Tree Genetics & Genomes (2020) 16: 57
https://doi.org/10.1007/s11295-020-01454-y

ORIGINAL ARTICLE

"A reference genome assembly and adaptive trait analysis of *Castanea mollissima* 'Vanuxem,' a source of resistance to chestnut blight in restoration breeding"

Margaret Staton[1] · Charles Addo-Quaye[2,3] · Nathaniel Cannon[2,4] · Jiali Yu[5] · Tetyana Zhebentyayeva[2] · Matthew Huff[1] · Nurul Islam-Faridi[6,7] · Shenghua Fan[8] · Laura L. Georgi[8,9] · C. Dana Nelson[8,10] · Emily Bellis[11] · Sara Fitzsimmons[9] · Nathan Henry[1] · Daniela Drautz-Moses[12] · Rooksana E. Noorai[13] · Stephen Ficklin[14] · Christopher Saski[15] · Mihir Mandal[16,17] · Tyler K. Wagner[2] · Nicole Zembower[2] · Catherine Bodénès[18] · Jason Holliday[16] · Jared Westbrook[19] · Jesse Lasky[11] · Frederick V. Hebard[9] · Stephan C. Schuster[12] · Albert G. Abbott[2,8] · John E. Carlson[2]

Received: 16 January 2020 / Revised: 29 June 2020 / Accepted: 10 July 2020 / Published online: 23 July 2020

Abstract

Forest tree species are increasingly subject to severe mortalities from exotic pests, pathogens, and invasive organisms, accelerated by climate change. Such forest health issues are threatening multiple species and ecosystem sustainability globally. One of the most extreme examples of forest ecosystem disruption is the extirpation of the American chestnut (*Castanea dentata*) caused by the introduction of chestnut blight and root rot pathogens from Asia. Asian species of chestnut are being employed as donors of disease resistance genes to restore native chestnut species in North America and Europe. To aid in the restoration of threatened chestnut species, we present the assembly of a reference genome for Chinese chestnut (*C. mollissima*) "Vanuxem," one of the donors of disease resistance for American chestnut restoration. From the de novo assembly of the complete genome (725.2 Mb in 14,110 contigs), over half of the sequences have been anchored to the 12 genetic linkage groups. The anchoring is validated by genetic maps and in situ hybridization to chromosomes. We demonstrate the value of the genome as a platform for research and species restoration, including signatures of selection differentiating American chestnut from Chinese chestnut to identify important candidate genes for disease resistance, comparisons of genome organization with other woody species, and a genome-wide examination of progress in backcross breeding for blight resistance. This reference assembly should prove of great value in the understanding, improvement, and restoration of chestnut species.

Keywords Chestnut · Genome · In situ hybridization · Disease resistance · Chestnut blight · Phytophthora root rot

Introduction

Chestnuts (*Castanea* species) are members of the Fagaceae (Order Fagales), whose members include many important forest tree species worldwide. The Fagaceae comprises eight genera containing over 1100 species (Kremer et al. 2007). These species are not only important to human industry (timber, pulp wood, furniture, cooperage and others) but are also dominant species in many of our forested ecosystems, providing food and shelter for wildlife as well as other important ecosystem services. Within *Castanea*, the seven recognized species—*C. crenata*, *C. dentata*, *C. henryi*, *C. mollissima*, *C. pumila*, *C. sativa*, and *C. seguinii* (Lang et al. 2007)—are

Margaret Staton, Charles Addo-Quaye and Nathaniel Cannon contributed equally to this work.

Submitting author Staton, Margaret

Communicated by M. Troggio

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11295-020-01454-y) contains supplementary material, which is available to authorized users.

✉ John E. Carlson
jec16@psu.edu

Margaret Staton
mstaton1@utk.edu

Extended author information available on the last page of the article

distributed mostly in the temperate regions of the world, with large indigenous ranges in North America, Europe, and Asia. Historically, these trees played critical roles in facilitating expansion and settlement of human populations into new territories and in many regions of the world are designated heritage trees under special laws of protection to ensure their survival.

Castanea species are well suited for addressing fundamental questions on the nature of host/pathogen genome coevolution and invasive pathogen biology in trees. For example, Chinese chestnut (*C. mollissima*) has coevolved with and has resistance to two major invasive pathogens, *Cryphonectria parasitica*(Murr.) Barr and *Phytophthora cinnamomi* Rands, both responsible for the complete demise of the susceptible American chestnut (*C. dentata*) as a dominant US forest species after their unintentional introduction to North America from Asia (Freinkel 2009; Anagnostakis 2012). Chinese chestnut and American chestnut can hybridize, and the resulting interspecies hybrid families segregate for disease resistance (Steiner et al. 2017) as well as other morphological and phenological traits (Diskin et al. 2006). Additionally, it is possible to obtain early flowering in these trees (within 2 years), substantially reducing generation time for basic and applied genetics approaches in heritage forest tree restoration programs (Baier et al. 2012). In this context, these species afford an excellent opportunity to advance our genetic understanding of the coevolution of host/pathogen complexes in forest trees as well as other traits. We have focused on the development of genetic and genomic resources in Chinese chestnut as a key tree species in *Castanea* that is currently used in several breeding programs as a donor of important traits including resistance to *C. parasitica* and *P. cinnamomi* in American chestnut (Kremer et al. 2007; Steiner et al. 2017).

Genome resources hold much promise for the restoration of forest tree species which have been, or are in the process of being, extirpated from their natural habitats by environmental threats imposed by exotic pests and pathogen, invasive organisms, and climate change. The extirpation of American chestnut from its natural range by the mid-twentieth century by the pathogens *P. cinnamomi* causing root rot (or ink disease) and *C. parasitica* Barr causing chestnut blight was recognized as the greatest environmental disaster of that time (Freinkel 2009). To address these and other environmental challenges, we have assembled a reference genome for Chinese chestnut (*Castanea mollissima*) to aid in the transfer of chestnut blight resistance loci from Chinese chestnut to American chestnut (*C. dentata*) through introgression (backcross breeding) and biotechnology approaches. We selected the cultivar "Vanuxem," one of the sources of resistance for the large-scale breeding program undertaken by The American Chestnut Foundation (Steiner et al. 2017) and a genotype used in genetic mapping efforts (Kubisiak et al. 2013). We demonstrate the utility of the genome with studies to understand the diversity and evolution of important host/pathogen complexes and other traits important in adaptation and response to climate change required for restoration of threatened chestnut species. This genomics-informed breeding for pathogen resistance may also serve as a model for genome-assisted species restoration efforts in other long-lived, undomesticated plant species.

Results

The Chinese chestnut genome assembly and structural features

An initial, de novo assembly of the genome, version V1.1, was produced in 2013 and released to the public in January 2014 as a browser and searchable database at the Hardwood Genomics website (www.hardwoodgenomics.org). A total of 13.7 Gb of 454 technology data (26.2 million reads) plus 46 Gb of Illumina MiSeq data (from 149.6 million reads) was produced. A de novo assembly placed 724 Mb in 41,260 scaffolds, which provided 91.2% overall coverage of the Chinese chestnut genome (estimated at 794 Mb by flow cytometry) (Kremer et al. 2007), with an N50 scaffold length of 39,580 bp, an L50 of 5019 scaffolds, and largest scaffold at 429,344 bp. The V1.1 scaffolds included 27,264 gaps, with an overall gap length of 13.5 Mb. A total of 36,478 gene models, and 38,146 transcripts and peptide sequences were predicted and annotated in the V1.1 assembly, which were also included for public access at the Hardwood Genomics database.

Improved, more contiguous de novo assemblies were then produced, based on co-localization with BAC-end sequences in the *C. mollissima* physical map [8], scaffolding with PacBio reads, and manual gap closing. This produced an improved hybrid assembly (V3.2) of 14,110 contig sequences with lengths up to 662 kb, encompassing 725.2 Mb of the genome with no substantial internal gaps and minimal sequence ambiguities. Mapping of DNA marker sequences from the integrated *C. mollissima* physical and genetic map (Kubisiak et al. 2013; Fang et al. 2013) validated that the V3.2 assembly accomplished close to complete coverage of the genome.

Contigs from the version 3.2 de novo assembly were initially anchored to chromosomal locations based on the order of DNA marker sequences on the chestnut research community's reference genetic linkage map for the Vanuxem genotype (Kubisiak et al. 2013). To increase the number of anchored contigs, DNA markers were integrated from 10 additional genetic maps: an expanded version of the original Vanuxem genetic map (Supplementary File 1 and 2), eight maps from three American chestnut backcross families

segregating for *P. cinnamomi* resistance (Zhebentyayeva et al. 2019), and a dense *Quercus robur* genetic map (Bodénès et al. 2016). Previous analysis of *C. mollissima* by *C. dentata* crosses and comparative genomics with *Q. robur* reveal highly conserved colinearity across chromosomes (Bodénès et al. 2012), lending confidence these maps will anchor contigs to the correct location. However, to prevent misplacements, these additional maps were filtered for regions consistent with the initial framework provided by the Vanuxem reference map. The set of anchored contigs, referred to as version 4.2, resulted in the screening of 13,195 genetic markers at 4618 unique genetic map positions, with which we uniquely anchored and ordered a total of 4040 contigs from the Chinese chestnut v3.2 de novo genome assembly. The summed anchored contig sequences for linkage groups range from 26.3 Mb (LG_L) to 59.9 Mb (LG_A), and totaling 412.8 Mb. This represents 57% of the total genome sequence length of 725.2 Mb. The long-standing linkage group lettering convention within the chestnut research community was maintained. A summary of the V4.2 anchored contig statistics is provided in Table 1.

Assembly validations

In situ assignment of anchored contigs to chestnut chromosomes

Individual linkage groups (LGs) were assigned to specific chestnut chromosomes by fluorescent in situ hybridization (FISH). From 2 to 8 markers per linkage group were chosen from the set of mapped markers on the Chinese chestnut reference genetic map (Kubisiak et al. 2013) that had been used to integrate the linkage groups from top to bottom on the Chinese chestnut BAC physical map (Fang et al. 2013). In addition, ribosomal DNA (18S–5.8S–26S; generally referred to as 45S and 5S rDNA) probes were used to identify their LG-specific cytological positions. For each linkage group and the corresponding marked regions of the physical map, BACs were selected as probes for FISH on Chinese chestnut root tip chromosome spreads (see Supplementary Table S1 for a full list of BAC clones selected). Since primary constrictions serve as cytologically visible landmarks for centromere positions, we were able to anchor the linkage groups to their respective chromosomes and determine the relationship of the linkage group to the long and short arms of each corresponding chromosome (Fig. 1). The 0 cM linkage map position was found to be associated with the short arm of nine chromosomes and the long arm of three chromosomes (LG_C, LG_G, and LG_L). The cytological analyses enabled a designation of six of the twelve Chinese chestnut LG-specific chromosomes (LG_A, _B, _C, _F, _G, and _I) as metacentric and/or near metacentric, four (LG_E, _H, _J, and _K) as near sub-metacentric, and two (LG_D and LG_L) as sub-metacentric chromosomes. Of the 54 BAC clones and two ribosomal DNA probes (45S rDNA and 5S rDNA) used in FISH, the cytological positions of all but three BAC clones were concordant with their expected linkage group position on the genetic map (Fig. 1). We observed the major 45S rDNA distally on the short arm of LG_H chromosome, but not the previously reported minor second locus (Ribeiro et al. 2011). A satellite (SAT) and nucleolus organizer region (NOR) were observed on the LG_H chromosome where the BAC H-C5 clone hybridized proximally to the 45S site (Fig. 1). One 5S rDNA site was located in the middle of the short arm of the LG_E chromosome. Representative cytological images showing examples of multiple BAC probe assignments by FISH to the *C. mollissima* LG_D chromosome are shown in Fig. 2, along with the corresponding locations of markers on the reference genetic linkage map.

Table 1 Chinese Chestnut assembly statistics

Assembly version	Assembly metrics
V3.2 de novo assembly	• Hybrid assembly of 454, Illumina, and PacBio sequences • 14,110 contigs covering 725.2 Mb • Longest contig: 663 kb • Contig N50 = 101,575 (L50 = 2098) • 30,832 high-quality gene models • BUSCO reported 1355 of 1440 expected single-copy complete genes within the *C. mollissima* genome • 98% of genetic markers included from Vanuxem consensus reference map (Kubisiak et al. 2013) • 99% of BAC-ends included from integrated genetic/physical map (Fang et al. 2013)
V4.2 LG-Anchored Contig Assembly	• 4040 contigs from V3.2 anchored to unique locations on *C. mollissima* linkage groups • Total of 412.8 Mb (57% of V3.2 de novo assembly) • Contig N50 = 144,434 (L50 = 931) • Anchored chromosome sequences range from 26.3 Mb (LG_L) to 59.9 Mb (LG_A) • Chromosome location assignments confirmed by FISH • 20,376 high-quality gene models in anchored contigs

