

SCIENTIFIC REPORTS

OPEN

Microsatellite Borders and Microsequence Conservation in *Juglans*

Aziz Ebrahimi¹, Samarth Mathur², Shaneka S. Lawson³, Nicholas R. LaBonte¹, Adam Lorch², Mark V. Coggeshall³ & Keith E. Woeste¹ 

Received: 9 January 2018

Accepted: 21 December 2018

Published online: 06 March 2019

Walnuts (*Juglans* spp.) are economically important nut and timber species with a worldwide distribution. Using the published Persian walnut genome as a reference for the assembly of short reads from six *Juglans* species and several interspecific hybrids, we identified simple sequence repeats in 12 *Juglans* nuclear and organellar genomes. The genome-wide distribution and polymorphisms of nuclear and organellar microsatellites (SSRs) for most *Juglans* genomes have not been previously studied. We compared the frequency of nuclear SSR motifs and their lengths across *Juglans*, and identified section-specific chloroplast SSR motifs. Primer pairs were designed for more than 60,000 SSR-containing sequences based on alignment against assembled scaffold sequences. Of the >60,000 loci, 39,000 were validated by e-PCR using unique primer pairs. We identified primers containing 100% sequence identity in multiple species. Across species, sequence identity in the SSR-flanking regions was generally low. Although SSRs are common and highly dispersed in the genome, their flanking sequences are conserved at about 90 to 95% identity within *Juglans* and within species. In a few rare cases, flanking sequences are identical across species of *Juglans*. This comprehensive report of nuclear and organellar SSRs in *Juglans* and the generation of validated SSR primers will be a useful resource for future genetic analyses, walnut breeding programs, high-level taxonomic evaluations, and genomic studies in *Juglandaceae*.

Walnuts (*Juglans*) are a genus of perennial trees and shrubs consisting of 21 species (Flora of North America) distributed in North America, South America, Eurasia, Central America, and the Caribbean¹. Several members of the genus are important sources of edible nuts and wood, including *Juglans regia* (Persian walnut) which is grown as a crop in every country of the world with a temperate climate, and *Juglans nigra* (black walnut), the most valuable North American hardwood. Other North American *Juglans* species include *Juglans major* (Arizona black walnut), which grows in hot, arid areas near the border between the United States (U.S.) and Mexico, and *Juglans cinerea* (butternut), which along with black walnut grows in the eastern deciduous forests of the U.S. All New World *Juglans* belong to section Rhysocaryon. Aside from *J. regia* (sect. Dioscaryon), Asian *Juglans* belong to section Cardiocaryon. Species in this section include *Juglans mandshurica*, which is native to Korea and China, and *Juglans ailantifolia*, native to Japan. Species hybrids within *Juglans* are common, often fertile and vegetatively vigorous, and used as rootstocks for nut production or for timber.

Plant genomes contain a large proportion of non-coding repetitive DNA, including transposable elements, retroelements, non-LTR retroelements, tandem repeats, long and short interspersed nuclear elements, and micro- and mini-satellites^{2,3}. Microsynteny among congeners for these types of repeated sequences is evidence for their conservation over evolutionary time and their role in genome evolution^{2,3}. The characterization of synteny and microsynteny among congeners or other phylogenetically related groups for genes, or groups of genes, is an important objective of comparative genomics⁴. Flanking regions of microsatellites (SSRs) exemplify a type of synteny for non-genic and non-repetitive sequences that is important because when flanking regions are shared across species, PCR primers can be designed that amplify (presumably) homologous regions. Cross-species amplification of SSRs is used often in studies of species for which no SSRs have been published but SSRs from relatives are available^{5,6}. The extent of genome-scale synteny and sequence conservation within genera at loci containing SSRs is not well understood⁷, although the sharing of LTR-retrotransposons and other repeat elements may be an important factor driving genome size variation across species and genera⁸. SSR markers are readily

¹Department of Forestry and Natural Resources, Hardwood Tree Improvement and Regeneration Center, Purdue University, 715 State Street, West Lafayette, IN, 47907, USA. ²Department of Biological Sciences, Purdue University, West Lafayette, IN, 47907, USA. ³USDA Forest Service, Northern Research Station Hardwood Tree Improvement and Regeneration Center 715 West State St., West Lafayette, IN, 47906, USA. Correspondence and requests for materials should be addressed to A.E. (email: aebrahi@purdue.edu) or K.E.W. (email: kwoeste@fs.fed.us)

Accession	Species	Total sequence ^a (reads (million))	Coverage depth (X)	nuSSR ratio ^b	nuSSRs loci ^c	cpSSRs loci ^d	mtSSRs loci ^d
'Chandler' ^e	<i>J. regia</i>	500	120	438,430	124,333	30	34
693	<i>J. regia</i> -1	82	12.63	13,236	2,808	30	34
694	<i>J. regia</i> -2	91	14.02	78,811	14,519	30	34
Lugar farm	<i>J. regia</i> -3	42	6.47	10,613	2,275	30	34
910	<i>J. mandshurica</i>	101	15.56	71,552	18,662	36	32
New Mexico	<i>J. major</i>	48	7.39	30,067	6,727	25	28
Purdue 1	<i>J. nigra</i>	131	20.18	34,927	9,029	12	26
OS-20	<i>J. cinerea</i>	73	11.25	20,650	4,587	16	32
1096	<i>J. ailantifolia</i>	61	9.40	64,193	14,597	24	36
654	<i>J. × intermedia</i>	5	1.0	1000	233	28	30
208	<i>J. × quadrangulata</i>	67	10.32	12,444	2,556	30	32
863	<i>J. regia</i> BC1	40	6.16	13,536	3,013	18	32
123 Rossville ^f	<i>J. × cinerea</i>	38	5.85	11,116	2,380	26	38
Transcriptome ^g	<i>J. regia</i>	0.7	—	23,596	6,379	—	—
Total		1,860		824,170	212,098	344	436

Table 1. Frequency of motifs and nuSSRs in all evaluated genomes. ^aMillion reads; ^bRatio of cumulative sequence length of all SSR to genome size; ^cNumber of SSR present in nuclear genome (loci with designed primers); ^dSSRs found in chloroplast and mitochondrial genomes; ^ePublished Persian walnut genome (*J. regia*); ^f*J. cinerea* × *J. ailantifolia* backcross; ^g*J. regia* transcriptome used for comparison.

identified from both transcriptome and whole genome data⁹. Nuclear SSRs (nuSSRs) exhibit codominance, high multiplex potential, and many have high levels of polymorphisms. These characteristics make nuSSRs preferable for high throughput mapping, population genetic analysis, and marker-aided plant improvement techniques such as marker-assisted selection in tree breeding programs¹⁰, genetic diversity and gene flow analysis^{11,12}, and quantitative trait loci (QTL) analysis¹³ in plant species.

