response	6.44e-05	43
innate immune response	6.99e-04	17
response to oomycetes	3.26e-02	6
protein phosphorylation	1.50e-12	53
cell death	3.953-05	14

Table 6. Gene ontology enrichment analysis for predicted chestnut genes with high (>0.9) interspecific F_{ST} values, based on analysis of the best *Arabidopsis* alignment for each predicted gene, using g:profiler software (Reimand et al. 2007).

GO term	p-value	genes (of 629 total)
response to endogenous stimulus	1.49e-07	85
response to abiotic stimulus	3.76e-05	87
response to stress	1.85e-05	133
hormone-mediated signalling pathway	8.69e-04	46
innate immune response	2.35e-03	24
reproductive system development	1.79e-02	52
shoot system development	1.52e-02	43

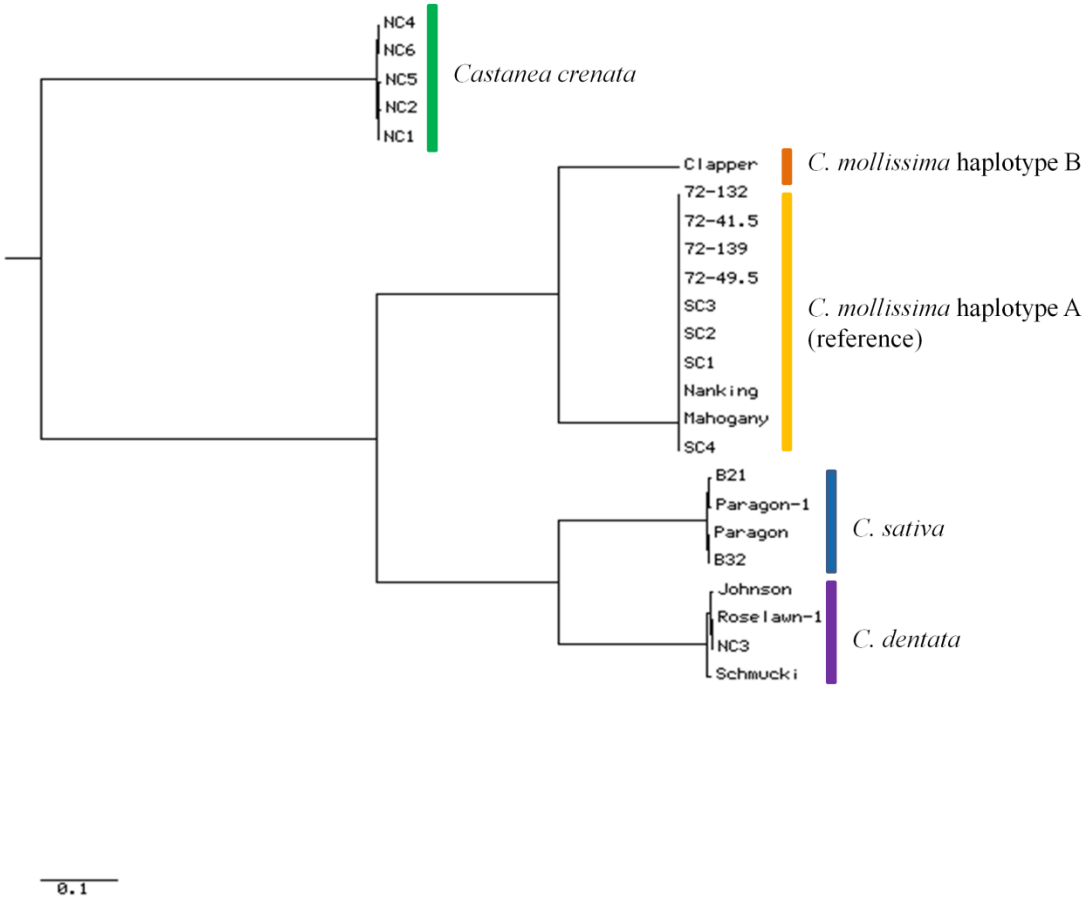

Figure 1. Maximum-likelihood tree constructed using SNP polymorphisms from assembled chloroplast genomes of 24 chestnut samples, showing two distinct chloroplast haplotypes of *Castanea mollissima* one *Castanea dentata* haplotype, a *Castanea sativa* haplotype from "Paragon" in its offspring, and a *Castanea crenata* haplotype in Korean-derived *C. mollissima* material.