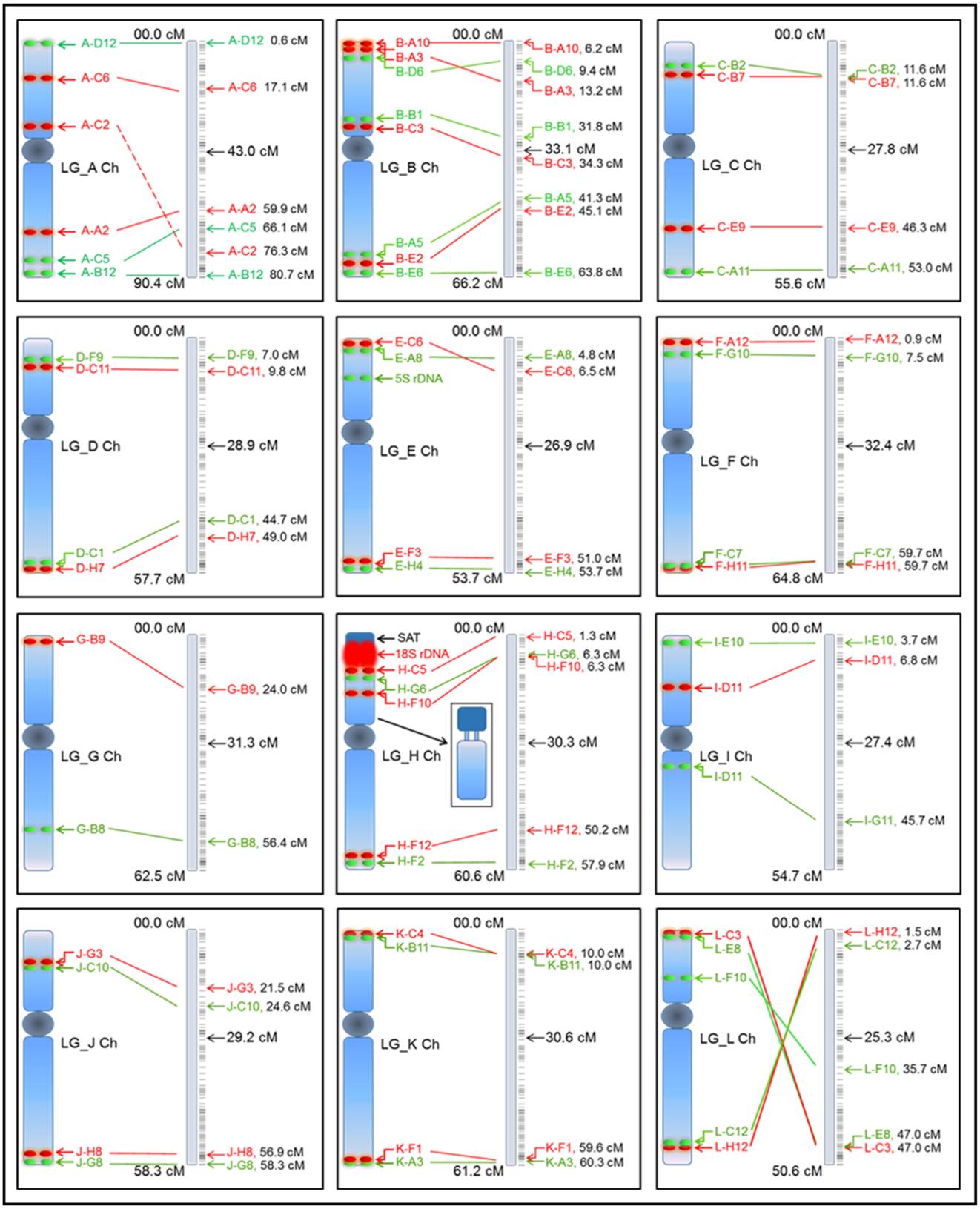

Fig. 1 Diagrammatic representation of BAC-FISH mapping results for assignment of *C. mollissima* chromosomes to their corresponding linkage groups. The putative position of the centromeres of each LG map is shown. Orientation of chromosomes is according to commonly accepted rules for karyotyping (short arm up); orientation of linkage groups follows the consensus genetic map (Kubisiak et al. 2013). The insert in the panel of LG_H shows the position of the satellite and rDNA expansion in the LG_H chromosome

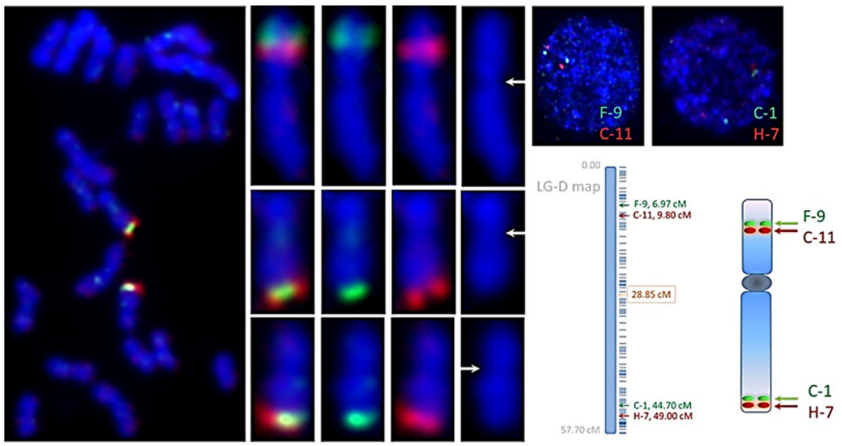

Fig. 2 Example of multiple-probe BAC-FISH mapping result for assignment of *C. mollissima* chromosomes linkage group D to its corresponding linkage group. FISH was conducted with four selected BAC clones (two from opposite ends of each arm) on Chinese chestnut chromosome spreads: (a) a complete root tip metaphase chromosome spread showed two BAC-FISH signals (BAC C1, 44.70 cM, green signal; BAC H7, 49.00 cM, red signal) located on the long arm of the LG_D homologs (chromosomes); (b1–4) an enlarged FISH image of a pro-metaphase chromosome with BAC clones (BAC F9, 6.97 cM, green signal; BAC C11, 9.80 cM, red signal) located on the short arm of the same chromosome; b1, a superimposed image from DAPI (blue chromosome background), FITC (green signal), and Cy3 (spectrum-orange/red signal) filters; b2, an image from DAPI and FITC filters; b3, an image from DAPI and Cy3 filters; b4, an image from DAPI filter, and it is the same for c1–4 and d1–4; and c1 and d1, enlarged images of the homologous pair of the LG_D metaphase chromosomes. The white arrows in b4, c4, and d4 showed the primary constriction (i.e., the centromere); (e) and (f) are from two interphase FISH nuclei showing the BAC-FISH signals; (g) a diagrammatic representation of the LG_D map; and (h) a diagrammatic representation the LG_D chromosome delineated by the primary constriction (centromere) and showed the short (S) arm and long (L) arm with respective BAC-FISH signals

Analyses of Chinese chestnut genome structural and functional features

Repetitive landscape in chestnut

A total of 1925 sequences were identified as repetitive elements within the *C. mollissima* assembly. Excluding rDNAs, the repeat sequences totaled 369,149,927 base pairs, or just over 50% of the genome. As shown in detail in Supplementary Table S2, most repetitive elements in the chestnut genome are interspersed repeats. The largest class of repetitive elements was in the "unclassified" category (25.09%), with the second most abundant class being retroelements (21.05%). Within retroelements, long terminal repeats (LTRs) were the most abundant form of repetitive element, consisting of 17.48% of the total genome. The most prevalent DNA transposon family identified was "Hobo-Activator."

Annotation statistics and quality assessment

Automated gene structure predictions utilized both deep RNAseq and protein homology information. Overall alignment of RNA reads to the assembly was high, with 95% or more of long Sanger-generated RNA read mapping. The statistics of alignments to the V3.2 genome assembly of RNAseq reads from several transcriptome projects can be found in Supplementary Table S3. The BRAKER gene-finding algorithm (Hoff et al. 2016) predicted a total of 50,911 genes for the assembly before filtering steps were taken. An evidence-based filtering protocol for genes supported by RNAseq and GenomeThreader gene models yielded 30,832 genes for the V3.2 de novo assembly, which is representative of gene number estimates for other Fagaceae species, such as *Q. suber* with 37,724 genes (Ramos et al. 2018) and *Q. robur* with 25,808 genes (Plomion et al. 2018). Of these high-quality gene models, 2085 were supported by homology with *Q. robur* only, 16,231 were supported by RNAseq only, and 12,518 were supported by both RNAseq and homology with *Q. robur*. The set of anchored contigs were then checked for presence of the gene models predicted for the complete v3.2 assembly. The v4.2 anchored contig set contained 20,376 of the filtered 30,832 genes models in the V3.2 Chinese chestnut de novo assembly. Thus, the chromosome-placed contigs currently include only 66% of putative gene models, suggesting that the remaining gene space may reside in contigs without available genetic map markers.

To examine the completeness of the annotation, the predicted genes were analyzed using the BUSCO software and database of single-copy orthologs for the group Embryophyta (Simão et al. 2015). BUSCO analysis reported that 1355 of the 1440 (93%) expected single-copy genes were complete and present within the *C. mollissima* genome. Of the complete BUSCOs, 1266 were single-copy genes, with the remaining 89 being present in more than one location in the genome. Of the unaccounted single-copy orthologs, 33 were fragmented ORFs and 52 were missing. With all but a few of the

single-copy genes present in the genome with full-length annotation, this indicates our assembly and annotation of the gene space is largely complete.

Functional annotations

As a first step in evaluating the evolution of the gene space in chestnut, we computed the shared orthogroups between the 30,835 chestnut gene models and a selection of model species with complete genome resources representing both woody tree and herbaceous plants. Chestnut predicted proteins were identified and clustered with the most current Arabidopsis (Lamesch et al. 2012), peach (International Peach Genome Initiative et al. 2013), poplar (Tuskan et al. 2006), and grape (Characterization TFPCFGG, The French–Italian Public Consortium for Grapevine Genome Characterization 2007) gene annotation sets to create ortholog groups. The resulting 16,687 orthogroups spanned 71.4% of 163,425 predicted proteins from the 5 species (for detailed results see Supplementary Table S4), with a mean size of 7 proteins per group. Only 212 species-specific orthogroups were obtained for chestnut, while 11,624 orthogroups had representatives in all 5 species. The analysis showed that the chestnut genome reference, as judged by the number of shared orthogroups, does not significantly differ from other closely and more distantly related species. Much of the historic interest in the genetics of chestnut has focused on resistance to the invasive pathogens that eliminated American chestnut as a dominant species of the eastern forests of North America. For this reason, we were particularly interested in the potential discovery of genes that underlie resistance to fungal or oomycete pathogens. The phenylpropanoid pathway produces the precursors to lignan and has been shown in previous studies (e.g., avocado (Engelbrecht and van den Berg 2013), eucalyptus (Cahill and McComb 1992)) to underpin stress response in trees to biotic and abiotic stressors and thus was of direct interest to this study (Table 2). For all genes examined from this pathway, chestnut had a similar number of copies of each gene to other plant species.

The NBS-LRR gene family

An examination of the NBS and LRR genes in *Q. robur* identified a major gene expansion in comparison to other woody plant species (Plomion et al. 2018). To examine if the chestnut genome shares this expansion, the orthogroups were updated to include the *Q. robur* protein sequences along with the original five species. A Pfam search using NBS and LRR motifs produced the following NBS-LRR gene family totals: *Chestnut: 300; Peach: 386; Vitis: 450; Poplar: 556; Oak: 874*. While our parameters for the Pfam search produced slightly different totals than previously reported for peach, grapevine, poplar, and oak, this result is consistent with the general observation that this gene family has experienced a major expansion in *Q. robur* (Plomion et al. 2018). In contrast, the NBS-LRR family of disease-resistant genes in the Chinese chestnut genome appears to be reduced to a number even lower than in the comparatively small peach genome (265 Mb) (International Peach Genome Initiative et al. 2013). While this strongly suggests a reduction in NBS-LRR genes in chestnut, there are a number of possible reasons this may be an underestimate of the total number of these genes in the genome, including high sequence similarities among NBS-LRR genes causing the assembly algorithm to collapse the copy number of tandemly repeated genes during the assembly process.

Genome structure

Since whole-genome sequences are available for several tree species, it was of interest to assess the level of genome preservation between chestnut and other species that have significant information on gene/trait associations. Strong preservation of genome organization may indicate gene/trait information can be translated across species and thus leverage the resources invested in one species to assist in knowledge gain in another. In this regard, we performed genome comparisons by shared one-to-one orthologs between chestnut (*C. mollissima*) and oak (*Q. robur*) which along with chestnut is within the Fagaceae family (Kremer et al. 2007; Kubisiak et al. 2013), and between chestnut and peach (*Prunus persica* Batsch.), a member of the Rosaceae for which there is rapidly increasing information on gene/trait associations (Aranzana et al. 2019). Overall, the ortholog mapping illustrates the high degree of macro-synteny at the whole chromosome level between chestnut and oak in agreement with previously reported genetic mapping studies (Kremer et al. 2007; Kubisiak et al. 2013; Ramos et al. 2018; Plomion et al. 2018) (Fig. 3a). Major blocks of synteny were also observed between the chestnut and peach genomes (Fig. 3b). This illustrates, as previously reported for oak (Plomion et al. 2018), that only a few chromosomal breaks and fusions may account for the differences in overall genome organization between the Fagaceae and Rosaceae families from their last common ancestor. A more detailed illustration of macro- and micro-syntenies for individual chestnut chromosomes with the oak and peach genomes are shown in Fig. S1 and S2. The individual alignments of chestnut chromosomes revealed that the chestnut chromosomes do contain some genes that do not fit the one-to-one chestnut-to-oak linkage group collinearity supported by the majority of genes (Fig. S2). While this may indicate small-scale chromosomal rearrangements, it is more likely an artifact of the ortholog identification process. Anchoring of additional chestnut contigs to chromosomes and improvement of both the oak and chestnut reference genomes will be needed to clarify these regions.

Table 2 Predicted numbers of lignin monomer pathway orthologous gene models in Chinese chestnut and model tree genomes

Gene family	Orthogroups	Arabidopsis	Chestnut	Peach	Populus	Grape
PAL	1	3	3	2	4	12
C4H	2	1	2	2	3	3
4CL	10	11	14	13	17	14
HCT	1	1	1	2	2	1
C3H	1	1	2	4	3	1
CCoAOMT	4	6	11	7	5	9
CCR	6	8	8	9	11	11
F5H	1	2	1	2	3	3
COMT	10	10	16	28	21	28
CAD	5	7	17	18	11	17

Signatures of selection in the blight resistance QTL *cbr1* region on linkage group B in *C. mollissima* vs *C. dentata*

The relative ease of hybridizing *C. mollissima* and *C. dentata* enabled previous QTL mapping studies (Kubisiak et al. 1997, 2013) to be conducted on the genetics of resistance to the blight disease in chestnut. With the availability of the *C. mollissima* genome, we were interested in knowing if genes in the mapped QTL regions might reveal signatures of selection differentiating American chestnut and Chinese chestnut, in support of the hypothesis that the blight pathogen had exerted selection pressure on Chinese alleles over the period of its coevolution with the host. If so, the affected genes would be the most suitable targets for resistance breeding. For the identification of selective sweep regions in the Chinese and American chestnut genomes, we focused on LG_B where mapping studies have demonstrated a significant QTL for blight resistance (Kubisiak et al. 1997, 2013). Statistical tests for departure from neutrality were calculated for two resequenced populations (five *C. dentata* and five *C. mollissima* genotypes) of American and Chinese chestnut. Two statistical parameters were calculated for each species: pairwise nucleotide diversity (?) and pooled Tajima's *D* (TajD). A 5-kb sliding window with step size of 1 kb was employed to allow gene-level resolution. The distribution of nucleotide diversity ratios in American vs Chinese chestnut (?Cden/?Cmol) plotted along the *C. mollissima* v4.2 LG_B is shown in Fig. 4a. Pairwise nucleotide diversity in *C. dentata* was higher than that for *C. mollissima* by 2 nucleotides per 5-