The primary role of SSRs in plant evolution is unclear, but studies have shown that the distribution of motif frequencies and microsatellite density across the genome is not equal among species¹⁴ and seemingly non-random¹⁵. The dominant occurrence of motif patterns, motif repeats, and specific sequences and lengths in plant genomes likely result from selection pressures applied on that specific motif during evolution¹⁶. Aside from their use in population genetics and breeding (described above), SSRs can also be used for studies of synteny among species within one family and for comparisons of genome organization and evolutionary relationships across species. Except for *Juglans nigra* and *Juglans regia*^{17–19}, few efforts have been made to generate SSR primer sets for other valuable *Juglans* species²⁰.

Our study utilized paired-end sequencing (Illumina Inc., San Diego, CA) to identify and evaluate SSRs in 12 *Juglans* genomes. We sequenced and assembled genomes of several *Juglans* species to identify nuclear (nuSSR), mitochondrial (mtSSR), and chloroplast (cpSSR) microsatellites and compare their frequency and distribution within and across various *Juglans* genomes. We also aimed to evaluate SSR-flanking region sequence similarity across walnut species. Finally, we sought to design unique SSR primers for amplification in a single species or homologous loci in multiple *Juglans* species.

Results and Discussion

Library preparation and paired end sequencing data, frequency of motif types. We used paired-end libraries to generate 977,379,442 reads (940 GB) of raw sequence data from all *Juglans* genotypes in our study. Quality control filtering yielded 787,083,476 reads (723 GB). Thus, 77% of data were used for downstream analyses (Table 1). The genomes of 12 walnut genotypes were sequenced and assembled using the draft nuclear genome, organellar genome, and transcriptome of *J. regia*²¹ as a reference to guide assemblies. Nuclear and organellar genomes of all 12 *Juglans* genotypes and the *J. regia* transcriptome were analyzed to identify simple sequence repeat (SSR) motifs. Motifs which repeated from two to four times and over a minimum length of 10 bp were selected for analysis (Supplementary Table S1). In *Juglans*, di-nucleotide and tri-nucleotide SSRs comprise 98% of total SSRs (Fig. 1). The AT/TA motif was the most frequent motif in the *J. regia* reference genome and across all of the 12 newly sequenced *Juglans* genomes (Fig. 2), but the frequency of this motif varied considerably; in *J. mandshurica*, *J. regia* and *J. major* it was 21%, 18%, and 4% respectively (Figs 1, 2, Supplemental Table S1). Our estimates of motif frequency were influenced by sequence quality and depth of coverage; for example, the frequency of a common motif was not consistent among different samples of *J. regia* or within sections of the genus. In general, the GC/CG motif was among the least represented of the dinucleotide repeats within *Juglans* genomes. The percentage of GC/CG motifs was relatively high in *J. nigra* (27%) from section *Rhysocaryon*, but much lower in another *Rhysocaryon*, *Juglans major* (10%). The frequency of GC/CG was surprisingly consistent in three *Cardiocalyx* species: *J. mandshurica* (17%), *J. cinerea* (17%), *J. ailantifolia* (17%), and among the samples of *J. regia* in section *Dioscaryon* (10%). Recent data indicated this specific repeat motif is rare in most hardwood trees, *Capsicum* species, *Arabidopsis*, rice, and wheat^{17,22}. In *Capsicum*, the AG/CT repeat was the most abundant SSR²²; this motif was also common in the *J. regia* reference, but not in the other *Juglans* genotypes (Fig. 2A),

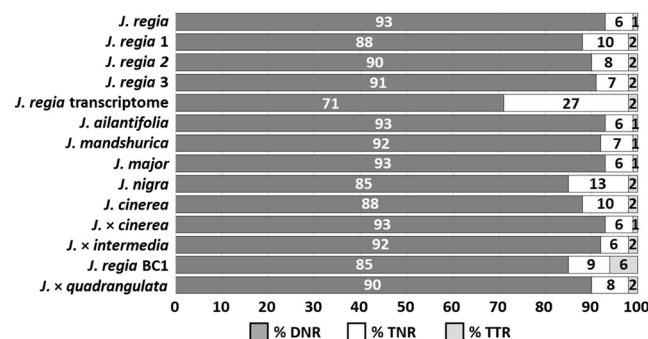


Figure 1. Frequency of SSR motifs in all examined genomes. (DNR, di-nucleotide; TNR, tri-nucleotide; TTR, tetra-nucleotide).

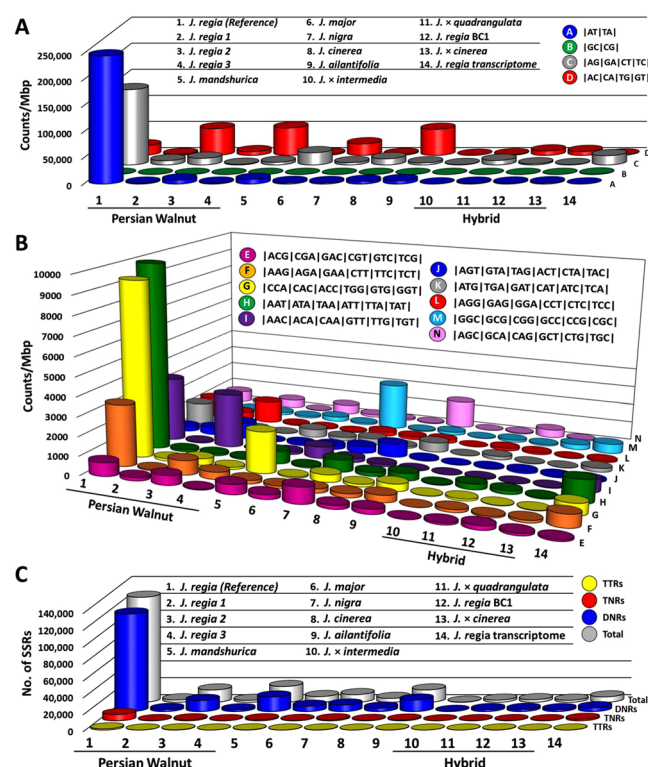


Figure 2. Motif type and frequency. (A) Abundance of di-nucleotide and (B) tri-nucleotide motifs. (C) Total numbers of nuSSRs within the nuclear genome of *Juglans* spp. and the Persian walnut (*J. regia*) transcriptome.

possibly because of low depth of coverage. The AAT/ATA motif was the most prominent tri-nucleotide motif found in *Juglans*. In monocots such as rice (*Oryza sativa*) and wheat (*Triticum spp.*), CCG/CGG comprised the greatest proportion of tri-nucleotide repeats²².

