



Comprehensive resequencing reveals new genetic polymorphisms applicable to wild and cultivated pecan

Jordan Brungardt¹ · Yanina Alarcon² ·
Maria J. Monteros³ · Warren Chatwin⁴ ·
Jennifer J. Randall⁵ · Clive H. Bock⁶

Received: 13 September 2025 / Accepted: 27 October 2025 / Published online: 15 December 2025
This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025

Abstract Pecan (*Carya illinoensis* (Wangenh.) K. Kock) is a high value nut crop cultivated in many parts of the world. Several types of genetic markers have been developed for pecan and are useful for understanding pedigrees of native trees and improved cultivars. Presented here is a publicly available 12.8 thousand marker-set developed from Illumina next-generation sequencing of sixty-three cultivars and 871 trees from a pecan provenance collection. The resulting information gives unprecedented resolution of the genetic relationship between wild pecan populations as well as that of improved cultivars.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10722-025-02651-1>.

J. Brungardt · C. H. Bock (✉)
USDA-ARS, Byron, GA, USA
e-mail: Clive.bock@usda.gov

Y. Alarcon · M. J. Monteros
Noble Research Institute, Ardmore, OK, USA

M. J. Monteros
Bayer Crop Science, Chesterfield, MO, USA

W. Chatwin
USDA-ARS, College Station, TX, USA

J. J. Randall
New Mexico State University, Las Cruces, NM, USA

C. H. Bock
USDA-ARS, Fort Pierce, FL, USA

Population statistics derived from the provenance collection will help direct conservation efforts for the species. The impacts of the identified genetic variants were also evaluated at the amino acid level using an in silico approach. In addition, the amplification-refractory mutation system was used to highlight an *ad-lib* method for developing molecular markers from the resequencing information to help identify unknown individuals.

Keywords Pecan · *Carya illinoensis* · GBS · Cultivar identification · Amplification-refractory mutation system · Genetic relatedness

Introduction

Pecan (*Carya illinoensis* (Wangenh.) K. Kock; *C. illinoensis*) is a hardy deciduous tree native to the United States and Mexico; however, there is debate as to how much influence prehistoric humans had on seed dissemination and pecan's native range (Reviewed in: Manning 1949). Today pecan is planted in many parts of the world for agricultural purposes due to its large, thin-shelled nuts that contain a nutritious and healthy nutmeat. The pecan nutmeat is a good source of protein, fatty acids, carbohydrates, minerals, dietary fiber, essential vitamins, and antioxidants (Kellett et al. 2019; USDA-FDC 2019). During particularly favorable conditions, revenue generated from pecan production in the United States

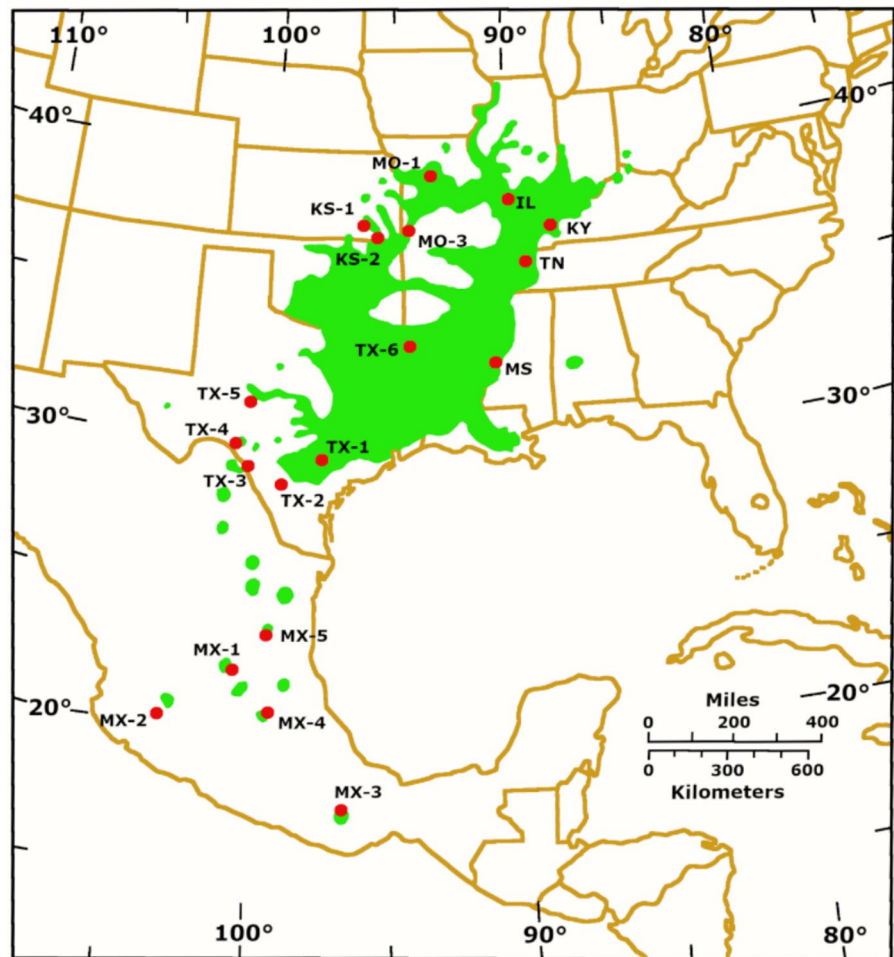
alone can exceed \$700 million (USDA-NASS; referring to 2017).

Genetic markers play a pivotal role in contemporary agriculture. The use of such markers can increase the efficiency of breeding programs, help identify genetic elements responsible for a given trait, and elucidate pedigrees of individuals. The first molecular-based investigations focusing on pecan were performed to determine the genetic diversity of the species and to help identify specific cultivars. Early molecular marker work on pecan was limited to examining only a handful of molecular traits at a time (Marquard 1987, 1991, R  ter 1999). Robert D. Marquard contributed much to this early work by examining protein variants. When examining cultivars for variations of a single enzyme, it was found that only one cultivar of sixty-five could be fully resolved (Marquard 1991). Marquard was also able to confirm/refute historical evidence of pecan pedigrees for several important cultivars by evaluating two proteins (Marquard 1987). Expanding on this body of knowledge, R  ter et al. (1999) assessed sixteen allozymes from 926 provenance trees, 377 wild trees, and 167 trees representing eighty-four cultivars to answer questions about genetic diversity of the species. The provenance trees used by R  ter et al., also used in the current work, is an extensive collection that originated from seed collected from throughout the species native range (Grauke et al. 1989). Two such sister collections were established with one at the USDA Southeastern Fruit and Tree Nut Research Laboratory (SEFTNRL) in Byron, Georgia and the other at the Brownwood pecan and hickory repository in College Station, Texas (part of the USDA-ARS National Plant Germplasm System). The Byron, GA provenance collection utilized in this study currently consists of 872 pecan trees. These trees are the progeny of ninety-two individuals collected from nineteen locations/provenances (Fig. 1). It is important to note that both provenance sister collections at the Georgia and Texas locations use the same designation system. For example, all trees labeled TX2-1 come from the second Texas pecan provenance and are all offspring of tree number one of that provenance.

