Received: 20 September 2021 | Accepted: 10 December 2021

DOI: 10.1002/aps3.11455

APPLICATION ARTICLE

Applications in Plant Sciences

Reference-free discovery of nuclear SNPs permits accurate, sensitive identification of *Carya* (hickory) species and hybrids

Robert A. Literman[1] | Brittany M. Ott[2] | Jun Wen[3] | L. J. Grauke[4] | Rachel S. Schwartz[5] | Sara M. Handy[1]

[1]Office of Regulatory Science, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA

[2]Office of Food Additive Safety, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA

[3]Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, D.C., USA

[4]United States Department of Agriculture (USDA)–Agricultural Research Service Pecan Breeding and Genetics, Somerville, Texas, USA

[5]Department of Biological Sciences, University of Rhode Island, Kingston, Rhode Island, USA

Correspondence

Robert A. Literman, Center for Food Safety and Applied Nutrition, Office of Regulatory Science, U.S. Food and Drug Administration, College Park, Maryland 20740, USA.
Email: Robert.Literman@fda.hhs.gov

Abstract

Premise: DNA-based species identification is critical when morphological identification is restricted, but DNA-based identification pipelines typically rely on the ability to compare homologous sequence data across species. Because many clades lack robust genomic resources, we present here a bioinformatics pipeline capable of generating genome-wide single-nucleotide polymorphism (SNP) data while circumventing the need for any reference genome or annotation data.

Methods: Using the SISRS bioinformatics pipeline, we generated de novo ortholog data for the genus *Carya*, isolating sites where genetic variation was restricted to a single *Carya* species (i.e., species-informative SNPs). We leveraged these SNPs to identify both full-species and hybrid *Carya* specimens, even at very low sequencing depths.

Results: We identified between 46,000 and 476,000 species-identifying SNPs for each of eight diploid *Carya* species, and all species identifications were concordant with the species of record. For all putative F_1 hybrid specimens, both parental species were correctly identified in all cases, and more punctate patterns of introgression were detectable in more cryptic crosses.

Discussion: Bioinformatics pipelines that use only short-read sequencing data provide vital new tools enabling rapid expansion of DNA identification assays for model and non-model clades alike.

KEYWORDS

Carya, hybrids, single-nucleotide polymorphisms (SNPs), species identification

Accurate species identification is a fundamental requirement for many biological research questions ranging from studies of ecology (Thomson et al., 2010) and systematics (Whitkus et al., 1994) to more applied questions, including sensitive detection of components in food products (Staats et al., 2016; Zhang et al., 2017a). Molecular species identification techniques (e.g., DNA barcoding, species-specific PCR primers, or protein assays) are especially valuable in cases where species cannot be easily distinguished morphologically, and they are required when existing physical samples are non-identifiable (e.g., from tissue punches, prepared food products, or pre-extracted DNA). When identifying animal species using DNA sequence data, a common practice when combating "food fraud" (Nehal et al., 2021), popular genetic markers such as the *COI* gene are often used to identify samples down to the genus or species level (Hebert et al., 2003; Handy et al., 2011). However, increased rates of hybridization and whole-genome duplication among plants have complicated the parallel search for a universal plant "barcode" (Mallet, 2005; Sémon and Wolfe, 2007; Hollingsworth, 2011), and research efforts have consequently shifted toward more focused, clade-specific strategies.

Plastid genomes are frequently used for molecular identification of plant species and studies of plant systematics (Martin et al., 2005). They are substantially smaller than their nuclear counterparts, and the relatively high conservation of gene content across groups greatly simplifies the process of identifying and aligning homologous markers (Soltis et al., 2013). Additionally, chloroplasts and other plastids are also found in high quantities in many plant tissues, making them especially valuable when identifying species from processed food products (Böhme et al., 2019). Many contemporary plastid-based species identification assays for use in food products or dietary supplements rely on a single locus or small sets of popular markers (i.e., barcodes) for differentiation (Amar, 2020; Intharuksa et al., 2020; Oyebanji et al., 2020). Advances in open science like the U.S. Food and Drug Administration's (FDA) GenomeTrakrCP project (Zhang et al., 2017b), which acts as a central repository for both sequence data and critical metadata for plastid genomes, are expanding what we know about plastid diversity and providing a larger set of potentially useful markers. However, due to both limited size and evolutionary constraints on essential plastid functions, the overall utility of even fully assembled plastid genomes to delineate species varies substantially among clades (Reginato et al., 2016; Loeuille et al., 2021).

Comparatively, nuclear genomes are often orders of magnitude larger than plastid genomes (Heslop-Harrison, 2017), and this increased size provides a substantially larger search area when searching for informative genetic regions. Advancements in high-throughput sequencing technology continue to lower barriers to generating genome-scale data for many individuals, and while plastid genomes are often (but not always) maternally inherited (Kuroiwa et al., 1982; Neale and Sederoff, 1989; Chat et al., 1999), biparental inheritance of nuclear DNA means that nuclear markers can also be applied to potential hybrid specimens. Nuclear data can be acquired through the sequencing of restriction enzyme digestion products (e.g., RAD-Seq) (Andrews et al., 2016), pre-amplified or bait-capture amplicon pools (e.g., reduced-representation sequencing, sequencing of ultraconserved elements) (Johnson et al., 2019), or through whole-genome sequencing (WGS) where minimal a priori locus selection is performed (Zhang et al., 2015). However, the increased data set sizes and added complexity of nuclear evolutionary processes also present challenges, including reliable identification and isolation of homologous nuclear loci from genome-scale data. In clades with substantial genomic resources (e.g., assembled reference genomes, annotated gene maps), extracting specific genes or loci of interest is relatively straightforward and these data sets can often be useful when analyzing closely related species. Unfortunately, many plant and animal groups lack even basic genomic resources, including species of allergenic or health concern. Analysis of genome-scale data from these understudied clades is therefore often limited to familiar sets of historically popular markers (Bell et al., 2017), resulting in a potentially dramatic underutilization of information-rich data sets. Thus, the development of reference-free approaches for the discovery of informative nuclear data will enable parallel expansion of species and hybrid identification pipelines for both model and non-model groups alike.

Here we present a novel bioinformatics pipeline for generating species-informative nuclear genetic markers without the need for a priori genomic resources, providing a roadmap for researchers working on a wide diversity of non-model organisms. Using a modified implementation of the SISRS bioinformatics pipeline (Schwartz et al., 2015), we identified samples from the genus *Carya* Nutt., an agronomically important plant genus of allergenic concern that includes pecan (*C. illinoinensis* (Wangenh.) K. Koch), hickories, and other edible and inedible species (Thompson and Grauke, 1991; López-Calleja et al., 2015). Using only low-coverage whole-genome sequence data (i.e., "genome skims"), we identified over 180,000 species-diagnostic single-nucleotide polymorphisms (SNPs) and, by leveraging publicly available sequence data (Huang et al., 2019), we were able to bolster these numbers to include over 1 million species-diagnostic SNPs. These high SNP counts supported accurate identification of *Carya* samples from as little as 7.5 Mbp of sequence data (~0.01× genome coverage). Additionally, we were able to use these data to identify several edible hybrids, including non-F_1 crosses with more cryptic origins. Taken together, we present a cost-effective, computationally efficient, and expandable pipeline for the generation of species-diagnostic markers that can be applied even in clades lacking any prior genomic characterization.

METHODS

All associated scripts and relevant output can be found in the companion GitHub repository: https://github.com/BobLiterman/Carya_SISRS_SNPs. A flowchart of the major steps is provided in Figure 1.

Generating genome skim data and acquiring companion sample data

We extracted DNA from vouchered plant leaf samples using the QIAGEN DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA) and quantified output with a Qubit 3 Fluorometer (Invitrogen, Carlsbad, California, USA). Between 60–100 ng of DNA was sheared using a Covaris M220 sonicator (Covaris, Woburn, Massachusetts, USA), targeting ~450 bp DNA fragments, and from this we prepared libraries using a HyperPrep kit (KAPA Biosystems, Wilmington, Massachusetts, USA) and KAPA dual-indexed adapters. These libraries were quantified using a Qubit 3 Fluorometer, sized using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California, USA), and sequenced on either Illumina MiSeq or HiSeq machines (Illumina, San Diego, California, USA). As part of the FDA GenomeTrakrCP project (Zhang et al., 2017b), all raw sequence data were deposited in the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (Project Accession: PRJNA325670).