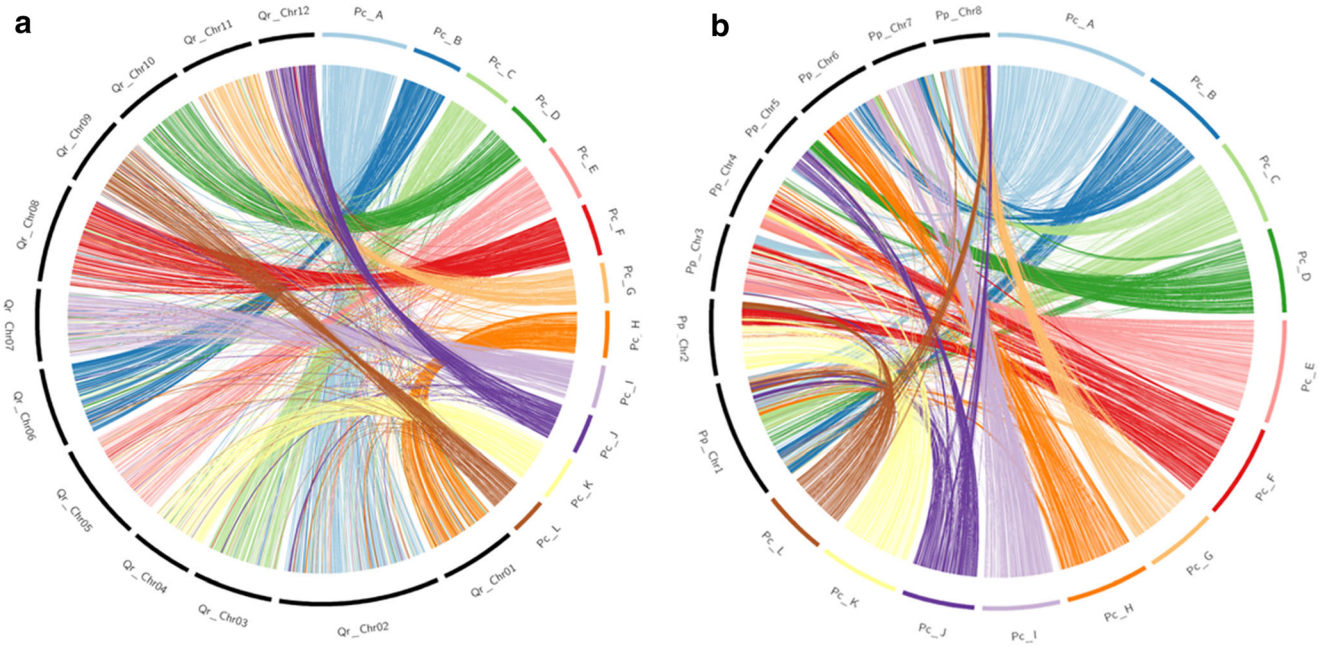

Fig. 3 Circos plot of alignments of orthologous genes in *C. mollissima* chromosomes (Pc_A-L) vs *Q. robur* chromosomes (Qr_1–12) (**a**) and *Prunus persica* chromosomes (Pp_1–8) (**b**). Alignment of the genomes was filtered to show only blocks supported by at least two collinear genes. Pc for pseudochromosome refers to the set of contigs anchored to LG locations

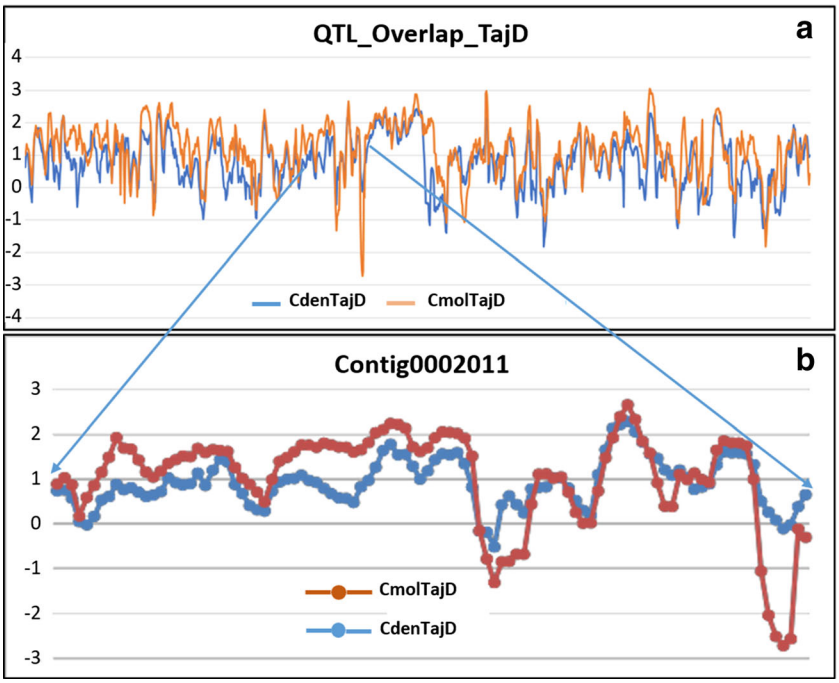

Fig. 4 Stepwise signature of selection analyses of a region of overlap in the LG_B QTL for blight resistance in two reference mapping studies. **a** A profile of TajD values determined from nucleotide diversity ratios in a QTL region on LG_B detected in mapping studies (Kubisiak et al. 1997, 2013) and defined as *cbr1* for *C. parasitica* resistance. The region of overlap between the QTLs consists of 13 contigs spanning a total of 1.122 Mb. **b** From the QTL region, the strongest differential signature of selection (negative TajD value in *C. mollissima* and positive in *C. dentata*) was identified within one gene in contig 0002011. This contig maps to EST-DNA marker CmSI0550 on the reference genetic map (Kubisiak et al. 2013). The *C. mollissima* profile suggests a very strong signature of selection at the end of this contig across a window of app. 6000 bp

kb window ($p < 0.001$, Wilcoxon signed-rank test). In total, 8 genomic regions (5 kb and larger) were detected as candidate regions under purifying or positive selection in *C. mollissima* (Tajima's $D < -2$ and ?Cden/?Cmol in upper 0.5% of the genome-wide distribution) (Table S5). One significant genome-wide outlier on LG_B, in contig 0002011, overlapped a QTL interval identified in two mapping studies (Kubisiak et al. 1997, 2013) (Fig. 4b). The strongly negative Tajima's *D* value overlaps with the predicted gene Cm_g21596, which has very strong BLASTn and BLASTp alignment scores (*E*-values of 0 and 75–96% identities) to Type I Inositol Polyphosphate 5-Phosphatase 1 genes in woody plants. UniProt describes the IP5P1 protein as "involved in the abscisic acid (ABA) signaling pathway (PubMed:12805629)." The IP5P1 gene's top gene ontology (GO) functional category is the biological process "abscisic acid-activated signaling pathway".

Signatures of selection in the *Phytophthora cinnamomi* resistance QTL region in *C. mollissima* vs *C. dentata*

The distribution of nucleotide diversity ratios in American and Chinese chestnut (?Cden/?Cmol and ?Cmol/?Cden) was plotted along the *C. mollissima* v4.2 chromosome LG_E. Pairwise nucleotide diversity in *C. dentata* was lower than that for *C. mollissima* by 1 nucleotide per 5-kb window ($p < 0.001$, Wilcoxon signed-rank test). Leveraging data from the previous QTL mapping studies of *P. cinnamomi* resistance in Chinese/American hybrid families (Zhebentyayeva et al. 2012, 2019; Olukolu et al. 2012; Santos et al. 2017), we ran statistical tests for natural selection across LG_E that had three strong QTL intervals associated with *P. cinnamomi* resistance. Using sequence-based markers from our mapping analyses and local blast alignment tools, we delineated the QTL intervals on the contigs anchored to Chinese chestnut LG_E chromosome and searched for genome-wide outliers from the selection scan in these QTL regions. In total 49 genomic regions (5 kb and larger) were detected as candidate regions under purifying or positive selection in the *C. mollissima* pool. Of these, 34 candidate regions were located within QTL intervals (Fig. 5). Similarly, 45 regions exhibited signatures of purifying or positive selection in the *C. dentata* pool, but these are located outside of QTL intervals for resistance to *P. cinnamomi*. Most of the loci potentially contributing to adaptation in Chinese chestnut to biotic stress caused by *P. cinnamomi* were annotated as genes involved in cell wall formation, transmembrane signaling and transport, posttranslational protein modification, and formation of reactive oxygen species (ROS) (Table S6).

Signatures of selection in phenology traits, the bud burst QTL region in Chinese chestnut, oak, and peach

Due to the extensive colinearity of deciduous tree genomes as highlighted above, we were able to perform genome comparative analysis for mapped QTL controlling budbreak in peach (Fan et al. 2010), oak (Scotti-Saintagne et al. 2004; Casasoli et al. 2006; Derory et al. 2010), and Chinese/American chestnut hybrids (Fan et al. 2020). This analysis revealed one common major colocalizing QTL region that in all three mapping analyses contributed with high significance to variation for budbreak in either floral buds (peach) or vegetative buds (oak and chestnut). This QTL was originally mapped in peach (Fan et al. 2010) and corresponds to the location of the Dormancy Associated MADS-box (DAM) genes in *Prunus* (Bielenberg et al. 2008). Here, we performed a comparative sequence characterization of this region among these three species utilizing the published genome sequences of peach, oak (*Q. robur*) and Chinese chestnut (Fig. 6a, Fig. S3). Results from this sequence comparative analysis reveal a high degree of preserved gene content and order among these species genomes; however, the *Fagaceae* species do not contain the tandem duplication of DAM genes that is characteristic of the peach genome. Due to the high degree of gene preservation in this region across many species, we hypothesized that genes in this region could show signatures of selection among chestnut species particularly since budbreak timing is a very selectable trait in fruit trees (Labuschagné et al. 2003; Campoy et al. 2011), and Chinese and American chestnuts show significant differences in the budbreak dates (Hebard 1994; Diskin et al. 2006). In order to determine if any of the genes in this common budbreak QTL demonstrated signatures of selection and if so, which genes, we calculated nucleotide diversity and Tajima's *D* values for LG_L in chestnut as outlined for the studies we performed on LG_E and LG_B, above. The analysis showed that across linkage group L, 43 loci were identified with negative TajD values in *C. mollissima* and with neutral or positive values in *C. dentata*, indicating potential regions under different scenarios of adaptation (purifying or balancing selection). However, within the QTL region, only the DAM gene ortholog in chestnut showed a signature of purifying or positive selection in the chestnut species comparison (Fig. 6c). This is consistent with the importance of its putative role in controlling dormancy and bud flush and the high heritability and selectability of this trait in fruiting trees.

Assessment of progress of the backcross breeding program on recovery of American chestnut genome

A multi-dimensional scaling (MDS) approach was used to visualize components of variation in sequence data among individuals representing several stages of introgression-based chestnut blight resistance breeding (Fig. 7). DNA sequence data from single-enzyme RADseq libraries were obtained for 48 BC3F2 progeny, 2 *C. dentata* parent trees, a *C. dentata* great-grandparent tree, six wild American chestnut trees, the *C. mollissima* genome reference genotype Vanuxem, and a clone of the original source of resistance for chestnut blight (for further descriptions of the genotypes used please refer to "Materials and methods"). Whole-genome mapping of sequence reads to the *C. mollissima* reference genome was used for SNP calling and genome-wide analysis of variation.

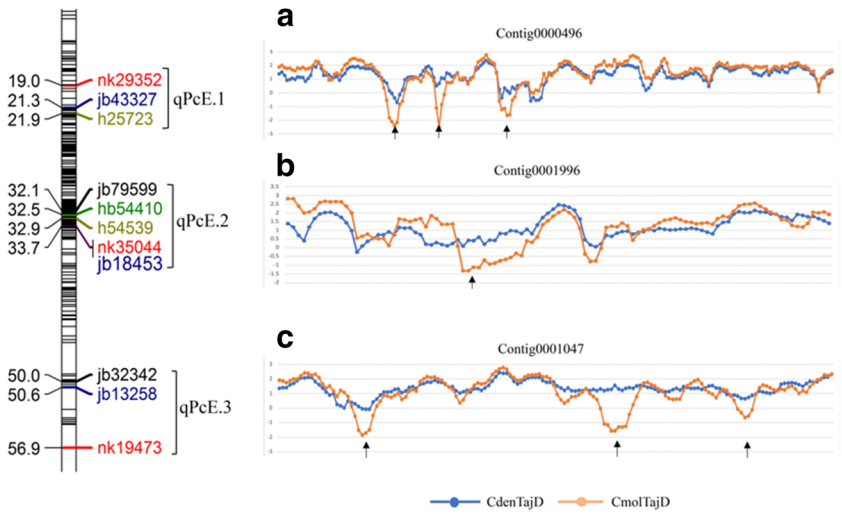

Fig. 5 Three QTL intervals for resistance to *P. cinnamomi* in linkage group E. The TajD profiles for the 3 QTL regions (profiles A, B, and C, respectively, for QTL qPcE.1, qPcE.2, and qPcE.3) are shown on the right. Representative contigs within QTL intervals are from genomic regions under purifying or negative selection in Chinese chestnut. TajD peaks of approximately −2 were considered most significant (identified with arrows), in which candidate genes for *P. cinnamomi* resistance were identified. Left panel—modified figure from Zhebentyayeva et al. (2019) representing fragment of the linkage map E with relative positions of DNA marker loci associated with candidate genes