As new technologies became available, marker-sets grew in complexity to give better resolution for identifying specific genotypes. In 2001, a set of 100 randomly amplified polymorphic DNA (RAPD) markers was developed from twenty-five primer

pairs by Conner and Wood (2001) to resolve forty-three cultivars of pecan. In a similar approach, Grauke et al. (2003) developed eleven primer pairs to amplify microsatellites for evaluating forty-eight pecan and hickory individuals. Monteros et al. 2017 developed a system for identifying sixty-two pecan cultivars from sixteen primer pairs using high resolution melt-curve analysis. Subsequent introduction of genome assemblies for pecan as well as the cost effectiveness associated with next generation sequencing (NGS) has increased the number of markers that are available by several orders of magnitude. Restriction-site associated DNA (RAD) and genotyping-by-sequencing (GBS) are two popular sequencing methods used for identifying polymorphic markers. Both methods employ the use of restriction enzymes (REs) during library preparation. Markers developed from these techniques are enriched in genomic regions associated with restriction-sites corresponding to the REs used during library preparation. This technology has been widely used for genotyping plant species including several close relatives of pecan (Ulaszewski, Meger et al. 2021, Ren et al. 2023). RAD-seq has previously been performed on 860 individuals of the *Carya* genus which resulted in 7,643 markers (Ishimori et al. 2021). Bentley et al. 2019 performed GBS on 108 pecan individuals, mostly cultivars, to identify 87,446 single nucleotide polymorphisms (SNPs) using a contig-level assembly (Jenkins et al. 2015). Several chromosome-level assemblies for pecan have recently been developed, which allows marker-sets to be more powerful when combined with analyses such as genome-wide association studies (GWAS) (Lovell et al. 2021). Using such an assembly, Wang et al. (2024) used GBS to identify 121,948 SNPs in 550 half-siblings of twenty-six open pollinated trees located in Texas. More recently Bhattarai et al (2025) performed GBS on a mapping population of 119 ‘Elliott’ and ‘USDA-ARS-Pawnee’ siblings generating markers specific to those cultivars. The aim of the current research is to identify novel polymorphisms in the pecan genome and to genotype the well-established provenance collection at the USDA-ARS Byron, GA location which is used for large scale phenotyping research. These markers will also aid in determining the relatedness among wild and cultivated pecan.

Fig. 1 Provenance collection maternal locations (shown in red). Pecan's native range is shown in green. Reproduced from Ruter et al. (1999) with corrections from Grauke et al. (2011)



Materials and methods

Plant materials

Leaf samples from 871 trees were collected from the pecan provenance collection located in Byron, GA (Grauke et al. 1989). These samples were collected from June to August of 2017. Leaf material was also collected for sixty-three individuals representing sixty-one historical cultivars from the pecan cultivar collection of the USDA-ARS National Plant Germplasm System – Brownwood located in College Station and Brownwood, TX. Leaf material for an ambiguous individual of interest was also collected from Louisiana (LSUunk). Cultivar samples for GBS were collected on September 15th, 2015. All samples consisted of a pooled sample of several young leaves, approximately $\frac{3}{4}$ expanded, that were collected from

each tree and placed into plastic bags prior to being put into coolers with ice. Specific information about individual trees used in the study are provided in Online Resource 1, tab taxaInfo. Both sets of samples were subsequently processed in the same manner. Samples were freeze dried for three days after storage at -80°C . A total of 15 mg of leaf tissue from each tree was ground with a Tissue Lyser (Qiagen, Valencia, CA) three times at 30 Hz for 1 min. each. Genomic DNA (gDNA) was extracted separately for each sample using a Qiagen Plant DNAeasy® Mini Kit according to the manufacturers protocol with minor modifications. Specifically, samples were centrifuged for 7 min. at 20 k-g after the lysis step, and only one elution step using 50 μL of buffer was used for DNA extraction. DNA was quantified using a Quant-iT PicoGreen dsDNA assay kit (Life Technologies Corp., OR, USA) on a Tecan Infinite M200

Pro microplate reader (Tecan Group Ltd., Männedorf, Switzerland) prior to diluting to 10 ng μL^{-1} .

Genotyping-by sequencing (GBS)

Genotyping was performed using a NextSeq™500 sequencing system (Illumina, CA, USA) at the Oklahoma Medical Research Foundation (OMRF) Genomics Facility using 75 base pair (bp) single-end reads. A protocol described by Elshire et al. (2011) using *ApeKI* was used for library preparation. Samples were sequenced on ten flow cells, each including a negative control. Computational analysis was performed on the USDA-ARS SCINet Ceres high performance computing cluster. Raw reads were first filtered with Trimmomatic v0.38 for reads having a mean q-score of 28 or greater (Bolger et al. 2014). Genetic polymorphisms/site modifications were identified using the Tassel5 v5.2.37 GBSv2 pipeline (Glaubitz et al. 2014). The 87MX3-2.11 genome, hereafter referred to as the Phytozome 573 genome or simply “573 genome”, was used as the reference genome (Lovell et al. 2021). Information required for assigning reads to individuals with Tassel5 can be found in the Online resource 1, tab barcodeInfo. The resulting sites were filtered with VCFtools v0.1.16 for a minimum depth of eight followed by sites missing from 60% of taxa/individuals (Danecek, Auton et al. 2011). An iterative site/taxa filter scheme was performed in an attempt to retain the most taxa and sites. This was performed with Tassel5 GUI by removing taxa missing 80% of sites, sites missing from 10% of taxa, and finally taxa missing 40% of sites. The VCFtools package was then used to remove sites with a minor allele frequency (MAF) of less than one percent and a Hardy–Weinberg equilibrium p-value of less than $1\text{E-}09$ as a method for removing sequencing errors (Gomes et al. 1999). The minimum MAF was chosen to keep minor alleles common to the smallest represented provenance (MX2). Kinship values were calculated using the normalized identity-by-descent (Normalized_IBS) function of Tassel 5 GUI. Trees in the provenance collection were considered mislabeled if an individual had a negative kinship value with a majority of trees from within its provenance. Mislabeled trees were removed from the variant call format (vcf) file resulting in the final marker-set, subsequently referred to as marker-set 867. A second marker-set was generated to evaluate the effect of

sequencing depth on the final number of sites present after using the same filters. This was done by removing four flow cells (P1 to P4), consisting of 384 individuals, with relatively low sequencing depth (Online Resource 2, Figure S1). This high-depth marker-set is subsequently referred to as marker-set 467. Both marker-sets are available in the supplemental material as vcf files (Online Resources 3 & 4; vcf reference, Danecek, Auton et al. 2011).

