









# Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding

John T. Lovell <sup>1,15</sup>✉, Nolan B. Bentley <sup>2,15</sup>, Gaurab Bhattarai<sup>3,15</sup>, Jerry W. Jenkins <sup>1,15</sup>, Avinash Sreedasyam <sup>1,15</sup>, Yanina Alarcon <sup>4</sup>, Clive Bock<sup>5</sup>, Lori Beth Boston<sup>1</sup>, Joseph Carlson<sup>6</sup>, Kimberly Cervantes<sup>7</sup>, Kristen Clermont<sup>8</sup>, Sara Duke<sup>9</sup>, Nick Krom<sup>4</sup>, Keith Kubenka<sup>10</sup>, Sujan Mamidi<sup>1</sup>, Christopher P. Mattison <sup>8</sup>, Maria J. Monteros <sup>4</sup>, Cristina Pisani<sup>5</sup>, Christopher Plott<sup>1</sup>, Shanmugam Rajasekar<sup>11</sup>, Hormat Shadgou Rhein<sup>7</sup>, Charles Rohla<sup>4</sup>, Mingzhou Song<sup>12</sup>, Rolston St. Hilaire<sup>13</sup>, Shengqiang Shu <sup>6</sup>, Lenny Wells<sup>14</sup>, Jenell Webber<sup>1</sup>, Richard J. Heerema <sup>12</sup>, Patricia E. Klein <sup>2</sup>, Patrick Conner<sup>14</sup>, Xinwang Wang<sup>10</sup>, L. J. Grauke <sup>10</sup>, Jane Grimwood <sup>1</sup>, Jeremy Schmutz <sup>1,6</sup>✉ & Jennifer J. Randall<sup>7</sup>✉

Genome-enabled biotechnologies have the potential to accelerate breeding efforts in long-lived perennial crop species. Despite the transformative potential of molecular tools in pecan and other outcrossing tree species, highly heterozygous genomes, significant presence-absence gene content variation, and histories of interspecific hybridization have constrained breeding efforts. To overcome these challenges, here, we present diploid genome assemblies and annotations of four outbred pecan genotypes, including a PacBio HiFi chromosome-scale assembly of both haplotypes of the ‘Pawnee’ cultivar. Comparative analysis and pan-genome integration reveal substantial and likely adaptive interspecific genomic introgressions, including an over-retained haplotype introgressed from bitternut hickory into pecan breeding pedigrees. Further, by leveraging our pan-genome presence-absence and functional annotation database among genomes and within the two outbred haplotypes of the ‘Lakota’ genome, we identify candidate genes for pest and pathogen resistance. Combined, these analyses and resources highlight significant progress towards functional and quantitative genomics in highly diverse and outbred crops.

<sup>1</sup>Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. <sup>2</sup>Department of Horticultural Science, Texas A&M University, College Station, TX, USA. <sup>3</sup>Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, GA, USA. <sup>4</sup>Noble Research Institute, Ardmore, OK, USA. <sup>5</sup>USDA Southeastern Fruit and Tree Nut Research Laboratory, Byron, GA, USA. <sup>6</sup>DOE Joint Genome Institute, Berkeley, CA, USA. <sup>7</sup>Department of Entomology, Plant Pathology and Weed Science, New Mexico State University, Las Cruces, NM, USA. <sup>8</sup>USDA-ARS Food Processing and Sensory Quality Research, New Orleans, LA, USA. <sup>9</sup>USDA-ARS Plains Area Administrative Office, College Station, TX, USA. <sup>10</sup>USDA Pecan Breeding and Genetics, College Station, TX, USA. <sup>11</sup>Arizona Genomics Institute, University of Arizona, Tucson, AZ, USA. <sup>12</sup>Department of Computer Science, New Mexico State University, Las Cruces, NM, USA. <sup>13</sup>Plant and Environmental Sciences, New Mexico State University, Las Cruces, NM, USA. <sup>14</sup>Department of Horticulture, University of Georgia-Tifton Campus, Tifton, GA, USA. <sup>15</sup>These authors contributed equally: John T. Lovell, Nolan B. Bentley, Gaurab Bhattarai, Jerry W. Jenkins, Avinash Sreedasyam. ✉email: [jl Lovell@hudsonalpha.org](mailto:jl Lovell@hudsonalpha.org); [jschmutz@hudsonalpha.org](mailto:jschmutz@hudsonalpha.org); [jrandall@nmsu.edu](mailto:jrandall@nmsu.edu)

While modern breeding has produced significant evolutionary bottlenecks in most major crops<sup>1,2</sup>, genetic diversity of many other economically and culturally important specialty crop species remains largely untouched. This is especially true for newly emerging, orphan, and long-lived perennial crops, which are often not amenable to accelerated breeding regimes<sup>3,4</sup>. The broad genetic diversity available in specialty crops will be crucial when adapting cultivars to new or changing pests, environmental conditions, and consumer demands.

Pecan (*Carya illinoensis*) is one such specialty crop. First transported outside its endemic range by Native Americans<sup>5</sup>, pecan is now cultivated on six continents<sup>6</sup>. While newly worldwide cropping will undoubtedly expose the species to a number of novel diseases and pests, pecan has co-evolved with many pests and pathogens in its endemic range; these include multiple species of *Phylloxera* (a genus of gall-forming aphid-like insects<sup>7</sup>) and other insect species that can significantly reduce yield<sup>6,8–10</sup>, and a genetically diverse phytopathogenic fungus (*Venturia effusa*) that causes scab disease<sup>11,12</sup>, which is the most economically damaging disease of pecan<sup>6,10,13</sup>. Despite a paucity of information on the cellular and genetic mechanisms responsible for susceptibility to these pests and pathogens, several resistant cultivars have been bred to mitigate some yield losses<sup>6,10</sup>.

Compared to the dramatic morphological evolution during domestication of many major crops<sup>12</sup>, modern pecan breeding efforts have thus far resulted in only modest improvements. For example, pecan nuts collected from prehistoric Native American archeological sites appear very similar to present-day cultivars<sup>5</sup>. This is due, in part, to the fact that traditional breeding efforts in pecan and other tree crops can take many decades. Consequently, the primary stocks used in contemporary pecan breeding were derived from crosses made from wild trees during the early twentieth century<sup>6</sup>. Nonetheless, modern pecan breeding has made some significant strides by selecting for genotypes with larger nut size<sup>14</sup>, higher nut quality, and tree tolerances of abiotic and biotic stresses<sup>15</sup>. Thus, the development of molecular markers for agronomic traits, which can be assayed early in life, will dramatically improve the speed, efficiency, and efficacy of selection<sup>3</sup> in long-lived perennial crops such as pecan.

Beyond their long lifespan, the outbred and highly diverse nature of pecan and many other tree crops can also complicate molecular breeding goals. For example, breeding programs in pecan and other tree species commonly seek traits originating from highly diverged populations, subspecies, or even related species<sup>16</sup>. Therefore, it is likely that some genes that could be targets for selection are simply not present in many genotypes. Such diversity, both within and among individuals, makes reliance on a single inbred ‘reference’ genotype untenable and necessitates a paradigm shift towards the use of multiple and outbred genomes.

Here, we construct and analyze four outbred de novo pecan genome assemblies and annotations as a step towards identifying candidate genes and molecular targets for accelerated breeding efforts in outbred and diverse crops. Our efforts to define gene presence-absence variation through a pan-genome annotation reveal evidence of widespread interspecific genomic introgressions. These introgressions and extensive gene content variation between meiotically homologous chromosomes provide a wealth of nut quality and biotic stress resistance genetic diversity that breeders can leverage to improve contemporary and future pecan nut production.

## Results

**Four pecan genomes provide a crucial resource for crop improvement.** The outbred nature of the pecan genome<sup>17</sup>

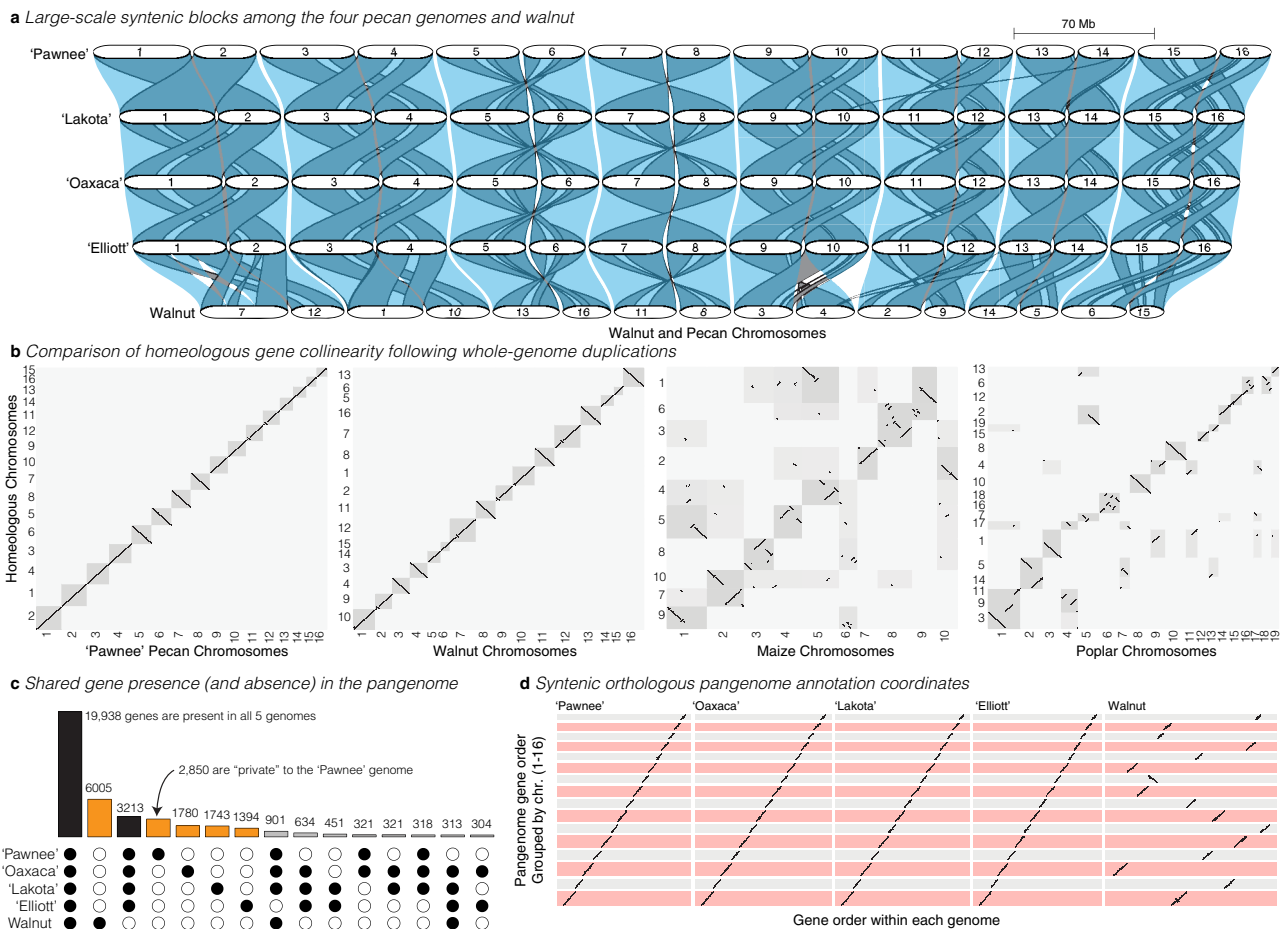
complicates genome assembly methods and efforts to leverage genome resources for breeding goals. For example, since genetic mapping in outbred perennial species typically uses F<sub>1</sub> breeding designs<sup>18</sup>, causal variants may segregate within pecan genomes. Therefore, it is crucial to generate genome assemblies of both meiotically homologous haplotypes in outbred diploids. Nonetheless, past genome sequencing efforts in outbred species have typically sought to represent a single haploid assembly for a genotype either via sequencing an inbred ‘reference’ genotype (e.g., B73 maize) or phasing a highly diverged F<sub>1</sub> hybrid assembly. However, pecan inbred pedigrees are neither biologically realistic (high inbreeding depression) nor practically feasible (long generation times).