Simple sequence repeat (SSR) motif frequency (among di-nucleotide, tri-nucleotide, and tetra-nucleotide motifs) decreased sharply with increased motif length in both the reference and the newly sequenced genomes (Fig. 1). For example, there were twice as many AT as AAT, and four times as many AG repeats as AAG (Supplementary Table S1). Di-nucleotide motifs account for 88.7% of nuSSRs while tri- and tetra-nucleotide motifs were significantly less abundant at 9.2% and 2%, respectively (Fig. 1). These data were similar to those from *Arabidopsis*, cucumber (*Cucumis sativus*), potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*)²², sorghum (*Sorghum bicolor*)¹⁶, and several hardwood tree genomes¹⁷. Simple sequence repeats (SSRs) with Tri-nucleotide motifs were the most common in monocot genomes²³. Evaluation of nuclear genome motif frequency variation in other species showed that *Arabidopsis* (*Arabidopsis thaliana*), cucumber (*Cucumis sativa*)²², pine (*Pinus taeda* L.)²⁴ and ten other hardwood tree species¹⁷ also showed a decline in motif numbers when motif lengths increased.

nuSSR primers in *Juglans* genomes. A total of 824,170 loci containing SSR were identified in this study. We designed 205,000 SSR primers from the nuclear genome and 6,000 from the transcriptome (Table 1). Total

↓ Database Query →	1	2	3	4	5	6	7	8	9	10	11	12
1 <i>J. regia</i> 1		95.14	95.08	92.88	92.06	92.7	92.85	92.79	92.61	94.6	94.37	92.75
2 <i>J. regia</i> 2	95.16		95.24	92.55	91.91	92.65	92.56	92.56	94.38	94.61	94.99	92.61
3 <i>J. regia</i> 3	95.1	95.28		92.5	91.59	92.75	92.51	92.34	94.13	94.7	94.77	92.44
4 <i>J. mandshurica</i>	92.86	92.59	92.37		91.31	92.27	92.78	94.11	91.79	93.09	92.39	93.77
5 <i>J. major</i>	92.02	91.9	91.54	91.21		93.82	93.04	91.21	92.5	92.9	92.6	92.4
6 <i>J. nigra</i>	92.58	92.63	92.66	92.16	93.81		92.72	92.11	93.84	93.16	93.57	92.32
7 <i>J. cinerea</i>	92.75	92.56	92.47	92.7	93.02	92.75		92.91	91.51	94.04	92.59	93.8
8 <i>J. ailantifolia</i>	92.78	92.58	92.18	94.16	91.29	92.22	93		91.76	93.13	92.21	93.86
9 <i>J. × intermedia</i>	92.71	94.36	94.46	91.88	92.59	93.94	91.51	91.83		93.87	96.36	92.04
10 <i>J. quadrangulata</i>	94.53	94.59	94.55	93.04	92.95	93.15	94	93.09	93.98		93.9	93.4
11 <i>J. regia</i> BC1	94.42	95.08	94.83	92.5	92.64	93.54	92.6	92.35	96	93.96		92.82
12 <i>J. × cinerea</i>	92.66	92.58	92.24	93.7	92.53	92.42	93.85	93.74	91.86	93.49	92.69	

Figure 3. Predicted flanking regions of SSRs from all the walnut species were compared for pairwise mean sequence similarity and mean e-value using BLASTn.

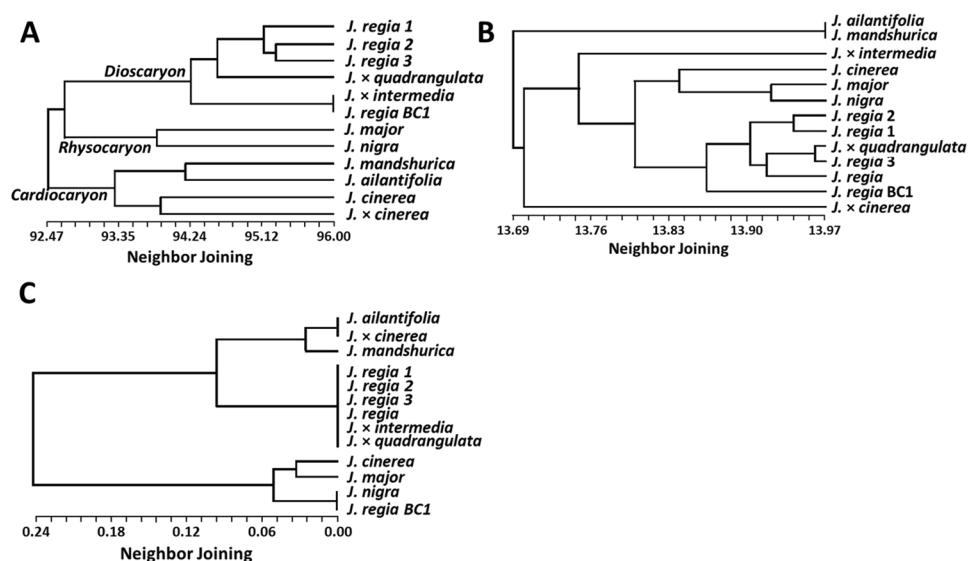


Figure 4. Phylogeny: (A) Phylogeny analyses based on mean similarity of all flanking region of nuSSRs obtained from BLASTN results. (B) Estimate of evolutionary divergence based on SSR-flanking region sequences. (C) Neighbor joining tree based on frequency of SSR motifs within the chloroplast.

SSR motif numbers from the reference genome and transcriptome were considerably higher than from other sequenced *Juglans* species (Fig. 2). For example, the number of SSRs in the *J. regia* reference genome was five times greater than that of *J. mandshurica* and seven times greater than *J. ailantifolia* (Table 1). The fewest SSRs were found in *J. × intermedia* (233) and *J. × cinerea* (2,380). Relatively few SSR loci were found in *J. major* and *J. nigra* (6,727 and 9,027) (Table 1). *J. nigra* had three times more reads than *J. major*, but only about 1.5 X more SSRs. *J. quadrangulata* had more reads than *J. ailantifolia* but *J. ailantifolia* had 5x more SSRs. We identified a total of 189,208 di-nucleotide, 13,826 tri-nucleotide, and 1,990 tetra-nucleotide SSR loci among all 12 *Juglans* genomes examined (Supplementary Table S2a,b). The compiled *Juglans* SSR primer database resulting from this research will provide a rich resource for *Juglans* markers, enable development of in-depth linkage maps, and allow the fine-mapping of QTLs.

Sequence similarities and synteny among SSR-flanking regions. We compared SSR flanking regions for sequence similarity and the presence of SNPs. BLASTN results indicated the pairwise mean sequence similarity of SSR flanking regions ranged from 91% to 96% across all *Juglans* genomes (Fig. 3). The species exhibiting the lowest average similarity to all other samples were *J. major* and *J. mandshurica* (91%, $e\text{-value} = 1.006\text{E-}10$) and the greatest pairwise similarity was between a *J. regia* backcross and *J. × intermedia* (96%). Similarity between pairs of *J. regia* samples was consistently about 95%. Similarity between all *J. regia* samples and related hybrids varied from 94% to 95%.

