JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12176

Research Article

Phylogenomics of polyploid *Fothergilla* (Hamamelidaceae) by RAD-tag based GBS—insights into species origin and effects of software pipelines

Zhe-Chen Qi[1,4], Yi Yu[1#], Xiang Liu[1], Andrew Pais[1], Thomas Ranney[2*], Ross Whetten[3*], and Qiu-Yun (Jenny) Xiang[1*]

[1]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27695, USA
[2]Department of Horticulture Sciences, North Carolina State University, Mills River, NC 28759, USA
[3]Department of Forestry & Environmental Resources, North Carolina State University, Raleigh, NC 27695, USA
[4]College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
[#]Present Address: Guangzhou Baiyunshan Zhongyi Pharmaceutical Company Limited, Guangzhou 510530, China
*Authors for correspondence. E-mails: qyxiang@ncsu.edu, ross_whetten@ncsu.edu, tom_ranney@ncsu.edu. Tel.:+ 919-515-2728. Fax: 9195153436.
Received 7 August 2015; Accepted 24 August 2015; Article first published online 9 September 2015

Abstract *Fothergilla* (Hamamelidaceae) consists of *Fothergilla gardenii* (4*x*) from the coastal plains of the southeastern USA, *F. major* (6*x*) from the piedmont and mountains of the same region, and a few allopatric diploid populations of unknown taxonomic status. The objective of this study was to explore the relationships of the polyploid species with the diploid plants. Genotyping by sequencing (GBS) was applied to generate genome-wide molecular markers for phylogenetic and genetic structure analyses of 36 accessions of *Fothergilla*. Sanger sequencing of three plastid and one nuclear regions provided data for comparison with GBS-based results. Phylogenetic outcomes were compared using data from different sequencing runs and different software workflows. The different data sets showed substantial differences in inferred phylogenies, but all supported a genetically distinct 6*x F. major* and two lineages of the diploid populations closely associated with the 4*x F. gardenii*. We hypothesize that the 4*x F. gardenii* originated through hybridization between the Gulf coastal 2*x* and an extinct (or undiscovered) 2*x* lineage, followed by backcrosses to the Atlantic coastal 2*x* before chromosome doubling, and the 6*x F. major* also originated from the "extinct" 2*x* lineage. Alternative scenarios are possible but are not as well supported. The origins and divergence of the polyploid species likely occurred during the Pleistocene cycles of glaciation, although fossil evidence indicates the genus might have existed for a much longer time with a wider past distribution. Our study demonstrates the power of combining GBS data with Sanger sequencing in reconstructing the evolutionary network of polyploid lineages.

Key words: *Fothergilla*, hybridization, illumina sequencing, phylogenomics, RAD-tag-based GBS, polyploidy.

Fothergilla L. (witch alder, Hamamelidacea) is a small genus of uncommon, deciduous shrubs found exclusively in woodland and swamps of the southeastern United States of America. The most recent taxonomic revision of *Fothergilla* (Weaver, 1969) recognized only two species: *F. gardenii* L. and *F. major* (Sims) Loddiges. However, variation in the genus is complex and as many as four species have been recognized in the past (e.g., *F. monticola* Ashe and *F. parvifolia* Keamey were once recognized and now were merged with *F. major*, and *F. gardenii*, respectively; see Weaver, 1969). *Fothergilla gardenii* is found in wet savannas and pocosins in the coastal plains of North Carolina, South Carolina, and Georgia (Weaver, 1969; Meyer, 1997; Weakley, 2008) and has been reported to be a tetraploid with $2n = 4x = 48$ (Weaver, 1969). This species is generally smaller in stature (3–10 dm) than *F. major* and is distinguished sometimes by smaller leaves ranging from 1.9 to 6 cm long and from 1.3 to 5.2 cm wide that are generally toothed only on the upper half and symmetric at the base. In contrast, *F. major* is found on more upland sites in the piedmont and mountains of North Carolina, South Carolina, Georgia, Alabama, Tennessee, and Arkansas (Weaver, 1969; Meyer, 1997; Weakley, 2008). *Fothergilla major* has been reported to be a hexaploid with $2n = 6x = 72$ (Weaver, 1969). This species generally is larger in stature (7–65 dm) than *F. gardenii* and is distinguished by larger leaves ranging from 2.5 to 13 cm long and 4.2 to 12.5 cm wide that generally are toothed from below the middle and conspicuously asymmetric at the base.

Although these two species have allopatric native ranges, they have been grown together in cultivation and will freely

hybridize. A recent cytological survey (Ranney et al., 2007) found that the majority of cultivars represented in commerce were pentaploid hybrids with ($2n = 5x = 60$) and a nothospecies *F.* ×*intermedia* Ranney & Fantz was proposed as the hybrid designation. However, no pentaploid hybrids have been reported in nature. Until recently, no diploid cytotypes of *Fothergilla* were known. Recent sampling of natural populations and cytometric analysis (Ranney et al., 2012) identified diploid plants in Florida, Georgia, South Carolina, and Alabama. Although the morphology of these plants is variable, and the subject of a separate study, they are generally more similar to *F. gardenii* in size and foliage traits than to *F. major.* Being isolated both geographically and cytogenetically from other *Fothergilla* species, these diploid plants may represent a new taxon that might have served as the parental species of *F. gardenii* and *F. major*. This recent discovery provides impetus to understand the evolutionary relationships and phylogeography of the genus.

Elucidating phylogenetic relationships and evolutionary processes in polyploid taxa has been a challenge in molecular systematics. Restriction–site Associated DNA (RAD) tags generated by next-generation sequencing (NGS) technology provide sensitive markers from many loci of the genome (Baird et al., 2008; Griffin et al., 2011; Rowe et al., 2011; Peterson et al., 2012) and overcome problems associated with extensive labor involved in cloning and Sanger sequencing of multiple-copy nuclear genes in polyploid organisms. RAD-tag sequencing has recently shown to be a valuable method in genotyping, evolutionary and ecological genomics, phylogeography, and phylogenetic studies of various organisms (Davey & Blaxter, 2010; Davey et al., 2013; Rubin et al., 2012; Eaton & Ree, 2013; Gagnaire et al., 2013; Recknagel et al., 2013; Chu et al., 2014), including polyploid species (Lu et al., 2013).

The objectives of this research was to conduct phylogenetic analyses of extant populations of *Fothergilla* using RAD-tags from genotyping by sequencing (GBS) to understand their evolutionary relationships and origins of the polyploidy species.

Material and Methods

Sampling and DNA extraction

Thirty six accessions collected from the natural distribution range of *Fothergilla* were included in the study. These accessions included 17 *F. major*, ten *F. gardenii*, and nine diploid plants (Table 1; Fig. 1). One accession of *Parrotiopsis jacquemontiana* (Decne.) Rehder and two *Hamamelis virginiana* L. were used as outgroups given their close relationships to *Fothergilla* (Xie et al., 2010). *Parrotiopsis* (Niedenzu) C. Schneid. is a monotypic genus from the western Himalayas ($2n = 2x = 24$) while *Hamamelis virginiana* is found through the eastern United States of America ($2n = 2x = 24$) (Weaver, 1969; Goldblatt & Endress, 1977; Li & Bogle, 2001). All accessions were grown at the Mountain Horticultural Crops Research and Extension Center in Mills River, NC. This sampling included every population that we could find by working with naturalists, botanists, and herbarium records. Both the 4*x* and 2*x* were quite rare. There was typically a small colony at each site for the 2*x* and 4*x* plants. The 2*x* and 4*x* plants were usually rhizomatous and possibly asexually spread at a given site. Fresh leaf samples were used for DNA extraction using the DNeasy Kit from Qiagen and those samples with high amounts of polysaccharides and proteins were further purified with Qiagen DNA Purification Kits (Qiagen, Inc., Valencia, CA, USA). The concentration and quality of each sample was assessed using a PicoGreen fluorescent dye assay (Life Technologies, ThermoFisher), UV absorbance measured on a Nanodrop spectrophotometer (ThermoFisher), and agarose gel tests prior to library preparation and sequencing.

Library preparation and sequencing

Purified genomic DNA (750 ng per sample) was digested with *Pst*I restriction enzymes followed by ligation of a barcoded adapter (P1, 4–10 bp barcode sequence, Table 1) and a common adapter (P2). The top-strand sequence of the P2 adapter, with an example barcode in bold and a *Pst*I compatible 3′ end, is as follows: GATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT**TGGTCAAG**TGCA. Digestion was performed in 20 μL with 10 μL of DNA extract, 8 U of *Pst*I, 7.2 μL water, and 2.0 μL NEB Buffer 4 at 37 °C for 2 h followed by 65 °C for 20 min. Ligation was performed in a 40 μL solution consisting of 20 μL restriction digest, 5 μL mix of barcoded and common adapters, 2 μL NEB Buffer 4, 4 μL ATP at 10 mmol/L, 8.5 μL water, and 200 U T4 DNA ligase. During ligation, samples were incubated for 2 h at 22 °C and were then incubated for 20 min at 65 °C. For both experiments, adapter-ligated fragments were proportionally pooled according to ploidy level (e.g., double the amount for 4*x* plants and triple the amount for 6*x* plants). Pooled fragments were selected for target sizes of 100–600 bp using the Agencourt AMPure Xp PCR Purification systems (Agencourt Bioscience Corp) in Experiment I (Run 1 in Table 1) and 300+/−50 bp using Pippin Prep (Sage Science Corp) in Experiment II (Run 2, Table 1). Selected fragments were amplified by PCR followed by purification with solid-phase reversible immobilization (Ampure XP beads) and quantification on a Bioanalyzer 2100 (Agilent Technologies). To test how GBS data from different taxon sampling scales, library preparation, and sequencing methods affect the phylogenetic results, we conducted two experiments. Experiment I included libraries of only 12 samples representing the different *Fothergilla* taxa and the outgroup *Parrotiopsis* (Table 1). Experiment II included newly prepared libraries of all 39 accessions, six of which represented library repeats of Experiment I (Table 1). The libraries from the two experiments were run independently. Run 1 included the 12 samples of Experiment 1 that was run in a lane of 96 samples. Run 2 included the 39 samples of Experiment II as well as nine of the 12 libraries of Experiment I and was run in a lane of 138 samples (48 *Fothergilla* samples as well as 90 samples of other taxa); the amount of DNA per library used in the pooled sample for Run 2 was adjusted to give an expected yield of half the number of reads of the other 90 samples of the same lane. The sequencing depth (number of reads per sample to be recovered) was planned to be lower than Run 1 but sufficient to recover enough loci for phylogenetic analyses based on results of Run 1. The libraries of each experiment were prepared for 100 bp paired end sequencing on Illumina 2000 Hi-seq at BGI (Philadelphia) via service of the Genomic Science Lab at NCSU.

Table 1 Information of plant materials included in the study. Values for 2C genomic size are means ± SEM. S13–S24 are sample ID of Run 1. F01–F48 are sample ID of Run 2. The number in parentheses indicates the repeated library of the same individuals. Voucher specimens are deposited at NCSC.