Recent assembly methodological improvements and sequencing technology have permitted highly contiguous genome assemblies of switchgrass<sup>19</sup> and several other outbred plant genotypes. Here, we have expanded upon these efforts—instead of collapsing two divergent haplotypes into a single haploid assembly, we sought to build diploid assemblies and capture both haplotypes in four outbred pecan genotypes. To this end, we selected four genotypes to represent the genetic diversity of pecan (‘Pawnee’<sup>20</sup>, ‘Lakota’<sup>21</sup>, ‘Elliott’<sup>22</sup>, and a wild collection from Oaxaca, Mexico ‘87MX3-2.11’, hereon ‘Oaxaca’<sup>23</sup>; see ‘Methods’ section and Supplementary Fig. 1 for details).

Three genomes (‘Oaxaca’, ‘Elliott’, and ‘Lakota’) were assembled using a whole-genome shotgun sequencing strategy combining PacBio single-molecule real-time (RS II, and SEQUEL I, 55.2–108.3 Gb raw long reads, 78.9×–135.3× coverage) and Illumina (HiSeq X Ten and 2500, 50–60× coverage) technologies (Fig. 1a, Table 1). Contigs along the primary haploid path were oriented, ordered, and joined together into 16 highly contiguous ( $N_{50} = 3.7\text{--}4.4$  Mb, Table 1) chromosome pseudomolecules using synteny and Hi-C indexed short-read sequencing. The remaining alternative haplotype contigs ( $n = 3,705\text{--}6,853$ ,  $N_{50} = 0.10\text{--}0.14$  Mb) represented the homologous sequence to the primary path. In total, the alternative haplotypes captured 64.5–76.1% of the total genomic sequence (Table 1). The missing sequence and relatively short contigs resulted from runs of homozygosity that are expected in breeding pedigrees and highly repetitive pericentromeric regions.

In contrast to these three genomes, the ‘Pawnee’ assembly was built with state-of-the-art PacBio circular consensus sequencing reads (‘CCS’ a.k.a. ‘HiFi’, mean coverage = 52.1×), ‘Pawnee’ is by far the most contiguous of the four assemblies (contig  $N_{50} = 26.5$  Mb) with 100% of primary path sequence assembled into chromosomes (Table 1). Crucially, the highly accurate CCS reads permitted the construction of haplotype-aware contigs even in homozygous and repetitive regions. Nearly 90% of the primary ‘Pawnee’ assembly size was captured in the highly contiguous (contig  $N_{50} = 2.9$  Mb) chromosome-scale alternative haplotype assembly (Table 1). We independently annotated each pecan genome assembly through homology and RNA-seq supported methods (Fig. 1a, Supplementary Table 1), which produced very complete annotations (BUSCO scores 94.4–97%). We leveraged these protein-coding DNA sequences to validate the accuracy and completeness of our assemblies.

The Juglandaceae family (walnut, hickory, pecan) experienced a whole-genome duplication (WGD) event ca. 60 million years (Myr) before present<sup>24</sup>, resulting in pairs of homeologous chromosomes with highly conserved paralogous gene order collinearity (i.e., synteny). Reanalysis of syntenic orthologous and paralogous gene blocks in the recently published walnut genome<sup>25</sup> and our ‘Pawnee’ assembly revealed a total of 26 large homeologous collinear gene blocks in walnut and 25 such blocks in pecan across 16 chromosomes. This represented an exceptional level of chromosomal evolutionary conservation (one



**Fig. 1 Comparative analysis of four de novo pecan genomes.** **a** A map of syntenic orthologous (transparent blue) and homeologous blocks (gray with black borders) among the four reference genomes and the walnut outgroup. Chromosomes are represented by white segments and are scaled to the same physical size (Mb: megabases) for all genomes. Orthologous chromosomes are stacked vertically and labeled accordingly. **b** Comparisons of the degree of synteny between homeologous chromosomes across the 'Pawnee', walnut, maize, and poplar genomes. The dotplots display the gene-rank-order positions of syntenic blastp hits along the main genome (x axis) and homeologous chromosomes (y axis). Chromosomal bounds are shaded by the total number of blast hits found between each pair of homeologous chromosomes. **c** Across the pan-genome, the vast majority of all genes are found in orthogroups that contain all four pecan genomes (bars shaded black); however, genes private to each genome (shaded orange) and, to a lesser degree, shared among >1 genome (gray) are also common. Filled circles represent presences in orthogroups; open circles are absences. **d** The high level of synteny between the pecan genomes and walnut allowed for simple pan-genome construction and gene ordering. Here, each point represents the location of a gene by its rank-order location within each de novo genome assembly (x axis) and the inferred syntenic position in the pan-genome (y axis). Source data are provided as a Source Data file.

rearrangement every 6.7 Myr in pecan, Fig. 1b). For context, comparative genomics using the same parameters revealed one rearrangement between homeologs every 800,000 years in poplar (WGD ~58 Myr ago<sup>26,27</sup>) and 490,000 years in maize (WGD ~12 Myr ago<sup>28,29</sup>; Fig. 1b). While shorter generation time in the progenitors of maize, and possibly poplar, certainly could have contributed to elevated chromosomal evolutionary rates, paralogous synteny in pecan represents a remarkable level of chromosomal stability over 60 M years. Crucially, such retained synteny offers an opportunity to validate genome completeness and contig ordering by comparing synteny between homeologous chromosomes. Since our pecan genomes were assembled agnostic to homeologous chromosome synteny, this level of conservation lends credence to the assertion that these four genomes were very complete and lack any major assembly errors.

**A pan-genome representation of pecan gene diversity.** Genome evolution and genetic diversity that underlie breeding targets can

arise from a diverse set of genetic and epigenetic changes including short insertions/deletions (INDELs), single nucleotide polymorphisms (SNPs), structural variants, and presence-absence variants (PAVs). Our four de novo genome assemblies and annotations permit the inference of each of these variant types through comparative analysis of a database of conserved and variable orthologous gene sequences among all genomes, a 'pan-genome annotation' (Fig. 1c, d). In clades with a history of whole-genome duplications such as Juglandaceae, pan-genome construction methods based solely on sequence homology are not sufficient for comparative genomics since paralogous sequences would likely pollute otherwise orthologous gene families. For example, 16.2% of genome-wide 'Pawnee' orthogroups contained homeologous gene pairs. To overcome this genome complexity, we constructed a synteny constrained orthologous pan-genome annotation (Supplementary Data 1 and Fig. 1), which simultaneously masked paralogous regions and condensed tandem arrays into a single orthologous path through multiple genomes. While offering a powerful method to reduce paralogous gene content in the pangenome, it is important to note that constraining to

**Table 1 Genome assembly and annotation statistics for each of the four genomes.**

Genomic features	'Oaxaca'	'Lakota'	'Elliott'	'Pawnee'
Assembly size (Mb) <sup>a</sup>	649.96	668.99	656.69	674.27
Number of scaffolds	298	261	431	16
Number of contigs	552	499	829	34
Gap content (%)	0.4%	0.4%	0.6%	0.0%
Contig N50 (Mb)	4.4	3.7	4.4	26.5
Genome in chromosomes (%)	98%	96.1%	95.5%	100%
Number of annotated genes	31,911	33,280	31,042	32,267
Average number of exons per gene	5.4	5.5	5.5	5.5
Repeat sequences (%)	46.5%	33.8%	32.3%	49.7%
Total alt. haplotype size (Mb) <sup>b</sup>	494.9	469.7	423.6	603.2
Number of alt. haplotype scaffolds	6,853	5,222	3,702	16
Number of alt. haplotype contigs	6,853	5,222	3,702	323
Alt. haplotype contig N50 (Mb)	0.13	0.10	0.14	2.90
Alt. genome size (% of main)	76.1%	70.2%	64.5%	89.5%

<sup>a</sup>Statistics extracted for the primary ('main', top section) assembly.  
<sup>b</sup>Alternative haplotype (alt.) are presented in the bottom five rows.

syntenic regions will ignore orthologs involved in very small chromosomal translocations. Overall, these minor translocations represent <0.4% of the genome.

Rooted against gene order of the 'Pawnee' genome and including walnut (*Juglans regia*)<sup>25</sup> as an outgroup, the pan-genome annotation contained 42,416 orthogroups, 21,196 of which were single-copy in all four pecan genomes (Fig. 1c). Among the four pecan genomes, the synonymous ( $K_s$ ) and non-synonymous ( $K_a$ ) nucleotide substitution rates of single-copy syntenic orthologs were fairly low (mean  $K_a \pm \text{SEM} = 0.0017 \pm 1.24 \times 10^{-5}$ ;  $K_s = 0.0042 \pm 2.4 \times 10^{-5}$ ; Supplementary Fig. 2 and Fig. 2a). This evolutionary conservation was exemplified by allergen proteins, which tend to be highly conserved in walnut and other tree nut crops<sup>30</sup>. Pecan allergic reactions are caused by immunoglobulin-E (*IgE*) recognition and binding of Car i 1, 2, and 4 allergen protein structures<sup>31</sup>. Like many of the orthogroups present in all four genomes, coding sequences of the Car i 1 and Car i 2 allergens were nearly identical between genotypes and *IgE* binding epitopes were conserved (Supplementary Fig. 3). We observed only a single amino acid substitution in the 'Oaxaca' Car i 1 and 'Elliott' Car i 2 alleles respectively. However, some unique differences were observed in the Car i 4 sequences among the cultivars, where 'Elliott', 'Oaxaca', and 'Lakota' shared a 33 amino acid exon, which may have differentiated the allergen profile of 'Pawnee'.

In contrast to the constrained coding sequence evolution of single-copy genes, we observed significant gene PAV among these relatively closely related genomes. Overall, 38.7% of orthogroups in the pan-genome ( $n = 13,010$ ) were incomplete, representing PAV among the pecan genomes (Fig. 1c and Supplementary Data 1). To dissect the differential roles of gene-model structural evolution, sequence deletion, and evidence-based gene model thresholding on PAV, we compared sequence similarity between genes present in one annotation and the syntenic unannotated genomic regions where absent genes should exist (Supplementary Table 2). Overall, a majority of the observed pan-genome PAV was driven by gene sequences that were unannotated yet similar to annotated sequences in alternative genomes. As observed previously<sup>32</sup>, such genes tended to be of low-quality barely

passing gene evidence score thresholds. However, 8,655 absent genes had no similar sequence within syntenic regions, indicating significant and diverse mechanisms of gene absence among our genomes.