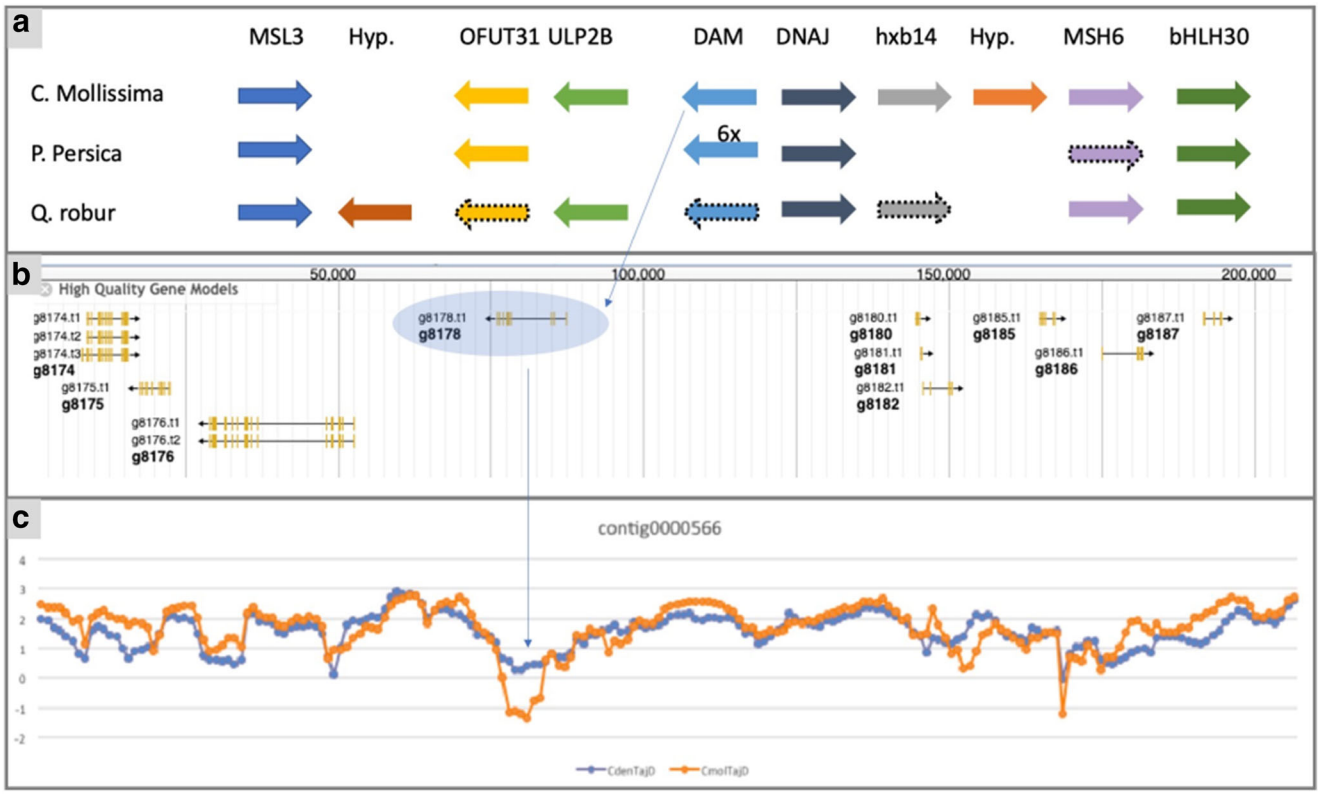

Fig. 6 The structural organization of the DAM gene-containing region (contig0000506) located in mapped QTL intervals of vegetative bud break in chestnut. **a** The genes in the region show conserved order in comparison to a region in oak (Chr09) and in peach (scaffold 1). Arrows denote specific annotated genes of the three species with shared colors indicating orthologs and conserved function noted above (MSL3, mechanosensitive ion channel protein 3; Hyp., hypothetical gene, function unknown; OFUT31, O-fucosyltransferase 31-like; ULP2B, ubiquitin-like-specific protease 2B; DNAJ, dnaJ homolog subfamily C member 14-like; hxb14, homeobox protein 14; MSH6, DNA mismatch repair protein 6-like; bHLH30, transcription factor bHLH30). Where arrows are not present, the gene is not found in this region; genes surrounded with a dotted line indicate the gene is not officially annotated but was found and manually annotated for this comparison. The *Prunus*-specific 6X-segmental duplication of the DAM genes is denoted with the text "6X". **b** The same region in chestnut with full gene models and alternative transcripts displayed in relative location on a genome browser. **c** The TajD values in this region demonstrating the differential signature of purifying or positive selection in the *C. mollissima* genome (orange line) vs the *C. dentata* genome (blue line), with the lowest point aligning with the position of the DAM gene (Cm_g8178)

MDS of SNP variants shows strong resolution between the species, the Clapper BC1 source of blight resistance, and the BC3F2 hybrid groups (Fig. 7). The distribution of variants among the BC3F2 trees placed most of the backcross progenies between the genome of the original "Clapper" BC1 genotype and existing *C. dentata* trees (Fig. 7), with many clustering closer to the *C. dentata* individuals. Thus, most hybrids sampled in the BC3F2 generation are intermediate in genome composition between the BC1 source of resistance and the recurrent American chestnut parents. However, a few of the BC3F2 individuals clustered with the American chestnut genotypes, indicating that they are more "American" in genome composition. We quantified the number of shared SNPs among the individuals to provide a first look at percentage of American and Chinese chestnut DNA in the hybrid group. We first removed SNPs found in both *C. mollissima* Vanuxem and any of the nine *C. dentata* accessions, then assessed the remaining SNPs in each BC2F3 individual. The majority of SNPs detected in each individual were shared with *C. dentata* (62 to 85%), with fewer shared with Vanuxem (0.7 to 5.2%) and an intermediate number not found in either (7.2 to 35.2%) (Supplementary File 3). This is in contrast to the BC1 Clapper, which shows a much higher number of shared SNPs with Chinese chestnut: 9.6% shared with Vanuxem, 80.7% shared with *C. dentata*, and 9.6% unique. As Vanuxem is not the parental source of Chinese chestnut DNA in Clapper or the BC3F2s, the unique SNPs likely represent parents not available for sequencing and further supports additional analysis of variation across the original species and The American Chestnut Foundation (TACF) breeding program to fully characterize introgression patterns. The MDS analysis also revealed a high level of variation among technical replicates of sequencing libraries for Clapper BC1 and at least 3 of the American chestnut parental genotypes ("Ort," "Joliett," and "LFR4T14"), as revealed by the MDS component 2 on the *y*-axis of Fig. 7. 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 variation among genomic DNA samples.

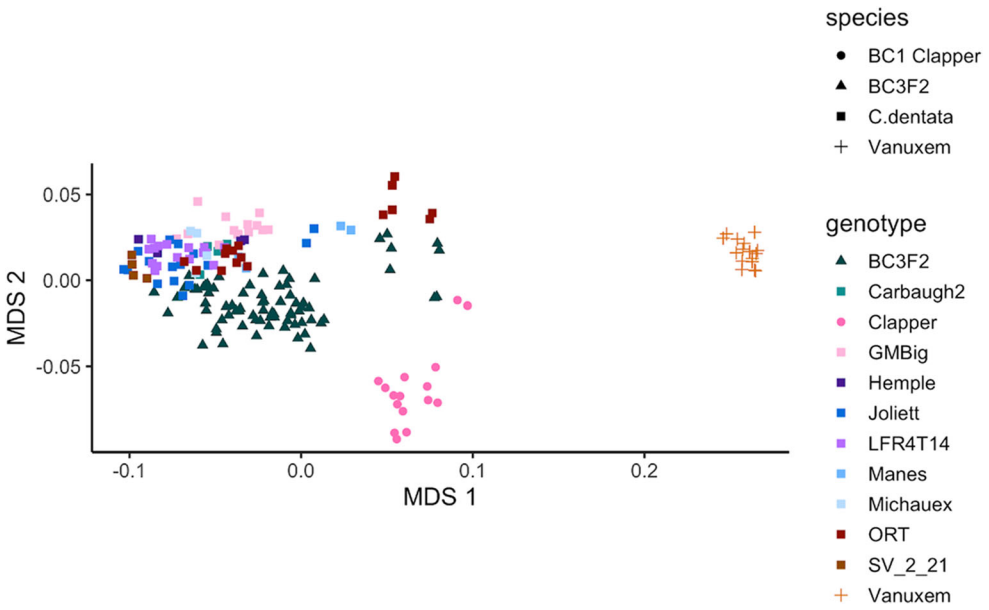

Fig. 7 Multidimensional scaling (MDS) analysis of SNP variation among sequence reads mapped to the Vanuxem reference genome. Variants predicted from mapping of reads for each individual and replicates were used to generate a genome-wide estimate of genomic distance among BC3F2 progeny relative to the Chinese reference sequence (Vanuxem), the blight resistance source (BC1 "Clapper"), American parents (Ort, Joliett), and other wild representatives of *C. dentata*. MDS of SNP variants shows strong resolution between species and hybrid groups as displayed on the *x*-axis (MDS1). Variation among samples within each group of trees is displayed on the *y*-axis for MDS2

Discussion

The tragic story of the American chestnut's demise is just one sobering example of what is becoming a recurring theme of the unintentional impact of human activity on our natural forest ecosystems. The success of future forest conservation and restoration efforts will increasingly rely on our ability to rapidly generate genetically improved and genetically diverse tree materials for forest replanting. As the foundation of a genomics toolbox, a genome sequence for a species provides the genetic architecture by which we can bridge genetic studies of traits to discovery of the underlying genes that control these traits. With this goal in mind, we developed a whole-genome sequence assembly of the Chinese chestnut Vanuxem. As an important genotype for resistance in ongoing breeding programs, we also demonstrate its utility in evolutionary and comparative genomics studies to advance our understanding of genome evolution and adaptation of species in the genus *Castanea* and to identify candidate genes for traits critically important to future conservation and restoration efforts.

The Chinese chestnut genome assembly and structural features

The initial draft de novo assembly of the genome, version V1.1, has supported many investigations and publications (Nelson et al. 2014; Staton et al. 2015; Pereira-Lorenzo et al. 2016; LaBonte et al. 2018; Tuskan et al. 2018). To better support basic research and restoration of the American chestnut, we then focused efforts for several years on developing more contiguous, chromosome-scale sequences. This proceeded through painstaking manual merging of contigs and gap closing until an assembly (V3.2) of 12,684 contig sequences spanning 783.4 Mb (98% of the estimated genome) was achieved. Placement of contigs to chromosome locations (version V4.2) was accomplished by genetic map marker-based anchoring 4040 of the V3.2 contigs. The overall placement of contigs was validated through cytological mapping (Figs. 1 and 2) and chromosome-scale sequence alignments to related tree genomes (Fig. 3). The V4.2 set of anchored contigs, although incomplete, does by virtue of organizing long contigs into chromosome-order provides a significant advancement in our ability to investigate genome organization and the evolution and genetic structure of important traits in chestnut, such as disease resistance, as well as in applications such as genome-wide selection.

Despite these efforts, our anchoring approach left gaps of unknown size and content between each contig, and the total number of placed contigs represented only 57% of the V3.2 de novo genome assembly of 783 Mb. By conservatively limiting genetic markers with sequence matches on multiple contigs to one chromosomal placement, we may have missed segmental duplications and/or repetitive non-coding regions. In addition, the availability of only short read sequences during the de novo stage of contig assembly might also have

limited the extent of repetitive DNA assembly. Improvement of the assembly might be achieved through the use of recent long-read technologies such as Nanopore (Madoui et al. 2015) or by scaffolding using chromatin-interaction data, such as Hi-C (Jiao and Schneeberger 2017). Difficulty in obtaining complete, correct genome assemblies at the chromosome level for undomesticated woody perennial plants such as chestnut may result from challenges in assembly inherent with obligate outcrossing plants with high levels of heterozygosity that cannot be reduced through inbreeding nor through creation of dihaploid individuals. The recently published *Quercus robur* genome (Plomion et al. 2018) utilized synteny with the *Prunus persica* genome to incorporate contigs and scaffolds that could not be placed with oak genetic map markers. This approach assumes that micro-level syntenies follow known macro-syntenies based on genetic maps, which may or may not always hold true. However, the hybrid synteny approach might complement the use of long-read technologies in future Chinese chestnut genome improvements.

Castanea comparative genomic analyses

Our comparisons of the chestnut genome with a selection of genomes that include the herbaceous model plant Arabidopsis, as well as woody vines and trees (grape, oak, peach, and poplar) confirmed that there have not been any recent genome duplications in chestnut, in keeping with previous reports for peach and oak. Our comparative analyses also confirmed the strong colinearity among genomes and gene content in tree species. This conservation of genome structure and information content after millions of years of species divergence suggests strong constraints on the evolution imposed in perennial plants of long-generation time and limited domestication (Luo et al. 2015; Staton et al. 2015; Groover and Cronk 2017). The conservation of genome structure may also be an underlying reason that gene flow is high and that interspecies hybridizations in natural stands of trees are so common. The conservation of genome structure and information content will permit the leveraging of information from model plants and among tree species to more rapidly advance our understanding of the many unique and fascinating features of long-lived tree species.

In hardwood forest trees, traditional single-family trait mapping and selections typically have not been done due to the long-generation times in these species. In contrast, fruit tree genetics is driven by domestication and orchard plantings and thus is a rich resource of genetic information on genes that control many aspects of fruit tree growth and development (Aranzana et al. 2019). As fruit trees are deciduous, many trait/gene associations are likely to translate to hardwood forest trees. Here we demonstrate that the high level of genome synteny between the peach, chestnut, and oak enables a comparative QTL analyses for traits central to sustaining forest trees in a rapidly changing environmental landscape. Our initial results indicate that knowledge of trait/gene associations in one tree species may be translated to other tree species, providing key information for genetic improvement of tree species with less genomic resources. Thus, we undertook searches to identify genes carrying signatures of selection within chestnut genomic regions known to be important in disease resistance and adaptation in chestnut and other trees.

Signatures of selection for blight resistance on linkage group B in *C. mollissima* vs *C. dentata*

Castanea species originated in eastern Asia, moved westward during the Tertiary period and currently exhibit a disjunct distribution pattern in eastern Asia and eastern North America (Lang et al. 2007). Asian chestnut species such as *C. mollissima* can be assumed to have evolved blight resistance as a result of continuous pressure from their endemic pathogen *C. parasitica*. Under such a hypothesis, American chestnut would not be expected to carry signatures of selection on the same genes as Chinese chestnut in genetically mapped blight resistance QTL intervals. Nonrandom fixation rates within genomic regions under selective pressure (i.e., selective sweeps) can be detected using a variety of population statistics (e.g., selection tests estimating nucleotide diversity and Tajima's *D* values). Reduction in the nucleotide diversity in these genes by elimination of susceptible alleles, excess of rare mutations, and negative TajD values can all result in signatures of loci under negative or positive selection. We utilized resequencing of five American and five Chinese chestnuts to identify these signatures. With the resulting dense genotyping data and scans conducted across 5-kb windows, the TajD statistic is likely robust with this number of samples (Tennessen et al. 2010).

Resistance to *C. parasitica* has been mapped in several interspecies hybrid Chinese × American families in both F_2 and backcross configurations (Kubisiak et al. 2013; Fan et al. 2020). From the QTL analyses, it appeared that variation in the resistance phenotype is attributable to a relatively small number of loci donated by the Chinese parent in the initial cross. One region in particular, located on linkage group B, has been reproducibly associated with the resistance. Leveraging the *C. mollissima* whole-genome sequence, we performed a comparative Tajima's *D* analysis across linkage group B in five *C. mollissima* and five *C. dentata* resequenced accessions and searched for contrasting evidence of selective sweeps between their genomes. Our results pinpointed only one gene in the core LG_B QTL region from the 10 resequenced individuals that showed a significant signature of purifying or positive selection in Chinese chestnut but not American. This gene model is a putative inositol polyphosphate-related phosphatase gene. Inositol phosphate signaling has been linked to multiple effects within plants, most notably abiotic stress (Kaye et al. 2011) and plant defense responses (Hung et al. 2014; Williams et al. 2015). As

transgenic studies can be performed in American chestnut, this candidate gene should be directly tested for its ability to confer resistance in future transgenic experiments.