We compared the similarity of each sample's SSR-flanking region sequences based on BLASTN results. Similarity scores were used for neighbor joining analysis, and the resulting dendrograms showed *Juglans* species sorted into three main groups consistent with their conventional assignment to sections within the genus (Fig. 4A).

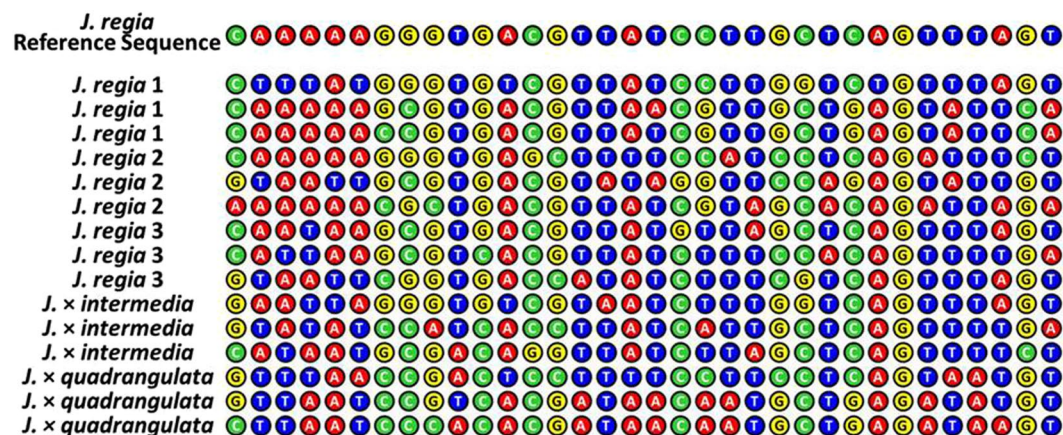


Figure 5. Genome-based SSR alignment. Typical alignment of SSR-flanking regions within *J. regia* and related hybrids).

A *Cardiocaryon* clade contained *J. ailantifolia* and *J. mandshurica*, *J. cinerea*, and *J. × cinerea*. *Juglans regia* genotypes and hybrids sorted into a *Dioscaryon* clade, while *J. nigra* and *J. major* sorted into a *Rhysocaryon* clade. Evolutionary analysis based on SSR-flanking region alignment (Fig. 4B) also sorted *Juglans* genotypes into their respective section with the exception that the closely-related *J. ailantifolia* and *J. mandshurica* could not be distinguished from one another²⁵. The placement of *J. × cinerea*, and *J. × intermedia* based on a Maximum Composite Likelihood model (MCL), which estimates evolutionary divergence, was not the same as their placement based on BLASTN results (Fig. 4A,B). The sample identified as *J. × cinerea* was a hybrid of unknown but probably complex pedigree that included butternut and Japanese walnut. In our analysis, it sorted onto a branch between *J. ailantifolia* and species in other sections of the genus (Fig. 4B). Based on chloroplast data (Fig. 4C), *J. × cinerea* sorted closely with *J. ailantifolia*, probably because the hybrid contains a *J. ailantifolia* chloroplast. Our results are in the agreement with phylogeny results obtained using ITS (Internal Transcribed Spacer) and matK (megakaryocyte-associated tyrosine kinase)²⁶. Our results showed that SSR flanking region sequences are a reliable method for phylogenetic evaluations of *Juglans* species and may represent an improvement over the marker systems used in previous studies because a large number of SSR flanking regions are available within genomes, are most often found in non-coding DNA (and so are likely to be neutral with respect to selection), and are scattered across the entire genome. About 16,000 flanking regions contributed to our phylogeny. Qi *et al.*²⁷ reported that the majority of the genome (91%) is covered by neither tandem repeats nor indels, but much of the remaining genome (about 8%) is associated with SSR flanking regions. If these flanking regions are distributed across a random 8% of the genome, they may be useful targets for future analyses utilizing fosmid libraries or segmented genomes.

Although our NGS data underwent adapter trimming before analysis, we cannot know how many SNPs in the flanking regions were true polymorphisms and how much of the sequence homology was homoplasious. Utilizing SNP calling with full coverage (100x) yields a 3% chance of having a read with SNP-like changes due to sequencing or assembly error²⁷. In our data, the percentage of flanking regions that were identical to *J. regia* was low and varied among species 0.3 to 1.1%; thus, we were unable to design primers that were completely conserved across all species studied (Figs 5 and 6).

Screening of SSR loci, development of unique shared nuSSR markers and genome duplication.

To identify primer pairs that amplify only a single locus, primer sequences which occurred at multiple locations in the genome were filtered out as not useful for most applications²⁸. We used a criterion of 100 percent identity in both primers for identifying duplicated loci, although it is likely that primers with less than 100% identity would cross-amplify non-target sites. Of the ~205,000 SSR-containing loci, we were able to design unique primers with >40% GC content within the flanking regions for ~60,000 loci. About 39,000 of these primer-containing loci were validated by e-PCR (Supplementary Table S4). The majority of validated single-locus primers identified in each species did not match a locus in any other *Juglans* species.

Of the 39,000 primer pairs, only 710 (1.8%) were 96–100% identical to a locus found in all *Juglans* (corresponding to a single nucleotide difference within a 20 bp primer), and no single primer sequence had a 100% match across all species studied (Fig. 6). In some cases, apparently identical loci in different species were similar enough in their flanking regions that we could design an upper primer with 100% identity and a lower primer with high homology (>95%). The use of wobble codes and adjustment of melting temperatures might make these primers practical as universal SSR primers for *Juglans*, or at least for all the species we evaluated. The greatest number of shared primer sets belonged to *J. ailantifolia* and *J. mandshurica*, with 37 shared primer pairs displaying 100% identity at a locus that was present only once in each species. Primers were designed for 63 regions conserved with 100% identity across at least two *Juglans* species (Supplementary Table S5). In all cases, complete primer sequence conservation was observed between only two species. Regions of shared homology with *J. regia*, the only species in the genus with a reference genome, could be useful in understanding genome evolution in *Juglans* and assist in comparative map development, especially if the SSRs associated with them are polymorphic in *J. regia*.