This pipeline leverages reference samples for comparison, so we acquired independently generated WGS data for each of the eight *Carya* species sequenced in this study (*C. aquatica* (F. Michx.) Nutt., *C. cathayensis* Sarg., *C. cordiformis* (Wangenh.) K. Koch, *C. illinoinensis*, *C. laciniosa* (F. Michx.) G. Don, *C. myristiciformis* (F. Michx.) Nutt., *C. ovata* (Mill.) K. Koch, and *C. palmeri* W. E. Manning) (Huang et al., 2019). These samples were downloaded from the European Nucleotide Archive

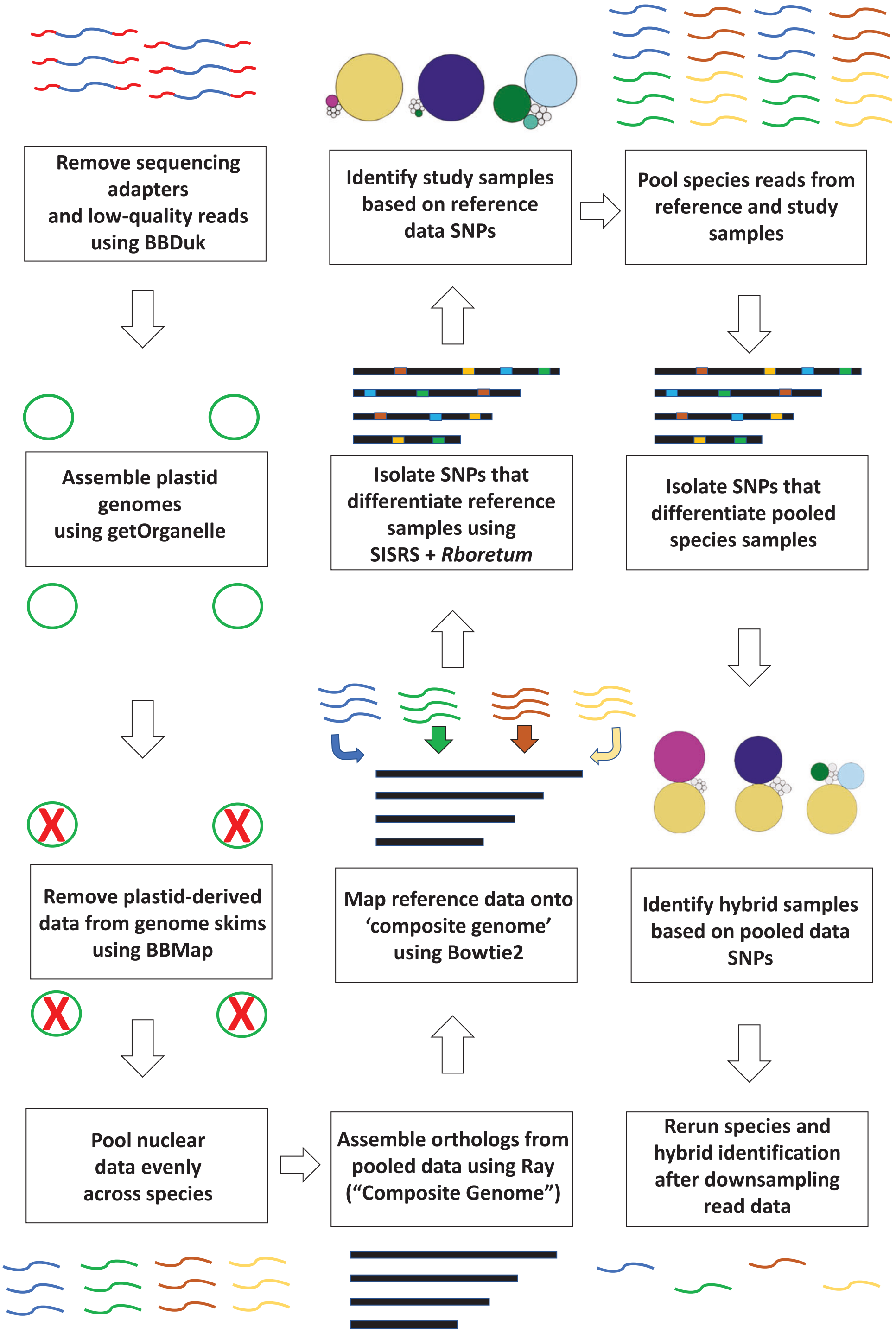
Remove sequencing adapters and low-quality reads using BBDuk
Identify study samples based on reference data SNPs
Pool species reads from reference and study samples
Assemble plastid genomes using getOrganelle
Isolate SNPs that differentiate reference samples using SISRS + *Rboretum*
Isolate SNPs that differentiate pooled species samples
Remove plastid-derived data from genome skims using BBMap
Map reference data onto 'composite genome' using Bowtie2
Identify hybrid samples based on pooled data SNPs
Pool nuclear data evenly across species
Assemble orthologs from pooled data using Ray ("Composite Genome")
Rerun species and hybrid identification after downsampling read data

FIGURE 1 A flowchart illustrating the major steps for generating species-informative SNPs from low-coverage genome skim data using SISRS

(Leinonen et al., 2010) and are referred to hereafter as the "companion data" (Appendix S1).

Pre-processing of sequencing data

We used the BBMap suite v. 38.86 (https://sourceforge.net/projects/bbmap/) and getOrganelle v.1.7.1 (Jin et al., 2020) to process all raw sequence data (i.e., study and companion data) into primarily nuclear-derived reads. First, we used *bbmerge* to (1) merge the paired-end reads (when possible), and (2) to detect sequencing adapters via paired-end overlap. After adapter trimming with *bbduk*, we generated plastid assemblies for each sample using getOrganelle, with *k*-mer values of 21, 45, 65, 85, and 105 and a maximum of 50 extension rounds; all plastid genome assemblies (complete and fragmented) were then pooled together. We trimmed and quality-filtered all reads with *bbduk*; we used a sliding window with a Q10-cutoff, and removed reads with a post-trimming minimum average quality below Q15 and/or a length less than 50 bp. To isolate nuclear reads from these trimmed data, any merged or unmerged reads that could be mapped onto the pooled "pan-plastid" data set using *bbmap* were removed from the read sets; we used the remaining quality-trimmed, nuclear-enriched reads in all downstream steps.

De novo identification of species-identifying markers for *Carya* species from genome-skim data

We used the SISRS pipeline (Schwartz et al., 2015) to identify regions of the *Carya* nuclear genome that were relatively well conserved across the eight study species (i.e., "SISRS orthologs"). As its sole input, SISRS takes WGS reads that have been pooled across species and uses them to perform a single de novo genome assembly. This results in the assembly of a set of genomic loci that are (1) present in the WGS data for most species, and (2) conserved enough among taxa to be assembled using pooled reads (and thus, conserved enough to compare among species). We assembled this "composite genome" using data from the 26 *Carya* specimens sequenced in this study (i.e., neither the hybrid samples nor companion data were used to generate orthologs). Based on a genome size estimate of 750 Mbp (Grauke et al., 2001; Huang et al., 2019), we subsampled bases from each species such that the final assembly depth was ~10× genomic coverage (i.e., 7.5 Gbp total; ~1 Gbp per species sampled evenly across specimens). By subsampling reads prior to assembly, regions of relatively high sequence conservation have sufficient depth for assembly while species-specific or poorly conserved regions will fail to assemble. We used Ray v.2.3.2-devel (Boisvert et al., 2010) to assemble the composite genome using the subsampled *Carya* nuclear reads, default parameters, and a *K* value of 31.

SISRS maps all of the reads from each sample (either from an individual specimen, or from specimens pooled by species) against the composite genome, removing any reads that map to multiple SISRS orthologs. That mapping information is then used to create a sample-specific copy of the composite genome. SISRS replaces non-specific bases with sample-specific bases within the SISRS orthologs, but only when two key conditions are met: (1) by default, sites must be covered by at least three reads (i.e., 3× coverage), and (2) there must not be variation within the sample (i.e., alleles must be fixed). However, due to the relatively low sequencing depths associated with genome skimming, requiring three reads of coverage to positively call a base would have been prohibitively constrictive; therefore, here we allowed site calling even when the data were derived from a single read. For sites covered by more than one read, any site that had within-sample variation was denoted as "N" and effectively treated as missing data. In this way, we generated SISRS ortholog data for each specimen individually. We repeated this analysis, grouping all samples from each species together to produce a separate data set of ortholog data by species. We generated three sets of species-level orthologs: (1) orthologs derived from the study samples, (2) orthologs derived from the companion data, and (3) orthologs generated after pooling the study and companion data.

