



UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI BIOSCIENZE

**PhD Course in Molecular and Cellular Biology
XXXV Cycle**

Insights into the Genetic Diversity of *Asimina triloba*:

A Study using Genome Assembly and Population

Genetics

James Friel

R12502

Scientific Supervisor:

Prof. Martin Kater

Co-Supervisor:

Prof. Aureliano Bombarely

Table of Contents

TABLE OF CONTENTS	1
TABLE OF FIGURES	5
TABLE OF SUPPLEMENTAL FIGURES	6
LIST OF TABLES	6
LIST OF SUPPLEMENTAL TABLES	6
LIST OF ABBREVIATIONS	7
ABSTRACT	9
PROJECT AIMS	10
ACKNOWLEDGEMENTS	12
CHAPTER 1 : INTRODUCTION	13
INTRODUCTION TO ASIMINA TRILOBA	13
<i>Species background</i>	13
<i>Flowering, reproduction, and fruit development</i>	14
<i>Commercial development</i>	16
<i>Medicinal and Pesticidal Uses</i>	20
<i>Genetic Research</i>	21
<i>Species Distribution</i>	25
GENOME SEQUENCING OF NON-MODEL SPECIES	29
<i>A brief history of sequencing</i>	29
<i>Next-Generation Sequencing</i>	33
<i>Third generation sequencing / long read sequencing</i>	34
ASSEMBLY, AND ANNOTATION IN NON-MODEL SPECIES	38
<i>Draft genome assembly</i>	38

<i>Genome Size</i>	40
<i>Genome assembly</i>	40
<i>Gap Filling and Scaffolding</i>	41
<i>Quality assessment</i>	42
<i>Repetitive element problems in the assembly</i>	43
<i>Annotation</i>	45
REDUCE REPRESENTATION APPROACHES TO POPULATION STRUCTURE ANALYSIS	47
CHAPTER 2 : DRAFT ASSEMBLY AND VIRGINIA STATE POPULATION	50
(MANUSCRIPT SUBMISSION).....	50
ABSTRACT	50
INTRODUCTION	51
MATERIALS AND METHODS.....	54
<i>Reference sampling and sequencing</i>	54
<i>Reference genome assembly</i>	54
<i>Virginia population sampling & GBS library construction</i>	56
<i>Read processing, mapping, filtering and variant calling</i>	62
<i>Clonal correction and population structure inference</i>	63
<i>Analysis of genetic variance</i>	64
RESULTS	65
<i>Pawpaw draft genome assembly and SNP calling</i>	65
<i>Presence of clones in the dataset</i>	69
<i>Population structure</i>	69
<i>Genetic diversity</i>	76
DISCUSSION.....	80
<i>Inference of population structure</i>	81
<i>River populations</i>	83
<i>Genetic diversity</i>	86
CONCLUSION	88

SUPPLEMENTAL MATERIAL.....	90
CHAPTER 3 : DE NOVO ASSEMBLY IN A NON-MODEL SPECIES.....	105
INTRODUCTION.....	105
<i>Genome sequencing and assembly in non-model species.....</i>	<i>105</i>
<i>Genome annotation in non-model species.....</i>	<i>107</i>
MATERIALS AND METHODS.....	109
<i>Sample material.....</i>	<i>109</i>
<i>HiFi Sequencing and assembly.....</i>	<i>109</i>
<i>Coverage.....</i>	<i>110</i>
<i>Genome assembly evaluation.....</i>	<i>110</i>
<i>Hi-C sequencing and scaffolding.....</i>	<i>111</i>
<i>RNA-sequencing.....</i>	<i>111</i>
<i>Repeat Masking.....</i>	<i>112</i>
<i>Gene prediction.....</i>	<i>113</i>
<i>Annotation quality assessment and gene clustering.....</i>	<i>114</i>
RESULTS AND DISCUSSION.....	115
<i>Sequencing coverage.....</i>	<i>115</i>
<i>CANU assembly.....</i>	<i>115</i>
<i>CANU haplotig purge.....</i>	<i>118</i>
<i>LJA assembly.....</i>	<i>118</i>
<i>Hifi-asm assembly.....</i>	<i>119</i>
<i>HiFi-asm and Hi-C.....</i>	<i>121</i>
ANNOTATION.....	125
<i>Repeat sequences.....</i>	<i>125</i>
<i>Gene model annotation.....</i>	<i>125</i>
<i>Gene cluster orthologues.....</i>	<i>130</i>
CONCLUSION.....	132
CHAPTER 4 : GENETIC DIVERSITY STUDY OF A. TRILOBA ACROSS ITS NATIVE RANGE.	133

INTRODUCTION	133
MATERIALS AND METHODS.....	134
<i>Sample material</i>	134
<i>GBS library construction</i>	135
<i>Raw read processing, mapping and variant calling</i>	136
<i>SNP filtering</i>	137
<i>Inference of population structure</i>	137
<i>Analysis of molecular variance (AMOVA)</i>	138
<i>Isolation-by-distance (IBD)</i>	139
<i>F-statistics and heterozygosity</i>	139
RESULTS AND DISCUSSION.....	140
<i>Sequencing and variant calling</i>	140
<i>SNP filtering</i>	142
<i>Inference of Population Structure</i>	159
<i>Genetic diversity</i>	170
CONCLUSIONS.....	175
SUPPLEMENTAL FIGURES.....	176
CHAPTER 5 : FINAL CONCLUSIONS.....	179
CONCLUSIONS.....	179
THE PAWPAW NETWORK	181
FINAL THOUGHTS AND FUTURE PERSPECTIVES	186
APPENDIX	188
BIBLIOGRAPHY.....	189

Table of Figures

FIGURE 1-1 <i>ASIMINA TRILOBA</i> SPECIES MORPHOLOGY	19
FIGURE 1-2 HISTORY OF GENOMIC SEQUENCING	32
FIGURE 1-3 PACBIO HIFI READS	37
FIGURE 2-1 VIRGINIA POPULATION STRUCTURE.....	72
FIGURE 2-2 JAMES RIVER BASIN POPULATION	75
FIGURE 2-3 PAWPAW NEIGHBOR-JOINING TREE	77
FIGURE 2-4 ISOLATION-BY-DISTANCE AND AMOVA RESULTS.....	78
FIGURE 3-1 MERQURY SPECTRA PLOTS	120
FIGURE 3-2 HI-C CONTACT HEATMAP	123
FIGURE 3-3 BLOB PLOT	124
FIGURE 3-4 ASSEMBLY REPEAT CONTENT	127
FIGURE 3-5 GENE MODEL AED COMPARISON	128
FIGURE 3-6 ORTHOLOGOUS CLUSTER ANALYSIS.....	131
FIGURE 4-1 MISSING DATA ASSESSMENT	145
FIGURE 4-2 POPULATION ADMIXTURE	160
FIGURE 4-3 POPULATION CLUSTERING AND MEMBERSHIP	161
FIGURE 4-4 NEIGHBOR-JOINING TREE	165
FIGURE 4-5 IBD AND AMOVA	169
FIGURE 4-6 DIVERSITY AMONG CLUSTERS	174
FIGURE 5-1 SAMPLING INSTRUCTIONS.....	182
FIGURE 5-2 PAWPAW NETWORK	183
FIGURE 5-3 PAWPAW T-SHIRT DESIGN.....	185

Table of Supplemental Figures

SUPPLEMENTAL FIGURE 2-1 MLG HEATMAP	90
SUPPLEMENTAL FIGURE 2-2 MERCURY SPECTRA PLOT	91
SUPPLEMENTAL FIGURE 2-3 STRUCTURE INFERENCE AND ADMIXTURE.....	92
SUPPLEMENTAL FIGURE 2-4 ADMIXTURE	93
SUPPLEMENTAL FIGURE 2-5 MINIMUM SPANNING NETWORK	94
SUPPLEMENTAL FIGURE 2-6 DAPC'S POSTERIOR MEMBERSHIP PROBABILITY	95
SUPPLEMENTAL FIGURE 4-1 RELATEDNESS HEATMAP.....	176
SUPPLEMENTAL FIGURE 4-2 DAPC GROUP OPTIMIZATION	177
SUPPLEMENTAL FIGURE 4-3 F_{ST} PAIRWISE COMPARISON BY STATE	178

List of Tables

TABLE 2-1 A. <i>TRILOBA</i> SAMPLE LOCATIONS	58
TABLE 2-2 DRAFT ASSEMBLY ASSESSMENT	67
TABLE 2-3 GENETIC DIVERSITY.....	79
TABLE 3-1 SEQUENCING COVERAGE.....	115
TABLE 3-2 DRAFT ASSEMBLY COMPARISON	117
TABLE 3-3 EVIDENCE CONTRIBUTION TO GENE PREDICTION	129
TABLE 4-1 SAMPLING LOCATIONS BEFORE AND AFTER QC.....	146
TABLE 4-2 GBS READ PROCESSING.....	147
TABLE 4-1 AMOVA RESULTS.....	168
TABLE 4-2 GENETIC DIVERSITY.....	170

List of Supplemental Tables

SUPPLEMENTAL TABLE 2-1 TABLE OF PCR PRIMERS AND ADAPTERS SEQUENCES	96
--	----

SUPPLEMENTAL TABLE 2-2 GBS READ PROCESSING	97
SUPPLEMENTAL TABLE 2-3 MLG OCCURRENCE AT SAMPLING SITES	103
SUPPLEMENTAL TABLE 2-4 AMOVA RESULTS.....	104

List of Abbreviations

Abbreviation	Definition
AED	Annotation edit distance
AFLP	Amplified fragment length polymorphism
AM	Appalachian mountain
ANGSD	Analysis of next generation sequencing data
BIC	Bayesian Information Criterion
BTI	Boyce Thomson Institutes
BUSCO	Benchmarking universal single-copy orthologs
BWA	Burrows-Wheeler alignment tool
CCS	Circular consensus sequencing
CDS	Protein-coding sequence
CLR	Continuous long read
DAPC	Discriminant analysis of principal components
ddNTP	2,3-dideoxynucleoside triphosphates
dNTP	Deoxynucleoside triphosphates
ESTs	Expressed sequence tags
F_{IS}	Inbreeding coefficients
F_{ST}	The fixation index
GBS	Genome-by-Sequencing
He	Expected heterozygosity
Hi-C	Chromatin-association/interaction analysis
HiFi	High-fidelity
HighBT	High bridge trail state park
Ho	Observed heterozygosity
HRB	Holston river basin
HWE	Hardy–Weinberg equilibrium
IBD	Isolation-by-distance
indels	Small insertions or deletions
ISSR	Intersimple sequence repeat marker
JRB	James river basin
KSU	Kentucky State University
LAI	LTR assembly index
LD	Linkage disequilibrium
LTR	Long terminal repeats
MLG	Multilocus genotypes
MSN	Minimum spanning network
NJ	Neighbor-joining

NRB	New river basin
OLC	Overlap-layout-consensus
ONT	Oxford nanopore technologies
PCA	Principal component analysis
QC	Quality controls
QV	Consensus quality
RADseq	Restriction site-associated DNA sequencing
RAPD	Random amplified polymorphic DNA
SMRT	Single-molecule real-time
SNAP	Semi-HMM-based nucleic acid parser
SNP	Single nucleotide polymorphism
SSR	Simple-sequence repeats
TGS	Third-generation sequencing
UPGMA	Unweighted pair-group mean clustering analysis
VCF	Variant file
VT	Virginia Tech
YRB	York river basin
ZMW	Zero-mode waveguide

Abstract

The objective of this PhD thesis was to understand the factors influencing the evolutionary history and distribution of genetic variation in the species *Asimina triloba*, commonly known as pawpaw. To achieve this, we have produced a high-quality reference genome by assembling a draft genome using PacBio's Sequel II long reads and polishing with Illumina short reads. We first used a genotype-by-sequencing (GBS) to genotype 124 individuals from 28 sites across the state of Virginia to produce a set of single nucleotide polymorphisms (SNPs). We then analysed the population structure and genetic diversity, revealing evidence of isolation by distance and an increase in geneflow over long distance by individuals along rivers.

Following on from this, an improved genome assembly was produced by incorporating HiFi, Hi-C, and RNA-seq data, resulting in an annotated, highly contiguous, and accurate genome assembly. The improved draft assembly provided a more comprehensive view of the genome structure and organization of *A. triloba*, allowing for the construction of pseudochromosomes and genome annotation, including the identification of repeat and transposable elements. An additional population genetics study was then carried out using the same GBS approach to analyse over 300 wild individuals collected from across North America. Our analysis provided insight into the overall diversity of the species and indicated that *A. triloba* may have experienced a bottleneck. This may have resulted from a reduced population size in its Pliocene refugia prior to the species' northern migration after the retreat of the Laurentide Ice Sheet. This research has important implications for understanding the mechanisms underlying genetic diversity in the species and conservation efforts.

Project Aims

The objective of this project was to study the genetic diversity and understand the factors influencing the evolutionary history and the distribution of genetic variation of the species *A. triloba*.

- **Aim 1:** Develop a high-quality reference genome for *A. triloba*.
- **Aim 2:** Characterization of the structure and diversity of *A. triloba* populations and their association to geographical factors.
 - Collect samples from across the native range of the species.
 - Prepare Genotype-by-Sequencing (GBS) libraries from collected wild samples and sequence using Illumina short read sequencing.
 - Use reference genome to call genetic variants and conduct population genetics studies to understand the diversity and factors influencing the structure, and evolution of the species.

The first aim of this project was to generate a high-quality reference genome to serve as a foundational resource for the rest of the study. The first reference genome was generated using a combination of long-read and short-read sequencing technologies. We compared the assembly performance of multiple assembler tools and pipelines before generating the draft assembly Astri041. Having a reference genome available is an invaluable tool as it significantly improves the number of variants, and thus the resolution available to study the genetic diversity of *A. triloba*. Later, we produced an improved assembly using a combination of highly accurate long-read HiFi sequencing with Hi-C for 3D conformation of chromatin to infer the genomic structure. In addition, we were able to annotate the reference genome using RNA-sequencing (RNA-seq) of leaf, fruit, and pericarp tissue. Gene prediction tools were trained using the RNA-seq data and protein sequences from closely related species.

The second aim of this project was to characterize the structure of *Asimina triloba* populations and their association to geographical and/or anthropological factors. To achieve this aim,

samples were collected from across the native range of the species. These samples were then prepared for GBS using Illumina short read sequencing. GBS is a powerful method for genotyping large numbers of samples at a relatively low cost. The resulting sequence data was used to call genetic variants, in this case single nucleotide polymorphisms (SNPs) These SNPs were used to conduct population genetics studies in an effort to understand the factors influencing the genetic diversity, the population structure, and the evolutionary history of the species. These studies included population genetic analyses such as Wrights F-statistics, PCA, STRUCTURE, ADMIXTURE, and other relevant population genetic analyses that provided a comprehensive understanding of the genetic diversity and structure of the *A. triloba* populations. These analyses were also used to test for association between genetic variation and geographic factors, such as potential physical barriers to, or corridors of increased geneflow.

Acknowledgements

I would like to express my deepest gratitude to Aureliano Bombarely and Siliva Manrique Uprí, this work would not have been possible without them. Their daily guidance, support, and encouragement over five years, two continents, and many countries, has sustained me from the beginning of my master's degree right through to the end of my Ph.D. journey. Their expertise and insights have been invaluable, but even more, I am extremely grateful for their friendship. I would also like to thank my supervisor Martin Kater for his support throughout the PhD, for giving me the opportunity to take part in this project and providing me with everything I needed to carry out this work.

To the members of my PhD committee, David Haak, Simona Masiero, and Iñaki Hormaza Urroz for their incredibly valuable time, feedback, and suggestions throughout my research. Over the past three years I have been very lucky to spend time working with people in the University of Milan, Virginia Tech (VT), IBMCP-CSIC Valencia, and BTI Cornell in New York. To all the colleagues and friends in these labs I would like to extend my sincere appreciation for their camaraderie, support, and helpful discussions. In particular, I want to express my gratitude to Ariel Heminger at VT for all the support preparing samples and to Suzy Strickler and her team at BTI for helping with the pawpaw genome and population analysis.

To everyone that ventured into forests, up mountains, and across rivers to collect pawpaw samples, without your help this project would never have succeeded.

Thank you all for your support, encouragement, and assistance. This thesis would not have been possible without your help.

Finally, I would like to acknowledge the funding agencies, university of Milan, Suzy Strickler, and Evofrulan h2020-msca-rise for their financial support of this research.

Chapter 1 : Introduction

Introduction to Asimina triloba

Species background

Asimina triloba [L.] Dunal (Pawpaw) is a member of the Annonaceae family (the Soursop family), the most diverse family in the early-divergent Magnoliid clade. The family is comprised of around 130 genera and more than 2500 species distributed globally in tropical to subtropical regions (Angiosperm Phylogeny Group, 2016; Erkens et al., 2022). A single exception to the tropical distribution is the *Asimina* genus which is by-far the most northerly representative of the family (Berry, 1916; Erkens et al., 2022). The *Asimina* genus is unique in the Soursop family as it is native to temperate regions of North America. The genus contains eight to ten species, and several possible hybrids, most of which are restricted to the state of Florida (Horn, 2015; Kral, 1960; Zimmerman, 1941). Species include *A. incana* (Woolly pawpaw), *A. longifolia* (Polecat-bush), *A. obovata*, *A. parviflora* (Small-flowered pawpaw), *A. pygmaea* (Dwarf pawpaw), *A. reticulata* (Netted pawpaw), *A. tetramera* (Four petal pawpaw), and *A. ×nashii* (Kral, 1960). All species are diploid, though some triploids have been observed in *A. triloba* (Bowden, 1949). The number of haploid chromosomes in *A. triloba* is not entirely clear. While $x = 7, 8, \text{ or } 9$ has been observed for many Annonaceae members (Okada and Ueda, 1984), early studies in pawpaw reported a haploid chromosomes number of 9, $2n = 2x = 18$ (Bowden, 1949, 1940; Kral, 1960; Locke, 1936). However, $x = 8$ has also been reported in members of the *Asimina* genus (Tanaka and Okada, 1972) including *A. triloba* (Ito and Mutsuura, 1956).

Pawpaw is a small broad-leaved deciduous tree, reaching a height of around 10 meters (Layne, 1996) [Figure 1-1(b and d)]. It is typically found in the understory of hardwood forests near

lakes and rivers where it grows in slightly acidic, deep, and well-drained alluvial soils, but also survives well in drier soils (Callway, 1992; Kral, 1960; Lagrange and Tramer, 1985). *A. triloba* is the most widespread of the Asimina genus; found in 26 states covering the entire eastern coast of the United States of America. It can be found in parts of Florida, as far west as Texas, and its range extends as far north as southern Ontario, Canada (Darrow, 1975; Fox, 2012), making it is the most northerly representative member of the Soursop family. Indeed, it is the only species in the entire Annonaceae family that is known to survive annual minimum temperatures as low as -28.8 °C (Kral, 1960).

Flowering, reproduction, and fruit development

In early spring, pawpaw trees produce dark brown flower buds on the previous year's woody growth (Ferrer-Blanco et al., 2022; Kral, 1960). The flowers are around 3-5 cm in diameter, emerging in April-May (Kral, 1960; Lagrange and Tramer, 1985; Layne, 1996). They have three sepals with an outer and inner set of three-lobed, maroon-coloured petals in conical arrangement typical in many other members of the Annonaceae [Figure 1-1c]. Pawpaw flowers are monoclinous with a globular androecium and a gynoecium comprised of 3 to 10 plicate unilocular carpels (Kral, 1960; Lampton, 1957; Losada et al., 2017; Willson and Schemske, 1980). The flowers have a strong protogynous dichogamy development where the stigma matures before anther dehiscence (Losada et al., 2017; Pomper and Layne, 2003), a feature common in other Annonaceae species (Gottsberger, 1999). It is often the case in plants that when stigmas are receptive before pollen is released from the anthers, self-pollination does not happen. Indeed, most cultivated and wild pawpaw trees are considered to be self-incompatible (Pomper and Layne, 2003; Willson and Schemske, 1980). Though there appears to be two barriers to self-compatibility in *A. triloba*, the temporal separation of gametes and a second unknown mechanism. This separation of gamete maturation timing makes pawpaw an obligate

out-crossing species, and thus, dependent on its pollinators for sexual reproduction and genetic recombination. Pollinator species for pawpaw flowers are thought to be carrion flies (Willson and Schemske, 1980) and beetles due to the dark and flesh like colour of the petals, combined with a yeasty or fetid smell. Such pollinators are typical of this flower phenotype, and are common pollinators of Annonaceae species (Goodrich et al., 2006; Gottsberger, 1999). Interestingly, in *A. triloba* the flowering cycle is longer than other family members, with the stigmatic receptivity lasting 15 days, something distinct even in other protogynous flowers of the magnoliid clade (Losada et al., 2017).

An important feature of the pawpaw tree is its ability to reproduce asexually; readily forming clonal patches (genets) by producing adventitious shoots, or “root suckers” from adventitious buds on its roots (Kral, 1960; Pomper et al., 2009). These tend to emerge close to the original stem (Botkins et al., 2012; Kral, 1960; Pomper et al., 2009), resulting in the species often being found in large genet clusters that can be comprised of a single or multiple genotypes (Botkins et al., 2012; Pomper et al., 2009; Willson and Schemske, 1980). In other clonally reproducing species a single genotype may live impressively long time, such as the quaking aspen (*Populus tremuloides*), thought to be around eleven thousand years old, while individual members (ramets or stems) may be shorted-lived (Barlow, 2001). While there are cultivated pawpaw trees over 100 years old, most wild trees are thought to only live around 25 years, no investigation has yet looked into the age of clonal pawpaw patches. Clonality in understory species is not uncommon, where it is likely a benefit in the low light conditions that can hinder the growth from seed (Groenendael et al., 1996). It is much easier to provide a supply of nutrient from an already established tree. This likely plays a large role in helping pawpaw maintain its long-term geographical presence on the edges of its suitable range.

While fruit production is reported to be very high in cultivated pawpaw trees (Layne, 1996), in the wild, the fruit set success varies by location with rates decreasing significantly in the most

northern parts of its range (Lagrange and Tramer, 1985). Willson and Schemske (1980), showed that in Illinois, the upper limits of the species' range, vast number of flowers were being produced in wild tress, but successful fruit set only occurred in approximately 0.4% of the flowers. Hand pollination increased the rate to 17%, indicating that an absence of pollinators in the colder northern climate may be an important factor contributing to the lower fruit production seen in the colder edges of the species' range.

However, when *A. triloba* has been successfully pollinated, it produces the largest fruit in North America. Pawpaw fruit are oblong-cylindric berries 3 to 15 cm long, developing as either single fruit or clusters, similar to bananas, and may contain 8 or even a many as 20 3 cm long dark flattened crustaceous seeds per fruit (Lagrange and Tramer, 1985; Layne, 1996) [Figure 1-1a]. Although fruit mass is typically in the range of 200–500 g (Layne, 1996), it is capable of reaching up to 1 kg (Darrow, 1975). The pawpaw fruit is very aromatic with ripe fruit having a creamy texture and a flavour combination that is a mixture of banana (*Musa×paradisiaca*), mango (*Mangifera indica*) (Layne, 1996), cherimoya (*Annona cherimola*) (Duffrin and Pomper, 2006) and pineapple (*Ananas comosus*) (McGrath and Karahadian, 1994). The fruit is highly nutritious, comparable to, or exceeding bananas in levels of vitamins and minerals per gram. When compared to the apples (*Malus domestica*) and oranges (*Citrus × sinensis*) its content is much higher in many minerals and essential amino acids (Duffrin and Pomper, 2006).

Commercial development

Within the Annonaceae family there are several agriculturally important species such as the cherimoya, sugar apple (*Annona squamosa*), (*Annona muricata* L.), and custard apple (*Annona reticulata*) (Hormaza, 2014; Losada et al., 2017). Significant commercial expansion of pawpaw as crop is still in its early stages (Callway, 1992; Layne, 1996; Peterson, 2003; Pomper

and Layne, 2003). When pawpaw fruit is ripe, the skin turns from green to brownish black while the flesh softens and turns a light cream or orange colour. At this point there are only 2–3 days until the fruit is over ripe. Over ripe pawpaw fruit are known to smell sickly sweet and unfavourable to customers (Pomper and Layne, 2003). However, the fruit is climacteric, meaning it continues to ripen away from the tree and it can be shipped to market locations prior to ripening. In addition, with refrigeration the fruit can last three weeks to a month at 4 °C (Ferrer-Blanco et al., 2022; Layne, 1996), with the ripening process continuing when returned to ambient temperature (Duffrin and Pomper, 2006). Brannan et al, (2015) analysed the phytochemical content of pawpaw pulp from ten varieties using a mass spectral characterization of phenolic acids and flavonoids. They reported that the predominant polyphenolic compounds found in both ripe and unripe pulp were three phenolic acids, protocatechuic acid hexoside, p-coumaroyl hexoside, and 5-Op-coumaroylquinic acid, and flavonols, particularly (–)-epicatechin, B-type procyanidin dimers and trimers. The authors considered the fruit to be a good source of phenolic acids, flavonoids, as well as procyanidins (condensed tannins which are potent antioxidants), and suitable for further processing to add to nutritional or flavour values to other food products. Indeed, there has been progress developing a market based on the pawpaw’s distinct flavour, with ice-cream, wines, beers, jams, and various pulp incorporated baked goods (Duffrin and Pomper, 2006; Layne, 1996; Pomper and Layne, 2003).

In 1994 Kentucky State University (KSU) was designated as a satellite repository of *A. triloba* for the U.S. Department of Agriculture (USDA), National Plant Germplasm System (NPGS). KSU now serves as a hub for development of pawpaw cultivars. However, interest in establishing *A. triloba* as a crop has extended beyond the US with research ongoing in several countries including Austria (Lehner et al., 2022), Italy (Bellini et al., 2003; Lolletti et al., 2021),

Romania (Tabacu et al., 2020), Belgium, Japan, Israel, Korea, Ukraine (Hrabovetska et al., 2006), Portugal, and China (Xinkun et al., 2021), all primarily focusing on cultivation and nutritional content. Despite market development efforts and the large fruit size, palatable flavour, and nutritional content, pawpaw has not yet seen the same successful market development as other Annonaceae members, or even as other wild fruit native of North America. This is in stark contrast to the global success of the blueberry (*Vaccinium sect. Cyanococcus*), whose domestication and market development began at the same time as pawpaw over 100 years ago (Peterson, 2003). While there are some established pawpaw farms, the fruit is still mainly sold at local markets and restaurants (Moore, 2015; Pomper and Layne, 2003). Part of the reason for this is the short shelf life and ease at which the fruit is bruised in transport; currently it seems unlikely that we will see pawpaw fruit as regular commodity in much of the US or the rest of the world unless and until post-harvest issues are managed. There is also a need for elite varieties with less seeds and consistent fruit size, market development, and improvements in the cultivation (Pomper and Layne, 2003).



Figure 1-1 *Asimina triloba* species morphology

Panel showing morphological features of *A. triloba* (pawpaw). **a**). Cluster of ripe pawpaw fruit. **b**). Mature pawpaw tree from BTI experimental orchard (approx. 6 meters tall) **c**). Pawpaw flower **d**) Pawpaw broad obovate leaf shape (a and c image sourced from Prof. Bombarely).

Medicinal and Pesticidal Uses

In addition to its nutritional value and potential as a crop, pawpaw twigs, bark, leaves and fruits contain several interesting Annonaceous acetogenins with medicinal and pesticidal applications (Johnson et al., 1996; McLaughlin, 2008; Ratnayake et al., 1992; Zhao et al., 1994). Annonaceous acetogenins are potent inhibitors of mitochondrial (complex I), as well as cytoplasmic (anaerobic) production of adenosine triphosphate (ATP) and the related nucleotides (McLaughlin, 2008). Indeed, many of the compounds have been shown to be potent cytotoxins with antitumor, pesticidal, antimalarial, anthelmintic, piscicide, antiviral and antimicrobial activity (McLaughlin, 2008), and anti-inflammatory activity (Nam et al., 2021). In particular, 3 compounds, bullatacin, bulletin, and bullanin have been shown *in vitro* to be highly effective against tumour cell lines (Zhao et al., 1994). Further, an *in vitro* analysis of the cell cycle phase distribution and expression of the apoptosis regulatory proteins BCL-2, BAX, caspase-3, and PARP showed anti-inflammatory effects of extracts from twigs, roots, and unripe fruits in arresting cell cycle and apoptosis of AGS and HeLa cells (Nam et al., 2021). This demonstrated the potential application of pawpaw phenols in the treatment of certain gastric and cervical cancers. The same group, Nam et al, (2019) also identified 17 phenolic components from pawpaw fruit that showed strong antioxidant and antimicrobial activities. A 95% ethanol extract of the ripe fruit inhibited effect against various microorganisms, in particular two species of bacteria known to cause health issues in the humans; *Corynebacterium xerosis* and *Clostridium perfringens* (Nam et al., 2019). Acetogenins extracts from pawpaw have been successfully applied to commercial products for the treatment of head lice, fleas, and ticks (McCage et al., 2002; McLaughlin, 2008) as well as ointments to treat oral herpes (HSV-1) and other skin conditions (McLaughlin, 2008).

Asimicin, another pawpaw phenol, has been shown to act as an effective pesticide on a number of pest species such as mosquito larvae (*Aedes aegypti* L.), blowfly larvae (*Colliphora vicina* Meig), two-spotted spider mite (*Tetranychus urticae* Koch), striped cucumber beetle (*Acalymma vittatum* F.), melon aphid (*Aphis gossyphii* Glover), Mexican bean beetle (*Epilachna varivestis* Mulsant), and a free-living nematode [*Caenorhabditis elegans* (Maupas) Dougherty] (Alkofahi et al., 1989). The production of pesticides derived from pawpaw have the additional benefit of being more environmentally friendly and biologically degradable. (Ratnayake et al. 1993).

Genetic Research

To date, there has been limited genetic study into *A. triloba*; focus has primarily been on the genetic diversity in cultivars and a small collection of individuals from wild populations curated at the KSU germplasm repository. One of the earliest investigations used a minisatellite M13 “fingerprinting” probe to evaluate genetic variation in wild pawpaw populations (Rogstad et al., 1991). Samples were collected from 16 sites across five states, with one to twenty-two individuals per site. The study described moderate to non-existing within-population variation and variation at the local and geographic scale varied but was low overall compared to other organisms like the quaking aspen. Clonal reproduction and inbreeding were suggested as the cause for such low to moderate variation in the species. However, in-breeding within pawpaw is considered to be rare (Willson and Schemske, 1980).

Genetic variation was next evaluated in 32 clones of cultivars and elite selected lines of pawpaw using 23 isozymes (Huang et al., 1997). Seven isozymes were polymorphic and nine polymorphic loci were identified. Nine of these loci and the 32 clones were used to generate 28 multi-locus isozymes which could be used to identify 24 varieties. Genetic differentiation was estimated to be within the average of similar long-lived woody perennials and higher than previously suggested in the previous paper (Rogstad et al., 1991). A subsequent study used

isozymes to examine genetic diversity within a collection of wild pawpaw trees from KSU (Huang et al., 1998). The collection comprised nine states from across the US and included 25 to 50 trees from each population. Genetic diversity was high among the sampled populations with the observed heterozygosity being higher than expected under a Hardy-Weinberg equilibrium (HWE). Partitioning with 17 polymorphic loci showed that as much as 88.2% of the genetic diversity was observed within populations. An unweighted pair-group mean clustering analysis (UPGMA) of genetic distance among the nine populations separated the southern (Georgia) and western (Illinois) state populations from the others (i.e., Kentucky, West Virginia, Indiana, Maryland, Pennsylvania, and New York). The authors analysis of principal components (PCA) revealed a similar clustering with the New York population separating along PC1.

Following this, random amplified polymorphic DNAs (RAPDs) were used to evaluate 19 polymorphic bands in an interspecific cross of PPF1-5 pawpaw [*A. triloba* (L.) Dunal.] x RET (*A. reticulata* Shuttlew.) displaying a Mendelian segregation (1:1 or 3:1) (Huang et al., 2000). These were then used to evaluate the genetic diversity in pawpaw populations from six states (Georgia, Illinois, Indiana, Maryland, New York, and West Virginia). The expected heterozygosity was $H_e = 0.25$ and the average genetic diversity within populations was $H_s = 0.26$. This accounted for 72% of the total genetic diversity, again showing the greatest diversity is within populations (Huang et al., 1998; Rogstad et al., 1991). Among populations, the genetic diversity was $D_{st} = 0.10$, for 28% of the total genetic diversity. A RAPD-based UPGMA dendrogram of Nei's genetic distances showed a slightly different clustering with Maryland, New York and Georgia in one clusters and Indiana, West Virginia, and Illinois in another. Subsequently, 71 pairs of RAPD markers were also used to "finger print" pawpaw cultivars and investigate their genetic diversity (Huang et al., 2003). Results of eight pairs of primers produced 14 polymorphic sites and Nei's genetic diversity analysis showed similar outcomes

to the previous RAPD results (Huang et al., 2000); expected heterozygosity of cultivars ($H_e = 0.28$) was similar to that of the wild populations sampled ($H_e = 0.25$). In addition to isozymes and RAPD markers, genetic diversity has been assessed by intersimple sequence repeat markers (ISSRs) (Pomper et al., 2003) using 10 ISSR markers from 19 pawpaw cultivars, with the authors stating that the percentage of polymorphic loci was $P = 80\%$. They showed a moderately high level of genetic diversity, $H_e = 0.36$ within the cultivars.

Amplified fragment length polymorphism (AFLP) markers have in the past been considered to be a reliable, high throughput, and cost-effective approach to genotyping (Hansen et al., 1999). In an attempt to provide a more detailed evaluation of the genetic diversity stored at the KSU pawpaw repository, the inheritance of the AFLP markers in interspecific crosses were determined and then used to construct the genetic linkage groups for pawpaw (Wang et al., 2005). Six linkage groups covering 206 centimorgans (cM) were identified in an interspecific cross of PPF1-5 pawpaw [*A triloba* (L.) Dunal.] x RET (*A. reticulata* Shuttlew.), the same cross used in the (2000) Huang et al, study. One hundred and thirty-four AFLP markers were used to evaluate the genetic diversity in eight wild populations, and in thirty-one cultivars and advanced selections. In the wild populations, the percentage of polymorphic loci was 79% and H_e was 0.245, in congruency with the previous studies. Additionally, most of the variation was again seen in within populations (81.3%). The authors were also able to show the effectiveness of AFLP markers by delineating between cultivars using only nine markers. The UPGMA dendrogram of Nei's genetic distances among the eight wild pawpaw populations based on dominant AFLP markers showed grouping with Indiana, New York, West Virginia, and Georgia in one group, and Maryland, Illinois, Virginia, and Pennsylvania in another, closely matching the group clustering observed using RAPD markers (Huang et al., 2000).

To estimate the prevalence of clonal reproduction in native pawpaw patches, leaves were collected from wild patches in three counties of Kentucky state (Pomper et al., 2009). ISSR-

PCR primers identified three polymorphic and six monomorphic markers in six of the patches. In three of the patches, no polymorphic sites were identified, indicating that the patch may be entirely composed of vegetatively reproduced individuals forming large genets. Another three patches showed the presence of polymorphic loci, indicating more than one genotype present at the site. Although very small in scale and using a limited number of markers, this study shows that during sampling of wild *A. triloba* populations it is important to consider sampling strategies and the effect of clonal representation in genetic diversity. Evaluation of genetic diversity typically relies on frequency of polymorphisms, and overrepresentation of clonal genotypes may skew results. This paper also showed that the prevalence of clonal reproduction in pawpaw is not well understood, a much broader study is needed to understand the occurrence, frequency and conditions for asexual reproduction in wild populations. In a following study, Botkins et al (2012), determined the genetic diversity and clonality displayed in seven pawpaw patches at several locations in Kentucky using DNA microsatellite markers, or simple sequence repeat (SSR). The goal of this work was not to estimate its prevalence but to determine if the clonality and genetic diversity of pawpaw patches had an impact on the ability of *A. triloba* to compete with local invasive species. Twenty-five trees from seven patches in the four different locations were analysed with four DNA microsatellite markers. Interestingly, no entirely clonal patches were found in their sample collection, and very few clonal individuals were found in any of the patches (Botkins et al., 2012). There was no significant difference in the presence or absence of invasive species in wild and control plots, stem density and shading appeared to be more important factors in competition with invasive species.