Sample ID	Voucher	Taxa	Ploidy	2C Genomic Size	Location (Co., State)	P1 Barcode	P2 Barcode	Sequencing plate
Run 2								
F01	2011-083-100	*Fothergilla* sp.	2*x*	1.70 ± 0.03	Okaloosa Co., FL	TGACGCCA	TGGTCAAG	Run 2
F02	2011-087-100	*F.* sp.	2*x*	1.78 ± 0.02	Baldwin Co., AL	GGTATA	TGGTCAAG	Run 2
F03	2011-088-100	*F.* sp.	2*x*	1.74 ± 0.00	Walton, Co., FL	GCCTACCT	TGGTCAAG	Run 2
F04	2011-168-100	*F.* sp.	2*x*	1.74 ± 0.05	Tattnall Co., GA	AACTGG	TGGTCAAG	Run 2
F05	2011-170-100	*F.* sp.	2*x*	1.75 ± 0.10	Emanuel Co., GA	CTCTCGCAT	TGGTCAAG	Run 2
F06	2011-171-100	*F.* sp.	2*x*	1.73 ± 0.02	Long Co., GA	CTCAT	TGGTCAAG	Run 2
F07	2011-178-100	*F.* sp.	2*x*	1.74 ± 0.02	Taylor Co., GA	GGAGTCAAG	TGGTCAAG	Run 2
F08	2012-060-100	*F.* sp.	2*x*	1.76 ± 0.01	Walton Co., FL	GAATGCAATA	TGGTCAAG	Run 2
F09	2012-084-001	*F.* sp.	2*x*	1.82 ± 0.04	Aiken Co, SC	CAGATA	TGGTCAAG	Run 2
F10	2001-047	*F. gardenii* 'Harold Epstein'	4*x*	3.51 ± 0.10	Undocumented	TCTTGG	TGGTCAAG	Run 2
F11	2011-085-100	*F. gardenii*	4*x*	3.69 ± 0.02	Richmond Co., NC	CACCA	TGGTCAAG	Run 2
F12	2011-096-100	*F. gardenii*	4*x*	3.64 ± 0.08	Carteret Co., NC	ATGAGCAA	TGGTCAAG	Run 2
F13	2011-097-100	*F. gardenii*	4*x*	3.57 ± 0.00	Hoke Co., NC	CAGAGGT	TGGTCAAG	Run 2
F14	2011-103-100	*F. gardenii*	4*x*	3.69 ± 0.00	Carteret Co., NC	ACGGTACT	TGGTCAAG	Run 2
F15	2011-123-003	*F. gardenii*	4*x*	3.68 ± 0.04	Richmond Co., NC	TGAAT	TGGTCAAG	Run 2
F16	2012-075-001	*F. gardenii*	4*x*	3.40 ± 0.01	Charleston Co, SC	TAGCAG	TGGTCAAG	Run 2
F17	2012-076-001	*F. gardenii*	4*x*	3.33 ± 0.16	Horry Co, SC.	GAAGTG	TGGTCAAG	Run 2
F18	2012-077-001	*F. gardenii*	4*x*	3.76 ± 0.05	Charleston Co, SC	GGTGT	TGGTCAAG	Run 2
F19	2012-078-001	*F. gardenii*	4*x*	3.61 ± 0.02	Effingham Co, GA	AATTAG	TGGTCAAG	Run 2
F20	2008-009-100	*F. major*	6*x*	5.27 ± 0.02	Dekalb Co., AL	CTTGA	TGGTCAAG	Run 2
F21	2011-082-100	*F. major*	6*x*	5.22 ± 0.12	Searcy Co, AR	GCGTACAAT	TGGTCAAG	Run 2
F22	2011-091-100	*F. major*	6*x*	5.40 ± 0.04	Oconee Co., SC	GCGCCG	TGGTCAAG	Run 2
F23	2011-092-002	*F. major*	6*x*	5.23 ± 0.11	Marshall Co., AL	CATAT	TGGTCAAG	Run 2
F24	2011-093-100	*F. major*	6*x*	5.29 ± 0.03	Blount Co., AL	ATCCG	TGGTCAAG	Run 2
F25	2011-105-100	*F. major*	6*x*	5.09 ± 0.05	Burke Co., NC	TAGCGGAT	TGGTCAAG	Run 2
F26	2011-112-100	*F. major*	6*x*	5.12 ± 0.02	Transylvania Co., NC	GGATA	TGGTCAAG	Run 2
F27	2011-121-100	*F. major*	6*x*	5.17 ± 0.01	Rutherford Co., NC	GGAACGA	TGGTCAAG	Run 2
F28	2011-122-100	*F. major*	6*x*	5.27 ± 0.06	Montgomery Co., NC	GCGTCCT	TGGTCAAG	Run 2
F29	2011-124-100	*F. major*	6*x*	5.15 ± 0.10	Orange Co., NC	ACGCGCG	TGGTCAAG	Run 2
F30	2011-131-100	*F. major*	6*x*	5.13 ± 0.05	Transylvania Co., NC	CAAGT	TGGTCAAG	Run 2
F31	2011-146-100	*F. major*	6*x*	5.36 ± 0.02	Walker Co., GA	GTGACACAT	TGGTCAAG	Run 2
F32	2011-147-100	*F. major*	6*x*	5.17 ± 0.05	Marshall Co., AL	CTTAG	TGGTCAAG	Run 2
F33	2011-163-100	*F. major*	6*x*	5.27 ± 0.01	Rutherford Co., NC	TATTCGCAT	TGGTCAAG	Run 2
F34	2011-164-100	*F. major*	6*x*	5.31 ± 0.01	Lumpkin Co., GA	CTAAGCA	TGGTCAAG	Run 2
F35	2011-169-100	*F. major*	6*x*	5.17 ± 0.17	Fulton Co., GA	ACAACT	TGGTCAAG	Run 2
F36	2012-065-006	*F. major*	6*x*	5.24 ± 0.28	Scott Co., TN	ACCAGGA	TGGTCAAG	Run 2

Continued

Table 1 Continued

Sample ID	Voucher	Taxa	Ploidy	2C Genomic Size	Location (Co., State)	P1 Barcode	P2 Barcode	Sequencing plate
F37	2012-081-100	*Hamamelis virginiana*	2x	2.06 ± 0.03	Santa Rosa Co., FL	GTCGCCT	TGGTCAAG	Run 2
F38	2012-089-001	*H. virginiana*	2x	2.10 ± 0.04	Monroe Co., AL	TCCGAG	TGGTCAAG	Run 2
F39	H2011-036-002	*Parrotiopsis jacquemontiana*	2x	1.59 ± 0.01	Undocumented	TATGT	TGGTCAAG	Run 2
F41 (F03), S16	2011-088-001	*F.* sp.	2x	1.74 ± 0.00	Walton, Co., FL	ACAAA	NA	Run 1 and 2
F42 (F12), S17	2011-096-001	*F. gardenii*	4x	3.64 ± 0.08	Carteret Co., NC	TTCTC	NA	Run 1 and 2
F43 (F12), S18	2011-096-002	*F. gardenii*	4x	3.64 ± 0.08	Carteret Co., NC	AGCCC	NA	Run 1 and 2
F47, S22	2011-123-001	*F. gardenii*	4x	3.68 ± 0.04	Richmond Co., NC	GCTTA	NA	Run 1 and 2
F48, S23	2011-123-002	*F. gardenii*	4x	3.68 ± 0.04	Richmond Co., NC	GGTGT	NA	Run 1 and 2
F49 (F15), S24	2011-123-003	*F. gardenii*	4x	3.68 ± 0.04	Richmond Co., NC	AGGAT	NA	Run 1
F44 (F29), S19	2011-124-001	*F. major*	6x	5.15 ± 0.10	Orange Co., NC	GTATT	NA	Run 1 and 2
F45, S20	2011-124-002	*F. major*	6x	5.15 ± 0.10	Orange Co., NC	CTGTA	NA	Run 1 and 2
F46, S21	2011-124-003	*F. major*	6x	5.15 ± 0.10	Orange Co., NC	ACCGT	NA	Run 1 and 2
F40 (F23), S13	2011-092-001	*F. major*	6x	5.23 ± 0.11	Marshall Co., AL	CGCTT	NA	Run 1 and 2
F50, S14	2011-092-002	*F. major*	6x	5.23 ± 0.11	Marshall Co., AL	TCACC	NA	Run 1
F51, S15	H2011-036-002	*P. jacquemontiana*	2x	1.59 ± 0.01	Undocumented	CTAGC	NA	Run 1

Determining ploidy level of samples

Knowledge of the ploidy levels of samples is important to library pooling for sequencing to minimize unequal coverage of reads from different samples sequenced in the same lane. The ploidy levels of samples are also crucial to interpretation of the phylogenetic results. To determine the ploidy level of the natural populations of *Fothergilla*, the genome size of each sample was measured using flow cytometry. Approximately 1 cm^2 of young leaf tissue was placed in a petri dish containing 500 μL of nuclei extraction buffer (CyStain ultraviolet Precise P Nuclei Extraction Buffer®; Partec, Mönster, Germany) and chopped finely with a razor blade. Solutions were pipetted through CellTricsTM (Partec), 50 μm, disposable filters. Then, 2 mL of a nucleotide staining buffer solution combined with 6 μL RNase A and 12 μL propidium iodide (CyStain PI absolute P; Partec) was added to the filtered solutions. Samples were incubated for 30 min at 4 °C. Nuclei were then analyzed with a flow cytometer (Partec PA II; Partec) with counts exceeding a minimum of 3000 cells per analysis. Two samples were analyzed for each accession. Holoploid, 2C genome size (i.e., DNA content of entire non-replicated chromosome complement irrespective of ploidy) was calculated based on an internal standard of a known genome size (*Pisum sativum* L. 'Ctirad', 2C DNA = 8.76 pg; Greilhuber et al., 2007). Ploidy levels were determined based on published relationships between genome sizes and ploidy levels for *Fothergilla* (Ranney et al., 2007).

GBS data processing

The raw sequences were first filtered by discarding reads with >10 bases of quality <20. The filtered reads were then split by barcodes (exact match required). We built data sets of haplotypes and SNPs, respectively for each experiment. Haplotypes/Tags were identified with a read depth of >5 (Run 1) and >2 (Run 2) called using both STACKS v 1.03 (Catchen et al., 2011) and TASSEL 3.0 (Glaubitz et al., 2014) pipelines on a Linux workstation. A read depth of two would be considered insufficient for discovery of SNP loci, but because the loci had been discovered in the first run, it was sufficient to detect a given haplotype. Furthermore, the haplotype data were generated for phylogenetic analyses and no homozygosity or heterozygosity at any individual loci needs to be assumed. Haplotypes were scored for presence and absence and converted into binary data matrices for phylogenetic analyses. Alleles were defined using the standard of haplotypes with one bp difference between alleles, incorporating error sequences with no more than two bp difference from the locus consensus. SNP calling was performed using the Universal Network-Enabled Analysis Kit (UNEAK) workflow (Lu et al., 2013) filtered for bi-variant orthologous loci. The UNEAK workflow was designed for identifying SNPs from GBS data of switchgrass (*Panicum virgatum* L.), a biologically complex species with multiple ploidy levels, a self-incompatible breeding system, and no reference genome. With the network filtering application, the UNEAK pipeline can remove paralogs and error tags through discarding complex networks of tag pairs and possible repeats (see Lu et al., 2013). We applied the network filter to identify and remove paralogs to improve the quality of the SNP data. The SNPs with higher coverage were more likely to be retained after the network filtering. For comparison, alternative pipelines *py*RAD (Eaton

