Final Report to the American Chestnut Foundation Research Program

PROJECT TITLE: "Assembly of the Mahogany genome using a new chromatin proximity approach."

PRINCIPAL INVESTIGATORS AND INSTITUTION:

John E. Carlson, Professor
Department of Ecosystem Science and Management
The Pennsylvania State University, University Park, PA. 16802

DATE: March 13, 2018

SUMMARY:

In this project we developed the first genome sequence for the important Chinese chestnut cultivar 'Mahogany.' This was accomplished, with the assistance of the Dovetail Genomics Company, using a new approach based on ordering genome fragment sequences into chromosome-scale sequences using genome-wide chromatin interaction data (aka 'Hi-C'). The resulting reference genome sequence consists of 12 sets of continuous sequences placed in their correct order for each linkage groups by comparison to C. mollissima genetic linkage maps. These 12 'pseudochromosomes' represent a reference of the ordered sequences for the Mahogany genome that can be used in TACF breeding programs for Genome-Wide-Selection.

RESULTS:

Objective 1. Conduct *de novo* assembly of existing short-read genome sequence data

To start the project, we delivered to the Dovetail Genomics company ("Dovetail") 164,360,173 Illumina sequence reads averaging 250 bp in length (i.e. app. 41Gbases of high quality data). An initial de novo assembly using this data was not satisfactory for Dovetail as their starting point. Thus, they provided 'matching' support ($11,750 in-kind contribution) to prepare a better de novo assembly. We provided additional Mahogany tissue (catkins) from which Dovetail staff prepared high-molecular DNA and new Illumina libraries from 510bp genomic DNA fragments, from which they generated 151.7 Gb of paired-end reads.

Dovetail assembled the 151.7 Gb of paired-end reads with our 41Gbases of single-end reads to obtain 152,840 contigs (N50 10.9 Kb) which further assembled into 51,675 Scaffolds (N50 19.6 Kb) totaling 507.1 Mbases. The genome assembly software predicted a genome size of 734 Mb. So, this assembly accounted for only 69% of the genome. A heterozygosity level of 39% was observed within the sequence data obtained. A quality check on the assembly was conducted using the BUSCO analysis, which queries for the presence of 303 known conserved eukaryotic single-copy genes. The BUSCO analysis identified 89% of the genes in the 51,675 scaffolds, of which 8% were duplicated. For a summary of the results, see table 1 below.

Objective 2. Produce "Chicago" chromatin interaction sequence data and conduct a new genome assembly.

To start this phase of the project, we collected fresh leaves from the Mahogany tree at the CAES' Graves nursery. A lower branch of the tree was covered with a thick tarp to keep the leaves in the complete dark for app. 36 hours. The youngest dozen etiolated leaves were

collected by JEC and Jack Swatt, wearing new lab gloves to prevent contamination. After being clipped from the branch, each leaf was immediately placed in a clean ziplock bag and flash-frozen in liquid nitrogen in a dewar, and then transferred directly into a 'charged' nitrogen dry shipper container designed to keep tissues frozen at liquid nitrogen temperature for up to 10 days. The frozen leaf tissues were shipped by overnight courier to the Dovetail Genomics company for isolation of very high molecular weight DNA from nuclei as chromatin. For the "Chicago" approach to generating chromatin-interaction data, the HMW DNA was cross-fixed to the chromatin proteins *in vitro*, i.e. in solution. Then the cross-linked (fixed) ends of DNA were purified from the chromatin, from which Illumina sequencing libraries were constructed and the paired ends sequenced, revealing sequences belonging to the same chromosome but at different sites on the chromosome.

App. 193M pairs of 150 bp sequences were produced from the Chicago library. The paired-end chromatin-interaction sequence data was then co-assembled with the de novo assembly using Dovetail's proprietary HiRise software. The chromatin interaction data provides long-range sequence position information that permits de novo contigs and scaffolds to be placed in the correct orders within chromosome-length assemblies. The result of the Chicago library and de novo co-assembly was a reduction in the number of scaffolds from 51K to 3,848, with a mean length of 2.1Mbases. This resulted in a **511.18 Mb assembly** representing 69% of the predicted total genome length of 734 Mb (85% of the non-repetitive DNA).