Pomper et al (2010) developed a set of SSR markers of finger printing of cultivars. Using their markers, they reported the genetic diversity as being low with observed heterozygosity $H_o = 0.68$ very close to the expected $H_e = 0.7$. In a subsequent study, the same SSR markers were

again used to study the genetic variation in old and new pawpaw cultivars (Lu et al., 2011). Eighteen polymorphic SSR loci were used to examine the level of unique genetic variation being used in the development of pawpaw cultivars. The authors were also able demonstrate an influence of SRR motif on allelic variation in pawpaw. Further, they showed that the greatest genetic diversity ($H_o = 0.69$) was in the older cultivars, something that might be expected as diversity is lost through selection and introgressions from a limited number of genotypes. However, KSU advanced selections were also shown to contain unique pawpaw germplasm that may be useful to enhance continued breeding. Many of these works have been carried out by a group of researchers based at, or collaborating with the KSU, with a focus on the genetic diversity of individuals in their collection. Regardless of the type of marker used, heterozygosity appears to only be slightly higher or close to HWE expectations. While the overall genetic diversity seems to vary from low to moderate. The collection at KSU has also been shown to provide an important genetic resource for further breeding and commercialization of *A. triloba* as a crop.

Species Distribution

The earliest evidence of *Asimina* in North America comes from fossils found in the Paleocene-Eocene Wilcox Group, an important geological group that comprises the Gulf of Mexico Basin, Mexico, and several states in south of the US such as Texas, Louisiana and Alabama (Berry, 1916). Fossilized foliage of a species called *A. eocenica* was found in Denver Basin of Colorado. A second species, *A. leiocarpa*, was recorded by a seed found in Mississippi, placed during the Denver formation. In total, there is 4–5 *Asimina* fossils, the oldest of which is from the early Eocene placing the genus in North America at least 52 million years ago (Mya). A more recent example was found much further north in the state New Jersey dating from the late Miocene, during the formation of the Bridgeton and in the interglacial beds of the Don Valley

in Ontario, and clearly resembled the *A. triloba* (Berry, 1916; Coleman, 1901). Those fossils found in the Don Valley are from interglacial warm beds (Coleman, 1901), a period dated to be around 0.125Mya (Karrow, 1990) meaning that pawpaw must have been in the area before the beginning of the Laurentide ice sheet 2.6 Mya.

It is thought that, until the Pleistocene (3-0.012 Mya), large mega fauna such as Mastodon or Megalonyx may have eaten the fruit of the pawpaw tree and could have been the primary method of seeds dispersal (Janzen and Martin, 1982). Since the extinction of these animals, pawpaw can be considered something of anachronism because it is not clear what animals have stepped into fulfil the role (Barlow, 2001). There are three viewpoints on the postglacial distribution of pawpaw, and each is thoroughly reviewed by Wykoff (2009). The perspectives essentially amount to; (1) the northward migration of pawpaw haven been entirely down to human activity (Keener and Kuhns, 1997), (2) it was absolutely not humans, migration north was by fruit floating along rivers and modern native fauna (Murphy, 2001), and (3), a complex evolutionary history that involved large mega fauna and followed later by human activities (Peterson, 1991).

Keener and Kuhns (1997) entirely dismiss the possibility of small mammals such as raccoons *Procyon lotor* (L.) Elliot, red foxes *Vulpes fulvus* (Desmarest), and opossums (*Didelphis virginiana*) transporting them any great distance, stating that the seeds are too large for these animals to swallow. The indigenous people of North America have a well-documented association with *A. triloba*. Indeed, from the year 1541 we have the written report of pawpaw from a member of the De Soto expedition through the South-Eastern United States in which a member of the expedition noted that the Native Americans were growing and eating pawpaw fruit in the Mississippi Valley region (Wykoff, 2009). The genus name, *Asimina*, is even derived from names used by indigenous people, "Assimin". Keener and Kuhns (1997) make reference to a number of ethnohistoric and archaeological evidences for pawpaw tree use in

prehistoric sites throughout the Southeast and in the Ohio Valley; and go on to suggest that the northernmost distribution of the pawpaw into Southern Ontario and Western New York, Ohio, and Michigan was attributable to Iroquois population movements. However, they were unable to find physical evidence of pawpaw seeds at Iroquoian archaeological sites but predicted that this would be the case in the future. To date, none have been found, nevertheless, in a recent paper from Wyatt et al (2021), the authors used SSR markers to demonstrate a close link between rare loci at Anthropogenic sites, once used by indigenous people and wild populations. They were able to identify rare alleles in pawpaw trees at historical indigenous people sites also those found in wild populations hundreds of kilometres away (mean = 723 km), a distance much greater than the distance travelled by any of the suggested mammalian dispersal agent. Murphy (2001), however, argued against Keener and Kuhns hypothesis, suggesting instead that the spread of pawpaw could have been driven mainly by other mammals that seem to be able to eat pawpaw fruits, including raccoons, squirrels, opossums, foxes, black bears (*Ursus americanus*), and white-tailed deer (*Odocoileus virginianus*). A concern here is the travel range of many of these species, for example the bear travels up to 10km/day (Wykoff, 2009). Deer which do travel much further, migrating around 80 km per year, but during a time where the fruit is not on the pawpaw tree. Wykoff (2009) explains that even understanding the movements of modern animals observed eating the fruit, it is still essential to know where, in a paleoenvironment, frugivorous and omnivorous species ranged during the time of pawpaw fruit ripening, and to include the related direction and distance to truly assess any potential zoochory candidate.

The pawpaw trees as far north as Ontario are potentially relict populations from a recent postglacial time where they would have been deposited by mammoths while they still roamed widely on the Allegheny plateau as well as on the lake plains (Peterson, 1991). Peterson (1991) hypothesized that after the last ice age, the same humans that hunted the mega fauna to

extinction could have saved the relic populations and become the primary vehicle for pawpaw seed dispersal. In many species, rivers are used as corridors of seed dispersal, joining separated communities (Berković et al., 2018; Cushman et al., 2014), a process known as hydrochory. Peterson stated hydrochory may have played a role but it was in all likelihood not a significant factor, partially because buoyancy of the seed is lost as water is imbibed (Wykoff, 2009).

The fossil record shows that *Asimina* was in parts of North America as early as the Eocene (52 Mya), reaching as far north as Ontario prior to the Laurentide ice sheet (2.6 Mya, ending 0.011 Mya). However, it seems unlikely that the species remained so far north during this time, more likely, the species was restricted to a small refugia in the Gulf of Mexico (Wyatt et al., 2020). During this time of glaciation, pawpaw's evolutionary zoochorous partners were lost. Conservative estimates place humans on North America prior to the retreat of the Laurentide ice sheet, sometime around 0.012 Mya, placing them in time to help pawpaw migrate north from its refugia in the Gulf of Mexico. A curiosity of the species not discussed by any previous investigator in relation to pawpaw's native range, is the presence of cold tolerance in a member of a tropical species. It is an important factor, as it is the only member of approximately 2,300 species to be able to tolerate conditions in the north of its range, even more, the species has so well adapted that it has a chilling requirement to produce flowers. The emergence timing of this trait may be an important indicator of natural or artificial selection in the species. It is not clear whether pawpaw had this ability prior to the last ice age or after. The Piacenzian stage of the Pliocene (2.6 to 3.6 Mya), the period before the ice age, was an interval of sustained global warmth with mean global temperatures of only 2 to 3 °C lower than today (de la Vega et al., 2020; Dowsett et al., 2013), meaning that the level of cold tolerance in current day pawpaw may not have been so high, or even required for it to be found in as far North as Ontario 2.6 Mya. An investigation into the mechanism of cold tolerance and its emergence may help to

elucidate the factors contributing to the species' northern migration. For now, the mechanisms of seed dispersal and the species distribution remain unclear.

Genome sequencing of non-model species

A brief history of sequencing

In 1953, the first biological molecule was sequenced by Frank Sanger (Sanger and Thompson, 1953a, 1953b), the protein sequence of insulin was determined by randomly fragmenting its two chains, deciphering each fragment, then overlapping the fragments to create a complete consensus sequence. The next big milestone in sequencing came eight years later in the 1960s for the first time, when the 76 bases long alanine tRNA from *Saccharomyces cerevisiae* was deciphered using a similar process employed to discern the first protein sequence; first RNAse A and RNAse T1 were used to fragment the RNA, the fragments pieces were separated by chromatography and electrophoresis, then sequential exonuclease digestion, and again an overlap consensus to conclude the sequence (Holley et al., 1965). This was a massive undertaking requiring 140 kg of yeast, and a team of five researchers working for three years to determine 76 nucleotides. Three years later, Wu (1968), with the use of a primer extension method was able to report on the 12 bases of the cohesive ends of bacteriophage lambda DNA. Then Gilbert and Maxam (1973) reported the 24 bases of an *Escherichia coli* lactose-repressor binding site by copying its DNA into RNA and sequencing the RNA fragments. A method that took two years to complete. Around the same time the sequencing of the *lac* repressor binding site was completed, Sanger and Coulson (1975) determined two sequences in bacteriophage ϕ X174. Their method used *E. coli* DNA polymerase I and DNA polymerase from bacteriophage T4 (Englund, 1972, 1971) along with different limiting nucleoside triphosphates and concurrent fractionation of the products according to size by ionophoresis on acrylamide

gels (Sanger and Coulson, 1975). The new rapid and simplified method known as ‘plus and minus’ required the preparation of four reactions of template DNA, DNA Pol I, a primer, and radiolabelled deoxynucleotides to produce new fragments of varying lengths. The four reactions were then split into ‘plus’ and ‘minus’ reactions. The ‘minus’ reactions contained three deoxynucleoside triphosphates (dNTPs) with DNA extension terminating once reaching the missing dNTP. In the ‘plus’ reaction only a dNTP was added to produce fragments of different lengths. The newly synthesized fragments of the eight reactions were loaded on to polyacrylamide gels and put onto X-ray film; the sequence could then be read directly off the resulting ladder. Only two years later Sanger and colleagues (1977) described a new breakthrough, an even faster and more accurate method which allowed for hundreds of bases to be deciphered in a single day. Initially called chain-termination method because it made use of the chain-terminating 2,3-dideoxynucleoside triphosphates (ddNTPs) to produce fragments of different lengths, it required four separate extensions of a tritium radiolabelled primers using DNA polymerase and a trace amount of one of the four ddNTPs. Fragments of differing length were produced by incorporating a chain-terminating ddNTP during the polymerase extension of the template DNA. Similar to the ‘plus or minus’ method, the fragments were measured by electrophoresis on polyacrylamide slab gels and X-rayed to deduce the sequence directly from a ladder image.

Not long after chain-termination sequencing was introduced, Staden (1979) proposed the idea now known as “shotgun sequencing”, also called random sequencing, which allowed for a much faster sequencing of larger genomes. This method used bacterial vectors clone to random fragments of a long DNA molecules. The fragments were then sequenced and resulting reads overlapped to generate an assembly. In 1981 Messing improved upon the method further by developing a single stranded M13 phage vector (Messing et al., 1981). Another milestone came

one year later when Sanger implemented the shotgun sequencing with M13 vector to assemble the 48,502 bp sequence of bacteriophage λ (Sanger et al., 1982). Smith and Hood (Smith et al., 1986, 1985) used fluorescent DNA primers to partially automate the Sanger method [Figure 1-2]. With the Human Genome Project (HGP) beginning in the 1980s and the growing commercial interest in sequencing, came more advances leading to a surge of new sequencing technologies being developed. An era of sequencing relying on rapidly advancing technology, collectively called Next-Generation Sequencing (NGS). For a detailed history and advancement of NGS see (Giani et al., 2020; Goodwin et al., 2016).

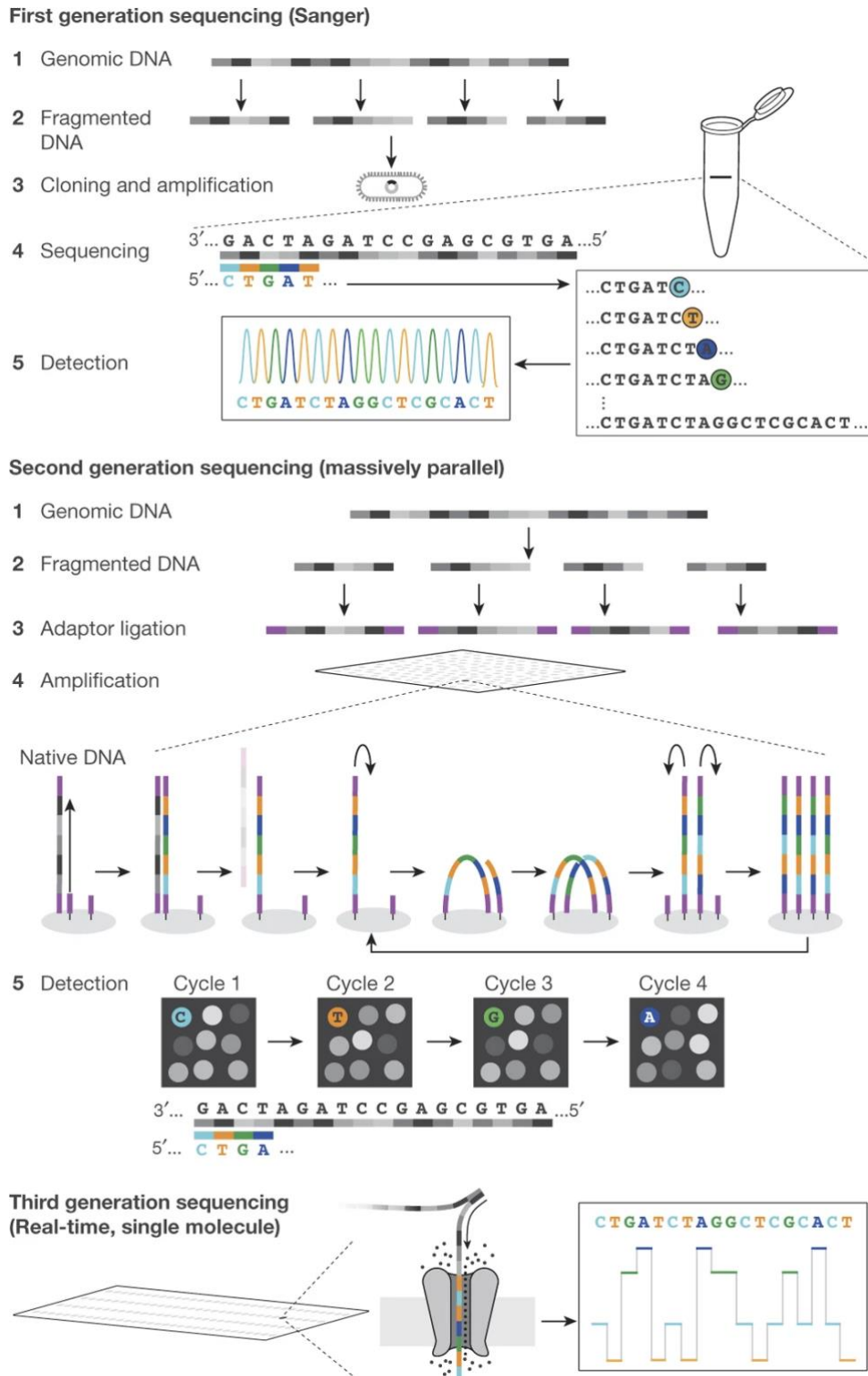


Figure 1-2 History of genomic sequencing

Schematic of first, second, and third generation sequencing methods. Image is adapted from Shendure et al., Figure 1 (Shendure et al., 2017).

Next-Generation Sequencing

NGS or massively parallel technologies (454, Solexa/Illumina, ABI Solid, Complete Genomics, and Ion Torrent) use different chemistries and approaches, but the major change from the first generation is a shift from bacterial cloning and electrophoresis to measure fragment lengths, towards a more inexpensive process of reading luminosity from reactions [Figure 1-2]. This is a method called ‘sequencing-by-synthesis’ (SBS) where light is produced from polymerase-mediated incorporation of fluorescently labelled nucleotides (Seo et al., 2005; Shendure et al., 2017). Another distinct shift in the NGS technologies was multiplexing, the use of complex libraries containing all template DNA fragments in one reaction instead of the previous single reaction per tube approaches. A ‘library’ is essentially a pool of amplified DNA fragments with sequencing platform specific adapter sequences attached. Although each method is highly variable, there are three steps found in almost all library preparations. First is the fragmentation of native DNA, then the annealing of adapters to DNA fragments, followed by amplification. Fragmentation may occur via physical, chemical or enzymatic methods such as acoustic shearing or digestion by restriction site specific endonucleases (Head et al., 2014; Marine et al., 2011). Depending on the sequencing platform used, specific barcode sequences are required; these take the form of short oligonucleotides attached to the end of the fragmented template DNA and are required to immobilize the DNA fragments onto a two-dimensional surface for amplification of the DNA template. To facilitate multiplexing, unique barcoded adapter sequences can be used to identify each individual sample (Head et al., 2014). NGS technologies can also be either long- or short-read sequencing platforms. Short-reads, typically around 300 base pairs (bp) in length, are the low-cost, high accuracy approach, very cost affective for population-level studies.

A draw-back to NGS is the need for amplification during the library preparation as replication can bias sequencing of some regions over others, cause loss of information, and even introduce copying errors (Shendure et al., 2017). The PCR amplification is also an issue for species having regions of high GC content, because these are not being efficiently amplified (Chen et al., 2013). This is a particular issue in Illumina short read system (Bentley et al., 2008; Chen et al., 2013). Another issue with NGS is the relatively short reads, even with long read sequencing, the read length can be insufficient to span the many repeat regions often found in plant genomes. This can lead to heavily fragmented genome assemblies with incorrectly collapsed regions, misassemblies, and introduced gaps (Goodwin et al., 2016; Salzberg and Yorke, 2005). A benefit of NGS is that it can be very effective in identifying small variants such as single-nucleotide variations (SNVs) and short indels, however, the larger structural variations (SVs) are not as easily detected (van Dijk et al., 2018). It is worth noting that while many of the NGS platforms have fallen out of use, Illumina's highly accurate and low-cost sequencing has remained popular as the preeminent short-read sequencing technology.

Third generation sequencing / long read sequencing

The third generation of sequencing began in the 2010s, with arrival of the new technologies that were capable of sequencing large single DNA molecules without the need for any prior amplification, such as single-molecule sequencing (SMS) but also real time sequencing of the DNA [Figure 1-2] (Schadt et al., 2010), meaning it vastly increased speeds. Since then, constant improvements and innovations in third-generation sequencing (TGS) technologies mean that it is now possible to sequence single molecules hundreds of kilobases (Kb) in length (Giani et al., 2020). Two defining technologies came first from Pacific Biosciences (PacBio) in 2011 with the release of their PacBio RS sequencer using 'single-molecule real-time' (SMRT) sequencing (Eid et al., 2009) and then in 2014 with the introduction of nanopore sequencing

by Oxford Nanopore Technologies (ONT) (Jain et al., 2015). The length of reads from PacBio varies, for example the RS system can produce reads around 1.5 Kb length, but this can increase to 50 Kb with PacBio Sequel I & II. However, with approximately 1 error per 10 nucleotides, long reads can contain a high error rate of 13–15% (Carneiro et al., 2012; Quail et al., 2012). Unlike NGS, the errors are not biased by the replication and CG content, PacBio errors were completely random and with multiple sequencing runs these can be corrected for via read-to-read error correction. The PacBio (SMRT) technologies use a closed, circular ssDNA template called a SMRTbell. This is created during the library preparation, by ligation of hairpin adaptors to both ends of a dsDNA molecule which circularizes the DNA to form the circular SMRTbell structure (Travers et al., 2010). The SMRTbell libraries are then loaded onto a SMRT cell containing an array of 10 microns-wide wells termed zero-mode waveguides (ZMWs). The upgraded PacBio RSII platform contained 150,000 ZMWs, but this number has increased to 1 million ZMWs for the newer Sequel platform and to 18 million ZMWs for Sequel II, massively increasing throughput and decreasing cost (Giani et al., 2020; van Dijk et al., 2018). The sequencing reaction takes place in each of the ZMW which contains an immobilized polymerase bound to a primer complementary to the hairpin adaptors (Eid et al., 2009). Here, the polymerase replicates the template DNA and incorporates γ -phosphate fluorescently labelled nucleotides producing a fluorescence signal when excited by a laser. The colour and duration of light emitted during the reaction is captured in real time by a camera. PacBio offers two types of the sequencing, circular consensus sequencing (CCS) and continuous long read (CLR) sequencing. The main difference is that with the CLR sequencing reads are produced by a polymerase which generates a long sequence called “polymerase read” or a CLR, from a single pass of the molecule. While in the CCS mode, multiple subreads are generated from multiple polymerase passes of a single SMRTBell template and are collapsed into a single high quality consensus sequence. The CLR reads are longer but contain a much higher error rate

(Rhoads and Au, 2015). In 2019 the sequencing accuracy of CSS sequencing was improved upon, producing new long high-fidelity (HiFi) reads with an impressive 99.9% accuracy and an average length of 13.5 Kb (Wenger et al., 2019) [Figure 1-3].

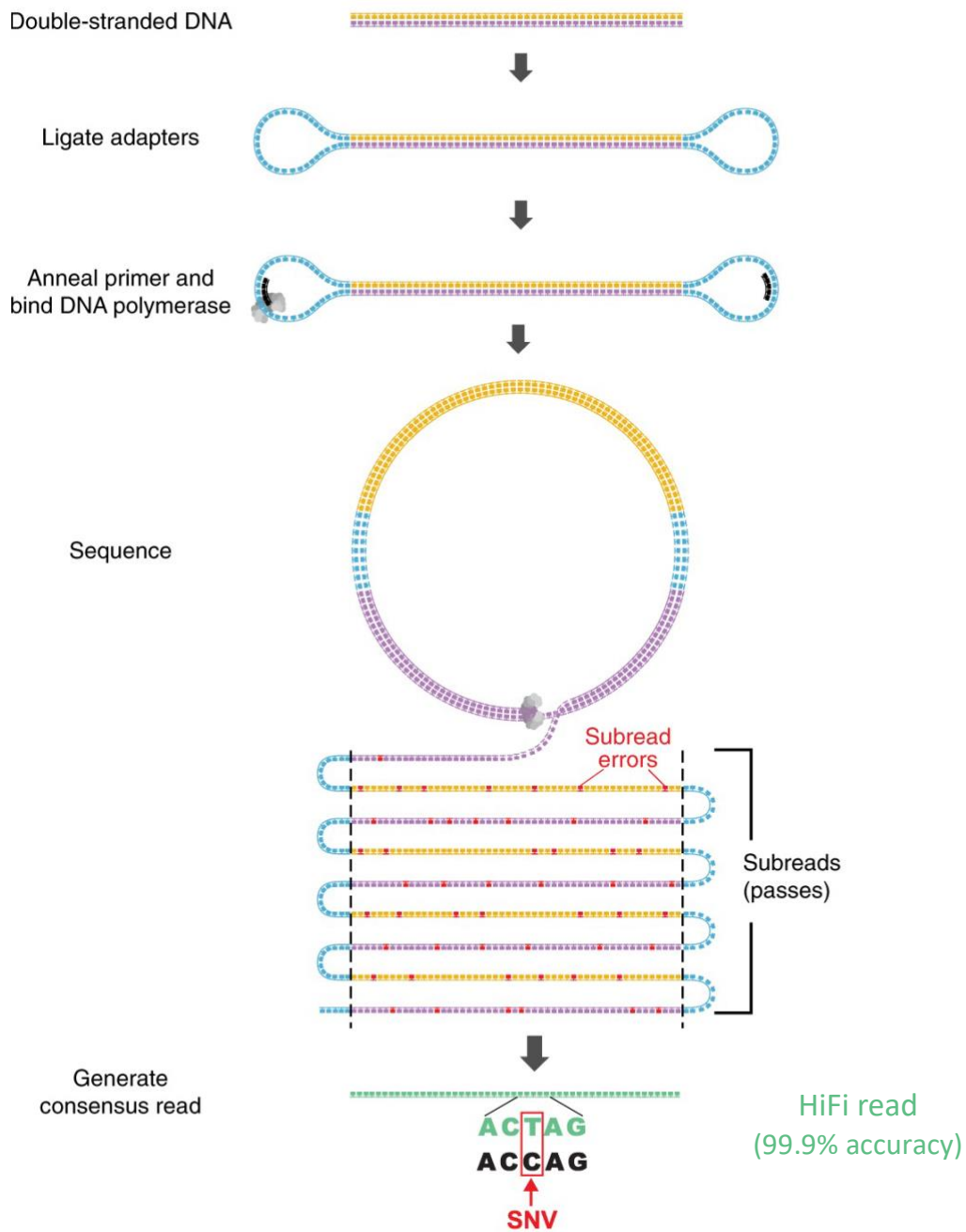


Figure 1-3 PacBio HiFi reads

The process of generating 99.9% accurate base calling HiFi reads produced from circular consensus sequencing (CCS). Figure is adapted from Wenger., et al., Figure 1 (Wenger., et al 2017).

The leading alternative to PacBio is the ONT. ONT utilizes a modified bacterial nanopore imbedded in an electrically resistant artificial lipid bilayer. A potential is applied to the lipid layers so that as a ssDNA molecule passes through the pore tunnel, it disrupts the current and generates distinct patterns in the change of current that can be used to read sequences (Jain et al., 2015). ONT can generate reads similar to SMRT sequencing, averaging around 10–20 Kb in length. However, an impressive feature of the ONT is its portability. For example, the MinION Nanopore device weighs only 100 g and can be run from the Universal Serial Bus (USB) port on a laptop (Quick et al., 2016). A draw-back of ONT, however, is that there appears to be a non-random sequencing bias that is difficult to correct for (Istace et al., 2017), and while there has been significant improvements in ONT, the error rate is still approximately 6% across their various platforms (Delahaye and Nicolas, 2021).

Assembly, and annotation in non-model species

Draft genome assembly

A draft genome assembly is the sequence of nucleotides inferred from sequencing data, and a *de novo* assembly is the process of generating a draft without prior information, the last being a difficult task. Eukaryotic genome sequencing and *de novo* assembly was once a very expensive and labour-intensive undertaking that could only be accessed by large consortia, but since the emergence of NGS and even more so with TGS, the process has become much more democratized (Giani et al., 2020). The optimum assembly is one that is as accurate as possible to the real sequence, and includes the correct nucleotide sequence for all chromosomes, along with an annotation map of the physical location of genetic elements [e.g. structural rearrangements, structural variants, repeat elements] (Ellegren, 2014). An accurate and complete genome is an important resource, revealing the content and arrangement of genomic material including the types and abundance of transposable elements, gene density, base

composition, noncoding RNAs, and nucleotide modifications. Information that is used in many downstream studies to genotype individuals, identify genes, plan gene manipulation experiments, for comparative genomics, transcriptomic, metabolic studies, and to interpret their results (Schuster, 2008). Because of the range of applications, a genome assembly may be used for, it is important that assemblies are as accurate as possible not only for the group generating them but for the wider scientific community who rely on the accuracy of the draft and its annotation to reliably perform analyses.

Significant advances in long-read sequencing and computer processing power have bought in a new era of sequencing, one in which a gapless telomere-to-telomere chromosome sequence can now be read and assembled (Kille et al., 2022). Though possible, it is still something difficult with only a limited number of species sequences reaching this standard thus far (Gonzalez de la Rosa et al., 2021; Hou et al., 2022; Miga et al., 2020; Xu et al., 2022). This is in part due to the difficulty in correctly assembling repeat regions [e.g. telomeres, centromeres, 5S rDNA clusters, and nucleolar organizer regions containing 45S rDNA.], particularly so in a *de novo* assembly, with such a level of completeness requiring ultra-long sequencing to span the entirety of the repeating sequence (Hou et al., 2022). A further complication in plants is that polyploidy is common place, the frequency of genome duplication events often result in large genomes complex with a high density of repeat regions (Pellicer et al., 2018). This can make assembly, especially haplotype phasing, very difficult. A certain amount of sequencing reads are required to cover the full length of the genome and to be able to correct for potential errors and biases introduced during the library preparation and sequencing platform but this amount varies with species and genome complexity (Jung et al., 2019).

While a telomere-to-telomere assembly is highly desirable and will provide the most use to the boarder scientific community, it is not often obtainable or even necessary for most analysis. It

is therefore necessary to consider genome size and complexity (repeat density, heterozygosity, ploidy level and CG content), and the completeness requirements for the project, as these factors in particular will affect the sequencing coverage, cost and overall quality of a *de novo* assembly project in a non-model species. (Angel et al., 2018; Jung et al., 2020, 2019).

Genome Size

As previously mentioned, it is important to measure the genome size and complexity to estimate the sequencing coverage needed but to also evaluate the completed draft assembly. An accurate way to determine the size of the genome in non-model species is to use flow cytometry, while k-mer frequency distribution can be used to obtain genome complexity (Li and Harkess, 2018). Flow cytometry compares propidium iodide-stained nuclei of the species of interest with a species of known genome size to simultaneously estimate genome size and ploidy level (Hare and Johnston, 2011). Alternatively, k-mer frequency uses the mean coverage of unique k-mers from raw Illumina DNA shotgun sequencing reads to infer genome size, perform repeat detection, and estimate heterozygosity (Li and Harkess, 2018). A certain amount of sequencing depth is required to cover the full length of the genome and to be able to correct for potential errors and biases introduced during the library preparation and sequencing platform. This varies across sequencing platforms and genomes, Jung et al. (2019) provides a more detailed review of coverage requirements. Additionally, databases for approximate genome sizes are available for plants from Kew Royal botanical gardens (<http://data.kew.org/cvalues>).

Genome assembly

The process of assembling a genome is computationally intensive to perform the assembly itself, but the process also requires access to significant storage capacity and memory to handle

the large volumes of data involved. The first step in assembling a genome is to generate consensus sequences, called contigs, there two common approaches: the de Bruijn graph (DBG) and the Overlap-Layout-Consensus (OLC). For a NGS based assembly, a DBG approach is better suited (Chaisson and Pevzner, 2008; Limasset et al., 2016). In this case the already short reads are decomposed into even shorter fragments (k-mers) of n nucleotides. All possible k-mers from one sequencing read are compared to all the possible k-mers from all other reads to form graphs of matching k-mers between reads. The rationale being that the graph can inform connections between reads and extend related reads into contigs. The OLC approach, introduced by Myers (2005), is better suited to the longer TGS reads because it is there is much larger sequence overlap (Li et al., 2012). A genome assembler using an OLC approach will run an “all vs all” comparison to identify and extend overlapping reads from the resulting graph. The approach is slower and more computationally intensive because it also requires a multisequence alignment to correct errors.

Once contigs have been generated, error corrections need to be performed to improve the accuracy of the assembly. The assembly process can provide potential sources of error such as insert/deletions, base calling errors, and misassemblies from poor read mapping, all need to be corrected for in a step known as ‘polishing’(Chu et al., 2017; Heydari et al., 2017). There are a numerous sequencing tools available, each with different approaches to assembling contigs and error correction, for a comprehensive reviews see (Chu et al., 2017; Giani et al., 2020; Heydari et al., 2017).

Gap Filling and Scaffolding

Depending on the sequencing platform, genome complexity, sequencing coverage and many other factors, a *de novo* assembly will likely be highly fragmented, limiting its usefulness for downstream analysis. After the initial contig assembly and correction (polishing), if enough

information about contig orientation and position is known, scaffolding can be performed to improve the assembly by extending contig length. By filling in gaps between contigs and linking them into scaffolds, the continuity of an assembly can be improved. However, even after the post processing polishing step, a *de novo* assembly will still likely contain multiple misassemblies (inversion and translocations), as well as having many remaining gaps. This can be a consequence of genome complexity, heterozygosity, polyploidy and repeats in particular, which are not handled well by many assembly pipelines (Ellegren, 2014). The contiguity of an assembly can be improved even further with supporting information from optical mapping methods (e.g., BioNano), linked-read technologies (e.g., 10X Genomics Chromium system), or chromatin-association/interaction analysis (Hi-C) (Lieberman-Aiden et al., 2009). Hi-C uses chromosome conformation capture, which involves crosslinking chromatin with formaldehyde, followed by digestion. The fragmented DNA is labelled with biotin and only the covalently linked fragments are re-ligated. These fragments are sequenced using Illumina, and each sequencing read contains details about the physical interaction of the chromatin, but also positional information, and this can be used to generate a map of long range physical interactions that can be used in conjunction with any assembly pipeline to correct misorientations, ordering, and extending scaffolds, allowing for near chromosome level assembly (Belton et al., 2012; Lieberman-Aiden et al., 2009).

Quality assessment

A completed draft assembly needs to be assessed for errors and to evaluate how successfully the original DNA sequence has been reassembled. By measuring contiguity, completeness, and accuracy it is possible to identify and remove potential issues in the assembly. The contiguity is a measure of composition of the draft assembly including number of contigs, contig length, and gaps between contigs essentially measuring how fragmented the assembly is. L90, N90,

L50 and N50 are standard contiguity metrics generated by ranking contigs by length and using the longest to report the least number of contigs (L) needed to cover either 90% (L90) or 50% (L50) of the genome and the number of bases (N) in the shortest of those contigs. The larger the L50 or L90 is, the more fragmented the assembly. The completeness of a genome is evaluated in terms of overall sequence completeness and gene space completeness. Overall completeness refers to how the size of the assembled sequence compares with the genome size estimated via flow cytometry and k-mer frequency distribution. Large variations here can indicate issues during assembly. The gene space completeness broadly refers the correct assembly of various genetic elements expected to be present in an assembly. A useful method is to search for the presence of benchmarking universal single-copy orthologs (BUSCO), a curated list of ancestrally conserved genes, and reporting the number of complete, fragmented, and missing genes to evaluate the assembly quality (Simão et al., 2015). The LTR Assembly Index (LAI) is a metric for estimating the completeness of the transposable element space by measuring the extent of assembled transposable elements (TE) regions in the genome assembly (Ou et al., 2018). A tool like Merqury can even be used to evaluate both accuracy and completeness (Rhie et al., 2020). Merqury can do this by decomposing the sequencing reads and the assembled genome into k-mer catalogues and comparing the two k-mer groups. By comparing the two, Merqury can calculate the error rate of bases in the assembly but not the reads. In an optimal assembly, all k-mers of sequencing reads should matches the k-mer of the assembly, indicating a good assembly of the sequencing data.

Repetitive element problems in the assembly

Repeats refers to interspersed repeats TE, as well as tandem copies of similar nucleotide sequences found throughout the genome. The potential issue with these homopolymeric sequences is the sequence similarity, because assembly tools are not able to distinguish between

them. This often results in mis-assemblies where regions of the genome which should be repeated in the draft assembly are instead collapsed into a single instance; alternatively, it could result in an incorrect multiplication of the repeats assembled (Phillippy et al., 2008). When an assembler is unable to identify the correct number of repeats it will stop extending the contigs at the border of the repeats, resulting in a more fragmented assembly (Chaisson et al., 2015). If a genome contains many repeat regions, then, using longer sequencing reads allows to avoid this issue by spanning beyond the length of the repeat sequences. Highly heterozygous species create an issue where the assembler tools will try to collapse the multiple allelic differences into a single consensus, producing a haplotype with alternative alleles or a two separate haplotypes (Pryszcz and Gabaldón, 2016). The problem for draft assembly confidence here is that some heterozygous regions may appear twice or not all, further fragmenting an assembly. The problem is already difficult in diploids but increases significantly with increasing ploidy levels. Not only does higher ploidy increase the number of allelic variants but whole-genome duplication events are often associated with genome rearrangement, atypical recombination, transposable element activation, meiotic/mitotic defects, and indels (Hufton and Panopoulou, 2009), which an assembler tool may arrange into the wrong subgenome. For this reason, 50% to 100% more sequence data might be required for assembly issues in polyploid and highly repetitive genomes (Jung et al., 2019). GC-content is also an issue during library preparation, as mentioned above, and in Illumina short read systems a GC bias can result in low or absence of sequencing coverage of those regions (Chen et al., 2013), again resulting in gaps and inaccurate haplotype phasing. HiFi sequencing can help to reduce GC bias by creating longer reads that span multiple GC-rich and GC-poor regions. This can help to increase coverage of GC-rich regions and provide more accurate sequence information from these areas. Additionally, HiFi sequencing can also reduce the effects of PCR amplification bias, which can be a major source of GC bias in short read sequencing.

Annotation

Annotating a genome is the process of ascribing the structural and functional roles to a genome sequence using evidence from closely related species and analysing sequence structure to predict coding regions (Salzberg, 2019). The process can be split into two categories: structural annotation, and functional annotation. The structural annotation involves identifying the location and structure of genes and other functional elements in the genome, while functional annotation involves identifying the function of those elements. Because of the difficulties involved for non-model species, the annotation is often confined to transcripts or protein-coding sequence (CDS) (Ekblom and Wolf, 2014). Structural annotation pipelines may use various tools and approaches but broadly fall into a two-phase process. The first phase is a computational phase where repetitive elements are identified, followed by either *ab initio* or evidence-based (expressed sequence tags (ESTs), homologous proteins, etc..) approaches to generate gene predictions. In the second phase, this information is synthesized into gene annotations (Yandell and Ence, 2012).