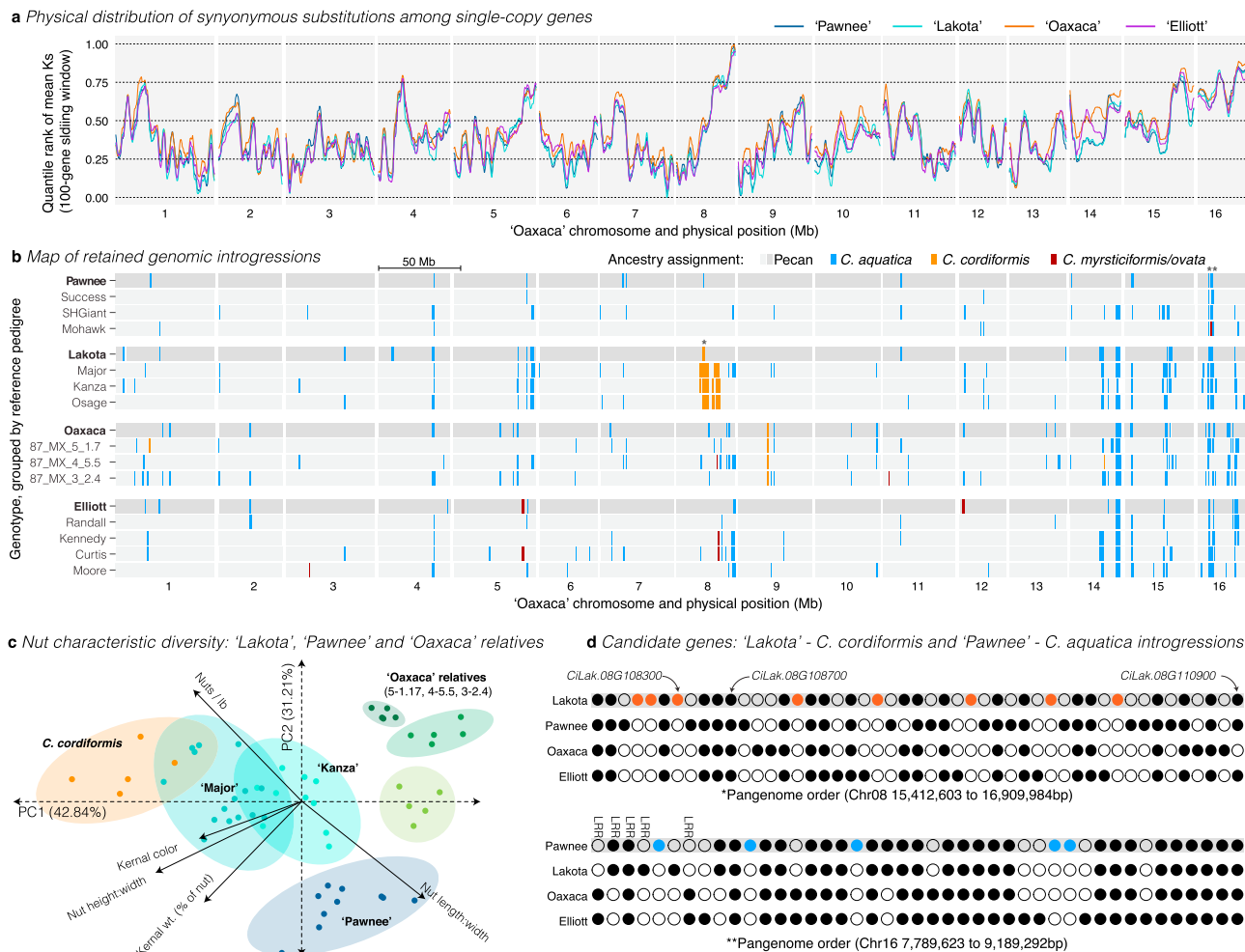
Among the four pecan genomes, we observed 3,889 blocks of five or more consecutive genes that were absent in one or more of the references. Many of these gaps represented true gene absences and demonstrated that multiple reference genomes offered a major improvement in gene content representation over a single pecan genome. Further, the ubiquity of large runs of genes that were unique to a single genotype ('private' genes) potentiated a role for independent genomic introgressions from distantly related gene pools into each of our reference genome lineages, a hypothesis we test below.

**Genomic introgressions as breeding targets for disease resistance.** In addition to conspicuous runs of PAV within each genome, we observed several physical regions of elevated divergence among the four genomes (Supplementary Fig. 2 and Fig. 2a). While a number of factors could cause these divergence peaks, ancient and contemporary hybridization and admixture offer one potential reason for the observed high level of PAVs, long runs of private genes, and regions of elevated nucleotide substitutions in each assembly. Indeed, there are records of historical pecan breeding incorporating progeny from bitternut (*C. cordiformis*) and other interspecific pedigrees<sup>16</sup>. Furthermore, morphological analysis of extant trees<sup>33</sup> and remains from pre-historical archeological sites<sup>5,34,35</sup> in Mexico found a strong affinity to *C. aquatica* and *C. myristiciformis*, indicating that ancient admixture between *Carya* species may have imbued pecan with desirable traits for human cultivation.

Given the complete sequence order of our assemblies, it was possible to track the positions and identities of genomic introgression blocks from these three related species into pecan breeding pedigrees. To estimate introgression proportions and positions, we resequenced multiple genotypes of each of these three potential admixing species (*C. cordiformis*, *C. aquatica* and *C. myristiciformis*) and defined admixture blocks by decoding SNP-based (38–69× coverage of Illumina 2 × 150 bp reads) posterior ancestry probabilities among the four reference genomes and three or four relatives of each reference genotype (Fig. 2b and Supplementary Table 3). By including multiple relatives for each genome, we were able to define high-confidence interspecific genomic introgressions as regions with non-pecan ancestry in all related genotypes. While these introgressions represented sequences from other species, it is important to note that some introgressions may have been derived from other unsampled species.

Overall, *C. aquatica* (or an unsampled related species) was the primary source of interspecific introgressions, representing 6.6–20.6 Mb (1.04–3.23%, Supplementary Table 3) of the entire genome sequence of the four reference genomes. These introgressions tended to be small and distributed regularly across the genome (Fig. 2b), indicating that *C. aquatica*'s hybridization history may have begun long before modern pecan breeding efforts. This seems particularly plausible given the largely sympatric geographic distributions of the two species. The physically discontinuous, yet high levels of *C. aquatica* ancestry likely contributed to the significantly elevated synonymous substitution rates on chromosome 5 (right arm), 14 (right arm), and 16 (Supplementary Fig. 2 and Fig. 2a).

In contrast to a putatively ancient and natural origin of admixture between pecan and *C. aquatica*, the vast majority of *C. cordiformis* ancestry was concentrated in a >7.5 Mb block on chromosome 8 derived from the 'Major'<sup>16</sup> cultivar and present



**Fig. 2 A map of interspecific genomic introgressions in four pecan genomes.** **a** Sliding window analysis of neutral site substitution rate ( $K_s$ ) within all single-copy orthogroups that were represented by all four genomes.  $K_s$  values were transformed to quantiles and a 100-gene sliding window was applied within each chromosome and genome. The resulting sliding window values are presented on a 0–1 scale where lower values represent the most similar regions across the physical genome (Mb; megabases). See Supplementary Fig. 2 for raw pairwise  $K_s$  values. Close-up pan-genome representations of two regions marked \* and \*\* are highlighted in **d**. **b** Genome ancestry maps of the four reference genomes and representative members of each pedigree. Posterior probabilities of ancestry for three primary hybridizing species were decoded into blocks (colors red, orange, blue) of  $\geq 500$  variants. The background probabilities of ancestry for the reference genomes and relatives respectively. **c** The large introgression in the ‘Major’ and ‘Kanza’ relatives of ‘Lakota’ appear to imbue phenotypic variation typical of *C. cordiformis* to these genotypes. 13 traits associated with nut yield and quality were assayed for a single *C. cordiformis* genotype (O2-COR-LA-BF1), ‘Pawnee’, two members of the ‘Lakota’ pedigree (‘Major’ and ‘Kanza’) and three genotypes from Mexico that may be related to ‘Oaxaca’. The 13 traits were reduced to five non-collinear ( $|r| < 0.75$ ) representatives and decomposed into the two major principal component axes (PC1, PC2), which collectively explained  $>74\%$  of the variation. For each genotype, we present the positions in PCA space and the 95% confidence ellipse. **d** Pan-genome gene representatives are shown for each unique orthogroup within two physical (base pairs, bp) introgression intervals. Circles represent presence (filled) or absence (open) for each genome (row) by orthogroup (column) in the introgression. The first row in each plot represents the genome into which an introgression was observed. Private orthogroups to that genome are colored following panel **b**. Three candidate genes in ‘Lakota’ and the dense region of leucine-rich repeat (LRR) genes are annotated along the top row of each map. Source data underlying Fig. 3a–c are provided as a Source Data file. Raw data associated with **d** can be found within the pangenome database in Supplementary Data 1.

among four genotypes related to ‘Major’ (Table 1, Fig. 2b; ‘Lakota’ [‘Mahan’  $\times$  ‘Major’], ‘Kanza’ [‘Major’  $\times$  ‘Shoshoni’], and ‘Osage’ [‘Major’  $\times$  ‘Evers’]). The presence of a single large introgression indicated (1) a recent origin, (2) positive selection to retain the introgressed region, and (3) purifying selection against all other *C. cordiformis* haplotypes across the genome. Additionally, while ‘Lakota’, ‘Major’, ‘Kanza’, and ‘Osage’ have overall pecan-like morphologies they possess other traits that cluster closely with *C. cordiformis* (Fig. 2c). Indeed, ‘Major’ (and to a lesser degree ‘Kanza’) also had the most similar nut characteristics to *C. cordiformis* of a subset of genotypes related to ‘Pawnee’, ‘Lakota’ and ‘Oaxaca’ (Fig. 2c).

‘Lakota’, ‘Major’ and other members of this pedigree are also known to have strong fungal and abiotic stress resistance<sup>16</sup>, traits that could be due to shared ancestry across the large introgression on chromosome 8. To explore this hypothesis, we examined the *C. cordiformis* introgression interval in ‘Lakota’, which was much narrower than the intervals in other members of its pedigree. Such introgression size reduction in a single generation indicated that recombinant gametes at the margins of this introgression were selected by breeders. The 1.41 Mb region contained 24 high-value candidate genes in the ‘Lakota’ genome (Supplementary Data 2), many of which had homologs in other species known to be involved in nutrient acquisition, plant development, and

defense responses including *SNF1*-related protein kinase and Leucine-Rich Repeat (LRR) receptors. The pan-genome database of this region contained 46 total orthogroups, eight of which were private only to ‘Lakota’ (Fig. 2d). The 17.4% of private genes unique to ‘Lakota’ represented a >4-fold enrichment in private gene content compared to the genome-wide average (Fisher’s exact test, odds ratio = 4.232,  $P=0.0012$ ), demonstrating evidence of non-pecan ancestry from both SNPs and PAV datasets.