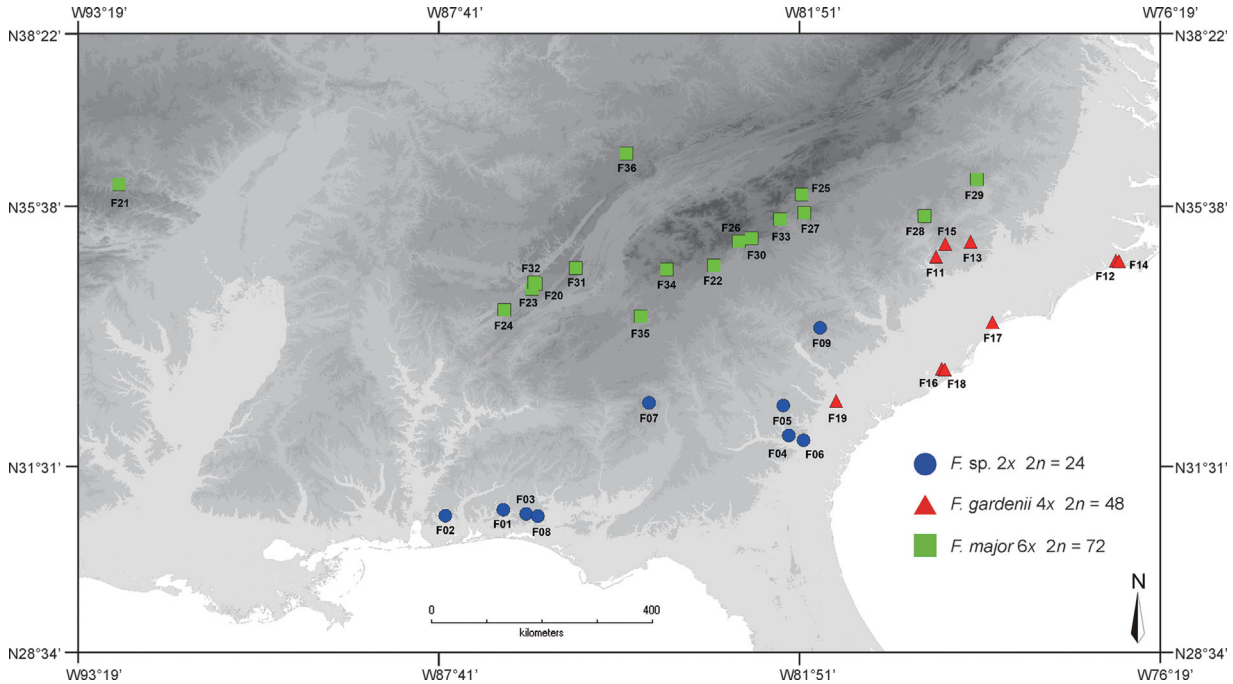

Fig. 1. Geographic distribution of 36 *Fothergilla* accessions sampled in the study. Blue circles represent 2*x F.* sp., red triangles represent 4*x F. gardenii*, and green squares represent 6*x F. major*. The USA elevation map was downloaded from the DIVA-GIS country level map website: http://www.diva-gis.org/gdata & visualized with DIVA-GIS 7.5 (Hijmans et al., 2012).

& Ree, 2013) and STACKS were used to call orthologous loci through alignment-clustering of fragments. *py*RAD (http://dereneaton.com/software/pyrad/) allows for the inclusion of indel variation in the alignment-clustering of fragments. This advantage improves identification of homology across highly divergent samples, more appropriate for RAD-tag studies at both above and below species level phylogenetic scales. SNPs calling using STACKS followed the procedure of Rubin et al. (2012) with user-specified parameters for required similarity within an individual and between individuals.

Phylogenetic analyses

To elucidate the evolutionary relationships of *Fothergilla* taxa, phylogenetic analyses of haplotype and SNPs data were conducted using neighbor-joining (NJ) (only for haplotype data) and maximum parsimony (MP) in PAUP* 4.0 (Swofford, 2003), maximum likelihood (ML) implemented in RAxML V.7.2.8 (Stamatakis et al., 2008) on the CIPRES cluster (Miller et al., 2010). Support of trees was estimated using bootstrap analyses of 500 replicates for NJ tree, 500 replicates with full heuristic search of 100 random taxon sampling for MP tree. ML bootstrap values were estimated from 500 or 1000 replicates. For MP analysis, heuristic search of 100 replicates of random taxon addition with character state unordered and equally weighted, Multitrees On, and other default settings. The same GTR+ I + G model, selected by Akaike Information Criteria (AIC) in *JMODELTEST* (version 2.1.4; Posada, 2008) for the SNP data, was applied for each ML analysis of SNP datasets. For haplotype datasets, binary data parameters were set in RAxML analysis.

For both the SNPs and the haplotype data, in order to reconstruct possible network-like evolutionary relationship among the species, SplitTree4 (Huson & Bryant, 2006) was used to generate the split network by implementing neighbor net analysis with variance of ordinary least squares. In all analyses, missing data were treated as unknown.

Structure analysis

We explored the genetic structure of each of the *Fothergilla* taxa using fastSTRUCTURE (Raj et al., 2014) that was modified from STRUCTURE 2.3.4 (Pritchard et al., 2000) for large scale SNP data. Structure software assigns individuals to populations using genotype data and a Bayesian statistic method. The analysis was performed with the filtered genotypes from STACKS analysis under an admixture model with correlated allele frequencies, a burnin of 200 000 generations, 2 000 000 MCMC iterations after burnin. Ten independent runs were performed with fastStructure for each *K* from 1 to 5. The optimum *K* value were determined by STRUCTURE Harvester (Earl & vonHoldt, 2012).

Sanger sequencing and phylogenetic analyses of plastid and nuclear loci

Sanger sequencing of three plastid DNA regions (*rps16*, *trnL-F* and *matK*) and a nuclear external transcribed spacer (ETS) were performed for the 39 accessions to compare with the GBS data on phylogenetic inference. PCR and sequencing primers and methods followed those used by Xie et al. (2010) in a phylogenetic study of Hamamelidaceae. Sequences were aligned using MUSCLE (Edgar, 2004), and checked by eye. Phylogenetic analyses were conducted for the plastid data, ETS data, and the combined data separately using ML implemented on RAxML V.7.2.8 (Stamatakis et al., 2008) and Bayesian Inference on MrBayes 3.2.2 (Ronquist &

Huelsenbeck, 2003; Ronquist et al., 2012) on the CIPRES server (Miller et al., 2010). We used Akaike Information Criteria (AIC) in *JMODELTEST* (version 2.1.4; Posada, 2008) to determine the appropriate substitution model for ETS, cpDNA and combined cpDNA-ETS dataset. HKY, HKY, and GTR+I were selected as the optimum substitution models for these three datasets respectively. ML was conducted with bootstrap of 500 replicates of full heuristic search and Bayesian Inference was conducted with four chains, 0.5 temp., 10 million generations sampling trees at every 1000 generation and a burn in of 20% with the selected models. Furthermore, we constructed haplotype networks for the plastid, ETS, and combined plastid-ETS sequences separately using TCS (version 1.21; Clement et al., 2000). The haplotype number and matrix used for TCS analysis were calculated by DNASP (version 5.0; Librado & Rozas, 2009). The haplotype networks were constructed under the 95% statistical parsimony criterion. Indels (gaps) were treated as single mutation events, and coded as substitutions (A or T).

Results

Results of flow cytometry

The 2C genomes sizes ranged from 1.70 to 1.82, 3.33 to 3.76, and 5.09 to 5.40 pg for *Fothergilla* sp. (diploids), *F. gardenii* (tetraploids), and *F. major* (hexaploid), respectively (Table 1), confirming the ploidy levels for these taxa based on published values. The outgroup species *Hamamelis virginiana* had 2C values of 2.06–2.10 pg, while that of *Parrotiopsis jacquemontiana* was 1.59 pg.

Illumina sequencing data

Tag numbers per sample recovered from STACKS analysis showed a positive relationship with ploidy levels ($R^2 = 0.4388$; Fig. S1 for Run 2; data not shown for Run 1), as expected. In Run 1 of the 12 samples, 15 787 Loci and 24 337 putative haplotypes were recovered. A majority of the loci (64.96%) were present in 1–3 samples, while 2064 loci (13.1%) are present in 85% of the samples (>10) (Table S1). For most of the loci, the number of haplotypes/alleles found is within the expected range, e.g., 1–2 for 2*x* samples, 1–4 for 4*x* samples, and 1–6 for 6*x* samples. Exceptions were found in only 25 loci, most of which had 1–2 extra alleles/haplotypes than expected (Table S2 for Run 1). These extra alleles are likely a result of sequencing errors, but could represent paralogous loci. Analyses with UNEAK called 3872 SNPs from the sequencing reads obtained from Run 1. For Run 2, more missing data were present due to the lower concentration of library DNA in the sample pool. Five samples had low read counts (fewer than 10 000 reads per sample).

A total of 15 344 haplotypes from 7845 loci were called from TASSEL, of which 2410 loci were polymorphic and 976 of the polymorphic loci were present in 30 or more of the 39 samples (Table S1). A total of 3789 Haplotypes of 933 loci were called from STACKS from the 36 *Fothergilla* accessions, of which 402 loci were present in >30 samples. At these 402 loci, none of the *Fothergilla* diploid individuals had more than two haplotypes/alleles per locus; loci with putative "triploid" haplotypes were found only in 4*x* or 6*x* plants except one locus in the outgroup *Parrotiopsis jacquemontiana* (Table S3). None of the loci had five or more haplotypes and only one locus with four haplotypes in a 6*x* sample (F35). In the full set of 933 loci from STACKS, four loci have three haplotypes in four diploid samples (Table S4). When combining STACKS haplotypes from Run 1 (48 samples) and Run 2 (12 samples), 1779 were present in all the 60 samples. A total of 2344 loci found in Run 1 were recovered in Run 2 in one or more pairs of the nine libraries sequenced twice, of which, 190 loci were recovered in all the nine library pairs loaded in both runs (Table S5).

For SNPs calls, a total of 838 bi-variant loci were recovered for the 39 new libraries of Run 2 from analysis using UNEAK. Among these, 165 loci were present in 29 or more samples. For SNPs called by STACKS, a total of 894 loci were common in two runs of the same nine libraries. After filtering missing data (removing 3 samples with low counts), a total of 345 SNPs common in two runs of all samples and 227 loci present in ≥30 samples. The GBS data are available on GenBank SRA (BioProject ID: PRJNA292973; http://www.ncbi.nlm.nih.gov/bioproject/292973; haplotype and SNP data matrices for the results presented below were submitted to TreeBase (Submission ID: 18134; available at http://purl.org/phylo/treebase/phylows/study/TB2:S18134).

Sanger sequencing

Sanger sequencing resulted in 2692 aligned base pair (bp) of plastid DNA data, of which 885 bp were from *matK*, 951 bp from *trnL-F*, 851 bp from *rps16*, and 321 bp from ETS data (for GenBank accessions of sequences, see Table S6). The polymorphic sites in these DNA regions were 9, 3 (including a 6-bp indel), 5, and 4, respectively. A total of 11 haplotypes of plastid DNA, five haplotypes of ETS, and 13 haplotypes of the combined plastid-ETS sequences were identified (Table 2). Most of the haplotypes occurred in the nine 2*x* samples, including six of the 11 plastid haplotypes, 4 of the 5 ETS haplotypes, and 8 of 13 plastid-ETS haplotypes. No haplotypes were shared between the 2*x* and polyploidy taxa but a majority of the 6*x* and 4*x* shared the most common haplotype that appear to be ancestral. All sequences were obtained by Sanger sequencing of PCR products directly including the nuclear region, as concerted evolution presumably homogenized the ETS.

