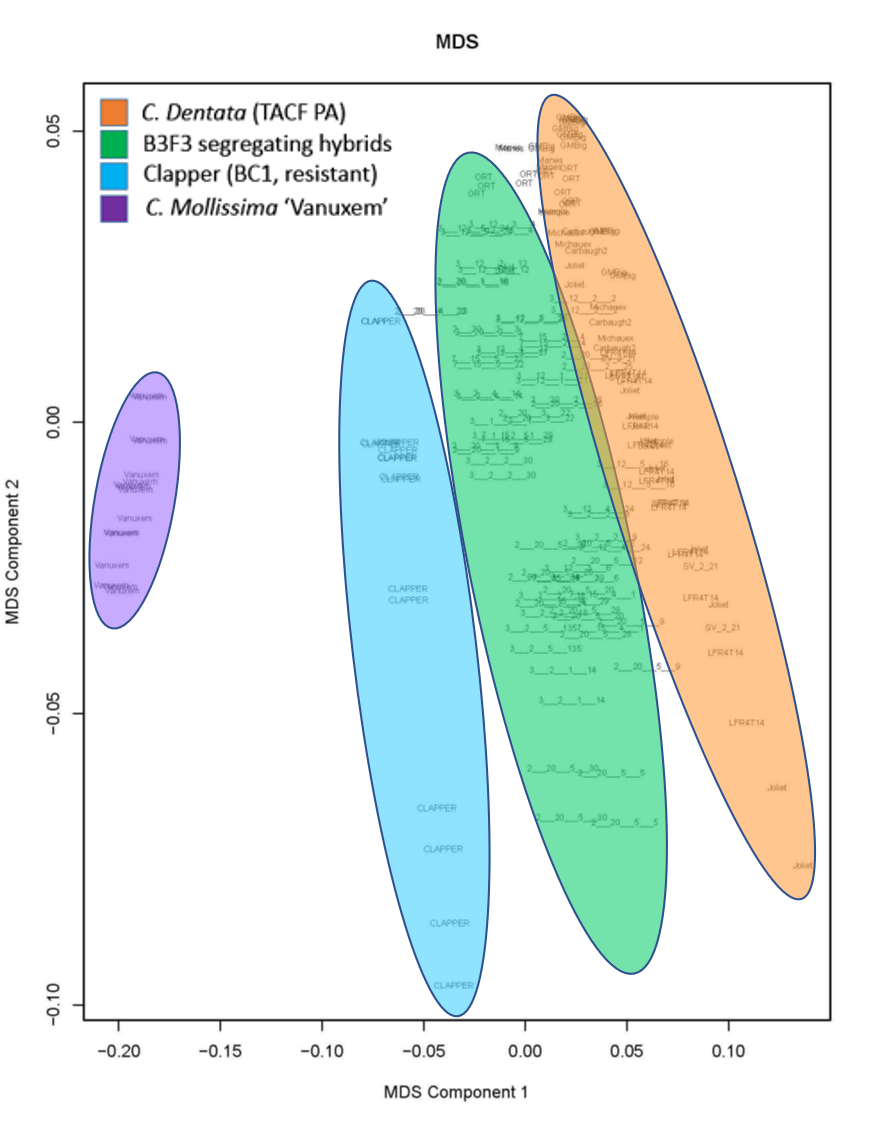
**Results**

Assessment of progress of the back-cross breeding program on recovery of American chestnut genome in the region of the cbr1 blight resistance QTL

DNA sequence data from single-enzyme RADseq libraries was obtained for 48 BC3F3 progeny from a reciprocal intercross of two BC3F2 trees (GR210 and CL287) from the Clapper BC1 pedigree and *C. dentata* parent trees ‘Ort’ and ‘Joliet.’ RADseq data was also obtained from six wild American chestnut trees, the *C. dentata* susceptible great-grandparent tree LFR4T14, the *C.* *mollissima* genomereference genotype ‘Vanuxem’ and a clone of the ‘Clapper’ BC1 tree which was the source of resistance for The American Chestnut Foundation back-cross breeding program. The *C. mollissima* ‘Vanuxem’ reference genome was used for whole-genome mapping of sequence reads and SNP calling. SNP data for a 1.4 Mb scaffold from the cbr1 blight-resistance QTL on linkage group B was used for preliminary analysis of variation among sequenced trees in that genome region.

**Figure XX.** SNP analysis of variation among sequence reads mapped to a 1.4 Mb window of the cbr1 QTL on linkage group B. Variants were predicted from mapping of reads for each individual and for digest replicates used to generate a regional estimate of genomic distance among BC3F3s relative to the Chinese reference sequence (Vanuxem), the blight resistance source (BC1 ‘Clapper’), and parental (Ort, Jolliet) and other wild representatives of *C. dentata*. Multi-dimensional scaling of SNP variants in the samples shows strong resolution between species and hybrid groups in our sample are shown on the x-axis (MDS1). Variation among samples within each group or trees are shown by the y-axis representing MDS2.