Signatures of selection for *Phytophthora cinnamomi* resistance on linkage group E in *C. mollissima* vs *C. dentata* Based on the global studies of mating type, *P. cinnamomi* is also hypothesized to have an Asiatic origin (Zentmyer 1988; Arentz 2017). In a similar manner to chestnut blight, coevolution between chestnut and this pathogen in Eastern Asia could have generated a strong selection pressure in *C. mollissima* to evolve defense mechanisms directly targeting *P. cinnamomi* and/or blocking host infection. The plant cell wall is the first barrier encountered by *P. cinnamomi* zoospores attempting to colonize chestnut roots. Alterations in cell wall structure can have significant impact on disease resistance (Miedes et al. 2014; Hamann 2015). Our tests showed that several cell wall-associated genes within QTL intervals may be under purifying or positive selection in the *C. mollissima* but not in *C. dentata*. These include a probable pectin methylesterase CGR2 (contig0000773_177500–180500) and a Golgi-localized type II membrane protein that has enzymatic activity toward pectin methyl-esterification. Two genes involved in phenylpropanoid metabolism, a probable 4-coumarate–CoA ligase 1 (contig0003825_11500–13500) and a flavanone 3-hydroxylase (contig0001240_19500–22500) may contribute to lignin biosynthesis and modification. Two tandem putative endo-1,3-beta-glucosidases (Cm_g9986 and Cm_g9987) are of interest as β-linked glucose polysaccharides are highly abundant in *Phytophthora* cell walls, with activity in modulating plant innate immunity (Robinson and Bostock 2014; Bacete et al. 2018), including resistance to *P. cinnamomi* in avocado rootstocks (van den Berg et al. 2018). Finally, a putative beta-fructofuranosidase, an ortholog of insoluble Cell Wall Invertase 1 (CWINV1) in Arabidopsis, may be subject to selection within the qPcE.1 region (Cm_g4432). This enzyme is ionically bound to the cell wall and was described as one of the key enzymes during plant pathogen/interactions (Tauzin and Giardina 2014; Veillet et al. 2016).

Membrane-localized receptors also may play a significant role in the defense against pathogens. Chemical substances associated with pathogen and/or cell wall degradation can modulate plant innate immune response upon recognition by receptors with varied extracellular domains (Raaymakers and Van den Ackerveken 2016). Several receptor genes may be subject to purifying or positive selection in *C. mollissima*, including genes in two LG-E contigs—a G-type lectin S-receptor-like serine/threonine-protein kinase G-LecRK in contig0000723_29500–34500, and a block of duplicated cysteine-rich receptor-like protein kinases (CRKs) in contig0001047. G-type-lectin-RLKs (Teixeira et al. 2018) were reported to be upregulated in roots of diploid strawberry, citrus rootstocks, and Japanese chestnut (*C. crenata*) infected with *P. cactorum, P. parasitica*, and *P. cinnamomi* zoospores, respectively (Serrazina et al. 2015; Toljamo et al. 2016; Naveed et al. 2019). A block of tandem duplicated CRKs potentially under purifying or negative selection (with negative Tajima's *D* values) was identified within the qPcE.3 (Fig. 5). CRKs are transmembrane proteins that exhibit ectodomains containing the cysteine-rich Domains of Unknown Function 26 (DUF26). They constitute a land plant-specific family of carbohydrate-binding proteins expanded through tandem duplications. Due to the presence of extracellular Cys-rich domains (C-X8-C-X2-C), the CRKs are potential targets for redox modifications and hypersensitive response associated with programmed cell death (Lee et al. 2017; Vaattovaara et al. 2019). CRKs were reported as overexpressed in soybean roots induced by *P. sojae* zoospores (Delgado-Cerrone et al. 2018) and in resistant *C. crenata* genotypes treated with *P. cinnamomi* zoospores (Naveed et al. 2019). In addition, our *C. mollissima* plants all had tandemly duplicated genes for ABC transporter C family member 8 (Cm_g6287 and Cm_g6288) which functions as a pump for glutathione S-conjugates in transmembrane transport (Kang et al. 2011).

The most striking difference in nucleotide diversity between *C. dentata* and *C. mollissima* was observed in the middle of chromosome E which colocalizes with qPcE.2, the most stable QTL detected over multiple years in progeny derived from two Chinese chestnut sources of resistance to *P. cinnamomi* (Zhebentyayeva et al. 2019). Two genes were annotated in this region—a putative ALA-interacting subunit 2, homolog of the ligand-effect modulator 3 (AT5G46150) in Arabidopsis that encodes a protein of unknown function with transmembrane activity; and an ortholog of ornithine delta-aminotransferase (KEGG:AT5G46180) involved in arginine and proline metabolisms and transcriptionally upregulated in response to osmotic stress and non-host disease resistance (Anwar et al. 2018). Though the role of these genes in defense response against *P. cinnamomi* is not clear, the neutrality tests using the whole-genome resequencing datasets *C. mollissima* and *C. dentata* highlighted the potential importance of putative ALA-interacting subunit 2, homolog, and ornithine delta-aminotransferase for adaptation of *C. mollissima* to pathogenic pressure by *P. cinnamomi*.

Signatures of selection in the bud burst QTL region in chestnut, oak, and peach Arguably from the standpoint of rapid climate change, the rate of evolution of tree genetic composition and the rapidly changing environmental factors pose one of the greatest challenges to adaptation of perennial trees. This is particularly true for phenological traits such as flowering and vegetative bud burst. From a number of studies in fruit and forest trees, a picture of the genetic control and evolution of the genes and gene networks that control the timing of floral and vegetative bud break is emerging (Cooke et al. 2012; Shim et al. 2014; Abbott et al. 2015). Comparative mapping

of QTL locations of budbreak loci among peach, chestnut, and oak and the availability of whole-genome sequences for each species enabled us to readily surmise that an orthologous genomic region was present in the major budbreak QTL of all three species. This region contains a single MADS-box transcription factor gene in oak and chestnut and a segmentally duplicated gene (six copies in Prunus) that has previously been characterized as a major floral bud dormancy and budbreak control gene in a number of fruiting tree species (Bielenberg et al. 2008; Campoy et al. 2011; Abbott et al. 2015; Liu et al. 2015) and in at least one fruit tree vegetative budbreak QTL as well (Gabay et al. 2018). Combining genomic sequence analyses, comparative QTL analyses and our TajD analyses of LG_L in chestnut, we hypothesize that the DAM gene-containing locus in these deciduous forest tree species is a major control locus for both vegetative and floral budbreak. Due to its central importance in regulating the timing of budbreak as seen in fruiting trees, the DAM gene differentially predisposes it to environmental and breeding selection pressures over those genes in this conserved region.

Assessment of progress of the backcross breeding program on recovery of American chestnut genome Genotyping by sequencing 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. Genotyping by sequencing 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 in the introgression process among progeny resulting from interspecies hybridization, which is often used in plant breeding to transfer disease resistance genes between species, such as TACF blight-resistance introgression program. The original aim of the backcrossing breeding program was to generate hybrids that inherited ~15/16ths of their genome from American chestnut yet retained alleles for blight resistance from Chinese chestnut. Recent genomic analyses suggest that there is a tradeoff between recovery of the American chestnut genome through backcrossing and blight resistance. The results imply that blight resistance may be a polygenic trait rather than controlled by a few major effect genes and that a larger proportion of the Chinese chestnut genome may need to be introgressed into the American chestnut background to generate hybrids that have adequate blight resistance for restoration (Westbrook et al. 2019).

Multidimensional scaling (MDS) graphical visualization of genome-wide SNP data among the TACF's parental and progeny generations (Fig. 7) placed the BC1 parent tree (Clapper) intermediate in sequence composition between the Chinese chestnut genome reference tree and the several American chestnut trees sequenced, which validated the experimental approach. The MDS graphical visualization also revealed substantial SNP variation among the selected BC3F2 progeny and placed most of them intermediate between the genomes of the original Clapper BC1 genotype and the *C. dentata* backcross parents "Ort" and "Joliett" and other American chestnut trees used as recurrent parents. Thus, the TACF breeding approach is moving the composition of the genome of backcross progeny toward the American chestnut genome as desired. In several of the BC3F2 individuals, the American chestnut genome has already been largely recovered. Future studies could incorporate additional backcross tree families and/or different genotyping methods to control for the technical variance component that we observed. Identification of the exact segments in the genome of blight-resistant BC3F2s individuals that are of *C. mollissima* origin could be better resolved when an American chestnut genome is also available.

Conclusion

Our goal of a chromosome-scale genome for the *C. mollissima* cultivar Vanuxem has progressed in a long, stepwise manner through several de novo assemblies, gap closings to reduce numbers of contigs, and finally anchoring of scaffolds to the chestnut research community's reference genetic linkage map. This is an important additional resource for the chestnut research community, particularly as Vanuxem is a part of a critical restoration breeding program. Our resource complements the recently reported de novo genome assembly of a wild *C. mollissima* in China (Xing et al. 2019) published during the preparation of this paper. Our extensive validations through genetic maps and cytology was not provided for that genome and a structural comparison of the two should be completed.

Although we have high confidence in the use of genetic maps to anchor contigs, we have yet to build comprehensive pseudochromosome sequences with few gaps. Nevertheless, the value of the genome resources that we report for *C. mollissima* in research and species restoration was documented through the identification of candidate genes for disease resistance and adaptation and the evaluation of progress in the backcross breeding program for transfer of blight resistance from Chinese chestnut to American chestnut. Signatures of selection studies based on Tajima's *D* analyses of contigs placed on LG_B, LG_E, and LG_L chromosomes identified candidate genes of high effect, including a putative inositol polyphosphate-related phosphatase gene for chestnut blight resistance, G-type lectin S-receptor-like serine/threonine-protein kinases and cysteine-rich receptor-like protein kinases for Phytophthora root rot resistance, and a MADS-box transcription factor gene for bud burst phenology. These candidate genes illustrate how information can be gained through a

genome-wide approach, to be followed by validations through gene expression and functional genomics studies, and by expansion to other approaches and other genotypes. Finally, we demonstrated the potential for our *C. mollissima* genome assembly in advancing genome-enabled breeding efforts to restore the threatened American chestnut species.

Materials and methods

Plant material

Reference tree

Leaves and twig tissues were collected from the blight-resistant Chinese chestnut *Castanea mollissima* genotype Vanuxem at The American Chestnut Foundation's farm in Meadowview, VA. Collections in summer 2011 were used for the version 1.1 gDNA assembly, while tissues collected in the summers of 2014 and 2016 were used for versions 2, 3.2, and 4.2 assemblies. The Vanuxem genotype was chosen for sequencing because it is expected to remain readily available to breeders and researchers. Also, the Vanuxem genotype was used as a parent in crosses for published genetic linkage maps (Kubisiak et al. 2013) and as the source DNA for the BAC libraries used in constructing the physical map and integrated genetic-physical map for Chinese chestnut (Fang et al. 2013). The cultivar Vanuxem was also chosen as the least heterozygous (50%) among several Chinese chestnut used as donors of genes for blight resistance (ranging from 52 to 64% in observed heterozygosity) within The American Chestnut Foundation's breeding program, as determined with 25 Simple Sequence Repeat loci (Kubisiak et al. 2013). Tissue samples were immediately snap-frozen in liquid nitrogen and then stored at −80 °C. DNA was extracted for 454 and Illumina genome sequencing from bud, cambial, and leaf tissues using a modified CTAB protocol (Clarke 2009). DNA was extracted for PacBio sequencing by the Arizona Genomics Institute from 36-h dark treated (tarp-shaded) leaf samples.

Diversity panel for QTL signatures of selection analyses

Twig and leaf samples were collected, and immediately snap-frozen in liquid nitrogen, in early spring of 2015, from five *C. dentata* genotypes and six *C. mollissima* genotypes. Tissue samples were provided by the Connecticut Agricultural Experiment Station (CAES), by The American Chestnut Foundation (TACF), and by the Pennsylvania Chapter of The American Chestnut Foundation (PENN). *C. dentata* accessions included 03denGMBCLEM, 04denTFACLEM, 05denALRPENN, 06denHROPENN, and 07denELLSUNY. *C. mollissima* accessions included 14molMHGCAES, 15molNKGTACF, 23molSPPCHNA, 27molFATPENN, 29molSTVPENN, and 28molGILPENN. Two of the five Chinese chestnut genotypes are known as "Mahogany" (14molMHGCAES) and "Nanking" (15molNKGTACF), which were used as donors of resistance to *P. cinnamomi* and *C. parasitica* in backcross breeding program by TACF. DNA was extracted for Illumina HiSeq library construction from the twig and/or leaf tissues using a modified CTAB protocol (Clarke 2009; Kubisiak et al. 2013).

Genome sequencing and assembly

Genomic DNA sequencing

Over 61 billion bases of genomic DNA sequence data were produced from a combination of Illumina MiSeq and Roche 454 Next Generation Sequencing platforms. This included twenty-one 454 FLX sequencing runs, producing 25,179,431 reads averaging 516 bp in length and totaling 13,175,668,630 bp of sequence. Also 915,895,342 bp of BAC-end sequences was obtained by 454 FLX paired-end sequencing of pools of BAC clones tiling the physical map of the Vanuxem cultivar to 1.5× depth (Fang et al. 2013). In addition, 9 runs of 250-bp paired-end reads of a 480-bp insert Illumina genomic DNA library on MiSeq machines produced 41,300,000,000 bp of sequence. The chestnut physical map minimum tiling path of BAC clones were also sequenced in 2 runs of 250 bp paired-end reads on the MiSeq, producing another 4,700,000,000 bp of sequence (Fang et al. 2013). Finally, two long insert libraries averaging 3000 bp and 8000 bp were prepared for 454 FLX sequencing, yielding 897,238 and 884,030 mate-pair reads averaging 500 bp per read, totaling 890,634,000 bp of sequence for use in scaffolding. Overall, the 454 FLX and MiSeq data totaled 60,982,197,972 bp of high-quality sequence data, representing app. 76× depth of coverage of the 794 Mb genome (Kremer et al. 2007).

Draft genome assembly and scaffolding

Ten hybrid assembly builds using the Newbler assembler versions 2.5, 2.6, and 2.8 (Roche) were conducted with various amounts and combinations of 454 and MiSeq data. The best hybrid assembly was obtained from the 7th assembly, using the heterozygosity option in Newbler v2.8. The total number of input reads was 89,135,536 (covering 36,739,712,156 bp), of which 77,421,025 reads were included in the final assembly. The assembly included input of 9,096,315 Illumina MiSeq paired reads (of which 5,192,637 paired reads were assembled into the same scaffolds at an average distance of 566 bp); along with a total input of 897,238 paired 454 reads from the 3-kb insert library (of which 529,560 paired reads assembled in the same scaffolds at an average distance of

1804 bp); and a total input of 884,030 paired 454 reads from the 8-kb insert library (of which 507,004 paired reads assembled in the same scaffolds at an average distance of 6076 bp).