Phylogenetic analyses

Analyses of haplotype data from Run 1 using MP and ML methods both resulted in a strongly supported clade of 6*x* samples and a close relationship of the 2*x* samples with the 4*x* samples that are also shown to be monophyletic when two samples (S17, S18) with relatively low counts due to low quantity of DNAs in the pool were removed (Fig. 2). Analyses of the SNP data from Run 1 resulted in a strongly supported 4*x* clade sister to the 2*x* samples and a basal polytomy of the 6*x* samples (tree not shown). The SNP data common in the nine repeated samples of Run 1 and Run 2 suggested the same relationships (if sample 6 with extensive missing data is excluded), but the support for the 4*x* clade was relatively low (Fig. 3). Analyses of the 15 344 haplotype data from TASSEL of 29 samples (10 samples having high amount of missing data removed) from Run 2 revealed a strongly supported 6*x* clade, and showed the 2*x* samples as a monophyletic group consisting of two strongly supported subclades, however

Table 2 ETS and plastid sequence polymorphisms for 13 haplotypes in 36 *Fothergilla* individuals

	Nucleotide position																							
	ETS				*matK*									*rps16*					*trnL-F*					
Haplotype	88	181	219	276	375	483	526	575	686	754	1026	1169	1185	1285	1574	1725	1845	1269	2137	2877	2893			
H1	G	T	T	C	T	C	C	C	T	A	C	A	G	.	C	T	G	A	A	1	T			
H2	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
H3	.	.	C	.	.	A	.	.	G	.	.	.	.	.	A	.	.	.	.	.	.			
H4	.	C	C	A	C	.	A	.	G	.	.	.	.	.	A	.	.	.	C	.	.			
H5	.	.	C	.	.	.	A	.	G	C	.	.	.	.	A	.	T	.	.	o	.			
H6	.	C	C	.	.	.	A		G	C	.	.	.	.	A	.	T	.	.	o	.			
H7	.	C	C	.	.	.	A	T	G	.	T	.	.	C	A	A	.	.	.	.	.			
H8	A	C	C	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.			
H9	A	C	C	.	.	.	A	.	G	.	.	.	A	.	A	.	.	.	.	.	.			
H10	A	C	C	.	.	.	A	.	G	.	.	.	.	.	A	.	.	.	.	.	.			
H11	A	C	C	.	.	.	.	.	G	.	.	.	.	.	A	.	.	.	.	.	.			
H12	A	C	C	.	.	.	A	.	G	.	.	.	.	.	A	.	.	.	.	.	G			
H13	A	C	C	.	.	G	A	.	G	.	.	C	.	.	A	.	.	.	.	.	.			

All sequences are compared to the reference haplotype H1. Numbers '0/1' in the sequences indicate absence/presence of length polymorphism. Note that poly-A or poly-T stretches were excluded from analyses. 1, TTTTTA.

with low support (61% bootstrap value). Relationships among the 2*x* and 4*x* samples remained unresolved (Fig. 4A). Analyses of the 3789 haplotype data from STACKS of Run 2 recognized the two strongly supported subclades of 2*x* samples, a strongly supported paraphyletic 4*x* samples with the 2*x* samples nested within, and a basal polytomy of the 6*x* samples except F20 (6*x*) (Fig. 4B). Analyses of the 838 bi-variant UNEAK SNPs from 39 samples of Run 2 recognized a monophyletic 6*x* group and an intermixed 2*x* and 4*x* groups, but both with low bootstrap values (Fig. 5). Analyses removing loci with missing data in >10 samples showed the same relationships as those revealed from the 838 loci (results not shown). Phylogenetic networks reconstructed using NeighborNet on SplitTree4 with SNPs from UNEAK further suggested the same results showing a clearly defined 6*x* lineage and the same two subclades of 2*x* nested among the 4*x* complex (Fig. 6A). When the SNP data were further filtered using Network Filter in UNEAK (removing complex networks of tag pairs and possible repeats), results from SplitTree4 showed the 2*x* subclades form a sister group distinct from the 4*x* complex (Fig. 6B). The consensus tree from the *py*RAD analysis supported a clade of 4*x* and 2*x* samples (Fig. 7); the 2*x* samples form a monophyletic group more closely related to 4*x* F15, F16, F18, and F19 within the clade.

Analyses of sequences from Sanger sequencing using ML and Bayesian methods produced similar trees with low resolution. The plastid data resolved the same two subclades of 2*x* samples although one of them did not include all the members (e.g., F04, F07 did not group with F05, F06, and F09; Fig. S2). The 2*x* subclade (F01, 02, 03, 08) from the Gulf coast was grouped with 4*x* F12, F13, and F14 in a clade with strong support (Fig. S2); the other 2*x* subclade (F05, F06, and F09) from the coastal plain was also well supported. Relationships among these two clades and the rest of the samples remained unresolved. The ETS tree recognized a strongly supported clade of all 2*x* samples (Fig. S3). The combined plastid-ETS data resolved a monophyletic group of the 2*x* samples with low support, containing the same two subclades as found in the plastid tree (Fig. S4). Haplotype networks showed results consistent with the phylogenetic trees. The ETS haplotype network showed a lineage of four haplotypes (H1–H4) from the 2*x* samples and all the 4*x* and 6*x* samples shared one haplotype (H5 in Fig. 8, A1) that was suggested to be ancestral based on the rooted ML haplotype geneology (Fig. 8, A2). The plastid haplotype network recognized a lineage of haplotypes (H1, H2, H6) from the Gulf coast 2*x* individuals (F01, F02, F03, F08) directly connects to the haplotype (H9) from three 4*x* individuals (F12, F13, F14) (Fig. 8, B1), similar to the plastid ML tree (Fig. S2). Three haplotypes from the Atlantic coast 2*x* individuals were recognized, each independently connecting, via two to four mutations, to the most common haplotype H8 shared by all the 6*x* (except two) and remaining 4*x* (except one), (Fig. 8, B1). These results are not in conflict with those from GBS.

Overall then, the exact alignments of the two 2*x* subclades and two 6*x* plants (F20, F22; Fig. 1) differed among trees. In the TASSEL haplotype tree (Fig. 4A), the UNEAK filtered SNP network reconstructed using SplitTree4 (Fig. 6B), *py*RAD SNP tree (Fig. 7), the ETS tree (Fig. S3), the combined plastid-ETS tree (Fig. S4), the two 2*x* subclades formed a monophyletic group. In contrast, The STACKS haplotype neighbor-joining

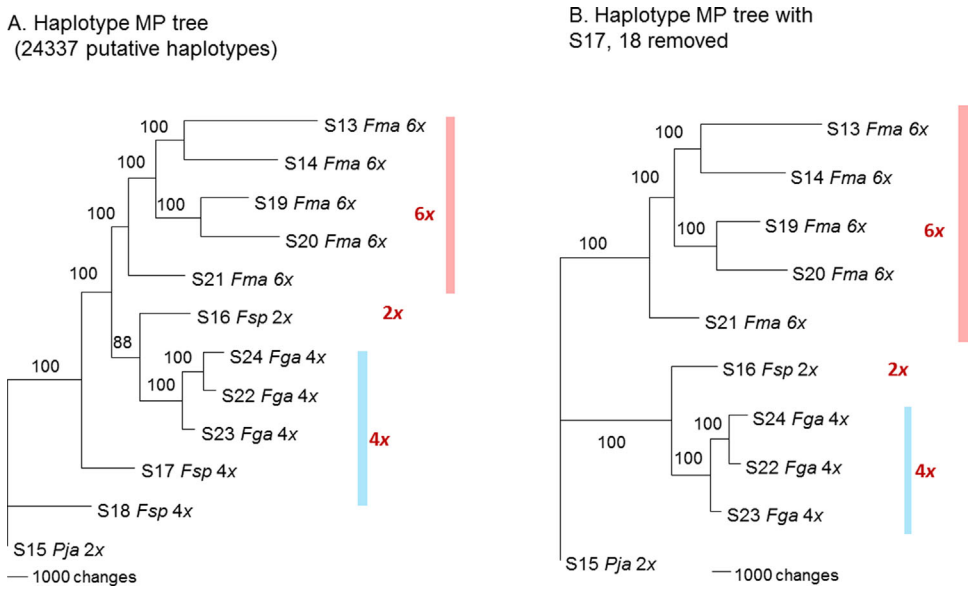

Fig. 2. Phylogenetic relationships inferred from maximum likelihood analyses of STACKS Haplotype and UNEAK SNPs from Run 1 of 12 samples. Samples denoted as S17 and S18 in Table 1 were removed in figure B due to relatively lower counts. Numbers on branches are values of bootstrap support. *Fga*, *Fothergilla gardenii*; *Fma*, *F. major*; *Pja*, *Parrotiopsis jacquemontiana*; *Fsp*, *F.* sp.

tree (Fig. S5) and the UNEAK unfiltered SNP network from SplitTree4 (Fig. 6A) supported a closer genetic relationships of the Atlantic coastal 2*x* lineage (F04, 05, 06, 09) with geographically adjacent Atlantic coastal 4*x* *F. gardenii* (Figs., 6 1A); the STACKS ML and NJ trees (Figs. 4B, S5) suggested the closest alignment of the Atlantic coastal 2*x* (except F04) with 4*x* F19 and the Gulf coastal 2*x* (including F04 from the Atlantic coast) with 6*x* F22 (Fig. 4B). The plastid gene tree from Sanger sequencing (Fig. S2) further supported the alignment of the Gulf coastal 2*x* lineage (F01, 02, 03, 08) with some of the Atlantic coastal 4*x* plants (F12, 13, 14) (Fig. S2). In the STACKS haplotype ML and NJ trees (Figs. 4B, S5), 6*x* F20 was resolved as a distinct lineage sister to the remainder of all samples. In other trees, this sample was either excluded (filtered out due to high amount missing data) or placed in the 6*x* group. The unusual placements of F20 and F22 were likely the result of homoplasy or stochastic error because the F22 did not contain the 2*x* genome based on STRUCTURE results.

Results of STRUCTURE analysis

Analysis of genotype data set generated by STACKs with STRUCTURE assigned the individuals to three subgroups corresponding to the 2*x*, 4*x*, and 6*x* taxa. The most probable number of subgroups was identified at $K = 3$ by STRUCTURE Harvester (Fig. S6). The genomes of the 4*x* plants were shown to contain a majority of the 2*x* genome (~3/4) and a small portion of the 6*x* genome (~1/4; Fig. 9).