Repetitive elements are sequences of DNA that are repeated multiple times within a genome. These elements can include transposable elements, tandem repeats, and satellite DNA. Repetitive elements play an important role in genome evolution and can have an impact on gene expression and regulation. However, repetitive regions are not at all well conserved among species and present unique challenges for gene prediction tools. It is therefore, essential that these regions of the genome assembly are “masked off” before predicting genes (Cantarel et al., 2008). Repeat families including long terminal repeats (LTR) can be predicted by tools like RepeatModeler (Flynn et al., 2020) to mask these regions of the genome with RepeatMasker (Tarailo-Graovac and Chen, 2009). When repeats have been successfully identified and masked off, an *ab initio* or evidence gene prediction process can begin.

Ab initio tools are those that predict the location and structure of genes and other functional elements in a genome without using any experimental data. These tools rely on the conservation of gene structure and function across different species. A popular *ab initio* gene prediction pipeline is BRAKER2 (Brůna et al., 2021), it uses the tool BRAKER to predict genes using AUGUSTUS (Stanke and Waack, 2003). But it can also be used to train gene predictors such as, AUGUSTUS, GeneMark-E (Lomsadze et al., 2014), and ProtHint (Mathebula, 2016).

Evidence-based annotation is the process of using experimental data to confirm and refine the predictions made by *ab initio* methods. This evidence can come from a variety of sources such as isoform sequencing (Iso-seq), protein sequencing, and gene models from related species, or transcripts for the target species if available. The more evidence that can be provide, the more accurate the annotation will be.

Annotation pipelines are a series of steps that are taken to annotate a genome. These pipelines often include a combination of *ab initio* prediction, evidence-based annotation, and can include functional annotation. Some popular pipelines include MAKER, MAKER-P, and MAKER-P-EVA (Cantarel et al., 2008; Holt and Yandell, 2011) which allow for integration of several evidences such the *ab initio* predictions from AUGUSTUS and Semi-HMM-based Nucleic Acid Parser (SNAP) (Korf, 2004), along with homologous protein models, *de novo* assembled transcript, and ESTs.

However, a poor-quality annotation will have negative effects on all further research projects relying on it to identify and target specific genomic features, so as with the assembly itself, it is important to also assess the quality of the annotation. Currently, it is still difficult to perform this assessment accurately. Some quality metrics including the number of gene models, exons per gene model, and the average lengths of genes, exons and transcripts can be informative but provide little assessment of the quality. The MAKER2 pipeline includes a measure called

annotation edit distance (AED) to evaluate how well an annotation agrees with overlapping aligned ESTs, mRNA-seq and protein homology data (Holt and Yandell, 2011). This quality evaluation can also be paired with a BUSCO estimation of the gene space completeness accounted for in the annotation itself, but also the contribution of evidence to the gene prediction from the ESTs, mRNA-seq and protein homology (Seppey et al., 2019).

Once the structural annotation is generated, it can then be used to assign functional annotation to coding regions of the genome. This can include identifying the proteins that are encoded by a gene, the regulatory regions that control gene expression, and the repeat elements that compose the genome. Relevant information on gene families and functional information can be gathered from model and annotation databases such as the Gene Ontology Consortium (Harris et al., 2004) or 'Kyoto encyclopedia of genes and genomes' (KEGG) (Kanehisa and Goto, 2000).

Reduce representation approaches to population structure analysis

Genetic variants are variations in the DNA sequence that occur within a population. These variations can include single nucleotide polymorphisms (SNPs), small insertions or deletions (indels), copy number variations (CNVs), and structural variations (SVs). Genetic variants can have a wide range of effects on an organism, from having no effect at all to causing genetic disorders, but they can also be used to genotype individuals by determining the specific genetic variants present in the genome of an individual. Reduce representation sequencing (RSS) approaches to genotyping is a method that reduces the cost and complexity of genotyping by focusing on a subset of the genome, fragmenting the DNA and sequencing a portion of the fragments. There are several commonly used methodologies; restriction site-associated DNA

sequencing (RADseq) (Baird et al., 2008) and genotyping-by-sequencing (GBS) (Elshire et al., 2011). RADseq methods (original RADseq, ddRAD, ezRAD) are all slightly different in their preparation but follow a similar approach. High quality genomic DNA is fragmented with one or more enzymes before addition of the Illumina sequencing adapters (oligos). The restriction digestion enzymes will produce a wide range of fragment sizes, and to produce fragments small enough for Illumina sequencing a size selection step is also necessary (Andrews et al., 2016). GBS is similar to RADseq, but instead of fragmenting the DNA with multiple restriction enzymes, it uses only one and requires a fragment size selection set (Elshire et al., 2011).

RRS is a highly cost-effective method of sequencing a large number of loci in multiple individuals. Barcode adapters are added to fragments before sequencing so that many unique samples can be sequenced on a single sequencing lane. This massively reduces the genotyping cost per individual but require additional bioinformatic processing skills before running any analysis (Elshire et al., 2011). The fields of ecological, evolutionary, and conservation genomics have benefited greatly from decreasing sequencing cost and the introduction of reduce representation methods. Previously, relatively small numbers of loci from microsatellites were used to infer population structure, but with the massive throughput of NGS and RRS approaches, also referred to as genotyping-by-sequencing, not to be confused the specific method of GBS by Elshire et al (2011). With RRS it is now possible to discover potentially thousands of polymorphic genetic markers (Andrews et al., 2016; Meger et al., 2019) providing useful insights into population structure, demography history, hybridization, genetic diversity, QTL mapping, and phylogeography (Andrews et al., 2016; Christiansen et al., 2021; Ravinet et al., 2016).

Another benefit of RRS is that thousands of variants can be called without the prior need for a reference genome (Emerson et al., 2010), but incorporating a reference genome will significantly improve the reliability of genotype calls and downstream analysis (Torkamaneh

et al., 2016). When a reference genome is available, millions of sequencing reads are aligned against the reference to identify genetic variants, usually SNPs (Davey et al., 2011). The choice of genetic markers used in population genetics can have significant impacts on inferences of population structure, genetic variation among populations, estimates of heterozygosity and overall genetic diversity (Arnaud-Haond et al., 2005; Hamrick and Godt, 1990; Meloni et al., 2013). Increasing resolution with cheap and highly polymorphic markers is particularly useful in population studies of clonally reproducing species (Arnaud-Haond et al., 2005), as in the case of the pawpaw tree.

Downstream analysis issues from RRS can arise from the potential for allele dropout, PCR biases, uneven coverage, genotyping errors, and skill level requirements that lead failure to identify non-independent and uninformative variants in linkage disequilibrium (LD) (Andrews et al., 2016; Christiansen et al., 2021; Lowry et al., 2017). Depending on the sequencing coverage, read depth and genome complexity, the density of available SNPs can be negatively affected. With a low density of markers, there is an increased risk in biased or erroneous analysis (Hoban et al., 2016; Whitlock and Lotterhos, 2015). Alternatively, low frequency but influential variants can be discarded, being indistinguishable from sequencing or base calling errors (Carson et al., 2014; Díaz-Arce and Rodríguez-Ezpeleta, 2019). Consequently, it is necessary to correctly filter called variants, often focusing on Hardy–Weinberg equilibrium (HWE), missing proportion (MSP) and minor allele frequency (MAF), to be able to call genotypes (Pongpanich et al., 2010; Teo et al., 2007). However, there are no strict rules for SNP filtering as each experiment with different species, RRS library method, and sequencing coverage will obtain a different set of markers. There is then a need for individual optimization and accurate reporting on filtering steps to be able to accurately reproduce results.

Chapter 2 : Draft assembly and Virginia state population

(Manuscript submission)

Population structure and geneflow influenced by waterways in tree species *Asimina triloba* (Pawpaw)

James Friel^{1a}, Alicia Talavera^{2a}, Silvia Manrique¹, Tomas Hasing³, Elijah Rinaldi³, David C. Haak³, José I. Hormaza^{4*}, Aureliano Bombarely^{5*}

¹ Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy

² Dipartimento di Biologia, Università di Pisa, 13-56126 Pisa, Italy

³ School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA

⁴ Instituto de Hortofruticultura Subtropical y Mediterránea “La Mayora”, Consejo Superior de Investigaciones Científicas (CSIC), 29750 Algarrobo-Costa, Málaga, Spain

⁵ Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas - Universitat Politècnica de València (IBMCP-CSIC-UPV), Valencia, Spain

^a These authors contributed equally to the manuscript

*

Abstract

Little is known about the role geographic features play in the genetic diversity of the fruit tree species *Asimina triloba* (Pawpaw). To address this, we have explored wild populations,

analysing 124 individuals from 28 patches across the state of Virginia, focusing sampling along the banks of the James River and from mountainous and lowland areas. Our analysis used a genotyping-by-sequencing (GBS) approach to call variants and revealed a coarse population structure with admixture throughout the state. Despite the homogenous nature of the population, we observed higher than expected levels of heterozygosity and a longitudinal isolation-by-distance (IBD) separation. Subpopulations could be clustered by the river basin from which they originated using principal component analysis (PCA). Further, we identified increased genetic similarity between sub-populations growing along a river, which supports the idea that hydrochory plays a major role in pawpaw's seed dispersal. We present the first draft genome assembly of *A. triloba* and provide an insight into the factors influencing dispersal and genetic diversity in pawpaw populations.

Introduction

The Annonaceae is a diverse pantropical family in the early-divergent Magnoliid clade, containing 130 genera and over 2000 species of trees, lianas and shrubs (Angiosperm Phylogeny Group, 2016). There are several economically important fruit-bearing trees from the Annonaceae family with a tropical or sub-tropical distribution, most of which are in the genus *Annona*. These include the cherimoya (*Annona cherimola* Mill.), sugar apple (*A. squamosa* L.), and soursop (*A. muricata* L.) (Hormaza, 2014). The lesser known *Asimina* genus is comprised of at least eight species, and several possible hybrids, all are native to North America (Callway, 1992; Horn, 2015; Kral, 1960). Perhaps the most peculiar species of *Asimina* is *Asimina triloba* [L.] Dunal (Pawpaw), which is the most widespread of the genus. It is indigenous to 26 states in the eastern United States, ranging from New York, and southern Michigan in the north, South to Northern Florida, and West to Eastern Texas, Nebraska, and

Kansas. Interestingly, wild patches have also been found as far north as Southern Ontario, Canada (Fox, 2012). The pawpaw fruit is described as having complex tropical flavour combination that is a mixture of banana, mango and cherimoya (Duffrin and Pomper, 2006). In the past 100 years, there has been a growing interest in domesticating and developing an industry around the fruit, similar to the highly successful market developed from the wild US blueberry (*Vaccinium angustifolium*) (Callway, 1992; Pomper and Layne, 2003). Pawpaw flowers have a dark marron colour and a fetid or sometimes yeast-like smell, typical of carrion fly or beetle pollinated species (Goodrich et al., 2006; Kral, 1960; Willson and Schemske, 1980). There appear to be two self-incompatible barriers, a temporal barrier in the form of strongly protogynous flowers and another unknown method (Lagrange and Tramer, 1985; Losada et al., 2017). Thus, pawpaw is an obligate outcrossing species, and is believed to favour asexual reproduction, which it achieves by means of root suckering (Willson and Schemske, 1980).

It is currently thought that seed dispersal was carried out by the large pre-historic mammals until the Pleistocene (3– 0.012Mya) which have since gone extinct (Janzen and Martin, 1982); and hypothesized that after the last ice age humans may have become the primary method (Keener and Kuhns, 1997; Peterson, 1991). Indigenous American people, who moved around in the current distribution range of the pawpaw, were known to grow and use the pawpaw and may have contributed to its current distribution during their northern migrations along trade routes (Keener and Kuhns, 1997). In opposition to this idea, Murphy (2001) suggested that human intervention was not required; floating seeds, and mammals such as raccoons (Cypher and Cypher, 1999), bears, and even possums could be filling the role of the now extinct mega fauna. The extent of genetic research on *A. triloba* to date has been limited to a handful of studies which made use of allozymes (Huang et al., 1998), Simple-sequence repeats (SSR), or Random Amplified Polymorphic DNA (RAPD) to assess clonality and genetic diversity in a

number of wild patches (Huang et al., 2000; Pomper et al., 2009; Tulowiecki, 2021; Wyatt et al., 2021, 2020) and cultivated varieties (Pomper et al., 2010). Using nine nuclear microsatellite loci across the known range, Wyatt et al. (2020) identified two populations that largely followed an east/west divide along the Appalachian Mountains. A further study, using the same SRRs and several purported anthropogenic sites, provided evidence for a human role in the species' movement across the current species distribution range by demonstrating a close link between rare loci at anthropogenic sites once used by indigenous people and wild pawpaw patches hundreds of kilometres away, suggesting selection and movement of pawpaw by people (Wyatt et al., 2021).

The mechanisms of seed dispersal and the species distribution, however, remain unclear. The pawpaw tree is wide spread, typically found in riparian habitats along creeks, small streams and rivers, but is also frequently found in a range of drier habitats (Horn, 2015; Kral, 1960; Lagrange and Tramer, 1985). In many species, geographical features such as rivers play an important role in shaping population structure (Blanchet et al., 2020; Muneeppeerakul et al., 2008); this has been seen in both plant (Geng et al., 2015; Looy et al., 2009; Schleuning et al., 2008) and animal species (Brunke et al., 2019; Ortiz et al., 2018). Rivers have the potential to fragment populations by creating a matrix of barriers (Chaput-Bardy et al., 2009; Zhang et al., 2007) or corridors of increased geneflow joining separated communities (Berković et al., 2018; Cushman et al., 2014; Wei et al., 2013). Many species are known to use sexual and asexual hydrochory, the dispersal of seeds and plants by water, to transfer genetic material long distances (Berković et al., 2018; Nilsson et al., 2010, 1991). Both the seeds and fruit of the pawpaw float, which has led to the suggestion that hydrochory may be a vehicle for distribution in pawpaw (Keener and Kuhns, 1997; Murphy, 2001; Peterson, 1991).

In this study, we have collected samples from across the state of Virginia, including sampling trees from multiple sites along the James River, which starts in the Appalachian Mountains and

flows 560 km to the coast, and constructed GBS libraries (Elshire et al., 2011) from the samples. The GBS reads were mapped to our *A. triloba* draft assembly, identifying 4,845 Single Nucleotide Polymorphisms (SNPs) in 124 individuals from 28 sites across the state of Virginia. These SNPs were used to explore the potential role and influence of geographical features such as rivers, and the geographical distance, have on gene flow in *A. triloba*.

Materials and Methods

Reference sampling and sequencing

The specimen selected for construction of the reference assembly was the oldest living pawpaw (>100 years) in collection of the Harvard Arboretum, accession 12708*A (Hormaza, 2014). Total genomic DNA was extracted from 1 g of fresh young leaves using the DNeasy Plant Minikit (Qiagen). The amount and the quality of the DNA was assessed with a NanoDrop One (ThermoFisher) and a Qubit 2.0 (Invitrogen) with the dsDNA HS Assay.

Prior to DNA sequencing, the genome size was estimated using flow cytometry. Flow cytometry was carried out following the same methodology as used in (Hasing et al., 2019).

Two aliquots of 1 ug of high-quality DNA were sent for sequencing to the GCB Facilities at the Duke University. The DNA sequencing consisted of two next generation sequencing methods: 1) Short reads with Illumina HiSeq 2500 paired-end 2x150 bp sequencing with an estimated insert size of 300 bp, and a sequencing coverage estimated of ~ 40X; 2) Long reads with PacBio Sequel with an estimated coverage of ~ 10X.

Reference genome assembly

Before assembly genome size was estimated with GenomeScope (Vurture et al., 2017) using Illumina short reads. PacBio Sequel reads were processed filtering out reads below 1 Kb with

Seqtk.v.1.2-r94 (Li, 2022). No adapter sequences were found with FastQC.v.0.11.5 (Andrews, 2010). Illumina reads were processed with a minimum quality of 30 and a minimum length of 50 bp with Fastq-mcf from the Ea-utils package v1.05 (Aronesty, 2013). The PacBio Sequel reads were assembled with Canu.v.2.2 (Koren et al., 2017) with the default parameters. The output was polished with Polca from the MaSuRCA package v.4.0.5 using the Illumina processed reads.

The quality, completeness, and contiguity of the assembly “Astri041” was evaluated before being used for read mapping. Assembly and contiguity stats, including N50 and N90, were calculated using a custom script available on Github; FastaSeqStats (Bombarely, 2022). Gene space completeness was assessed using benchmarking universal single-copy orthologs (BUSCO) (Simão et al., 2015). BUSCO.v.5 with the eudicot_db10 data set was used to search for 2326 orthologous genes expected to be present in all eudicot species. A k-mer based approach was also used to evaluate completeness and the overall quality of the assembly using Merqury.v1.3 (<https://github.com/marbl/merqury>). Merqury takes as input the reads used in the assembly decomposed into a dataset of k-mers. For the Astri041 assembly of 0.8 Gb, the recommend k-mer size of 20 was used to construct the required meryl datasets for the Illumina short reads. Merqury was then able to evaluate how the sequencing reads had been incorporated into the final assembly generating a completeness score (i.e. a phred-scaled consensus quality (QV) score) along and copy number spectra plots to visually inspect the assembly for un-assembled reads and artificial duplications (Rhie et al., 2020). Finally, the continuity of repetitive sequences was assessed with an LTR assembly index (LAI) score used using LTRretriever.v.2.8 (Ou and Jiang, 2018).

Virginia population sampling & GBS library construction

A. triloba samples were collected from wild trees at 11 sites comprising of 28 patches with average of five individuals per patch from across the state of Virginia, in the US. GPS locations along with the elevation of each patch was recorded Table 2-1. Samples were stored at -20 °C until extraction and GBS library preparation. Total genomic DNA was extracted from the 134 Virginia samples from fresh leaves and GBS library preparation was carried out following a protocol adapted for (Elshire et al., 2011). The 134 samples were divided into two sets to account for the 96 unique “barcode” sequences available. This meant that two libraries were prepared using the following approach. DNA concentration was quantified using a Thermo Invitrogen Qubit 2.0 Fluorometer and a 10 µl aliquot of 10 ng of prepared DNA. Digestion of samples was carried out using restriction digestion enzyme ApeKI (New England Biolabs, Ipswich MA), during 2 hrs at 75 °C in a T100 Thermocycler (Bio-Rad Laboratories, Inc). A sample specific “barcode” and a common Illumina adapter sequence were ligated with T4 ligase (New England Biolabs, Ipswich MA) to sticky ends. During the ligation step, samples were incubated at 22 °C for 1 h and heated to 65 °C for 30 minutes in a T100 Thermocycler. 5 µl of each ligated sample was pooled into a single library. The resulting library was then amplified in a 50 µl solution containing 5 µl of pooled DNA library, 1X Taq Master Mix (New England Biolabs), and 12.5 pmol each of PCR primer [Table S2-1] containing complementary sequences for amplifying the fragments of DNA with ligated adapters. The PCR conditions were as follows; a primer step of 5 min at 72 °C; 98 °C for 30 s; 25 cycles of 98 °C for 30 s, 65 °C for 30 s; 72 °C for 30 s; and a final extension step at 72 °C for 5 min. Each library was purified using a Monarch PCR & DNA Cleanup Kit (New England Biolabs) and 1 µl was used for the quality evaluation and selection of fragment sizes using a BluePipin (Sage Science). A library was considered suitable for sequencing if adapter dimers were minimal (~128 bp in length) and the majority of the others DNA fragments were between 170 to 350 bp.

Each of the libraries were sequenced by BGI genomics with one lane of HiSeq2500 Illumina system as 2x150 bp. The sequences of the barcode adapters and the second, or “common” adapter sequence was shared among all samples and consisted of an ApeKI-compatible sticky end [Table S2-1].

Table 2-1 A. *triloba* sample locations

Pawpaw sample names and collection sites. This table contains the list individual trees sampled across Virginia. Table includes individual sample name, patch name, latitudinal and longitudinal coordinates, along with altitude and the area name of each sample location.

name	patch	lat	long	Altitude	Location	River Basin
AndyLT-1-1	AndyLT.1	37.46	-80.01	453	Andy Lane Trail	New River Basin
AndyLT-1-2	AndyLT.1	37.46	-80.01	453	Andy Lane Trail	New River Basin
AndyLT-1-3	AndyLT.1	37.46	-80.01	453	Andy Lane Trail	New River Basin
AndyLT-1-4	AndyLT.1	37.46	-80.01	453	Andy Lane Trail	New River Basin
AndyLT-1-5	AndyLT.1	37.46	-80.01	453	Andy Lane Trail	New River Basin
AndyLT-2-1	AndyLT.2	37.46	-80.01	441	Andy Lane Trail	New River Basin
AndyLT-2-2	AndyLT.2	37.46	-80.01	441	Andy Lane Trail	New River Basin
AndyLT-2-3	AndyLT.2	37.46	-80.01	441	Andy Lane Trail	New River Basin
AndyLT-2-4	AndyLT.2	37.46	-80.01	441	Andy Lane Trail	New River Basin
AndyLT-2-5	AndyLT.2	37.46	-80.01	441	Andy Lane Trail	New River Basin
AndyLT-3-1	AndyLT.3	37.45	-80.01	415	Andy Lane Trail	New River Basin
AndyLT-3-2	AndyLT.3	37.45	-80.01	415	Andy Lane Trail	New River Basin
AndyLT-3-3	AndyLT.3	37.45	-80.01	415	Andy Lane Trail	New River Basin
AndyLT-3-4	AndyLT.3	37.45	-80.01	415	Andy Lane Trail	New River Basin
AndyLT-3-5	AndyLT.3	37.45	-80.01	415	Andy Lane Trail	New River Basin
AndyLT-4-1	AndyLT.4	37.45	-80.01	429	Andy Lane Trail	New River Basin
AndyLT-4-2	AndyLT.4	37.45	-80.01	429	Andy Lane Trail	New River Basin
AndyLT-4-3	AndyLT.4	37.45	-80.01	429	Andy Lane Trail	New River Basin
AndyLT-4-4	AndyLT.4	37.45	-80.01	429	Andy Lane Trail	New River Basin
AndyLT-4-5	AndyLT.4	37.45	-80.01	429	Andy Lane Trail	New River Basin
AndyLT-5-1	AndyLT.5	37.46	-80.01	418	Andy Lane Trail	New River Basin
AndyLT-5-2	AndyLT.5	37.46	-80.01	418	Andy Lane Trail	New River Basin
AndyLT-5-3	AndyLT.5	37.46	-80.01	418	Andy Lane Trail	New River Basin
AndyLT-5-4	AndyLT.5	37.46	-80.01	418	Andy Lane Trail	New River Basin
AndyLT-5-5	AndyLT.5	37.46	-80.01	418	Andy Lane Trail	New River Basin
DevilB-1-1	DevilB.1	36.82	-82.62	482	Devil Bathtub	Holston River Basin
DevilB-1-2	DevilB.1	36.82	-82.62	482	Devil Bathtub	Holston River Basin
DevilB-1-3	DevilB.1	36.82	-82.62	482	Devil Bathtub	Holston River Basin
DevilB-1-4	DevilB.1	36.82	-82.62	482	Devil Bathtub	Holston River Basin
DevilB-1-5	DevilB.1	36.82	-82.62	482	Devil Bathtub	Holston River Basin
DevilB-2-1	DevilB.2	36.82	-82.64	525	Devil Bathtub	Holston River Basin
DevilB-2-2	DevilB.2	36.82	-82.64	525	Devil Bathtub	Holston River Basin
DevilB-2-3	DevilB.2	36.82	-82.64	525	Devil Bathtub	Holston River Basin
DevilB-2-4	DevilB.2	36.82	-82.64	525	Devil Bathtub	Holston River Basin
DevilB-2-5	DevilB.2	36.82	-82.64	525	Devil Bathtub	Holston River Basin
DevilB-3-1	DevilB.3	36.81	-82.65	588	Devil Bathtub	Holston River Basin
DevilB-3-2	DevilB.3	36.81	-82.65	588	Devil Bathtub	Holston River Basin

DevilB-3-3	DevilB.3	36.81	-82.65	588	Devil Bathtub	Holston River Basin
DevilB-3-4	DevilB.3	36.81	-82.65	588	Devil Bathtub	Holston River Basin
DevilB-3-5	DevilB.3	36.81	-82.65	588	Devil Bathtub	Holston River Basin
FairyS-1-1	FairyS.1	36.80	-80.11	302	Fairy Stone State Park	New River Basin
FairyS-1-2	FairyS.1	36.80	-80.11	302	Fairy Stone State Park	New River Basin
FairyS-1-3	FairyS.1	36.80	-80.11	302	Fairy Stone State Park	New River Basin
FairyS-1-4	FairyS.1	36.80	-80.11	302	Fairy Stone State Park	New River Basin
FairyS-1-5	FairyS.1	36.80	-80.11	302	Fairy Stone State Park	New River Basin
FairyS-2-1	FairyS.2	36.80	-80.11	343	Fairy Stone State Park	New River Basin
FairyS-2-2	FairyS.2	36.80	-80.11	343	Fairy Stone State Park	New River Basin
FairyS-2-3	FairyS.2	36.80	-80.11	343	Fairy Stone State Park	New River Basin
FairyS-2-4	FairyS.2	36.80	-80.11	343	Fairy Stone State Park	New River Basin
FairyS-2-5	FairyS.2	36.80	-80.11	343	Fairy Stone State Park	New River Basin
FallRP-1-1	FallRP.1	37.19	-80.32	410	Fall Ridge Preserve	New River Basin
FallRP-1-2	FallRP.1	37.19	-80.32	410	Fall Ridge Preserve	New River Basin
FallRP-1-3	FallRP.1	37.19	-80.32	410	Fall Ridge Preserve	New River Basin
FallRP-1-4	FallRP.1	37.19	-80.32	410	Fall Ridge Preserve	New River Basin
FallRP-1-5	FallRP.1	37.19	-80.32	410	Fall Ridge Preserve	New River Basin
HardR-1-1	HardW.1	37.75	-78.41	64	Hardware River Wildlife Area	James River Basin
HardR-1-1	HardW.1	37.75	-78.41	64	Hardware River Wildlife Area	James River Basin
HardR-1-3	HardW.1	37.75	-78.41	64	Hardware River Wildlife Area	James River Basin
HardR-1-4	HardW.1	37.75	-78.41	64	Hardware River Wildlife Area	James River Basin
HardR-2-1	HardW.2	37.75	-78.41	82	Hardware River Wildlife Area	James River Basin
HardR-2-2	HardW.2	37.75	-78.41	82	Hardware River Wildlife Area	James River Basin
HardR-2-3	HardW.2	37.75	-78.41	82	Hardware River Wildlife Area	James River Basin
HardR-2-4	HardW.2	37.75	-78.41	82	Hardware River Wildlife Area	James River Basin
HardR-2-5	HardW.2	37.75	-78.41	82	Hardware River Wildlife Area	James River Basin
HardR-3-1	HardW.3	37.74	-78.41	83	Hardware River Wildlife Area	James River Basin
HardR-3-2	HardW.3	37.74	-78.41	83	Hardware River Wildlife Area	James River Basin
HardR-4-1	HardW.4	37.74	-78.41	77	Hardware River Wildlife Area	James River Basin
HardR-4-2	HardW.4	37.74	-78.41	77	Hardware River Wildlife Area	James River Basin
HardR-4-3	HardW.4	37.74	-78.41	77	Hardware River Wildlife Area	James River Basin
HardR-4-4	HardW.4	37.74	-78.41	77	Hardware River Wildlife Area	James River Basin
HardR-4-5	HardW.4	37.74	-78.41	77	Hardware River Wildlife Area	James River Basin
HighBT-3-1	HighBT.1	37.31	-78.39	98	High Bridge Trail State Park	James River Basin
HighBT-3-2	HighBT.1	37.31	-78.39	98	High Bridge Trail State Park	James River Basin
HighBT-4-1	HighBT.1	37.31	-78.39	98	High Bridge Trail State Park	James River Basin
HighBT-4-2	HighBT.1	37.31	-78.39	98	High Bridge Trail State Park	James River Basin
HighBT-4-3	HighBT.1	37.31	-78.39	98	High Bridge Trail State Park	James River Basin
JamesR-1-2	JamesR.1	37.55	-77.51	38	James River Park	James River Basin
JamesR-1-3	JamesR.1	37.55	-77.51	38	James River Park	James River Basin
JamesR-1-4	JamesR.1	37.55	-77.51	38	James River Park	James River Basin
JamesR-1-5	JamesR.1	37.55	-77.51	38	James River Park	James River Basin

JamesR-2-1	JamesR.2	37.55	-77.51	44	James River Park	James River Basin
JamesR-2-2	JamesR.2	37.55	-77.51	44	James River Park	James River Basin
JamesR-2-3	JamesR.2	37.55	-77.51	44	James River Park	James River Basin
JamesR-2-4	JamesR.2	37.55	-77.51	44	James River Park	James River Basin
JamesR-2-5	JamesR.2	37.55	-77.51	44	James River Park	James River Basin
NatB-2-1	NatB.2	37.63	-79.55	303	Natural Bridge	James River Basin
NatB-3-1	NatB.3	37.63	-79.55	301	Natural Bridge	James River Basin
NatTun-1-1	NatTun.1	36.70	-82.74	518	Natural Tunnel State Park	Holston River Basin
NatTun-1-3	NatTun.1	36.70	-82.74	518	Natural Tunnel State Park	Holston River Basin
NatTun-1-4	NatTun.1	36.70	-82.74	518	Natural Tunnel State Park	Holston River Basin
NatTun-1-5	NatTun.1	36.70	-82.74	518	Natural Tunnel State Park	Holston River Basin
NatTun-2-1	NatTun.2	36.71	-82.74	536	Natural Tunnel State Park	Holston River Basin
NatTun-2-2	NatTun.2	36.71	-82.74	536	Natural Tunnel State Park	Holston River Basin
NatTun-2-3	NatTun.2	36.71	-82.74	536	Natural Tunnel State Park	Holston River Basin
NatTun-2-4	NatTun.2	36.71	-82.74	536	Natural Tunnel State Park	Holston River Basin
NatTun-2-5	NatTun.2	36.71	-82.74	536	Natural Tunnel State Park	Holston River Basin
NatTun-3-1	NatTun.3	36.70	-82.74	513	Natural Tunnel State Park	Holston River Basin
NatTun-3-2	NatTun.3	36.70	-82.74	513	Natural Tunnel State Park	Holston River Basin
NatTun-3-3	NatTun.3	36.70	-82.74	513	Natural Tunnel State Park	Holston River Basin
NatTun-3-4	NatTun.3	36.70	-82.74	513	Natural Tunnel State Park	Holston River Basin
NatTun-3-5	NatTun.3	36.70	-82.74	513	Natural Tunnel State Park	Holston River Basin
StartP-1-1	StartP.1	37.59	-79.39	204	Start Point	James River Basin
StartP-1-2	StartP.1	37.59	-79.39	204	Start Point	James River Basin
StartP-1-3	StartP.1	37.59	-79.39	204	Start Point	James River Basin
StartP-1-4	StartP.1	37.59	-79.39	204	Start Point	James River Basin
StartP-1-5	StartP.1	37.59	-79.39	204	Start Point	James River Basin
StartP-2-1	StartP.2	37.59	-79.39	209	Start Point	James River Basin
StartP-2-2	StartP.2	37.59	-79.39	209	Start Point	James River Basin
StartP-2-3	StartP.2	37.59	-79.39	209	Start Point	James River Basin
StartP-2-4	StartP.2	37.59	-79.39	209	Start Point	James River Basin
StartP-2-5	StartP.2	37.59	-79.39	209	Start Point	James River Basin
StartP-3-1	StartP.3	37.60	-79.39	215	Start Point	James River Basin
StartP-3-2	StartP.3	37.60	-79.39	211	Start Point	James River Basin
StartP-3-3	StartP.3	37.60	-79.39	215	Start Point	James River Basin
StartP-3-4	StartP.3	37.60	-79.39	214	Start Point	James River Basin
TexasB-3-1	TexasB.3	37.53	-77.47	18	Texas Beach	James River Basin
TexasB-3-2	TexasB.3	37.53	-77.47	18	Texas Beach	James River Basin
TexasB-3-3	TexasB.3	37.53	-77.47	18	Texas Beach	James River Basin
TexasB-3-4	TexasB.3	37.53	-77.47	18	Texas Beach	James River Basin
TexasB-3-5	TexasB.3	37.53	-77.47	18	Texas Beach	James River Basin
TexasB-4-1	TexasB.4	37.53	-77.47	16	Texas Beach	James River Basin
TexasB-4-3	TexasB.4	37.53	-77.47	16	Texas Beach	James River Basin
TexasB-4-4	TexasB.4	37.53	-77.47	16	Texas Beach	James River Basin

TexasB-4-5	TexasB.4	37.53	-77.47	16	Texas Beach	James River Basin
TwinL-1-1	TwinL.1	37.17	-78.28	153	Twin Lakes State Park	James River Basin
TwinL-1-2	TwinL.1	37.17	-78.28	153	Twin Lakes State Park	James River Basin
TwinL-1-3	TwinL.1	37.17	-78.28	153	Twin Lakes State Park	James River Basin
TwinL-1-4	TwinL.1	37.17	-78.28	153	Twin Lakes State Park	James River Basin
TwinL-1-5	TwinL.1	37.17	-78.28	153	Twin Lakes State Park	James River Basin
YorkR-1-1	YorkR.1	37.41	-76.72	22	York River State Park	York River Basin
YorkR-1-2	YorkR.1	37.41	-76.72	22	York River State Park	York River Basin
YorkR-1-3	YorkR.1	37.41	-76.72	22	York River State Park	York River Basin
YorkR-1-4	YorkR.1	37.41	-76.72	22	York River State Park	York River Basin
YorkR-1-5	YorkR.1	37.41	-76.72	22	York River State Park	York River Basin

Read processing, mapping, filtering and variant calling

Raw GBS reads were first de-multiplexed using GBSX_v1.3 (Herten et al., 2015). The raw reads were next processed with FASTQ_MCF v1.05 (Aronesty, 2013) to remove Illumina adapters, and low quality, and/or short reads. A minimum phred-scaled quality score of 30 and minimum read length of 50 bases was applied to all reads. The reference assembly was then indexed using the Burrows-Wheeler Alignment Tool (BWA) v.0.7.17-r1188t (Li, 2013) prior to mapping. The processed reads were mapped to the indexed reference genome using BWA default parameters for all options with the exception of a seed length of 24 bp to improve mapping quality scores. The BWA mapped reads were output in an unsort SAM format which were then sorted and converted the binary form, BAM, using SAMTOOLS.v1.7 (Li et al., 2009). Once sorted, the bam files were merged into a single bam file with BAMADDRG (<https://github.com/ekg/bamaddrg>). Variants were called with FREEBAYES.v.1.3.1-16-g85d7bfc (Garrison and Marth, 2012) using a custom script to use multiple threads and increase variant calling speed; MultiThreadFree-Bayes, (<https://github.com/aubombarely/GenoToolBox/tree/master/SNPTools/MultiThreadFreeBayes>). The resulting variant file (VCF) output was filtered with VCFTOOLS v0.1.15 (Danecek et al., 2011) using the following parameters; retain only biallelic SNPs, remove indels, a minimum read depth of 5 with a minimum mean depth of 20, a minimum SNP QC of 30, no missing data in any sample (--max-missing 1), a MAF of 0.05. After filtering with VCFTOOLS the remaining variants were filtered for linkage disequilibrium (LD) with PLINK.v.1.90b4 (Purcell et al., 2007). SNPs in LD were selected and removed based on a LD using independent pairwise filtering with a 10 Kb window, a variant shift count of 5 and r^2 value of 0.2. Variations on several of these parameters were tested during filtering to assess the performance while retaining the most samples, after which 9 samples were removed due to high levels of missing data. Number of reads mapped, sites, and variants called are listed in Table S2-2.

Clonal correction and population structure inference

Population genomic analyses were carried out in R studio with R programming language version 4.1.2, all R scripts used in the following methodologies can be accessed at <http://githubpagedetails.com>.

Because pawpaw has been shown to propagate vegetatively, clonality was tested across the sample group, in particular the overrepresentation of multilocus genotypes (MLG) was evaluated. To do this, the VCFTOOLS option `--relatedness2` which infers a pairwise probability of relatedness between samples in the vcf file using the KING relationship inference algorithm (Manichaikul et al., 2010) and produces a phi score between 0 and 0.5 for each pairwise comparisons between all samples. This output was used to produce a matrix of clonality in R with the package GGLOT2.v.3.3.5 (Wickham, 2016) [Figure S2-1] and identify potential clones. This was cross validated using the R package POPPR.v.2.9.3 (Kamvar et al., 2014) function `clonecorrect()` which attempts to identify and remove duplicated MLG.