Final de novo genome assembly

The de novo genome assembly was improved through gap closing and contig merging prior to building chromosome-length sequences. For this, 6.8 Gb PacBio sequence data was generated by the Arizona Genomics Institute from flash-frozen etiolated leaves collected directly from the Vanuxem ramet at The American Chestnut Foundation's farm in Meadowview, VA, that was previously sampled for 454 and Illumina sequencing. Filtering of low-quality reads, removal of short reads, and sequence correction using Illumina reads and the de novo contig sequences, yielded 2 Gb of high-quality long sequence reads. The PacBio reads were error corrected using CLC Genomics Workbench (Qiagen) and pooled with high-quality consensus sequences generated from mapping 454 and MiSeq reads against a set of assembled contigs. Mapping and consensus generation were done using the CLC Read-Mapping and Consensus Sequence tools. The pooled PacBio reads and consensus contigs were meta-assembled using the overlap layout consensus algorithm of the De Novo Assembly tool in Geneious. This meta-assembly produced approximately 72 k contigs greater than 1 kb with a max of 594 kb, spanning 812 Mb. The meta-assembly in Geneious was followed by multiple rounds of contig joining/gap filling using the long-read algorithm of the Join Contigs tool in CLC Genomics Workbench. Successive rounds of contig joining were performed, where each round was used as input a set of long reads generated by re-mapping MiSeq and 454 reads against the new consensus contigs, saving unmapped reads, and using those as input for PacBio read correction. In this way, the "gaps" in the assembly were closed by segments of PacBio reads corrected by unmapped (i.e., gap-spanning) short 454 and MiSeq reads. Joined contigs were then "cleaned" by mapping a comprehensive set of Vanuxem short read data (454SE, 454MP, MiSeq, whole-genome and transcriptome data) and generating consensus sequences using a majority-rule condition to minimize ambiguities in the final sequences. Contigs were split at points where total read depth was low (<5×).

Contig anchoring

To place contigs in order along chromosomes, the sequence contigs from the final de novo assembly were first anchored to 10 linkage maps and then merged into one assembly per linkage group using LPmerge (Endelman and Plomion 2014). The 10 linkage maps include three *C. mollissima*-only maps: original consensus linkage map (Kubisiak et al. 2013), the updated linkage map (Supplementary File 1 and 2), and a map based on a F2 mapping population (Kubisiak et al. 1997) updated with additional SSR and SNP markers that were recently scored on this population. Six single parent maps (HB2_1, HB2_2, NK4_1, NK4_2, JB1_1, JB1_2) were also included. These were generated using RADseq SNP markers scored in three hybrid backcross families (HB2 = AD98 × KY115, NK4 = CG61 × NCDOT, and JB1 = "Cranberry" × JB197) (Zhebentyayeva et al. 2019). The tenth map is from the congeneric species *Q. robur* as reported by Bodénès et al. (2016)). Marker sequences from these maps were located on sequence contigs by BLASTN, and markers with more than one equivalently good location were discarded. The contigs were then anchored to their chromosomal position based on merging the maps with LPmerge using default parameters.

Chestnut cytogenetics

Chromosome preparation for cytology

Actively growing root tips were excised from Chinese chestnut seedlings growing in potting soil. The excised root tips were pretreated with an aqueous solution of a-monobromonaphthalene (0.8% *v*/v) and/or 2.5 mM 8-hydroxyquinoline for 1.5 h or 3 h, respectively, in the dark to accumulate prophase and metaphase stages for FISH, and then fixed in 4:1 (95% ethanol: glacial acetic acid). Fixed root tips were treated with cell wall-degrading enzymes (40% (v/v) Cellulase (C2730, Sigma), 20% (v/v) Pectinase (P2611, Sigma), 40% (v/v) 0.01 M citrate buffer, pH 4.5, 2% (*w*/*v*) Cellulase RS (Yakult Pharmaceutical, Tokyo, Japan), 1% (w/v) Macerozyme (Yakult Pharmaceutical), and 1.5% (w/v) Pectolyase Y23 (Kyowa C24, Chemical Co., Osaka, Japan), and the chromosome spreads were prepared as described previously (Jewell and Islam-Faridi 1994). For meiocyte pachytene chromosome spreads, emerging Chinese chestnut (genotype Vanuxem, ramet AD274) male flower buds (catkins) were harvested and placed in 3:1 ethanol:glacial acetic acid fixative and then transferred to fresh 3:1 fixative containing 1% Polyvinyl Pyrrolidone (BP431-100, Fisher Scientific, USA). Emerging anthers from the flower buds were isolated and squashed under a 22 × 22 mm glass coverslip to provide the best anthers for pachytene analysis, and then intermittently heated over an alcohol burner and tapped with forceps to spread the chromosomes. Slides with good chromosome spreads were stored at −80 °C for future use.

Probe labeling and FISH

Probe DNAs (BAC clones from the Chinese chestnut physical map (Fang et al. 2013, Table S1) including 18S-28S rDNA and 5S rDNA probes were labeled with biotin-16-dUTP (Biotin Nick Translation Mix, Roche, USA) and/or

digoxigenin-11-dUTP (Dig Nick Translation Mix, Roche, USA) following the manufacturer's instructions.

Fluorescence in situ hybridization (FISH)

FISH with BAC clones on somatic metaphase of different cultivars of Chinese chestnut and pachytene chromosome spreads of the Chinese chestnut Vanuxem genotype was carried out as described previously (Jewell and Islam-Faridi 1994; Islam-Faridi et al. 2002, 2009). Probe hybridization sites were detected with Cy3-conjugated streptavidin (Jackson ImmunoResearch Laboratories, USA) for biotin-labeled probes and FITC-conjugated anti-digoxygenin (Roche, USA) for digoxigenin-labeled probes. The FISH preparations were mounted with Vectashield containing DAPI (Vector Laboratories, USA) to prevent photobleaching the fluorochromes. Digital images were recorded using an epi-fluorescence microscope (AxioImager M2, Carl Zeiss, Germany) with suitable filter sets (Chroma Technology, USA) and a Cool Cube high-performance CCD camera, and processed with ISIS V5.1 (MetaSystem Inc., USA) and Adobe Photoshop CS v8 (Adobe System, USA).

Repeat masking and repetitive sequence annotations

Repeat masking was performed by first running RepeatModeler v1.0.10 using "ncbi" as the engine to identify novel repeats (Smit et al. 2015). Ribosomal RNAs were removed from the RepeatModeler output. Remaining novel repeats were concatenated with the RepBase plant repeat library (Bao et al. 2015), and input to RepeatMasker v4.0.6 with the parameters to ignore low complexity regions (−nolow) and to softmask repeats from the contigs (−xsmall) (Smit et al. 2015). The "ProcessRepeats" software that comes with RepeatMasker was used to correctly classify the repeats from the RepBase plant library with the flagship species eudicotyledons.

RNA sequencing, filtering and QC

Total RNA was extracted from leaf, petiole, twig, bark, and root tissues of the Vanuxem Chinese chestnut reference cultivar, from twig and leaf samples of the Nanking Chinese chestnut cultivar, and from roots of 25 cultivar Nanking seedlings which had been challenged with the *Phytophthora cinnamomi* root-rot pathogen, using either Spectrum™ Plant Total RNA Kit (Sigma-Aldrich) or Qiagen RNeasy protocol (Barakat et al. 2009). Total RNA samples were converted to HiSeq Illumina cDNA libraries using the TruSeq® Stranded Total RNA Library Prep kit for Plants. Quality checks of RNA and library preparations were conducted by micro-capillary electrophoresis on the 2100 BioAnalyzer (Agilent Genomics). BioAnalyzer RIN value of 8.0 was used as minimum quality scores. The RNAseq libraries were selected for 250 bp insert sizes, and pooled prior to multiplex sequencing on a Illumina HiSeq 2000, producing 150-bp paired-end reads. Sequencing was conducted in rapid mode at the Pennsylvania State University Genomics Core Facility. Reads were demultiplexed into separate forward and reverse read FASTQ files using the barcoded adaptor sequences. The RNAseq reads were trimmed of adaptors and further filtered for base quality using the CLC Genomics Workbench (Qiagen) tools, selecting reads with a minimum quality score of 0.01 and minimum length of 100 bp. Two to three gigabases of high-quality RNAseq data was produced in each direction for each of the libraries. The filtered reads were aligned to the Vanuxem reference genome to check for alignment quality using STAR v2.5.3a with default alignment parameters (Dobin et al. 2013) prior to use in gene model predictions and validations.

Gene prediction

BRAKER2 was used to identify gene models in the contigs (Hoff et al. 2016). It was run with the soft-masked version of the contigs using two lines of evidence for training: alignments of *C. mollissima* RNAseq reads and alignments of the protein sequences from the *Q. robur* genome version PM1N, selected because it is the closest fully sequenced reference genome with high-quality gene models (Plomion et al. 2018). The RNAseq reads were aligned with STAR v2.5.3a using default alignment parameters (Dobin et al. 2013), and the proteins were aligned to the chestnut contigs with GenomeThreader v1.7.0 using default alignment parameters (Gremme et al. 2005).

Gene model predictions were sorted into high quality and low quality by two criteria. First, expression evidence was examined. Using the gff file from BRAKER and the RNAseq alignments from STAR, the number of reads per predicted gene was assessed using HTSeq (Anders et al. 2015). Next, homology to *Q. robur* proteins was determined. A reciprocal BLASTp analysis was run to identify chestnut genes with a likely ortholog in *Q. robur* (Camacho et al. 2009). Any gene model with at least 100 RNAseq aligned reads and/or a reciprocal best hit to a *Q. robur* gene was retained as high quality; the rest were placed in the low-quality category.

Quality assessment and functional annotation

Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis was used to compare the gene model to 1440 common orthologs across all embryophytes (Simão et al. 2015). To predict function, genes were annotated by BLAST searches (Camacho et al. 2009) to SWISS-PROT and TrEMBL protein databases (e-value <1e−5) (Bairoch and

Apweiler 1998), InterProScan sequence searches with Gene Ontology results parameter (Jones et al. 2014), and ghostKOALA searches against the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2016).

Comparative genome structure analyses

Orthogroups were built by OrthoFinder v2.2.7 (Emms and Kelly 2015) using the chestnut proteins and proteins from *Arabidopsis thaliana* version TAIR10 (Lamesch et al. 2012), *Prunus persica* v2.1 (Verde et al. 2017), *Populus trichocarpa* v3.1 (Tuskan et al. 2006), *Vitis vinifera* v2.1 (Characterization TFPCFGG, The French–Italian Public Consortium for Grapevine Genome Characterization 2007), and *Quercus robur* version PM1N (Characterization TFPCFGG, The French–Italian Public Consortium for Grapevine Genome Characterization 2007; Plomion et al. 2018). The first four were downloaded from Phytozome (Goodstein et al. 2012), and the latter from the Oak Genome Sequencing website (Plomion et al. 2018). Orthogroups containing genes from the monolignol biosynthetic pathway were flagged if they contained the genes annotated from that pathway reported in *Arabidopsis thaliana* (Raes et al. 2003) and/or *Populus trichocarpa* (Shi et al. 2010). The genes from poplar were converted from genome annotation version 1.1 to genome annotation version 3.1 by searching for the older gene name in Phytozome, and if not found, BLAST against the new version 3.1 genes. The version 3.1 gene was only accepted for BLAST results with at least 95% identity. NBS-LRR genes were identified by searching the chestnut genes with the Pfam model NB-ARC (PF00931.22) with HMMER v3.2.1 (Eddy 2011).

Circos plots were built using Circos version 0.69–6 (Krzywinski et al. 2009). The Circos map of contigs to the Kubisiak reference genetic map (Kubisiak et al. 2013) used the same BLAST results used to anchor the contigs to chromosomes. The Circos map of contigs to the *Quercus robur* and *Prunus persica* genomes were built using orthologs. The orthologs were identified by orthogroups with a single chestnut member gene and a single other species member gene in order to exclude gene families. Further filtering of the orthologs was performed to ensure only linkages with at least two points of agreement were retained. This was done by examining each ortholog in chestnut against the closest upstream and downstream ortholog. If either the upstream or downstream ortholog did not match the target genome on the same chromosome within 10 Mb, the ortholog was discarded.

QTL selection signal analysis methods

Next-generation Illumina sequence was conducted as described in Genomic DNA sequencing. Sequences of five accessions representing each American and Chinese chestnut species were aligned against reference Chinese chestnut (Vanuxem) genome assembly v.3.2 as described below in the Materials and methods section Assessment of progress of the back-cross breeding program on recovery of American chestnut genome. Sorted and indexed bam files were generated for each linkage group and "unmapped" contigs individually. The ANGSD software version 0.920 was used to calculate the allele frequency spectrum, obtain a maximum likelihood estimate of the unfolded site frequency spectrum (SFS) (Nielsen et al. 2012), estimate pairwise nucleotide diversity, and perform tests for selection based on Tajima's *D* coefficient (Tajima 1989), which compares the number of pairwise differences to the number of segregating sites (Korneliussen et al. 2014). Population genetic statistics were estimated for sliding 5-kb windows along each linkage group with a step size of 1 kb. A window of 5 kb was selected based on the average size of gene models in a gff3 file. Output summary tables generated for *C. mollissima* and *C. dentata* were used to export Tajima's *D* statistics and to calculate integrative indices of nucleotide diversity, i.e., ratios of πCden/πCmol and πCmol/πCden. We considered windows with πCden/πCmol in the upper 0.5% of the empirical distribution across 12 linkage groups as candidate regions under selection. QTL intervals for traits of interest (bud emergence, resistance to *P. cinnamomi* and *C. parasitica*) were delineated using sequence-based markers associated with QTLs (Zhebentyayeva et al. 2019; and Kubisiak et al. 2013, respectively) and local BLAST® (https://www.ncbi.nlm.nih.gov/books/NBK279690/) against the *C. mollissima* v3.2 contigs. Results were plotted using *ggplot2* in R (Wilkinson 2011).

To identify regions of homology among chestnut, oak, and peach, the DAM gene from peach, along with other surrounding genes were BLASTed against the oak (version PM1N) (Plomion et al. 2018) and peach (v1.0-r1) gene sets (International Peach Genome Initiative et al. 2013). Peach genome v1.0-r1 from Phytozome was used for this analysis because it was manually annotated for the DAM genes and represents the most accurate gene model annotation of the DAM region. Where homology was identified that was not among the originally annotated gene models, the region was manually annotated using the chestnut protein and fgenesh+ (Solovyev 2004).