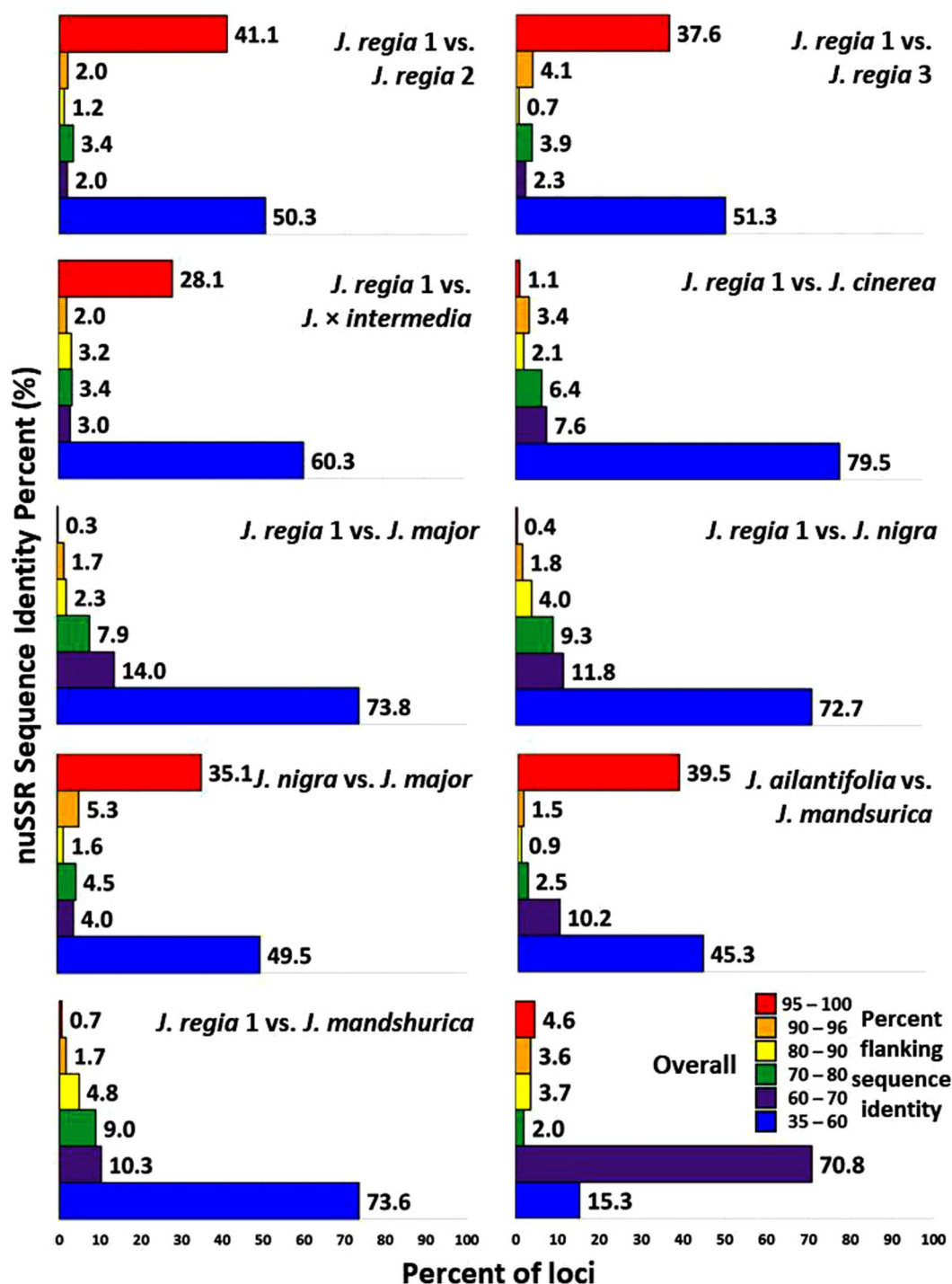


Figure 6. Percent flanking sequence identity. Percent of identical flanking-SSR sequences across *Juglans* genomes. Overall is the comparison of all genotypes versus all other genotypes.

Based on pairwise comparisons among *J. regia* genomes, about half of the SSR regions in Persian walnut are evolving, and the other half remain highly conserved (Fig. 6). The proportion is the same, more or less, when comparisons are made within each section (for example, in *J. ailantifolia* versus *J. mandshurica* or *J. major* vs. *J. nigra*), but comparisons across sections (*J. regia* vs. *J. mandshurica* or *J. regia* vs. *J. major* or *J. cinerea* or *J. nigra*) showed that about 75% of the flanking regions have low similarity (30–60% identity). Across sections, very few loci (less than 1%) show high identity. This demonstrates why it is difficult to identify primers that will amplify well across the entire genus. Heterologous amplification has been demonstrated in *Juglans*, although the cross-species utility of primers varies by locus and by species⁶. The success of heterologous amplification using SSR primers depends generally upon the evolutionary distance between the original species and the tested species, with decreasing success as genetic distance increases²⁹.

The fossil record places the radiation of the *Juglandaceae* to the Paleocene³⁰ (approximately 56 to 66 MYA) and whole genome duplication may have occurred at this time, when the haploid number of chromosomes changed from 8 to 16. Within *Juglans*, it is unknown what percent of the genome is highly conserved across the genus. Factors leading to conservation, divergence or loss of flanking regions of SSRs in *Juglans* are probably similar to those in other species, but in general they are not well understood. The size of the genome, and number of retro-transposons and other mobile elements (which is not known for *Juglans*), may have an important role in genome duplication. Nevertheless, it seems clear that at least 710 (of 60,000) SSR loci in *Juglans* genomes are older than the most recent speciation in their lineage, as they were found across all *Juglans*. Interestingly, about 708 loci were duplicated in *J. regia*, i.e., for these 708 SSRs there was 100% homology in the flanking region at two loci in the *J. regia* genome (Supplementary Table S8). These 708 loci may be so ancient as to predate the genome duplication event for *Juglans*. It is also possible their sequence identity is homoplasious. In some cases, the number of identical flanking regions was as great as 19, but duplicated or multiple copies of a flanking region was generally rare. The frequency of multiple copies of a flanking sequence in a genome was greater in *J. nigra* than in *J. regia* or *J. cinerea*, and considerably greater in *J. mandshurica* and *J. ailantifolia* than in other species we analyzed (Supplementary Table S8). A low percent of the genome was a shared flanking region with 100-percentage identity across 12 *Juglans* genomes; perfect sequence identity was found between two species only. Shared loci between two species most likely originated with the common ancestor of the two species in which they were found, and unique SSRs must have arisen later in evolutionary time. The relationship between SSR duplication within a genome and conservation of SSRs across genomes within a genus is not understood but could ultimately shed light on genome evolution. We suggest that loci with sufficient identity to share primers across two *Juglans* species probably arose prior to speciation, about 1 to 3 million years ago.

Motif frequency in cpSSRs and mtSSRs. The extent of within species variation in chloroplasts and mitochondria in *Juglans* is poorly described, although there is some data based on reasonable sample sizes and sample distribution³¹. The reference genome and twelve additional walnut genotypes sequenced in this study revealed 230 chloroplast and 330 mitochondrial SSRs in the walnut genomes. Total density of SSRs in the chloroplast genome was significantly lower than mitochondrial genomes ($p = 0.001$). There are no cpSSRs in *Juglans* chloroplasts known to be polymorphic within a species, emphasizing the low level of genetic variation found within *Juglans* chloroplasts. All chloroplast polymorphisms identified within *Juglans* species so far are indels or SNPs, many of which result in restriction site differences^{32,33}. Many of these polymorphisms are in SSR-rich regions, but the documented polymorphism is not within the SSR. Comparison of all *Juglans* genomes showed both cpSSRs and mtSSRs exhibited very slow rates of evolution, which means that in *Juglans*, cpSSRs are good tools for high-level taxonomic evaluations^{34–38}.