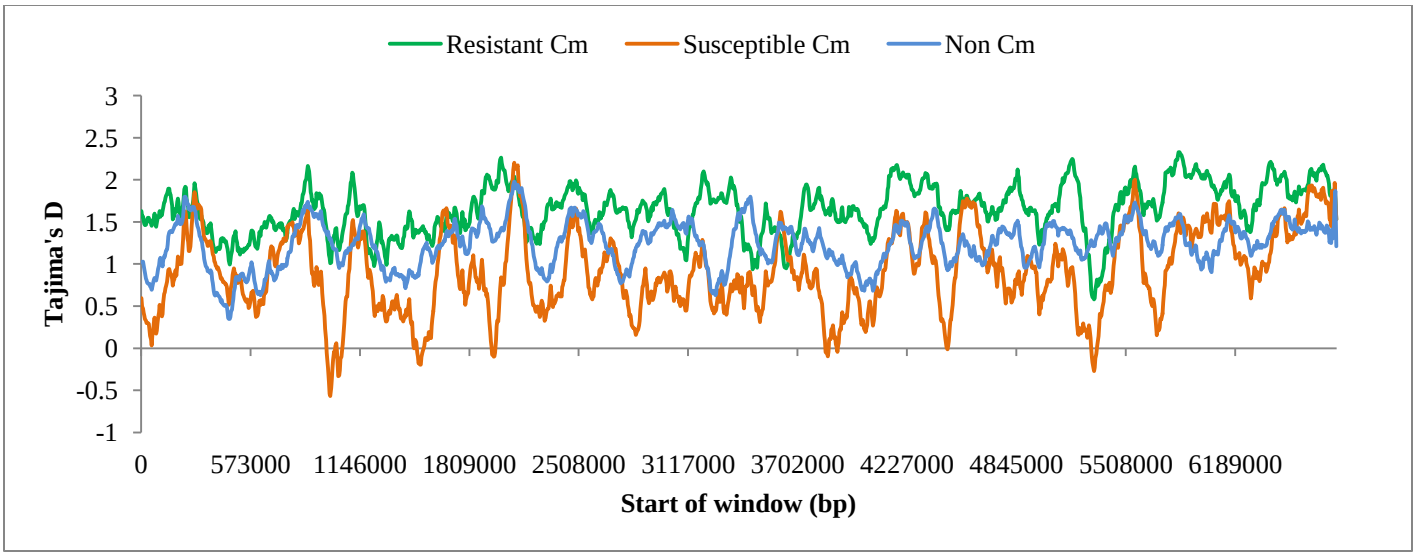

Figure 2. Plot of Tajima's D statistic for cbr1, averaged over 90,000 bp windows with 3,000 bp steps.