Downstream analysis

Multidimensional scaling (MDS) was performed by using the distance matrix and MDS functions of Tassel5. Results were visualized with RStudio v2024.9.0.375 (Posit 2024) using the ggplot package (Wickham 2016). Cultivar kinship values were compared to relatedness values generated by Bentley et al. (2019) using the Spearman method, ‘cor.test’ base statistics function, of R. The number of populations (K) for the provenance collection was accessed using the fastStructure v0.0.0 chooseK function evaluating K five through twenty-four (Raj et al. 2014). Measures of genetic diversity π (π), θ (θ), and Tajima’s D were calculated with Tassel5 using the 867 marker-set, while F-statistics (F_{st}) for provenances and regions were calculated with VCFtools and the 867 marker-set (Danecek, Auton et al. 2011). Individuals making up the Northern region included all provenances except for those from the Texas and Mexico provenances. Genome-wide effects of SNPs were analyzed with the program snpEff (Cingolani et al. 2012). This was performed by generating a snpEff database using the 573 genome and 573 genome annotation files (Lovell et al. 2021). Analysis was run on primary transcripts only using Bash commands and seqtk v1.2 to modify database input files accordingly (Li 2012). In silico digestion of the 573 genome was performed with HiCEXplorer v3.7.6 for evaluating REs to be used in future sequencing efforts (Ramírez et al. 2018).

ARMS marker development

For amplification refractory mutation system (ARMS) reactions, leaf material was collected from trees at the Byron, GA USDA-ARS location in 2024. Leaf material came from selected blocks (Historical, Cape Fear/Elliott, Old Chestnut, and

Provenance blocks). Material was collected into 1.5 µL microfuge tubes, placed into liquid nitrogen, and subsequently stored at -80 °C. The material was then freeze-dried overnight. Isolation of gDNA was performed using a cetyltrimethylammonium bromide (CTAB) method developed by Porebski et al. (1997). PCR reactions pertaining to ARMS reactions used *Taq* based Thermo Fisher Scientific (Waltham, MA) 2X PCR Master Mix. Information about primers used in ARMS reactions can be found in the Online Resource 1, tab ARMSprimers. Development of ARMS primers and amplification conditions follow methods developed by Stephen Little (1995). High-quality SNPs identified by GBS which contained SNPs within 30 bp were excluded from consideration for marker development. Rare SNPs for trees of interest were identified for the cultivar ‘Desirable’ and the provenance TN1-1 unique identifying number (UID) 497 individual by first generating a VCF file containing an arbitrarily small diverse subset of cultivar and provenance trees (Online Resource 1, tab smAbitrarySet). The VCFtools *vcf-contract* function was subsequently used to find novel genotypes and rare alleles for the trees of interest.

Results

The number of reads produced from Illumina sequencing was approximately 4.451 billion consisting of 338.3 gigabases. Raw sequences are available through the National Center of Biotechnology Institute’s Sequence Read Archive (NCBI-SRA) as bioproject accession PRJNA1208079. The number of barcoded reads assigned to individuals was significantly different between flow cells P1-P4, those excluded from marker-set 467, compared to that of flow cells P5-P10 (Least Significant Difference test; $P < 0.05$; Online Resource 2, Figure S1). The average number of reads assigned to individuals in marker-set 867 was approximately 3.85 million and increased to around 4.2 million per individual for set 467. The total number of polymorphisms identified by Tassel5 was around 1.1 million when using all flow cells as input, with a modest reduction of 7.5% when using only the flow cells that had higher sequencing depth (P5-P10). The effect of each filtering step on the number of markers and individuals for both marker-sets is presented in Table 1. The greatest percentage of sites removed from the filtering process was due to removing sites with a minor allele frequency of less than 1% (a 93% and 91% reduction for marker-sets 867 and 467 respectively). Site modifications in the

Table 1 Statistics for variant call format (VCF) files after passing various filters

Filter	Large set (Flow cells P1-P10)		High-depth set (Flow cells P5-P10)	
	Ind.	Sites	Ind.	Sites
Tassel5 output	944	1,102,236	560	1,019,544
Depth 8 & site missingness >40%	944	567,447	560	576,897
Iterative taxa/site missingness	916	339,093	546	402,850
Sites w/ minor allele freq. <1%	916	22,586	546	37,874
HWE 1E-9	916	18,730	546	34,347
Site missingness >4%	916	12,820	537	29,975
Mislabeled trees removed	888	12,820	514	29,975

The high-depth set is a subset of individuals with greater sequencing depth. The yellow, orange, and green boxes represent statistics for marker-sets 462, 467, and 867 respectively.

freq., frequency, w/, with, HWE, Hardy–Weinberg equilibrium, Ind., individuals

867 and 467 marker-sets after filtering totaled 12,820 and 29,975 respectively. For both marker-sets, 99.6% of the polymorphisms reported were SNPs. The mean allelic depth per marker per individual was greater for the initial 467 marker-set output compared to that of the 867 marker-set at 25.5 versus 22.4 respectively; however, after filtering was complete, this statistic favored the 867 marker-set at 71.2 while the 467 marker-set had an allelic depth of only 56.5. Not surprisingly, a high proportion of sites in set 867 are found in set 467 (94.4%).

For marker-set 867 the average distance between markers was 49.5 kilobases (kb) with the largest gap between markers being on chromosome thirteen at 1.88 megabases (Mb) (Online Resource 2, Figure S2). Both these statistics decreased for set 467 with the average distance between markers being 21.2 kb and the largest gap between markers being 1.52 Mb on chromosome eight. Of the 30k markers in set 467, 12,580 have gaps greater than 1 kb and 5,167 are 10 bp or closer to neighboring markers. For marker-set 867 the number of variants falling within exons was 6,625 which are predicted to cause thirty-nine nonsense and 3,383 missense mutations. Of these, the program snpEff predicted there to be 355 high-impact variants corresponding to sixty-two primary transcripts. This number increased to 668 high-impact variants (125 primary transcripts) in the 467 marker-set. The high-impact variants for set 467 are listed in the Online Resource 1, tab 467highImp. When accounting for polymorphisms that are closely linked to predicted genes, those within 1.5 kb, in the 573 genome the 867 marker-set accounts for 18.8% of the genes (Online Resource 2, Table S1). This increases to 31.9% for the 467 marker-set.

When evaluating kinship values twenty-eight provenance trees were classified as mislabels (Online Resource 1, tabs kinshipMatrix & putative_mislabels). No particular pattern was observed for samples identified as being mislabeled, with twelve of the nineteen provenances having mislabel representatives. No large consecutive blocks of trees in the provenance grove were represented as mislabels; however, almost half of the trees identified as a mislabel had a neighboring tree within the block row with a provenance matching that of the mislabel's suggested provenance. No particular patterns for mislabels were present on library plates. Based on the relatedness results, it appears that incorrect placards on some of

the provenance trees were responsible for some mislabeling. Two trees identified as being mislabeled, 18d26MX5d1 and 25d24MX1d2, were correctly identified to their orchard map designated provenance through evaluation of highest mean kinship values while having placards that matched the labels of the data set. One individual appears to have obtained a clerical error somewhere between sample collection and library sequencing. For example, mean kinship highest value, orchard map designation, and placard are all in agreement for the individual 11d7TX4d4 as being a part of provenance MX4-4 rather than TX4-4.