In addition to the chromosome 8 *C. cordiformis* introgression, there were a number of other high-confidence introgressions that appeared in multiple related genotypes (Fig. 2b and Supplementary Fig. 2). For example, chromosome 5 and 16 harbored introgressions from *C. myristiciformis* into ‘Elliott’ and *C. aquatica* into the ‘Pawnee’ pedigree, respectively. For each of these regions, we queried the pan-genome and extracted the synteny constrained orthogroups within each focal genome annotation (Supplementary Data 2). The introgressed region on chromosome 5 was characterized by plant signaling genes (there are no less than 10 cell wall receptor kinases) and cell wall defense genes including lignin biosynthesis genes (4 genes), cellulose synthase, and inositol oxygenase, which involved in cell wall polymerization. The region of *C. aquatica* introgression into ‘Pawnee’ on chromosome 16 contained nine LLR receptor serine/threonine kinase genes from five unique orthogroups (Fig. 2d). The apparent overabundance of defense-related genes within introgression regions hints at a possible adaptive role for introgressions in both pecan breeding and wild populations.

### Induced gene networks in a pathogen susceptible cultivar.

Biotic stress tolerance is a major breeding objective in many crops, but especially in long-lived tree species where pests and disease incidence varies across years and locations<sup>10,36</sup>. A temporally and spatially variable pathogen composition can obfuscate breeding values, and subsequently, reduce the efficacy of traditional breeding efforts. Given these constraints, generating molecular targets for resistance to specific pathogens can dramatically accelerate crop improvement outcomes<sup>37–39</sup>. For example, pecan scab (caused by the phytopathogenic fungus *V. effusa*) produces black circular lesions that can reduce yield and nut quality, and if not controlled, can cause crop failure<sup>40</sup>. *V. effusa* is composed of multiple pathotypes each capable of infecting a relatively small subset of pecan cultivars<sup>41</sup>. Most benign *V. effusa*-pecan cultivar interactions result in the arrest of fungal growth shortly after cuticular penetration, whereas virulent interactions result in abundant intercellular hyphal growth and sporulation<sup>42</sup>. Natural populations of pecan present the host with a diverse and evolving host, limiting the buildup of virulent races. In contrast, pecan orchards composed of replicated stands of only a few cultivars promote the accumulation of pathogenic strains<sup>41,42</sup>. In recent years several major industrial pecan cultivars, including the most widely planted cultivar in the southeastern U.S. (‘Desirable’), have become more susceptible to scab infection<sup>10</sup>.

To understand susceptibility in ‘Desirable’ and the landscape of short-term gene-expression plasticity to *V. effusa*, we compared transcript abundance in leaf tissue inoculated with the scab isolate ‘De-Tif-11’ compared to the control treatment (Supplementary Fig. 4) through sequencing of RNA extracted across three biological replicates at 24 h post inoculation (Supplementary Table 4). While we did not generate a genome assembly and annotation for ‘Desirable’, the phylogenetic dispersion of the four pecan genomes covers much of the pecan diversity whereby ‘Pawnee’ and ‘Desirable’ share a grandparent. Of the 32,267 genes in the ‘Pawnee’ reference, 194 genes were differentially expressed ( $|\log_2$  fold-change $|\geq 1.5$  and FDR-adjusted  $P$ -value < 0.05)

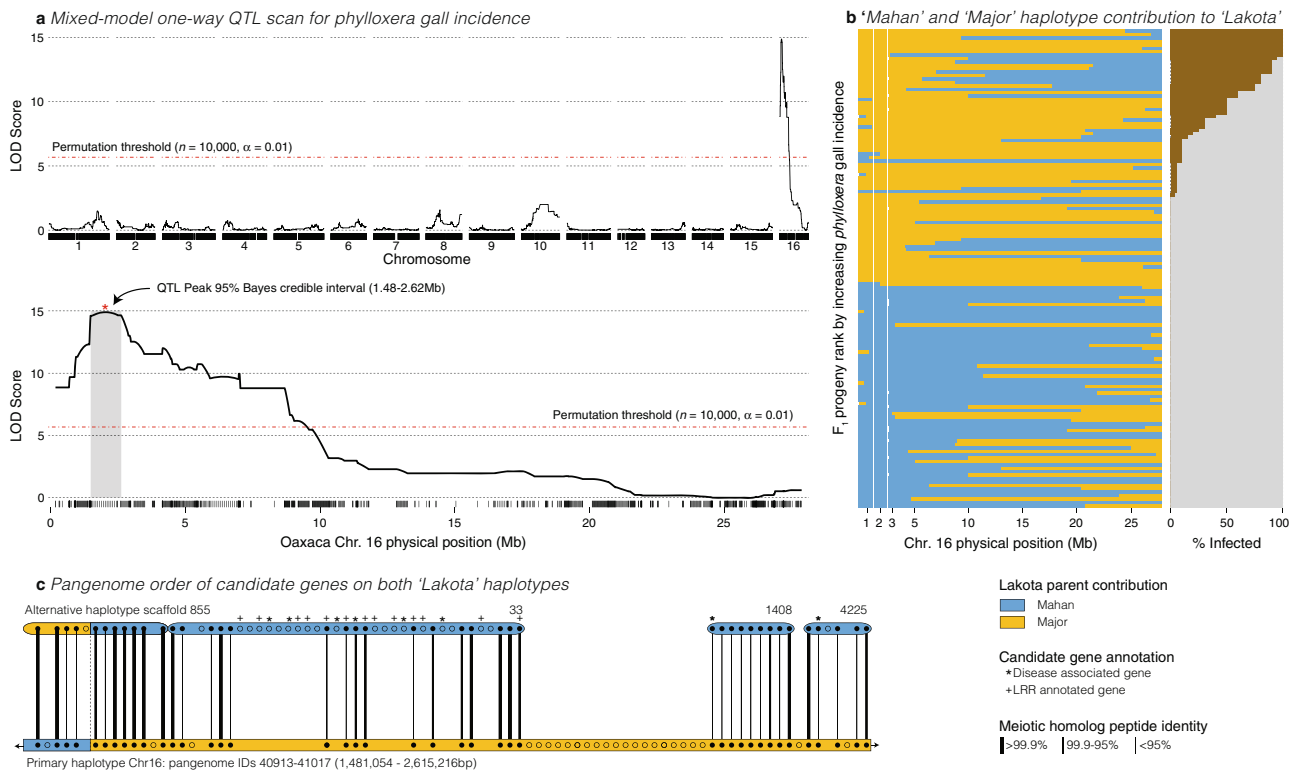
between control and inoculated tissue, showing strong evidence of molecular phenotypic plasticity to the fungal pathogen treatments (Supplementary Fig. 4 and Supplementary Data 3). While gene ontology (GO) term enrichments from such differential expression analysis can be vague and imprecise, GO enrichments in this experiment were clear (Supplementary Table 5): by far the most significant terms were ‘response to wounding’ (downregulated genes) and ‘response to chitin’ (upregulated genes). Other significantly enriched terms were heavily biased towards stress responses and oxidation–reduction status. Since chitin is the primary trigger of plant responses to fungus<sup>43,44</sup>, and redox status is crucial to plant defense responses<sup>45,46</sup>, these differentially expressed genes offer a set of targets to explore host susceptibility to *V. effusa*.

### PAV within genomes to target candidate genes in outbred pedigrees.

While genome-informed molecular and genetic diversity exploration can document potentially important breeding targets, these efforts lack causality at a per-locus level; however, linkage-based quantitative trait locus (QTL) mapping can identify sequences in linkage disequilibrium with causal variants. Due to pecan’s long generation times and inbreeding depression, breeding programs have utilized pseudo-testcross ( $F_1$ ) mapping strategies<sup>47–49</sup> to identify causal variants that segregate within the genomes of one or both parents. We applied this mapping strategy to test the genetic basis of phylloxera leaf gall incidence caused by feeding of the larva of aphid-like phylloxera insects (order Hemiptera) among 143 2-year old ‘Lakota’ × ‘Oaxaca’  $F_1$  saplings (Supplementary Data 4) at a nursery in Somerville, TX. Linkage phases of 11,489 loci (Supplementary Data 5) that were heterozygous in ‘Lakota’ and homozygous in ‘Oaxaca’ were defined by the parent of origin (‘Mahan’ or ‘Major’) based on comparison of the phased marker positions in common with a previous restriction-enzyme site associated sequencing of ‘Mahan’ and ‘Major’<sup>48,49</sup>.

QTL mapping revealed a single large peak on chromosome 16 (Fig. 3a and Supplementary Data 5). Given a left-skewed phenotypic distribution, the peak logarithm of the odds (LOD) score of 14.8 approached the maximum possible value in this experimental design. Indeed, all but two of the individuals with the ‘Mahan’ haplotype at the peak QTL position (2.021 Mb, Fig. 3a) were completely free of phylloxera galls, while all highly susceptible genotypes inherited the ‘Major’ haplotype (Fig. 3b).

To define the candidate genes that may explain phylloxera gall incidence in this population, we explored the syntenic pan-genome region in ‘Lakota’ that corresponded to the 95% Bayes credible QTL interval from positions 1,481,054–2,615,216 bp in the chromosome 16 sequence in the ‘Oaxaca’ assembly (Fig. 3c and Supplementary Table 6). Since causal variants in  $F_1$  experimental crosses are heterozygous within the parents, we ranked candidate genes by the level of divergence between the primary and alternative haplotype of the outbred ‘Lakota’ genome assembly. Overall, the ‘Lakota’ genomic interval contained 40 genes only found in the primary assembly. The bulk of these genes resided within a block where no homologous alternative sequence was assembled. These likely represent homozygous regions that were less likely to contribute to the QTL. However, 22 gene models were only found in the alternative assembly (Fig. 3c, unfilled circles in alternate sequence) and represented higher-confidence PAV since both haplotypes across these loci were assembled but genes were only annotated on one sequence. Finally, 12 genes were found in both assemblies but with peptide identities of <98% between haplotypes. Based on this logic, the aforementioned 22 and 12 gene model groupings were made high-priority candidate genes.



**Fig. 3 Analysis of a major QTL for phylloxera resistance.** **a** Quantitative Trait Locus (QTL) scans, controlling for genomic background via the leave-one-chromosome-out method for % phylloxera gall incidence. This experiment was conducted once at a single time point. Since the phenotype is non-normal, we determine the significance of QTL peaks via 10,000 permutations. The full genome and a close-up visualization of chromosome 16 are presented along the physical position (Mb: megabases) of the ‘Oaxaca’ genome assembly. The 95% confidence interval surrounding the QTL peak is shaded. **b** As evidenced by very high LOD scores for a 140-genotype population, there is an extremely strong haplotype structure at the peak QTL (between the vertical white bars), where all but two individuals that inherited the ‘Mahan’ haplotype from ‘Lakota’ have no evidence of phylloxera galls (gray horizontal bars in the plot to the right), while all individuals with >50% phylloxera gall incidence retained the ‘Major’ haplotype at the QTL peak region (brown horizontal bars indicate % incidence). **c** To define candidate genes, we queried the pan-genome within the physical bounds (base pairs, bp) of the QTL interval. All unique genes in this interval were projected onto the alternative haplotype; those contigs where >50% of the projected genes were derived from the candidate interval were extracted and aligned to the primary haplotype. Orthologous genes between the two haplotypes are connected by a solid line, the thickness of which is scaled by % identity between the two protein sequences. Presence-absence variant (PAV) genes without a projected ortholog are represented by open circles. Homologs of the genes in the interval were queried in model systems and qualified by whether annotations indicated a disease-related function or a leucine-rich repeat (LRR) motif. Finally, the haplotypes were coded by whether they were derived from the ‘Mahan’ or ‘Major’ parents of ‘Lakota’. Source data underlying **c** are provided as a Source Data file. Raw data associated with **a**, **b** can be found in Supplementary Data 5.