Inference of population structure was carried out using the r package LEA.v.3.6.0 (Frichot and François, 2014). LEA function `snmf()` performs a Bayesian clustering on MLG data with FASTSTRUCTURE (Raj et al., 2014), a faster, more resource efficient method of performing STRUCTURE (Pritchard et al., 2000). This method assumes Hardy-Weinberg equilibrium (HWE) and linkage equilibrium between loci within population. LEA estimated admixture coefficients to produce STRUCTURE-like outputs and infer the mostly likely genetic clusters based on allele frequency and clustering probability using sparse Non-Negative Matrix Factorization algorithms meaning overrepresentation of MLG can influence results. Estimated population admixture at suggested K values from LEA was compared to ADMIXTURE.v.1.3.0 (Alexander et al., 2009) run with 5-fold cross-validation.

Additionally, population structure was inferred using non-model based principal component analysis (PCA) and discriminant analysis of principal components (DAPC) implemented in the ADEGENET.v.2.1.5 (Jombart et al., 2010; Jombart and Ahmed, 2011) package. Both were used as alternatives to the Bayesian approach of STRUCTURE as they do not make any prior assumptions about the population which are not applicable to a clonally reproducing species. PCA relies on genetic distance to form clusters and summarize the variation between and within clusters. DAPC uses sequential K-means and model selection to infer genetic clusters and has a greater focus on summarizing between cluster variation. DAPC function *optim.a.score()* was used to choose the appropriate number of PCs to analyse for maximum variance while avoiding an overfit. DAPC allows for best subpopulation assignment using Bayesian Information Criterion (BIC).

A Neighbor-joining (NJ) tree based on allele frequencies evaluated with the R package POPPR and visualized using the package APE.v.5.6-1 (Paradis and Schliep, 2019). The dendrogram was produced from obtained genetic distances calculated using the fraction of different sites between samples with *bitwise.dist()* and run with 100 bootstrapping support. In addition, the pairwise distances between haplotypes were used to construct a distance matrix for minimum spanning network (MSN) analysis, to visualize the relationship between individuals. This was also implemented in POPPR.

Analysis of genetic variance

An evaluation of the variation in the population clusters, identified using the above methods of structure inference, was performed using analysis of molecular variance (AMOVA) (Excoffier et al., 1992) as implemented in the package POPPR (Kamvar et al., 2014). The analysis was run with the distance matrix produced from the VCF file and a table partitioning the data into

different stratifications. AMOVA was carried out with the clone correction option to avoid the influence of potential clones in the dataset. Analysis was validated using the function *randtest()* with 999 permutations in the ADE4.v.1.7-18 package (Dray and Dufour, 2007) to estimate strata variation significance.

To ascertain whether geographic separation has a biologically significant role on population structure an isolation-by-distance (IBD) analysis was carried out. This was achieved by generating geographic distances and genetic distance matrices using ADEGENET function *dist.genpop()* and performing a Mantel test with the R package ADE4.v.1.7-18 (Dray and Dufour, 2007). IBD was then run with 999 permutations and simulated p-value of 0.001 in order to explore the effects of genetic drift within identified clusters.

F-statistics of genetic diversity estimates were run with the R package POPGENOME.v.2.7.5 (Pfeifer et al., 2014) including, number of segregating sites, nucleotide diversity, Watterson's theta and Tajima's D. Heterozygosity, fixation index (F_{ST}) and inbreeding coefficients (F_{IS}) were estimated for assigned populations using DARTR.v.2.0.3 (Gruber et al., 2018). For cross comparison, heterozygosity was additionally calculated directly from bam files to include all sites using ANGSD.v.0.940 (Korneliussen et al., 2014).

Results

Pawpaw draft genome assembly and SNP calling

A k-mer size estimation with genomescope (Vurture et al., 2017) estimated the genome size to around 1.04 Gb with heterozygosity of 0.4, with flow cytometry estimating 0.98 Gb. GenomeScope analysis indicated heterozygosity was 0.04% with a repeat content of 51.42%.

The draft *A. triloba* assembly (Astri041) was 0.84 Gb. This was generated using CANU v2.2 (Koren et al., 2017) with PacBio Sequel sequencing for the initial assembly and using Illumina HiSeq2500 reads to polish the consensus. The quality and completeness of the assembly was assessed prior to read mapping. This step involved four metrics [**Error! Reference source not found.**], contiguity stats, gene space completeness with BUSCO v5, a k-mer completeness assessment using Merqury [Supplemental Figure S2-2] and assessment of repetitive sequences with LAI. Total assembly size is 845,748,466 bp. The assembly is composed of 8,300 scaffolds with an L50 of 4, and N50 of 432,660 bp. Gene space completeness of the final assembly was estimated using the BUSCO eudicot_db10 dataset containing a set of 2,326 ancestral eudicot specific genes. BUSCO analysis provides an estimation of assembly quality by looking for the presence or absence of these genes in the assembly. 92.3% of core genes were present in the assembly where 3.8% were duplicated, 3.1% fragmented and 4.6% missing. Finally, the k-mer-based assessment tool Merqury v1.3 (<https://github.com/marbl/merqury>), was used to map the Illumina short reads back to the completed reference genome to estimate completeness (93.6%) and quality (35.4). LAI score was of 11.4.

Table 2-2 Draft assembly assessment

Table shows the results of the four completeness and assembly quality metrics (assembly stats, genespace completeness, k-mer mapping, and LAI used to evaluate the Astri041 genome.

Assembly stats	Astri041
Assembly size (Gb)	0.84
Scaffolds	8,300
Longest seq (Mb)	7.15
Shortest seq (bp)	222
Average seq length (Mb)	0.10
L90, number of seq	2,334
N90 (Mb)	0.07
L50, number of seq	23
N50 (Mb)	432.7
% BUSCO complete	92.30
% BUSCO duplicated	3.80
% BUSCO fragmented	3.10
% BUSCO missing	4.60
Merqury Completeness	93.60
Merqury QV	35.40
Merqury Error	2.89E-04
LAI index	11.43

After assessment of the draft assembly, the raw reads were processed and mapped to the indexed Astri041 assembly. An average of 9.5 million reads (SD = 5.7, Max = 31.5, Min = 1.5) were mapped to the reference genome from each of the 134 GBS samples. Of those, an average of 7.4 million (81%, SD = 4.2, Max = 22.6, Min = 1.3) correctly mapped to the Astri041 reference genome. The total number of sites in the merged bam file was 504,887. Reduced representation methods such as GBS can result in large amounts of missing data across samples (Elshire et al., 2011); thus, during the filtering of SNPs, a cut-off of 75% missing data was initially set to remove samples with the highest amount of missing sites while retaining the greatest number of usable samples with an appropriate number of SNPs for analysis. The number of informative SNPs required while maintaining as many individuals as possible was evaluated by performing FASTSTRUCTURE; PCA and NJ-tree analysis on various SNP datasets were generated from applying a variety of filtering parameters and evaluating the point where increasingly strict cut offs stopped affecting the results of these preliminary tests. This resulted in the removal of 9 low quality samples which were missing 75% of possible sites. Once these were removed, no missing data was allowed in the remaining samples, keeping only sites shared across all remaining samples. This resulted in retention of 9,142 SNPs. To reduce likelihood of any association between SNPs, thinning of these variants was achieved by pruning SNPs that appeared to be in LD using independent pairwise filtering; this final filtering left 4,845 purportedly independent SNPs for population analysis. Supplementary table S2-2 contains a summary of raw data processing, mapping, and variant calling. After final filtering, the remaining 124 individuals represented pawpaw patches from the eastern to western borders of Virginia, with a distance of approximately 542 km between the two most distant sample locations.

Presence of clones in the dataset

The KING relationship inference algorithm revealed (Manichaikul et al., 2010) that several samples had high Phi scores of 0.4–0.5 [Figure S2-1], confirming the presence of clones in our dataset. Cross-validation with the ADEGENT package clonal correction function found that none of the MLGs were similar enough to remove them from the sample data. In any case, all 124 individuals and 4,845 SNPs were retained in the analysis, but to minimize any possible effects, corrections were made when possible, and care was taken when inferring biological and demographic meaning by using multiple analysis. Supplemental Table S2-3 contains a list of individuals per patch with potential clones, based on VCFTOOLS's KING relatedness estimation.

Population structure

An initial inference of population structure was obtained by performing STRUCTURE analysis. In the Supplemental Figure S2-3a, the cross-entropy ranged from 0.7 to the lowest value of 0.39 at $K = 24$, dropping quickly and plateauing around 0.40. Other possible K values were observed at less distinct minima of $K = 10$ (cross-entropy: 0.48), $K=15$ (cross-entropy: 0.42), and $K = 27$ (cross-entropy: 0.39). The level of admixture between individuals were estimated using two methods, the first was an ADMIXTURE-like output from FASTSTRUCTURE and then the second a cross-validation using the ADMIXTURE algorithm by Alexander, Novembre, and Lange (2009). ADMIXTURE-like ancestry assignment was run for $K = 24$, the lowest cross-entropy value (0.39) [Figure S2-3c]. Next cross-validation with ADMIXTURE was estimated for lower K values (2–9) [Figure S2-4]; with this approach, there was high level of admixture across the whole sample population.

Principal components analysis revealed a more heterogenous pattern, with the first three PC axes (PC1, PC2 & PC3) accounting for only 15% the total variance. There was a clear separation of individuals matching their geographic distribution laterally across Virginia and, perhaps more interestingly, PC2 (5% of total variance) and PC3 (4.5% of total variance), appeared to be clustering of individuals from the same river basin [Figure 2-1b]. These were the Holston River Basin (HSB), the New River Basin (NRB), James River Basin (JRB), and the York River Basin (YRB). Additionally, individuals from along the James River clustered more tightly together than with others geographically closers, despite being as much as ~190 km straight-line or ~285 km river flow distance between the furthest samples along the river [Figure 2-1a].

To explore this further, James River patches and two outgroups (Devil's bathtub, and York River) were subset and another PCA was performed on these individuals [Figure 2-2b]. In this case there was an even clearer clustering of the James River associated patches, with a single patch from the James River Park clustering more closely to the York River patch ~67 km away and not on the same river.

Subsequently, DAPC was run retaining 100 axes in the discriminant analysis and the Bayesian Information Criterion (BIC) indicated the lowest value was at 28 clusters (BIC range of 600 to 225) [Figure S2-3b].

Next, DAPC was run to assess the clusters observed in the PCA. River basins appeared to have the strongest influence on population structure, this was assayed by running the DAPC on assigned river basin groups retaining six PCs. This number was suggested as the optimal number of PCs to retain in order to maximize variance, by running the ADEGENET function *optim.a.score()*. DAPC clustering showed a distinct clustering by river basin [Figure 2-1c]. DAPCs posterior membership was also performed, running K clustering of K = 4–7 [Figure

S2-6]. The posterior membership assignment at $K = 4$ showed that HRB, JRB, and YRB could form a single cluster. But at $K = 5-7$, we could see clear separation of each river basin. However, in all K clusters we could see that JRB and YRB are single group and the NRB is the most diverse group.

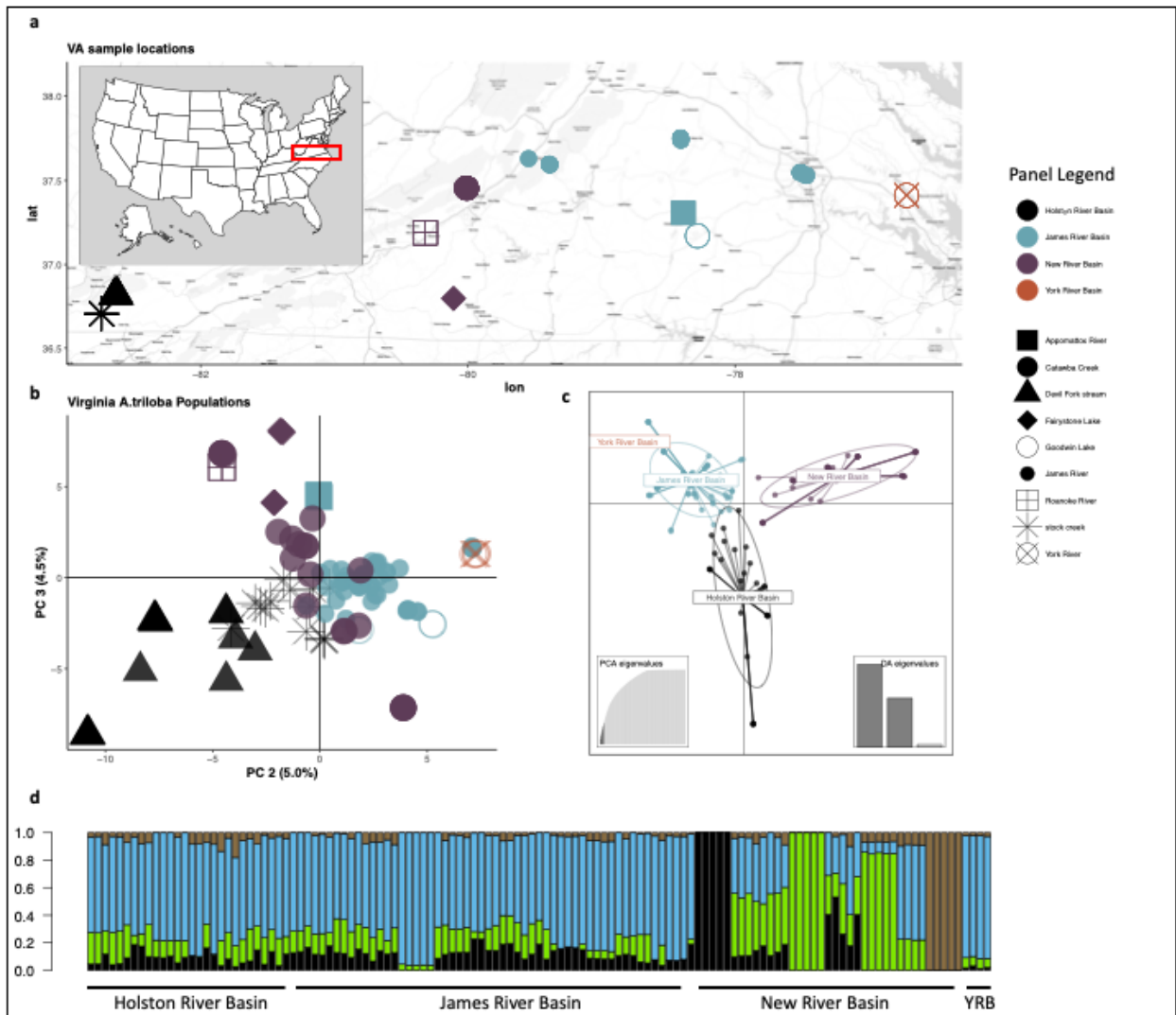


Figure 2-1 Virginia population structure

Virginia population structure. Panel showing sampling location of Virginia pawpaw sample patch locations and estimates of population structure. **a)** Map of sample locations from across the state of Virginia. River Basin is denoted by colour and the river/creek that each patch is located on or near is indicated by shape (panel legend). **b)** Principal component analysis (PCA) of individual samples. Figure shows PC 2 and 3 accounting for 9.5% of total variation. Individuals are labelled in the same manner as 1a. **c)** DAPC scatterplot showing clustering of individuals. **d)** Admixture of individuals a $K = 4$. As estimated by FASTSTRUCTURE. YRB: York River Basin.

The genetic relationship of Virginia haplotypes was further investigated by generating a Neighbor-joining (NJ) tree using a Nei's (Nei, 1972) distance matrix and Euclidian distances matrix. The tree was run with 100 bootstrapping replicates. The NJ tree [Figure 2-3] was comprised of three major clades that, apart from four individuals from the Natural Tunnel area, in Southwestern Virginia, broadly resembled the geographic sampling locations by river basin, with NRB samples nested inside the JRB and individual patches clustering together. While the bootstrapping support for the patches grouping was >70 , the large clades at base of the tree were not all supported with bootstrapping values of <50 in all cases.

A complementary approach to NJ trees, minimum spanning network (MSN), was also generated using the POPPR package with a distance matrix of dissimilarity and Euclidean distances [Figure S2-5]. This network analysis indicated the lowest distances between many of the same individuals purported clones identified using the KING relationship inference algorithm. The MSN linked individuals by their patch with a distance of 0.002 but did not separate the patches into any distinct clusters. Instead, there was high relatedness between almost all samples.

Isolation-By-Distance (IBD) with Mantel test was run with Nei's distances (Nei, 1972) and Euclidean geographic distances of individual samples locations, with a simulated p-value of 0.001. In the scatterplot [Figure 2-4a], we saw distinct patches of discontinuities indicating the presence of an impact of distance on geneflow.

Hierarchical analysis of genetic diversity was carried out by running an AMOVA to detect and analyse the molecular variance within and among populations independent of HWE assumptions (Meirmans, 2006). The AMOVA was run with Monte-Carlo significance testing and clone correction (detailed results can be found in Table S2-4), and AMOVA significance

testing [Figure 2-4b]. The results indicated that 0.9% of the variation was found between subpopulations, 106% was within samples, and -6.7% was between samples with subpopulation. The negative number could suggest that no population structure was present between samples, or alternatively that the variation within groups was much greater than the variation among groups (Meirmans, 2006).

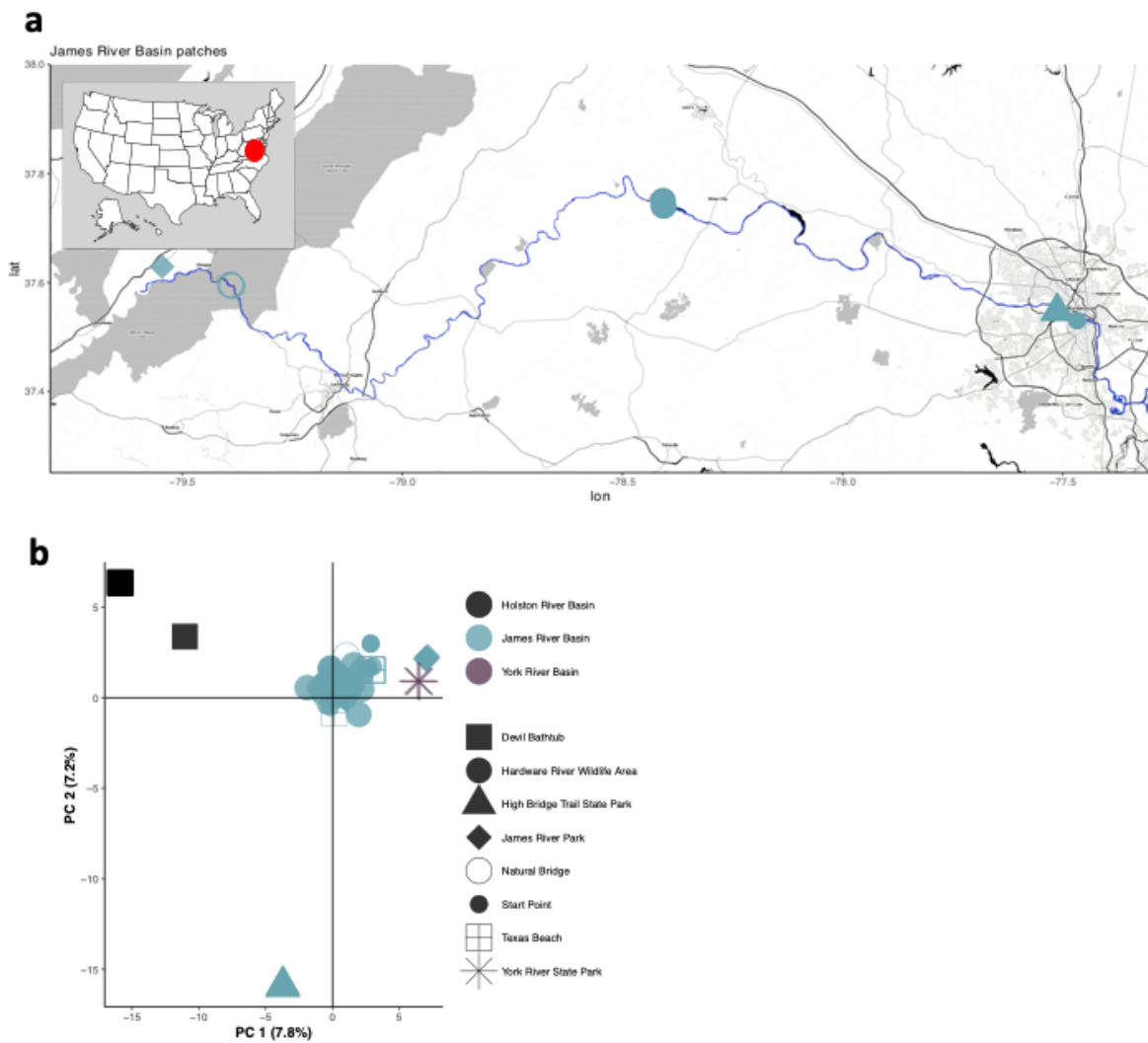


Figure 2-2 James River Basin population

a) Map indicating sample locations from along the James River (highlighted in dark blue). Only sample sites along the river are indicated, other sites not on the river or sample sites used as out groups from other river basins, e.g., Devil's Bathub and York are shown on this map. **b)** Principal component analysis (PCA) clustering analysis of individuals from the entire JRB and representatives of the HRB and YRB as out groups.

Genetic diversity

The heterozygosity of the Virginia population was calculated with two approaches, DARTR and Analysis of Next Generation Sequencing Data (ANGSD). Key differences between the two methods are that DARTR uses only the variants remaining after filtering to estimate the genotype likelihoods for each individual at each site, which is then used to calculate the observed heterozygosity, while ANGSd uses all possible variant and invariant sites from raw data, applying a Bayesian method to estimate the genotype probabilities for each individual at each site. First, using the package DARTR to evaluate the genetic variance in the total population the observed heterozygosity (H_o) 0.32 was higher than the HWE expected (H_e) 0.26 [Table 2-3].

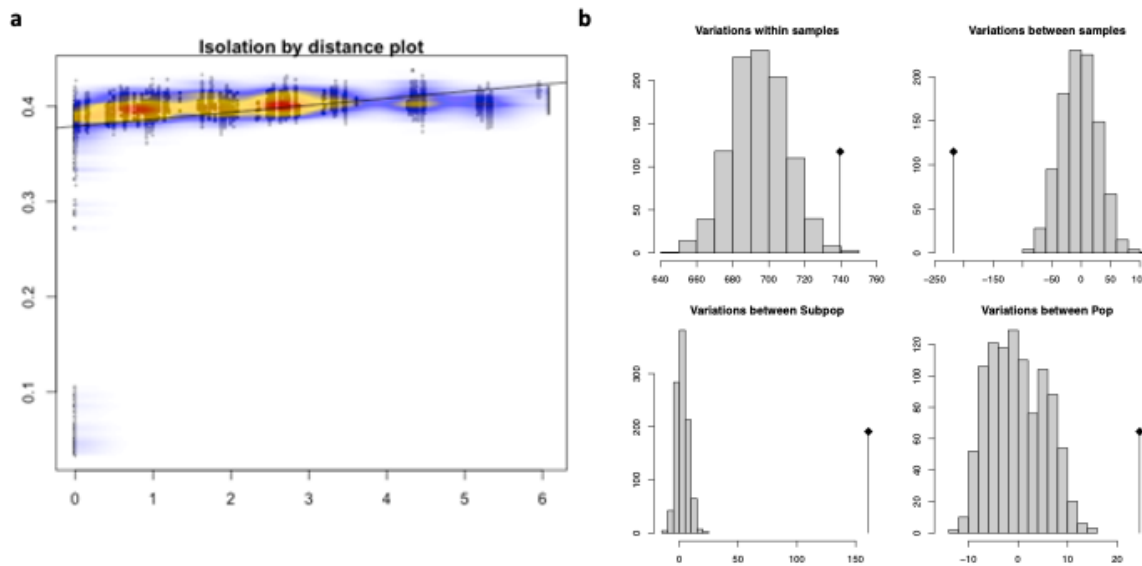


Figure 2-4 Isolation-by-distance and AMOVA results

Evaluation of genetic diversity and population interactions **a)** Cloud plot of Mantel tested Isolation-By-Distance in the total sample population. Simulated p-value: 0.001. **b)** Histograms of AMOVA estimates of variation contribution to total population diversity. AMOVA was run with 999 permutations and individuals assigned to random strata to avoid influence of clones. Black dot indicates observed variation while histogram indicated expected values.

In a cross validation using ANSGD, the observed heterozygosity (H_o) was 0.07 [Table 2-3]. The observed heterozygosity at river and river basin level was explored using DARTR as they appeared to be important factors in defining subpopulations. The observed (0.30–0.35) was higher than expected (0.15–0.29) little to no variation in the H_o across all assigned groups [Table 2-3]. To further understand the scale of variation between river basin sub-populations, segregating sites and Tajima’s D values were estimated. The number of segregating sites was between 4,300–4,543 (1,467 for YRB) with a nucleotide diversity range of 1,315–1,364 (813 for YRB) and, for all groups, Tajima’s D values were all in the range of 2 indicating a population undergoing balancing selection [Table 2-3]. A final comparison of these subpopulations was carried out via F_{ST} . The values when comparing the river basin groups ranged from 0.05 to 0.2 [Table 2-3], indicating little to no variation between groups.

Table 2-3 Genetic diversity

Summary table of genetic diversity analyses. Table contains results from several summary diversity statistics. All analyses were performed with individuals assigned to river basin group: New River Basin (NRB), Holston River Basin (HRB), James River Basin (JRB), York River Basin (YRB). Heterozygosity estimated with river, and river basin groups using DARTR. Segregating sites, nucleotide diversity, Tajima's D, and Watterson's Theta, estimated using R packages POPGENOME. Pairwise F_{ST} estimations carried out using R package DARTR.

R Package	n.biallelic sites	4845			
ANGSD	Ho	0.07			
DARTR	RIVER BASIN	AVG Ho	AVG He	n.IDV	
	NYB	0.32	0.27	37	
	HRB	0.33	0.18	28	
	JRB	0.32	0.27	56	
	YRB	0.30	0.24	4	
	RIVER	AVG Ho	n.IDV	n.IDV	
	Roanoke River	0.31	0.16	5	
	Catawba Creek	0.32	0.27	23	
	Goodwin Lake	0.32	0.24	5	
	Devil Fork stream	0.33	0.27	14	
	Appomattox River	0.35	0.18	5	
	Fairystone Lake	0.33	0.20	9	
	York River	0.30	0.16	4	
James River	0.32	0.29	46		
stock creek	0.32	0.29	14		
POPGENOME	Group	<i>NYB</i>	<i>HRB</i>	<i>JRB</i>	<i>YRB</i>
	n. segregating sites	4252	4300	4543	1467
	nucleotide diversity	1315.90	1350.06	1364.08	813.04
	Tajima's D	1.79	1.60	1.99	2.40
	Watterson's Theta	872.29	936.08	858.59	565.79
DARTR	Group Fst	<i>NRB</i>	<i>HRB</i>	<i>JRB</i>	<i>YRB</i>
	<i>NRB</i>	NA	0.06	0.05	0.20
	<i>HRB</i>	0.06	NA	0.05	0.20
	<i>JRB</i>	0.05	0.05	NA	0.16
	<i>YRB</i>	0.20	0.20	0.16	NA
DARTR	Ho	0.32			
	Hs	0.26			
	Ht	0.30			
	Dst	0.03			
	Htp	0.31			
	Dstp	0.04			
	F_{ST}	0.11			
	F_{STP}	0.14			
	F_{IS}	-0.20			
Dest	0.06				

Discussion

In this study, the genetic diversity of *A. triloba* trees from the state of Virginia, were assessed on a population level using a GBS approach mapping the sequencing reads to the first draft assembly of the species. Our results indicated three main features of the Virginia pawpaw population: (i) it is highly panmictic; (ii) geneflow is increased along river and streams; and (iii) it contains high levels of observed heterozygosity and a weak pattern of isolation by distance. We show that not only are pawpaw trees along a shared river more closely related despite hundreds of kilometres of separation, but that pawpaw trees found within a river basin share higher genetic similarity than those in a separate river network. Our study provides the first evidence of rivers and waterways as preferential corridors for geneflow in *A. triloba*, likely due to hydrochory.

Our *de novo* reference genome Astri041 represents the first draft assembly for the species *Asimina triloba*. It is slightly less than the estimated 1 Gb and is missing only around 4% of 2,326 expected ancestral eudicot genes. BUSCO completion scores indicating 92.3% percentage of core gene were correctly assembled, which is sufficient for variant based studies.

Clonality appears common among the sampled population, as could be expected for the species (Willson and Schemske, 1980). However, despite efforts to avoid sampling multiple ramets (clones) from genets (clonal patch), identical or near identical MLGs were present in all but three of the wild populations collected, indicating multiple sampling of single genets, some of which may be over 10 meters in diameter. Studies of clonally propagating species are prone to erroneous inferences on the genetic structure of their populations and linked to assumptions of low diversity, likely due the use of low-power markers lacking sufficient resolution (Arnaud-

Haond et al., 2005; Hamrick and Godt, 1990). Many of the methods used can be affected by high levels of clonality by overrepresentation of MLGs, resulting in error inducted by linkage disequilibrium (LD) in some haplotypes and erroneous allele frequencies. Analysis of populations with mixed clonal/sexual reproduction seems to have negligible effects on most genetic diversity analytics (Balloux et al., 2003). However, to minimize any possible effects, corrections were made when possible, and care was taken when inferring biological and demographic meaning.

Inference of population structure

Understanding population structure is key to identifying the barriers or the preferred pathways for gene flow in a species. Moreover, a population's structure can help to reveal the interaction between geneflow and genetic drift and also inform conservation efforts by highlighting genetically diverse subpopulations (Cushman et al., 2014; Grady et al., 2011). In our preliminary analysis of the Virginia state population with FASTSTRUCTURE, there were an estimated 24 clusters based on the approximated individual ancestry coefficients and allele frequencies, $K = 24$ [Figure S2-3a]. Similarly, the lowest BIC value was observed at 28 [Figure S2-3b], both methods estimating a value at, or close to, the number of sample sites, which may be due to the occurrence of a limited number of multilocus genotypes present at some of the sites causing the grouping by sample sites. Another possible reason could be a lack of genetic diversity across the state with a limited genetic variation among sample patches. It is highly unlikely that these results represent the true number of ancestral populations, as pawpaw is an outcrossing species. The lack of clear structure may also be related to high levels of geneflow across the state with genetic drift applying equally to individuals (Lawson et al., 2018). Indeed, this is somewhat supported by a previous study of *A. triloba*'s using microsatellites sampling across the entire native range (Wyatt et al., 2020), where the authors identified a lack of genetic

structuring and suggested there were two populations with high levels of admixture in each. These two populations emerged as the species migrated north from the now submerged Gulf of Mexico, separating into east and west populations by the Appalachian Mountains as pawpaw migrated north. Our sample sites in the state of Virginia fit well within the previously identified eastern population and there may not have been enough time since to accumulate significant levels variation within the population.

We also attempted to estimate structure via NJ tree [Figure 2-3]. This produced a tree with little to no support for the lowest branches and well supported upper branches with individuals largely grouping by patch. This could indicate high levels of geneflow in the population or that the population has only recently undergone an expansion. The minimum spanning network analysis resulted in a compact tree network of closely related individuals, with no strong separation of any of the individuals or groups from any other, thus supporting the notion of high levels of geneflow across the state. This supposes a freely inbreeding (panmictic) population, similar to the conclusions of Huang (1998). Here, the authors used an allozyme analysis from nine states and identified high (72%) levels of within population genetic variation, with little to no variation seen among populations indicating a freely breeding population. High variation within populations can be expected in insect pollinated species with fragmented distribution as seen in pawpaw (Hamrick et al., 1992). Indeed, this was the case in wild service tree (*Sorbus torminalis*), a species that, like pawpaw, is an obligate out crossing species capable of clonal reproduction by clonal suckers; the majority of variation (61.78%) was found to be within populations (Belletti et al., 2008), likely a result of a small pollination range from the flies not spanning multiple patches.

The results of the analysis of the molecular variance (AMOVA) showed significant variation between river basin populations and between subpopulations of sample patches [Figure2-4b].

High variance between pawpaw patches in different river basins may indicate that there is some resistance to gene flow when individuals are under the influence of different river networks and water levels (Chaput-Bardy et al., 2009; Cushman et al., 2014).

River populations

In addition to the above methods, clustering was performed using PCA and DAPC. Using these approaches, it was possible to see clustering of individuals by geographical features and distance. The analysis of principal components clustered by patches and by proximity to rivers. However, there was little genetic diversity to delineate between the clusters. Indeed, the variance captured in PCs 1, 2, and 3 only accounted for around 15% of the total variation present. Interestingly, we observed three patterns of clustering; first, along PC 2 on the y axis [Figure 2-1b], a distinct pattern of the separation of patches east to west was identified. Secondly, we observed the clustering of individuals by river. In particular, sample sites from five locations (Hardware River wildlife area, James River Park, Natural Bridge, Start Point, and Texas beach) which are all on the James River or a tributary feeding into the river, clustered closely together. This is a distance of around 190 km “as the crow flies” or ~285 km following the flow of the river [Figure 2-2a]. This is interesting because it is thought that bears, raccoons, or even humans might have replaced large extinct mega faunas role in the dispersal of pawpaw seeds, and while there is evidence of human influence on distribution (Wyatt et al., 2021), here we show evidence of rivers as a significant vehicle for geneflow in pawpaw. Potentially, this could come from the emergence of new ramets from seeds floating down river (hydrochory), crossing with existing trees in new locations, leading to increased genetic similarity observed along a waterway and indeed across the entire Virginia population. This is often the case in riparian plant species, which are known to take advantage of this dispersal method (Berković et al., 2018; Levin et al., 2003; Lopez, 2001; Nilsson et al., 2010, 1991). The third patten of

clustering, was observed in individuals on a single river basin. Many of the rivers within these basins are not directly linked by an interconnected complex web of rivers and streams flowing to one main water body. Instead, they fragment the landscape and may be contributing to a slight decreased rate of geneflow between individuals (Chaput-Bardy et al., 2009; Cushman et al., 2014). The level of branching complexity of the river and angle of stream bifurcation has been shown to affect gene flow by altering the proximity of new individuals establishing downstream (Chaput-Bardy et al., 2009). Our genetic structuring analysis suggests that river topography is affecting gene flow in pawpaw, but is not strong enough to fully isolate patches. The clustering observed with DAPC [Figure 2-1c] and its posterior membership probability assignment [Figure S2-6] both strongly support the river basins as a factor influencing the genetic diversity in the Virginia population. In particular, the posterior membership probability assignment at $K = 4$ (the same number of river basins) [Figure S2-6] also makes clear there is a distinct genetic divergence between river basin groups in pawpaw. Though it is not immediately clear if we are seeing an increase in geneflow between individual members of a river basin, or a separation acting to decreased geneflow between patches outside the river basin. There may be decreased interaction between individuals because of the landscape fragmentation, leading to some weak isolation effect (Chaput-Bardy et al., 2009; Cushman et al., 2014). A similar pattern was observed in pawpaw where samples sites in close proximity, but on separate watersheds, had the greatest pairwise G_{st} values (Wyatt et al., 2020) and a separation of populations by a “watershed effect” was also reported in the species *Veratrum woodii* (Zomlefer et al., 2018).

Taking into account these river basin clustering (HRB, NRB, JRB, and YRB) seen in the PCA and DAPC, the ADMIXTURE-like analysis was run again with a K value = 4 [Figure 2-1d] corresponding to the number of river basin clusters. The shared ancestry was higher in populations which were either along a shared river source or physically close together as in the

case of the HRB samples. The NRB contained the most distinct and diverse populations. This is notable because although part of the same river basin, there is no one river connecting the patches as seen in the JRB. These results were compared to the output of ADMIXTURE algorithm, testing $K = 2:9$ [Figure S2-4]. Comparing the ancestry estimation of both methods at various K values, several patches (i.e., FairyS, AndysLT, HighBT and DevilB) were identified as being more distinct or isolated than others. The high bridge trail state park samples stand out within the JRB as being a highly distinct patch. In the absence of any obvious physical barrier or distance to individuals from the JRB or NRB, it is very possible that this patch was introduced from another area by either humans or other animals. Unfortunately, without samples from outside the state, we were not able to test this. However, it is clear that rivers play an important role in the genetic diversity of the species, and hydrochory has been suggested as a potential method of seed dispersal for pawpaw (Peterson, 1991). So, to explore this further, a subset containing only the individuals found along the James River (JR), a geographically close outgroup (YorkR), and a second geographically distant outgroup (DevilB) were analysed again by PCA. The second PCA carried out on the JR subpopulation showed an even tighter clustering of all individuals growing along the river [Figure 2-2b]. Although here it could also be seen that, according to PC 2 (7.2%), the High bridge trail state park (HighBT) patch is more distantly related to the rest of the JR subpopulation than even the more physically distant Devils bathtub samples. The HighBT patch is on the edge of the JRB boundary with the NRB and is the only sample site not near to, or directly on the banks of the James River. Indeed, this patch is approximately 40 km south of the JR, while the nearest patch at the Twin Lakes is around 60 km south and groups closely with the rest of the JRB patches, displaying the previously mentioned “watershed effect” (Wyatt et al., 2020; Zomlefer et al., 2018).