Identifying *Carya* species and hybrids using SNPs

Our species identification pipeline leverages sequence data from a set of reference samples to serve as an identification guide, and we classified the study *Carya* species samples using the companion data as the reference. To create the reference data set, we used SISRS to identify sites among the companion-derived orthologs where substitutions were only observed in one of the eight species (i.e., species-informative SNPs). Identifying reliable SNPs for species identification is complicated when sequence data are missing for certain taxa, so we only considered SNPs where there was a positive base call for all species (i.e., no missing data).

To classify each of the *Carya* species samples, for each sample we extracted all sites from the specimen-specific orthologs that (1) could be genotyped for that test sample and (2) were also in the list of species-informative SNPs derived from the companion data. We used the Rboretum package (https://github.com/BobLiterman/Rboretum) in R v.4.0.2 (R Core Team, 2021) to match the signal from each of the test sample SNPs (e.g., A, G, C, T, or an indel/gap) against the companion reference data. For each test sample, we tabulated (1) the total number of sites that overlapped with classifier SNPs associated with each species, and of those (2) the number of SNPs whose fixed allele correctly matched with the reference species. We used a modified Z-score test (Leys et al., 2013) to compare the proportion of matching-to-nonmatching SNPs from each of the reference species (i.e., sample X had fixed bases at 100 SNP positions informative for *C. illinoinensis* and 75 had the matching allele [75%]), highlighting species from the companion data set where a disproportionately high proportion of SNPs matched the test sample (i.e., the most likely species match). The modified Z-score test is a median-based outlier test that enables robust comparisons of proportions when group sizes are small (e.g., among eight species), and we interpreted statistical significance using a Bonferroni-corrected alpha of 0.05 ($\alpha = 0.05/8$ species ~ $P_{\text{Threshold}} = 6.25\text{E}^{-3}$).

For characterization of the *Carya* hybrid specimens, we generated a second set of species-identifying SNPs using the pooled *Carya* species data set (i.e., SNPs with consistent signal across all specimens in the study and companion data sets). The hybrid specimens were classified based on the proportion of SNPs that matched with each of the eight potential parental species, and all comparisons and statistical assessments were performed just as for the species samples.

To more finely assess the nature of the sites used for hybrid classification, we performed an additional read mapping analysis for all hybrid crosses between *C. illinoinensis* and another *Carya* species; however, these analyses included sites where more than one allele was present (i.e., not only fixed sites). Based on the parental species involved in the cross, we specifically queried the read mapping data for sites that acted as species-diagnostic SNPs for either of the parental species (i.e., for a cross between *C. illinoinensis* and *C. aquatica*, we queried hybrid reads mapped to species-informative SNPs for either species). For all sites in each sample, we assessed (1) read coverage data, (2) whether the site was homozygous or heterozygous, and (3) whether one or both alleles matched either parental species.

To assess the impact of sequencing depth on the robustness and reliability of SISRS-based species and hybrid identification, we simulated lower sequencing depths for all study samples (species and hybrid specimens). Using the *reformat* tool from BBMap, we randomly subsampled reads from each specimen to final depths of 0.5×, 0.25×, 0.1×, 0.05×, 0.025×, and 0.01× (i.e., 375 Mbp down to 7.5 Mbp of read data per sample) and re-ran the classification pipeline as described above.

Leveraging existing genomic resources for *Carya*

Our SISRS ortholog data were generated in the absence of reference genome data, but leveraging such genomic resources when they are available allows for more contextualized results. To that end, we used Bowtie2 (Langmead and Salzberg, 2012) to map the *C. illinoinensis* SISRS orthologs generated using the pooled data set onto the NCBI *C. illinoinensis* genome assembly (C.illinoinensisPawnee_v1; Genbank Accession: GCA_018687715.1). We binned SISRS orthologs into three categories: (1) those that could be mapped uniquely onto the reference genome, (2) those that mapped to multiple genomic locations, and (3) those that could not be mapped at all. The pecan reference genome is assembled into chromosomes, and we used this information to visualize our SNP results in a chromosomal context. All species-informative SNPs derived from uniquely mapping SISRS orthologs were binned into one of 16 chromosomal data subsets, and for all species and hybrid samples we calculated the per-chromosome proportion of sites matching each reference species. These results provide an interpretive lens through which to view the genome-averaged results, but chromosome-level results were not analyzed statistically due to reduced SNP counts associated with data partitioning.

RESULTS

Sequence data processing

We generated genome skim data (i.e., low-coverage Illumina whole-genome sequencing data) for eight *Carya* species (*C. aquatica*, *C. cathayensis*, *C. cordiformis*, *C. illinoinensis*, *C. laciniosa*, *C. myristiciformis*, *C. ovata*, and *C. palmeri*) and 13 specimens of putative *Carya* hybrid crosses. Post-trimming base counts for individual specimens ranged from 415 Mbp to 2.01 Gbp per sample (Appendices S1, S2), and pooling samples within species resulted in species-level data sets containing 839 Mbp to 4.06 Gbp of trimmed read data per species (Appendices S1, S2). In addition to the samples sequenced as part of this study, for one additional specimen from each *Carya* species we also acquired companion data consisting of publicly available sequence data generated as part of an independent study (Huang et al., 2019). The companion samples contained 4.7–7.7 Gbp per species after trimming (Appendices S1, S2).

Plastid-derived reads made up less than 10% of reads from any specimen (Appendix S2). Based on a nuclear genome size estimate of 750 Mbp (Grauke et al., 2001; Huang et al., 2019), removal of plastid data resulted in trimmed nuclear read depths of 0.54–2.53× for the study species specimens, 0.57–1.61× for the study hybrid specimens, and 6.18–10.1× for the specimens from the companion data (Appendix S2). Pooling study samples within species resulted in nuclear read depths of 1.12–5.42× per species, and sequencing depths for the pooled species data set (study data + companion data) ranged from 7.82–14.6× per species (Appendix S2).

Genome skim data alone are sufficient to generate over 160 Mbp of *Carya* nuclear data

We used the SISRS bioinformatics pipeline (Schwartz et al., 2015) to identify useful regions of the *Carya* nuclear genome for species discrimination. Using just the *Carya* species genome skim data generated in this study (i.e., without using data from hybrids or the higher-depth companion data), we assembled 820,000 SISRS orthologs that totaled 169 Mbp of data (Appendix S3). We mapped WGS reads from each specimen and species pool back onto the composite genome to call sample- or species-specific bases. By default, SISRS only calls bases with no within-sample variation (i.e., fixed alleles within specimens/species); this resulted in positive base calls for 23–67 million sites among the *Carya* species samples, and 29–50 million sites for the *Carya* hybrid samples (13.7–39.7% of all possible sites; Appendix S4). Species-level orthologs were also generated from the study samples (47–81 million sites per species), the companion data (89–104 million sites per species), and the pooled data (100–114 million sites per species).

SISRS data sets yield over one million species-identifying SNPs for *Carya*

From the three sets of species-level orthologs, we used SISRS and the *Rboretum* package (https://github.com/BobLiterman/Rboretum) in R v.4.2 (R Core Team, 2021) to identify all sites where variation was only seen in one of the eight *Carya* species (i.e., species-informative SNPs). Total SNP yields were 188,000, 1.31 million, and 1.40 million for the study, companion, and pooled data sets, respectively (Appendix S5). Relative to SNP counts in the study samples (10,000–60,000 SNPs per species; Appendix S5), for any given species the companion data (which were sequenced to a higher depth) yielded 267–734% more SNPs (47,000–476,000 SNPs per species; Appendix S5).

In comparison to the companion data alone, pooling the study data and companion data led to SNP gains of 2.3–13.8% for half of the *Carya* species (*C. cathayensis*, *C. cordiformis*, *C. myristiciformis*, and *C. palmeri*; Appendix S5); conversely, pooling resulted in 2.3–10% fewer SNPs for the other four species (*C. aquatica*, *C. illinoinensis*, *C. laciniosa*, *C. ovata*; Appendix S5). In each data set, sites that differentiated *C. cathayensis* (Chinese hickory) from the North American *Carya* species comprised over 31–38% of all SNPs (Appendix S5); conversely, *C. laciniosa* and *C. ovata* yielded the fewest SNPs across data sets (3–8% of all SNPs; Appendix S5).