Analysis of the functional domains (Pfam and PANTHER) and UniProt descriptions of the high-priority candidate genes identified 20 putative plant immune response genes. The largest grouping of these included a series of 13 LRR motif-related genes that were present at variable copy numbers in all the reference sequences but were least numerous in the primary ‘Lakota’ sequence which represented the ‘Major’ derived haplotype across this interval. Consistent with LRR-induced phylloxera resistance, studies in other systems have mapped aphid resistance to LRRs in *Capsicum baccatum*<sup>50</sup> and cucumber<sup>51</sup>. While this analysis does not define a single candidate gene, it does inform future efforts to characterize the mechanism of phylloxera susceptibility in pecan by prioritizing experiments of differential LRR-gene-mediated phylloxera gall resistance. Additionally, these candidate variants between haplotypes of the outbred ‘Lakota’ genome provide target loci for marker-assisted selection to improve phylloxera resistance in several major breeding pedigrees of pecan.

**Discussion**

Traditional tree breeding strategies require observation of production-associated traits; yet in species with long juvenile

growth periods like pecan, these data can take years to observe. Conversely, marker-assisted, and genomic selection can be implemented prior to plant maturity, dramatically improving the speed and efficacy of selection in long-lived perennial species breeding programs.

The majority of genome-enabled breeding efforts rely on mapping short sequences against a single haploid reference genome. However, outside of a handful of modern domesticated crops, many plant species (including pecan and other tree crops) are characterized by tremendous genetic diversity and often have a history of interspecific hybridization or polyploidy. In these cases, high-value breeding targets may not be easily verifiable when contrasted to a single reference. For example, the candidate genes for the phylloxera susceptibility QTL described here represented a complex genomic region with tandem arrays and extensive presence-absence variation within a single parental genome. Without additional genomic assemblies, annotations and comparisons, mapping short reads to a reference would not have been sufficient to determine the extent of molecular evolution in this region.

Multiple genome assemblies have been recently constructed for many species, revealing a previously undiscovered level of intra- and interspecific genetic exchange<sup>52</sup>. It is becoming clear that

such large-effect evolutionary events are common in nature and may be a source for accelerated selection of high-value breeding targets, especially in emerging and perennial crops. Furthermore, much of the genetic variation in wild and outbred species exists as highly diverged heterozygous haplotypes within individuals. As such, the single-reference-genome paradigm is not sufficient for functional genomics in such systems, not only because heterozygous gene sequence variation cannot be captured by a haploid genome assembly, but also because gene content is highly variable among genotypes. Combined, our approach that integrated comparative and quantitative genomics among multiple outbred de novo genomes shed light into an evolutionary system that would have been poorly represented by a single haploid genome.

## Methods

**Sequenced genotypes and pedigrees.** Four pecan genotypes were selected for complete genome sequencing. ‘Oaxaca’ was collected in September 1987 as an open-pollinated nut accession from its mother tree near Zaachila, Oaxaca, Mexico<sup>23</sup>. ‘Oaxaca’, had higher homozygosity than the other three genotypes. As such, and because of a fairly independent pedigree from the other cultivars, we chose to use ‘Oaxaca’ as the reference to which short reads were mapped for introgression and genetic mapping purposes (see below). The other three are commercial cultivars. ‘Elliott’ was a seedling of unknown parentage, but possibly descended from trees collected by John Hunt in Mexico during the Mexican war of 1848<sup>22</sup>. While not a heavy nut producer, ‘Elliott’ produces a relatively small, round nut of very high kernel quality and exhibits a high level of resistance to scab. ‘Lakota’ was released in 2007 after extensive testing and was derived from the controlled cross of ‘Mahan’ × ‘Major’ in 1964<sup>21</sup>. ‘Lakota’ pecan trees have excellent tree strength, can produce large yields, exhibit early nut maturation, and have excellent scab resistance. ‘Pawnee’ was the progeny of a controlled cross of ‘Mohawk’ × ‘Starking Hardy Giant’ (‘SHGiant’ in Fig. 2b) in 1963 and released in 1985<sup>20</sup> and is notable for its early nut harvest and excellent resistance to the yellow aphid complex<sup>53</sup>. ‘Lakota’ and ‘Pawnee’ share ‘Mahan’ as an ancestor; ‘Mahan’ is the mother of ‘Lakota’ and the maternal grandfather of ‘Pawnee’. To place these genotypes in the context of a sample of pecan genetic diversity, we calculated principal components (Supplementary Fig. 1b) derived from the marker-based genetic distance matrix presented in Bentley et al.<sup>48</sup>. The following genotypes were excluded from Bentley et al.’s<sup>48</sup> distance matrix prior to PCA calculation due to either not being a pecan genotype, having low coverage or being a clonal replicate: ‘Major’, ‘Jones Hybrid’, ‘Abbott Thinsell’, ‘89-XBR-RDM-1’, ‘Nielsen Ovata’, ‘92-AQU-TX-2’, ‘09-CAT-ZH-45’, ‘02-COR-LA-BF-2’, ‘Ring Palmeri’, ‘Clark II’, ‘CloneA-Ramet1’, ‘CloneA-Ramet2’, ‘CloneA-Ramet3’, ‘CloneB-Ramet1’, ‘CloneB-Ramet2’, ‘CloneB-Ramet3’, ‘CloneC-Ramet1’, ‘CloneC-Ramet2’, and ‘CloneC-Ramet3’.

**Genome sequencing and assembly.** Leaf tissue was collected from extant trees at the College Station, TX, USA orchard (30.51°N, 96.44°W): CSHQ13-4 (‘Pawnee’), CSX8-4 (‘Lakota’), CSP1-30 (‘Oaxaca’) and CSV16-10 (‘Elliott’). High molecular weight DNA was extracted for all genomes from young leaves using the protocol of Doyle and Doyle<sup>54</sup> with minor modifications. Size was validated by pulsed-field gel electrophoresis.

The ‘Oaxaca’ genome was assembled and polished with MECAT v1<sup>55</sup> using 78.94× PACBIO coverage (average read length of 12,163 bp), and the resulting assembly was polished using QUIVER v2.2.2<sup>56</sup>. Misjoins in the assembly were identified using a combination of previously published 774-marker genetic map<sup>49</sup> and HiC scaffolding. A total of nine misjoins were identified in the polished assembly. The main genome consisted of 552 contigs assembled into 298 scaffolds that contained 647.4 Mb of sequence (contig  $N_{50}$  = 4.4 Mb, scaffold  $N_{50}$  = 42.3 Mb). Scaffolds were oriented, ordered, and joined together into chromosome pseudomolecules using a combination of the Hi-C scaffolds and genetic markers<sup>49</sup>. A total of 298 joins were applied to the broken assembly to form the final assembly consisting of 16 chromosomes, with a total of 97.98% of the assembled sequence contained in the chromosomes. Assembling a diploid genome in an outbred individual requires a computational step to distinguish a primary and alternative haplotype. While we define the primary haplotype as that with the most contiguity, regions that collapse to a single haplotype due to homozygosity or repeat content may introduce overlapping chromosomal regions that must be represented as a single-copy haplotype without duplicate copies being unnecessarily repeated. To resolve minor overlapping regions on contig ends, adjacent contig ends were aligned to one another using BLAT v36<sup>57</sup> and a total of 44 adjacent alternative haplotypes were identified in the joined contig set and were collapsed using the longest common substring between the two haplotypes. Heterozygous SNPs and INDELS that represented phasing errors were corrected using the 65.01× raw PACBIO data. A total of 119,268 (5.5% of the 2,152,592 heterozygous SNPs/INDELS) were corrected.

‘Lakota’ and ‘Elliott’ genomes were assembled in an identical manner to ‘Oaxaca’, except that syntenic markers with ‘Oaxaca’ were used to identify misjoins

and joins instead of a de novo genetic map. The syntenic markers consisted of 57,706 1 kb unique, non-repetitive regions extracted from ‘Oaxaca’ sequences, with a minimum spacing between markers of 20 kb. Assembly and polishing were conducted following the ‘Oaxaca’ genome, with PacBio coverage (125.01×/135.33× ‘Lakota’/‘Elliott’, respectively; average read length of 11,488/8,835 bp); 61/11 misjoins and 298/4 contig joins were identified with Hi-C and syntenic markers. 60/96 alternative haplotypes were collapsed, and a total of 461,327/56,589 heterozygous SNPs/INDELS phasing errors were corrected with the raw PACBIO data. The Lakota genome contained 669.0 Mb of sequence in scaffolds with a contig and scaffold  $N_{50}$  of 3.7 and 41.6 Mb, respectively, and 99.8% of the main genome assembled into scaffolds >50 kb. The Elliott genome contained 652.7 Mb of sequence in scaffolds with a contig and scaffold  $N_{50}$  of 4.4 and 41.2 Mb, respectively, and 99.4% of the main genome assembled into scaffolds >50 kb.

The ‘Pawnee’ main assembly was performed with HifiAsm v0.5<sup>58</sup> using 52.12× CCS coverage (mean read length of 20,869 bp), and the resulting assembly was polished using RACON v0.5<sup>59</sup>. As above, misjoins in the assembly were identified using a combination of 58,192 1 kb unique, non-repetitive syntenic sequences derived from the V1 ‘Lakota’ release, and Hi-C scaffolding using the JUICER v1.5.6<sup>60</sup> pipeline. A single misjoin was identified in the polished assembly. Scaffolds were then oriented, ordered, and joined together using a combination of the Hi-C scaffolds and syntenic markers. A total of 18 joins were applied to the broken assembly to form the final assembly consisting of 16 chromosomes, with a total of 100% of the assembled sequence contained in the chromosomes. Heterozygous SNP/INDEL phasing errors were corrected using the 52.12× CCS data. A total of 559 (0.01% of the 5,428,928 heterozygous SNPs/INDELS) were corrected. Additionally, homozygous SNPs and INDELS were corrected in the release sequence using 50× of Illumina reads (2 × 150, 400 bp insert). The Pawnee primary genome assembly contained 674.3 Mb of sequence in scaffolds with a contig and scaffold  $N_{50}$  of 26.5 and 44.7 Mb, respectively. The alternative haplotype genome assembly contained 603.2 Mb of sequence in scaffolds with a contig and scaffold  $N_{50}$  of 2.9 and 40.0 Mb respectively.