Genetic diversity

A Mantel test on Nei's distance (Nei, 1972) and Euclidean geographic distances found evidence of isolation by distance in our sample population [Figure 2-4a]. This stepping stone clustering seen in Figure 2-4a indicates that, while there is only a weak overall impact, there is evidence for decreased gene dispersal over distance (Wright, 1943). This is supported by the AMOVA results indicated a level distinction between individuals separated by river basins and patches. Although it might not be strictly a 'watershed effect', as the pollination method of pawpaw may leave it vulnerable to the isolation by distance. Pawpaw is thought to be pollinated by beetles and flies (Goodrich et al., 2006; Kral, 1960; Willson and Schemske, 1980). These are considered weak fliers which cannot transfer pollen over long distances. Furthermore, fruit set success in pawpaw can be very low, with as little as 0.41% success in the middle of its range (Lagrange and Tramer, 1985; Willson and Schemske, 1980), so transfer of seeds may be aided by rivers but happening at low frequency. In addition to this, clonality in wild populations is a common occurrence (Botkins et al., 2012; Pomper et al., 2009). It is therefore possible that the IBD is a result of the limited range of its pollinators, increasing local genetic similarity, or due to potential clones in our dataset.

The DARTR estimation of H_o 0.32 and H_e 0.26 is lower than the species-level genetic diversity reported in Wyatt (2020) (0.53). However, direct comparisons between such estimates are not possible given the different statistical software used; and even more problematic in such direct comparisons is the differences in quantity and biological qualities in the markers being used. For example, SNPs tend to be bi-allelic and can be found in non-coding regions in linkage equilibrium, while microsatellites are more polymorphic and found in coding regions (Tsykun et al., 2017). Further, both types of markers are affected differently by mutation rates and genetic drift (Fischer et al., 2017; Li et al., 2002). However, what is comparable is the

heterozygosity levels expected under HWE, versus the observed heterozygosity (H_o). This finding is in agreement with the earlier studies using microsatellites (Botkins et al., 2012; Wyatt et al., 2020) and RAPDs (Huang et al., 2000). The observed heterozygosity was seen to be consistent among the river basin groups and among the river groups and, in all comparisons, the H_o was higher than the H_e (Table 2–3). In all river basin groups, the H_o were around 0.32 compared to the H_e of around 0.24. For river groups the results were near identical with a H_o of 0.32 and H_e of 0.23 showing that the heterozygosity in all populations is maintained a much higher levels than expected under HWE. Furthermore, the positive Tajima’s D values in all river basin groups may be due to balancing selection taking place during sexual reproduction. Again, excluding the YRB because of low levels of sampling, F_{st} estimates between the river basin groups were 0.05– 0.06 [Table 2-3]. The genetic diversity within populations and genetic differentiation among populations can be influenced by many and varying factors including, demographic history, gene dispersal methods, and reproductive approaches (Hamrick et al., 1992; Hamrick and Godt, 1997, 1990). Both heterozygosity and allelic diversity at each locus are thought to increase during clonal reproduction (Balloux et al., 2003), and have been shown in *Ruta macrocarpa* to contribute to an overall increase in genetic diversity (Meloni et al., 2013). Conversely, clonal reproduction is thought to decrease the genetic diversity among populations (Balloux et al., 2003). The mixed reproductive approaches of *A. triloba* appear to have contributed to a population with high levels of heterozygosity but little among population diversity. It is important to keep in mind however, that we have only sampled from one state, representing a small fraction of the whole population, and genetic diversity on a local level might vary when looking over the full native range.

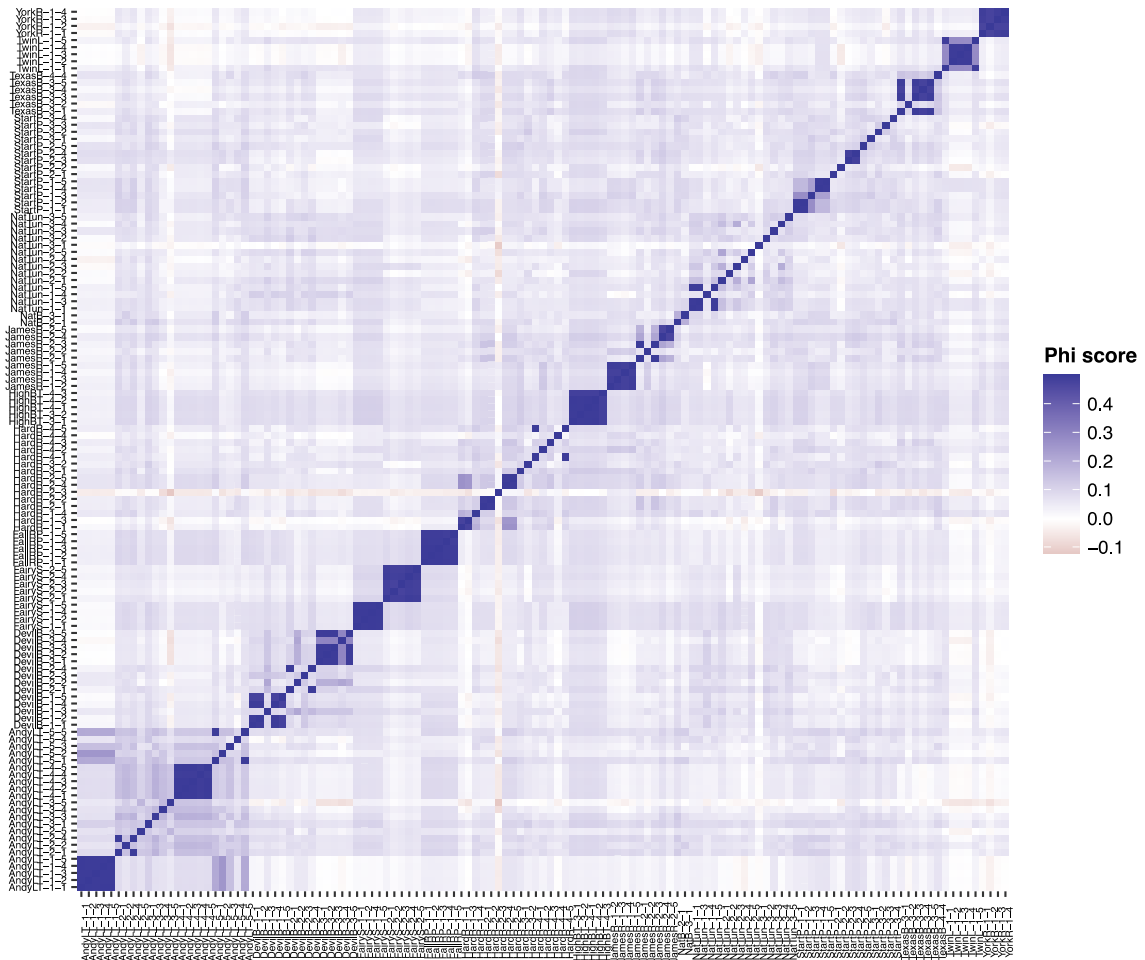
Conclusion

In conclusion, our study provides new and comprehensive insights into the genetic structure and diversity of the Virginia pawpaw population. By generating the first *de novo* genome assembly for this species using long and short read sequencing, we have established a valuable resource for future genetic and genomic studies. Our findings reveal that the heterozygosity of the Virginia population is much higher than expected under HWE, suggesting a balance of clonal reproduction and high levels of gene flow acting to maintain genetic variation post-Holocene expansion. However, we also observed a weak pattern of isolation by distance across the state, and clear evidence that river basins have a significant influence on population structure, suggesting that gene flow is, under certain conditions, restricted. Specifically, our results suggest that pawpaws growing physically close to each other are more likely to interbreed, with seeds moving longer distances across the state when facilitated by waterways. Additionally, the low genetic diversity and high panmixia observed in the sample population highlights the importance of conservation efforts for this species. Our research provides valuable insights into the genetic structure of this species and will inform future conservation, management strategies and aid in the development of breeding markers.

Acknowledgements

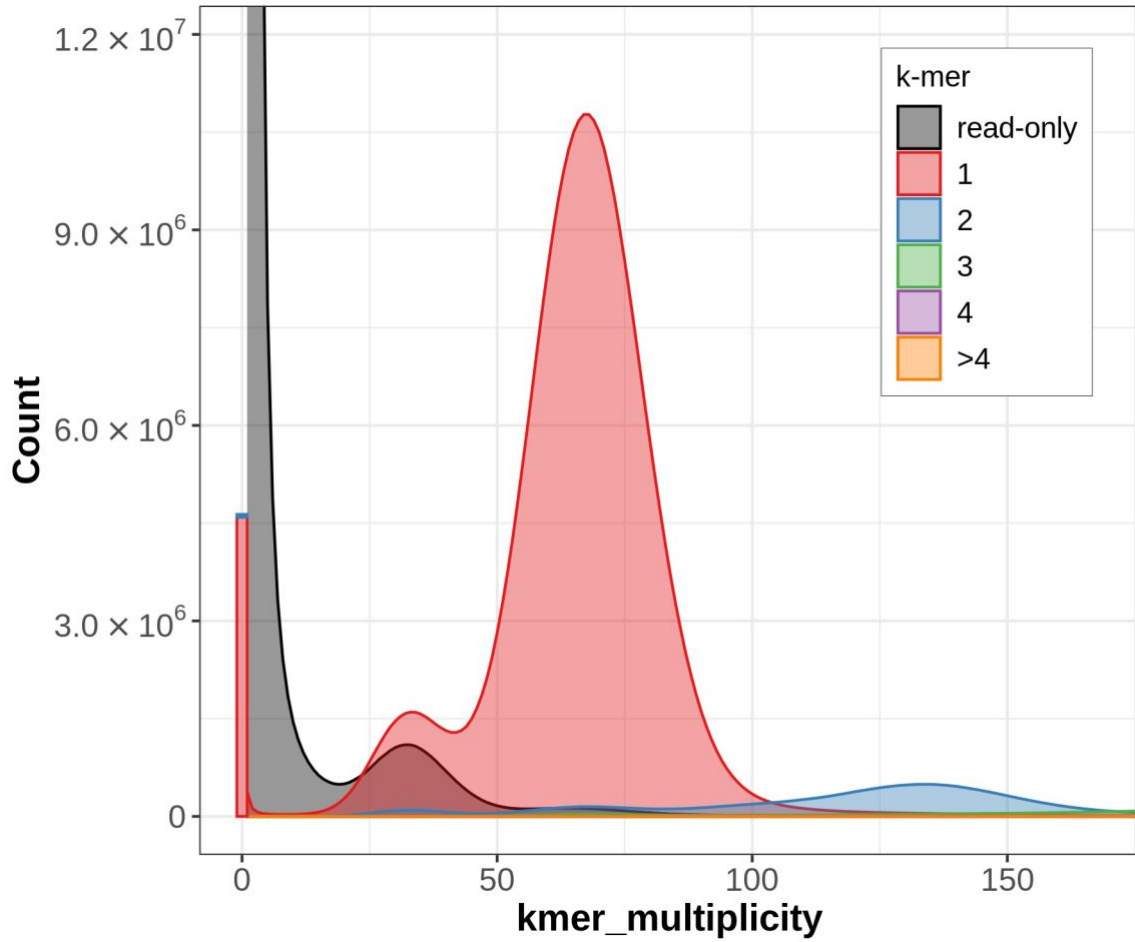
Authors kindly thank Lisa Rotasperti and Giovanna Giudicelli for their help with sample collection. For advice and guidance on population analysis, we are very grateful to Jacob Landis and Suzy Strickler.

Supplemental Material



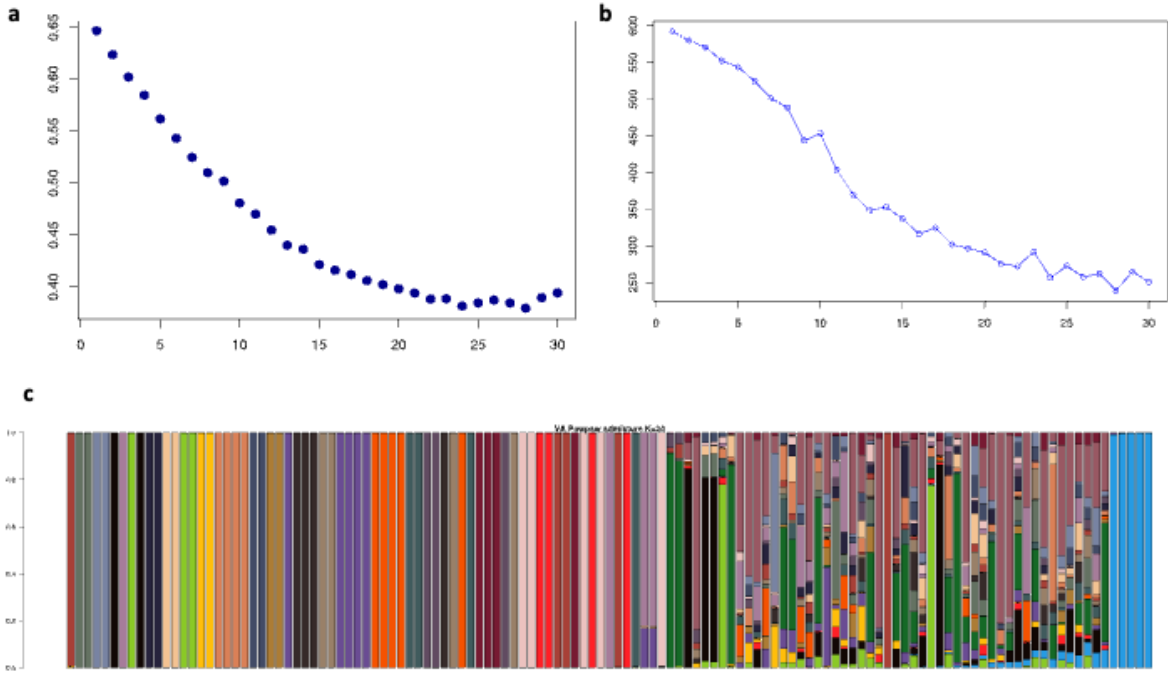
Supplemental Figure 2-1 MLG Heatmap

Heatmap showing the KING estimated phi values for each pairwise genotype comparison. Values close to 0.5 indicate high genetic similarity.



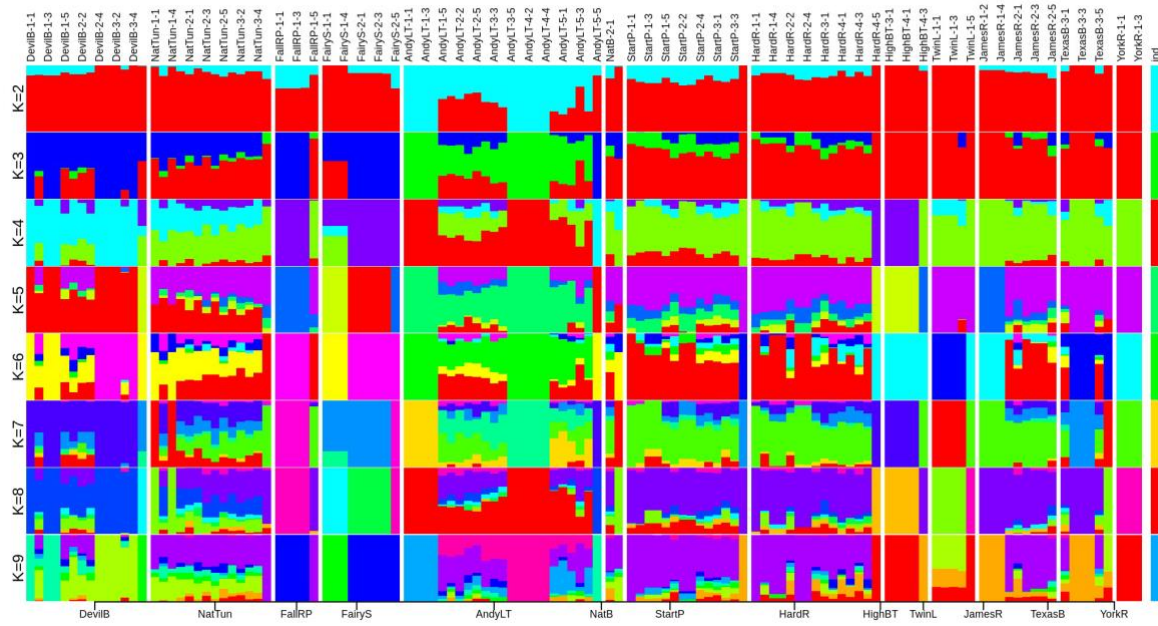
Supplemental Figure 2-2 Merqury spectra plot

Merqury spectra_asm plot showing the occurrence of k-mers in the assembly. X-axis represents the genome coverage. Y-axis represents k-mer coverage. Peaks are coloured by their occurrence in the genome, grey = in read only, not in part of the assembly, red = single copy, blue = two copies, etc.



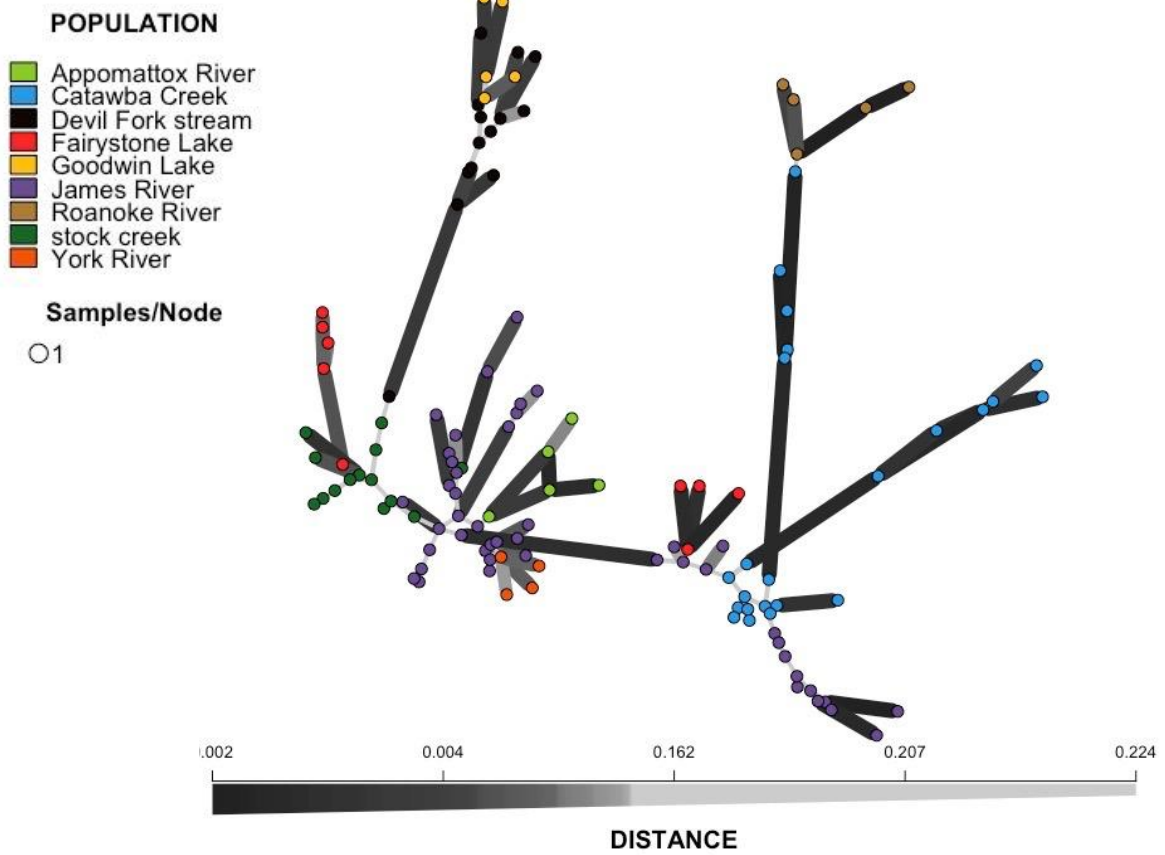
Supplemental Figure 2-3 Structure Inference and admixture

Panel showing inference of population structure and levels of admixture. **a)** population structure estimation by FASTSTRUCTURE. X-axis shows ancestry coefficients and Y-axis shows the number of populations. **b)** shows a population number estimation using a bayesian information criterion (BIC). Both **a&b** show the most likely number of clusters is around 24. **c)** Admixture between individuals at the $K = 24$



Supplemental Figure 2-4 ADMIXTURE

Figure shows admixture at K = 2:9.



Supplemental Figure 2-5 Minimum spanning network

Figure shows the minimum spanning network (MSN) with individuals coloured by the river they were closest to. Genetic distance is indicated by colour (grey to black) and thickness of connecting bar.

Supplemental Table 2-1 Table of PCR primers and adapters sequences

Table of PCR primers and adapters sequences used in the GBS library preparation

PCR primer 5'	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PCR primer 5'	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCT
unique barcode	5-ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxx and 5-CWGyyyAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
common Illumina adapter	5'-CWGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
common Illumina adapter	5'-CTCGGCATTCTGCTGAACCGCTCTTCCGATCT.
	“xxx” and “yyy” are the barcode and barcode complement, respectively

Supplemental Table 2-2 GBS read processing

Table of GBS data processing. Raw sequencing reads mapped to the reference genome Astri041. Quality control and the overall mapping performance of raw reads including the quality of the reads, the number of sites, and the number of SNPs that could be called.

name	raw	processed	mapped	% mapped	sites	% sites
AndyLT-1-1	8795444	8072256	5881818	72.86461183	182337	36.11
AndyLT-1-2	10674318	9575442	6099132	63.69556622	208277	41.25
AndyLT-1-3	11644000	10396324	6042791	58.12430432	200157	39.64
AndyLT-1-4	19118070	17448132	11901544	68.21099244	210658	41.72
AndyLT-1-5	16297450	14299238	7009208	49.01805257	175460	34.75
AndyLT-2-1	11501146	10406326	8770824	84.28357905	150272	29.76
AndyLT-2-2	6702582	6062822	4885155	80.57559664	144194	28.55
AndyLT-2-4	14457536	13171562	10443993	79.29198526	195246	38.67
AndyLT-2-5	15062362	13598930	10141906	74.57870582	203543	40.31
AndyLT-3-1	12491256	11342828	11274317	99.39599719	230721	45.69
AndyLT-3-3	1612000	1466174	1263746	86.19345316	110141	21.81
AndyLT-3-4	17708106	15729770	10614061	67.47753464	197350	39.08
AndyLT-3-5	6901810	6376380	5148179	80.73827156	180956	35.84
AndyLT-4-1	12709486	11457788	8484019	74.04587168	184938	36.62
AndyLT-4-2	13669828	12214326	8233316	67.40704317	188533	37.34
AndyLT-4-3	13358210	12016936	9044738	75.26659042	188568	37.34
AndyLT-4-4	3819648	3454244	2574896	74.54296801	134501	26.63
AndyLT-4-5	8553614	7852140	6805136	86.66600443	184049	36.45
AndyLT-5-1	16545900	14617860	7598342	51.97985204	174249	34.51
AndyLT-5-2	10749214	9756626	7415043	76.00007421	179224	35.49
AndyLT-5-3	4150178	3741018	3282285	87.73774946	149966	29.7

AndyLT-5-4	11166864	10180938	9102805	89.41027831	208786	41.35
AndyLT-5-5	14872796	13342584	8597640	64.4375932	201978	40
DevilB-1-1	10872692	9837350	9241782	93.94584924	240429	47.62
DevilB-1-2	14943342	13357662	6218005	46.55009986	196869	38.99
DevilB-1-3	5415346	4833672	4583835	94.83132078	206689	40.93
DevilB-1-4	6727576	6079292	5853214	96.2811788	202304	40.06
DevilB-1-5	2154532	1964554	1865955	94.98110004	141006	27.92
DevilB-2-1	8484772	7624392	7082422	92.89162992	224340	44.43
DevilB-2-2	11604812	10360280	9158106	88.39631747	224958	44.55
DevilB-2-3	6035614	5395920	4895223	90.7208224	201270	39.86
DevilB-2-4	8312586	7504690	6930790	92.35278206	194544	38.53
DevilB-3-1	7683218	6893000	5993891	86.95620194	193507	38.32
DevilB-3-2	15147966	13664490	11896752	87.0632713	225404	44.64
DevilB-3-3	12073484	10890868	9947258	91.33576865	233760	46.29
DevilB-3-4	13456302	12051736	10551401	87.55088064	199813	39.57
DevilB-3-5	5744650	5173388	4922733	95.15491589	190270	37.68
FairyS-1-1	15293048	13358696	5345451	40.01476641	161157	31.91
FairyS-1-2	15335718	13544486	6128762	45.24912942	177125	35.08
FairyS-1-4	35774252	31505550	11945928	37.91690036	218977	43.37
FairyS-1-5	14683206	13021590	7474784	57.40300532	196079	38.83
FairyS-2-1	10144080	8886082	3527150	39.69297155	144750	28.66
FairyS-2-2	4017772	3585156	1914449	53.39932209	126249	25
FairyS-2-3	5975588	5391700	2554295	47.37457574	125426	24.84
FairyS-2-4	27697056	24507792	9535486	38.9079767	158515	31.39

FairyS-2-5	9617270	8371284	2742937	32.7660249	129005	25.55
FallRP-1-1	25040788	22042228	7981857	36.21166154	202306	40.06
FallRP-1-2	21794358	19417072	9035179	46.53213935	186834	37
FallRP-1-3	30735580	27050014	10903320	40.30800132	196619	38.94
FallRP-1-4	9203532	8046262	3168378	39.37701755	149456	29.6
FallRP-1-5	12468794	10807092	2962494	27.41249913	145502	28.81
HardR-1-1	3744766	3310574	3238682	97.82841284	167620	33.19
HardR-1-3	2008970	1811822	1767179	97.53601623	134080	26.55
HardR-1-4	5029428	4499662	4383733	97.42360648	182061	36.05
HardR-2-1	6262840	5730298	5427136	94.7094898	189246	37.48
HardR-2-2	9140034	8456142	8215553	97.15486093	215896	42.76
HardR-2-3	8124542	7384956	7225159	97.8361821	195940	38.8
HardR-2-4	5638164	5212756	5159269	98.9739209	195136	38.64
HardR-2-5	14250248	13088352	12753202	97.4393262	251133	49.74
HardR-3-1	10392556	9509556	9440730	99.27624381	245685	48.66
HardR-3-2	7663746	6927182	6688028	96.54760045	217304	43.04
HardR-4-1	16288324	14794344	14504406	98.0402105	254128	50.33
HardR-4-2	15299764	14096538	13984216	99.20319443	265513	52.58
HardR-4-3	16913332	15550936	15117009	97.20964063	263680	52.22
HardR-4-4	11097420	10150478	10097802	99.48104907	230984	45.74
HardR-4-5	6202904	5780822	5734466	99.19810712	196602	38.93
HighBT-3-1	9035960	8099626	7674329	94.74917731	244782	48.48
HighBT-3-2	7271526	6639232	6477791	97.56837839	219267	43.42
HighBT-4-1	12772248	11546874	10181127	88.17214945	233453	46.23

HighBT-4-2	24771542	22820724	22622483	99.13131152	219659	43.5
HighBT-4-3	2575104	2357420	2142121	90.86717683	142764	28.27
JamesR-1-2	3875758	3552180	3396418	95.61503077	195328	38.68
JamesR-1-3	4065664	3669736	3552024	96.79235782	233936	46.33
JamesR-1-4	2554290	2349764	2167098	92.22619804	164332	32.54
JamesR-1-5	5898876	5374780	5196601	96.68490617	210570	41.7
JamesR-2-1	16720862	15127902	14487737	95.76831606	261408	51.77
JamesR-2-2	5016790	4574288	4218035	92.21183712	201848	39.97
JamesR-2-3	7457174	6624960	6519596	98.4095904	230083	45.57
JamesR-2-4	3568010	3223968	3215865	99.74866376	181132	35.87
JamesR-2-5	1986184	1730520	1655376	95.65772138	184910	36.62
NatB-2-1	3855230	3531504	3073815	87.03982779	162867	32.25
NatB-3-1	1725140	1576546	1557122	98.76793953	140761	27.87
NatTun-1-1	6208722	5608434	5510078	98.24628408	214044	42.39
NatTun-1-3	8372738	7437782	7319147	98.40496804	229122	45.38
NatTun-1-4	9983578	9062484	8404069	92.73471821	235933	46.72
NatTun-1-5	3646222	3304672	3182956	96.31685081	170738	33.81
NatTun-2-1	7579214	6924660	6427856	92.82558277	216414	42.86
NatTun-2-2	7203702	6572104	6355016	96.69682647	223613	44.28
NatTun-2-3	9729282	8769702	8353068	95.24916582	221579	43.88
NatTun-2-4	19713432	17977636	17443458	97.02865271	294998	58.42
NatTun-2-5	12296654	10782204	9828162	91.15169774	233585	46.26
NatTun-3-1	12759116	11811810	11463195	97.0485895	241610	47.85
NatTun-3-2	6916224	6307044	6220702	98.63102271	204092	40.42

NatTun-3-3	12085094	11015628	9864839	89.55312398	212741	42.13
NatTun-3-4	9674618	8820464	8634471	97.89134676	232099	45.97
NatTun-3-5	11506854	10545440	10148785	96.23861119	230785	45.71
StartP-1-1	23483214	21359662	19080037	89.32742943	283413	56.13
StartP-1-2	22101510	20107416	16865283	83.87593413	262186	51.92
StartP-1-3	20430226	18548936	15591375	84.05536037	228442	45.24
StartP-1-4	19476070	17687282	15927783	90.05217987	262336	51.95
StartP-1-5	17997862	16553990	14895867	89.98354475	260398	51.57
StartP-2-1	16293368	14668298	14281219	97.36111852	259096	51.31
StartP-2-2	5745568	5223932	4649800	89.00958129	173621	34.38
StartP-2-3	2404964	2176134	1847506	84.89854026	137095	27.15
StartP-2-4	5059524	4613766	4324985	93.74088326	170893	33.84
StartP-2-5	20069418	18131752	18160211	100.1569567	241540	47.84
StartP-3-1	9045356	8139344	7828538	96.18143673	203551	40.31
StartP-3-2	15517076	14056882	13999633	99.5927333	249182	49.35
StartP-3-3	7172866	6495152	5996844	92.32800095	184489	36.54
StartP-3-4	9251078	8353862	8286953	99.19906506	205464	40.69
TexasB-3-1	10534512	9366756	9254335	98.79978725	229301	45.41
TexasB-3-2	9477026	8380638	8352011	99.65841503	235506	46.64
TexasB-3-3	4843286	4401670	4345230	98.7177594	188844	37.4
TexasB-3-4	1889968	1715898	1705707	99.40608358	145923	28.9
TexasB-3-5	3806150	3460452	3455893	99.8682542	173455	34.35
TexasB-4-4	3796446	3466052	3539092	102.1072967	175587	34.77
TwinL-1-1	7085656	6414994	3940175	61.4213357	153372	30.37

TwinL-1-2	20572964	18660510	11404362	61.11495345	205544	40.71
TwinL-1-3	7961040	7125916	3890105	54.59094662	149388	29.58
TwinL-1-4	9338834	8355942	4840532	57.92921971	142551	28.23
TwinL-1-5	5080658	4605010	3172710	68.89691879	145946	28.9
YorkR-1-1	8998068	7933560	3170442	39.96241284	135466	26.83
YorkR-1-2	3869968	3506744	1465807	41.79965803	106931	21.17
YorkR-1-3	5922718	5461620	3122696	57.17527034	127528	25.25
YorkR-1-5	13764148	12325566	6610718	53.63419416	170789	33.82
All	1328872140	9576713.472	930012490	81.94921569	504887	38.9396

Supplemental Table 2-3 MLG occurrence at sampling sites

Table summarizes the number of multi-locus-genotypes (MLG) at each sampling patch. Estimates using VCFTOOLS relatedness function which using the KING algorithm to perform pairwise comparisons of individuals. MLG numbers based on pairwise estimates of 0.45 to 0.5.

Patch	MLGs/Patch	Individuals/Patch	Nucleotide Diversity
AndyLT.1	1	5	806.02
AndyLT.2	3	4	1270.79
AndyLT.3	4	4	1279.14
AndyLT.4	1	5	838.87
AndyLT.5	4	5	1292.67
DevilB.1	3	5	1084.47
DevilB.2	3	4	1319.00
DevilB.3	3	5	923.00
FairyS.1	2	4	885.68
FairyS.2	2	5	838.38
FallRP.1	3	5	803.44
HardW.1	2	3	1126.80
HardW.2	2	5	1237.49
HardW.3	2	2	1403.50
HardW.4	2	5	1307.87
HighBT.1	2	5	898.56
JamesR.1	2	4	853.18
JamesR.2	3	5	1193.98
NatB.2	2	2	1504.00
NatTun.1	3	4	1496.00
NatTun.2	2	5	1205.89
NatTun.3	5	5	1324.44
StartP.1	5	5	1345.93
StartP.2	5	5	1166.73
StartP.3	4	4	1284.78
TexasB.3	3	5	1364.68
TexasB.4	1	1	1522.00
TwinL.1	3	5	991.33
YorkR.1	2	4	813.04

Supplemental Table 2-4 AMOVA results

Summary of AMOVA results from individuals assigned to strata (Basin/Patches and) with 999 permutations. Significance testing carried out via Monte-Carlo Simulation of variation.

	Basin/Patches	Df	Sum Sq	Mean Sq	
POPPR - AMOVA					
Variations Between Pop		3	9,230.07	3,076.69	
Variations Between Subpop Within Pop		26	42,380.27	1,630.01	
Variations Between samples Within Subpop		95	28,877.13	303.97	
Variations Within samples		125	92,451.50	739.61	
Total		249	172,938.96	694.53	
Sigma				%	
Variations Between Pop			24.53	3.47	
Variations Between Subpop Within Pop			160.83	22.74	
Variations Between samples Within Subpop			-217.82	-30.80	
Variations Within samples			739.61	104.59	
Total variations			707.16	100.00	
Statphi				Phi	
Phi-samples-total				-0.05	
Phi-samples-Subpop				-0.42	
Phi-Subpop-Pop				0.24	
Phi-Pop-total				0.03	
Monte-Carlo					
Permutation number: 999		Obs	Std.Obs	Alter	Pvalue
Variations Within Samples		739.61	2.90	less	0.997
Variations Between Samples		-217.82	-6.79	greater	1
Variations Between Subpop		160.83	32.16	greater	0.001
Variations Within Pop		24.53	4.44	greater	0.001

Chapter 3 : De novo assembly in a non-model species

Introduction

Genome sequencing and assembly in non-model species

A reference genome provides a powerful resource to investigate the genomic, transcriptomic, proteomic, and metabolomic landscape of a species (Schuster, 2008), but also provides insight to its evolutionary history (Lamichhaney et al., 2012; Montero-Mendieta et al., 2017) and even species level genetic diversity (Etherington et al., 2020; Lamichhaney et al., 2012). Since the early days of sequencing in the 1950's and 60's (Holley et al., 1965; Sanger and Thompson, 1953a, 1953b; Wu and Kaiser, 1968), many advances have been made to reduce the prohibitive financial cost, time and effort required to sequence and assemble a reference genome (Giani et al., 2020). The ultimate goal of any genome assembly project is to produce a highly accurate draft, representing the truest version of the real nucleotide sequence, including correct super scaffolding of pseudochromosome molecules, along with annotated genetic elements [e.g. structural rearrangements, structural variants, repeat elements and transposable elements] (Ellegren, 2014). Regardless of the impressive sequencing advancements over the past 50 years, generating a complete, accurate, and annotated high-quality assembly still poses significant challenges for anyone attempting to generate a *de novo* assembly for non-model species. Without prior information on the species in question or a closely related species, successful assembly can be even more challenging. There are many considerations to be wary of during each step of the process and careful quality controls are needed at every step. Although an annotated telomere-to-telomere assembly is always desirable because it provides the broadest range of downstream applications, it is not an affordable or even necessary target. More important considerations for a non-model *de novo* assembly project is the length of genome, its structural complexity (repeat density, heterozygosity, ploidy level and CG content), and how complete the final assembly draft needs to be for the current project. These factors

inform the sequencing coverage, the sequencing platform, the overall cost and final quality of the assembly (Angel et al., 2018; Jung et al., 2020, 2019).