After removing mislabels, MDS analysis showed provenances clustering nicely, with cultivar individuals interspersed mostly within the KS, MX5, TX2, TX3, and TX6 provenance clusters (Fig. 2A). In general, MDS analysis showed broad clustering into three groups; one group comprised primarily of Mexico provenance trees, another of Texas provenance trees minus the TX6 provenance, and the last cluster with the remaining United States provenances and the TX6 provenance (Fig. 2B). As mentioned before, most trees from the Mexico provenances cluster to a unique group with a notable exception for the MX5 provenance, which had two distinct sub-clusters, one of which was interspersed throughout an area dominated by cultivar trees (Fig. 2C). The top principal components for the MDS analysis explains 56.8% and 32.0% of the variance in the data. When examining northern provenances via MDS analysis only two provenances show distinct clustering, KY1 and MO1, with some of the progeny of KY1-1 maternal lineage clustering to the broader northern group (Online Resource 2, Figure S3). Generally, there does not appear to be a pattern between provenance latitude and MDS clustering/genetic distance. One possible exception to this are provenances bordering the Mississippi river (IL1, TN1, and MS1; Online Resource 2, Figure S4). Despite the general lack of correlation between latitude and clustering, provenances located closer to one another geographically typically cluster closer to each other.

Kinship values generated by this research are highly corroborative with historical information for different cultivars (Reviewed in: Conner and Wood 2001). For example, 'Burkett' is said to be a native tree coming from Texas, 'Giles' from Kansas, 'Major' from Kentucky, and 'Starking Hardy Giant' (SHG) from Missouri. Of these 'Burkett', 'Giles', and 'SHG'

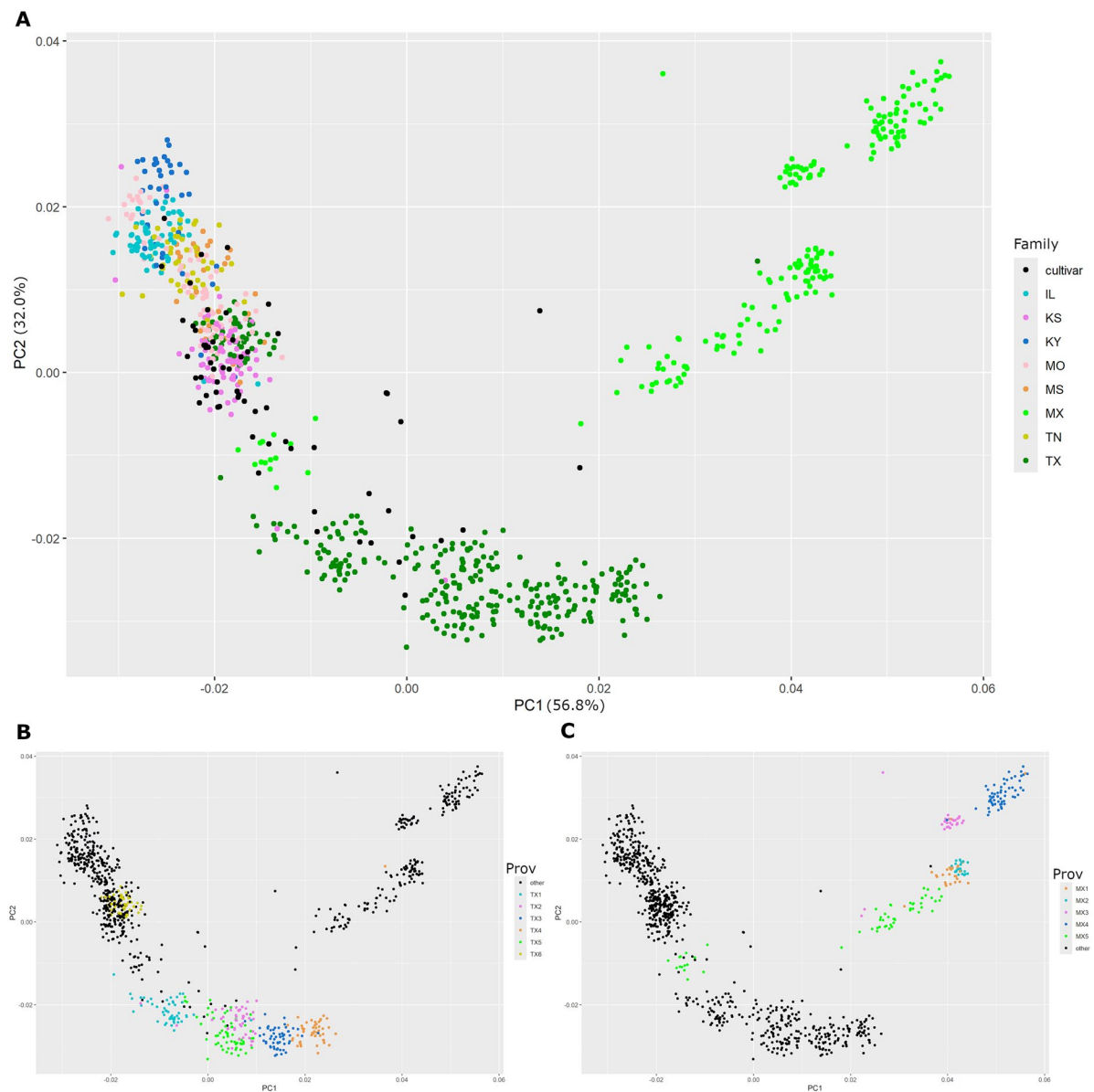
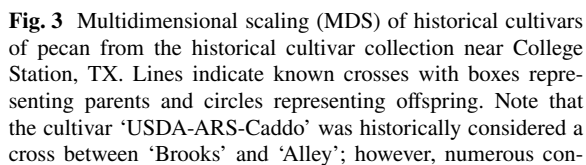


Fig. 2 Multidimensional scaling (MDS) for Byron, GA provenance and College Station, TX cultivar trees. Mislabels removed from analysis. MDSs with emphasis on region level (A), Texas provenances (B), and Mexico provenances (C).

have the highest mean kinship values with provenances that aligns with this information, TX5 (0.052), KS2 (0.056), and MO1 (0.081) respectively; however, ‘Major’ had a higher kinship value with the Illinois provenance (0.120) compared to that of the Kentucky provenance (0.079). In addition to this, ‘Western Schley’ was derived from open pollinated ‘San Saba’ in San Saba, Texas (Crane, Reed et al. 1937). Not surprisingly ‘Western Schley’ has the highest mean

provenance kinship with TX5 (0.113), which is San Saba, Texas’s geographically closest provenance.

Known crosses of pecan are also represented accurately in MDS analysis with offspring situated between their parents (Fig. 3). One exception to this is ‘USDA-ARS-Caddo’. This cultivar is said to be a cross between ‘Brooks’ and ‘Alley’; however, both Conner and Wood’s (2001) results from RAPD analysis, and more recent GBS data by Bentley et al.