Discussion

Influences of experimental repeats, software, taxon sampling, and data type

RAD-tag/GBS data have shown to be a useful tool in evolutionary biology and genetics (reviewed in Davey & Blaxter, 2010, Davey et al., 2013; McCormack et al., 2013; Buerkle & Gompert, 2013; Peterson et al., 2012). The method has been applied to identification of QTL and genome-wide associate mapping (Amores et al., 2011; Chutimanitsakun et al., 2011; Baxter et al., 2011; Pfender et al., 2011; Gompert et al., 2012), phylogenetics and phylogeography (Emerson et al., 2010; Rubin et al., 2012; Eaton & Ree, 2013; Cariou et al., 2013; Nadeau et al., 2013; Pante et al., 2015), and population genetics/genomics

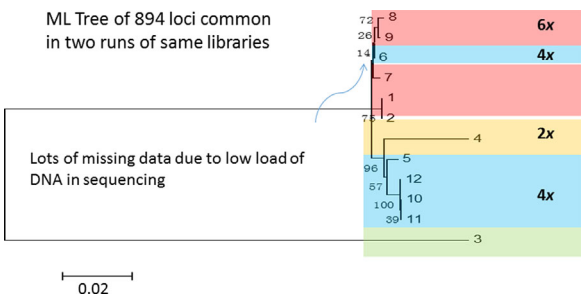

Fig. 3. Phylogenetic relationships inferred from maximum likelihood analysis of common SNPs from 894 loci in repeated samples of the same libraries loaded in both Run 1 and Run 2. SNPs were called from STACKS pipeline. Numbers on branches are values of bootstrap support. Numbers on branch terminals represent the combination of the same individual library sequenced in the two Runs. Number 1 represents S13 and F40; Number 2 represents S14 and F50; Number 3 represents S15 and F51; Number 4 represents S16 and F41; Number 5 represents S17 and F42; Number 6 represents S18 and F43, Number 7 represents S19 and F44, Number 8 represents S20 and F45, Number 9 represents S21 and F46; Number 10 represents S22 and F47, Number 11 represents S23 and F48, Number 12 represents S24 and F49. Samples of S17 and F43 and outgroup were removed due to low counts. Corresponding taxon names of S and F numbers are referred to Table 1.

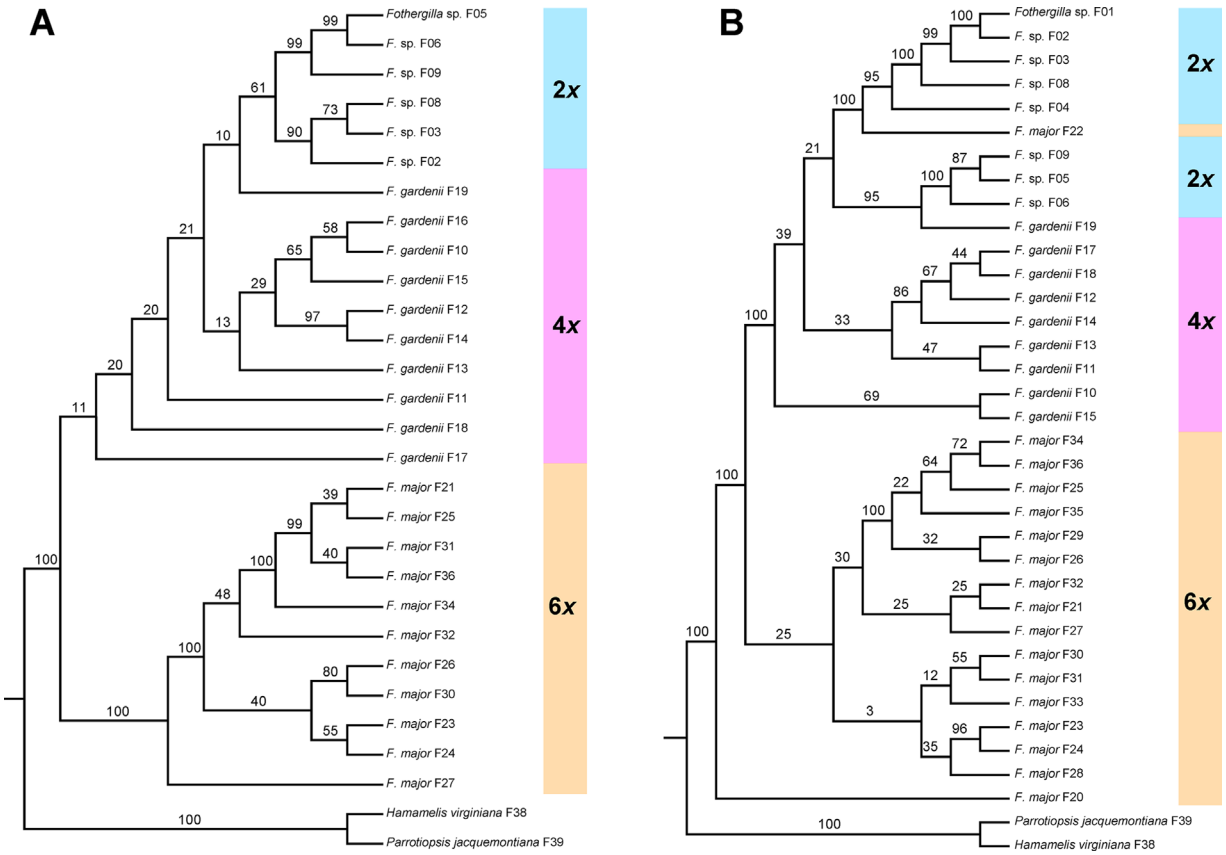

Fig. 4. Phylogenetic trees resulting from analyses of haplotypes from Run 2 using maximum likelihood method. Numbers on branches are values of bootstrap support. **A,** TASSEL Haplotypes (29 samples; ten samples with extensive missing data removed). **B,** STACKS haplotypes (36 samples; three with extensive missing data removed).

(Andersen et al., 2012; Hohenlohe et al., 2012a, 2012b; Gompert et al., 2012). Although RAD-seq and RAD-tag based GBS data have shown to be promising for phylogenetic studies of recently diverged taxa, few studies have tested how repeatable are phylogenetic outcomes from data obtained from different library preparations of the same samples, from different sequencing runs of the same libraries, different taxon sampling scales, different software pipelines, and different data types. Therefore, in our study, we attempted to discern the impacts of these variables on *Fothergilla* phylogeny. By repeating the sequencing of some of the same libraries, using the same samples in different library preparations, and by comparing haplotype and SNP data called using four different software pipelines (STACKS, TASSEL, UNEAK, *py*RAD), we found substantial differences in the data amount of haplotypes and SNPs obtained from different software. However, there were large number of haplotypes and SNPs shared between the two runs from the same software pipeline (e.g., STACKS), even the second run was prepared for much lower reads/tag sequencing. Furthermore, the phylogenetic patterns revealed were quite consistent among the data sets in recognizing a distinct 6*x F. major* clade, two well-supported subclades of the 2*x* plants (one from the Atlantic coastal plain and the other from the Gulf coast of the southeastern United States of America (e.g., Figs. 6, 7). Although the support for the 6*x* clade and alignment of the 2*x* plants with the 4*x* plants are not always high, the pattern repeatedly emerged in all data sets and phylogenetic methods used. However, variation in tree topologies derived from different data sets did exist, largely involving the relationships of the Atlantic coastal plain diploid lineage, being sister to the Gulf coastal diploid lineage or closer to the geographically adjacent 4*x* plants (e.g., Figs. 6–8). The discrepancy of the relationships revealed by different datasets suggested that different software may produce data biased toward certain portions of the polyploid genomes that tracked different aspects of their evolutionary histories. For example, the unfiltered UNEAK SNPs and STACKS haplotypes may have been biased toward one parental 2*x* genome of the 4*x* plants (genome portions of "a" and "b" in Fig. 10), the 2*x* genome of the Atlantic coastal plain lineage, resulting in the close association of the 4*x* and Atlantic coastal 2*x* plants in the phylogeny (Fig. 6A). On the other hand, the data from *py*RAD SNPs, filtered UNEAK SNPs, and the TASSEL haplotypes may have biased more toward the other 2*x* genome, an extinct or undiscovered parental genome, of the 4*x* plants (genome portion "c" in Fig. 10; referred to as "extinct" hereafter) that was evolutionary divergent from the extant 2*x* lineages, leading to the recovering of the monophyly of the two extant diploid lineages (Figs. 4A, 6B, 7). This argument is supported by the plastid DNA data whose gene tree revealed additional aspects of the evolutionary history of *Fothergilla* not apparent in the GBS trees; that is, a closer relationship of some of the Atlantic

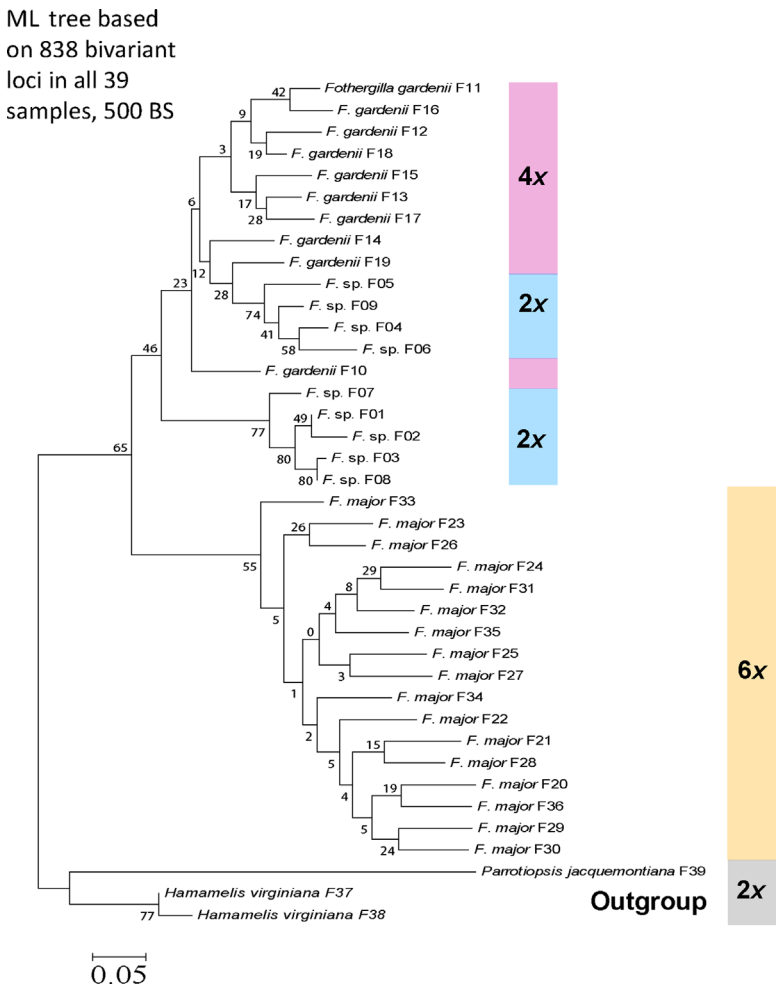

Fig. 5. Phylogenetic trees resulting from maximum likelihood analyses of UNEAK SNPs of Run 2. Bootstrap support ≥50% are shown on branches and bootstrap support <50% are not shown.