A multi-dimensional scaling (MDS) approach was used to visualize components of variation in the mapped sequence data (Figure XX). Multi-dimensional scaling of SNP variants in the sample (MDS component 1 on the x-axis) shows strong resolution between species and hybrid groups. The distribution of variants among the BC3F3 trees sampled places these progeny roughly equidistant between the genome of the original ‘Clapper’ BC1 source of blight resistance and existing *C. dentata* trees, (Figure XX). This indicates that the *cbr1* QTL for blight resistance is intermediate in genome sequence composition between the BC1 source of resistance and the recurrent American chestnut genome. It is also appears that the technical replication of RE digest libraries has revealed a diverse population structure among the BC3F3 generation trees (from the MDS component 2 on the y-axis of figure XX). This diversity may be due to the incomplete and random nature of the process of restriction fragmentation of whole genome DNA for the RADseq approach, as well as the variation among F3 trees in the results of the introgression process itself.

**Discussion:**

Assessment of progress of the back-cross breeding program on recovery of American chestnut genome in the region of the cbr1 blight resistance QTL

Genotyping by sequencing (GBS) is a cost-effective method to identify variants across the genomes of large sample sizes of individuals, for which genome assemblies are available to use as a reference. The RADSeq version of GBS uses the presence of a restriction digest cutting site less than one read length distance from a given SNP variant, and therefore the distribution of discovered variants should be highly sensitive to differences in the genome sequences among individuals. This approach has great power for resolving differences from introgression among progeny resulting from inter-species hybridization, which is often used in plant breeding to transfer disease resistance genes between species. The MDS graphical visualization of the SNP data for one large genome sequence scaffold correctly placed the Clapper BC1 parent tree exactly intermediate in sequence composition between the Chinese chestnut genome reference tree and the several American chestnut trees sequenced. This validated the approach taken to produce the genome sequence data and the MDS approach taken to analyze and present the results. The MDS graphical visualization of the SNP variation also revealed substantial SNP variation in the blight-resistance QTL *cbr1* among the selected BC3F3 progeny and placed their haplotype sequence in the *cbr1* region to be intermediate between those of the original Clapper BC1 parent and *C. dentata* parents ‘Ort’ and ‘Joliet’ and wild American chestnut trees. Thus the backcross breeding approach is moving the composition of the genome of backcross individuals towards the American chestnut genome as desired. However, the breeding program evidently still requires either more intensive selection using whole-genome based selection strategies and/or more rounds of back-crosses to American chestnut parent trees.

Our sequence variation results are preliminary given the analysis was limited to an app, 1 Mb region of the genome, to 48 BC3F3 trees, and to one sequencing approach. This future studies should incorporate additional backcross trees and/or different library construction methods to control for the technical variance component, and/or be conducted either at the computationally more challenging whole-genome level to identify segments in the genome of BC3F3s individuals that are outside the resistance blight QTL which are be default most likely to be of *C. mollissima* origin. It is anticipated that a more extensive genome-wide analysis of SNP variation in blight resistant individuals in backcross intercross generations will reveal a more complicated picture of introgression results, but one that is independent of the selection for blight resistant individuals.

**Materials and Methods:**

Assessment of progress of the back-cross breeding program on recovery of American chestnut genome in the region of the cbr1 blight resistance QTL

DNA was extracted using a modified CTAB protocol (74) from twig and/or leaf tissues from the following individuals:a clone of the BC1 ‘Clapper’ which was the source of resistance for The American Chestnut Foundation back-cross breeding program; 48 BC3F3 progeny from a reciprocal intercross of two resistant trees (GR210 and CL287) in the BC3F2 generation; *C. dentata* parent trees ‘Ort’ and ‘Joliet’; six wild American chestnut trees (Carbaugh, Hemple, Manes, Michaux, Stone Valley, and Glade Mountain Big); *C. dentata* susceptible great-grandparent LFR4T14; and the *C.* *mollissima* reference individual ‘Vanuxem’. In 2014**,** the Cornell University Biotechnology Resource Center**,** constructed RADSeq libraries foreachsample using single restriction enzyme (PstI) digests, and then produced single-end 100bp sequences from the restriction fragment libraries using an Illumina HiSeq 2500. Sequencing was performed in multiplex pools that included 6-8 replicates of restriction fragment libraries for parental trees and the reference genotype, and single aliquots for all other libraries.

The *C. mollissima* ‘Vanuxem’ reference genome was used for whole-genome mapping of sequence reads and SNP calling (with bioinformatics support from the Genomic Diversity Facility at Cornell University). Preliminary analysis of SNP variants focused on mapping results for the longest sequence in the v3.2 de novo assembly - a 1.4 Mb scaffold in the blight resistance QTL region on linkage group B (cbr1) from the Vanuxem reference genome. Multi-Dimensional Scaling (MDS) was used to graphically display the SNP variation results among samples.