For all genomes, contigs containing telomeric sequence were identified using the (TTTAGG)<sub>n</sub> repeat, and care was taken to ensure that contigs terminating in this sequence were properly oriented in the production assembly.

**Genome annotation.** Our gene annotation pipeline leveraged both homology and RNA sequencing evidence to build high-confidence gene models. Transcript assemblies were generated from 2 × 150 paired-end Illumina RNA-seq reads using PERTRAN (Lovell et al.<sup>32</sup>; see Supplementary Table 1 for library coverage, read counts, and other metadata). RNA-seq transcript assemblies and ESTs were aligned to the genome assemblies with PASA v2.0.2<sup>61</sup>. Repetitive DNA elements were identified de novo with RepeatModeler v2.0.1<sup>62</sup>. Loci were determined by transcript assembly alignments or EXONERATE v2.4.0<sup>63</sup> alignments of proteins from *Arabidopsis thaliana*<sup>64</sup>, *Populus trichocarpa*<sup>27</sup>, soybean<sup>65</sup>, *Oryza sativa* (var Kitaake)<sup>66</sup>, *Sorghum bicolor*<sup>67</sup>, *Setaria viridis*<sup>68</sup> and Swiss-Prot<sup>69</sup> proteins to repeat soft-masked genomes using RepeatMasker v4.1.0<sup>70</sup>. Alignment extensions of up to 2,000 bp were permitted on both strands unless the extension overlapped with another locus on the same strand. Homology-based gene model prediction was accomplished via FGENESH v3.1.1/ FGENESH\_EST v2.6<sup>71</sup>, and GenomeScan v1.0<sup>72</sup>. EST and protein support scores, and down-weighting by overlaps with repetitive regions, were used to determine and select the highest-scoring predictions for each locus. PASA was subsequently used to improve gene models by adding UTRS, splice junctions, and alternative transcripts. The transcripts were selected if its Cscore (homology and coverage weighted gene model score) and protein coverage were ≥0.5, unless >20% of the CDS overlapped with repeats, in which case the Cscore threshold was increased to ≥0.9 and homology coverage to >70%. Finally, gene models with protein were annotated with >30% PFAM TE domains were removed.

**Comparative genomics.** To infer paralogous collinear blocks, we ran orthofinder v2.3.11<sup>73</sup> on pairwise diamond v0.9.36<sup>74</sup> blast-like hits pruned to the top two-bit score hits per gene for each pairwise combination of pecan (‘Pawnee’ v1.1), English walnut (Chandler v2.0<sup>25</sup>), maize (RefGen v4<sup>28</sup>) and poplar (*Populus trichocarpa* v3.1<sup>27</sup>). The self-blast hits were pruned to cases where both query and target genes were members of the same orthogroups, then to synteny via MCScanX<sup>75</sup> (–m = 50, –s = 10) and dbScan v1.1–5<sup>76</sup> (radius = 50, min. hits = 10). The number of homeologous collinear blocks were determined as the number of MCScanX breakpoints for each non-redundant combination of off-diagonal (not self-hit chromosomes) chromosome pairs and corrected by the base number of chromosomes in each comparison.

We built the pan-genome annotation using GENESPACE<sup>32</sup>. In short, GENESPACE accomplishes synteny constrained orthology inference across multiples species permitting variable ploidy by parsing protein similarity scores into syntenic blocks and runs orthofinder<sup>73</sup> on synteny constrained blast results. The resulting block coordinates and syntenic orthogroups give high-confidence anchors for evolutionary inference. The five-genome pan-genome annotation (with *J. regia* v2.1<sup>25</sup> as the outgroup) was constructed using default settings (minimum block size (b) = 10, radius / gaps (g) = 20, n. hits / gene / haploid genome = 1). Each orthogroup in the pan-genome representation was a transformation of orthogroup- and synteny constrained blast hits. The pan-genome order and



chromosome ID were taken hierarchically, where each orthogroup was positioned by the most likely syntenic position against the ‘Pawnee’ genome. In the case of orthogroups with a single-copy gene in ‘Pawnee’, the pan-genome location was simply the rank-order location of that gene in the ‘Pawnee’ annotation. For orthogroups with multiple members within a genome, the inferred pan-genome position was taken as the location of the most central gene, calculated as the gene with the highest summed blast bit score across the within-genome blast hits. Ties were broken by physical centrality (closest to the median position of the orthogroup) then gene length. For orthogroups without a representative in ‘Pawnee’, the mean syntenic position of the representative member of each genome was taken as the initial position. Molecular evolution statistics ( $K_a$ ,  $K_c$ ) were calculated from multiple CDS MAFFT v7.470<sup>77</sup> alignments for each single-copy orthogroup in the pan-genome and subsequent analysis in Seqnr 4.2-5<sup>78</sup>.

To define candidate variants between haplotypes within each genome, we projected the closest representative of each pan-genome orthogroup against the alternative haplotype assembly of each genome using gmap v2020-06-30<sup>79</sup>. Protein blast databases between each primary and alternative haplotype annotation were parsed to find the midpoint syntenic location of each alternative haplotype contig. Protein sequences were aligned for each orthologous sequence pair and the percent identity was calculated as  $100 \times (\text{identical positions}) / (\text{aligned positions} + \text{internal gap positions})$ .

It is important to note that, while necessary to compare orthologous sequences within user-defined coordinates among genomes, constraining to synteny may induce a slight reduction in precision of all genome-wide orthogroups. This is because small translocations (<min block size) will not be captured as syntenic regions. We checked this by extracting all 1:1 reciprocal best scoring diamond<sup>74</sup> hits (RBHs) from the blast-like database. Overall, we observed 131,869 unique pairwise RBHs. Of these 131,025 were in the syntenic network, revealing very little loss of precision when constraining to synteny.

Previous comparisons of sequences underlying annotation-based presence-absence variation<sup>32</sup> have found that complete sequence deletions rarely underlie regions that lack a gene model (absences) in PAV orthogroups. More commonly, syntenic absences contain similar or nearly identical sequences that did not satisfy the criteria for calling a gene model. In some cases, these are ‘low-confidence’ genes that barely passed a threshold in the first place. Alternatively, mutations in introns, splice sites or other key positions can reduce evidence for a gene model below a threshold even if the coding sequence is identical. To test for these various patterns of gene absences, we extracted the longest CDS among genes present in an orthogroup and aligned that sequence against the assembly of the genomes containing syntenic absences with gmap<sup>79</sup>, allowing only a single best match and outputting a psl-formatted text file. The psl file was parsed to only alignments on the syntenic chromosome of the orthogroup and percent identity ( $\# \text{ mismatches} / (\# \text{ mismatches} + \# \text{ matches})$ ) and percent coverage ( $\# \text{ matches} / \text{CDS length} \times 100$ ) were calculated. The resulting alignments were categorized as ‘very similar’ (>99% sequence coverage,  $\geq 95\%$  sequence identity) ‘diverged’ (75–99% sequence coverage, 75–95% sequence identity), or ‘absent’ (0–75% sequence coverage or 0–75% sequence identity). Gene counts are summarized in Supplementary Table 2.

**Genomic introgressions.** A total of 30 DNA samples, extracted using the Qiagen DNeasy Plant kit (Qiagen, Inc., Valencia, CA), were resequenced at a median depth of 55 $\times$  (range 38 $\times$ –214 $\times$ , Supplementary Table 7), encompassing the four reference genomes, 13 of their relatives, five ‘true pecan’ genotypes that were known to have little or no interspecific admixture, and eight outgroup samples (*C. cordiformis* = 2, *C. aquatica* = 3, *C. myristiciformis* = 3). The samples were sequenced using Illumina HiSeq paired-end sequencing (2  $\times$  150 bp) at the HudsonAlpha Institute for Biotechnology (Huntsville, AL). The reads were mapped to the ‘Oaxaca’ assembly using bwa-mem v0.7.12<sup>80</sup>. The resulting bam file was filtered for duplicates using Picard v2.19.0 (<http://broadinstitute.github.io/picard/>). Multi-sample SNP calling was accomplished with SAMtools v1.9<sup>81</sup> mpileup (-Q 20 -d 500) and Varscan v2.4.3<sup>82</sup> with a minimum coverage of 8 and a minimum alternate allele count of 4.

To infer the position and identity of genomic introgressions, we pruned the SNP dataset to sites with a minor allele count of  $\geq 3$ , no missing data, and maximum linkage disequilibrium  $r^2 \leq 0.999$  within 100-SNP windows via bcftools v1.9<sup>83</sup>. The pruned vcf was transformed into reference allele counts (0/1/2). Proportion of ancestry ( $P_0$ : *C. aquatica* = 0.021,  $P_1$ : *C. cordiformis* = 0.014,  $P_2$ : *C. myristiciformis* = 0.075,  $P_3$ : pecan = 0.890) was inferred with SNPRelate<sup>84</sup>.

To infer positions and ancestry of introgression regions, we ran Ancestry\_HMM v0.94<sup>85</sup>, which leverages allele frequencies in putative parental populations to determine regions of likely introgressions in a test population via a hidden Markov model. For the Ancestry\_HMM run, we assumed a recent history of introgression and subsequent backcrossing to true pecan (-p 0 5 1 -p 1 5 1 -p 2 5 1 -p 3 5 1 -p 3 4 .5 -p 3 3 0.25 -p 3 2 0.125 --ne 1000 --tmax 5 -e 1e-3 --tolerance .01 -g) where population 3 (-p 3) is the true pecan and the three potential introgressing species are populations 0–2. Posterior probabilities were converted into hard calls of the most likely genotype, and genotype blocks were calculated by iteratively culling runs of identical calls from two- to 500-marker blocks.

**Differential expression to scab inoculation.** The commercial pecan cultivar, ‘Desirable’ was used for scab fungal inoculation experiments. Thirty grafted 1-year-

old potted trees were split into two groups (15 trees in each): the control group was mock-inoculated with sterilized diH<sub>2</sub>O while the other group was sprayed until run off with a conidial suspension of scab isolate De-Tif-11 (1  $\times$  10<sup>6</sup> conidia/mL). Trees were placed in a humidity room (cooler with power off, overhead light, and several humidifiers running, 24–27 °C) to maintain free moisture on leaf surfaces for 48 h. Trees were removed and placed in a warehouse with diffuse overhead light provided by interspersed clear ceiling panels (12 h day length, ambient humidity, 20–29 °C) for the remainder of the experiment. Both control and treatment groups were divided into 3 subgroups of 5 trees to serve as replicates. At 24 h post inoculation, 2 leaflets from each tree were collected and frozen with liquid nitrogen. Thus, for each group, there were 3 replicates each containing 10 leaflets (2 each from 5 seedlings). The 24 h time point was chosen to both control for diurnal/circadian gene-expression regulatory patterns and capture the early molecular responses to the presence of the fungus that may be critical in understanding host susceptibility.