SISRS SNP data sets facilitate genome-scale identification of *Carya* species

For every *Carya* species sample, we (1) identified sites from each sample that had a fixed base in a species-informative position, and (2) calculated the proportion of those SNPs that matched with the species-informative allele from the companion data. For all *Carya* species samples, the highest proportion of matching alleles derived from the correct species of record, and these results were all significantly greater than the median proportions among species (all $Z > 76$; all $P < 2.2\mathrm{E}^{-16}$; Figure 2, Appendices S6, S7). Using subsampled read depths of 0.5–0.01× (i.e., 375 Mbp down to 7.5 Mbp of read data per sample) had no impact on the accuracy or statistical interpretation of species identification for any *Carya* species sample (Appendix S8).

Except for *C. laciniosa* and *C. ovata*, 67.2–93.7% of SNPs were concordant with the reference species among samples (Figure 2, Appendices S6, S7), and identifications were based on 15,500–279,000 matching SNPs per sample (Appendix S7). Although *C. laciniosa* and *C. ovata* samples were accurately identified (i.e., the highest proportion of SNPs matched the appropriate species), both the number of matching SNPs (1600–3300 SNPs per sample; Appendix S7) and the proportion of total SNPs carrying a matching signal were lower (18–23% of *C. laciniosa*/*C. ovata* SNPs matched the reference samples; Figure 2, Appendices S6, S7). The second highest proportion of SNPs in *C. laciniosa* samples derived from *C. ovata* (8.2–10.5%; Figure 2, Appendices S6, S7), while the second highest proportion of SNPs in *C. ovata* derived from *C. laciniosa* (7.3–8.0%; Figure 2, Appendices S6, S7).

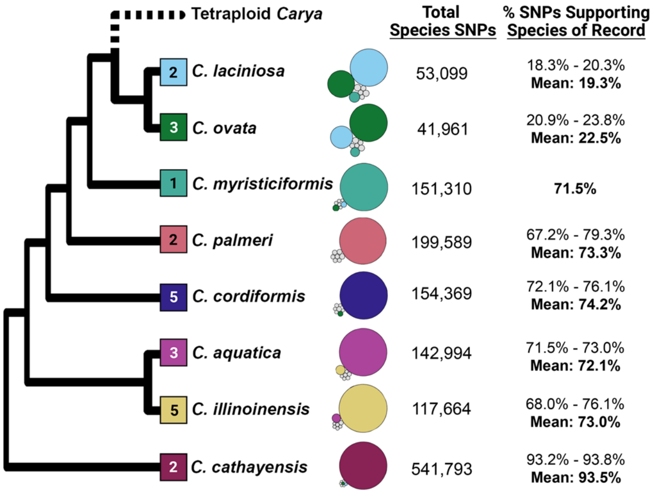

FIGURE 2 Species-informative SNP counts for *Carya*, and the per-specimen proportion of SNPs matching the species of record. We used SISRS to generate 169 Mbp of de novo, *Carya*-conserved nuclear loci and mapped read data from (1) this study, (2) an external companion data set, and (3) a pooled set of data onto the orthologs. We isolated sites where variation was restricted to a single species (pooled SNP counts shown here). We identified our 23 *Carya* samples using the companion data as a reference (specimen count noted in phylogenetic tree tips) by calculating the proportion of sites from each sample matching each reference species, and identifying species with a significantly higher contribution using a modified Z-score test. SNP data from a representative specimen are displayed using bubble plots, where the size of the circle is scaled to the proportion of matching SNPs per species, and with circle colors matching those from the phylogenetic tree if SNP enrichment was significant after Bonferroni correction. The highest proportion of matching SNPs corresponded to the species of record for all specimens, and all results were significantly higher than the data set median (all $Z > 76$; all $P < 2.2\mathrm{E}^{-16}$). Statistical results were robust to low sequencing coverage, with all specimens positively identified after sampling reads down to 0.01× genomic coverage (7.5 Mbp per specimen)

Hybrid identification from low-coverage genome skims

Just as for the *Carya* species samples, we classified *Carya* hybrid specimens based on the proportion of SNPs from each sample that matched with each of the reference samples; however, for hybrid detection the species-identifying SNPs were generated using the pooled data set as opposed to the companion data alone. Hybrid specimens are expected to have two alleles at each of the species-diagnostic positions (i.e., one allele from each parent, which by rule will be different from each other), but 96–99% of species-informative SNPs per sample were covered by five or fewer reads and 15–59% were covered by only a single read (Appendix S9); thus, for 89–95% of SNPs per hybrid sample, we observed only one allele (Appendix S10).

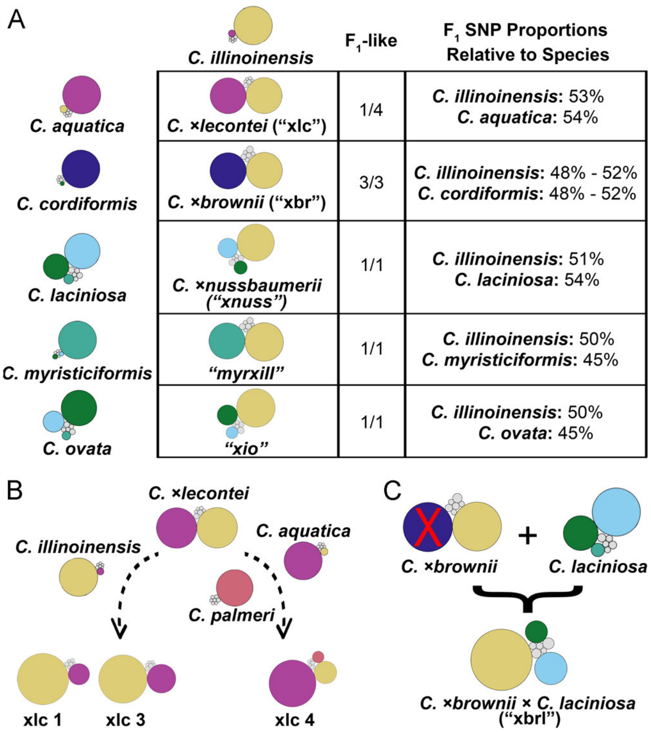

FIGURE 3 Identification of hybrid *Carya* using SNPs derived from genome skims. We mapped genome skims for 13 putative *Carya* hybrids against the de novo nuclear loci generated in this study and compared alleles falling in species-informative positions to alleles fixed in both our data and the companion data. (A) For the 10 crosses between pecan (*C. illinoinensis*) and one other *Carya* species, the two highest ratios of matching SNPs corresponded to the correct pair of parental species, and all results were significantly higher than the data-wide median based on a modified Z-score test (all $Z > 14.8$; all $P \le 2.99\text{E}^{-50}$). SNP ratios for seven of these 10 crosses were suggestive of an F_1 cross, with matching SNP proportions among hybrids hovering around 50% that of full species samples. (B) Three of the crosses between pecan and *C. aquatica* (*C.* ×*lecontei*) had SNP signatures suggesting backcrossing following F_1 hybridization, with two samples ("xlc 1" + "xlc 3") showing predominantly pecan-specific SNPs and the other ("xlc 4") displaying more *C. aquatica* sites, along with a significant enrichment of SNPs associated with *C. palmeri* ($Z = 9.6$; $P = 4.6\text{E}^{-22}$). (C) In a putative cross between *C. laciniosa* and *C.* ×*brownii* (pecan × *C. cordiformis*), no significant *C. cordiformis* signature was detected ($Z = -0.86$; $P = 0.19$). *Carya laciniosa* SNPs were detected at levels around 50% of the full species samples ($Z = 13.7$; $P = 7.7\text{E}^{-43}$), and SNP ratios resembled those from the F_1 *C.* ×*nussbaumerii* (pecan × *C. laciniosa*)

Despite this, for all crosses between pecan and one other *Carya* species, the two highest-matching SNP proportions always corresponded to the correct pair of parental species (Figure 3, Appendices S11–S13), identifications were supported by 10,300–37,200 matching SNPs per cross, and enrichment of parental SNPs was statistically significant in all samples (all $Z > 14.8$, all $P \le 2.99\text{E}^{-50}$; Appendix S13). Subsampling reads down to 0.01× genomic coverage had no impact on the accuracy or statistical interpretation of hybrid identification for any of these crosses (Appendix S14).