Di-, tri-, and tetra-nucleotide motifs were detected in both chloroplast and mitochondrial genomes, although their presence and frequency varied considerably among species. Motifs longer than three nucleotides in organellar genomes were generally rare and were termed “complex”. There were between three and six loci per chloroplast with complex nucleotide repeats, depending on walnut species (Supplemental Table S1). Complex motifs present in chloroplasts were unique for each species (Fig. 4C, Supplemental Table S1). The total numbers of motifs identified in chloroplast and mitochondrial genomes were small compared with nuSSRs but differences in the numbers of motifs among species and differences in the presence or absence of motifs among species may be great enough that they could be used for understanding plastid evolution (Supplementary Tables S1, S6, S7). The cpSSRs studied were rich in AT motifs. Among dinucleotide SSRs, AT/TA repeats were the most common (46%). Trinucleotide SSRs (ATT/ATA) were also present, but they were rare (1 to 3 motifs per genome). Trinucleotide variations appeared to be species dependent, with 1 to 2 motifs identified per species. *J. regia* and *J. mandshurica* displayed two ATT repeats and one ATA repeat, but *J. nigra*, *J. major*, and *J. cinerea* had only the two ATT repeats. *J. ailantifolia* and *J. × cinerea* (which likely contains a *J. ailantifolia* chloroplast) had only one ATT repeat.

All *J. regia* and *J. regia* hybrids (*J. × intermedia* and *J. × quadrangulata*) had similar repeat motifs in their chloroplast, and for this reason they formed a distinct clade (Fig. 4C). For example, the ATAAA/TTTATA motif was only found in *J. regia* and hybrids of *J. regia* (*J. × intermedia* and in *J. × quadrangulata*), as expected, assuming the female parent of the hybrid was *J. regia*. The GATAA motif was found in *J. ailantifolia*, *J. mandshurica*, and *J. × cinerea* only, but the AATA motif was found in these species and *J. cinerea* (Fig. 4C, Supplemental Table S1). *J. × cinerea* was not joined with *J. cinerea*, which means the *J. × cinerea* sample contained a *J. ailantifolia* chloroplast, which is commonly observed (Fig. 4C). Chloroplast motifs were similar in *J. major*, *J. nigra*, and the *J. regia* backcross, with the exception that the *J. major* chloroplast did not contain a TAAA motif. Differences in read depth may have resulted in the apparent absence of motifs. The *J. regia* backcross was joined with *J. nigra* because *J. nigra* was used as a female in the first cross (Supplemental Table S1). Yi-heng *et al.*³⁹ reported ATAAA motifs in *J. regia* and *J. sigillata* and AAGAT repeat motifs in *J. cathayensis*, *J. hopeiensis* and *J. mandshurica*. Whether these motifs are found in all lineages of these species is not yet established.

Chloroplast and mitochondrial SSRs were monomorphic, in sharp contrast with the nuclear genomes (Supplemental Table S1). In all 12 mitochondrial genomes studied, tri-nucleotide repeats were the most prevalent (42%), followed by tetra-nucleotide (35%) and di-nucleotide (23%) motifs. Complex motifs were absent from most species' mitochondria, but complex motifs were present in the mitochondria of a few species (e.g., AGCA and TATC) (Supplemental Table S1). Most mitochondrial genome SSR motifs were conserved across all studied genomes; however GAA repeats were not found in *J. major*, *J. cinerea* or *J. nigra*. The remaining walnut species contained GAA motifs in addition to TCTT. The TAAA motif was found in *J. ailantifolia*, *J. cinerea*, *J. × cinerea* and *J. mandshurica* only. Although no complex motif was common to all the mitochondrial genomes we studied, there were several mtSSR motifs that were absent from some species and present in all the others.

Data Archiving. All primer sequences were included as supplementary files.

Conclusion

We developed databases containing lists of nuclear, chloroplast and mitochondrial loci containing SSRs from a total of 12 genotypes of six *Juglans* species. The depth and density of our marker database will assist researchers seeking to fill gaps in linkage-based genetic maps and improve the resolution of plant breeding approaches. Pairwise similarity of SSR flanking regions reflected known phylogeny. The further development of flanking sequences as sequence tagged sites would increase their utility. In particular, the maternally inherited cpSSRs showed high levels of sequence conservation and are useful for high-level taxonomic evaluations, including analyses of geographical origin, identification of distinct genetic lineages, and studies of dispersal. Mitochondrial motif patterns exhibited a lack of sequence diversity but showed variability in the number of repeats across species.

Material and Methods

A flow chart of the methods process is available in the Supplementary files (Supplementary Fig. S1).

Sample collection and DNA extraction. Tissue samples of 12 *Juglans* genotypes representing six species and four species hybrids were provided by the Hardwood Tree Improvement and Regeneration Center at Purdue University (HTIRC; www.htirc.org) (Table 1). Genomic DNA was extracted using a CTAB-based extraction method⁴⁰.

Library preparation, genomic sequencing, and sequence assembly. We constructed DNA sequencing libraries for each of the 12 walnut taxa and sequenced them using a single lane of paired-end reads at the Purdue Genomics Core (<https://www.purdue.edu/hla/sites/genomics/>) using the Illumina HighSeq2500 (Illumina Inc., San Diego, CA). Raw sequence reads were trimmed using Trimmomatic⁴¹. Contiguous sequence fragments were assembled *de novo* for each species with SOAP-denovo⁴² using trimmed and filtered reads. Whole genome, transcriptome, and organellar sequences from the Persian walnut genome were downloaded from the Walnut Genome Database (http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/Reju/)²¹. Chloroplast genomes were constructed by assembling short reads to the Persian walnut chloroplast reference sequence²¹ using BWA⁴³. Duplicates were flagged and sorted using the Picard tools software⁴⁴ and SNPs were called using the Haplotype Caller tool from the Genome Analysis Toolkit (GATK)⁴⁵. A similar approach was used to assemble the mitochondrial genomes.