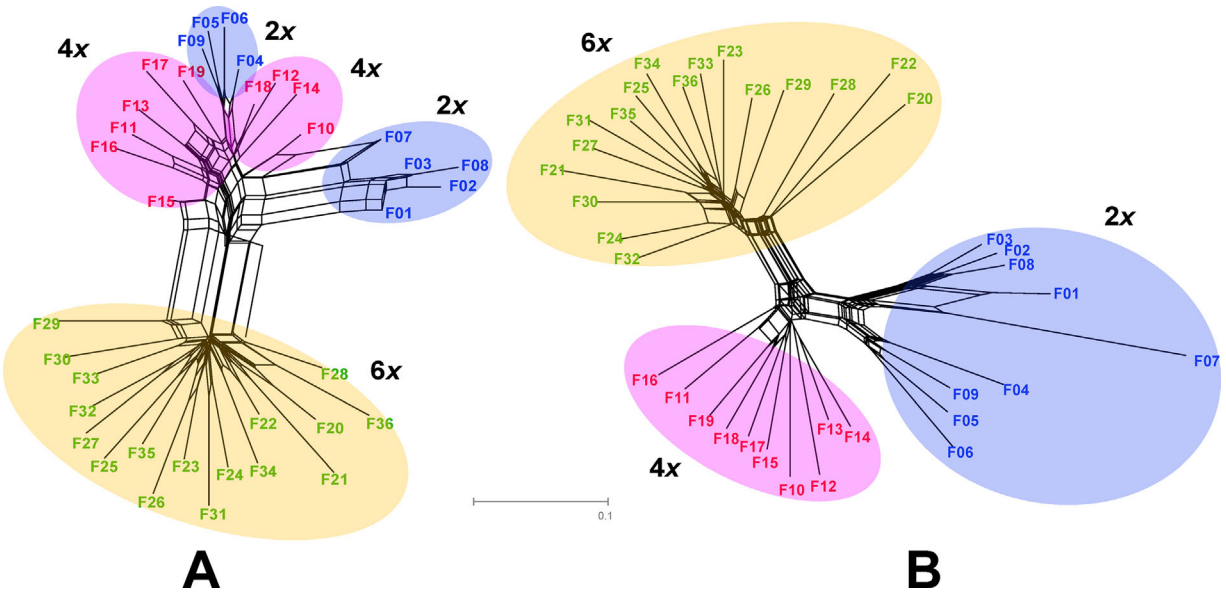

Fig. 6. Phylogenetic networks generated by NeighborNet method using SplitTree 4. **A,** 838 SNPs dataset generated by UNEAK pipeline. **B,** 165 SNPs dataset filtered from 838 SNPs dataset using Network Filter in UNEAK package.

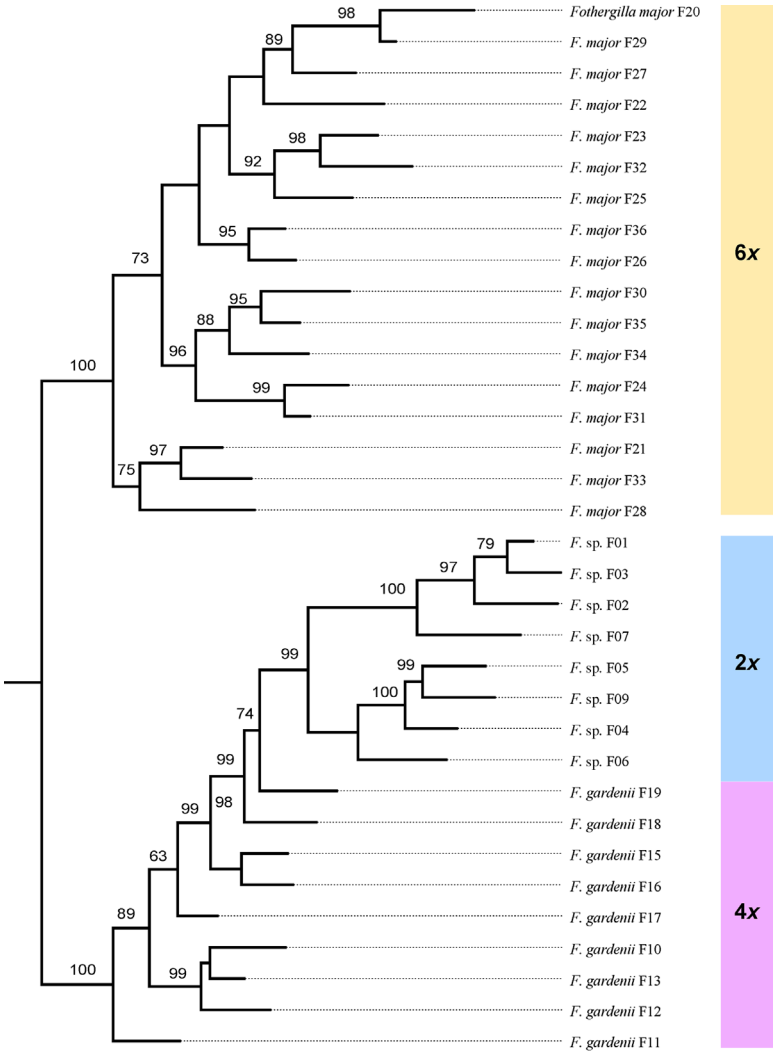

Fig. 7. Phylogenetic tree resulting from clustering analysis of data from *py*RAD Clustering Orthologues with 157 621 sites. Bootstrap support ≥50 are shown on branches. Outgroups were removed from the tree due to the long branch connecting to the ingroup.

costal 4*x* plants with the Gulf coastal 2*x* lineage rather than with the Atlantic coastal 2*x* lineage (Figs. 8, S2), suggesting maternal contribution of the Gulf coastal 2*x* lineage to the origins of the Atlantic coastal 4*x* *F. gardenii* via hybridization. Monophyly of the 2*x* plants supported by the ETS data (Fig. S3) can be explained by concerted evolution of ETS fixing the ETS type from the "extinct" 2*x* lineage with the "c" genome. Therefore, for lineages without reticulation, we suggest that GBS haplotype and SNP data from these different software pipelines are robust to produce similar phylogenetic results, but applying different data types and different software may help to uncover additional evolutionary histories not apparent from analysis of one data set, especially in polyploid lineages.

Origins and evolution of polyploid *Fothergilla* species

The phylogenetic patterns and results from STRUCTURE analysis clearly support the progenitor-derivative relationships between the diploid and tetraploid taxa (Figs. 2–9, S2–S5). The STRUCTURE result also supports the contribution of the 6*x* species to the 4*x* species that was shown to carry approximately ¾ of the 2*x* genome and ¼ of the 6*x* genome. Based on these and the phylogenetic results, we hypothesize that the 4*x* species evolved from genome doubling of a 2*x* genome that was a hybrid containing ¾ genome of the extant 2*x* lineages ("a" and "b" portions in Fig. 10) and ¼ of an "extinct" 2*x* lineage ("c" portion in Fig. 10). The F_1 was potentially derived from at least two hybridization events between the Gulf coastal 2*x* lineage (once maternal and once paternal) and the unknown, more divergent 2*x* lineage (Fig. 10). This F_1 likely backcrossed to the Atlantic 2*x* plants as pollen receivers to form a genome containing the original F_1 plastid genome, ¾ of the extant 2*x* genome and ¼ of the "extinct" 2*x* genome that was the progenitor of the 6*x* *F. major*, followed by genome doubling to form the 4*x* plants. This scenario explains the genome relationships revealed by STRUCTURE analysis among the 2*x*, 4*x*, and 6*x* taxa and the phylogenetic relationships revealed by the GBS data and the plastid DNA sequence data. These data showed a combined genome of the 2*x* and 6*x* in the 4*x* genome biased toward the 2*x* (Fig. 10) and at least two major distinct plastid DNA types in

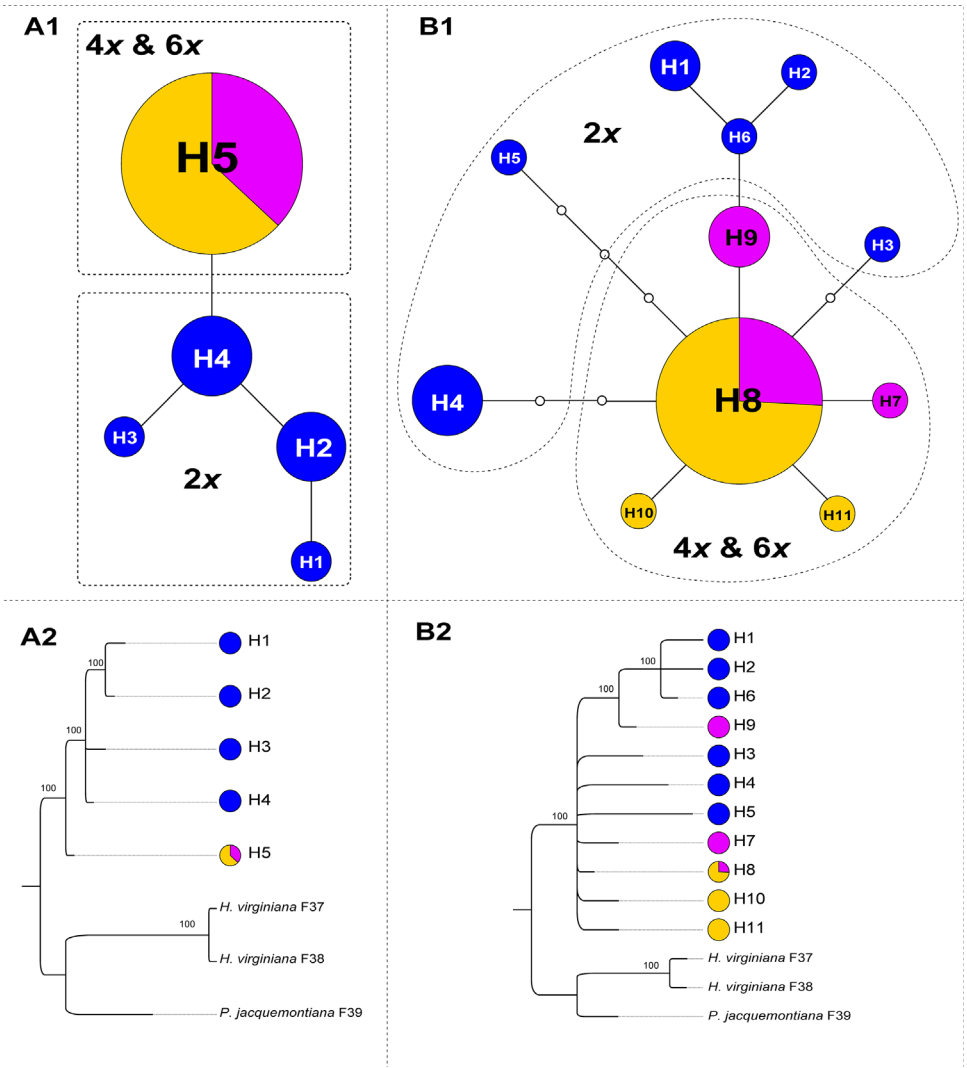

Fig. 8. Plastid and ETS haplotype networks and genealogies reconstructed from TCS and ML analyses, respectively. **A1, B1**, The 95% statistical parsimony networks of five ETS haplotypes (**A1**) and 11 plastid (*matK*, *rps16*, *trnL-F*) haplotypes (**B1**). Blue indicates haplotypes occurring in the 2*x* samples; purple indicates haplotypes occurring in the 4*x* samples while orange indicates haplotypes occurring in the 6*x* samples. The size of circles corresponds to the frequency of each haplotype. Colored area in each haplotype is proportional to the individual numbers of each species. The small open circles represent extinct or missing haplotypes. **A2, B2,** Maximum likelihood phylograms of ETS (**A2**) and plastid (**B2**) haplotypes. Bootstrap values >50% are indicated above the branches. Haplotype sequences are presented in Table 2.

the 4*x* species with one (Fig. S2; H9 in Fig. 8, B1) closely associated with that of the Gulf coast diploid lineages (H1, H2, H6 in Fig. 8B1) and the other(s) (H7, H8 in Fig. 8) more divergent and likely ancestral, to those of the extant diploid lineages. The phylogenetic and STRUCTURE results largely supported the hypothesis that the extant diploid and tetraploid lineages were likely not involved in the origin of the hexaploid species. The data from STRUCTURE and plastid DNA haplotype network suggested, instead, that the "extinct" 2*x* lineage contributing to the origin of the 4*x* *F. gardenii* was likely the progenitor of the 6*x* taxon *F. major*. It is possible that hexaploid *F. major* was formed by two successive autopolyploidy events of the "extinct" 2*x* lineage, intervened by a backcross of the autotetraploid 4*x* to the 2*x* parent (Fig. 10). This "extinct" 2*x* parent shared by the 4*x* and 6*x* species likely had an ETS haplotype H5 (Fig. 8A) that was fixed in the 4*x* and 6*x* species as a result of concerted evolution and a plastid haplotype H8 that was ancestral and shared by a majority of the 4*x* and 6*x* (Fig. 8B). The "extinct" 2*x*, the Gulf coastal 2*x*, and the Atlantic coastal 2*x* lineages were likely once sympatric in the south (e. g., in the Florida Panhandle of the Gulf coast) allowing for hybridization followed by polyploidy and subsequent allopatry, possibly during the Pleistocene glaciation cycles.