Crosses between pecan and *C. cordiformis* (*C.* ×*brownii*; "xbr"), *C. laciniosa* (*C.* ×*nussbaumerii*; "xnuss"), *C. myristiciformis* ("myrxill"), and *C. ovata* ("xio"), along with one of the four crosses with *C. aquatica* (*C.* ×*lecontei*; "xlc"), displayed SNP ratios suggestive of a first-generation (i.e., F_1) cross between two species (Figure 3A, Appendices S11, S12). In these samples, SNP proportions matching pecan (35.0–38.8% of SNPs) corresponded to 48.0–53.2% of the mean value among species samples of pecan, and SNP proportions for the non-pecan parent ranged from 45.2–54.0% of their mean species counterparts (i.e., half as many parental SNPs are detected in these crosses; Figure 3A, Appendices S11–S13). The other three *C.* ×*lecontei* crosses displayed more skewed SNP ratios, with "xlc 1" and "xlc 3" containing more pecan-specific SNPs (2.8–5.8× that of *C. aquatica* SNPs; Figure 3B, Appendices S11–S13), while "xlc 4" was more enriched for *C. aquatica* SNPs (4.4× that of *C. illinoinensis*; Figure 3B, Appendices S11–S13). In addition to signal from pecan and *C. aquatica*, "xlc 4" also showed low-level enrichment of SNPs matching *C. palmeri* (3.76% matching SNPs; $Z = 9.59$, $P = 4.60\text{E}^{-22}$; Figure 3B, Appendices S11–S13).

In the putative cross of *C. laciniosa* and *C.* ×*brownii* ("xbrl"), *C. cordiformis* SNPs matched the reference data set at the third lowest percentage of all species, suggesting nonsignificant enrichment (0.47% of sites; $Z = 0.86$, $P = 0.193$; Figure 3C, Appendices S11–S13). For the parent-offspring pair of samples described as pecan crossed with *C.* ×*laneyi* (*C. cordiformis* × *C. ovata*; "xila"), no significant *C. ovata* signature was detected in either sample (0.52% and 0.46% matching SNPs; both $Z < 0.12$, both $P > 0.43$; Appendices S11–S13), and while the third highest proportion of matching SNPs from each sample was from *C. cordiformis* (0.53% and 0.57%), neither value was significantly higher than the species-wide median (both $Z < 0.38$, both $P > 0.35$; Appendices S11–S13).

Leveraging genomic resources allows chromosome-level investigation

While no reference genome was used in the SISRS composite genome assembly, 494,000 SISRS orthologs (60.3% of all assembled contigs) mapped to one unique location on the NCBI *C. illinoinensis* reference genome (GenBank accession: GCA_018687715.1), and these orthologs covered 14.8% of the genome (Appendix S3). The pecan reference genome was assembled into chromosomes, and we leveraged this information to visualize the chromosome-level distribution of species-informative SNPs by classifying samples just as before, but with the following differences: (1) only uniquely mapping orthologs were used and (2) the data subset was based on which pecan chromosome the ortholog mapped onto. This chromosome-level analysis was performed on all samples (Appendix S15), and we highlight notable findings from hybrid crosses below.

In concordance with the genome-scale results, one of the four *C.* ×*lecontei* crosses showed an even distribution of *C. illinoinensis* and *C. aquatica* SNPs across chromosomes ("xlc 2"; Figure 4, Appendix S15), while the other three showed uneven SNP biases toward one species, albeit with ratios that varied greatly among chromosomes (Figure 4, Appendix S15). The *C. palmeri* enrichment detected in "xlc 4" at the genome scale was also seen at the chromosome scale; signal matching *C. palmeri* was found on fewer than 0.78% of SNPs per chromosome in the other three *C.* ×*lecontei* crosses, but it was detected at 3.0–4.4% of SNPs across chromosomes in "xlc 4" (Figure 4, Appendix S15).

Genome-scale results for *C. cordiformis* enrichment in the two *C.* ×*laneyi* samples failed to detect significant contribution, but chromosome-level analysis revealed that both the parent and offspring samples share a specific enrichment of *C. cordiformis* SNPs on chromosome 8 (Figure 4, Appendix S15). While fewer than 1% of SNPs on any other chromosome carried *C. cordiformis* signal, 7.84% and 6.12% of the

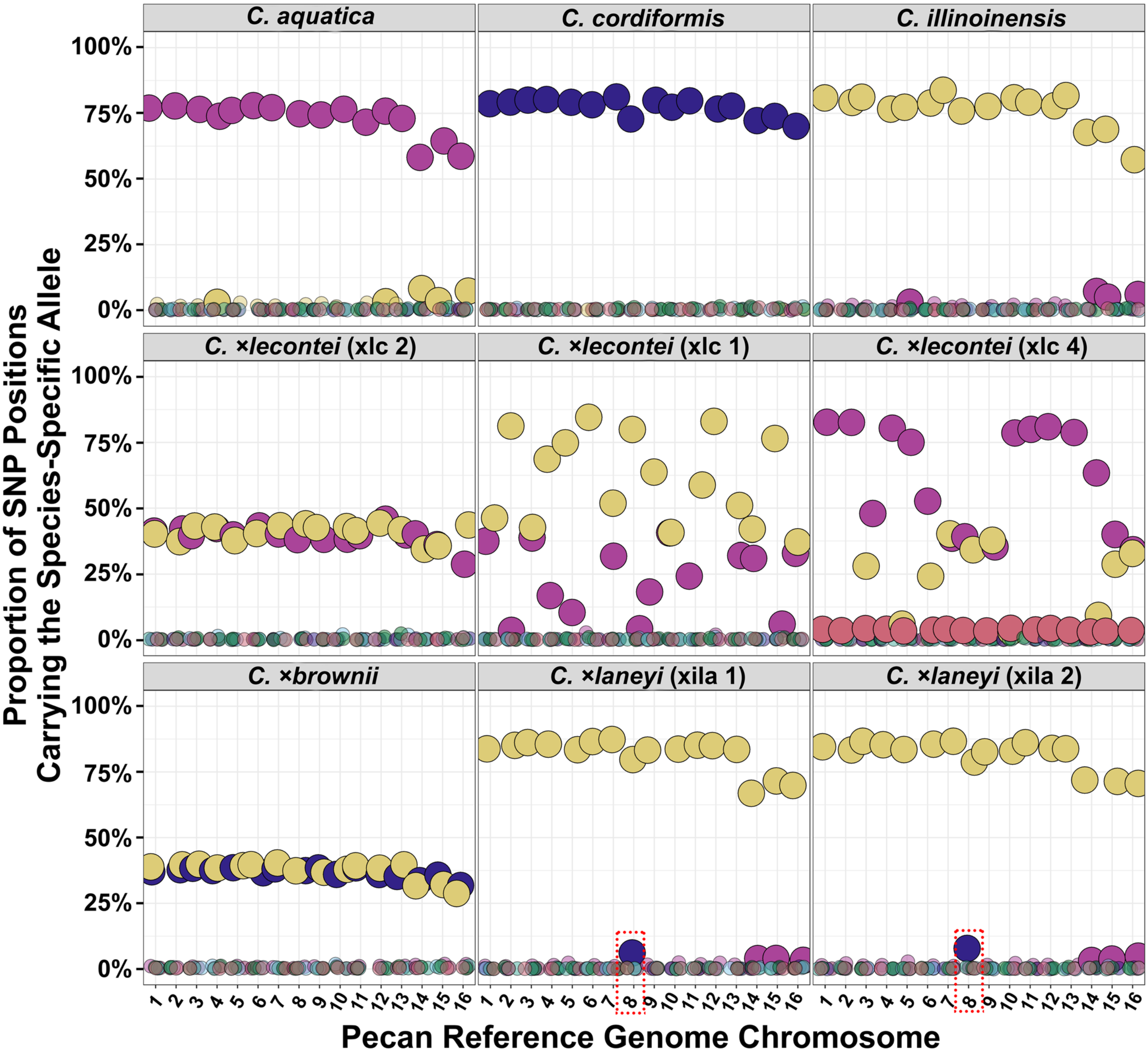

FIGURE 4 Visualizing SNP identification results in a chromosomal context. We mapped our de novo ortholog data onto the *Carya illinoinensis* reference genome, isolated uniquely mapping sequences, and binned orthologs based on which chromosome the ortholog mapped onto. For each chromosome, the proportion of species-informative SNPs matching the reference data sets was calculated for each sample. Top row: Representative species samples for *C. aquatica*, *C. cordiformis*, and *C. illinoinensis* show that the vast majority of SNPs on each chromosome correspond to the species of record. Second row: Crosses between *C. illinoinensis* and *C. aquatica* (*C.* ×*lecontei*) displayed F_1-like signal ("xlc 2"), as well as signal indicative of post-hybridization backcrossing with either *C. illinoinensis* ("xlc 1") or *C. aquatica* ("xlc 4"). The *C. palmeri* enrichment in "xlc 4" detected at the genome scale was also seen at the chromosome level, where 3.0–4.4% of *C. palmeri*-specific SNP positions across chromosomes matched the reference data set (salmon-colored points), ratios higher than all samples aside from *C. palmeri*. Bottom row: While F_1-like *C. cordiformis* signal was detected in all *C.* ×*brownii* (pecan × *C. cordiformis*) samples, a parent-offspring pair of crosses between pecan and *C.* ×*laneyi* (*C. cordiformis* × *C. ovata*) show only slight enrichment on chromosome 8 of *C. illinoinensis*, and no significant enrichment of *C. ovata* on any chromosome