(2019), suggests ‘Schley’ as a more likely parent of ‘USDA-ARS-Caddo’ compared to ‘Alley’. The data generated for the present research also suggests this, with ‘Brooks’ and ‘Schley’ having the highest kinship values with ‘USDA-ARS-Caddo’, 0.562 and 0.593 respectively, compared to that of ‘Alley’ (0.207).

to differences in anatomical features. Furthermore, Monteros et al. (2017) showed these individuals to be genetically different. ‘Dependable’ individuals of the historical collection in Byron, GA also show dramatic differences in fruit anatomy compared to each other and the ARM04 reaction consistently shows a third genotype associated with this cultivar of pecan (Table 2). ‘Dependable’ 8_3 and ‘USDA-ARS-Shoshoni’ show close proximity in MDS, have few private alleles (seventeen and twelve), and a high kinship value of 1.03). It is likely ‘Dependable’ 8_3 is a mislabel as ‘USDA-ARS-Shoshoni’ has high kinship values with its known parents ‘Odom’ and ‘Evers’. The two ‘Jackson’ individuals sequenced for this work have also been shown to have different genotypes with the ‘JacksonBW’ individual considered to be true to the cultivar (Unpublished Communication: Grauke 2009).

Table 2 Comparing genotyping-by-sequencing (GBS) and amplification-refractory mutation system (ARMS) genotypes

Reaction	USDA-ARS-Caddo	CapeFear	Desirable	Elliott	USDA-ARS-Pawnee	147_TX6-4	170_MX4-2	262_IL1-2	284_TX1-5	497_TN1-1	560_KY1-3	680_MX5-3	702_MX2-1	719_TX4-1	Dep6-2	Dep8-3	Dep17-2
ARM01	0/0	0/0	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/1 NA	0/1	0/0	0/1
ARM03	0/0	1/1	0/0	0/0	1/1	0/0	1/1	1/1	0/1	0/1	0/1	0/0	0/1	0/1	0/0	0/1	0/0
ARM04	0/1 0/0	0/0	1/1	0/0	0/1	0/0	1/1	0/1	0/0	0/0	0/0	0/0 0/1	0/0	0/1	0/1	0/1	1/1
ARM06	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/0	0/0 0/1	0/0	0/0	0/0	0/0	0/0
ARM07	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/1	1/1	0/1 0/0	0/0	0/0	0/0	0/0	0/0
ARM09	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/1	0/0	0/0	0/0	0/0	0/0	0/0
ARM12	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0
												#		*	*	**	

Reference and alternate alleles are depicted as zero and one respectively. Boxes outlined in red indicate ARMS-GBS genotype calls that are in disagreement with the ARMS genotype in subscript. Some markers did not give ARMS results for certain individual-marker combinations (not available, NA). For cultivars GBS data was generated from College Station, TX trees and ARMS data from Byron, GA trees. Labels for provenance trees are unique identifier (UID) numbers followed by provenance and mother tree.

#Individual identified as a mislabel through GBS data. *'Dependable' trees from the College Station, TX cultivar collection (GBS only data). **'Dependable' tree from Byron, GA historical collection (ARMS only data)

When comparing kinship values for cultivars sequenced in both Bentley et al. (2019) and this work, there is a high degree of correlation between the two sets of data (Online Resource 1, tab corrBentley). The lowest correlation between these data is that of 'Major' ($\rho=0.174$); however, 'Major' kinship values with offspring for the current work is what one would expect ('Lakota' = 0.570 and 'Kanza' = 0.545). Cultivar data generated for this work are generally in agreement with historical information and previous genotyping endeavors, including those cultivars with questionable lineages. The 'Dependable' and 'Jackson' information generated for this work should be used with caution.

Population statistics give important information on population structure that can help guide conservation efforts. The number of populations for best describing model complexity for the provenance collection as determined by fastStructure was eight to thirteen. Several population genetics statistics for the trees sequenced for this research are presented Table 3. Measures of diversity varied from one statistic to another. The fixation index (F_{ST}) is a measure

of genetic variance between two populations while nucleotide diversity (π) is a measure of genetic diversity within a population (Nei and Li 1979; Weir and Cockerham 1984). At the regional level, Texas trees had the most alternate alleles followed by the Northern region and lastly the Mexico region (12,046, 11,606, and 9,330 respectively). The provenance with the highest amount of nucleotide diversity was KS2 ($\pi=0.158$). This aligns with the Rüter et al. (1999) assessment of 16 loci for the provenance collection in regards to mean number of alleles per locus. The provenance with the lowest nucleotide diversity was MX3 ($\pi=0.0887$). When assessing the provenance collection by broader regions, the Northern region had the highest nucleotide diversity ($\pi=0.1622$) and Mexico trees had the lowest ($\pi=0.1332$). Despite the low nucleotide diversity present in most Mexico provenances, MX2 and MX3 provenances had the highest genetic variance compared to the overall population (Weighted F_{st} 0.1229 and 0.1225 respectively). The TX6 provenance in contrast had the lowest genetic variance compared to the overall population (Weighted F_{st} 0.0259).

Table 3 Population genetics statistics for the pecan provenance collection planted at Byron, GA and historical cultivars planted in College Station and Brownwood, TX

Provenance/region	N	π bp ⁻¹	θ bp ⁻¹	TajimaD	F _{st_mean}	F _{st_weighted}
IL	60	0.1574	0.1575	-0.0028	0.0344	0.0425
KS1	37	0.1569	0.1729	-0.3523	0.0243	0.0316
KS2	40	0.1580	0.1766	-0.3984	0.0220	0.0273
KY1	37	0.1434	0.1563	-0.3153	0.0466	0.0698
MO1	35	0.1452	0.1502	-0.1280	0.0438	0.0618
MO3	39	0.1567	0.1719	-0.3338	0.0220	0.0273
MS1	28	0.1563	0.1777	-0.4776	0.0225	0.0296
MX1	27	0.1244	0.1265	-0.0667	0.0460	0.0751
MX2	22	0.1021	0.0857	0.7875	0.0637	0.1229
MX3	26	0.0887	0.1065	-0.6700	0.0563	0.1225
MX4	58	0.1027	0.0839	0.8111	0.0683	0.1137
MX5	57	0.1443	0.1344	0.2660	0.0340	0.0430
TN1	47	0.1567	0.1684	-0.2572	0.0268	0.0340
TX1	50	0.1494	0.1500	-0.0158	0.0340	0.0463
TX2	47	0.1511	0.1533	-0.0536	0.0286	0.0360
TX3	55	0.1463	0.1400	0.1633	0.0394	0.0509
TX4	48	0.1364	0.1323	0.1141	0.0508	0.0763
TX5	61	0.1509	0.1503	0.0129	0.0315	0.0406
TX6	53	0.1577	0.1644	-0.1484	0.0221	0.0259
Northern	323	0.1622	0.1428	0.4254	0.0283	0.0367
Texas	314	0.1596	0.1491	0.2207	0.0196	0.0234
Mexico	190	0.1332	0.1255	0.2004	0.0430	0.0617
Cultivars	61	0.1575	0.1674	-0.2134	0.0271	0.0333