Objective 3. Produce "Hi-C" DNA sequence and assemble the "Chicago" scaffolds into chromosome-scale sequences.

From another subset of the frozen leaf tissues that we collected from the Mahogany tree, the Dovetail staff isolated nuclei, but conducted the cross-linking of DNA to chromatin proteins *in vivo*, i.e. within the intact nuclei, prior to DNA isolation. This is the "Hi-C" approach to generating chromatin-interaction data, which provides paired-end fragments of DNA on the same chromosome that resided physically close to each other within the tightly-packed interphase nucleus. The cross-linked DNA ends were purified from the nuclei and Illumina sequencing libraries constructed. App. 247M pairs of 150 bp sequences were produced from the Hi-C library. The paired-end chromatin-interaction sequence data was co-assembled with the Chicago-based assembly using the HiRise software, which further reduced the number of genome sequence pieces from 3,848 to 3,180 scaffolds. The average length of the Hi-C scaffolds was 43.Mb, and overall **the assembly covered 511.84 Mb** (70% of total predicted genome length; 85% of non-repetitive DNA).

A quality check on this final version of the assembly was conducted by BUSCO analysis. Of the 303 eukaryotic genes queried, 88% of the expected single-copy genes were identified in the 3,180 scaffolds, of which 8% were duplicated, 5% were fragmented, and 12% were missing.

Objective 4. The HiRise genome assembly will be validated by comparison to genetic linkage maps and the Vanuxem reference genome.

Dr. Tatyana Zhebentyayeva conducted several analyses to determine the quality and extent of the Hi-C assembly of the Mahogany genome. She found that most of the Kubisiak (2013) reference genetic linkage map for C. mollissima was covered by the 12 largest scaffolds. These 12 scaffolds collectively contain 506 Mbases, which is 99% of the total Hi-C genome assembly. Table 2, below, prepared by Dr. Tatyana Zhebentyayeva, shows the lengths of each of the 12 scaffolds and which Linkage Group each scaffold corresponds to. Table 2 also shows

how much of the Vanuxem genotype *de novo* genome assembly that we have been able to anchor to the same reference genetic linkage groups through tedious identification of genetic marker sequences from the linkage map in the scaffold sequences. These results confirm that 12 largest scaffolds from the Mahogany Hi-C assembly can indeed be considered 'pseudo-chromosome' sequences. we have , in ongoing research by our genomics group (Drs. Zhebentyayeva, Nelson, Abbott, Staton, and Carlson labs) to build 'pseudo-chromosome' sequences for Chinese chestnut cv Vanuxem.

Dr. Zhebentyayeva also confirmed that the 12 Mahogany chromosome-length sequences include essentially all of the app 1200 expressed gene based markers on the reference genetic map. This indicates that the pseudo-chromosome assembly probably covers most of the gene space in Chinese chestnut, but that the repetitive DNA regions of the genome are still mostly missing from the 12 chromosome-scale sequences. Tatyana did also observe, however, that the orders of neighboring genes was flipped within the 12 Mahogany pseudo-chromosomes in comparison to the orders of corresponding gene markers on the high density genetic map, in quite a few places. Further studies will be required to determine which orders of gene pairs are correct, the map or the Hi-C genome assembly.

Another means of validating the genome assembly is by comparison to highly complete model genomes. Our model for rosid trees is the relatively small and simple peach genome. Dr. Staton's group conducted an alignment of the 12 Mahogany chromosome-length scaffold sequences, using the BLAST algorithm, to the 8 peach chromosomes. Figure 1 below shows a visualization of the result of the alignments, both collectively for the entire genomes (panel A) and for the Mahogany genome to each peach chromosome. These 'Circos' plots clearly demonstrate that most of the peach genome is represented by regions of homology in the Mahogany genome, and vice versa. Furthermore, the plots indicate that the overall structures of the peach and Chinese chestnut genomes are similar, with gross differences that can be accounted for by relatively few rearrangements or reshuffling of larger segments of the peach chromosomes.