SNPs on chromosome 8 matched the *C. cordiformis* reference allele in the parent and offspring, respectively (Figure 4, Appendix S15), a pattern seen in no other samples. *Carya laciniosa* signal was found on 4.8–15.0% of SNPs across chromosomes in the *C.* ×*brownii* × *C. laciniosa* sample ("xbrl"), but in agreement with the genome-scale results, fewer than 0.87% of sites on any chromosome displayed *C. cordiformis*-specific alleles (Appendix S15); these results closely overlap with the *C.* ×*nussbaumerii* (*C. illinoinensis* × *C. laciniosa*) sample (Appendix S15).

DISCUSSION

Here we present a reference-free bioinformatics pipeline that can convert genome skims (i.e., low-coverage, short-read, whole-genome sequencing data) into an ortholog data set containing millions of useful SNPs. Without the aid of a reference genome or any a priori gene annotation data, we leveraged these SNPs to identify species and hybrid specimens from the genus *Carya* (pecan and hickories) with high sensitivity and accuracy. This type of de novo locus and site-selection strategy provides a roadmap for the rapid expansion of molecular species identification assays for many non-model groups, even when starting data are of low-quality or in trace amounts (Zimmer and Wen, 2015; Liu et al., 2019), as is often the case with prepared foods and dietary supplements (Llongueras et al., 2013).

SISRS provides a genome-scale perspective on *Carya* identification

DNA-based identification of allergenic species is a critical step in many food safety protocols (Puente-Lelievre and Eischeid, 2018), with finer discriminatory power required in cases like hazelnuts where cultivars within species can vary significantly in their allergenicity (Ribeiro et al., 2020). Nearly all DNA-based identification pipelines rely on the ability to identify shared regions of DNA that contain useful variation (Graybeal, 1994; Townsend, 2007), but while technological advancements and reduced costs associated with high-throughput DNA sequencing have led to rapid increases in WGS data availability, the ultimate utility of any WGS data will typically scale to match the available genomic resources for the clade. For example, the National Collection of Genetic Resources for Pecans and Hickories (NCGR-*Carya*) developed a set of ~90,000 *Carya*-informative SNPs by mapping WGS data from multiple *Carya* species onto the same pecan reference genome (Bentley et al., 2019). Comparatively, by first generating genus-informative ortholog data de novo using SISRS (Schwartz et al., 2015), here we identified over one million diagnostic SNPs that facilitated accurate species and hybrid identification even after WGS data were subsampled down to ~0.01× genomic coverage (~7.5 Mbp per specimen). Parallel work has shown that these large SNP data sets are also useful for phylogenetic analyses (Literman and Schwartz, 2021) and for identifying trace amounts of adulterant in botanical mixes (Hunter et al., 2021). While these pipelines rely on next-generation sequencing data as input, the same SISRS orthologs could also be filtered post hoc to identify loci with higher relative densities of species-informative SNPs, thereby facilitating the straightforward development of species-diagnostic PCR primers or probe-based identification strategies (e.g., TaqMan Real-Time PCR Assay [Applied Biosystems, Waltham, Massachusetts, USA]).

Supplementing our experimental data with other publicly available data (Huang et al., 2019) led to increased SNP yields for some species (i.e., one data set supplementing another), while other data sets saw a reduction in size indicative of a purging of SNPs that would not be considered reliable species markers. Taken together, these results suggest that the robustness of any such assay will be highest with (1) enough WGS data per individual to maximize allele capture (total amounts will scale with estimated genome sizes), and (2) multiple individuals per species to reduce the impact of sampling error when assigning species-informativeness to alleles.

Large nuclear SNP data sets permit hybrid identification and inspection of parental contribution

Wild and cultivated pecan readily hybridize with other members of the *Carya* genus (Thompson and Grauke, 1991), and any practical molecular diagnostic assay for the clade must be sensitive to this as a number of these 'hican' crosses find their way into consumer goods. The most common edible hybrids are between pecan and the congener species shagbark hickory (*C. ovata*; 'Henke's Hican') and shellbark hickory (*C. laciniosa*; 'Nussbaumer's Hickory'), while crosses between pecan and bitternut hickory (*C. cordiformis*; *C.* ×*brownii*) are more commonly used as wood for cooking or lumber (Thompson and Grauke, 1991). Although hybrid samples have been identified using non-DNA methods such as biochemical profiling of tree bark (Likhanov et al., 2020) and the application of geometric morphometrics techniques (Strom et al., 2020), DNA-based methods are more common, including analysis of length polymorphisms in simple sequence repeat (SSR) markers (Hanson et al., 2020), analysis of amplified products from restriction enzyme digestion (i.e., sequence-characterized amplified region [SCAR] markers) (Anuntalabhochai et al., 2007), and even through the application of convolutional neural networks where SNPs are re-encoded as binary images and analyzed using deep-learning algorithms (Blischak et al., 2021). However, when paired with genome-scale data sets, we have found that simple SNP profiling (i.e., binary allele matching) is a computationally tractable yet statistically robust method for hybrid identification even at low sequencing depths, and this method requires no bait capture, enzymatic digestion, or specialized processing steps prior to sequencing. While we employed SISRS to identify orthologs and informative SNPs, other bioinformatics pipelines can also be used, such as STACKS (Catchen et al., 2011; Rochette and Catchen, 2017), a reference-free method for generating orthologous data and SNPs from whole-genome sequence data that has been used to identify hybrid fish specimens based on RAD-Seq data (Hohenlohe et al., 2011).

In addition to accurately identifying the major genetic contributors to hybrid samples, the SNP data generated here provided both qualitative and quantitative information regarding the nature of their hybridity. For all but three of the crosses between pecan and one other *Carya* species, the genome-wide ratios of SNPs supporting either parental species were half that of the full species counterparts (i.e., compared to a pecan and the respective crossing species, these samples had half as many species-specific SNPs for either), suggesting an F_1-like cross. Furthermore, mapping these results onto the pecan reference genome revealed that these SNPs were also evenly spread across chromosomes. Conversely, more complex patterns of hybridization were detected in three of the four crosses between pecan and *C. aquatica* (*C.* ×*lecontei*), where genome-wide SNP ratios suggested subsequent backcrossing following initial hybridization, and with corresponding chromosome-scale patterns that were far less even than their F_1-like counterparts.

In a purported cross between *C.* ×*brownii* and *C. laciniosa*, both genome- and chromosome-scale results were concordant in not detecting significant evidence of *C. cordiformis* contribution, with SNP patterns closely overlapping those of the *C.* ×*nussbaumerii* (pecan × *C. laciniosa*) sample. Phenotypic inspection reinforced genotypic findings, as post-hoc inspection of nut morphology from this sample also matched more closely with *C.* ×*nussbaumerii*. Although genome-scale results failed to detect significant signal enrichment of either *C. cordiformis* or *C. ovata* in a set of crosses involving *C.* ×*laneyi* (*C. cordiformis* × *C. ovata*), chromosome-scale analysis revealed that these samples (which were a parent-offspring pair) both shared a block of *C. cordiformis* SNPs on chromosome 8. The vestigial nature of SNPs restricted to a single chromosome implies a more complex origin for this cross, but notably the nut morphology of these samples does reflect traits specific to *C. cordiformis*. Taken together, these results suggest that loci on the eighth chromosome may be a fruitful target in understanding how these samples differ from typical pecan.

CONCLUSIONS

In this study, we provide a reference-free, genome-scale perspective on species identification using only low-coverage, whole-genome sequencing reads (i.e., genome skims). These genome skim data alone were sufficient to identify over 180,000 species-diagnostic SNPs, and supplementation with external data resulted in over one million SNPs. These large SNP pools facilitated species and hybrid identification even when data were artificially downsampled to 0.01× genomic coverage. Together, this reference-free species and hybrid identification pipeline provides a valuable resource for researchers developing diagnostic tools in non-model systems, avoiding the need for the financially and computationally costly development of high-level genomic resources.

