AN IMPROVED CHINESE CHESTNUT GENOME

John E. Carlson[1], Margaret E. Staton[2], Charles Addo-Quaye[1,3], Nathaniel Cannon[1], Tetyana Zhebentyayeva[4], Nurul Islam-Faridi[5], Jiali Yu[2], Matthew Huff[2], Shenghua Fan[6], Anna O. Conrad[6], Stephan C. Schuster[1,7], Albert G. Abbott[6], Jared Westbrook[8], Jason Holliday[9], C. Dana Nelson[10], Laura Georgi[11], and Frederick V. Hebard[11]

SUMMARY

The introduction of the chestnut blight fungus (*Cryphonectria parasitica*) to North America in 1904 devastated American chestnut populations. Asian Chestnut species, which evolved resistance to the sympatric chestnut blight fungus, are being used as donor species for the transfer of resistance genes to *C. dentata* and *C. sativa* via hybridization. In the United States, Chinese chestnut (*Castanea mollissima*) genotypes are the source of blight-resistance genes for introgression into American chestnut by backcross breeding. To better understand the genetic basis of blight resistance and to provide tools for chestnut breeding, we sequenced the genome of Chinese chestnut, with a particular focus on the major blight resistance quantitative trait locus (QTL). A draft genome (v.1.1) covering app. 90 percent of the genome of The American Chestnut Foundation's Chinese chestnut cultivar "Vanuxem" was released to the public in January 2014. Recombinant DNA clones covering the three major blight resistance QTL were also sequenced to great depth. Over 780 genes were identified in the 3 blight resistance QTLs, including 15 known "defense response" genes. We are developing a new version of the Chinese chestnut genome with chromosome-scale assemblies of genome scaffolds anchored to the 12 chestnut linkage groups. This will serve as a reference for genome-wide selection in advanced generations of backcross breeding programs, and for basic research on genome structure and function in woody plants.

GOALS

The overarching goals of the public Chinese chestnut genome project were to:

(1) Construct a complete genome sequence for *Castanea mollissima*

(2) Identify candidate genes for *Chryphonectria parasitica* (chestnut blight) resistance.

(3) Provide tools to accelerate breeding and restoration of *C. dentata* (American chestnut)

PROGRESS

Version 1 of the Chinese Chestnut Genome

The first version of the genome assembly was produced for the TACF cultivar Vanuxem (fig. 1). Derived from 60 Gb of 454 and Illumina NGS sequence data, the assembly comprised 724.4 Mb in 41,270 scaffolds which averaged app. 40,000 bp in length (table 1). This represented 91 percent of the predicted size of the Chinese chestnut genome, based on estimates from previous

[1] Schatz Center for Tree Molecular Genetics, Pennsylvania State University, University Park, PA 16802, (jec16@psu.edu).
[2] University of Tennessee, Knoxville, TN 37996.
[3] Division of Natural Sciences and Mathematics, Lewis-Clark State College, Lewiston, ID 83501.
[4] Clemson University, Clemson, SC 29634.
[5] USDA Forest Service, Saucier, MS 39574.
[6] Forest Health Research and Education Center, University of Kentucky, Lexington, KY 40506.
[7] Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technology University, Singapore 639798.
[8] The American Chestnut Foundation, Meadowview, VA 24361.
[9] Virginia Polytechnic and State University, Blacksburg, VA 24061.
[10] Forest Health Research and Education Center, Southern Research Station, USDA Forest Service, Lexington, KY 40546.
[11] Retired from The American Chestnut Foundation, Meadowview, VA 24361.

Citation for proceedings: Nelson, C. Dana; Koch, Jennifer L.; Sniezko, Richard A., eds. 2020. Proceedings of the Sixth International Workshop on the Genetics of Host-Parasite Interactions in Forestry—Tree Resistance to Insects and Diseases: Putting Promise into Practice. e-Gen. Tech. Rep. SRS–252. Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station. 170 p.

Figure 1—Fred Hebard and Laura Georgi with a clone of the cultivar "Vanuxem" at Meadowview, VA.