SSR primer pipeline. We used a modified Perl script from Staton *et al.*¹⁷ to identify microsatellites in our *Juglans* genomes. Only SSR loci containing perfect repeat units of 2–4 nucleotides were utilized for analysis. We set a minimum-length criterion for SSR analysis; 8–40 repeats for di-nucleotide SSRs, 7–30 repeats for tri-nucleotide SSRs, and 6–20 repeats for tetra-nucleotide SSRs. Our nuSSR primers met the following parameters: a product length of 100–200 bp, a primer size from 18 to 25 bp, annealing temperatures between 55–60 °C, and 40–60% GC content. Simple sequence repeats (SSRs) flanking regions were masked to exclude low complexity regions using Dustmasker⁴⁶, and primers were designed using Primer3 (v2.3.5)⁴⁷. Designed SSR primers were denoted ‘nuclear SSRs’ (nuSSRs) to distinguish them from motif sequences. Organellar and transcriptomic genome analyses utilized the bioinformatics pipeline described in Staton *et al.*¹⁷. Organellar SSR primers were designed using Primer 3 in batch mode⁴⁸. Basic patterns (di-, tri-, and tetra-nucleotide) were identified in all studied genomes. Complex patterns were species-specific, and matched those previously reported for the chloroplast genome³⁹. Unique motifs within the chloroplast genome, and the frequency of di-, tri-, and tetra-nucleotide repeats were counted based on the number of each motif in each species (Supplementary Table S1). Motif frequency values for each genotype were used to perform statistical and pairwise distance analyses of the genotypes using Ward’s method⁴⁹. Phylogenetic analyses based on motif frequency for chloroplast genome were computed using Ntsys⁵⁰.

Similarity of SSR flanking regions among *Juglans* genomes. We performed one-on-one comparisons of all SSR-flanking region sequences identified using BLASTN (BLAST Command Line Applications User Manual 2016). Simple sequence repeat (SSRs) flanking sequences in all species were compared pairwise for sequence similarity and pairwise percent similarity for all comparisons were stored in a database. We then calculated the mean sequence similarity and mean *e*-value for all BLAST hit results. Mean sequence similarity between SSRs of different species was an indication of how concordant genetic markers and primer sequences of different walnut species were to each other. Pairwise similarity between species for all loci was used for neighbor joining. Mean *e*-values represent the probability of finding the same sequence in the database by chance, and are directly related to query sequence size. Shorter sequences have higher probabilities of occurring randomly based on mean sequence similarities (Fig. 3). Phylogenetic analysis was performed based on sequence similarity of SSR-flanking regions using Ntsys software⁵⁰. To find SNP variation (Figs 5 and 6) and the level of heterozygosity across *Juglans* genomes, SSR flanking regions were aligned with MEGA6⁵¹. Mean distance within and between groups, genome heterozygosity and SNP variation were calculated with MEGA6.

To estimate the evolutionary divergence between sequences based on SSRs flanking regions, the number of base substitutions per site between sequences was measured. Analyses were conducted using the Maximum Composite Likelihood model (MCL). All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA6⁵¹, which generated a matrix of mean distances among genomes which was analyzed using neighbor joining to represent the phylogeny. Phylogenetic analysis based on frequency of SSR motifs within the chloroplast was performed based on Neighbor joining methods and a tree was drawn using Ntsys software⁵⁰.

Filtering of unique SSRs and electronic polymerase chain reaction (e-PCR) validation. Primer sets were designed for all SSRs identified within the 12 *Juglans* genomes. To avoid untargeted amplification, primer sequences were matched against respective assembled scaffold sequences using BLASTN, and only the primer sets exactly matching a unique genomic region were retained. Primer sequences (forward or reverse) that aligned to multiple regions were filtered out²⁸. The filtered primer pairs were designated 'unique SSRs' and were used for further downstream analysis. All remaining primer pair sets were sorted by their repeat number (>10 repeats) and GC content (40–60%). These unique primer sets were then validated by e-PCR⁵². Although we focused on primers that amplified a single locus, we also evaluated the frequency with which some primer sets could amplify multiple loci with the goal of understanding genome organization.

Filtering shared SSRs among *Juglans* species. Primer pairs that were shared between two or more species were identified using custom Perl scripts and ClustalW (hybrids were not included in this analysis)⁵³. All previously filtered forward and reverse primer sequences from each species were pairwise matched (>35% sequence identity). Retained primer sets showed 100% sequence identity with primer sets from a different species and displayed the same SSR-flanking region within both species. These SSR regions were considered 'conserved' between species.