The first issue with non-model species is that without prior investigations, the genome size maybe unknown. Further, sourcing tissue might require sampling in remote locations or using preserved samples, all of which can affect the quality of the DNA extracted. This can be an issue for long read platforms that rely on long strands of intact template DNA (Etherington et al., 2020). Additionally, sample tissue from non-model species often originates from highly heterozygous wild individuals which can complicate the assembly unless sufficient sequencing cover is provided (Etherington et al., 2020; Zerbino and Birney, 2008). A common first step in any non-model sequencing project is to measure genome size by flow cytometry, and k-mer frequency distribution for genome complexity and heterozygosity (Li and Harkess, 2018).

Once the size of the genome and approximate GC content and complexity is known, a sequencing platform or combination of platforms need to be chosen. Short-read sequencing was once the most accurate with low error rates but the short reads increase the difficulty when assembling complex and repetitive sequences, potentially resulting in a heavily fragmented genome with incorrectly collapsed regions, misassemblies, and introduced gaps (Goodwin et al., 2016; Salzberg and Yorke, 2005). PacBio's Sequel and Oxford Nanopore platforms provide longer reads capable of covering larger genomic regions but with an error rate of around 10–15%, discussed in more detail in chapter 1. A successful approach to minimize the drawbacks of each approach has been to combine both methods using the more accurately called short reads to correct errors and long reads to resolve longer stretches of sequence (Chen et al., 2020; Jiménez-Ruiz et al., 2020; Tomé et al., 2022). A weakness in this strategy is that short-read NGS platforms rely on template amplification during the library preparation which can introduce replication bias in some regions, cause loss of information, and copying errors to the

sequences relied on for error correction (Shendure et al., 2017). Further, high GC regions are often not sufficiently replicated during PCR amplification (Bentley et al., 2008; Chen et al., 2013). In 2019 Pacific Bioscience optimized their circular consensus sequencing (CCS) to produce new long high-fidelity (HiFi) reads, improving the accuracy of their single-molecule real-time (SMRT) sequencing to an impressive 99.8% accuracy with an average read length of 13.5 Kb (Wenger et al., 2019), an approach that avoids the pitfalls of both NGS and third-generation sequencing (TGS) methods.

There are many *de novo* assembler tools and pipelines that can be used for non-model organisms but proper testing of multiple tools is needed to find the optimum approach for each project (Bankevich et al., 2022; Jung et al., 2020, 2019a). Indeed, this is also the case with annotation and transcript assembly tools which may not be optimized for the species of interest (Duarte et al., 2021; Yandell and Ence, 2012).

Genome annotation in non-model species

Annotating a genome requires evidence from closely related species and structural information from closely related species to predict and assign coding sequences to the draft genome (Salzberg, 2019). Approaches to annotation vary but broadly follow either an *ab initio* prediction approach without prior information, or using all available evidence [transcripts, expressed sequence tags (ESTs), homologous proteins], to train gene predictions (Yandell and Ence, 2012). The more evidence that can be provide the more accurate the annotation is likely to be. However, in non-model species, only limited evidence might be available and the annotation restricted to transcripts or protein-coding sequence (CDS) generated from additional RNA-sequencing data (Ekblom and Wolf, 2014).

In a first step, the repetitive regions must be “masked off” because they are highly varied across species and thus pose difficulty for gene predictors (Cantarel et al., 2008). Repeat families can be identified in the draft genome using tools like RepeatModeler (Flynn et al., 2020) and used to mask these regions with RepeatMasker (Tarailo-Graovac and Chen, 2009). Popular predictor pipelines, like BRAKER2 (Brůna et al., 2021) and MAKER2 (Holt and Yandell, 2011), are then used to train gene predictors such as GeneMark-E (Lomsadze et al., 2014), AUGUSTUS (Stanke and Waack, 2003), Semi-HMM-based Nucleic Acid Parser (SNAP) (Korf, 2004) and ProtHint (Mathebula, 2016) to improve their predicted gene models. Once gene predictions are made using the supplied evidence, they are synthesized to generate a final set of annotated genes. From there, it is then possible to use gene family information to assign biological meaning to coding regions of the genome for use in further studies (Harris et al., 2004; Kanehisa and Goto, 2000). However, as with the quality of the genome, a poor-quality or low confidence annotation will have negative effects on all projects using the assembly and annotation and so the quality and reliability of both needs to be carefully evaluated.

The goal of this genome assembly and annotation project is to determine the complete DNA sequence of *A. triloba* genome, improving on the previous version (Chapter 2), and to annotate, and describe the features of the genome such as genes, regulatory regions, and repeats. This information can be used to understand the biology of evolutionary history of the pawpaw tree as well as providing a resource to study the genetic diversity across the geographical range of the species.

Materials and Methods

Sample material

Plant material for both the PacBio High fidelity (HiFi) (Pacific Biosciences, Menlo Park, California) and Illumina (San Diego, California) Chromatin-association/interaction analysis (Hi-C) were donated from the Harvard Arboretum. The leaf tissue was collected from a tree over 100 years old, the oldest living pawpaw tree in the collection, accession number 12708-A (origin Grand Rapids, Michigan, planted 1903).

For RNA-sequencing, fresh leaf, pericarp, and fruit tissue were collected from a pawpaw tree at the Boyce Thomson Institutes (BTI) experimental orchard in Ithaca, New York.

HiFi Sequencing and assembly

Leaf tissue was sent to the Icahn School of Medicine at Mount Sinai (NY, USA) for DNA extraction and library preparation. Sequencing was performed on a PacBio Sequel II at the Mount Sinai genomics centre. PacBio HiFi reads were generated at the sequencing facility from circular CCS. The first step in the assembly was to test the performance of several assembler tools.

Three *de novo* assemblies were generated to test the performance of three different assemblers run with default parameters Canu.v2.2 (Koren et al., 2017), Hifi-asm.v0.16.0 (Cheng et al., 2021), La Jolla Assembler.v02 (LJA) (Bankevich et al., 2022). All three assembly tools were run with default parameters, the only changes were with Canu and consisted in selecting HiFi assembly with Hicanu using the option -pacbio-hifi.

The HiFi-asm draft assembly was extracted from the haplotype-resolved processed uniting graph by converting the GFA to FASTA following the instructions in the HiFi-asm manual.

Coverage

Sequencing coverage was calculated by $N \times L/G$ where (G) is the genome length, (N) is the number of sequencing reads, (L) and the average read length. Mapping coverage was calculated by mapping HiFi reads to the completed assembly with MINIMAP2 (Li, 2018) and using SAMTOOLS.v1.7 (Li et al., 2009) and BEDTOOLS.v2.29.0 (Quinlan and Hall, 2010) to estimate mapping depth and breadth.

Genome assembly evaluation

Draft assembly quality was assessed on three characteristics: completeness, accuracy, and contiguity. Completeness was evaluated on assembly genome size compared to the flow cytometry estimates. BUSCO v.5.0.0 (Seppey et al., 2019; Simão et al., 2015) with the embryophyte_db10 database was used to measure gene space completeness, while transposable element space completeness was estimated with long terminal repeats LTRretriever.v2.9.0 (Ou and Jiang, 2018). The incorporation of the reads into the assembly was evaluated with Merqury.v1.1.0 (Rhie et al., 2020) by k-mer catalogue comparison generated from the PacBio HiFi reads using the recommend k-mer size of 20 base-pairs (bp) for a 0.85 Giga-bases (Gb) assembly. Merqury also measured the sequence accuracy, providing a phred-scaled consensus quality (QV) score and consensus error rate for each draft assembly, then generating copy number spectra plots. Contiguity stats, including L50, N50, L90 and N90, were calculated using a custom script available on Github; FastaSeqStats (Bombarely, 2022). The draft assembly was scanned using Blobtools.v1.1.1 (Laetsch and Blaxter, 2017) to identify sequences not originating from *A. triloba*.

Hi-C sequencing and scaffolding

Prior to DNA sequencing, the genome size was estimated using flow cytometry. Flow cytometry was carried out following the same methodology as used in (Hasing et al., 2019). Leaf material was sent to Phase Genomics (Seattle, Washington) for DNA extraction, Hi-C library preparation with restriction digestion enzyme MboI to digest chromatin, and sequencing on Illumina NovaSeq 6000. A total of 24.26 Gb of Hi-C reads with a 27X coverage was generated and used to anchor contigs to scaffolds. This involved converting the raw Hi-C fragments into a contact map (3D positional information on adjacent genomic regions) using Juicer.v.1.6 (Durand et al., 2016), so that this map could then be used by 3D-DNA to orientate and anchor contigs. To do this, the draft assembly was indexed with BWA v.0.7.17-r1188t (Li, 2013), then the Hi-C read pairs were aligned to the indexed assembly with Juicer, before removing duplicates and near-duplicates to produce a list of Hi-C contacts. This was then used by 3D-DNA pipeline.v.201008 (Dudchenko et al., 2017) to anchor assembled contigs into chromosome length scaffolds. Juicebox.v.1.11.08 was then used to visually inspect the scaffolds for any ordering errors (translocations) or orientation errors (inversions) produced during the assembly of Astri106.

RNA-sequencing

Total RNA was extracted using the RNeasy power plant kit from Qiagen (Hilden, Germany) following the manufacturer protocol. Briefly, 50 mg of each tissue was added to an individual 2 ml PowerBead tube containing 600 µl MBL/ β -mercaptoethanol solution. Tissue lysis was carried out in a centrifuge (Centrifuge 5425 Eppendorf) run at 13,000 xg for 2 mins at room temperature. Next ~600 µl of the supernatant was transferred to each new 2 ml tube and 150 µl of solution IRS was added. Tubes containing the supernatant were left to incubate for 5 mins at 4 °C before being centrifuged (Centrifuge 5425 Eppendorf) at 13,000 xg for a further 2 mins.

650 μ l of the supernatant was transferred to new 2 ml collection tubes carefully avoiding the pellet. 650 μ l of solution PM3 and 650 μ l of solution PM4 were added to each tube and vortexed. Each solution was then added to MB RNA Columns and centrifuged at 13,000 xg for 1 min, repeating 650 μ l loads at a time until all remaining solution passed through the columns. 600 μ l of PM5 was added to each column and centrifuged at 13,000 xg for 1 min. The columns were then placed in new 2 ml collection tubes and 600 μ l of solution PM4 was added before being centrifuged at 13,000 xg for 1 min. Flowthrough was discarded and centrifuged again at 16,000 xg for 2 mins. RNA was then eluted in 50 μ l of RNase-free water and stored at -80 °C. RNA was quantified by nanodrop, and fragments size and quality were measured by Agilent 4200 TapeStation. RNA samples were sent to Novogene (Sacramento, California) for sequencing.

Repeat Masking

Annotation of the HiFi and Hi-C assembly was carried out using a semi-automatic consensus approach which involved three main steps; identification and masking of non-coding regions, gene prediction, and finally assigning of the biological meaning to the predicted genes. In the first step, *de novo* repeat family identification, and modelling were carried out using RepeatModeler.v.2.0.3 (Flynn et al., 2020). LTR were identified with the LTR discovery pipeline included in RepeatModeler. Repeat sequences identified via *de novo* tools may also include protein-coding genes and can need curation to remove these from the repeat library before genome masking. To manually perform this curation step, a blastx was performed on the repeat family library against the UniProtKB/Swiss-Prot protein sequence database using DIAMOND.v.0.9.24.125. Repeat sequences with high sequence homology to protein-coding genes in the UniProtKB/Swiss-Prot database were removed from the library. A “soft” masked version of the draft was generated by RepeatMasker.v.4.0.9 (Smit et al., 2015) using the

RepeatModeler output to locate and mask repeats in the assembly sequence. The results of RepeatMasker were analysed using the R package repeat R.v.0.1.0 to assess the composition of repeats for each contig (Winter, 2022).

Gene prediction

Gene prediction was optimised by combining and comparing two approaches. We first ran the BRAKER2.v.2.1.5 pipeline (Brůna et al., 2021) to predict genes with the AUGUSTUS predictor (Stanke and Waack, 2003) and to train AUGUSTUS to *A. triloba*, then we ran MAKER.v.3.01.02 (Cantarel et al., 2008) using the trained AUGUSTUS gene models with additional evidence. Both approaches are described in more detail below.

Ab initio gene prediction was done using BRAKER2.v.2.1.5 pipeline with GeneMark-EX (Lomsadze et al., 2014), AUGUSTUS, and ProtHint.v.2.6.0 (Mathebula, 2016). BRAKER2 can integrate RNA-seq spliced alignment information from bam files and homologous protein sequences to improve gene prediction. To produce the bam files, the RNA-seq reads were mapped to the assembly by running STAR.v.2.7.5a (Dobin, 2023) with default settings. A FASTA file containing homologous protein sequences from 5 Magnoliidae species was compiled from unpublished and published resources. The species used were *Cinnamomum micranthum* (Chaw et al., 2019), *Liriodendron chinense* (Chen et al., 2019), *Magnolia biondii* (Dong and Liu, 2020), *Persea americana* (unpublished data), and the closely related Annonaceae species *Annona muricata* (Strijk et al., 2021).

Additionally, MAKER.v.3.01.02 pipeline, an evidence-based method of gene prediction, was run for comparison. The AUGUSTUS gene prediction model for *A. triloba* was trained during the *ab initio* BRAKER pipeline. MAKER allowed for integration of the trained AUGUSTUS predictor and Semi-HMM-based Nucleic Acid Parser (SNAP) (Korf, 2004) along with

homologous protein models, and RNA-seq assembled transcripts. SNAP was trained by successive runs by first by providing expressed sequence tags (EST) and homologous protein models with Hidden Markov models (HMM) created from iterative SNAP runs. Protein evidence was provided by the same Magnoliidae dataset used with BRAKER, while ESTs were generated from the transcripts assembled by StringTie.v.2.1.5 (Pertea et al., 2015) using RNA-seq reads mapped to draft assembly with STAR.v2.7.5a and the predicted using TransDecoder.v.5.5.0 (Hass and Papanicolaou, 2016). After training of SNAP, the MAKER pipeline was run incorporating the soft-masked draft genome from RepeatMasker, the Magnoliidae protein dataset, HMM from the SNAP training, AUGUSTUS with *A. triloba* specific models trained by the previous BRAKER2 predictions and StringTie transcripts from RNA-seq data. The final MAKER annotation was synthesized from all overlapping genes predicted by each evidence source.

Annotation quality assessment and gene clustering

Annotation quality of both the BRAKER and MAKER pipelines were evaluated by annotation edit distance (AED) comparing annotation with overlapping aligned ESTs, mRNA-seq and protein homology data. BUSCO.v.5 was also used to evaluate the gene space completeness of the predicted genes in the annotation, and compared with the completeness of those from ESTs, mRNA-seq and protein homology to evaluate the contribution of each evidence to the gene prediction.

The MAKER pipeline annotation was found to produce the best quality and most complete annotation and was used to carry out orthologous gene clustering with OrthoVenn2 (Xu et al., 2019), comparing the modelled *A. triloba* genes with closely related *A. muricata* and *P. americana*.

Results and Discussion

Sequencing coverage

For the second *A. triloba* assembly, a third-generation sequencing (TGS) approach was taken, combining PacBio HiFi with chromosomal contact information from Illumina Hi-C for improved scaffolding. Raw sequencing data consisted of 44.8 Gb of HiFi reads generated from PacBio CCS reads for a 49.8x sequencing coverage [Table 3-1]. Contig scaffolding was improved with genome contact maps generated from 24.26 Gb Illumina Hi-C data with a 27x sequencing coverage.

Table 3-1 Sequencing coverage

Table summary of the sequencing data and genome coverage

Sequencing report	PacBio Hifi	Hi-C
Reads	4,948,538	161,746,572
Total size sequenced (Gb)	44.8	24.26
Average size (Kb)	9,045.09	150
Longest read (Kb)	465,820	150
Coverage	49.80	27

CANU assembly

The first assembly from Canu produced an assembly with a total length of 1.71 Gb, approximately 62% larger than the 1.05 Gb estimated by GenomeScope and more than double the 0.84 Gb of the Astri041 assembly (see Chapter 2). The assembly was comprised of 8,882 contigs with an average contig length of 0.2 Mb, an L90 of 3,964 sequences, and an N90 of 0.05 Mb. The assembly gene space completeness evaluated with benchmarking universal single-copy orthologs (BUSCO) reported 94.6% completeness for 2,326 orthologous single copy genes. This included 64% single copies of genes, but 30.5% of the complete BUSCO genes present were duplicated [Table 3-2]. A Merqury k-mer comparison also indicated a

highly complete assembly from Canu. According to Merqury, the draft assembly was 99.2% complete with a high-quality score of 69.5 and a low error rate of 1.13E-07. As seen with BUSCO, Merqurys spectra plots [Figure 3-1] indicated that significant portion of the read k-mers were duplicated in the assembly. Difficulty resolving heterozygous and repeat regions may be the reason the assembly is a lot larger than Astri041.

Table 3-2 Draft assembly comparison

Table summary of the draft assembler performance. Canu, La jolla, Hifiasm were run with only HiFi data. Astri105 used HiFi and Hi-C. The final assembly ‘Astri106’ corresponds to the version ‘Astri105’ after removal of contamination sequences.

		Primary assembly				Final assembly
		Canu	La Jolla	Hifiasm	Hifiasm Hi-C (Astri105)	Hifiasm Hi-C (Astri106)
Assembly tool						
FASTASTATS <i>Assembly summary</i>	Assembly size (Gb)	1.70	1.60	0.90	0.90	0.85
	Contigs	8,882	9,212	992	614 (Scaffolds)	294(Scaffolds)
	Longest seq (Mb)	53.2	12	157.5	161.7	161.66
	Shortest seq (bp)	4,494	1,502	6,538	1,000	4120
	Average seqlength (Mb)	0.2	0.1	0.9	1.5	2.92
	L90, number of seq	3,964	3,709	9	8	7
	N90 (Mb)	0.05	0.05	28.2	63.6	92.55
	L50, number of seq	43	342	4	4	4
	N50 (Mb)	10.3	2.9	107.6	107.1	107.11
BUSCO	% Complete	94.6	93.1	94.7	92.6	98.80
	% Single	64.1	65.1	91.4	89.2	
	% Duplicated	30.5	28	3.3	3.4	2
	% Fragmented	1.9	2.8	1.7	30	0.60
	% Missing	3.5	4.1	3.6	4.4	0.60
Merqury	Completeness	99.2	99.1	93.1	93.44	93.44
	QV	69.5	55.4	61.7	61.95	61.95
	Error	1.13E-07	2.90E-06	6.69E-07	6.38E-07	6.38E-07
LTR retriever	LTR Assembly Index				13.57	13.57

CANU haplotig purge

Regions that have not been correctly collapsed during the assembly process can later be identified and removed using the tool `Purge_dups.v.0.0.3` (Guan et al., 2020). `Purge_dups` uses sequence similarity and read depth to identify and remove both haplotigs and heterozygous overlaps introduced during assembly. After running `Purge_dups` on the Canu draft, the sequence length was reduced from 1.71 Gb to 1.50 Gb, and there was a significant improvement in the contiguity of the assembly, reducing the number of contigs from 8,882 to 1,058. The average contig size increased and the N90 steeply decreased to 297 Mb. BUSCO scores very similar to the primary unprocessed Canu draft [Table 3-2]. The increased contiguity and only a small drop in the number of duplicated BUSCO genes, suggests that Canu was struggling to collapse heterozygous haplotigs into a single consensus. Evidently, `purge_dups` did not remove many repeats, and a possible explanation could be because of the way Canu handles HiFi sequencing reads, focusing on haplotypes diversity during the assembly (Nurk et al., 2020), which may cause a problem for `purge_dups` haplotype cut off threshold. This is because `Purge_dups` maps reads to the genome draft with `Minimap2` (Guan et al., 2020; Li, 2018) and then calculates read-depth for haploid and diploid coverage to select a threshold to separate the two. The threshold might be inaccurate in a HiFi assembly, resulting in a failure to identify and remove duplicate regions.

LJA assembly

The LJA assembly produced an assembly similar in size to the 1.50 Gb long haplotig purged Canu assembly. However, it was more fragmented than the Canu assembly, with 9,212 contigs, longest contig being 12 Mb, almost five times shorter than the longest Canu contig. The assembly had an L90 of 3,709, and a similar N90 of 0.05 Mb. BUSCO completeness was of 93.1%, with 65.1% of single copies and 28% duplicated. K-mer catalogue comparison with

Mercury showed an equal assembly completeness at 99.1%, a quality score of 55.4, and a slightly higher error rate (2.90E-06) [Table 3-2]. The spectra plots [Figure 3-1] show a similar representation of duplicated read k-mers in the assembly k-mer catalogue. It appears that while both assemblies are highly complete, the excessive genome size and duplicated BUSCO genes in both assemblies indicated a difficulty with the consensus calling of the repeats and highly heterozygous loci resulting in uncollapsed regions being added to the assembly length.

Hifi-asm assembly

The third assembler, Hifi-asm, produced the closest assembly to the size of the flow cytometry estimate and the Astri041 draft (Chapter 2). The Hifi-asm draft was 0.89 Gb long, roughly half the size of the other two. It was also the least fragmented, containing only 992 contigs, nine times less contigs than the LJA assembly. The longest contig was 157.2 Mb long, much larger than in any of the other assemblies, with an average length of 0.9 Mb. Interestingly, the L90 was 9 with a N90 of 28.2 Mb. Since pawpaw is said to have a chromosome number of $2n = 2x = 18$ (Bowden, 1940; Kral, 1960; Locke, 1936), HiFi-asm appears to have produced at least a near chromosome level assembly without the need for a contact map. The BUSCO score for the assembly was similar to others in overall completeness with 94.7% of expected genes present. A clear difference can be seen in the occurrence of the single and duplicated genes. In contrast to the previous two assemblies, the Hifi-asm draft contained 91.4% single copy, and only 3.3% duplicated genes. Analysis with Mercury indicated that this was the least complete assembly (93.1% completeness). Overall, quality was similar with a QC score of 61.7 and an error rate of 6.69E-07. The copy number spectra plots for the Hifi-asm showed a significant difference to the previous two assemblies in terms of duplicated k-mers [Figure 3-1].

Comparing all three without any further processing, the Hifi-asm draft appeared to be the most complete while also being the most accurate in size when compared with the flowcytometry

estimate and the previous Astri041 assembly (Chapter 2). Further, the size is the most comparable to the closely related *Annona muricata*, who's 0.65 Gb genome was recently assembled using Illumina, PacBio, 10X, BioNano, and Hi-C sequencing reads (Strijk et al., 2021). Although the overall completeness is lower slightly according to BUSCO and Merqury, it is likely that the larger amounts of duplication in the Canu and LJA assemblies are contributing to their completeness. Hifi-asm attempts to resolve duplicated segments during contig assembly, this could be the reason only one out the three drafts did not have high levels of duplicated single copy ancestral genes.

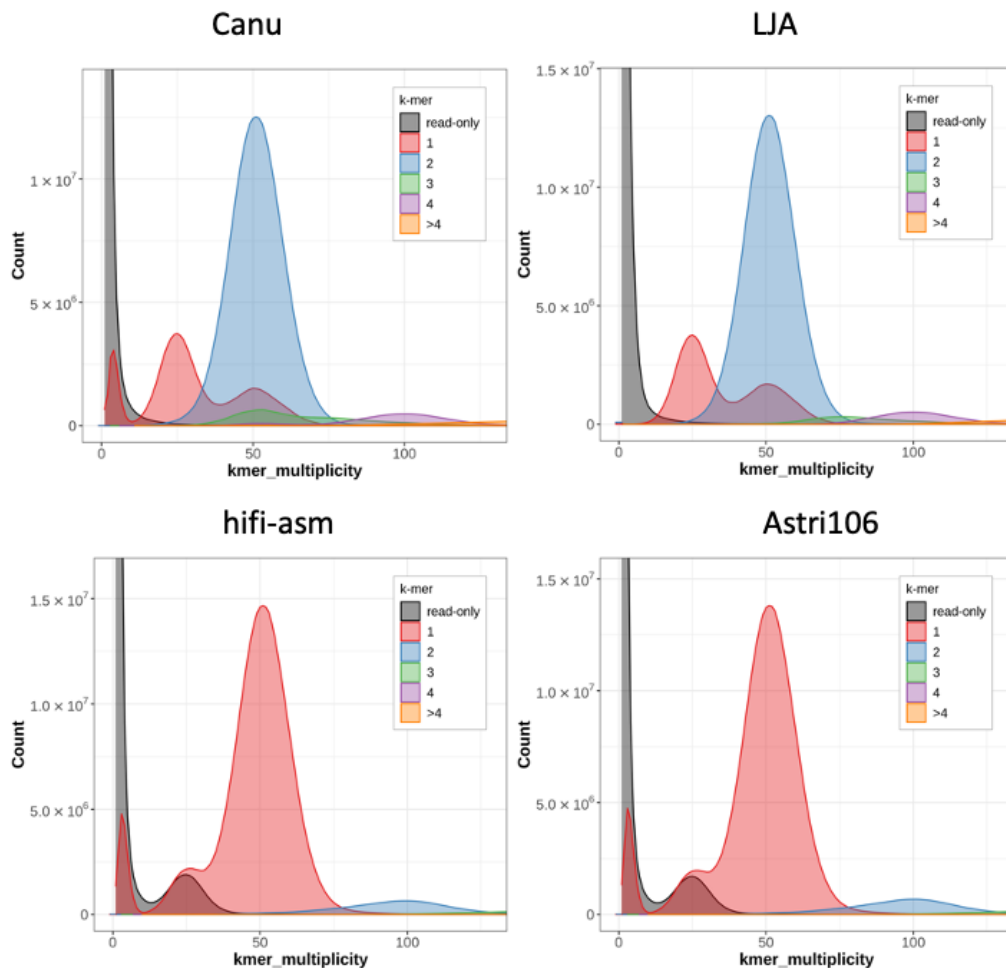


Figure 3-1 Merqury Spectra plots

Merqury spectra_asm plots showing the occurrence of k-mers in the assembly. X-axis represents the genome coverage. Y-axis represents k-mer coverage. Peaks are coloured by their occurrence in the genome, grey = in read only, not in part of the assembly, red = single copy, blue = two copies, etc. Each assembly is a draft using only the HiFi data. Assembler tool is noted above each plot.

HiFi-asm and Hi-C

Considering all descriptive and qualitative metrics, the HiFi-asm draft was selected for scaffolding with Hi-C as it had the highest quality; reaching a near chromosome-level contiguity, least duplicated, low error rates, and closely matched the expected size closely.

Scaffolding with Hi-C data followed the 3D-DNA pipeline (Dudchenko et al., 2017) and resulted in the Astri105 assembly. This version of the HiFi-asm assembly was the same size (0.89 Gb), but the number of contigs reduced from 992 to 614 scaffolds, the longest sequence increased to 161.7 Mb, and shortest sequence reduced to 1,000 Mb. This showed a significant reordering of the contigs informed by the Hi-C contact map during the scaffolding process. The L90 number also reduced to eight and the N90 almost tripled, signifying a greater contiguity in the overall assembly. This level of contiguity was comparable to recent *A. muricata* assembly with Hi-C scaffolding (Strijk et al., 2021). Mapping HiFi reads to the Astri105 genome with Minimap2 showed a mean mapping coverage of 49.47x and a mapping coverage >20x for 91.42% of the assembly, implying a good support of the consensus calls.

The contact map for the draft HiFi/Hi-C assembly was visualised with Juicebox [Figure 3-2] and we identified eight large chromosomes and a ninth much smaller than the rest. The mean chromosome size was 102 Mb, but the ninth was only 18.58 Mb, and approximately one-third the size of the next smallest chromosome. The diminutive size of chromosome number nine was so distinct from the others that it appeared to be a potential contamination. We ran blobtools to compare each assembled contigs sequence similarity to known sequences in the NCBI database [Figure 3-3]. Chromosome 9 and several small contigs returned hits for an unknown species in the phylum Ascomycota, indicating a fungal contamination in the assembly. Any sequences indicated by Blobtools as of fungal origin were removed to produce the final assembly version Astri106, containing eight chromosomal pseudomolecules. The removal of all sequences reduced the length of the assembly from 0.89 Gb to 0.85 Gb, much

more similar to the previous Illumina/PacBio assembly Astri041 (0.84 Gb). The total number of scaffolds reduced from 614 to 294, with eight superscaffolds in total (pseudomolecules) indicating a highly contiguous draft.

What was particularly interesting was that the number of haploid chromosomes was reported to be nine, $2n = 2x = 18$, in three different early studies (Bowden, 1949, 1940; Kral, 1960; Locke, 1936), and one group from the time period reported eight chromosomes (Ito and Mutsuura, 1956). The reason for the discrepancy previous reports is unclear but both Locke and Bowden mentioned the difficulty in obtaining a clear image of the chromosomes, with Locke (1936) stated that it was difficult to count the number of chromosomes, remarking that although it was not very clear, during certain phases of cell division there appeared to be nine chromosomes; and Bowden (1949) mentioned some difficulties in obtaining a clear image of all chromosomes because of their arrangement. Bowden's observation of two large chromosomes and two very small ones is also noteworthy, as we have assembled eight chromosomes of roughly equal length. Early cytological studies may have encountered difficulties in discerning chromosome number, or there might be some natural variation in the number and size of chromosomes in pawpaw trees, but here we can show a definitive number.

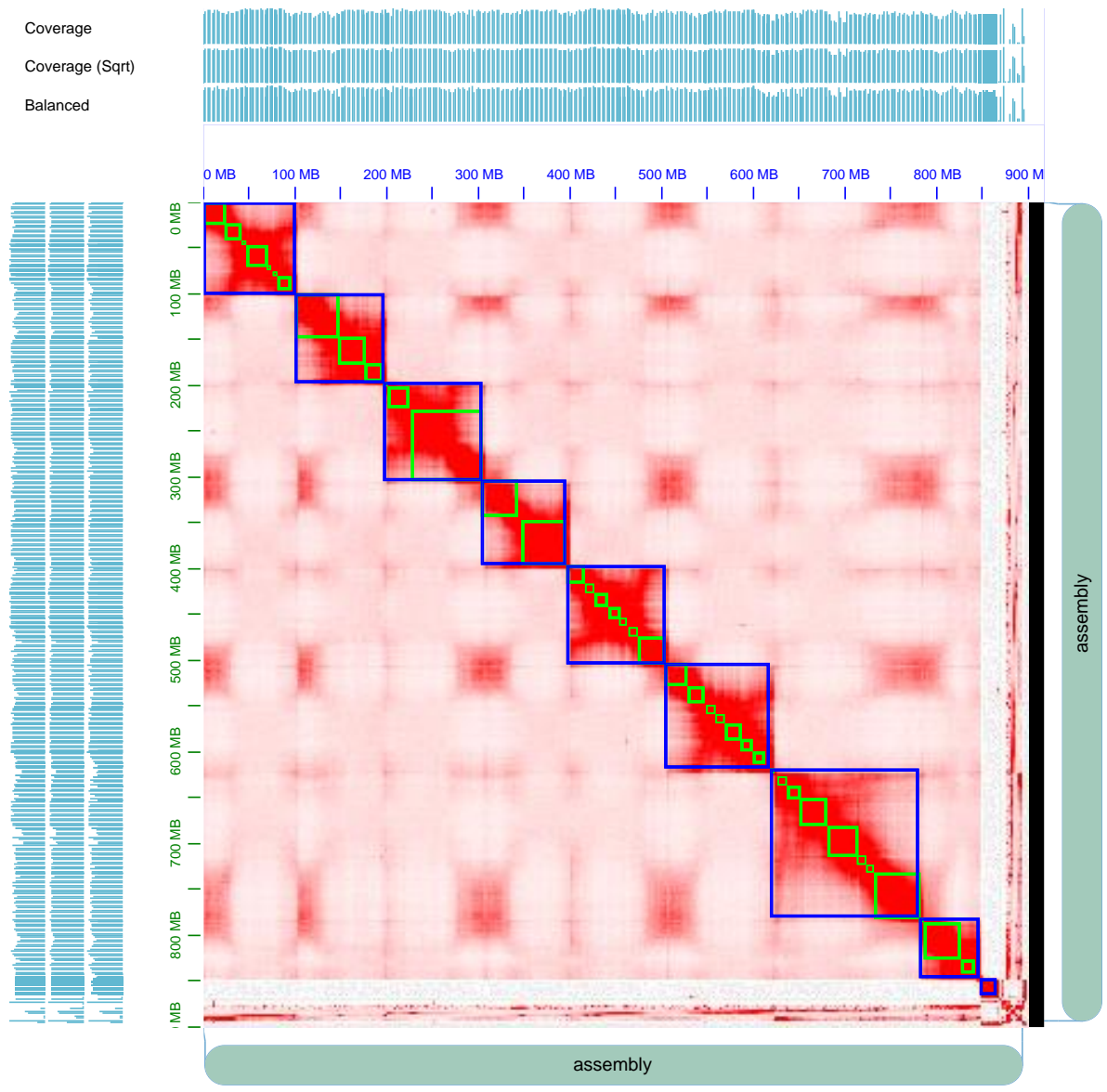


Figure 3-2 Hi-C contact heatmap

Heatmap visualization of Hi-C contact map using Juicebox. Blue squares highlight assembled chromosomes. Green squares show the scaffolds.

Astri105.blobtools.plot.Astri105_blobplot.blobDB.json.bestsum.phylum.p8.span.100.blobplot.bam0

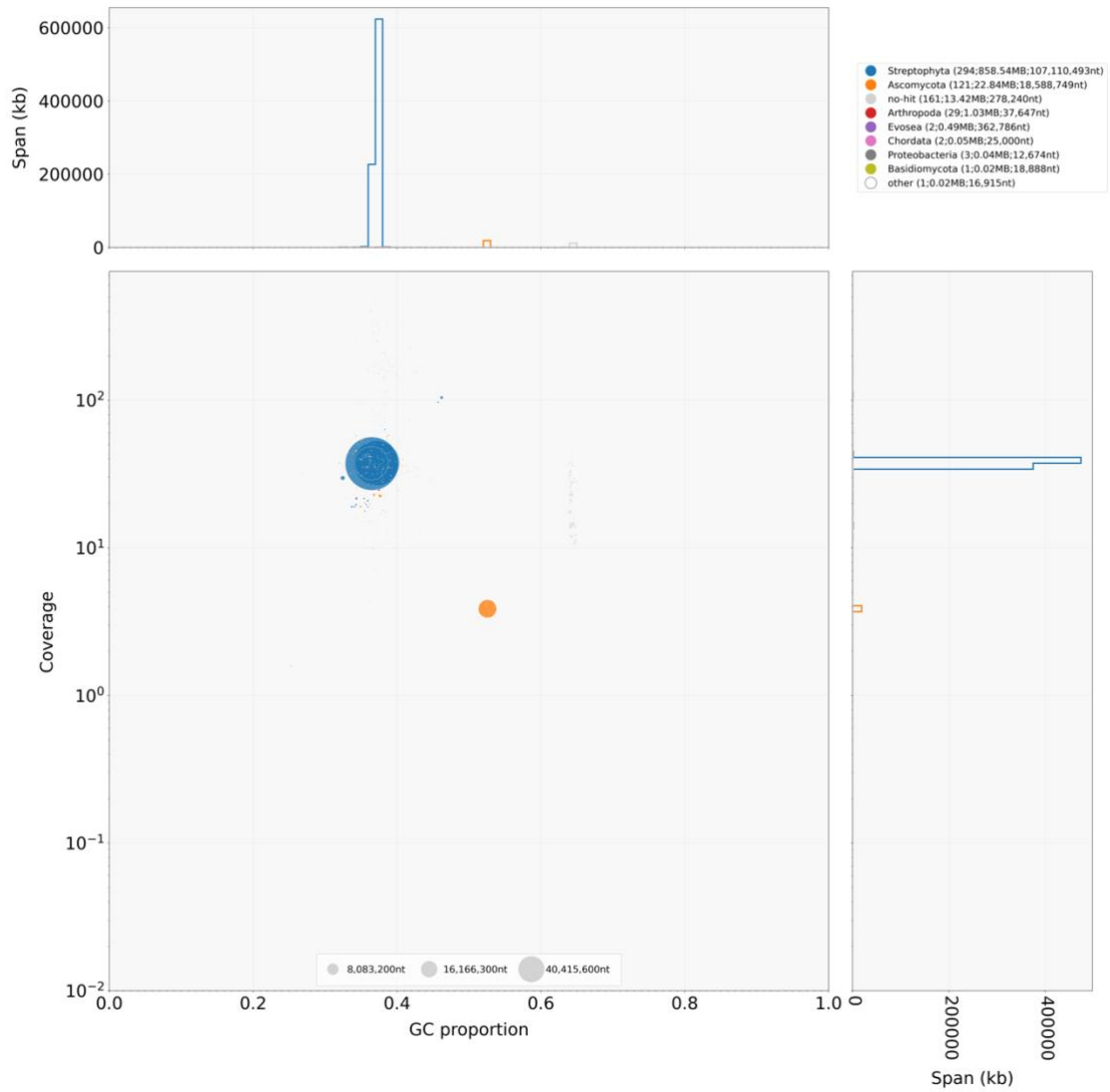


Figure 3-3 Blob plot

Blob plot of phyla identified in the HiFi sequencing reads. Large blue dot is in central plot and blue histograms represent a Streptophyta, this is the expected plant species. The orange dot and histograms represent the contaminated reads from an Ascomycota phylum member

Annotation

Repeat sequences

The pawpaw genome was composed of 40.48% identified repeats, less than the 54.47% of the *A. muricata* (soursop) genome (Strijk et al., 2021), or the 63.81% repeat composition of the *Liriodendron chinense* genome (Chen et al., 2019). The most abundant repeats were the LTRs, these comprised 39.19% of the assembly, less than the 41.28% of soursop and the 56.25% in *L. chinense* genome. Among the LTR retrotransposons, Gypsy elements are the most abundant (27.29%) followed by copia (9.62%). The DNA transposon element content is 0.52%, markedly less than soursop (7.29%) or *L.chinense* (12.67%) [Figure 3-4b].