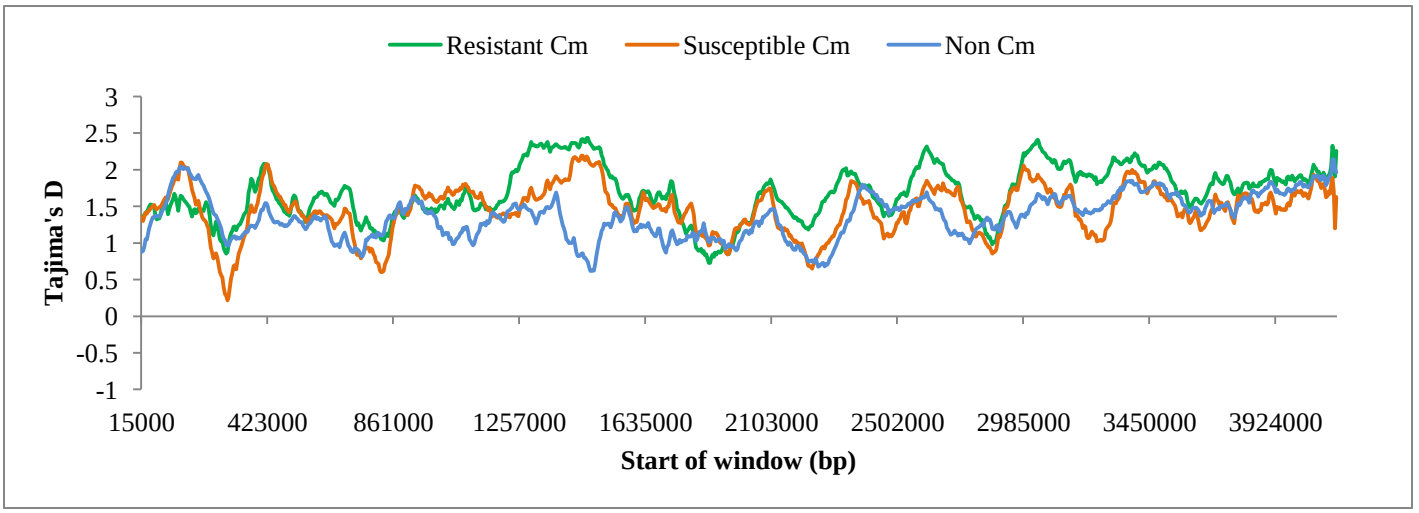

Figure 3. Plot of Tajima's D statistic for cbr2, averaged over 90,000 bp windows with 3,000 bp steps.

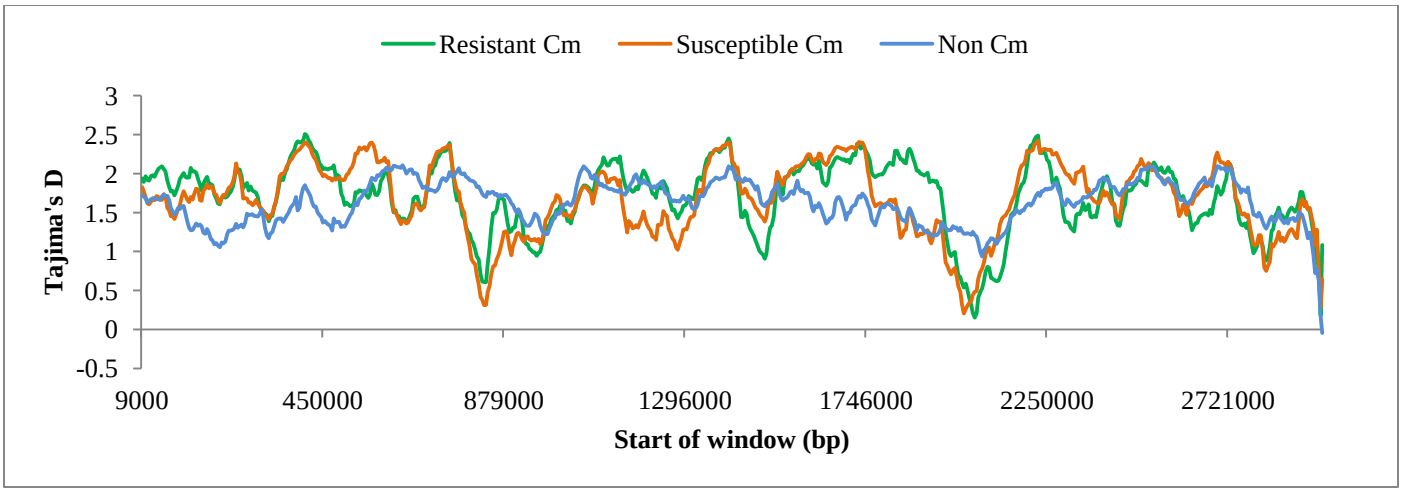

Figure 4. Plot of Tajima's D statistic for cbr3, averaged over 90,000 bp windows with 3,000 bp steps.