Objective 5. Deliver the validated Mahogany genome assembly to TACF and collaborators, along with locations of genes.

The entire Mahogany genome assembly, including the results for the *de novo*, Chicago, and Hi-C steps, have been copied to the Staton lab server for further analyses, including gene finding and gene model annotations. In addition, the 12 Mahogany Hi-C pseudo-chromosome scaffold sequences have recently been provided to Jeremy Schmutz's JGI group at the HudsonAlpha Institute. The JGI staff will conduct comparisons of our Mahogany genome assembly with the more complete American chestnut genome assembly to help determine which segments from the Mahogany genome may be missing, and if there are any major structural differences between the genomes, and perhaps how well the American chestnut gene annotations might be carried over to Chinese chestnut.

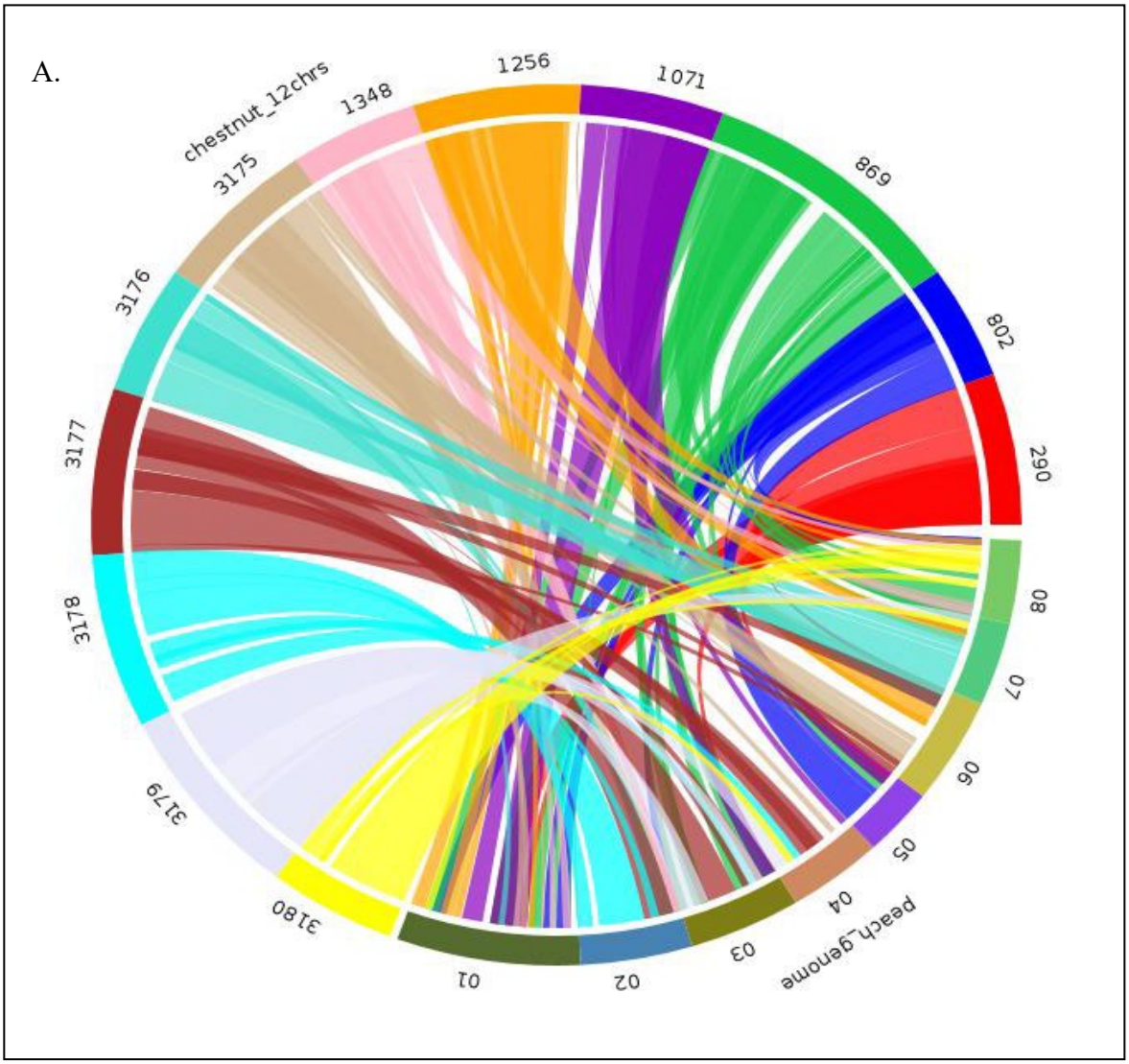
Table 1. Mahogany Version 1 Chromosome Sequencing and Assembly Statistics