Regions or states (encompassing multiple provenances) are given at the bottom of the table with the Northern region being comprised of all trees not part of the Texas or Mexico provenances

For the development of ARMS molecular markers, rare alleles for individuals of interest were first identified from a small diverse set of trees. This identified 141 loci with unique genotypes, forty-six of which had a unique allele, for the ‘Desirable’ individuals. For the highly scab resistant individual TN1-1 UID 497, 129 loci were identified with unique genotypes, thirty-one of which had a unique allele. Once developed, ARMS reactions were generally consistent in giving correct genotype

calls compared to the 867 marker-set developed from NGS data (Table 2); however, the addition of internal controls sometimes resulted in non-specific amplification or absence of amplification for the ARMS amplicon regardless of the internal control used (data not shown). In rare cases, the ARMS was unable to call a genotype (Table 2, UID719 ARM01 reaction). Results for a typical ARMS reaction are shown in the Online Resource 2, Fig. S5.

Discussion

The development of a robust set of molecular markers to evaluate a highly diverse group of pecan trees for the current study represents the highest resolution analysis of genetic relatedness for pecan as a species. There has been discussion whether *C. illinoensis* originated in ancient forests in Mexico and moved northward following retreating glaciers or if the species originated in the northern part of its range and moved south along the Mississippi river. For northern provenances there seems to be no spatial pattern associated with genetic relatedness apart from provenance proximity. Analysis by Rüter et al (1999) came to a similar conclusion. Admixture of populations due to human activities could be responsible for some of this ambiguity. This is certainly possible as fossilized pecan has been associated with human artifacts dating as far back as 6100 B.C. (Chadderdon 1983). Also, historical records describe at least one case of pecan being introduced into a new area by humans. A paper by Manning (1949) referenced that many trees in Jaumave, Tamaulipas Mexico had been introduced from Texas. Indeed, information for the current study shows the Jaumave provenance MX-5 trees as having a distinct sub-cluster which maps much closer to the Texas provenance's compared to the rest of the Mexico provenances (Fig. 2 B&C).

The identification of mislabels in the provenance collection highlights the great care that needs to be taken when working with large populations/datasets as they can have dramatic effects on downstream applications. The time at which a tree became a mislabel will greatly affect the difficulty in elucidating its actual parentage. Mislabel genotype data produced for the current research could first be validated via low-throughput sequencing methods such as Sanger sequencing or ARMS of rare alleles. Data not matching would indicate sample mix-up during or after leaf collection for GBS sequencing. The tree UID-680 in the current work is an example of this scenario (Table 2). GBS could be performed again on such individuals and genotype information merged into the marker-sets developed here. Re-sequencing information in agreement with the current data, however, would indicate a mix-up sometime prior to leaf sample collection. Ease of assigning mislabels to the correct provenance will depend largely on the amount of genetic uniqueness of the provenance they belong to.

This could be relatively easy in some cases (Online Resource 2, Fig. S6). Transplanting much of the provenance collection from a high-density plot to a lower-density plot in the mid 2000's may have contributed to some mislabeling given the large number of trees in the collection. Simple errors made during planting, orchard mapping, sample collection, library preparation, well loading, and data analysis could all account for less obvious cases of mislabeling.

Several major goals of any GBS study include determining relatedness, identifying polymorphisms that result in non-synonymous mutations, and to generate high-density marker-sets for GWAS. A larger number of markers having greater genomic coverage aligns with each of these goals. Both GBS research previously performed on pecan produced more markers compared to this work (Bentley et al. 2019; Wang et al. 2024). Several factors could account for this. High-throughput sequencing projects are often a balance between sequencing depth and cost of sequencing. As the number of individuals to be sequenced increases, sequencing depth may become less of a priority due to financial limitations. With this in mind, the present research sequenced almost twice the number of individuals as the larger Wang et al. (2024) study. Re-sequencing with longer read lengths, as opposed to the 75 bp single end reads used here, could drastically increase the number of polymorphisms identified. A number of GBS investigations in plants have generated marker-sets several orders of magnitude greater as those generated for the current research (Zhang et al. 2021; Patiranage et al. 2022; Wang et al. 2022). This is most likely due to fewer numbers of individuals being genotyped and less stringent filters being used. Several investigations using filters similar to those used here have resulted in comparable numbers of markers (Juliana et al. 2019, Ulaszewski, Meger et al. 2021).

Many GBS methodologies depend on restriction digestion of genetic material to enrich for markers within genomic regions that are gene-rich. Genotyping-by-sequencing protocols using different restriction enzymes (RE) can have drastic effects on the number and location of markers detected. This is due to the inability of REs to digest certain parts of the genome into fragments of sizes that are suitable for NGS. In silico digestion of the genome-of-interest is therefore an important tool for evaluating the potential efficiency a given GBS restriction enzyme has for

identifying markers (Fu et al. 2016). Results for in silico digestion of the 573 genome with different REs can be found in the online resource 2, Table S1. This suggests certain REs are better suited for producing genetic markers from the pecan genome compared to others. For example, the theoretical level of gene representation produced by *ApeKI*, the RE used in this work, is around six times greater than that of the *PstI* RE used in previous pecan GBS endeavors (Bentley et al. 2019; Wang et al. 2024). Bentley et al. (2019) were able to mitigate for lower levels of sequenceable fragments by adding a sonication shearing step after RE digestion. The marker-sets developed for this research were substantially less efficient at identifying site modifications associated with genes compared to the theoretical potential of *ApeKI*, with a minimum difference of 52.3%. This could indicate that greater sequencing depth is required to realize *ApeKI*'s full potential. Markers for this work were developed with site completeness in mind, which is amiable to performing GWAS. As a result, many sites with true alternate alleles could have been removed from the final marker-sets. Less filtered precursors could be a better representation of the theoretical potential of *ApeKI*. For example, the percentage of gene-associated site modifications for marker-set 462 starts to approach the theoretical gene representation of *ApeKI* with only a 20.9% difference. Regardless, marker-sets 467 and 867 did exceed the theoretical potential of *PstI* with at least a 5% increase in genes represented.

A phenotype of particular interest in pecan is resistance to the fungal disease pecan scab, caused by *Venturia effusa* (*V. effusa*). Scab is the most important disease of pecan in hot and humid climates such as those found in the southeastern United States. Countless marker-sets developed from NGS data generated for the present work showed few reproducible GWAS results when comparing four years of Byron, GA provenance scab phenotype data (Bock et al. 2016; Bock et al. 2020; GWAS results not shown). The marker-sets evaluated used numerous combinations of filtering as well as exclusion of cultivar variant information. Lack of reproducibility could be due to the population make-up of *V. effusa* being different from one year to the next as the species has been shown to have different pathovars (Conner 2002). A sexual cycle leading to additional genetic variants could also complicate matters (Charlton et al. 2020). Additional scab ratings along with detailed records of *V. effusa* population dynamics in the

provenance collection could be used along with markers developed here to elucidate mechanisms responsible for successful defense to specific pathovars of *V. effusa*. Several individuals from the provenance collection have been identified as being highly resistant and highly susceptible to scab development across four different years of scab rating (Brungardt et al. 2024). Of the 3,445 moderate to high-impact site variants identified by the software snpEff for marker-set 867, no alternate allele was found to be exclusive to either of these groups (data not shown).