References

- Freeman, C. C. & Reveal, J. L. Flora of North America. St. Louis: *Missouri Botanical Garden*, 5, 492–496 (2005).
- Kidwell, M. G. & Lisch, D. R. Transposable elements, parasitic DNA, and genome evolution. *Evolution*. **55P**, 1–24 (2001).
- Youn, W. S. *et al.* Identification of repetitive DNA sequences in the *Chrysanthemum boreale* genome. *Sci. Hort.* **236**, 238–243 (2018).
- Ishii, T. & McCouch, S. R. Microsatellites and microsynteny in the chloroplast genomes of *Oryza* and eight other Graminae species. *Theor. Appl. Genet.* **100**, 1257–1266 (2000).
- Moore, S. S. *et al.* The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. *Genomics* **10**(3), 654–660 (1991).
- Ross-Davis, A. & Woeste, K. E. Microsatellite markers for *Juglans cinerea* L. and their utility in other Juglandaceae species. *Conserv. Genet.* **9**(no. 2), 465–469 (2008).
- Chen, X., Cho, Y. & McCouch, S. Sequence divergence of rice microsatellites in *Oryza* and other plant species. *Mol. Genet. Genomics*. **268**(3), 331–343 (2002).
- Macas, J. *et al.* In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabaeae. *PLoS One* **10**(11), e0143424 (2015).
- Kumar, S., Shah, N., Garg, V. & Bhatia, S. Large scale *in-silico* identification and characterization of simple sequence repeats (SSRs) from *de novo* assembled transcriptome of *Catharanthus roseus* (L.) G. Don. *Plant. Cell. Rep.* **33**, 905–918 (2014).
- Ashkani, S. *et al.* SSRs for marker-assisted selection for blast resistance in rice (*Oryza sativa* L.). *Plant Mol. Biol. Rep.* **30**, 79–86 (2012).
- Ebrahimi, A., Fatahi, R. & Zamani, Z. Analysis of genetic diversity among some Persian walnut genotypes (*Juglans regia* L.) using morphological traits and SSRs markers. *Sci. Hort* **130**(1), 146–151 (2011).
- Ebrahimi, A., Zarei, A., Lawson, S., Woeste, K. E. & Smulders, M. J. M. Genetic diversity and genetic structure of Persian walnut (*Juglans regia*) accessions from 14 European, African, and Asian countries using SSR markers. *Tree Genet Genomes* **12**, 114, <https://doi.org/10.1007/s11295-016-1075-y> (2016).
- Di Pierro, E. A. *et al.* A high-density, multi-parental SNP genetic map on apple validates a new mapping approach for outcrossing species. *Horticulture research* **3**, 16057 (2016).
- Cavagnaro, P. F. *et al.* Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). *BMC Genomics*. **11**, 569, <https://doi.org/10.1186/1471-2164-11-569> (2010).
- Morgante, M., Hanafey, M. & Powell, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**, 194–200 (2002).
- Sonah, H. *et al.* Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. *PLoS One* **6**, e21298, <https://doi.org/10.1371/journal.pone.0021298> (2011).
- Staton, M. *et al.* Preliminary Genomic Characterization of Ten Hardwood Tree Species from Multiplexed Low Coverage Whole Genome Sequencing. *PLoS One* **10**, e0145031, <https://doi.org/10.1371/journal.pone.0145031> (2015).
- Topçu, H. *et al.* Development of 185 polymorphic simple sequence repeat (SSR) markers from walnut (*Juglans regia* L.). *Sci. Hort* **194**, 160–167 (2015).
- Woeste, K., Burns, R., Rhodes, O. & Michler, C. Thirty polymorphic nuclear microsatellite loci from black walnut. *J Hered.* **93**, 58–60 (2002).
- Takezaki, N. & Nei, M. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144**(1), 389–399 (1996).
- Martínez-García, P. J. *et al.* The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J.* **87**, 507–532 (2016).
- Cheng, J. *et al.* A comprehensive characterization of simple sequence repeats in pepper genomes provides valuable resources for marker development in *Capsicum*. *Sci. Rep.* **6**, 18919, <https://doi.org/10.1038/srep18919> (2016).
- Kantety, R., La Rota, M., Matthews, D. & Sorrells, M. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biol.* **48**, 501–510, 10.1023/A: 1014875206165 (2002).
- Wegrzyn, J. L. *et al.* Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* **196**, 891–909 (2014).
- Mu, X. Y., Sun, M., Yang, P. F. & Lin, Q. W. Unveiling the identity of wenwan walnuts and phylogenetic relationships of Asian *Juglans* Species Using Restriction Site-Associated DNA-Sequencing. *Front Plant Sci.* **8**, 1708 (2017).
- Stanford, A. M., Harden, R. & Parks, C. R. Phylogeny and biogeography of *Juglans* (Juglandaceae) based on matK and ITS sequence data. *Am. J. Bot.* **87**, 872–882 (2000).
- Qi, J., Chen, Y., Copenhaver, G. P. & Ma, H. Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proc. Natl. Acad. Sci. USA* **111**(27), 10007–10012 (2014).
- Nowakowski, A. J., Willoughby, J. R., DeWoody, J. A. & Donnelly, M. A. Polymorphic microsatellite loci for a neotropical leaf-litter frog (*Craugastor bransfordii*) characterized through Illumina sequencing. *Conserv. Genet. Resour.* **6**(3), 697–8 (2014).
- Downey, S. L. & Iezzoni, A. F. Polymorphic DNA markers in black cherry (*Prunus serotina*) are identified using sequences from sweet cherry, peach, and sour cherry. *J. Amer. Soc. Hort. Sci.* **125**.1, 76–80 (2000).
- Manchester, S. R. Early History of the Juglandaceae. *Plant. Syst. Evol.* **50**(162), 231 (1989).
- Hoban, S. M., McCleary, T. S., Schlarbaum, S. E. & Romero-Severson, J. Geographically extensive hybridization between the forest trees American butternut and Japanese walnut. *Biol. Lett.*, pp.rsbl-2009 (2009).

32. Aradhya, M. K., Potter, D., Gao, F. & Simon, C. J. Molecular phylogeny of *Juglans* (Juglandaceae): a biogeographic perspective. *Tree Genet Genomes* **3**(4), 363–378 (2007).
33. Bai, W. N., Liao, W. J. & Zhang, D. Y. Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. *New Phytol.* **188**, 892–901 (2010).
34. Provan, J., Powell, W. & Hollingsworth, P. M. Chloroplast Microsatellites: New Tools for Studies in Plant Ecology and Evolution. *Trends Ecol. Evol.* **16**, 142–147 (2001).
35. Wang, H. L. *et al.* Developing conserved microsatellite markers and their implications in evolutionary analysis of the *Bemisia tabaci* complex. *Sci. Rep.* **4**, 6351, <https://doi.org/10.1038/srep06351> (2014).
36. Tian, A. G. *et al.* Characterization of soybean genomic features by analysis of its expressed sequence tags. *Theor. Appl. Genet.* **108**, 903–913 (2004).
37. Subramanian, S. *et al.* Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol.* **4**, R13, <https://doi.org/10.1186/gb-2003-4-2-r13> (2003).
38. Souframanien, J. & Reddy, K. S. *De novo* assembly, characterization of immature seed transcriptome and development of genic-SSR markers in black gram [*Vigna mungo* (L.) Hepper]. *PLoS One* **10**, e0128748, <https://doi.org/10.1371/journal.pone.0128748> (2015).
39. Yi-heng, H., Woeste, E. K. & Zhao, P. Completion of the chloroplast genomes of five Chinese *Juglans* and their contribution to chloroplast phylogeny. *Front. Plant. sci.* **7** (2016).
40. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* **19**, 11–15 (1987).
41. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120 (2014).
42. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18, <https://doi.org/10.1186/2047-217X-1-18> (2012).
43. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25, <https://doi.org/10.1186/gb-2009-10-3-r25> (2009).
44. Wysoker, A., Tibbetts, K. & Fennell, T. Picard tools version 1.90. (2013).
45. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
46. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A. fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).
47. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic. Acids. Res.* **40**, e115–e115, <https://doi.org/10.1093/nar/gks596> (2012).
48. You, F. M. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC bioinformatics.* **9**, 253 (2008).
49. Anderberg, M. R. Cluster analysis for application. Academic Press, Inc., New York. (1973).
50. Rohlf, F. J. NTSYSpc numerical taxonomy and multivariate analysis system version 2.0 user guide. Applied Biostatistics Inc., Setauket. (1998).
51. Tamura, K. *et al.* MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2012).
52. Shyu, C., Foster, J. A. & Forney, L. J. Electronic polymerase chain reaction (EPCR) search algorithm. *Proceedings of the IEEE 1st Bioinformatics Conference* **1**, 338 (2002).
53. Larkin, M. A. *et al.* ClustalW and ClustalX version 2.0. *Bioinformatics* **23**, 2947–2948, <https://doi.org/10.1093/bioinformatics/btm404> (2007).

Acknowledgements

The authors thank James McKenna for his comments on previous version of this manuscript and providing samples. Funding provided by Hardwood Improvement and Regeneration Center and the USDA Forest Service. Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the US Department of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable.

Author Contributions

A.E. conceived, designed and conducted the experiments, analyzed the results and wrote the main manuscript text; S.M. and N.L. were consulted and helped to analyze the results; S.L. revised, the manuscript, drew the figures and tables, and funded the genome sequencing data; M.C. and A.L. revised the manuscript; K.E.W. helped conduct the experiments and performed, and revised the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-39793-z>.

Competing Interests: The authors declare no competing interests.

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) 2019