METRIC	Illumina Sequencing and *De Novo* Assembly	Assembly with "Chicago" chromatin data	Assembly with "HiC" chromatin data
Assembly Stats:	• 164,360,173 reads at 250 bp • 535,528,520 pe reads at 130 bp • best fit: 55 k-mer size, 39% heterozygosity, 734Mb genome size • 152,840 contigs (N50 10.9 Kb) • 51,675 scaffolds (N50 19.6 Kb) • **507.1 Mb assembled** (69% total)	• 193M read pairs; 2x150 bp • contigs N50 11.01 Kb • 3,848 scaffolds • L50/N50: 67 scaffolds; 2.1Mb • L90/N90: 295 scaffolds; 401Kb • **511.18 Mb assembled** (69% total; 85% of non-repeat DNA)	• 3,180 scaffolds total • L50/N50: 5 scaffolds; 43.Mb • L90/N90: 11 scaffolds; 32.8 Mb • Longest scaffold 68,930,353 bp • **511.84 Mb assembled** (70% total; 85% of non-repetitive DNA) • 149,065 gaps, covering 3.46%
Gene models:	Of 303 expected single copy genes: • 195 found single copy (64%) • 25 duplicated (8%); • 38 fragmented (8%); 45 missed (19%) • Total # genes not determined	Of 303 single copy genes: • 226 found single copy (75%) • 26 duplicated (8.5%); • 16 fragmented (5%); 35 missed (12%) • Total # genes not determined	Of 303 single copy genes: • 227 found single copy (75%) • 25 duplicated (8%); • 14 fragmented (5%); 37 missed (12%) • Total # genes not determined
Pseudo-chromosomes	• not attempted • L50/N50 = 7,299 scaffolds; 20 Kb • L90/N90 = 28,550 scaffolds; 4 Kb	• not attempted	• 12 chromosome-length sequences; covering complete reference genetic linkage map, but missing repeat DNA

Table 2. assignments of Mahogany and Vanuxem genome pseudochromosome sequence assemblies with the Chinese chestnut Genetic Linkage Groups.

Linkage Groups	Mahogany Scaffold number	Mahogany Scaffold Length (nt)	Vanuxem Pseudo-chromosome lengths (nt)
LG_A	869	68,930,353	64,357,929
LG_B	1256	41,097,837	43,210,214
LG_C	290	40,207,821	37,471,317
LG_D	1071	37,205,925	34,610,428
LG_E	3179	53,013,531	42,797,072
LG_F	3177	46,124,846	40,838,100
LG_G	3180	32,837,157	30,178,183
LG_H	3175	42,962,653	34,186,704
LG_I	3176	33,700,751	32,344,349
LG_J	802	31,017,659	30,132,902
LG_K	3178	45,236,068	33,874,491
LG_L	1348	33,629,445	28,317,178
	total=	505,964,046	452,318,867

Figure 1. Mahogany genome structural comparisons with the peach genome.
Panel A – Circos plot aligning all peach (Pp) and Mahogany chromosome sequences



Panel B - Circos plot aligning each of the 8 individual peach chromosomes with all 12 of the Mahogany chromosome sequences (scaffolds).