Gene model annotation

A set of 53,904 genes models and 246,511 exons were constructed by BRAKER with support from mapped RNA-sequencing reads and protein evidence. BUSCO evaluation of the predicted gene models identified 85% of expected genes with 71.1% single copy genes and 14.1% duplicated genes present. MAKER, on the other hand, derived a set of 32,440 genes and 165,884 exons supported by the BRAKER trained AUGUSTUS, and additional evidence mentioned above. BUSCO scores for the MAKER showed 82.7% of single copy ancestral genes were correctly modelled, similar to BRAKERs 85.6%, but with considerably less duplicated gene models, only 2.9%. AED evaluation of both gene sets was over 0.5 for more than 80% of the predicted proteome [Figure 3-5]. This shows that there is a low congruency between the overlapping evidences provided and each annotation (Holt and Yandell, 2011; Yandell and Ence, 2012), suggesting there are several issues with automated annotation and may require manual curation to improve. However, the poor AED scores for the annotations produced by BRAKER, MAKER, and the fourth run of SNAP, were near identical which may point more towards insufficient evidence provided for gene prediction. To explore this, a

BUSCO analysis of genes modelled using only single contributing evidence provided to MAKER was performed. The results revealed ESTs contributed to only 40% of expected ancestral gene being identified and the annotation of the genome would likely be improved by providing further RNA-sequencing.

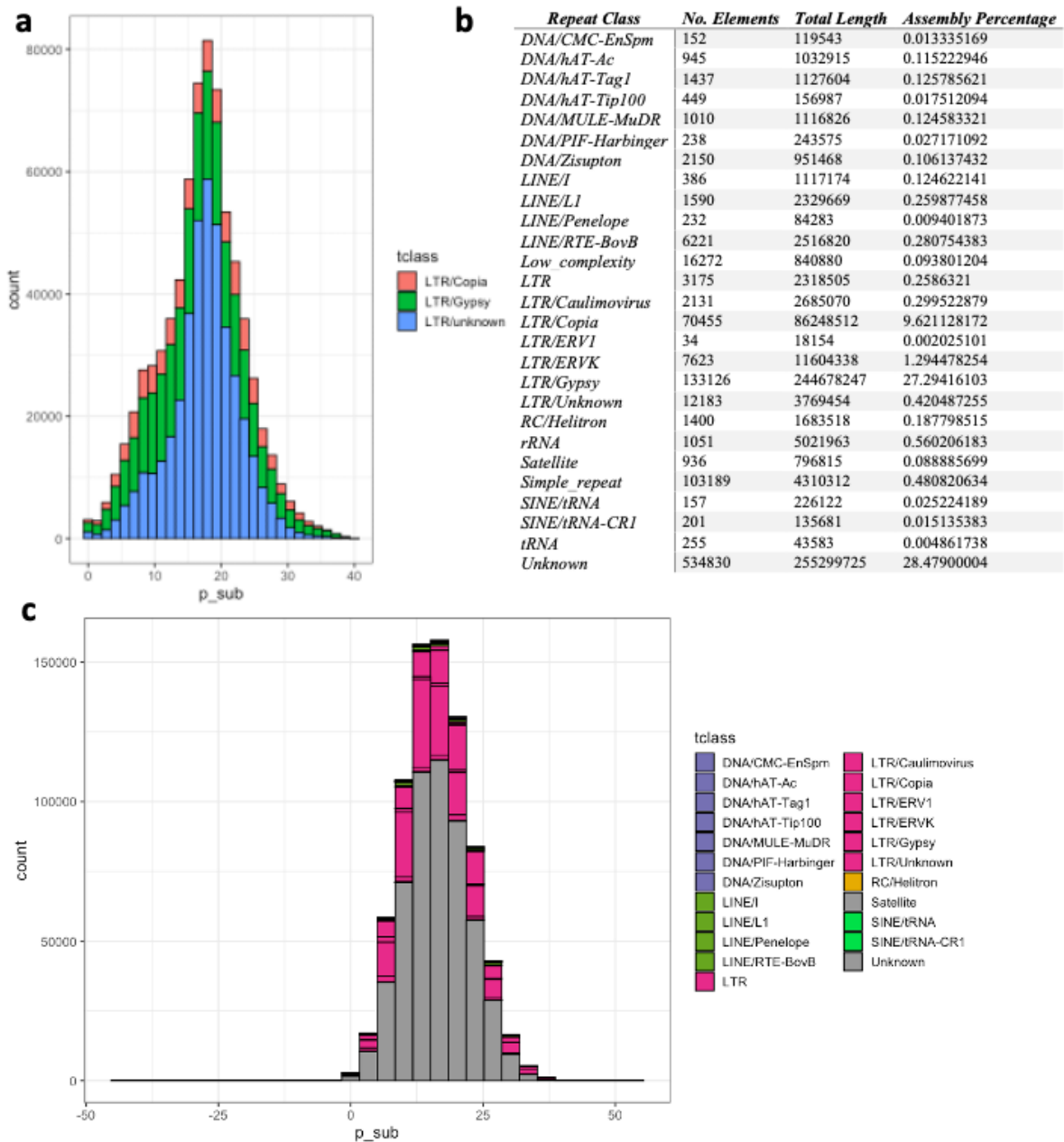


Figure 3-4 Assembly repeat content

Panel showing the repeat content of the assembled genome. **a)** Histogram of most common repeat types by occurrence **b)** Table summarizing elements by length and proportion of the assembly. **c)** Histogram of all repeat classes by occurrences.

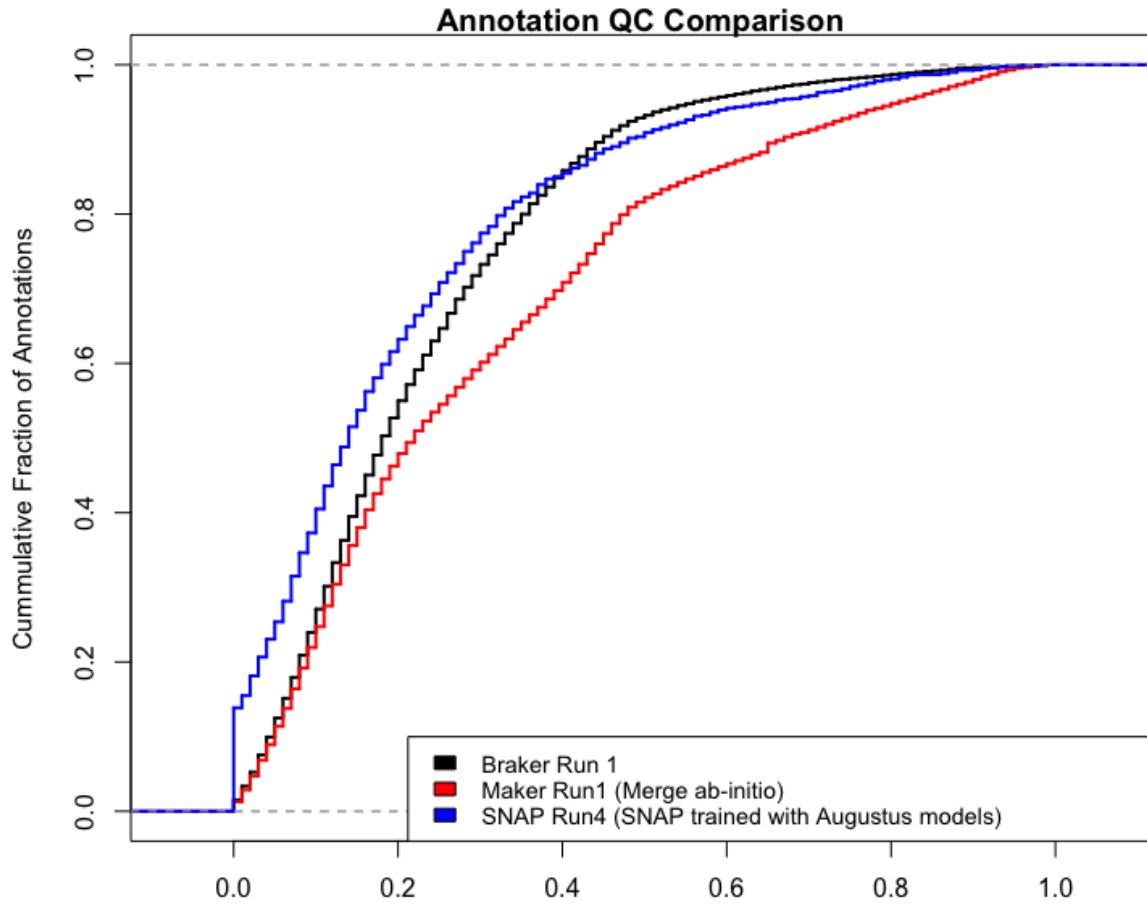


Figure 3-5 Gene model AED comparison

Annotation edit distance (AED) comparisons. X-axis is percentage of the total annotation. Y-axis is the agreement of annotation model with the evidence provided for prediction of the annotation. The evaluation is based on the number of changes (edits) that are needed to transform the sample annotation into the reference annotation (evidence provided). This includes adding, deleting, and substituting annotation elements, such as words and phrases. The lower the edit distance, the closer the sample annotation is to the reference annotation.

Table 3-3 Evidence contribution to gene prediction

Table summarising annotation models by each predictor and the contribution of each type of evidence to MAKER. **Top rows:** show benchmarking universal single-copy orthologs (BUSCO) evaluations of genes predicted by BRAKER, MAKER, AUGUSTUS, and semi-HMM-based Nucleic Acid Parser (SNAP). The non-overlapping column refers to MAKER consensus models but includes the non-overlapping predictions. **Middle rows:** BUSCO gene space completeness scores for models predicted by MAKER using only the evidence indicated. **Bottom rows:** Summary of predictions by BRAKER, MAKER, and StringTie.

Ancestrally conserved gene by model		Braker	Maker	non-overlapping	Augustus	Snap
BUSCO	% Complete	85.2	85.6	90.3	38.6	46.8
	% Single	71.1	82.7	70.5	35.4	44.7
	% Duplicated	14.1	2.9	19.8	3.2	2.1
	% Fragmented	8.1	5.7	5.4	23.9	23.8
	% Missing	6.7	8.7	4.3	37.5	29.4
Evidence contribution		Stringtie Transcript	est2genome (Stringtie fasta)	prot2geno	prot2geno (Annona muricata)	
BUSCO	% Complete	74.4	40.5	86.3	76.8	
	% Single	52.2	38.2	83	73	
	% Duplicated	22.2	2.3	3.3	3.8	
	% Fragmented	10.3	24.2	7.2	9.5	
	% Missing	15.3	35.3	6.5	13.7	
Genetic elements by gene prediction model		BRAKER	MAKER	StringTie		
	CDS	246,498	163,369			
	exon	246,511	165,884	197,394		
	gene	53,904	32,440			
	transcript	56,584	32,440	40,374		
	Transcript / gene	1.05	1			

Gene cluster orthologues

An orthologous cluster analysis of the predicted *A. triloba* genes was run and comparing pawpaw gene clusters with *A. muricata* and *P. americana*, using OrthoVenn2 (Xu et al., 2019). OrthoVenn2 identified 906 gene clusters unique to *A. triloba*, 2,183 shared with *A. muricata*, and 903 with *P. americana* [Figure 3-6]. Orthologous genes are clusters of genes descended from a common ancestor (Altenhoff et al., 2013; Xu et al., 2019), and so it is unsurprising to see a higher number of shared clusters between the much closer related *A. muricata*, a species in the same family, than with *P. americana*, a member of a closely related order. Within the unique *A. triloba* gene clusters there may be genes related to the species adaption to colder temperate climates, however, a deep exploration of the pathways or processes these genes participate in will need to wait until more RNA-seq can be provided to produce an improved annotation.



Figure 3-6 Orthologous cluster analysis

Venn diagram of *A. triloba* orthologous gene clusters shared with *A. muricata* and *P. americana*

Conclusion

In conclusion, we have presented an updated *de novo* assembly of the *A. triloba* genome using HiFi sequencing reads and Hi-C contact map. Our assembly demonstrates high accuracy and completeness, with a contig N50 of over 100 Mb and over 90% of the genome covered by eight scaffolds. The use of Hi-C contact maps in the assembly process helped to improve the continuity of the genome and the accuracy of the assembly. This new version of the *A. triloba* genome represents an improvement over the previous version of the genome in terms of accuracy and completeness and provides a valuable resource for future genetic, and genomic studies of pawpaw. Further, it contributes to the study of evolution in Annonaceae, and Magnoliids, a very underrepresented group.

The use of HiFi sequencing reads and Hi-C contact maps in the assembly process highlights the power of these technologies for improving the quality of *de novo* assemblies.

Chapter 4 : Genetic diversity study of *A. triloba* across its native range.

Introduction

The study of genetic diversity within a species is crucial for understanding the evolutionary processes affecting the species and can also have practical applications in conservation and breeding programs. In this chapter, we present a comprehensive analysis of the genetic diversity of *A. triloba* using a combination of genotyping by sequencing and the improved genome assembly from chapter 3.

Asimina triloba [L.] Dunal (Pawpaw) is a member of the Annonaceae family (the Soursop family), a family around 130 genera and more than 2,500 species distributed across tropical to subtropical regions (Angiosperm Phylogeny Group, 2016; Erkens et al., 2022). The *Asimina* genus is unique as it the most northerly representative of the family (Berry, 1916; Erkens et al., 2022). The genus contains eight to ten species, and several possible hybrids, most of which are restricted to the state of Florida (Horn, 2015; Kral, 1960; Zimmerman, 1941).

A. triloba is the most widespread of the *Asimina* genus; found in 26 states covering the entire Eastern coast of the United States of America, its range extending into Southern Canada (Darrow, 1975; Fox, 2012). It is thought to have migrated north from a refugia in the Gulf of Mexico after the last glacial maximum (Wyatt et al., 2020). For a more detailed review of the species see Chapter 1. North America was dominated by icy temperatures for the 2.4 Ma of the Pleistocene, followed by the Holocene, a warm inter-glacial period. Under these drastically different and shifting conditions, many tree species of North America evolved and adapted, leading to a geographical differentiation of many tree populations (Hewitt, 2004; Ony et al., 2021; Roberts and Hamann, 2015).

Genome-by-sequencing (GBS) is a cost-effective and efficient method for genotyping large numbers of individuals, allowing for the identification of single nucleotide polymorphisms (SNPs) and other genetic variations within a population (Elshire et al., 2011) that can be used to study the evolutionary history of a species. The quality of the variants called from GBS reads can be significantly improved by the use of the high quality reference genome, increasing the number of informative SNPs available for population studies (Bohling, 2020; Friel et al., 2021).

Using the GBS and the improved genome assembly data, we conducted a detailed analysis of the genetic diversity within a population of *A. triloba* trees from across its native range. Our results revealed a low level of genetic diversity among populations, but we also show that the population was divided into two distinct genetic clusters, which likely reflect the different routes taken during *A. triloba*'s post-glacial migration. Our study confirms previous findings and provides new insights into the genetic diversity of *A. triloba*.

Materials and methods

Many methodologies applied in this study are the same or similar to those used in Chapter 2, to avoid needless repetition these will only be briefly described here with a reference to the relevant section in chapter 2 materials and methods section for more detailed information.

Sample material

Genetic data for the analysis came from a total of 362 leaf tissue samples [Table 4-1]. This included the raw Illumina paired reads of samples previously studied Virginia populations [see chapter 2] and new samples collected from Boyce Thompson Institute's experimental orchard in Ithaca, New York. Additional samples were collected from wild populations by volunteers

and sent to Virginia Tech (VT), Virginia, for DNA extraction and library preparation. This network of volunteers allowed us to collect a large number of samples from diverse geographical locations, despite the logistical challenges and travel restrictions posed by the pandemic. To enlist their help collecting samples we provided everyone with a set of instructions to collect and preserve the samples during shipping to the lab in VT. After sequencing, mapping, variant calling quality controls (QC) a total of 329 samples remained, including pawpaw trees from 19 different USA states, and Ontario, Canada [Table 4-1] and samples of *Asimina parviflora* to use as an outgroup.

GBS library construction

Samples from volunteers were shipped with tissue paper between each leaf to prevent mold build up during shipping and stored at -20 °C until extraction. Total genomic DNA was extracted from 1 g of leaf using the DNeasy Plant Minikit (Qiagen). DNA quantity and quality was assessed with a NanoDrop One (ThermoFisher) and a Qubit 2.0 (Invitrogen) with the dsDNA HS Assay. GBS library preparation was following a protocol adapted from (Elshire et al., 2011). Preparation steps are described in chapter 2. In brief, DNA concentration was quantified using a Thermo Invitrogen Qubit 2.0 Fluorometer and 100 ng aliquots of each sample. Subsequently, the aliquots of DNA were fragmented by digestion with restriction digestion enzyme ApeKI (New England Biolabs, Ipswich MA), during 2 hrs at 75 °C in a T100 Thermocycler (Bio-Rad Laboratories, Inc). The 225 samples were divided into 3 libraries to assign each sample one of the 96 available unique Illumina barcodes. Sample specific “barcode” and common adapter sequences were ligated to sticky ends with T4 ligase (New England Biolabs, Ipswich MA) during a 1 hr incubation at 22 °C followed by a 65 °C heating step for 30 min in a T100 Thermocycler. 5 µl of each ligated sample was pooled into a single library and amplified by PCR. PCR steps are described in chapter 2. Following amplification,

the three libraries were purified with a Monarch PCR & DNA Cleanup Kit (New England Biolabs). Fragment size selection in the range of 170–350 bp was then implemented via BluePipin (Sage Science). Libraries were sequenced at BGI Genomics Cambridge, MA, with one lane of HiSeq2500 Illumina system as 2x150 bp.

Raw read processing, mapping and variant calling

Samples from the two separate sequencing projects were merged into a single data set for this population analysis. The Virginia only population from the previous GBS library preparation, here after called VApop, and samples combining individuals from across the native range and the BTI experimental orchard, here after called NRpop. To incorporate both data sets while minimizing potential “batch” effects (Tom et al., 2017), the demultiplexed raw reads from VApop and NRpop data sets were processed together (Tpop).

Processing of the raw sequencing reads were performed as described in chapter 2 materials and methods section. In summary, the raw reads were de-multiplexed with GBSX.v1.3 (Herten et al., 2015). Next, Illumina adapters, low quality, and short reads were removed with FASTQ_MCF v1.05 (Aronesty, 2013). After cleaning, the processed reads were mapped to the *A. triloba* reference genome Astri105. Before mapping, Astri105 was indexed with the Burrows-Wheeler Alignment Tool (BWA).v0.7.17-r1188t (Li, 2013) with default settings. BWA outputs the reads with mapping information into SAM format files which were converted to BAM and sorted by coordinates via samtools view and samtools sort respectively SAMTOOLS.v1.7 (Li et al., 2009). The sorted BAM files, one per sample, were then merged into a single BAM file with BAMADDRG (Garrison, 2022). SNPs were then called from the merged BAM with FREEBAYES.v.1.3.1-16-g85d7bfc (Garrison and Marth, 2012).

SNP filtering

SNPs were filtered with VCFTOOLS.v.0.1.15 (Danecek et al., 2011) keeping only biallelic SNPs, removing indels, removing SNPs a minimum read depth less than five and with a mean depth less than 20. A minimum SNP QC of 30 was also applied. Loci with >10% missing data were also removed. Due to high amounts of missing data across individuals, SNP data sets were prepared with no missing data, 5% missing data and 35% missing data to assay performance by principal component analysis (PCA) following recommendations in Yi and Latch (2022). All SNP datasets were filtered for linkage disequilibrium (LD) with PLINK.v1.90b4 (Purcell et al., 2007). LD filtering involved removing SNPs based on a 10 Kb window with independent pairwise filtering and a variant shift count of 5 and r^2 value of 0.2. Mapping statistics are detailed in Table 4-2.

Inference of population structure

Potential clones were previously identified within the Virginia population, this was also examined across the combined dataset Tpop using the same KING relationship inference algorithm (Manichaikul et al., 2010) method run via the relatedness2 option in VCFTOOLS. The resulting phi score for each pairwise comparisons was then converted into a matrix to produce a heatmap of clonality in R with the package GGLOT2.v.3.3.5 (Wickham, 2016) [Figure S4-1].

As with the previous study in Chapter 2, inference of population structure was approached using multiple methods, STRUCTURE, PCA and discriminant analysis of principal components (DAPC) and Neighbor-joining (NJ). This was done to provide a more robust estimation by cross-validation with differing approaches in an attempt to avoid potential biases

that can be introduced when evaluating allele frequency, or assuming a HWE. FASTSTRUCTURE (Raj et al., 2014), a more resource efficient version of the STRUCTURE Bayesian multilocus genotype clustering (Pritchard et al., 2000) was run as part of the R package LEA.v.3.6.0 (Frichot and François, 2014). Additionally, admixture coefficients were also produced from the FASTSTRUCTURE outputs of LEA. Both PCA and DAPC methods of clustering do not require a population to be in Hardy-Weinberg equilibrium (HWE) and do not make any prior assumptions about the population. Both analyses are implemented in the ADEGENET.v.2.1.5 R package (Jombart et al., 2010; Jombart and Ahmed, 2011). The most appropriate number of principal components retained to analyze for maximum variance while avoiding an overfit were found by using function *optim.a.score()*. DAPC was then run retaining an optimal number of PCs (11), with the default number of discriminate analysis axes (100).

As in chapter 2 a Neighbor-joining (NJ) tree based on allele frequencies was generated with the R package POPPR.v.2.9.3 (Kamvar et al., 2014). Genetic distances were calculated using the fraction of different sites between samples with *bitwise.dist()* and run with 100 bootstrapping support. Visualisation of the NJ tree was achieved using the package APE.v.5.6-1 (Paradis and Schliep, 2019) and Figtree.v1.4.4 (Ramaut, 2018).

Analysis of molecular variance (AMOVA)

An analysis of molecular variance (AMOVA) can be a powerful tool that can help support a hypotheses of population structure due to clonal reproduction or isolation without making assumptions about HWE (Excoffier et al., 1992). An AMOVA was performed using the R package POPPR.v.2.9.3 (Kamvar et al., 2014) comparing the ‘East’, ‘West’, and ‘Cultivar’

assigned groups. An AMOVA was executed with and without the clone correction option to explore a potential influence of clones on population structure. Variation significance was validated using ADE4.v.1.7-18 (Dray and Dufour, 2007) using the function *randtest()* with 999 permutations.

Isolation-by-distance (IBD)

The presence of any geographic separation influencing observed population structure was assayed by an Isolation-by-distance (IBD) analysis. To perform the IBD, a matrix of geographic distances and genetic distances was generated using ADEGENET.v.2.1.8 (Jombart, 2008) function *dist.genpop()*. A Mantel test was then performed on the matrix using the R package ADE4.v.1.7-18 (Dray and Dufour, 2007) with 999 permutations and a simulated p-value of 0.001.

F-statistics and heterozygosity

Wright's F-statistics and heterozygosity were calculated to analyse the effects of genetic drift within the three identified groups. The F-statistics of genetic diversity estimates were performed using the R package; these included F_{ST} , inbreeding coefficients (F_{IS}), as well as the number of segregating sites, nucleotide diversity, Watterson's theta and Tajima's D POPGENOME.v.2.7.5 (Pfeifer et al., 2014). F_{ST} was also calculated on the assigned groups using DARTR.v.2.7.2 (Gruber et al., 2018) for cross validation.

The observed (H_o) and expected (H_e) heterozygosity for each of the three assigned groups (East, West, Cultivar) was calculated with two available software tools for comparison, these were Analysis of Next Generation Sequencing Data (ANGSD) (Korneliussen et al., 2014) which uses the R package DARTR (Gruber et al., 2018). DARTR uses a genlight object

containing the filtered SNPs while ANGSD uses bam files containing all unfiltered sites to estimate the site frequency spectrum and the probability of each genotype at each site. The observed heterozygosity (H_o) is then calculated as the proportion of sites that are heterozygous. DARTTR calculates the genotype likelihood for each loci under HWE for each individual in the population and averaging it to give the expected heterozygosity. These genotype likelihoods are then used to infer the most likely genotypes and calculate the observed heterozygosity as the proportion of sites that are heterozygous.

Results and Discussion

Sequencing and variant calling

Sampling material from wild populations from such a large area as pawpaw's native range (26 states and parts of southern Canada) is a difficult task, particularly during the limited time frame of this project. To be able to continue with the project it was necessary to develop a network of people already in the US and with the knowledge and access to wild pawpaw trees. To this end the Pawpaw Network was established, this was a group of volunteers spread out across the many US states and who had either worked commercially, or on an ecological bases with the tree. It also included hobbyists, foragers, and gardeners all of whom had enough familiarity to correctly identify the species. Thanks to the help of the Pawpaw Network, we were able to collect a total of 362 individual samples from 22 states covering almost the entire native range of the species [Table 4-1]. However, after sequencing, mapping, and variant calling quality controls, 33 low-quality samples needed to be removed and a total of 329 samples remained, including pawpaw trees from 19 different USA states, and Ontario Canada. It is unfortunate that we had to lose so many samples, but this is something of a risk with a GBS approach due to the fragmentation step in the library preparation. The enzymatic

fragmentation of genomic DNA can lead to unequal sequence representation in the sequencing library. In addition, the library preparation can introduce errors from pipette techniques and form PCR biases (Chen et al., 2013; Elshire et al., 2011; Friel et al., 2021; Shendure et al., 2017). Another concern, for two main reasons, is the quality of the samples used. First, because high-molecular weight is needed at the start of the library preparation, and degraded DNA will be randomly fragmented before the digestion step in the library preparation. This results in an increase of smaller fragments that are removed during the size selection step. Secondly, because contamination from cell necrosis and possible fungal and bacterial infection of decaying material can introduce impurities that reduced the DNA extraction quality and the efficiency of the restriction enzymes, leading to partial digestion and reduced genome fragmentation. Since the sampling by both the research team and the volunteers were conducted under different conditions, during different times of the growing season, with shipping times varying by state, the overall condition of the samples was not consistent, with some samples already beginning to decay. Due to travel and time restrictions on the project, sequencing was attempted with all samples regardless of quality variations, and likely accounts for the number of low-quality samples that were subsequently removed.

The raw sequencing reads were demultiplexed and filtered by length and base calling quality before combining all samples (BTI, Harvard Arboretum, and all wild samples not from Virginia) with the processed Virginia samples used in Chapter 2. The combined samples were mapped to the new HiFi and Chromatin-association/interaction analysis (Hi-C) reference genome *Astri105*. On average there was around 10.1 million reads per sample with an average of 6.5 million reads (72.4%) mapping to the genome. However, the standard deviation was 8.6 million due to the variation in sample quality, the low-quality samples being removed at various proceeding filtering steps. The total number of shared sites in the merged bam file was

5,170,794, significantly more than the number of sites in the previous study (504,887) using the Astri041 reference genome and Virginia only data set.

SNP filtering

The sample, TN_CumbR-4-6 for example, had only 14,954 mapped reads compared to the average of 6.5 million and so was discarded. This represents a sample that may have had highly fragmented or degraded DNA at the time of library preparation, resulting a low number of reads mapped to the reference. Indeed, restriction site-associated DNA sequencing (RADseq) libraries produced with low-quality DNA templates have been seen to generate more non-random missing data, as much as 60%, than is likely from allele dropout due to natural mutation (Rivera-Colón et al., 2021). Given that both RADseq and GBS are two very similar reduced representation methods which rely on restriction digestion enzymes to fragment template DNA before sequencing (Andrews et al., 2016; Baird et al., 2008; Elshire et al., 2011), it is reasonable to assume the same has occurred in our GBS libraries.

Thus, samples like this needed to be removed because of the probability that they would have much less coverage of the genome, a lower number of mapping sites, and a greater percentage of missing data reducing the overall number of variants available for analysis after all SNP filters are applied, also, they might introduce non-random missing data; potentially biasing the analysis. Considering this, we removed several samples with significantly low mapping results before calling and filtering SNPS with VCFTOOLS. However, the data set still contained a high proportion of poorly sequenced individuals with many variants present in less than 50% of the population. Once all filters listed in material and methods, with no allowance for missing data, were applied, around 300 SNPs remained. Potentially these may have been informative enough to differentiate between individuals and detect population structure, yet, increasing the

number can significantly raise the probability of being able to correctly assign a particular individual to its origin group (Turakulov and Easteal, 2003). We observed that the number of variants most significantly dropped when no missing data filter was applied. In the case where all individuals were retained, this filter reduced the hundreds of thousands of SNPs to approximately one hundred.

However, it is worth noting that the best practice for inclusion or exclusion of missing data, missing loci, or poorly sequenced individuals during population or phylogenetic studies are a matter of some debate (Arnold et al., 2013; Huang and Knowles, 2016; Wilkinson, 1995; Yi and Latch, 2022). Huang and Knowles (2016) state that overly conservative filtering results in a loss of information, both from removing individuals, and “a biased representation of the mutation spectrum among screened loci”. Further, in the case of large *Drosophila* RADseq, data matrices having higher portion of missing data was not shown to adversely affect phylogenetic analysis (Rubin et al., 2012). While others found that missing data did impact the inference of population structure, and suggested tuning the optimal filtering parameters to suit each study (Arnold et al., 2013; Wright et al., 2019).

In our case, retaining a large set of variants while keeping as many individuals as possible certainly meant allowing more missing data. But a high percentage of non-random missing data can affect the individual group assignment and admixture estimates. By assaying various allowances of missing data it is possible to evaluate the effect this uneven haplotype sampling has on the group assignment (Yi and Latch, 2022). Therefore, several rounds of filtering and removing individuals were preformed to optimize the filtering steps and maximize the total number of informative SNPs available for the greatest genetic resolution while retaining the greatest number of individual pawpaw samples. We found that the best practice was to remove samples with extremely poor mapping quality then remove all loci appearing in less than 10%

of the data set, before applying a minimum read depth, mean depth, minimum QC filters, and LD. We then generated multiple VCF files with 0% missing data allowance resulting in 8,759 SNPs, 5% missing data for 5,811 SNPs and 35% missing data for 347 SNPs. These were compared by PCA as recommended by Yi and Latch (2022). In doing so, we could see that there were two clear clusters regardless of missing data allowance [Figure 4-1]. 0% missing data allowed for PCs 1 and 2 to explain 10% of total variance, while the 5% allowance explained 8% total variation and 35% allowance explained 7.9%. All samples were coloured using a gradient of missing data percentage for that individual prior to filtering, this allowed for the group assignment comparison, showing where an individual clustered with varying missing data allowance. We could see that the overall clustering was not affected by percentage of missing data. However, the location within the cluster did vary for some samples. For example, individuals from a patch Ohio moved further along the PC2 away from their cluster with inclusion of missing data, while cultivars spread out more within their cluster when no missing data was allowed. Although, in all cases, PC2 only accounted for 2–3% of total variation and this loss of accuracy is not likely to affect the overall structure of the population. The results of this preliminary testing showed that the filtering choices may have impacted on inferences about individual assignment and admixture estimations, as previously suggested (Yi and Latch, 2022), but it is unlikely to have had a significant impact on overall group comparison analyses. After filter optimisation, we proceeded with a 5% missing data allowance and retained 329 individuals with 5,811 SNPs.

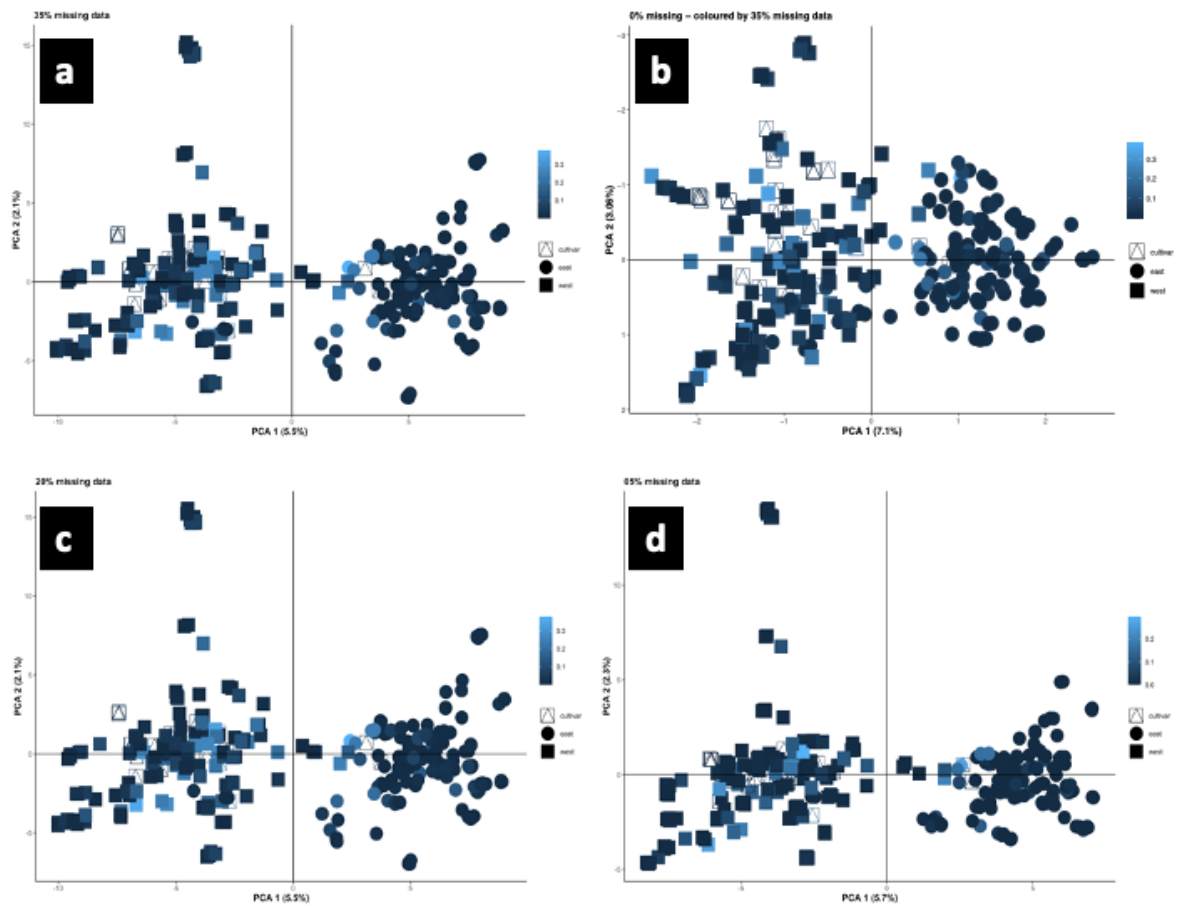


Figure 4-1 Missing data assessment

Panel showing the comparison of clustering performed by PCA using dataset containing various amount of missing data (**a** = 35%, **b** = 0%, **c** = 20%, **d** = 0.5%). To compare the impact of missing data on clustering across all datasets, we coloured each sample by the percentage of missing data it contained, 35% missingness is set as the limit.