ACKNOWLEDGMENTS

This research was partially funded by the National Science Foundation (#1812201 to R.A.L.). We acknowledge the New York Botanical Garden (sample acquisition); Alexandra Greenbaum (U.S. Food and Drug Administration), DNA4Technologies, and the University of Maryland School of Medicine Institute for Genome Sciences (sequencing); Amanda Windsor (U.S. Food and Drug Administration) (library preparation); and Chun Su (Smithsonian Institution) (DNA extraction from herbarium specimens). B.M.O. and R.A.L. were supported by the Oak Ridge Institute for Science and Education (ORISE), and B.M.O. was supported by the Joint Institute for Food Safety and Applied Nutrition (JIFSAN; #FDU001418 to B.M.O.).

AUTHOR CONTRIBUTIONS

B.M.O., J.W., and S.M.H. initiated the focal idea of this study to identify *Carya* species using DNA. J.W. and L.J.G. collected samples and all supporting metadata. R.A.L performed all computational analyses. R.A.L prepared the first draft of this manuscript with input from S.M.H., J.W., and L.J.G. S.M.H. oversaw all aspects of study. All authors approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

All raw sequence data were deposited in the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (Project Accession: PRJNA325670; https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA325670). All associated scripts and relevant output can be found in the companion GitHub repository (https://github.com/BobLiterman/Carya_SISRS_SNPs).

REFERENCES

Amar, M. H. 2020. *ycf1-ndhF* genes, the most promising plastid genomic barcode, sheds light on phylogeny at low taxonomic levels in *Prunus persica*. *Journal of Genetic Engineering and Biotechnology* 18: 42.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17: 81–92.

Anuntalabhochai, S., S. Sitthiphrom, W. Thongtaksin, M. Sanguansermsri, and R. W. Cutler. 2007. Hybrid detection and characterization of *Curcuma* spp. using sequence characterized DNA markers. *Scientia Horticulturae* 111: 389–393.

Bell, K. L., V. M. Loeffler, and B. J. Brosi. 2017. An *rbcL* reference library to aid in the identification of plant species mixtures by DNA metabarcoding. *Applications in Plant Sciences* 5: 1600110.

Bentley, N., L. J. Grauke, and P. Klein. 2019. Genotyping by sequencing (GBS) and SNP marker analysis of diverse accessions of pecan (*Carya illinoinensis*). *Tree Genetics & Genomes* 15: 8.

Blischak, P. D., M. S. Barker, and R. N. Gutenkunst. 2021. Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. *Molecular Ecology Resources* 21: 2676–2688.

Böhme, K., P. Calo-Mata, J. Barros-Velázquez, and I. Ortea. 2019. Review of recent DNA-based methods for main food-authentication topics. *Journal of Agricultural and Food Chemistry* 67: 3854–3864.

Boisvert, S., F. Laviolette, and J. Corbeil. 2010. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology* 17: 1519–1533.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait. 2011. *Stacks*: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* 1: 171–182.

Chat, J., L. Chalak, and R. J. Petit. 1999. Strict paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in intraspecific crosses of kiwifruit. *Theoretical and Applied Genetics* 99: 314–322.

Grauke, L. J., H. J. Price, and J. S. Johnston. 2001. Genome size of pecan as determined by flow cytometry. *HortScience* 36: 814.

Graybeal, A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. *Systematic Biology* 43: 174–193.

Handy, S. M., J. R. Deeds, N. V. Ivanova, P. D. N. Hebert, R. H. Hanner, A. Ormos, L. A. Weigt, et al. 2011. A single-laboratory validated method for the generation of DNA barcodes for the identification of fish for regulatory compliance. *Journal of AOAC International* 94: 201–210.

Hanson, E., H. Zhou, S. P. Tallury, X. Yang, D. Paudel, B. Tillman, and J. Wang. 2020. Identifying chromosomal introgressions from a wild species *Arachis diogoi* into interspecific peanut hybrids. *Plant Breeding* 139: 969–976.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. DeWaard. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences* 270: 313–321.

Heslop-Harrison, J. S. 2017. Plant genomes. *In* B. Thomas, B. G. Murray, and D. J. Murphy [eds.], Encyclopedia of applied plant sciences, 2nd ed. Academic Press, Oxford, United Kingdom.

Hohenlohe, P. A., S. J. Amish, J. M. Catchen, F. W. Allendorf, and G. Luikart. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* 11: 117–122.

Hollingsworth, P. M. 2011. Refining the DNA barcode for land plants. Proceedings of the National Academy of Sciences, USA 108: 19451–19452.

Huang, Y., L. Xiao, Z. Zhang, R. Zhang, Z. Wang, C. Huang, R. Huang, et al. 2019. The genomes of pecan and Chinese hickory provide insights into *Carya* evolution and nut nutrition. *GigaScience* 8: giz036.

Hunter, E. S., R. Literman, and S. M. Handy. 2021. Utilizing big data to identify tiny toxic components: Digitalis. *Foods* 10: 1794.

Intharuksa, A., Y. Sasaki, H. Ando, W. Charoensup, R. Suksathan, K. Kertsawang, P. Sirisa-Ard, and M. Mikage. 2020. The combination of ITS2 and *psbA-trnH* region is powerful DNA barcode markers for authentication of medicinal *Terminalia* plants from Thailand. *Journal of Natural Medicines* 74: 282–293.

Jin, J.-J., W.-B. Yu, J.-B. Yang, Y. Song, C. W. dePamphilis, T.-S. Yi, and D.-Z. Li. 2020. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* 21: 241.

Johnson, M. G, L. Pokorny, S. Dodsworth, L. R. Botigue, R. S Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68: 594–606.

Kuroiwa, T., S. Kawano, S. Nishibayashi, and C. Sato. 1982. Epifluorescent microscopic evidence for maternal inheritance of chloroplast DNA. *Nature* 298: 481–483.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.

Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, et al. 2010. The European nucleotide archive. *Nucleic Acids Research* 39: D28–D31.

Leys, C., C. Ley, O. Klein, P. Bernard, and L. Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49: 764–766.

Likhanov, A. F., R. I. Burda, S. N. Koniakin, and M. S. Kozyr. 2020. Identifying species and hybrids in the genus juglans by biochemical profiling of bark. *Modern Phytomorphology* 14: 27–34.

Literman, R., and R. Schwartz. 2021. Genome-scale profiling reveals noncoding loci carry higher proportions of concordant data. *Molecular Biology and Evolution* 38: 2306–2318.

Liu, J., W. Li, J. Wang, D. Chen, Z. Liu, J. Shi, F. Cheng, et al. 2019. A new set of DIP-SNP markers for detection of unbalanced and degraded DNA mixtures. *Electrophoresis* 40: 1795–1804.

Llongueras, J. P., S. Nair, D. Salas-Leiva, and A. E. Schwarzbach. 2013. Comparing DNA extraction methods for analysis of botanical materials found in anti-diabetic supplements. *Molecular Biotechnology* 53: 249–256.

Loeuille, B., V. Thode, C. Siniscalchi, S. Andrade, M. Rossi, and J. Rubens Pirani. 2021. Extremely low nucleotide diversity among thirty-six new chloroplast genome sequences from *Aldama* (Heliantheae, Asteraceae) and comparative chloroplast genomics analyses with closely related genera. *PeerJ* 9: e10886.

López-Calleja, I. M., S. de la Cruz, I. González, T. García, and R. Martín. 2015. Market analysis of food products for detection of allergenic walnut (*Juglans regia*) and pecan (*Carya illinoinensis*) by real-time PCR. *Food Chemistry* 177: 111–119.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology and Evolution* 20: 229–237.

Martin, W., O. Deusch, N. Stawski, N. Grünheit, and V. Goremykin. 2005. Chloroplast genome phylogenetics: Why we need independent approaches to plant molecular evolution. *Trends in Plant Science* 10: 203–209.

Neale, D. B., and R. R. Sederoff. 1989. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. *Theoretical and Applied Genetics* 77: 212–216.

Nehal, N., B. Choudhary, A. Nagpure, and R. K. Gupta. 2021. DNA barcoding: A modern age tool for detection of adulteration in food. *Critical Reviews in Biotechnology*: 1–25.