ARMS markers were specifically developed to detect successful crosses of the pecan scab resistant provenance tree UID-497 with susceptible cultivars. The ARMS markers developed here will help identify these progenies. Development of a mapping population with UID-497 as a parent will help identify genetic components that are responsible for resistance to *V. effusa*. Additionally, markers developed for this work could help to fine map already developed mapping populations of pecan (Bhattarai et al. 2025). Similar molecular markers can be developed from the 867 and 467 marker-sets for identifying other trees of interest and their progeny. As an example of the 467 marker-sets versatility, the melt-curve analysis system for identifying cultivars described by Monteros et al. (2017) was not able to resolve the cultivars 'Philema' and 'Desirable' using sixteen different markers. In contrast, the 467 marker-set has sixty-seven loci, each of which are able to differentiate the two either through ARMS or another genotyping system. Limitations of the ARMS system may require the use of Sanger sequencing or qPCR for genotyping a marker of interest. Failure of ARMS to genotype particular individuals could be due to the presence of extremely rare alleles within the primer binding regions. The number of trees to be genotyped may or may not justify using ARMS for genotyping. While primer development for ARMS is technically more demanding compared to those used in Sanger sequencing and qPCR, the combination of qualitative results and requirement of less specialized equipment will make ARMS an attractive genotyping system for some.

Conclusion

Enhanced understanding of genetic diversity and improved genetic resources will play an essential

part in developing pecan cultivars that are drought resistant, salt tolerant, have higher and more consistent yield, and are more resistant to pests and disease. The current study gives the highest resolution to date of the genetic relatedness among wild and cultivated pecan. It is also the first large scale genetic information available for the cultivars ‘Carter’, ‘USDA-ARS-Cheyenne’, ‘Humble’, ‘Jenkins’, ‘USDA-ARS-Mohawk’, ‘Philema’, ‘Podsednik’, ‘Syrup Mill’, and ‘Witte’. This work shows that most provenances exhibit distinct genetic characteristics despite potential introduction of pecan into new areas by humans. These genetic resources can also guide germplasm conservation efforts. The genome site modifications identified in this work represent the first large-scale, publicly available marker-set that can be used in a wide array of applications such as improving primer design, marker-assisted selection, and as an essential part of GWAS when using the trees genotyped. The polymorphic information can be used to make inferences about closely related species as well. Furthermore, post-hoc development of ARMS markers was performed to highlight a cost-effective method for detecting SNPs present in novel individuals.

Acknowledgements The authors would like to thank Dr. Larry J. Grauke and Dr. Bruce Wood for their work developing the pecan provenance collections. High performance computing resources used for this work were provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D. This article reports the results of research only. Mention of a trademark or proprietary product is solely for the purpose of providing specific information and does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products that may also be suitable.

Authors’ contributions JJB, experimental design, ARMS development, data analysis, final manuscript write-up YA, experimental design, sample collection, data analysis, preliminary manuscript write-up MJM, experimental design, sample collection, manuscript editing, funding acquisition WC, experimental design, manuscript editing JJR, experimental design, manuscript editing, funding acquisition CHB, experimental design, manuscript editing, funding acquisition.

Funding JJB and CHB were funded by the USDA-ARS projects 6042–21220-013–000-D and 6042–21220-014–000-D. WC was funded by the USDA-ARS project 3091–21000-046–000-D. MJM and YA were funded by NIFA-Specialty Crops Research Initiative “Coordinated Development of Genetic Tools for Pecan” (Grant no. 2016–51181-25408), The Oklahoma Department of Agriculture, Food and Forestry, and the Noble Research Institute. JJR was funded by the

NIFA-Specialty Crops Research Initiative “Coordinated Development of Genetic Tools for Pecan” (Grant no. 2016–51181-25408) and NIFA-Specialty Crops Research Initiative “Trees for the future: Coordinated development of genetic resources and tools to accelerate breeding of geographic and climate adapted pecan trees” (Grant no. 2022–51181-38332). Additional support was provided by the Genomics Core Facility at Noble Research Institute, LLC for genotyping.

Data availability Raw sequence files can be found at the National Center of Biotechnology Institute’s Sequence Read Archive (NCBI-SRA) as bioproject PRJNA1208079. The 467 and 867 marker-sets are available in vcf format as Online Resources 3 and 4.

Declarations

Conflict of interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bentley N, Grauke LJ, Klein P (2019) Genotyping by sequencing (GBS) and SNP marker analysis of diverse accessions of pecan (*Carya illinoensis*). *Tree Genet Genomes* 15(1):8
- Bhattarai G, Cao S, Bentley N, Klein P, Bock CH, Pisani C, Conner PJ (2025) High-density linkage mapping and identification of quantitative trait loci associated with scab resistance in pecan. *J Amer Soc Hort Sci* 150(2):86–103
- Bock CH, Grauke LJ, Conner P, Burrell SL, Hotchkiss MW, Boykin D, Wood BW (2016) Scab susceptibility of a provenance collection of pecan in three different seasons in the Southeastern United States. *Plant Dis* 100(9):1937–1945
- Bock CH, Alarcon Y, Conner PJ, Young CA, Randall JJ, Pisani C, Grauke LJ, Wang X, Monteros MJ (2020) Foliage and fruit susceptibility of a pecan provenance collection