Table 1—Sequencing and assembly statistics for the first genome assembly, v1.1

METRIC	Assembly 1.1, 2014 (available at hardwoodgenomics.org)
genomic DNA sequence	**60 Billion nucleotides** of (75X depth) • 454 sequence of 500 bp DNA fragments • Illumina paired-end sequence (2Kb & 8Kb)
Assembly of contigs:	• 323,611 contigs (2 to 1.87 Kb) • **41,270 scaffolds** • 724 Mbases assembled • 90.5% genome coverage • Scaffolds 2K to 429K bases
Gene models:	• **36,478 genes** predicted
Resistance QTL	• Three blight-resistance QTL sequenced, assembled, and genes identified
Physical Map Coverage:	• 92% of BAC sequences mapped to scaffolds • Ns (gaps) in scaffolds = 2 %

cytometry studies. A total of 36,146 likely genes were identified. The quality and contiguity of the version 1.1 genome for Chinese chestnut proved suitable for genome-wide studies of genetic diversity, gene expression, DNA marker development, and evolutionary studies. LaBonte et al. (2018) used the version 1.1 genome to detect signatures of selection for genes related to flower phenology and development, fruit maturation, and secondary metabolism, and genes potentially impacted by domestication in the Chinese Chestnut genome. We conducted initial genome-wide comparative genome structure and evolution analyses, which revealed the levels of similarities with genes from other plants, and the extent to which large blocks of syntenies are shared between chestnut and model genome such as peach, grape, and poplar (fig. 2).

Version v1.1 of the Chinese chestnut genome was released to the public as a genome browser in January 2014 at the website https://hardwoodgenomics.org/chinese-chestnut-genome, developed and curated by Margaret Staton, along with browsers for the three chestnut-blight resistance QTL. For the blight resistance QTL, sets of overlapping BAC recombinant DNA clones that physically spanned each QTL region were separately sequenced and assembled into a total of 395 scaffolds covering 13.8 Mb (table 2). Over 1,900 genes were found in the QTLs, including 194 known stress-response genes, from which 15 candidate genes for blight resistance were selected for further study (Fang et al. 2013). Annotations of the QTL gene sequences were used to conduct GO functional category analyses. From