Oyebanji, O. O., E. C. Chukwuma, K. A. Bolarinwa, O. I. Adejobi, S. B. Adeyemi, and A. O. Ayoola. 2020. Re-evaluation of the phylogenetic relationships and species delimitation of two closely related families (Lamiaceae and Verbenaceae) using two DNA barcode markers. *Journal of Biosciences* 45: 96.

Puente-Lelievre, C., and A. C. Eischeid. 2018. Development and evaluation of a real-time PCR multiplex assay for the detection of allergenic peanut using chloroplast DNA markers. *Journal of Agricultural and Food Chemistry* 66: 8623–8629.

Reginato, M., K. M. Neubig, L. C. Majure, and F. A. Michelangeli. 2016. The first complete plastid genomes of Melastomataceae are highly structurally conserved. *PeerJ* 4: e2715.

Ribeiro, M., J. Costa, I. Mafra, S. Cabo, A. P. Silva, B. Gonçalves, M. Hillion, et al. 2020. Natural variation of hazelnut allergenicity: Is there any potential for selecting hypoallergenic varieties? *Nutrients* 12: 2100.

Rochette, N. C., and J. M. Catchen. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols* 12: 2640–2659.

Schwartz, R. S., K. M. Harkins, A. C. Stone, and R. A. Cartwright. 2015. A composite genome approach to identify phylogenetically informative data from next-generation sequencing. *BMC Bioinformatics* 16: 193.

Sémon, M., and K. H. Wolfe. 2007. Consequences of genome duplication. *Current Opinion in Genetics and Development* 17: 505–512.

Soltis, D. E., M. A. Gitzendanner, G. Stull, M. Chester, A. Chanderbali, S. Chamala, I. Jordon-Thaden, et al. 2013. The potential of genomics in plant systematics. *Taxon* 62: 886–898.

Staats, M., A. J. Arulandhu, B. Gravendeel, A. Holst-Jensen, I. Scholtens, T. Peelen, T. W. Prins, and E. Kok. 2016. Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry* 408: 4615–4630.

Strom, D. M., N. F. Bendik, D. A. Chamberlain, J. A. Watson, and J. M. Meik. 2020. Phenotypic variation in endangered Texas salamanders: Application of model-based clustering for identifying species and hybrids. *Diversity* 12: 297.

Thompson, T. E., and L. J. Grauke. 1991. Pecans and other hickories (*Carya*). *Acta Horticulturae* 290: 839–906.

Thomson, R. C., I. J. Wang, and J. R. Johnson. 2010. Genome-enabled development of DNA markers for ecology, evolution and conservation. *Molecular Ecology* 19: 2184–2195.

Townsend, J. P. 2007. Profiling phylogenetic informativeness. *Systematic Biology* 56: 222–231.

Whitkus, R., J. Doebley, and J. F. Wendel. 1994. Nuclear DNA markers in systematics and evolution. *In* R. L. Phillips and I. K. Vasil [eds.], DNA-based markers in plants. Advances in Cellular and Molecular Biology of Plants, vol. 1. Springer, Dordrecht, the Netherlands.

Zhang, N., J. Wen, and E. A. Zimmer. 2015. Congruent deep relationships in the grape family (Vitaceae) based on sequences of chloroplast genomes and mitochondrial genes via genome skimming. *PLoS ONE* 10: e0144701.

Zhang, N., D. L. Erickson, P. Ramachandran, A. R. Ottesen, R. E. Timme, V. A. Funk, Y. Luo, and S. M. Handy. 2017a. An analysis of *Echinacea* chloroplast genomes: Implications for future botanical identification. *Scientific Reports* 7: 216.

Zhang, N., P. Ramachandran, J. Wen, J. A. Duke, H. Metzman, W. McLaughlin, A. R. Ottesen, et al. 2017b. Development of a reference standard library of chloroplast genome sequences, GenomeTrakrCP. *Planta Medica* 83: 1420–1430.

Zimmer, E. A., and J. Wen. 2015. Using nuclear gene data for plant phylogenetics: Progress and prospects II. Next-gen approaches. *Journal of Systematics and Evolution* 53: 371–379.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

Appendix S1. Data sources and whole-genome sequencing data for each sample used in this study. Reads were adapter- and quality-trimmed using BBTools as described in the manuscript. Only genome skims from *Carya* species samples were used in construction of the composite genome.

Appendix S2. Results of separating nuclear- and plastid-derived reads for each sample. Reads from each specimen were mapped against a pool of assembled plastid data, and non-mapping reads were designated as nuclear.

Appendix S3. Assembly statistics for the composite genome assembled de novo by SISRS using Ray, along with information about mapping those SISRS contigs onto the *Carya illinoinensis* reference genome.

Appendix S4. Reads from each specimen and pooled sample were mapped against the SISRS composite genome, and only sites with a fixed allele were officially "called."

Appendix S5. For each species of *Carya*, we identified all species-informative SNPs, or sites in the composite genome where that species differed from all other *Carya species*. We break down which of these SNPs are singleton in nature, as well as which can be positively mapped to the reference genome.

Appendix S6. Using the companion data as a reference, we identified our 23 *Carya* samples by calculating the proportion of sites from each sample matching each reference species, and identifying species with a significantly higher contribution using a modified Z-score test. SNP data for all specimens are displayed using bubble plots, where the size of the circle is scaled to the proportion of matching SNPs per species, and with circle colors matching those from the phylogenetic tree in Figure 1 if SNP enrichment was significant after Bonferroni correction. The highest proportion of matching SNPs corresponded to the species of record for all specimens, and all results were significantly higher than the data set median (all Z > 76; all $P < 2.2\text{E}^{-16}$).

Appendix S7. Statistical results for species identification of entire genome skim samples. Each sample was scored based on the number of species-informative SNPs that matched each reference sample in the companion data. The highest matching ratio for each specimen is noted by boldfaced red text, while all other results that were significantly enriched after Bonferroni correction are noted with yellow highlighting.

Appendix S8. Statistical results for species identification of downsampled genome skim samples. Each downsampled data set (0.01–0.5× genomic coverage) was scored based on the number of species-informative SNPs that matched each reference sample in the companion data. The highest matching ratio for each specimen is noted by boldfaced red text, while all other results that were significantly enriched after Bonferroni correction are noted with yellow highlighting.

Appendix S9. Read-mapping coverage for all sites called in hybrid samples that could support either of the putative parental species, showing ~95% of sites covered by four or fewer reads.

Appendix S10. A breakdown of sites called with one, two, or more alleles for all sites called in hybrid samples that could support either of the putative parental species.

Appendix S11. Using the pooled data as a reference, we characterized our 13 *Carya* hybrid samples by calculating the proportion of sites from each sample matching each reference species, and identifying species with a significantly higher contribution using a modified Z-score test. SNP data for all specimens are displayed using bubble plots, where the size of the circle is scaled to the proportion of matching SNPs per species, and with circle colors matching those from the phylogenetic tree in Figure 1 if SNP enrichment was significant after Bonferroni correction.

Appendix S12. A heatmap showing the breakdown of matching species-informative SNPs per species for each hybrid specimen. Species with a significant SNP enrichment in each sample are denoted with larger, bold font. Numbers in red font indicate species with specific enrichment in the sample that are not part of the canonical cross.

Appendix S13. Statistical results for hybrid identification of entire genome skim samples. Each sample was scored based on the number of species-informative SNPs that matched each reference sample in the pooled data. The two highest matching ratios for each specimen are noted by boldfaced red text, while all other results that were significantly enriched after Bonferroni correction are noted with yellow highlighting.

Appendix S14. Statistical results for hybrid identification of downsampled genome skim samples. Each downsampled data set (0.01–0.5× genomic coverage) was scored based on the number of species-informative SNPs that matched each reference sample in the pooled data. The two highest matching ratios for each specimen are noted by boldfaced red text, while all other results that were significantly enriched after Bonferroni correction are noted with yellow highlighting.

Appendix S15. Visualizing SNP identification results in a chromosomal context. We mapped our de novo ortholog data onto the *Carya illinoinensis* reference genome, isolated uniquely mapping sequences, and binned orthologs based on which chromosome the ortholog mapped onto. For each chromosome, the proportion of species-informative SNPs matching the reference data sets was calculated for each sample.

How to cite this article: Literman, R. A., B. M. Ott, J. Wen, L. J. Grauke, R. S. Schwartz, and S. M. Handy. 2022. Reference-free discovery of nuclear SNPs permits accurate, sensitive identification of *Carya* (hickory) species and hybrids. *Applications in Plant Sciences* 10(1): e11455. https://doi.org/10.1002/aps3.11455