Total RNA was isolated and purified from the leaf tissues using the Norgen Plant/Fungi Total RNA Purification Kit (Norgen Biotek Corp., Tharold, Ontario, Canada). 150-bp paired-end sequencing was performed using Illumina HiSeq platform (Illumina, San Diego, CA). Raw reads were checked for quality with FastQC v0.11.8 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), adapter trimmed, and filtered for quality and length with Trimmomatic v0.36<sup>86</sup> with default parameters. Processed reads were aligned against the SILVA rRNA database for eukaryotes using Bowtie2 v2.3.4.1<sup>87</sup> to remove any rRNA reads present. Unaligned paired reads were recovered and aligned against the ‘Pawnee’ reference genome via STAR v2.7<sup>88</sup> with default parameters. Read counts per gene were obtained using HTseq v0.9.1<sup>89</sup>. Linear differential gene-expression analysis was performed via Wald contrasts with DESeq2 v1.28.1<sup>90</sup>. Differentially expressed genes were defined as those with Benjamin-Hochberg adjusted contrast  $P$ -value  $\leq 0.05$  and  $|\log_2 \text{ fold-change}| \geq 1.5$  (Supplementary Data 3). Differentially expressed genes were subjected to gene ontology enrichment analysis using Fisher’s exact test in topGO v2.40.0<sup>91</sup>. GO terms were considered significant with Fisher’s exact test of <0.05 (Supplementary Table 6).

**‘Lakota’  $\times$  ‘Oaxaca’ mapping population creation and phenotyping.** Controlled cross progeny were generated by multiple teams using pollen collected from the ‘Oaxaca’ ortet in Byron, GA, and applying it to receptive flowers on multiple cloned accessions of ‘Lakota’ during the spring of 2016 and 2017. Progeny nuts were assigned individual numbers, measured, stratified, and planted in pots in a Brownwood TX, greenhouse in March of 2018. In June of 2018, after diameter and height measurements, the progenies were sampled for DNA analysis and randomized in racks in a pecan scab screening nursery at the NCGR-Carya in Somerville, TX stratified by an orchard of origin. Seedlings were transplanted in March, 2019, into nursery rows that maintained their randomized positions.

Progeny were monitored for various traits including gall incidence in 2019. The species of *Phylloxera* observed was determined from photographs of the galls compared to verified specimens (Michele R. Warmund, personal communication). While small numbers of pecan leaf phylloxera galls (*Phylloxera notabilis*) were observed, the vast majority of galls had morphologies indicating southern pecan leaf phylloxera (*Phylloxera russellae*). Given some ambiguities in the systematics of the gall-forming pests, we have opted to refer to the incidence of galls due to aphid-like insects as ‘phylloxera’ here and elsewhere.

Phylloxera gall incidence was monitored by a single-trained rater from 21 to 24 October 2019 by counting or estimating the number of galls on the most affected leaf (worst phylloxera) and the percent of leaves on the seedling showing any galls (percent phylloxera). The incidence of phylloxera galls in ‘Lakota’ has not been well characterized to our knowledge. Historical documentation shows an unusually high phylloxera susceptibility in ‘Mahan’. However, both ‘Mahan’ and ‘Major’ were noted to have progeny with variable levels of phylloxera gall incidence<sup>92,93</sup>. While phylloxera typically only results in cosmetic damage, the presence of such a powerful QTL in the commercially important pedigree of ‘Lakota’ makes this locus of interest to breeders and researchers interested in understanding and controlling for more economically significant insect pests such as pecan stem phylloxera (*Phylloxera devastatrix*), yellow pecan aphid (*Monelliopsis pecanis*), blackmargined pecan aphid (*Monellia caryella*), and black pecan aphids (*Melanocallis caryaefoliae*).

**Mapping population read mapping and variant detection.** Genomic DNA was isolated using a CTAB based method<sup>94</sup> modified for pecan<sup>48</sup> to extract from approximately 150 mg of tender foliar tissue. Samples were RNase treated and cleaned using the Zymo Genomic Clean and Concentrator Kit (Zymo Research, Irvine, CA). The successful control of pollination was confirmed by the presence of rare alleles contributed by ‘Oaxaca’ at SSR loci Ga39<sup>95</sup> and/or Wga242<sup>96</sup> at which ‘Oaxaca’ is homozygous<sup>23</sup>. In order to generate high-density genetic maps, 143 progeny confirmed at Ga39 and Wga242 were selected for resequencing.

Genetic linkage maps totaling 1,196 cM in length were calculated from 11,491 heterozygous SNP loci segregating in the ‘Lakota’ genome. Due to the relatively high homozygosity of ‘Oaxaca’, linkage maps for ‘Oaxaca’ were not generated. Sequencing reads were mapped to the ‘Oaxaca’ v1.1 reference sequence and variants detected using the pipeline described in Bentley et al.<sup>48</sup> for calling markers from GBS data with the following modifications; the CLC Genomics Workbench

(Qiagen, Germantown, MD) version 12.0.2 was used for mapping. Paired-end reads were trimmed and processed using *Trim Reads 2.3* with the following parameters: quality trim was set to 0.05 with an ambiguous limit of 2, automatic read-through trimming was used, and the first 10 nucleotides and the last 3 nucleotides were removed. The reference sequence used was ‘Oaxaca’ v1.1. Read mapping parameters were modified so that the insertion and deletion cost = 6, insertion open cost was 6, insertion extend cost = 1, deletion open cost = 6, deletion extend cost = 1, and minimum read length required to match the reference = 85%. After read mapping and prior to variant detection, the sequencing reads were locally realigned with 3 passes using the CLC function *Local Realignment 1.2*. The variants detected via this pipeline with minor allele frequencies < 0.05, heterozygous call frequencies > 0.8, missing call frequencies > 0.1, or where more than two alleles were observed were not tested as part of the linkage analysis. Additionally, 1% of loci with abnormally high or low read depths were discarded. SNP markers were named based on the sequence and position in the ‘Oaxaca’ reference sequence. This pipeline was also used to reanalyze the GBS sequencing data from Bentley et al.<sup>48</sup> to call markers in ‘Major’ and ‘Mahan’ (the parents of ‘Lakota’) and determine the origin of the ‘Lakota’ haplotypes.

Informative SNPs were defined as those where ‘Oaxaca’ alternative and primary alignments were monomorphic and polymorphisms existed between ‘Lakota’ primary and alternative alleles. Progeny genotypes were used to phase the informative markers from each chromosome into two clusters following Bentley et al.<sup>49</sup>. After clustering, the historical GBS data of ‘Mahan’ and ‘Major’ from Bentley et al.<sup>48</sup> was used to define the marker phases so that at phase one loci the alternate allele was derived from ‘Mahan’ and at phase to loci the alternate allele was derived from ‘Major’. Markers were subset to one ‘Lakota’ informative testcross marker per phase and 25,000 bp bin prioritizing the markers that demonstrated the greatest agreement with the mean haplotype observed 10 SNPs upstream and downstream of the position. Visualization and manual curation were used to remove remaining loci that demonstrated clear patterns of disagreement with local patterns of recombination.

**Linkage map calculation QTL mapping.** Linkage maps and marker/trait associations were calculated in R/qtl2 v0.24<sup>97</sup> with the subset markers input as a backcross population. Framework linkage maps were calculated using *est\_map* using the Kosambi function with an error.prob of 0.0165 (Supplementary Data 5). Kinship between samples was calculated using *calc\_kinship* and the leave-one-chromosome-out (LOCO) method. Trait associations were calculated using *scan1* with a step of 0.1 and an LMM model. Tracking meiotic recombination in ‘Lakota’ was accomplished with 11,489 SNP loci where ‘Lakota’ was heterozygous and ‘Oaxaca’ was homozygous (‘pseudo-testcross’ loci; Supplementary Data 5). QTL Bayesian credible (95%) confidence intervals were calculated in R/qtl v1.47-9 and projected onto the physical position of the Oaxaca genome.

**Candidate gene identification.** To document polymorphisms between the ‘Lakota’ sequence candidate genes and the other three genomes (Oaxaca, Elliott, Pawnee), we carried over all unique pan-genome gene annotations onto the alternative haplotypes and projected each alternative haplotype contig’s physical positions onto each main haplotype reference sequence. We extracted single-nucleotide and structural variants from aligned orthologous sequences. Unalignable genes in the middle of contigs were defined as ‘absent’ while genes without alternative orthologs in regions that lacked an alternative haplotype contig were assumed to be too homozygous for alternative contig assemblies. Candidate genes were determined to be disease associated by manually evaluating the Uniprot knowledgebase (<https://www.uniprot.org/>) to determine if an orthologue of the best matching *Juglans regia* or *Arabidopsis thaliana* gene had been described as likely to be related to plant immune response functions. Genes containing PFAM motifs PF00931, PF08263, or PF13855 and/or Panther domain PTHR11017 were identified as possible LRR sequences.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this article is available as a Supplementary Information file. Genome assembly and annotation have been deposited in GenBank under BioProjects PRJNA680555 (‘Oaxaca’), PRJNA680556 (‘Pawnee’), PRJNA680557 (‘Lakota’), and PRJNA680558 (‘Elliott’). Genomes and annotations are also available through phytozome: Pawnee, Elliott, Lakota, and Oaxaca. RNA sequencing reads for annotation and fungus-induced gene expression have been deposited under SRA BioProject PRJNA680537. See Supplementary Tables 1, 4, and 7 as well as Supplementary Data 4 for RNA and DNA resequencing short reads SRA identifiers. Resequencing reads for the ‘Lakota’ × ‘Oaxaca’ genetic map were deposited on SRA under BioProject number PRJNA679828. Source data are provided with this paper.

Received: 22 January 2021; Accepted: 7 June 2021;