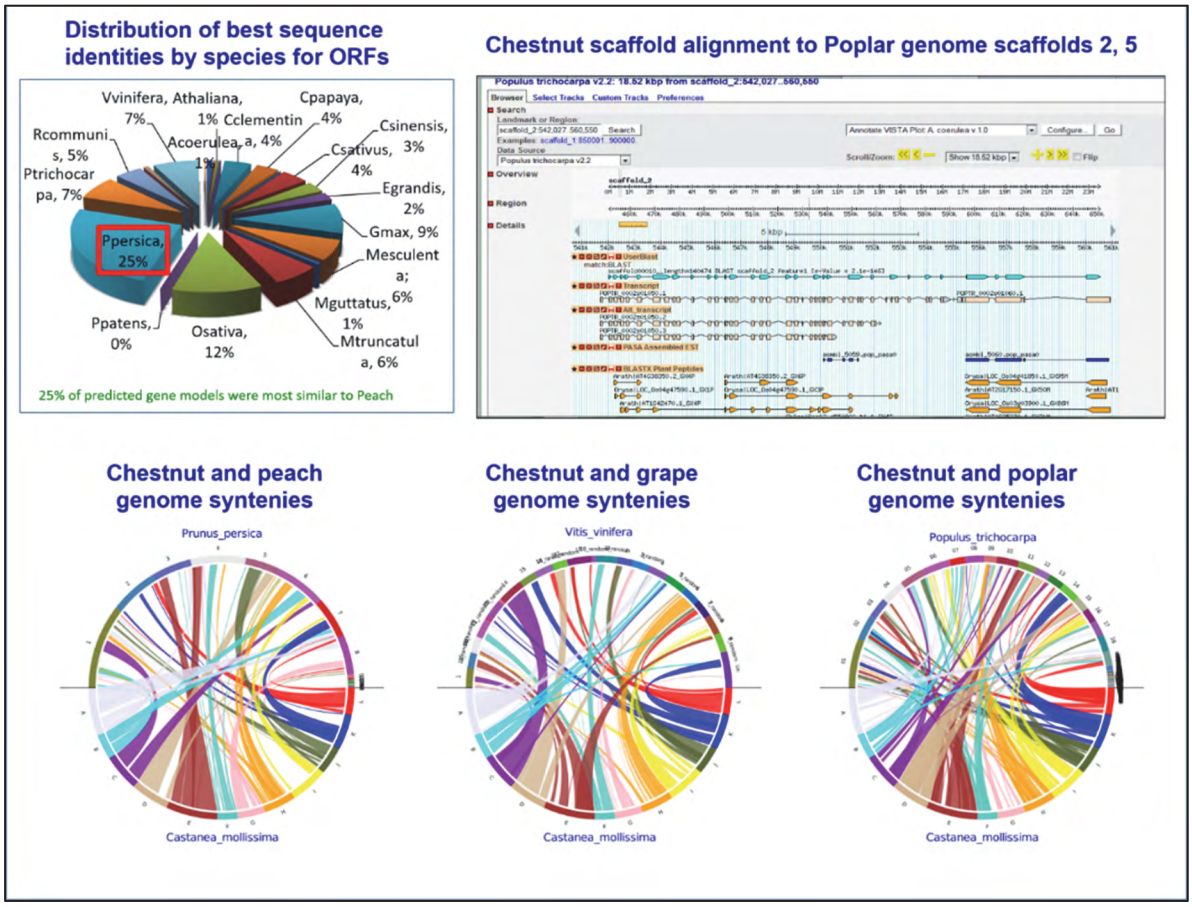

Figure 2—Examples of synteny analyses to model plant genomes and plant gene sequence databases.

Table 2—Chestnut blight-resistance QTL assembly and gene content information

QTL	LG	# scaffolds	# bases / QTL	Total # Genes	# Genes / Mbases	# Stress-response genes
cbr1	**B**	214	6.77Mb	994	151	98
cbr2	**F**	128	4.12Mb	548	137	58
cbr3	**G**	53	2.99Mb	410	139	38

the genes classified as disease- or stress-response, 15 of the most likely candidate genes for chestnut blight resistance were identified (table 3) for future transgenic studies. The genome and the QTL browsers have had thousands of visits from across the globe, for gene searches and downloads of genome and QTL scaffolds, genes, transcripts, and predicted protein data.

Improvements to the Chinese Chestnut Genome

For application of the chestnut reference genome in Genome-Wide-Selection to advance back-cross breeding and disease resistance introgression programs, such as the TACF is conducting, chromosome-scale sequences assemblies of scaffolds are required. For the past 4 years,

Table 3—Fifteen candidate disease resistance genes in Chinese chestnut blight-resistance QTL selected for future functional studies

Seq. Name	Seq. Description	Closest matching NCBI nr protein
cbr1_scaffold114-gene-0.3-mRNA-1	transcription factor tga1	Transcription factor TGA1 (Vitis vinifera)
cbr1_scaffold134-gene-0.0-mRNA-1	cc-nbs-lrr resistance protein	Putative disease resistance protein RGA3 (Vitis vinifera)
cbr1_scaffold16-gene-0.12-mRNA-1	rna recognition motif-containing protein	PREDICTED: DAZ-associated protein 1-like (Vitis vinifera)
cbr1_scaffold17-gene-0.29-mRNA-1	beta-hydroxyacyl-acp dehydratase	predicted protein (Populus trichocarpa)
cbr1_scaffold28-gene-0.12-mRNA-1	transcription factor tga1	TGA transcription factor 1 (Populus tremula x Populus alba)
cbr1_scaffold32-gene-0.28-mRNA-1	14-3-3-like protein gf14 lambda	hypothetical protein ARALYDRAFT_496774 [Arabidopsis lyrata subsp. lyrata)
cbr1_scaffold4-gene-0.38-mRNA-1	multicatalytic endopeptidase complex	proteasome subunit alpha type-7 (Vitis vinifera)
cbr1_scaffold61-gene-0.11-mRNA-1	disease resistance protein at4g27190-like	PREDICTED: disease resistance protein At4g27190-like (Vitis vinifera)
cbr2_scaffold29-gene-0.6-mRNA-1	cc-nbs-lrr resistance protein	cc-nbs-lrr resistance protein (Populus trichocarpa)
cbr2_scaffold34-gene-0.3-mRNA-1	feronia receptor-like kinase	Serine/threonine-protein kinase PBS1, putative (Ricinus communis)
cbr2_scaffold3-gene-0.42-mRNA-1	Protein	PREDICTED: MLO protein homolog 1-like (Glycine max)
cbr2_scaffold5-gene-0.9-mRNA-1	transferring glycosyl	transferase, transferring glycosyl groups, putative (Ricinus communis)
cbr3_scaffold1-gene-1.1-mRNA-1	histone-lysine n-methyltransferase ashh2-like	PREDICTED: uncharacterized protein LOC100245350 (Vitis vinifera)
cbr3_scaffold1-gene-1.19-mRNA-1	set domain protein	PREDICTED: uncharacterized protein LOC100245350 (Vitis vinifera)
cbr3_scaffold28-gene-0.8-mRNA-1	cysteine proteinase rd19a	Cysteine proteinase RD19a (Arabidopsis thaliana)