Assessment of progress of the backcross breeding program on recovery of American chestnut genome

DNA was extracted using a modified CTAB protocol (Clarke 2009) from twig and/or leaf tissues from the following individuals: a clone of the BC_1 "Clapper" which was a source of blight resistance that The American Chestnut Foundation backcross breeding program started with (Hebard 2005); 48 BC3F2 progeny from a reciprocal intercross of two BC_3F_1 trees with pedigrees "Joliet" x GR210 and "Ort" x CL287 (NB: "Ort" and "Joliet" were American chestnut trees, while GR210 and

CL287 were BC2 trees from crosses of American chestnut trees with the Clapper BC1 tree); the American chestnut parent trees "Ort" and "Joliet;" six wild American chestnut trees (Carbaugh, Hemple, Manes, Michaux, Stone Valley, and Glade Mountain Big); the susceptible American chestnut great-grandparent tree LFR4T14; and the *C. mollissima* reference individual Vanuxem. The Cornell University Biotechnology Resource Center constructed RADseq libraries for each sample using single restriction enzyme (*Pst*I) digests, and then produced single-end 100-bp 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.

Raw reads were trimmed by skewer (Jiang et al. 2014) with 30 bp as a minimum read length. Trimmed reads were then aligned to *C. mollissima* reference genome v3.2 using BWA-MEM in Burrows-Wheeler Aligner (Li and Durbin 2009). MarkDuplicate from Picard v2.20.2 broadinstitute 2019) was used to remove PCR artifacts. Remaining mapped reads were input into Genome Analysis Tool Kit (McKenna et al. 2010) to call variants (Auwera et al. 2013). To reduce the false positive variants, strand bias was filtered out using "VariantFiltration" in GATK on the variant with a FisherStrand score of more than 30, with a sliding window of 35 bp and a cluster size of 3 SNPs. Multidimensional scaling (MDS) was generated by VCFtools-0.1.14 (Danecek et al. 2011) and PLINK-1.07 (Purcell et al. 2007) to graphically display the SNP variation results among samples. Comparison of SNPs shared among individuals and species was assessed using vcf-compare from vcftools v1.16 (Danecek et al. 2011).

Acknowledgements This project was funded by the Forest Health Initiative (https://foresthealthinitiative.org/) through grant # 137RFP# 2008-011 to JEC. Support was also provided by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture grant 2016-67013-24581 to The American Chestnut Foundation. Additional support was provided through several grants-in-aid to JEC from The American Chestnut Foundation and to JEC and MES through the USDA National Institute of Food and Agriculture Federal Appropriations under Project PEN04532 (Accession number 1000326) and NE-1833, respectively. Construction of saturated genetic maps and root RNAseq dataset was partially supported by the Foundation for the Carolinas. Bioinformatics was supported by National Science Foundation (NSF) Award #1444573, "Standards and Cyberinfrastructure that Enable 'Big-Data' Driven Discovery for Tree Crop Research" (MES; PI Main). Emily Bellis was supported by NSF Postdoctoral Research Fellowships in Biology Grant No. 1711950. Rooksana Noorai was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM109094.

We would like to thank Webb Miller, Ronald Sederoff, John Davis, Claude dePamphilis, Joshua Der, and Nicholas Wheeler for their enthusiastic guidance and assistance. We thank Qi Sun, Computational Biology Service Unit, Life Sciences Core Laboratories Center at Cornell University for conducting the initial MDS analysis.

Authors' contributions Margaret Staton contributed to bioinformatics, data analysis, and website development activities, including gene functional and structural annotation and comparative genomics, as well as being a major contributor to project design, obtaining grant support, and writing the manuscript; Charles Addo-Quaye conducted the sequence data QC and did all of the initial de novo sequence assemblies with Illumina and 454 sequence data and post-assembly analyses resulting in version 1.1 released to the public in January 2014; Nathaniel Cannon planned and conducted the hybrid de novo genome assemblies, conducted manual contig gap filling and merging, conducted the initial contig anchoring and whole-genome analyses, and participated in preparation of the manuscript; Jiali Yu assisted in anchoring scaffolds to chromosome positions, comparative genomics, and RNA sequence mapping and assembly and in manuscript writing; Tetyana Zhebentyayeva analyzed physical map contigs, selected BACs for FISH, constructed six saturated genetic maps for genome assembly, delineated QTL intervals for resistance to *P. cinnamomi*, participated in Tajima's *D* and nucleotide diversity analyses, identified candidate genes, and contributed to writing; Matthew Huff contributed to the gene functional analysis and structural annotation and comparative genomics; Nurul Islam-Faridi conducted all in situ hybridization experiments and analysis; Shenghua Fan provided reference genetic map for genome assembly and QTL interval analysis for blight resistance in linkage group B for the coevolution study; Laura L. Georgi provided plant materials and phenotypic data, conducted crosses and inoculations, and was a PI on the USDA grant; C. Dana Nelson performed contig anchoring to the genetic maps, supervised the chestnut genetic linkage mapping and QTL analyses, and assisted in writing the manuscript; Emily Bellis produced and analyzed Tajima's *D* data; Nathan Henry provided bioinformatics support, assisted in the website development and curation, and conducted the data analyses for figures and tables in the manuscript; Daniela I Drautz-Moses conducted the DNA and RNA sequencing and advised on sample collection and sequence analysis; Rooksana Noorai assembled and annotated the chestnut root transcriptome, leading to candidate gene identification; Stephen Ficklin provided data and analyses for the integrated genetic-physical map, BAC-end sequences, which led to the initial genome assemblies, as well as assisting in initial web portal development; Christopher Saski led BAC library construction, provided BAC-end sequence data, and provided oversight of and insights from the integrated genetic-physical map construction; Mihir Mandal prepared multi-tissue RNAs from Chinese chestnut seedlings, constructed cDNA libraries for sequencing, and assisted in RNA sequence data analyses; Tyler K Wagner and Nicole Zembower provided technical support in the lab and field throughout the project; Catherine Bodénès performed genetic linkage mapping and QTL analyses for bud burst on *Quercus*; Jason Holliday provided project design and RNA sequencing resources; Jared Westbrook led the USDA project supporting the hybrid de novo assembly and gap filling, as well as providing input on current TACF breeding and chestnut restoration efforts; Jesse Lasky helped plan analyses and contributed to interpreting results and writing the paper; Frederick Hebard led the TACF backcross breeding program, provided plant materials and phenotypic data, conducted crosses and inoculations, and was PI on the USDA grant; Stephan Schuster contributed to planning the sequencing and assembly approach, supervised the DNA and RNA sequencing, and assisted in proposal writing; Albert G Abbott advised on genetic mapping, participated in Tajima's *D* analyses, organized the writing team, and made major contributions to manuscript preparation; John E Carlson obtained funding, provided overall project design and leadership, contributed to data analyses, and helped to prepare the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Competing interests The authors declare that they have no competing interests.

Data archiving statement The chestnut genome versions 1.1, 3.2, and 4.2 are available at https://hardwoodgenomics.org. The website contains links to download the contigs, the anchored contig locations, and predicted genes, transcripts, and proteins and associated functional annotations. A J-Browse implementation for the whole genome is located at the URL https://hardwoodgenomics.org/tools/jbrowse/?data=chinese_chestnut. The raw and assembled sequences are also available at the NCBI BioProject No. PRJNA46687.

References

Abbott AG, Zhebentyayeva T, Barakat A, Liu Z (2015) The genetic control of bud-break in trees. Adv Bot Res:201–228

Anagnostakis SL (2012) Chestnut breeding in the United States for disease and insect resistance. Plant Dis 96:1392–1403

Anders S, Pyl PT, Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169

Anwar A, She M, Wang K, Riaz B, Ye X (2018) Biological roles of ornithine aminotransferase (OAT) in plant stress tolerance: present progress and future perspectives. Int J Mol Sci 19. https://doi.org/10.3390/ijms19113681

Aranzana MJ, Decroocq V, Dirlewanger E, Eduardo I, Gao ZS, Gasic K, Iezzoni A, Jung S, Peace C, Prieto H, Tao R, Verde I, Abbott AG, Arús P (2019) *Prunus* genetics and applications after de novo genome sequencing: achievements and prospects. Horticulture Research 6:58

Arentz F (2017) Phytophthora cinnamomi A1: an ancient resident of New Guinea and Australia of Gondwanan origin? For Pathol 47:e12342

Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43

Bacete L, Mélida H, Miedes E, Molina A (2018) Plant cell wall-mediated immunity: cell wall changes trigger disease resistance responses. Plant J 93:614–636

Baier K, Maynard C, Powell W (2012) Early flowering in chestnut species induced under high-Intensity, high-dose light in growth chambers. J Amer Chest Found 26:8–10

Bairoch A, Apweiler R (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res 26: 38–42

Bao W, Kojima KK, Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11

Barakat A, DiLoreto DS, Zhang Y et al (2009) Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. BMC Plant Biol 9:51

Bielenberg DG, Wang Y(E), Li Z et al (2008) Sequencing and annotation of the evergrowing locus in peach [*Prunus persica* (L.) Batsch] reveals a cluster of six MADS-box transcription factors as candidate genes for regulation of terminal bud formation. Tree Genet Genomes 4:495–507

Bodénès C, Chancerel E, Gailing O, Vendramin GG, Bagnoli F, Durand J, Goicoechea PG, Soliani C, Villani F, Mattioni C, Koelewijn H, Murat F, Salse J, Roussel G, Boury C, Alberto F, Kremer A, Plomion C (2012) Comparative mapping in the Fagaceae and beyond with EST-SSRs. BMC Plant Biol 12:153

Bodénès C, Chancerel E, Ehrenmann F, Kremer A, Plomion C (2016) High-density linkage mapping and distribution of segregation distortion regions in the oak genome. DNA Res 23:115–124

broadinstitute broadinstitute/picard. In: GitHub. https://github.com/broadinstitute/picard. Accessed 19 Dec 2019

Cahill DM, McComb JA (1992) A comparison of changes in phenylalanine ammonia-lyase activity, lignin and phenolic synthesis in the roots of *Eucalyptus calophylla* (field resistant) and *E. marginata* (susceptible) when infected with *Phytophthora cinnamomi*. Physiol Mol Plant Pathol 40:315–332

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Campoy JA, Ruiz D, Egea J, Rees DJG, Celton JM, Martínez-Gómez P (2011) Inheritance of flowering time in apricot (*Prunus armeniaca* L.) and analysis of linked quantitative trait loci (QTLs) using simple sequence repeat (SSR) markers. Plant Mol Biol Report 29:404–410

Casasoli M, Derory J, Morera-Dutrey C, Brendel O, Porth I, Guehl JM, Villani F, Kremer A (2006) Comparison of quantitative trait loci for adaptive traits between oak and chestnut based on an expressed sequence tag consensus map. Genetics 172:533–546

Characterization TFPCFGG, The French–Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Clarke JD (2009) Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. Cold Spring Harb Protoc 2009: db.prot5177

Cooke JEK, Eriksson ME, Junttila O (2012) The dynamic nature of bud dormancy in trees: environmental control and molecular mechanisms. Plant Cell Environ 35:1707–1728

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Delgado-Cerrone L, Alvarez A, Mena E, Ponce de León I, Montesano M (2018) Genome-wide analysis of the soybean CRK-family and transcriptional regulation by biotic stress signals triggering plant immunity. PLoS One 13:e0207438

Derory J, Scotti-Saintagne C, Bertocchi E, le Dantec L, Graignic N, Jauffres A, Casasoli M, Chancerel E, Bodenes C, Alberto F, Kremer A (2010) Contrasting relations between diversity of candidate genes and variation of bud burst in natural and segregating populations of European oaks. Heredity 105:401–411

Diskin M, Steiner KC, Hebard FV (2006) Recovery of American chestnut characteristics following hybridization and backcross breeding to restore blight-ravaged *Castanea dentata*. For Ecol Manag 223: 439–447

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157

Endelman JB, Plomion C (2014) LPmerge: an R package for merging genetic maps by linear programming. Bioinformatics 30:1623–1624

Engelbrecht J, van den Berg N (2013) Expression of defence-related genes against Phytophthora cinnamomi in five avocado rootstocks. S Afr J Sci 109:1–8

Fan S, Bielenberg DG, Zhebentyayeva TN, Reighard GL, Okie WR, Holland D, Abbott AG (2010) Mapping quantitative trait loci associated with chilling requirement, heat requirement and bloom date in peach (*Prunus persica*). New Phytol 185:917–930

Fan S, Georgi L, Hebard FV, et al (2020) Mapping QTLs for blight resistance and morphological and phenological traits in chestnut (*Castanea* spp.). (in prep)

Fang G-C, Blackmon BP, Staton ME, Nelson CD, Kubisiak TL, Olukolu BA, Henry D, Zhebentyayeva T, Saski CA, Cheng CH, Monsanto M, Ficklin S, Atkins M, Georgi LL, Barakat A, Wheeler N, Carlson

JE, Sederoff R, Abbott AG (2013) A physical map of the Chinese chestnut (*Castanea mollissima*) genome and its integration with the genetic map. Tree Genet Genomes 9:525–537

Freinkel S (2009) American chestnut: the life, death, and rebirth of a perfect tree. Univ of California Press

Gabay G, Dahan Y, Izhaki Y, Faigenboim A, Ben-Ari G, Elkind Y, Flaishman MA (2018) High-resolution genetic linkage map of European pear (*Pyrus communis*) and QTL fine-mapping of vegetative budbreak time. BMC Plant Biol 18:175

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186

Gremme G, Brendel V, Sparks ME, Kurtz S (2005) Engineering a software tool for gene structure prediction in higher organisms. Inf Softw Technol 47:965–978

Groover A, Cronk Q (eds) (2017) Comparative and evolutionary genomics of angiosperm trees. Springer, Cham

Hamann T (2015) The plant cell wall integrity maintenance mechanism—concepts for organization and mode of action. Plant Cell Physiol 56:215–223

Hebard FV (1994) Inheritance of juvenile leaf and stem morphological traits in crosses of Chinese and American chestnut. J Hered 85:440–446

Hebard FV (2005) The backcross breeding program of the American Chestnut Foundation. In Proc. of Restoration of American Chestnut to Forest Lands Conference. Steiner, K.C. and J.E. Carlson (eds.)