Alternatively, the origin of the 4*x* species may also be explained by autoploidy of the extant 2*x* followed by introgression of the 6*x* into the 4*x*, presumably during cycles of the Pleistocene glaciation where all of the cytotypes could have been sympatric in the south. This could have occurred through multiple pathways: (i) The 6*x* and 2*x* plants could

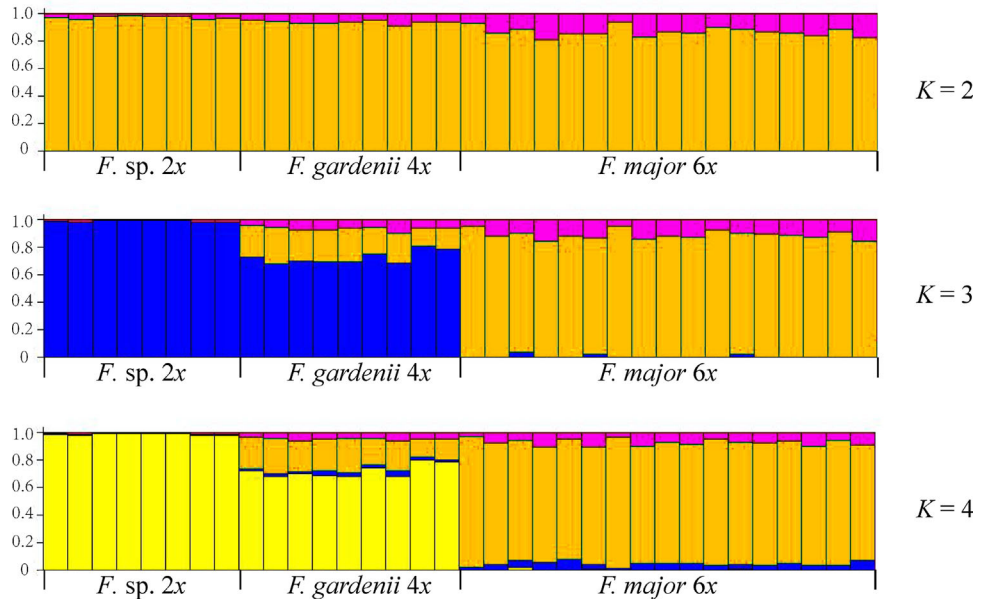

Fig. 9. Genetic structure of the *Fothergilla* 36 samples for $K = 2$ to $K = 4$. Each individual is represented by a vertical bar, partitioned into *K* segments representing the amount of assignment of its genome in *K* clusters as shown by different colors. Genotype matrix was generated by STACKS pipeline.

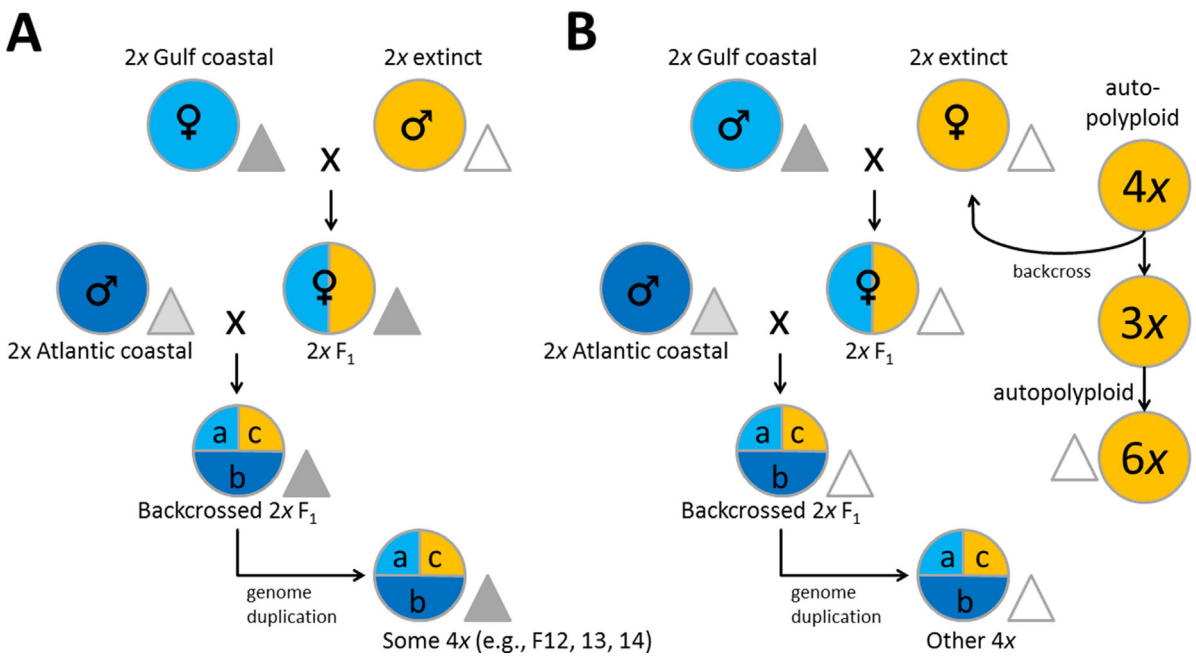

Fig. 10. Schematic diagram illustrating the hypothesized origins of *Fothergilla gardenii* (4*x*) and *F. major* (6*x*) based on results from phylogenetic analyses, STRUCTURE, haplotype genealogy, and flow cytolometry. 2*x* Gulf coastal: F01, F02, F03, F08; Atlantic coastal: F04, F05, F06, F07, F09. Colored circles represent nuclear genome with dark blue and light blue represent the two extant 2*x* genomes and yellow represents the "extinct" 2*x* genome. Dark gray, light gray, and open triangles represent the plastid genomes of the Gulf coastal, Atlantic coastal, and "extinct" 2*x* lineages, respectively. **A,** hypothesized hybridization events resulting in the 4*x* plants of F12–F14 with a nuclear genome more similar to the Atlantic coastal 2*x* plants but a plastid genome more similar to the Gulf coastal 2*x* plants. **B,** hypothesized hybridization events resulting in (left) the 4*x* plants other than F12–F14 that have a nuclear genome more similar to the extant 2*x* plants but a plastid genome more similar to the "extinct" 2*x* plants, and (right) the 6*x* plants having a nuclear and plastid genome derived from the "extinct" 2*x* parent. The plastid genome is shared with a majority of the 4*x* plants.

have hybridized to create F_1 4*x* plants that no longer exist. These F_1 4*x* plants then backcrossed/interbred with the larger pool of 4*x F. gardenii.* (ii) The 6*x* and 4*x* plants could have hybridized to create F_1 5*x* plants as has been documented in garden settings (Ranney et al., 2007). Although pentaploids typically have low fertility, they can potentially serve as a reproductive bridge and have been reported to produce occasional 2*x* gametes in other genera (Vorsa et al., 1987). In doing so, 5*x* plants could have backcrossed to 4*x F. gardenii* producing tetraploids. However, both scenarios would likely result in 4*x* complexes with a wide variation in the proportion of the 6*x* genome in the 4*x* plants. This does not seem to be the case based on the STRUCTURE results. These scenarios are also not supported due to the lack of 5*x* hybrids in nature.

The origin of hexaploid species might have involved, at a point, a third "extinct" and more divergent 2*x* parent (yellow genome in STRUCTURE results; Fig. 9) whose genome signal was subsequently reduced by gene conversion or other mechanisms. If so, this process likely occurred rapidly as the low haplotype diversity and low phylogenetic diversity within the 6*x F. major* suggest relatively recent origin of the species. However, the vast majority of the haplotype loci recovered from the 6*x* plants have only two alleles (Table S3), supporting a single 2*x* ancestor. Alternatively a strong selection for one parental genome (green genome in Fig. 9) had occurred. At present, we have no evidence to support this argument.

The genetic data typically show that the 6*x F. major* represents one large, divergent clade with no particular subgrouping corresponding to the current distribution (e.g., mountains and Piedmont) or plant form (e.g., plants from mountain locations tend to be larger growing, while some lowland plants are smaller stature and more rhizomatous). These observations further suggest extensive and on-going gene flow within this species and that the variations in phenotype may result from small genotypic variation or environmental differences, e.g., perhaps cooler temperature and greater humidity on the mountains led to the larger stature.

The origin of *Fothergilla* can be traced back to the early Eocene (49–50 million years ago) according to the earliest fossil leaf record of the genus, *Fothergilla malloryi* Radtke, Pigg et Wehr described from the Republic flora of northeastern Washington State (Radtke et al., 2005). Other fossil leaf species of *Fothergilla* were also reported from younger beds, in the Oligocene of North America and several Neogene Asian localities (i.e., the Oligocene Bridge Creek flora of Oregon and of Kazakhstan by Meyer & Manchester, 1997, the Miocene of Shantung, China by Hu & Chaney, 1940, and Japan by Suzuki, 1961). This evidence indicates that *Fothergilla* had an ancient origin and was once more widely distributed from western North American extending to Asia and eastern North America. Extinction in western North America and Asia presumably resulted in today's narrow distribution of *Fothergilla* that is restricted to southeastern North America (Weaver, 1969). The eastern North American lineage would have diversified at the 2*x* level and likely took refugium in the Florida Panhandle Gulf coast during the glaciation, where hybridization occurred and all cytotypes were formed. Interglacial warming possibly have allowed the 6*x* and 4*x* taxa to expand ranges northward, adapted to the Appalachian Mountains, Piedmont, and Atlantic Coast. The much greater diversity in ETS (four out of five) and plastid haplotypes (6 out of 11) in the nine diploid plants similarly supports an ancient origin of the genus (Table 2; Fig. 8).

There is little debate that hybridization and polyploidy play an important role in speciation and angiosperm diversification (Stebbins, 1985; Soltis et al., 2009; Soltis & Soltis, 2009). The profound effects of Pleistocene glacial cycles on species evolution and promoting hybridization and polyploid speciation have also been well recognized (Dufresne & Hebert, 1997; Paun et al., 2006; Casazza et al., 2012; Peirson et al., 2013). Our study adds another example to the evidence, demonstrating the roles of hybridization and polyploidy in the diversification and evolution of *Fothergilla*. The combination of genome wide GBS data with Sanger sequencing of specific markers from plastid and nucleus provided evidence revealing detailed reticulate evolutionary history of the genus involving multiple past hybridization events and polyploidy.