- to scab, caused by *Venturia effusa*. CABI Agric Biosci 1(1):19
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120
- Brungardt J, Alarcon Y, Shiller J, Young C, Monteros MJ, Randall JJ, Bock CH (2024) Transcriptome profile of pecan scab resistant and susceptible trees from a pecan provenance collection. *BMC Genomics* 25(1):180
- Chadderdon, M. F. (1983). Baker cave, Val Verde county, Texas: The 1976 excavations, Center for Archaeological Research. 13.
- Charlton ND, Yi M, Bock CH, Zhang M, Young CA (2020) First description of the sexual stage of *Venturia effusa*, causal agent of pecan scab. *Mycologia* 112(4):711–721
- Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6(2):80–92
- Conner PJ (2002) A detached leaf technique for studying race-specific resistance to *Cladosporium caryigenum* in pecan. *J Amer Soc Hortic Sci* 127(5):781–785
- Conner PJ, Wood BW (2001) Identification of pecan cultivars and their genetic relatedness as determined by randomly amplified polymorphic DNA analysis. *Journal of the American Society for Horticultural Science* 126(4):474–480
- Crane, H. L., C. A. Reed and M. N. Wood (1937). Nut breeding. Yearbook of agriculture, 1937. G. Hambige, United States Department of Agriculture: 827–887.
- Danecek P, Auton A, Abecasis C, Albers E, Banks M, A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin and G. Genomes Project Analysis (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6(5):e19379
- Fu YB, Peterson GW, Dong Y (2016) Increasing genome sampling and improving SNP genotyping for genotyping-by-sequencing with new combinations of restriction enzymes. *G3 Genes/genomes/genetics* 6(4):845–856
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9(2):e90346
- Gomes I, Collins A, Lonjou C, Thomas NS, Wilkinson J, Watson M, Morton N (1999) Hardy–Weinberg quality control. *Ann Hum Genet* 63(6):535–538
- Grauke LJ, Payne JA, Wood BW (1989) North American pecans: a provenance study. *Annu Rep North Nut Grow Assoc* 80:124–131
- Grauke LJ, Iqbal MJ, Reddy AS, Thompson TE (2003) Developing microsatellite DNA markers in pecan. *J Amer Soc Hortic Sci* 128(3):374–380
- Grauke LJ, Mendoza-Herrera MA, Miller AJ, Wood BW (2011) Geographic patterns of genetic variation in native pecans. *Tree Genet Genomes* 7(5):917–932
- Grauke, L. J. (2009).
- Ishimori M, Takanashi H, Fukami K, Cervantes K, Nagano AJ, Kajiya-Kanegae H, Grauke LJ, Tsutsumi N, Randall J, Iwata H (2021) Genome-wide RAD-Seq analysis revealed subpopulation structures of the pecan (*Carya illinoensis*) germplasm collection and their relationship to geographical distribution patterns, International Society for Horticultural Science (ISHS). Belgium, Leuven
- Jenkins J, Wilson B, Grimwood J, Schmutz J, Grauke LJ (2015) Towards a reference pecan genome sequence, International Society for Horticultural Science (ISHS). Belgium, Leuven
- Juliana P, Montesinos-López OA, Crossa J, Mondal S, González Pérez L, Poland J, Huerta-Espino J, Crespo-Herrera L, Govindan V, Dreisigacker S, Shrestha S, Pérez-Rodríguez P, Pinto Espinosa F, Singh RP (2019) Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. *Theor Appl Genet* 132(1):177–194
- Kellett ME, Greenspan P, Gong Y, Pegg RB (2019) Cellular evaluation of the antioxidant activity of U.S. pecans [*Carya illinoensis* (Wangenh.) K. Koch]. *Food Chem.* <https://doi.org/10.1016/j.foodchem.2019.04.103>. (1873–7072 (Electronic))
- Li, H. (2012). Seqtk, toolkit for processing sequences in FATA/Q formats.
- Little S (1995) Amplification-refractory mutation system (ARMS) analysis of point mutations. *Curr Protoc Hum Genet* 7(1):9.8.1–9.8.12
- Lovell JT, Bentley NB, Bhattacharai G, Jenkins JW, Sreedasyam A, Alarcon Y, Bock C, Boston LB, Carlson J, Cervantes K, Clermont K, Duke S, Krom N, Kubenka K, Mamidi S, Mattison CP, Monteros MJ, Pisani C, Plott C, Rajasekar S, Rhein HS, Rohla C, Song M, Hilaire RS, Shu S, Wells L, Webber J, Heerema RJ, Klein PE, Conner P, Wang X, Grauke LJ, Grimwood J, Schmutz J, Randall JJ (2021) Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. *Nat Commun* 12(1):4125
- Manning WE (1949) The genus *Carya* in Mexico. *J Arnold Arbor* 30(4):425–432
- Marquard RD (1987) Isozyme inheritance, polymorphism, and stability of malate dehydrogenase and phosphoglucose isomerase in pecan. *J Am Soc Hortic Sci* 112(4):717–721
- Marquard RD (1991) Inheritance of phosphoglucose isozymes in pecan. *HortScience* 26(9):1213–1214
- Monteros, M. J., Y. Alarcon and C. Rohla (2017). Methods for identifying pecan tree cultivars. U. S. o. America. United States of America. US-20170268072-A1: 1–46.
- Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* 76(10):5269–5273
- Patiranage DSR, Rey E, Emrani N, Wellman G, Schmid K, Schmöckel SM, Tester M, Jung C (2022) Genome-wide association study in quinoa reveals selection pattern typical for crops with a short breeding history. *Elife* 11:e66873
- Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol Biol Rep* 15(1):8–15

- Posit (2024). RStudio: integrated development environment for R, Posit Software, PBC.
- Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197(2):573–589
- Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* 9(1):189
- Ren J, Wang Y-A, Zhou X-K, Xie K-W, Han F-J, Peng H, Liu X-Y (2023) Analysis of genetic diversity and development of a core collection in walnut (*Juglans regia* L.) germplasm native to china via genotyping-by-sequencing. *Agronomy*. <https://doi.org/10.3390/agronomy13071786>
- Rüter B, Hamrick JL, Wood BW (1999) Genetic diversity within provenance and cultivar germplasm collections versus natural populations of pecan (*Carya illinoensis*). *J Hered* 90(5):521–528
- Ulaszewski, B., J. Meger and J. Burczyk (2021) "Comparative analysis of SNP discovery and genotyping in *Fagus sylvatica* L. and *Quercus robur* L. using RADseq, GBS, and ddRAD Methods." *Forests* 12 <https://doi.org/10.3390/f12020222>.
- USDA-FDC. (2019). "Food Data Central; Nuts, pecans, dry roasted, without salt added." Retrieved 11.23.24, 2024, from <https://fdc.nal.usda.gov/food-details/170183/nutrients>.
- USDA-NASS. "Statistics by subject: Pecans, utilized, in shell - production, measured in \$." Retrieved 10.20.22, 2022, from https://www.nass.usda.gov/Statistics_by_Subject/result.php?EB7A7D83-A799-32A3-836F-0C54A8513C38§or=CROPS&group=FRUIT%20%26%20TREE%20NUTS&comm=PECANS.
- Wang J, Ye H, Zhou H, Chen P, Liu H, Xi R, Wang G, Hou N, Zhao P (2022) Genome-wide association analysis of 101 accessions dissects the genetic basis of shell thickness for genetic improvement in Persian walnut (*Juglans regia* L.). *BMC Plant Biol* 22(1):436
- Wang X, Stein L, Black M, Kubenka K, Randall J, Ding C (2024) Phenotypic diversity and population structure of Pecan (*Carya illinoensis*) collections reveals geographic patterns. *Sci Rep* 14(1):18592
- Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6):1358–1370
- Wickham H (2016) ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York
- Zhang M-Y, Xue C, Hu H, Li J, Xue Y, Wang R, Fan J, Zou C, Tao S, Qin M, Bai B, Li X, Gu C, Wu S, Chen X, Yang G, Liu Y, Sun M, Fei Z, Zhang S, Wu J (2021) Genome-wide association studies provide insights into the genetic determination of fruit traits of pear. *Nat Commun* 12(1):1144

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