we worked on producing such an improved, chromosome-scale version of the Chinese chestnut genome. The approach taken to accomplish this involved merging assembled sequence contigs and scaffolds from the v1.1 genome using longer genome sequences, followed by the anchoring of the larger scaffolds to positions on genetic linkage maps based on alignment to DNA marker sequences defining loci on the genetic maps (fig. 3). Taking this approach, we used long PACBio sequences (over 10 Kb in length) to bridge contigs into scaffolds and close gaps within scaffolds, reducing the number of genome scaffolds to 12,684, covering 784 Mb of genome sequence (table 4). By aligning sequences from BAC clones distributed across the physical length of the genome (Fang et al 2013), we estimated that app. 98 percent of the estimated genome size was included in the new set of scaffolds. Margaret Staton's group used our RNA sequence resources for chestnut to find and annotate 30,832 high quality gene models in the new assembly. Pseudo-chromosome sequences were assembled by anchoring 4,099 of the scaffolds to DNA markers in the reference genetic linkage map for chestnut (Kubisiak et al. 2013) (fig. 3D), and the integrated genetic-physical map for Chinese chestnut (Fang et al., 2013) (fig. 3A). The pseudo-chromosome sequences accounted for 421.3Mb, representing about 60 percent of the full genome.

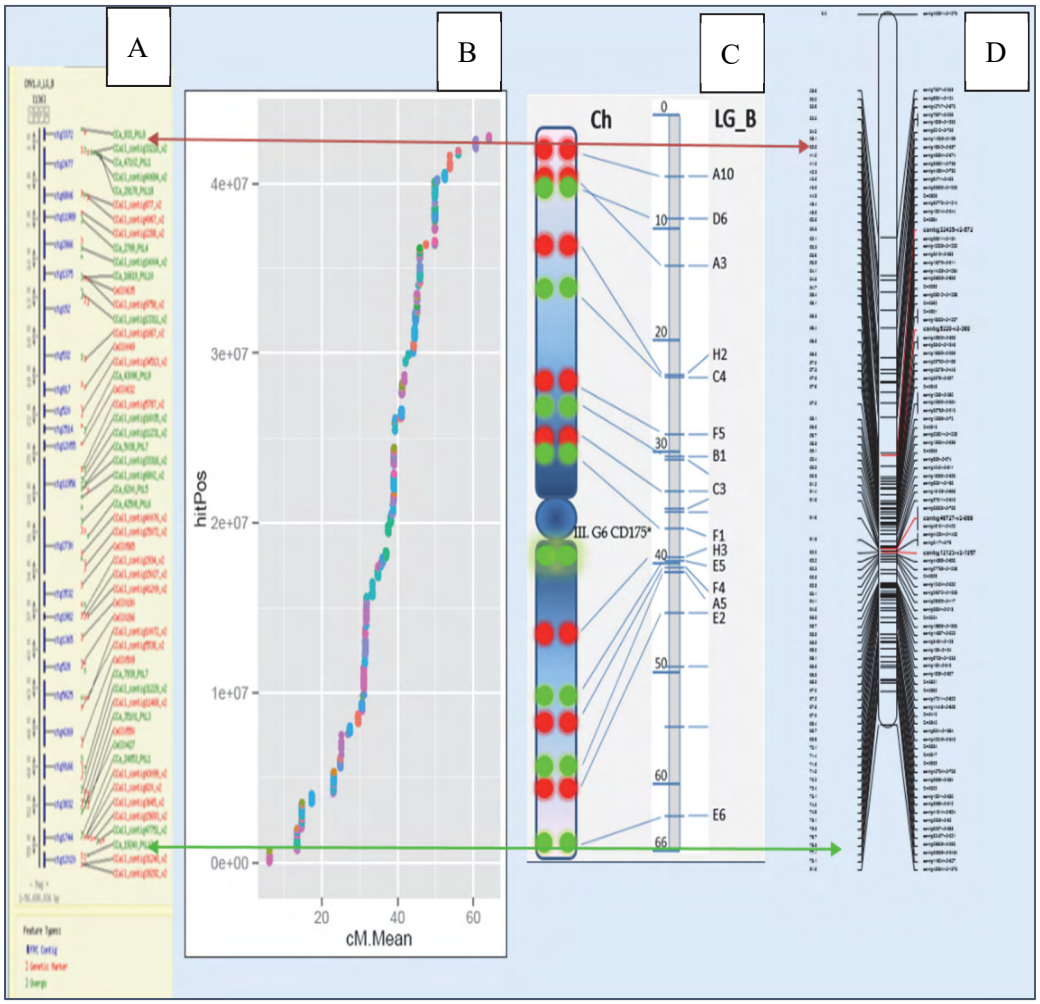

Figure 3—Illustration of approach taken in building pseudo-chromosome sequences by alignment and anchoring to genetic, physical, and cytology maps. (A) Example of physical map integrated with linkage group, (B) a chromosome-scale assembly of scaffolds, (C) cytogenetic map of linkage group DNA markers, and (D) chestnut genetic linkage map alignment example.