In conclusion, the genetic evidence from RAD-tag-based GBS data supports a divergent hexaploid species *Fothergilla major*, a diploid taxon *Fothergilla* sp. (to be named) consisting of two allopatric lineages, and a tetraploid species *Fothergilla gardenii*. Although variation of loci retrieved from different software pipelines, different experimental runs, and repeated libraries existed and was substantial likely due to biases to different portions of the genome, the total data in combination with the Sanger sequencing of a few plastid and nuclear loci permitted us to reconstruct the evolutionary network of *Fothergilla* that likely involved at least two hybridization events of an extant diploid with an "extinct", divergent diploid lineage, as well as backcrosses and genome doubling.

Acknowledgements

Extensive field collections were completed by Ron Miller and Rick Lewandowski with additional assistance from Tom Patrick, Scott Walker, Kelly Oates, Ray Head, Jon Lindstrom, Fred Spicer, Ewin Jenkins, Clarence Towe, Andy Whipple, and Amira Ranney. We also thank Nathan Lynch, Joel Mowrey, and Will Kohlway for technical support. The PB403/503 2014 class at NCSU generated some of the plastid and ETS sequences. NCSU Genomic Science Lab provided Illumina sequencing service. Funding was provided by the North Carolina Agricultural Research Service, Mt. Cuba Center, USDA-ARS Woody Landscape Germplasm Repository, and the Birmingham Botanical Gardens. The authors of this article have no conflict of interest.

References

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. 2011. Genome evolution and meiotic maps by massively parallel DNA sequencing: Spotted gar, an outgroup for the teleost genome duplication. *Genetics* 188: 799–808.

Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Félix M-A, Kruglyak L. 2012. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics* 44: 285–290.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3: e3376.

Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD, Blaxter ML. 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* 6: e19315.

Buerkle CA, Gompert Z. 2013. Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology* 22: 3028–3035.

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution* 3: 846–852.

Casazza G, Granato L, Minuto L, Conti E. 2012. Polyploid evolution and Pleistocene glacial cycles: A case study from the alpine primrose *Primula marginata* (Primulaceae). *BMC Evolutionary Biology* 12: 56.

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

Chu ND, Kaluziak ST, Trussell GC, Vollmer SV. 2014. Phylogenomic analyses reveal latitudinal population structure and polymorphisms in heat stress genes in the North Atlantic snail *Nucella lapillus*. *Molecular Ecology* 23: 1863–1873.

Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistué L, Corey A, Filichkina T, Johnson EA, Hayes PM. 2011. Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics* 12: 4.

Clement M, Posada D, Crandall KA. 2000. TCS: A computer program to estimate gene genealogies. *Molecular Ecology* 9: 1657–1660.

Davey JW, Blaxter ML. 2010. RADSeq: Next-generation population genetics. *Briefings in Functional Genomics* 9: 416–423.

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. 2013. Special features of RAD sequencing data: Implications for genotyping. *Molecular Ecology* 22: 3151–3164.

Dufresne F, Hebert PD. 1997. Pleistocene glaciations and polyphyletic origins of polyploidy in an arctic cladoceran. *Proceedings of the Royal Society B: Biological Sciences* 264: 201–206.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4: 359–361.

Eaton DA, Ree RH. 2013. Inferring phylogeny and introgression using RAD-seq data: An example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62: 689–706.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA* 107: 16196–16200.

Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L. 2013. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution* 67: 2483–2497.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9: E90346.

Goldblatt P, Endress PK. 1977. Cytology and evolution in Hamamelidaceae. *Journal of the Arnold Arboretum* 58: 67–71.

Gompert Z, Parchman TL, Buerkle CA. 2012. Genomics of isolation in hybrids. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367: 439–450.

Greilhuber J, Temsch EM, Loureiro JCM. 2007. Nuclear DNA content measurement. In: Doležel J, Greilhuber J, Suda J eds. *Flow cytometry with plant cells: Analysis of genes, chromosomes and genomes*. Weinheim, Germany: Wiley-VCH. 67–101.

Griffin PC, Robin C, Hoffmann AA. 2011. A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. *BMC Biology* 9: 1–18.

Hijmans RJ, Guarino L, Bussink C, Mathur P, Cruz M, Barrentes I, Rojas E. 2012. DIVA-GIS 7.5. A geographic information system for the analysis of species distribution data [online]. Manual available from http://www.diva-gis.org (accessed 22 July 2015).

Hohenlohe PA, Bassham S, Currey M, Cresko WA. 2012a. Extensive linkage disequilibrium and parallel adaptive divergence across three spine stickleback genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367: 395–408.

Hohenlohe PA, Catchen J, Cresko WA. 2012b. Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. In: Pompanon F, Bonin A eds. *Data production and analysis in population genomics*. New York: Humana Press. 235–260.

Hu HH, Chaney RW. 1940. Miocene flora from Shantung province, China. Part I. *Carnegie Institution of Washington. Contributions to Paleontology Publication* 507: 1–147.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.

Li J, Bogle AL. 2001. A new suprageneric classification system of the Hamamelidoideae based on morphology and sequences of nuclear and chloroplast DNA. *Harvard Papers in Botany* 5: 499–515.

Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451–1452.

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Geneticx* 9: e1003215.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* 66: 526–538.

Meyer FG. 1997. *Fothergilla*. In: Flora of North America Editorial Committee eds. *Flora of North America North of Mexico*. New York: Oxford University Press. 3: 365–366.

Meyer HW, Manchester SR. 1997. *The Oligocene Bridge Creek Flora of the John Day Formation*. Oregon: University of California Press.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans (USA)*, November 2010. 1–8. doi: 10.1109/GCE2010.5676129

Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, Baxter SW, Blaxter ML, Mallet J, Jiggins CD. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology* 22: 814–826.

Pante E, Abdelkrim J, Viricel A, Gey D, France S, Boisselier M-C, Samadi S. 2015. Use of RAD sequencing for delimiting species. *Heredity* 114: 450–459.

Paun O, Stuessy TF, Hörandl E. 2006. The role of hybridization, polyploidization and glaciation in the origin and evolution of the apomictic *Ranunculus cassubicus* complex. *New Phytologist* 171: 223–236.

Peirson JA, Dick CW, Reznicek AA. 2013. Phylogeography and polyploid evolution of North American goldenrods (*Solidago* subsect. *Humiles*, Asteraceae). *Journal of Biogeography* 40: 1887–1898.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RAD-seq: An inexpensive method for de novo SNP

discovery and genotyping in model and non-model species. *PLoS ONE* 7: e37135.

Pfender W, Saha M, Johnson E, Slabaugh M. 2011. Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theoretical and Applied Genetics* 122: 1467–1480.

Posada D. 2008. jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* 25: 1253–1256.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Radtke MG, Pigg KB, Wehr WC. 2005. Fossil *Corylopsis* and *Fothergilla* leaves (Hamamelidaceae) from the Lower Eocene flora of Republic, Washington, USA, and their evolutionary and biogeographic significance. *International Journal of Plant Sciences* 166: 347–356.

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* 197: 573–589.

Ranney TG, Lynch NP, Fantz PR, Cappiello P. 2007. Clarifying taxonomy and nomenclature of *Fothergilla* (Hamamelidaceae) cultivars and hybrids. *HortScience* 42: 470–473.

Ranney TG, Miller R, Lewandowski R, Xiang J. 2012. Discovery of a new diploid cytotype of *Fothergilla*. *HortScience* 47: S367.

Recknagel H, Elmer KR, Meyer A. 2013. A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3: Genes, Genomes, Genetics* 3: 65–74.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61: 539–542.

Rowe H, Renaut S, Guggisberg A. 2011. RAD in the realm of next-generation sequencing technologies. *Molecular Ecology* 20: 3499–3502.

Rubin BE, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS ONE* 7: e33394.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336–348.

Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. *Annual Review of Plant Biology* 60: 561–588.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the raxml web servers. *Systematic Biology* 57: 758–771.

Stebbins GL. 1985. Polyploidy, hybridization, and the invasion of new habitats. *Annals of the Missouri Botanical Garden* 72: 824–832.

Suzuki K. 1961. The important and characteristic Pliocene and Miocene species of plants from the southern part of the Tohoku district, Japan. *Science Reports of the Faculty of Arts and Science, Fukushima University* 10: 1–95.

Swofford DL. 2003. *PAUP*: Phylogenetic analysis using parsimony (and other methods), Version 4*. Sunderland: Sinauer Associates.

Vorsa N, Jelenkovic G, Draper A, Welker W. 1987. Fertility of 4*x* x 5*x* and 5*x* x 4*x* progenies derived from *Vaccinium ashei/corymbosum* pentaploid hybrids. *Journal of the American Society for Horticultural Science (USA)* 112: 993–997.

Weakley A. 2008. *Flora of the Carolinas, Virginia, and Georgia, and surrounding areas*. Chapel Hill: UNC Herbarium, North Carolina Botanical Garden, University of North Carolina at Chapel Hill.

Weaver RE Jr. 1969. Studies in the North American genus *Fothergilla* (Hamamelidaceae). *Journal of the Arnold Arboretum* 50: 599–619.

Xie L, Yi T-S, Li R, Li D-Z, Wen J. 2010. Evolution and biogeographic diversification of the witch-hazel genus (*Hamamelis* L., Hamamelidaceae) in the Northern Hemisphere. *Molecular Phylogenetics and Evolution* 56: 675–689.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12176/suppinfo:

Fig. S1. Relationship of ploidy levels and tag/haplotype counts from STACKS.

Fig. S2. P Phylogenetic tree resulting from ML analysis of *matK-trnL-F-rps16* sequences of plastid genome with RAxML 7.2.8. Numbers on branches are values of bootstrap support. Blue branches represent 2x samples, purple branches represent 4x, and yellow branches represent 6x samples.

Fig. S3. Phylogenetic tree resulting from analysis of nuclear ETS sequences using ML method on RAxML 7.2.8. Numbers on branches are values of bootstrap support. Blue branches represent 2x samples, purple branches represent 4x, and yellow branches represent 6x samples.

Fig. S4. Phylogenetic tree resulting from analysis of combined ETS- *matK-trnL-F-rps16* from Sanger sequencing using RAxML. Bootstrap support $\geq$50 are shown on branches. Blue branches represent 2x samples, purple branches represent 4x, and yellow branches represent 6x samples.

Fig. S5. Phylogenetic trees resulting from analyses of STACKS haplotypes from Run 2 using Neighbor-Joining method. Numbers on branches are values of bootstrap support (36 samples; three with extensive missing data removed).

Fig. S6. Plot of Delta K and K values from STRUCTURE runs of genotype matrix generated by STACKS pipeline using Run 2 GBS data. K represents genetic clusters.

Table S1. Distribution of loci abundance among accessions/individuals called from STACKS.

Table S2. Frequency of number of haplotypes per locus in each sample of Run 1 called from STACKS.

Table S3. Number of putative "triploid" loci in each individual sample in the STACKS haplotype dataset from Run 2.

Table S4. Haplotype loci from Run 2 called from STACKS that have three haplotypes in the diploid set of samples.

Table S5. Frequency of haplotypes called from STACKS detected in both sequence runs of the nine libraries prepared for Run 1.

Table S6. GenBank accession numbers of sequences obtained from Sanger sequencing.