Published online: 05 July 2021

## References

- Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L. & Gaut, B. S. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* **95**, 4441–4446 (1998).
- Tanksley, S. D. & McCouch, S. R. Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**, 1063–1066 (1997).
- Lemmon, Z. H. et al. Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat. Plants* **4**, 766–770 (2018).
- Naylor, R. L. et al. Biotechnology in the developing world: a case for increased investments in orphan crops. *Food Policy* **29**, 15–44 (2004).
- Hall, G. D. Pecan food potential in prehistoric North America. *Econ. Bot.* **54**, 103–112 (2000).
- Grauke, L. J., Wood, B. W. & Harris, M. K. Crop vulnerability: Carya. *HortScience* **51**, 653–663 (2016).
- Wells, L. Pecan *Phylloxera* (and UGA pecan hotline information). *UGA Pecan Extension* (2015).
- Ring, D. R., Grauke, L. J., Payne, J. A. & Snow, J. W. Tree species used as hosts by Pecan Weevil (Coleoptera: Curculionidae). *J. Econ. Entomol.* **84**, 1782–1789 (1991).
- Harris, M. K., Hunt, K. L. & Cognato, A. I. DNA identification confirms Pecan Weevil (Coleoptera: Curculionidae) infestation of Carpathian Walnut. *J. Econ. Entomol.* **103**, 1312–1314 (2010).
- Thompson, T. E. & Conner, P. J. in *Fruit Breeding* Vol. 10, 771–801 (Springer US, 2011).
- Bock, C. H., Young, C. A., Stevenson, K. L. & Charlton, N. D. Fine-scale population genetic structure and within-tree distribution of mating types of *Venturia effusa*, cause of Pecan Scab in the United States. *Phytopathology* **108**, 1326–1336 (2018).
- Olsen, K. M. & Wendel, J. F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
- Bock, C. H., Brennenman, T. B., Wood, B. W. & Stevenson, K. L. Challenges of managing disease in tall orchard trees – pecan scab, a case study. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resource*. **12**, 1–18 (2017).
- Conner, P. J. Performance of 19 pecan cultivars and selections in Southern Georgia. *HortTechnology* **24**, 407–412 (2014).
- Thompson, T. E. & Grauke, L. J. Genetic resistance to scab disease in Pecan. *HortScience* **29**, 1078–1080 (1994).
- Grauke, L. J., Klein, R., Grusak, M. A. & Klein, P. The forest and the trees: applications for molecular markers in the repository and pecan breeding program. *Acta Horticulturae* **1070**, 109–126 (2015).
- Thompson, T. E. & Conner, P. J. in *Fruit Breeding* 771–801 (Springer, 2012).
- Williams, C. G. in *Molecular Dissection of Complex Traits* (ed. Paterson, A. H.) 81–94 (CRC Press, 1998).
- Lovell, J. T. et al. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* **590**, 438–444 (2021).
- Thompson, T. E. & Hunter, R. E. Pawnee pecan. *HortScience* **20**, 776 (1985).
- Thompson, T. E., Grauke, L. J. & Reid, W. ‘Lakota’ Pecan. *HortScience* **43**, 250–251 (2008).
- Grauke, L. J. Family trees: roots & resilience. *Pecan South* **52**, 12–21 (2019).
- Wang, X. et al. Chloroplast genome sequences of *Carya illinoensis* from two distinct geographic populations. *Tree Genet. Genomes* **16**, 859 (2020).
- Luo, M.-C. et al. Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. *BMC Genomics* **16**, 707 (2015).
- Marrano, A. et al. High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome. *GigaScience* **9**, 1959 (2020).
- Harikrishnan, S. L., Pucholt, P. & Berlin, S. Sequence and gene expression evolution of paralogous genes in willows. *Sci. Rep.* **5**, 292 (2015).
- Hofmeister, B. T. et al. A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial *Populus trichocarpa*. *Genome Biol.* **21**, 287 (2020).
- Schnabe, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
- Swigonova, Z. Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923 (2004).
- Smeekens, J. M., Bagley, K. & Kulis, M. Tree nut allergies: Allergen homology, cross-reactivity, and implications for therapy. *Clin. Exp. Allergy* **48**, 762–772 (2018).
- Elizur, A. et al. Clinical and molecular characterization of Walnut and Pecan Allergy (NUT CRACKER Study). *J. Allergy Clin. Immunol.* **8**, 157–165.e2 (2020).
- Lovell, J. T. et al. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat. Commun.* **9**, 5213 (2018).
- Voorhies. Just where Mexican pecans originate. *Am. Nut J.* (1931).
- Manning, W. E. The genus *Carya* in Mexico. *J. Arnold Arbor.* **30**, 425–432 (1949).

35. Stone, D. E. Affinities of a Mexican endemic, *Carya palmeri*, with American and Asian hickories. *Am. J. Bot.* **49**, 199–212 (1962).
36. Cole, J. R. & Gossard, A. C. Stuart pecan found to be susceptible to scab in Mississippi. *Plant Dis. Rep.* **40**, 156 (1956).
37. Troggio, M. et al. Apple, from genome to breeding. *Tree Genet. Genomes* **8**, 509–529 (2012).
38. Naidoo, S., Slippers, B., Plett, J. M., Coles, D. & Oates, C. N. The road to resistance in forest trees. *Front Plant Sci.* **10**, 1 (2019).
39. Liu, J. J. et al. Limber pine (*Pinus flexilis* James) genetic map constructed by exome-seq provides insight into the evolution of disease resistance and a genomic resource for genomics-based breeding. *Plant J.* **98**, 745–758 (2019).
40. Bock, C. Challenges of managing disease in tall orchard trees—pecan scab, a case study. *CAB Rev.* **12**, (2017).
41. Conner, P. J. & Stevenson, K. L. Pathogenic variation of *Cladosporium caryigenum* isolates and corresponding differential resistance in Pecan. *HortScience* **39**, 553–557 (2004).
42. Demaree, J. B. Behavior of *Cladosporium effusum* (WINT). *J. Agric. Res.* **38**, 363 (1929).
43. Pusztahelyi, T. Chitin and chitin-related compounds in plant–fungal interactions. *Mycology* **9**, 189–201 (2018).
44. Gong, B.-Q., Wang, F.-Z. & Li, J.-F. Hide-and-seek: chitin-triggered plant immunity and fungal counterstrategies. *Trends Plant Sci.* **25**, 805–816 (2020).
45. Frederickson Matika, D. E. & Loake, G. J. Redox regulation in plant immune function. *Antioxid. Redox Signal.* **21**, 1373–1388 (2014).
46. González-Bosch, C. Priming plant resistance by activation of redox-sensitive genes. *Free Radic. Biol. Med.* **122**, 171–180 (2018).
47. Beedanagari, S. R., Dove, S. K., Wood, B. W. & Conner, P. J. A first linkage map of pecan cultivars based on RAPD and AFLP markers. *Theor. Appl. Genet.* **110**, 1127–1137 (2005).
48. Bentley, N., Grauke, L. J. & Klein, P. Genotyping by sequencing (GBS) and SNP marker analysis of diverse accessions of pecan (*Carya illinoensis*). *Tree Genet. Genomes* **15**, 403 (2019).
49. Bentley, N. et al. Linkage mapping and QTL analysis of pecan (*Carya illinoensis*) full-siblings using genotyping-by-sequencing. *Tree Genet. Genomes* **16**, 403 (2020).
50. Sun, M. et al. Aphid resistance in *Capsicum* maps to a locus containing LRR-RLK gene analogues. *Theor. Appl. Genet.* **133**, 227–237 (2019).
51. Liang, D. et al. QTL mapping by SLAF-seq and expression analysis of candidate genes for aphid resistance in cucumber. *Front. Plant Sci.* **7**, 174 (2016).
52. Todesco, M. et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**, 602–607 (2020).
53. Skrivaneck, S., Grauke, L. J., Martin, D., Thompson, T. E. & Harris, M. Relative susceptibility of pecan germplasm to blackmargined aphid. *Southwest. Entomol.* **38**, 33–40 (2013).
54. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
55. Xiao, C.-L. et al. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072 (2017).
56. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
57. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
58. Cheng, H. et al. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
59. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
60. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
61. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
62. Smit, A. & Hubley, R. RepeatModeler Open-2.0. <http://www.repeatmasker.org> (2010).
63. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
64. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2011).
65. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
66. Jain, R. et al. Genome sequence of the model rice variety KitaakeX. *BMC Genomics* **20**, 905 (2019).
67. Paterson, A. H. et al. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
68. Mamidi, S. et al. A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nat. Biotechnol.* **38**, 1203–1210 (2020).
69. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).
70. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.1. <http://www.repeatmasker.org> (2013–2015).
71. Salamov, A. A. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
72. Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816 (2001).
73. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, E9 (2015).
74. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
75. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
76. Hahsler, M., Piekenbrock, M. & Doran, D. dbscan: fast density-based clustering with R. *J. Stat. Softw.* **91**, 1–30 (2019).
77. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
78. Charif D, Lobry J. in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, series Biological and Medical Physics Biomedical Engineering (eds Bastolla, U., Porto, M., Roman, H. & Vendruscolo, M.) 207–232 (Springer Verlag, 2007).
79. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
81. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
82. Koboldt, D. C. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
83. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
84. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
85. Corbett-Detig, R. & Nielsen, R. A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLoS Genet.* **13**, e1006529 (2017).
86. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
87. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
88. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).
89. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
90. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
91. Alexa, A. & Rahnenfuhrer, J. *topGO: Enrichment Analysis for Gene Ontology. R Package Version 2.40.0.* <https://bioconductor.org/packages/topGO/> (2021).
92. Calcote, V. R. Southern pecan leaf phylloxera (Homoptera: Phylloxeridae): clonal resistance and technique for evaluation. *Environ. Entomol.* **12**, 916–918 (1983).
93. Calcote, V. R. et al. *Resistance of Pecan Clones to Phylloxera devastatrix Pergande and P. russellae Stoezel. Special Publication 63–69* (Georgia Agricultural Experiment Stations, 1985).
94. Doyle, J. in *Molecular Techniques in Taxonomy* 283–293 (Springer, 1991).
95. Grauke, L. J., Iqbal, M. J., Reddy, A. S. & Thompson, T. E. Developing microsatellite DNA markers in pecan. *J. Am. Soc. Hortic. Sci.* **128**, 374–380 (2003).
96. Grauke, L. J., Mendoza-Herrera, M. A., Miller, A. J. & Wood, B. W. Geographic patterns of genetic variation in native pecans. *Tree Genet. Genomes* **7**, 917 (2011).
97. Broman, K. W. et al. R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. *Genetics* **211**, 495–502 (2019).

## Acknowledgements

The pecan trees used in this study are part of a large genetically diverse collection housed at the National Collection of Genetic Resources, Pecans and Hickories, USDA-ARS Pecan Breeding and Genetics, Somerville, and Brownwood, TX, USA, and a Provenance collection located at the USDA-ARS Southeastern Fruit and Tree Nut Research Laboratory in Byron, Georgia, USA. Contact X. Wang ([xinwang.wang@usda.gov](mailto:xinwang.wang@usda.gov)) for access to germplasm used in this study. The work was funded by USDA NIFA SCRI-2016-51181-25408. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy

under Contract No DE-AC02-05CH11231. The work conducted by the U.S. Department of Agriculture Crop Germplasm Research Unit is supported by CRIS projects 3091-21000-039-00D and 3091-21000-042-00D. This research was supported by funds from the USDA-ARS project funds to CPM (CRIS project 6054-43440-046-00D) and by an ARS Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) (DE-SC0014664) and USDA. We wish to thank Michelle R. Warmund for her assistance in identifying the species of *Phylloxera* observed across the 'Lakota' × 'Oaxaca' mapping population.

### Author contributions

J.T.L., N.B.B., G.B., K.C., C.M., M.M., R.H., P.E.K., L.J.G., J.S., and J.J.R. wrote the manuscript with contributions from all authors. J.T.L., N.B.B., G.B., J.W.J., A.S., S.M., C.P., K.C., M.J.M., Y.A., C.B., N.K., C.M., M.S., J.C., S.S., P.E.K., J.G., J.R., and J.S. conducted genome/statistical analyses. J.J.R., J.S., C.M., P.C., L.W., N.B.B., R.S.H., P.E.K., Y.A., L.J.G., C.R., X.W., M.J.M., K.K., M.S., R.H., C.B., and C.P. designed and executed experiments. P.C., L.W., N.B.B., L.J.G., X.W., and K.K. conducted phenotyping and field research. J.J.R., H.S.R., K.C., Y.A., J.W., S.R., L.B., and N.B.B. conducted molecular and lab work.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-24328-w>.

**Correspondence** and requests for materials should be addressed to J.T.L., J.S. or J.J.R.

**Peer review information** *Nature Communications* thanks Bruno Contreras-Moreira, Nathaniel Street, Lihong Xiao and other, anonymous, reviewers for their contributions to the peer review of this work. Peer review reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021