Table 4—Statistics for assembly improvements and initial anchoring of scaffolds to genetic maps

METRIC	Chromosome-scale assembly of the Chinese chestnut genome
Contig Assembly	• **421.3Mb (**4,099 contigs) assembly of scaffolds anchored to *C. mollissima* chromosomes (genetic linkage groups) • Anchored sequence per chromosome ranged from 25.9Mb (LG_L) to 60.2 Mb (LG_A) • 303.9Mb (10,011 contigs) unanchored contig sequences • 57.8Mb estimated gaps (based on estimated genome size of 794Mb)
Gene models:	• 30,832 high quality gene models identified in V3.2 • 20,770 high quality gene models in anchored contigs
Validations:	• BUSCO reported 1,355 of 1,440 expected single-copy genes are complete and present within the *C. mollissima* genome • Alignments of BAC end sequences from the *C. mollissima* physical map confirmed order of contigs in the pseudochromosomes

Gene positions and overall genome organization of the assembly were determined and compared to the genomes of other related trees and model plant systems. A manuscript on this chromosome-scale assembly (table 4) was published as a preprint on April 22, 2019 (Staton et al. 2019). Submission of a peer-reviewed manuscript is planned, following efforts to anchor more scaffold to increase gene and genome coverage in the pseudo-chromosomes.

FUTURE RESEARCH DIRECTIONS

The current assembly, although an incomplete draft, does by virtue of chromosome-scale sequences provide a significant advancement in our ability to investigate genome organization and the evolution and genetic structure of important traits such as disease resistance, as well as applications such as genome-wide selection. Future improvement of the assembly may be achieved through the use of more recent long-read technologies, such as Nanopore (Madoui et al. 2015), and/or scaffolding with chromatin-interaction data, such as Hi-C (Jiao and Schneeberger 2017). However even these approaches may result in less than full-genome assembly given the challenges of high heterozygosity levels and an inability to generate dihaploid individuals in Chinese chestnut. The recently published *Quercus robur* genome (Plomion et al. 2018) utilized synteny with the *Prunus persica* genome to order incorporate contigs and scaffolds that had not been assembled *de novo* nor scaffolded with oak genetic map markers. This approach assumes that micro-level syntenies follow known macro-syntenies based on genetic maps, which may not always hold true. However, the hybrid synteny approach could also complement long-read technologies in future Chinese chestnut genome improvements.

Data Availability

The contigs, scaffolds, and pseudochromosome sequences are available at the NCBI BioProject No. PRJNA46687, and are also available for download and query at the Hardwood Genomics Project website (https://www.hardwoodgenomics.org/genomes).

FUNDING

Funding was provided by the Forest Health Initiative through grant # 137RFP#2008-011 to JEC. Support was also provided by the USDA National Institute of Food and Agriculture grant 2016-67013-24581 to The American Chestnut Foundation, grants-in-aid to JEC from The American Chestnut Foundation, and to JEC through the USDA National Institute of Food and Agriculture Federal Appropriations under Project PEN04532, Accession number 1000326. Bioinformatics was supported by NSF Award #1444573 to MS (Main, PI).

REFERENCES

Fang, G.C.; Blackmon, B.P.; Staton, M.E. [et al.]. 2013. A physical map of the Chinese chestnut (*Castanea mollissima*) genome and its integration with the genetic map. Tree Genetics & Genomes. 9(2): 525–537.

Jiao, W.B.; Schneeberger, K. 2017. The impact of third generation genomic technologies on plant genome assembly. Current Opinion in Plant Biology. 36: 64–70.

Kubisiak, T.L.; Nelson, C.D.; Staton, M.E. [et al.]. 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(2): 557–571.

LaBonte, N.R.; Zhao, P.; Woeste, K. 2018. Signatures of selection in the genomes of Chinese chestnut (Castanea mollissima Blume): the roots of nut tree domestication. Frontiers in Plant Science. 9: 810.

Madoui, M.A.; Engelen, S.; Cruaud, C. [et al.]. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics. 16(1): 327.

Plomion, C.; Aury, J.M.; Amselem, J. [et al.]. 2018. Oak genome reveals facets of long lifespan. Nature Plants. 4(7): 440–452.

Staton, M.; Addo-Quaye, C.; Cannon, N. [et al.]. 2019. The Chinese chestnut genome: a reference for species restoration. BioRxiv. 615047.