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769

Hung C-Y, Aspesi P Jr, Hunter MR et al (2014) Phosphoinositide-signaling is one component of a robust plant defense response. Front Plant Sci 5:267

International Peach Genome Initiative, Verde I, Abbott AG et al (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45:487–494

Islam-Faridi MN, Childs KL, Klein PE, Hodnett G, Menz MA, Klein RR, Rooney WL, Mullet JE, Stelly DM, Price HJ (2002) A molecular cytogenetic map of sorghum chromosome 1. Fluorescence *in situ* hybridization analysis with mapped bacterial artificial chromosomes. Genetics 161:345–353

Islam-Faridi MN, Nelson CD, DiFazio SP et al (2009) Cytogenetic analysis of *Populus trichocarpa*–ribosomal DNA, telomere repeat sequence, and marker-selected BACs. Cytogenet Genome Res 125:74–80

Jewell DC, Islam-Faridi N (1994) A technique for somatic chromosome preparation and C-banding of maize. The Maize Handbook:484–493

Jiang H, Lei R, Ding S-W, Zhu S (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15:182

Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36:64–70

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240

Kanehisa M, Sato Y, Morishima K (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428:726–731

Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, Martinoia E (2011) Plant ABC transporters. Arabidopsis Book 9:e0153

Kaye Y, Golani Y, Singer Y, Leshem Y, Cohen G, Ercetin M, Gillaspy G, Levine A (2011) Inositol polyphosphate 5-phosphatase7 regulates the production of reactive oxygen species and salt tolerance in *Arabidopsis*. Plant Physiol 157:229–241

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15:356

Kremer A, Casasoli M, Barreneche T et al (2007) Comparative genetic mapping in Fagaceae. In: Kole CR (ed) Genome Mapping & Molecular Breeding in plants, Vol. 7: Forest trees. Springer, Heidelberg, pp 161–187

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645

Kubisiak TL, Hebard FV, Nelson CD, Zhang J, Bernatzky R, Huang H, Anagnostakis SL, Doudrick RL (1997) Molecular mapping of resistance to blight in an interspecific cross in the genus castanea. Phytopathology 87:751–759

Kubisiak TL, Nelson CD, Staton ME, Zhebentyayeva T, Smith C, Olukolu BA, Fang GC, Hebard FV, Anagnostakis S, Wheeler N, Sisco PH, Abbott AG, Sederoff RR (2013) A transcriptome-based genetic map of Chinese chestnut (Castanea mollissima) and identification of regions of segmental homology with peach (Prunus persica). Tree Genet Genomes 9:557–571

LaBonte NR, Zhao P, Woeste K (2018) Signatures of selection in the genomes of Chinese chestnut (*Castanea mollissima* Blume): the roots of nut tree domestication. Front Plant Sci 9

Labuschagné IF, Louw JH, Schmidt K, Sadie A (2003) Budbreak number in apple seedlings as selection criterion for improved adaptability to mild winter climates. HortScience 38:1186–1190

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The Arabidopsis information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1202–D1210

Lang P, Dane F, Kubisiak TL, Huang H (2007) Molecular evidence for an Asian origin and a unique westward migration of species in the genus *Castanea* via Europe to North America. Mol Phylogenet Evol 43:49–59

Lee DS, Kim YC, Kwon SJ, Ryu CM, Park OK (2017) The Arabidopsis cysteine-rich receptor-like kinase CRK36 regulates immunity through interaction with the cytoplasmic kinase BIK1. Front Plant Sci 8:1856

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Liu Z, Zhu H, Abbott A (2015) Dormancy behaviors and underlying regulatory mechanisms: from perspective of pathways to epigenetic regulation. Advances in Plant Dormancy 75–105

Luo M-C, You FM, Li P, Wang JR, Zhu T, Dandekar AM, Leslie CA, Aradhya M, McGuire PE, Dvorak J (2015) Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. BMC Genomics 16:707

Madoui M-A, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM (2015) Genome assembly using nanopore-guided long and error-free DNA reads. BMC Genomics 16:327

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Miedes E, Vanholme R, Boerjan W, Molina A (2014) The role of the secondary cell wall in plant resistance to pathogens. Front Plant Sci 5:358

Naveed ZA, Huguet-Tapia JC, Ali GS (2019) Transcriptome profile of Carrizo citrange roots in response to *Phytophthora parasitica* infection. J Plant Interact 14:187–204

Nelson CD, Powell WA, Maynard CA, et al (2014) the forest health initiative, american chestnut (*castanea dentata*) as a model for forest tree restoration: biological research program. acta horticulturae 179–189

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One 7:e37558

Olukolu BA, Nelson CD, Abbott AG (2012) Mapping resistance to Phytophthora cinnamomi in chestnut (Castanea sp.). In: In: Sniezko, Richard A.; Yanchuk, Alvin D.; Kliejunas, John T.; Palmieri, Katharine M.; Alexander, Janice M.; Frankel, Susan J., tech. coords. Proceedings of the fourth international workshop on the genetics of host-parasite interactions in forestry: Disease and insect resistance in forest trees. Gen. Tech. Rep. PSW-GTR-240. Albany, CA: Pacific Southwest Research Station, Forest Service, US Department of Agriculture. p. 177. p 177

Pereira-Lorenzo S, Costa R, Anagnostakis S, et al (2016) Interspecific hybridization of chestnut. Polyploidy and hybridization for crop improvement Boca Raton 377–407

Plomion C, Aury J-M, Amselem J, Leroy T, Murat F, Duplessis S, Faye S, Francillonne N, Labadie K, le Provost G, Lesur I, Bartholomé J, Faivre-Rampant P, Kohler A, Leplé JC, Chantret N, Chen J, Diévart A, Alaeitabar T, Barbe V, Belser C, Bergès H, Bodénès C, Bogeat-Triboulot MB, Bouffaud ML, Brachi B, Chancerel E, Cohen D, Couloux A, da Silva C, Dossat C, Ehrenmann F, Gaspin C, Grima-Pettenati J, Guichoux E, Hecker A, Herrmann S, Hugueney P, Hummel I, Klopp C, Lalanne C, Lascoux M, Lasserre E, Lemainque A, Desprez-Loustau ML, Luyten I, Madoui MA, Mangenot S, Marchal C, Maumus F, Mercier J, Michotey C, Panaud O, Picault N, Rouhier N, Rué O, Rustenholz C, Salin F, Soler M, Tarkka M, Velt A, Zanne AE, Martin F, Wincker P, Quesneville H, Kremer A, Salse J (2018) Oak genome reveals facets of long lifespan. Nat Plants 4:440–452

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Raaymakers TM, Van den Ackerveken G (2016) Extracellular recognition of Oomycetes during biotrophic infection of plants. Front Plant Sci 7:906

Raes J, Rohde A, Christensen JH, van de Peer Y, Boerjan W (2003) Genome-wide characterization of the lignification toolbox in Arabidopsis. Plant Physiol 133:1051–1071

Ramos AM, Usié A, Barbosa P, Barros PM, Capote T, Chaves I, Simões F, Abreu I, Carrasquinho I, Faro C, Guimarães JB, Mendonça D, Nóbrega F, Rodrigues L, Saibo NJM, Varela MC, Egas C, Matos J, Miguel CM, Oliveira MM, Ricardo CP, Gonçalves S (2018) The draft genome sequence of cork oak. Sci Data 5:180069

Ribeiro T, Loureiro J, Santos C, Morais-Cecílio L (2011) Evolution of rDNA FISH patterns in the Fagaceae. Tree Genet Genomes 7:1113–1122

Robinson SM, Bostock RM (2014) β-Glucans and eicosapolyenoic acids as MAMPs in plant-oomycete interactions: past and present. Front. Plant Sci 5:797

Santos C, Nelson CD, Zhebentyayeva T, Machado H, Gomes-Laranjo J, Costa RL (2017) First interspecific genetic linkage map for *Castanea sativa* x *Castanea crenata* revealed QTLs for resistance to *Phytophthora cinnamomi*. PLoS One 12:e0184381

Scotti-Saintagne C, Bodénès C, Barreneche T et al (2004) Detection of quantitative trait loci controlling bud burst and height growth in *Quercus robur* L. Theor Appl Genet 109:1648–1659

Serrazina S, Santos C, Machado H, Pesquita C, Vicentini R, Pais MS, Sebastiana M, Costa R (2015) Castanea root transcriptome in response to *Phytophthora cinnamomi* challenge. Tree Genet Genomes 11

Shi R, Sun Y-H, Li Q, Heber S, Sederoff R, Chiang VL (2010) Towards a systems approach for lignin biosynthesis in *Populus trichocarpa*: transcript abundance and specificity of the monolignol biosynthetic genes. Plant Cell Physiol 51:144–163

Shim D, Ko J-H, Kim W-C, Wang Q, Keathley DE, Han KH (2014) A molecular framework for seasonal growth-dormancy regulation in perennial plants. Hortic Res 1:14059

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Smit AFA, Hubley R, Green P (2015) RepeatMasker Open-4.0. 2013–2015

Solovyev V (2004) Statistical approaches in eukaryotic gene prediction. Handbook of Statistical Genetics

Staton M, Zhebentyayeva T, Olukolu B, Fang GC, Nelson D, 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 16:744

Steiner KC, Westbrook JW, Hebard FV, Georgi LL, Powell WA, Fitzsimmons SF (2017) Rescue of American chestnut with extraspecific genes following its destruction by a naturalized pathogen. New For 48:317–336

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tauzin AS, Giardina T (2014) Sucrose and invertases, a part of the plant defense response to the biotic stresses. Front Plant Sci 5:293

Teixeira MA, Rajewski A, He J, Castaneda OG, Litt A, Kaloshian I (2018) Classification and phylogenetic analyses of the *Arabidopsis* and tomato G-type lectin receptor kinases. BMC Genomics 19:239

Tennessen JA, Madeoy J, Akey JM (2010) Signatures of positive selection apparent in a small sample of human exomes. Genome Res 20:1327–1334

Toljamo A, Blande D, Kärenlampi S, Kokko H (2016) Reprogramming of strawberry (*Fragaria vesca*) root transcriptome in response to *Phytophthora cactorum*. PLoS One 11:e0161078

Tuskan GA, Difazio S, Jansson S et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & gray). Science 313:1596–1604

Tuskan GA, Groover AT, Schmutz J, DiFazio SP, Myburg A, Grattapaglia D, Smart LB, Yin T, Aury JM, Kremer A, Leroy T, le Provost G, Plomion C, Carlson JE, Randall J, Westbrook J, Grimwood J, Muchero W, Jacobson D, Michener JK (2018) Hardwood tree genomics: unlocking woody plant biology. Front Plant Sci 9:1799

Vaattovaara A, Brandt B, Rajaraman S, Safronov O, Veidenberg A, Luklová M, Kangasjärvi J, Löytynoja A, Hothorn M, Salojärvi J, Wrzaczek M (2019) Mechanistic insights into the evolution of DUF26-containing proteins in land plants. Commun Biol 2:56

van den Berg N, Christie JB, Aveling TAS, Engelbrecht J (2018) Callose and β-1,3-glucanase inhibit *Phytophthora cinnamomi* in a resistant avocado rootstock. Plant Pathol 67:1150–1160

Veillet F, Gaillard C, Coutos-Thévenot P, La Camera S (2016) Targeting the AtCWIN1 gene to explore the role of invertases in sucrose transport in roots and during Botrytis cinerea infection. Front Plant Sci 7

Verde I, Jenkins J, Dondini L, et al (2017) The peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. BMC Genomics 18

Westbrook JW, Zhang Q, Mandal MK, et al (2019) Genomic selection analyses reveal tradeoff between chestnut blight tolerance and genome inheritance from American chestnut (*Castanea dentata*) in (*C. dentata*x Prunus) x *C. dentata* backcross populations

Wilkinson L (2011) ggplot2: elegant graphics for data analysis by WICKHAM, H. Biometrics 67:678–679

Williams SP, Gillaspy GE, Perera IY (2015) Biosynthesis and possible functions of inositol pyrophosphates in plants. Front Plant Sci 6:67

Xing Y, Liu Y, Zhang Q, Nie X, Sun Y, Zhang Z, Li H, Fang K, Wang G, Huang H, Bisseling T, Cao Q, Qin L (2019) Hybrid de novo genome assembly of Chinese chestnut (*Castanea mollissima*). Gigascience 8. https://doi.org/10.1093/gigascience/giz112

Zentmyer GA (1988) Origin and distribution of four species of Phytophthora. Trans Br Mycol Soc 91:367–378

Zhebentyayeva T, Chandra A, Abbott AG, et al (2012) Genetic and genomic resources for mapping resistance to *Phytophthora cinnamomi* in chestnut. In: V International Chestnut Symposium 1019. pp 263–270

Zhebentyayeva TN, Sisco PH, Georgi LL, Jeffers SN, Perkins MT, James JB, Hebard FV, Saski C, Nelson CD, Abbott AG (2019) Dissecting resistance to *Phytophthora cinnamomi* in interspecific hybrid chestnut crosses using sequence-based genotyping and QTL mapping. Phytopathology 109:1594–1604

Affiliations

Margaret Staton[1] · Charles Addo-Quaye[2,3] · Nathaniel Cannon[2,4] · Jiali Yu[5] · Tetyana Zhebentyayeva[2] · Matthew Huff[1] · Nurul Islam-Faridi[6,7] · Shenghua Fan[8] · Laura L. Georgi[8,9] · C. Dana Nelson[8,10] · Emily Bellis[11] · Sara Fitzsimmons[9] · Nathan Henry[1] · Daniela Drautz-Moses[12] · Rooksana E. Noorai[13] · Stephen Ficklin[14] · Christopher Saski[15] · Mihir Mandal[16,17] · Tyler K. Wagner[2] · Nicole Zembower[2] · Catherine Bodénès[18] · Jason Holliday[16] · Jared Westbrook[19] · Jesse Lasky[11] · Frederick V. Hebard[9] · Stephan C. Schuster[12] · Albert G. Abbott[2,8] · John E. Carlson[2]

1 Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, TN 37996, USA

2 Schatz Center for Tree Molecular Genetics, Pennsylvania State University, University Park, PA 16802, USA

3 Division of Natural Sciences and Mathematics, Lewis-Clark State College, Lewiston, ID 83501, USA

4 Department of Biology, Southern Utah University, Cedar City, UT 84322, USA

5 Genome Science and Technology, University of Tennessee, Knoxville, TN 37996, USA

6 USDA Forest Service, Southern Research Station, College Station, TX 77843-2474, USA

7 Dept. of Ecology and Conservation Biology, Dept. of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843, USA

8 Forest Health Research and Education Center, USDA Forest Service, Southern Research Station, Lexington, KY 40546, USA

9 The American Chestnut Foundation, Meadowview, VA 24361, USA

10 Southern Institute of Forest Genetics, USDA Forest Service, Southern Research Station, Saucier, MS 39574, USA

11 Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

12 Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, 50 Nanyang Ave, Singapore 639798, Singapore

13 Clemson University Genomics & Bioinformatics Facility, Clemson University, Clemson, SC 29634, USA

14 Department of Horticulture, Washington State University, Pullman, WA 99164, USA

15 Plant and Environmental Sciences Department, Clemson University, Clemson, SC 29634, USA

16 Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

17 Department of Biology, Claflin University, Orangeburg, SC 29414, USA

18 UMR Biodiversité Gènes et Communautés, French National Institute for Agricultural Research (INRA), 69 route d'Arcachon, 33612 CESTAS Cedex, France

19 The American Chestnut Foundation, Asheville, NC 28804, USA