Table 4-1 Sampling locations before and after QC

Summary table of all sampled locations and those remaining informative high-quality samples after all strict application of quality controls (QC). Here, QC includes all controls and filtering steps applied at all stages from the processing of the raw reads (base calling errors, read length, etc.) to variant calling filters (percentage of missing data, minimum read depth, linkage disequilibrium, etc.).

	Before QC	After QC
Total	362	329
States	22	20
Wild	320	295
Harvard Arboretum	6	5
VA Nursery	3	1
BTI Orchard	33	28
Alabama	3	3
Arkansas	1	1
Florida	1	1
Illinois	1	1
Indiana	37	35
Kansas	1	0
Kentucky	6	6
Louisiana	8	7
Maryland	12	12
Missouri	1	1
Michigan	13	11
New Jersey	5	5
New York	1	1
North Carolina	5	5
Ohio	57	44
Oklahoma	3	3
Ontario	2	2
Pennsylvania	27	23
South Carolina	1	1
Tennessee	13	9
Virginia	137	134
West Virginia	1	0

Table 4-2 GBS read processing

Read processing and mapping results. All samples included at the beginning of the project are shown. Raw: The total number of raw sequencing reads after demultiplexing. Pprocessed: The total number of reads after minimum base calling score 30 and sequence length of 50 bp filter applied. Mmapped: The total number of the processed reads that successfully mapped to the reference genome. %mapped: The percentage of the processed reads that mapped to the reference genome. Ssites: The total number of sites where processed reads have mapped to in the reference genome. %sites: The percentage of mapping sites shared with all other individuals.

Sample name	raw	processed	mapped	% mapped	sites	% sites
AAHU122279A-1	5,277,382	4,912,750	3,816,444	77.68	232,412	4.49
AAHU122279A-2	8,385,193	7,125,773	7,001,602	98.26	299,219	5.79
AAHU12708A	2,340,315	2,024,307	1,992,799	98.44	199,572	3.86
AAHU14394B	8,260,852	7,054,649	6,969,162	98.79	281,797	5.45
AAHU20591A	12,164,643	10,454,206	10,273,486	98.27	335,920	6.5
AAHUAA2522013A	10,072,212	8,417,320	8,305,844	98.68	302,254	5.85
BTI_1-2	33,001,572	29,488,252	10,200,858	34.59	577,749	11.17
BTI_1-23	3,061,558	2,849,470	1,927,189	67.63	146,875	2.84
BTI_1-68	100,886	91,750	58,740	64.02	21,223	0.41
BTI_1-7-1	14,913,814	13,927,452	13,512,194	97.02	207,795	4.02
BTI_1-7-2	14,280,454	13,343,998	6,757,860	50.64	347,141	6.71
BTI_10-35	20,616,750	19,187,734	9,232,201	48.12	405,005	7.83
BTI_11-13	11,409,816	10,385,648	4,112,982	39.6	314,604	6.08
BTI_11-5	6,951,776	6,457,926	5,428,876	84.07	198,862	3.85
BTI_2-10	3,652,076	3,473,980	3,068,466	88.33	151,702	2.93
BTI_2-3	18,616,574	16,638,436	7,979,172	47.96	418,493	8.09
BTI_2-54	13,879,150	12,869,580	8,440,063	65.58	294,883	5.7
BTI_3-11	4,326,004	4,022,728	3,553,532	88.34	182,175	3.52
BTI_3-21	21,203,188	20,135,460	16,650,185	82.69	316,492	6.12
BTI_3-3	14,335,802	12,892,928	7,095,968	55.04	383,510	7.42
BTI_4-2	14,013,304	12,991,590	8,494,775	65.39	295,412	5.71
BTI_5-5	4,867,454	4,484,834	3,664,677	81.71	189,403	3.66
BTI_7-90	21,103,164	20,040,250	16,700,730	83.34	308,226	5.96
BTI_8-20	10,164,632	9,418,774	7,721,962	81.98	273,519	5.29
BTI_8-58	4,188,844	3,976,210	3,570,263	89.79	155,298	3

BTI_9-47	14,644,708	13,580,480	10,943,751	80.58	273,547	5.29
BTI_9-58	3,451,704	3,268,406	2,614,779	80	157,625	3.05
BTI_Middletown	12,844,850	11,565,124	3,485,235	30.14	403,745	7.81
BTI_Mitchell	9,248,536	8,301,494	2,481,048	29.89	244,080	4.72
BTI_Munser	11,448,154	10,533,546	5,926,307	56.26	301,901	5.84
BTI_NC-1	14,587,582	13,528,202	10,883,601	80.45	274,289	5.3
BTI_Overleese	10,831,350	10,173,684	9,335,483	91.76	179,820	3.48
BTI_PA_Golden	2,210,572	2,043,516	1,288,866	63.07	157,464	3.05
BTI_Shenandoah	10,427,004	9,355,834	1,626,299	17.38	257,703	4.98
BTI_Sunflower	400,192	377,016	268,299	71.16	52,854	1.02
BTI_Taylor	18,484,018	16,517,576	7,944,909	48.1	418,399	8.09
BTI_Taytwo	12,088,424	10,992,592	4,500,078	40.94	312,849	6.05
BTI_Wells	7,826,488	7,391,510	7,237,041	97.91	163,468	3.16
BTI_Wilson	6,153,994	5,710,064	4,842,059	84.8	221,657	4.29
EdLandNurs01	12,277,734	10,111,477	9,994,836	98.85	287,471	5.56
EdLandNurs02	1,694,041	1,440,495	1,419,098	98.51	181,557	3.51
EdLandNurs03	3,431,953	2,866,810	2,819,760	98.36	237,080	4.58
VA_AndyLT-1-1	8,795,444	8,072,256	6,193,516	76.73	285,571	5.52
VA_AndyLT-1-2	10,674,318	9,575,442	6,565,121	68.56	347,213	6.71
VA_AndyLT-1-3	11,644,000	10,396,324	6,606,276	63.54	356,044	6.89
VA_AndyLT-1-4	19,118,070	17,448,132	12,675,513	72.65	420,458	8.13
VA_AndyLT-1-5	16,297,450	14,299,238	7,937,092	55.51	402,635	7.79
VA_AndyLT-2-1	11,501,146	10,406,326	9,062,267	87.08	240,695	4.65
VA_AndyLT-2-2	6,702,582	6,062,822	5,077,901	83.75	215,419	4.17
VA_AndyLT-2-3	2,825,098	2,640,890	1,596,955	60.47	192,034	3.71
VA_AndyLT-2-4	14,457,536	13,171,562	10,867,426	82.51	315,420	6.1
VA_AndyLT-2-5	15,062,362	13,598,930	10,718,920	78.82	371,404	7.18
VA_AndyLT-3-1	12,491,256	11,342,828	11,295,489	99.58	249,606	4.83
VA_AndyLT-3-2	15,835,428	13,823,048	5,088,108	36.81	339,993	6.58
VA_AndyLT-3-3	1,612,000	1,466,174	1,295,776	88.38	130,276	2.52
VA_AndyLT-3-4	17,708,106	15,729,770	11,326,010	72	366,252	7.08

VA_AndyLT-3-5	6,901,810	6,376,380	5,374,882	84.29	252,504	4.88
VA_AndyLT-4-1	12,709,486	11,457,788	8,912,400	77.78	325,412	6.29
VA_AndyLT-4-2	13,669,828	12,214,326	8,744,981	71.6	338,784	6.55
VA_AndyLT-4-3	13,358,210	12,016,936	9,502,475	79.08	334,961	6.48
VA_AndyLT-4-4	3,819,648	3,454,244	2,712,219	78.52	195,162	3.77
VA_AndyLT-4-5	8,553,614	7,852,140	6,984,533	88.95	258,545	5
VA_AndyLT-5-1	16,545,900	14,617,860	8,486,313	58.05	373,186	7.22
VA_AndyLT-5-2	10,749,214	9,756,626	7,781,198	79.75	298,710	5.78
VA_AndyLT-5-3	4,150,178	3,741,018	3,353,315	89.64	186,058	3.6
VA_AndyLT-5-4	11,166,864	10,180,938	9,291,760	91.27	276,585	5.35
VA_AndyLT-5-5	14,872,796	13,342,584	9,223,186	69.13	376,429	7.28
VA_DevilB-01-1	10,872,692	9,837,350	9,330,920	94.85	282,462	5.46
VA_DevilB-01-2	14,943,342	13,357,662	7,148,265	53.51	461,188	8.92
VA_DevilB-01-3	5,415,346	4,833,672	4,623,962	95.66	233,984	4.53
VA_DevilB-01-4	6,727,576	6,079,292	5,891,287	96.91	227,934	4.41
VA_DevilB-01-5	2,154,532	1,964,554	1,880,909	95.74	157,616	3.05
VA_DevilB-02-1	8,484,772	7,624,392	7,162,353	93.94	264,804	5.12
VA_DevilB-02-2	11,604,812	10,360,280	9,308,491	89.85	286,779	5.55
VA_DevilB-02-3	6,035,614	5,395,920	4,963,236	91.98	237,108	4.59
VA_DevilB-02-4	8,312,586	7,504,690	7,016,526	93.5	236,065	4.57
VA_DevilB-02-5	1,750,224	1,586,066	1,501,647	94.68	138,979	2.69
VA_DevilB-03-1	7,683,218	6,893,000	6,113,277	88.69	246,042	4.76
VA_DevilB-03-2	15,147,966	13,664,490	12,126,183	88.74	306,986	5.94
VA_DevilB-03-3	12,073,484	10,890,868	10,082,906	92.58	291,635	5.64
VA_DevilB-03-4	13,456,302	12,051,736	10,744,255	89.15	264,968	5.12
VA_DevilB-03-5	5,744,650	5,173,388	4,960,853	95.89	215,974	4.18
VA_FairyS-1-1	15,293,048	13,358,696	6,355,643	47.58	406,108	7.85
VA_FairyS-1-2	15,335,718	13,544,486	7,116,190	52.54	422,187	8.16
VA_FairyS-1-3	6,282,500	5,513,080	1,639,549	29.74	215,762	4.17
VA_FairyS-1-4	35,774,252	31,505,550	14,413,493	45.75	679,509	13.14
VA_FairyS-1-5	14,683,206	13,021,590	8,244,072	63.31	400,028	7.74

VA_FairyS-2-1	10,144,080	8,886,082	4,199,814	47.26	309,457	5.98
VA_FairyS-2-2	4,017,772	3,585,156	2,142,061	59.75	204,843	3.96
VA_FairyS-2-3	5,975,588	5,391,700	2,931,253	54.37	240,459	4.65
VA_FairyS-2-4	27,697,056	24,507,792	11,487,193	46.87	602,927	11.66
VA_FairyS-2-5	9,617,270	8,371,284	3,442,095	41.12	296,941	5.74
VA_FallRP-1-1	25,040,788	22,042,228	9,898,366	44.91	548,030	10.6
VA_FallRP-1-2	21,794,358	19,417,072	10,592,677	54.55	490,351	9.48
VA_FallRP-1-3	30,735,580	27,050,014	12,949,409	47.87	585,240	11.32
VA_FallRP-1-4	9,203,532	8,046,262	3,778,075	46.95	312,679	6.05
VA_FallRP-1-5	12,468,794	10,807,092	3,961,306	36.65	370,722	7.17
VA_HardR-1-01	3,744,766	3,310,574	3,258,492	98.43	184,500	3.57
VA_HardR-1-02	756,218	681,672	672,662	98.68	106,481	2.06
VA_HardR-1-03	2,008,970	1,811,822	1,780,278	98.26	148,660	2.87
VA_HardR-1-04	5,029,428	4,499,662	4,411,881	98.05	200,202	3.87
VA_HardR-2-1	6,262,840	5,730,298	5,497,327	95.93	219,198	4.24
VA_HardR-2-2	9,140,034	8,456,142	8,262,486	97.71	242,496	4.69
VA_HardR-2-3	8,124,542	7,384,956	7,256,440	98.26	218,254	4.22
VA_HardR-2-4	5,638,164	5,212,756	5,177,536	99.32	213,725	4.13
VA_HardR-2-5	14,250,248	13,088,352	12,818,543	97.94	284,126	5.49
VA_HardR-3-1	10,392,556	9,509,556	9,462,163	99.5	263,925	5.1
VA_HardR-3-2	7,663,746	6,927,182	6,725,568	97.09	242,240	4.68
VA_HardR-4-1	16,288,324	14,794,344	14,576,169	98.53	283,676	5.49
VA_HardR-4-2	15,299,764	14,096,538	14,024,000	99.49	288,051	5.57
VA_HardR-4-3	16,913,332	15,550,936	15,226,586	97.91	295,012	5.71
VA_HardR-4-4	11,097,420	10,150,478	10,125,750	99.76	249,013	4.82
VA_HardR-4-5	6,202,904	5,780,822	5,751,777	99.5	212,605	4.11
VA_HighBT-1-1	9,035,960	8,099,626	7,731,832	95.46	277,789	5.37
VA_HighBT-1-2	7,271,526	6,639,232	6,509,004	98.04	242,854	4.7
VA_HighBT-1-3	12,772,248	11,546,874	10,370,018	89.81	307,933	5.96
VA_HighBT-1-4	24,771,542	22,820,724	22,709,849	99.51	250,779	4.85
VA_HighBT-1-5	2,575,104	2,357,420	2,173,953	92.22	165,260	3.2

VA_JamesR-1-2	3,875,758	3,552,180	3,425,401	96.43	214,731	4.15
VA_JamesR-1-3	4,065,664	3,669,736	3,577,955	97.5	254,013	4.91
VA_JamesR-1-4	2,554,290	2,349,764	2,193,986	93.37	182,630	3.53
VA_JamesR-1-5	5,898,876	5,374,780	5,233,869	97.38	232,167	4.49
VA_JamesR-2-1	16,720,862	15,127,902	14,618,002	96.63	301,414	5.83
VA_JamesR-2-2	5,016,790	4,574,288	4,269,669	93.34	227,595	4.4
VA_JamesR-2-3	7,457,174	6,624,960	6,557,991	98.99	248,876	4.81
VA_JamesR-2-4	3,568,010	3,223,968	3,233,055	100.28	195,459	3.78
VA_JamesR-2-5	1,986,184	1,730,520	1,669,121	96.45	202,855	3.92
VA_NatB-2-01	3,855,230	3,531,504	3,134,560	88.76	191,794	3.71
VA_NatB-3-01	1,725,140	1,576,546	1,566,714	99.38	153,008	2.96
VA_NatTun-1-1	6,208,722	5,608,434	5,547,645	98.92	234,089	4.53
VA_NatTun-1-3	8,372,738	7,437,782	7,357,529	98.92	249,387	4.82
VA_NatTun-1-4	9,983,578	9,062,484	8,509,978	93.9	277,465	5.37
VA_NatTun-1-5	3,646,222	3,304,672	3,212,799	97.22	189,474	3.66
VA_NatTun-2-1	7,579,214	6,924,660	6,515,525	94.09	253,898	4.91
VA_NatTun-2-2	7,203,702	6,572,104	6,408,153	97.51	249,082	4.82
VA_NatTun-2-3	9,729,282	8,769,702	8,436,079	96.2	254,973	4.93
VA_NatTun-2-4	19,713,432	17,977,636	17,568,419	97.72	333,588	6.45
VA_NatTun-2-5	12,296,654	10,782,204	9,981,551	92.57	290,131	5.61
VA_NatTun-3-1	12,759,116	11,811,810	11,535,280	97.66	276,058	5.34
VA_NatTun-3-2	6,916,224	6,307,044	6,254,786	99.17	225,218	4.36
VA_NatTun-3-3	12,085,094	11,015,628	10,039,952	91.14	283,793	5.49
VA_NatTun-3-4	9,674,618	8,820,464	8,678,180	98.39	258,356	5
VA_NatTun-3-5	11,506,854	10,545,440	10,223,364	96.95	265,949	5.14
VA_StartP-1-01	23,483,214	21,359,662	19,404,786	90.85	375,325	7.26
VA_StartP-1-02	22,101,510	20,107,416	17,342,839	86.25	404,924	7.83
VA_StartP-1-03	20,430,226	18,548,936	16,020,111	86.37	362,809	7.02
VA_StartP-1-04	19,476,070	17,687,282	16,215,945	91.68	357,461	6.91
VA_StartP-1-05	17,997,862	16,553,990	15,212,941	91.9	357,838	6.92
VA_StartP-2-01	16,293,368	14,668,298	14,378,373	98.02	290,999	5.63

VA_StartP-2-02	5,745,568	5,223,932	4,748,093	90.89	214,764	4.15
VA_StartP-2-03	2,404,964	2,176,134	1,899,767	87.3	165,576	3.2
VA_StartP-2-04	5,059,524	4,613,766	4,384,864	95.04	200,004	3.87
VA_StartP-2-05	20,069,418	18,131,752	18,232,202	100.55	259,645	5.02
VA_StartP-3-01	9,045,356	8,139,344	7,900,303	97.06	232,024	4.49
VA_StartP-3-02	15,517,076	14,056,882	14,055,484	99.99	268,277	5.19
VA_StartP-3-03	7,172,866	6,495,152	6,085,389	93.69	220,699	4.27
VA_StartP-3-04	9,251,078	8,353,862	8,329,227	99.71	223,599	4.32
VA_TexasB-3-1	10,534,512	9,366,756	9,304,285	99.33	250,642	4.85
VA_TexasB-3-2	9,477,026	8,380,638	8,388,367	100.09	251,568	4.87
VA_TexasB-3-3	4,843,286	4,401,670	4,367,964	99.23	204,969	3.96
VA_TexasB-3-4	1,889,968	1,715,898	1,715,119	99.95	158,530	3.07
VA_TexasB-3-5	3,806,150	3,460,452	3,471,902	100.33	187,040	3.62
VA_TexasB-4-1	950,408	878,924	906,619	103.15	128,041	2.48
VA_TexasB-4-3	965,496	888,402	915,924	103.1	129,382	2.5
VA_TexasB-4-4	3,796,446	3,466,052	3,555,025	102.57	188,437	3.64
VA_TexasB-4-5	1,057,940	968,656	1,000,737	103.31	131,477	2.54
VA_TwinL-1-1	7,085,656	6,414,994	4,278,009	66.69	271,221	5.25
VA_TwinL-1-2	20,572,964	18,660,510	12,411,020	66.51	446,239	8.63
VA_TwinL-1-3	7,961,040	7,125,916	4,323,926	60.68	288,268	5.57
VA_TwinL-1-4	9,338,834	8,355,942	5,288,620	63.29	276,336	5.34
VA_TwinL-1-5	5,080,658	4,605,010	3,361,201	72.99	214,714	4.15
VA_YorkR-1-1	8,998,068	7,933,560	3,755,437	47.34	285,395	5.52
VA_YorkR-1-2	3,869,968	3,506,744	1,730,869	49.36	195,464	3.78
VA_YorkR-1-3	5,922,718	5,461,620	3,435,649	62.91	232,912	4.5
VA_YorkR-1-4	3,310,900	2,938,572	1,173,026	39.92	177,761	3.44
VA_YorkR-1-5	13,764,148	12,325,566	7,331,092	59.48	361,150	6.98
AL_HuntsV-1-1	14,704,388	13,754,828	5,153,412	37.47	469,093	9.07
AL_HuntsV-1-2	30,274,436	28,182,112	20,756,724	73.65	535,499	10.36
AL_HuntsV-1-3	18,707,826	17,164,264	10,342,468	60.26	504,451	9.76
AR_OuNF-1-1	11,464,672	10,629,610	7,437,616	69.97	330,432	6.39

FL_Leeburg-1-1	10,595,018	9,740,196	8,901,736	91.39	302,783	5.86
IL_Champ-1-1	6,903,610	6,307,204	4,750,382	75.32	271,274	5.25
IN_BrownC-1-1	6,737,476	6,023,454	3,436,630	57.05	255,973	4.95
IN_MucCP-2-1	19,458,126	17,864,494	7,798,988	43.66	473,466	9.16
IN_MucCP-2-2	7,002,988	6,395,852	4,795,172	74.97	273,196	5.28
IN_MucCP-2-3	26,443,110	24,338,120	7,118,882	29.25	711,507	13.76
IN_MucCP-2-4	19,373,772	17,786,622	7,776,263	43.72	472,716	9.14
IN_MucCP-2-5	3,920,894	3,432,558	1,530,359	44.58	208,427	4.03
IN_WillCMP-3-1	822,462	716,498	385,650	53.82	81,657	1.58
IN_WillCMP-3-2	7,266,186	6,540,646	4,167,787	63.72	281,642	5.45
IN_WillCMP-3-3	2,066,922	1,851,872	1,444,611	78.01	157,068	3.04
IN_WillCMP-3-4	12,814,270	11,509,314	5,071,269	44.06	473,164	9.15
IN_WillCMP-3-5	22,961,954	20,659,458	13,872,262	67.15	461,115	8.92
IN_CrookedC-4-1	12,639,396	11,356,036	8,817,374	77.64	341,934	6.61
IN_CrookedC-4-2	4,557,196	4,150,774	2,903,518	69.95	247,436	4.79
IN_CrookedC-4-3	12,277,626	11,101,498	8,325,308	74.99	353,593	6.84
IN_CrookedC-4-4	41,026,170	37,365,464	19,612,681	52.49	631,129	12.21
IN_CrookedC-4-5	12,417,962	11,349,658	9,669,137	85.19	325,836	6.3
IN_CrookedC-4-6	7,277,008	6,553,300	4,144,587	63.24	281,593	5.45
IN_Springhill-5-1	18,830,800	17,114,748	14,767,050	86.28	313,696	6.07
IN_Springhill-5-2	1,122,128	1,024,448	849,970	82.97	125,579	2.43
IN_Springhill-5-3	22,859,028	20,566,040	13,820,528	67.2	460,694	8.91
IN_Springhill-5-4	3,889,402	3,537,778	2,546,791	71.99	194,966	3.77
IN_Springhill-5-5	3,108,268	2,753,698	1,848,766	67.14	185,447	3.59
IN_Springhill-5-6	8,364,636	7,628,532	6,316,581	82.8	261,388	5.06
IN_Springhill-5-7	3,903,920	3,548,754	2,993,626	84.36	201,098	3.89
IN_Springhill-5-8	12,292,856	11,084,382	8,293,190	74.82	353,246	6.83
IN_CentralP-6-1	22,675,998	20,016,766	11,074,463	55.33	478,973	9.26
IN_CentralP-6-2	18,134,554	16,019,542	7,023,860	43.85	444,011	8.59
IN_CentralP-6-3	586,760	502,448	193,218	38.46	64,167	1.24
IN_CentralP-6-4	11,199,048	9,729,730	4,015,633	41.27	307,668	5.95

IN_CentralP-6-5	3,994,118	3,514,036	1,651,663	47	213,597	4.13
IN_HolidayP-7-1	5,657,162	4,768,862	1,512,782	31.72	221,010	4.27
IN_HolidayP-7-2	15,966,302	13,845,278	4,210,032	30.41	390,882	7.56
IN_HolidayP-7-3	24,533,718	21,749,952	10,133,218	46.59	472,436	9.14
IN_HolidayP-7-4	40,683,808	35,582,230	5,550,663	15.6	567,167	10.97
IN_HolidayP-7-5	39,337,128	33,905,586	12,119,076	35.74	568,107	10.99
KY_DupreeNP-1-1	3,084,488	2,633,302	1,622,871	61.63	152,835	2.96
KY_DupreeNP-1-2	4,260,932	3,713,940	1,809,987	48.73	182,247	3.52
KY_DupreeNP-1-3	5,895,790	5,127,510	1,817,469	35.45	209,156	4.04
KY_DupreeNP-1-4	13,357,740	12,093,770	6,189,504	51.18	359,646	6.96
KY_DupreeNP-1-5	9,164,212	8,208,112	5,455,279	66.46	205,535	3.97
LA_Hoges-1-1	4,493,310	4,148,456	3,614,406	87.13	221,864	4.29
LA_Tamar-2-1	15,291,308	14,117,064	13,560,361	96.06	313,531	6.06
LA_Layf-3-1	54,661,970	48,970,456	10,389,235	21.22	634,500	12.27
LA_Layf-4-1	34,128,990	30,719,074	15,431,513	50.23	612,060	11.84
LA_TangR-5-1	2,182,270	1,994,984	1,975,820	99.04	218,452	4.22
LA_OuachPB-6-1	9,488,112	8,162,202	7,839,524	96.05	255,766	4.95
LA_ChiSP-7-1	4,609,588	4,231,328	4,194,153	99.12	217,175	4.2
LA_Tunica-8-1	6,487,454	6,026,014	5,823,844	96.65	251,619	4.87
MD_Balti-1-1	1,953,472	1,700,536	1,502,153	88.33	115,882	2.24
MD_Balti-1-2	4,246,474	3,799,432	3,332,294	87.71	166,509	3.22
MD_WyeM-2-1	12,080,584	10,798,926	10,069,897	93.25	278,672	5.39
MD_WyeM-2-2	10,346,046	9,278,034	8,364,405	90.15	295,178	5.71
MD_WyeM-2-3	5,472,568	4,802,596	4,312,809	89.8	203,891	3.94
MD_WyeM-2-4	6,122,908	5,559,446	5,141,258	92.48	214,198	4.14
MD_WyeM-2-5	4,552,560	4,147,036	4,047,171	97.59	189,040	3.66
MD_WyeM-2-6	2,501,970	2,228,302	2,157,675	96.83	154,420	2.99
MD_WyeM-2-7	6,815,536	5,891,894	5,246,627	89.05	231,615	4.48
MD_WyeM-2-8	11,573,136	10,364,472	9,646,787	93.08	274,431	5.31
MD_WyeM-2-9	10,000,934	9,152,358	7,674,882	83.86	293,163	5.67
MD_WyeM-2-10	5,562,226	5,088,240	4,588,148	90.17	216,134	4.18

MI_YankeeSP-1-1	24,604,944	23,561,012	10,022,905	42.54	527,426	10.2
MI_YankeeSP-1-2	1,253,578	1,206,030	885,259	73.4	114,196	2.21
MI_YankeeSP-1-3	10,423,138	9,578,194	9,196,555	96.02	255,649	4.94
MI_YankeeSP-1-4	7,929,874	7,251,476	6,846,631	94.42	220,121	4.26
MI_YankeeSP-1-5	7,212,444	6,676,970	5,924,657	88.73	244,266	4.72
MI_YankeeSP-1-6	9,013,818	8,234,428	6,496,383	78.89	302,989	5.86
MI_YankeeSP-1-7	50,770	45,744	30,199	66.02	12,249	0.24
MI_YankeeSP-1-8	3,735,670	3,365,262	3,244,572	96.41	180,438	3.49
MI_YankeeSP-1-9	9,863,456	9,003,432	7,268,339	80.73	282,180	5.46
NC_DutchC-1-1	24,900,306	22,796,730	10,614,260	46.56	413,849	8
NC_DutchC-1-2	20,286,942	18,795,948	6,816,155	36.26	456,028	8.82
NC_DutchC-1-3	12,615,750	11,960,596	4,245,239	35.49	406,296	7.86
NC_DutchC-1-4	22,001,976	20,510,972	5,376,487	26.21	585,207	11.32
NC_DutchC-1-5	6,918,980	6,326,922	4,595,344	72.63	289,746	5.6
NJ_Bridge-1-1	4,518,110	3,999,736	2,726,967	68.18	228,523	4.42
NJ_Bridge-1-2	4,967,472	4,529,946	3,925,717	86.66	232,075	4.49
NJ_Bridge-1-3	5,325,824	4,876,136	4,515,519	92.6	219,690	4.25
NJ_Bridge-1-4	7,916,962	7,214,024	7,083,580	98.19	265,910	5.14
NJ_Bridge-1-5	10,776,398	9,846,820	8,996,717	91.37	287,382	5.56
OH_NatureT-1-1	70,520	65,150	15,943	24.47	6,840	0.13
OH_MoonT-2-1	18,404,970	16,636,576	3,527,400	21.2	375,339	7.26
OH_Cleves-3-1	6,079,912	5,471,138	3,215,378	58.77	257,444	4.98
OH_Cleves-3-2	7,177,644	6,534,364	5,118,226	78.33	259,494	5.02
OH_Cleves-3-3	8,228,490	7,371,768	5,295,757	71.84	271,895	5.26
OH_Cleves-3-4	32,828	29,368	21,865	74.45	9,739	0.19
OH_Cleves-3-5	13,867,972	12,445,652	7,959,842	63.96	348,639	6.74
OH_Cleves-3-6	131,690	116,750	51,135	43.8	20,340	0.39
OH_Cleves-3-7	18,501,664	16,488,984	9,452,539	57.33	375,389	7.26
OH_Cleves-3-8	9,846,018	9,087,054	5,070,090	55.79	288,605	5.58
OH_Cleves-3-9	33,827,346	31,631,426	19,670,690	62.19	555,876	10.75
OH_Cleves-3-10	8,420,856	7,504,066	3,534,412	47.1	281,129	5.44

OH_Cleves-3-11	10,513,026	9,815,010	7,825,179	79.73	301,628	5.83
OH_Cleves-3-12	5,926,910	5,457,930	3,949,238	72.36	215,805	4.17
OH_Cleves-3-13	33,419,290	30,253,338	16,443,593	54.35	412,475	7.98
OH_Cleves-3-14	26,585,014	24,595,414	16,521,245	67.17	358,510	6.93
OH_Cleves-3-15	28,555,468	26,374,708	14,849,198	56.3	364,684	7.05
OH_Cleves-3-16	33,285,104	30,130,780	16,388,182	54.39	411,816	7.96
OH_Cleves-3-17	14,660,540	14,143,824	12,147,857	85.89	244,023	4.72
OH_Cleves-3-18	28,420,784	26,247,926	14,804,732	56.4	364,327	7.05
OH_Cleves-3-19	6,170,826	6,124,624	5,091,279	83.13	112,103	2.17
OH_Cleves-3-20	3,322,708	3,135,274	2,982,974	95.14	197,280	3.82
OH_PleaT-4-1	21,053,610	19,873,118	6,180,477	31.1	335,868	6.5
OH_PleaT-4-2	5,663,970	5,367,146	3,705,922	69.05	210,109	4.06
OH_PleaT-4-3	32,687,090	30,437,702	24,692,890	81.13	425,995	8.24
OH_PleaT-4-4	13,565,872	12,467,346	10,298,580	82.6	329,711	6.38
OH_PleaT-4-5	7,552,532	6,997,116	6,065,626	86.69	244,619	4.73
OH_PleaT-4-6	7,704,708	7,645,822	1,531,884	20.04	234,837	4.54
OH_PleaT-5-1	38,179,896	34,644,652	6,229,644	17.98	584,442	11.3
OH_PleaT-5-2	36,688,674	34,180,598	19,257,826	56.34	588,703	11.39
OH_PleaT-5-3	13,135,530	12,190,752	9,094,133	74.6	305,757	5.91
OH_PleaT-5-4	21,026,030	19,847,202	6,204,858	31.26	338,001	6.54
OH_PleaT-5-5	31,303,594	29,437,926	5,843,369	19.85	486,577	9.41
OH_PleaT-5-6	11,645,710	10,718,084	9,352,340	87.26	302,558	5.85
OH_PleaT-5-7	5,123,758	4,882,654	2,571,899	52.67	210,024	4.06
OH_PleaT-6-1	14,542,796	13,068,920	3,437,132	26.3	333,546	6.45
OH_PleaT-6-2	13,712,944	12,603,340	10,366,680	82.25	333,387	6.45
OH_PleaT-6-3	12,471,684	11,541,424	10,896,192	94.41	289,993	5.61
OH_PleaT-6-4	12,608,804	11,576,030	9,596,925	82.9	329,573	6.37
OH_PleaT-7-1	10,297,386	10,101,396	3,424,841	33.9	202,529	3.92
OH_PleaT-7-2	38,442,094	34,882,940	6,252,797	17.93	585,917	11.33
OH_PleaT-7-3	54,413,386	47,636,490	6,727,870	14.12	578,829	11.19
OH_PleaT-7-4	13,386,662	12,792,362	4,344,508	33.96	278,892	5.39

OH_PleaT-7-5	7,157,474	6,568,906	2,036,394	31	229,238	4.43
OH_PleaT-7-6	11,506,876	10,551,328	4,193,833	39.75	291,265	5.63
OH_PleaT-8-1	21,644,184	20,140,860	7,337,656	36.43	355,712	6.88
OH_PleaT-8-2	5,694,430	5,625,406	2,311,242	41.09	179,805	3.48
OH_PleaT-8-3	11,199,464	10,454,172	9,371,394	89.64	235,011	4.54
OH_PleaT-8-4	4,257,392	4,116,722	915,594	22.24	169,363	3.28
OH_PleaT-8-5	5,715,650	5,259,820	5,061,627	96.23	211,655	4.09
OH_PleaT-9-1	16,856,432	15,354,640	2,406,336	15.67	405,347	7.84
OH_PleaT-9-2	10,115,194	9,335,750	8,584,334	91.95	264,588	5.12
OH_PleaT-9-3	806,934	789,680	205,896	26.07	63,467	1.23
OH_PleaT-9-4	13,702,354	12,745,706	3,044,934	23.89	270,714	5.24
OH_PleaT-9-5	98,570	95,120	23,954	25.18	8,543	0.17
OH_PleaT-9-6	308,418	292,192	66,573	22.78	23,406	0.45
OK_TaliSD-1-1	2,490,330	2,288,298	1,325,725	57.93	172,383	3.33
OK_TaliSD-1-2	12,282,244	11,674,956	5,091,687	43.61	348,342	6.74
OK_TaliSD-1-3	10,918,408	10,016,928	5,609,155	56	364,431	7.05
PA_DeadM-1-1	3,500,488	3,232,294	2,867,560	88.72	177,920	3.44
PA_DeadM-1-2	3,921,382	3,529,920	3,110,807	88.13	188,966	3.65
PA_DeadM-1-3	168,484	146,530	102,521	69.97	33,555	0.65
PA_FriendL-10-1	5,231,110	4,768,398	3,372,045	70.72	235,704	4.56
PA_FriendL-10-2	15,515,940	14,121,306	9,733,245	68.93	312,900	6.05
PA_VenU-11-1	13,028,342	11,816,190	9,100,719	77.02	336,401	6.51
PA_VenU-11-2	6,413,898	5,924,062	5,396,696	91.1	231,616	4.48
PA_VenU-11-3	4,841,046	4,426,700	3,951,852	89.27	222,426	4.3
PA_BoyceMP-2-1	7,649,578	6,991,348	6,538,317	93.52	256,053	4.95
PA_BoyceMP-2-2	5,702,874	5,183,814	4,733,774	91.32	237,352	4.59
PA_BoyceMP-2-3	4,345,316	3,922,986	3,349,742	85.39	196,826	3.81
PA_Mayview-3-1	6,275,434	5,737,454	5,028,770	87.65	256,381	4.96
PA_Mayview-3-2	3,776,898	3,439,748	3,061,195	88.99	186,837	3.61
PA_Mayview-3-3	24,550,190	22,450,108	19,902,002	88.65	357,430	6.91
PA_Nadine-4-1	13,908,286	12,814,726	12,028,565	93.87	320,574	6.2

PA_ChartP-5-1	3,989,186	3,585,678	3,238,546	90.32	194,307	3.76
PA_ChartP-5-2	6,086,582	5,573,732	4,336,388	77.8	260,850	5.04
PA_ChartP-5-3	5,935,648	5,451,780	5,052,324	92.67	213,804	4.13
PA_ChartP-5-4	4,152,898	3,857,648	3,755,954	97.36	227,117	4.39
PA_ChartP-5-5	1,220,054	1,113,290	1,040,003	93.42	167,501	3.24
PA_ChartP-5-6	10,065,746	8,902,696	2,242,407	25.19	291,148	5.63
PA_RiceL-6-1	2,818,318	2,600,000	2,404,506	92.48	145,489	2.81
PA_LowerF-7-1	9,260,372	8,239,104	4,639,372	56.31	297,973	5.76
PA_LowerF-8-1	41,400	38,032	34,667	91.15	13,717	0.27
PA_CSX-9-1	11,638,368	10,697,500	7,810,665	73.01	245,445	4.75
PA_CSX-9-2	960,210	816,130	692,655	84.87	121,706	2.35
PA_CSX-9-3	4,001,428	3,521,376	2,470,514	70.16	207,346	4.01
SC_Cheraw-1-1	5,600,388	5,098,738	4,213,661	82.64	241,454	4.67
TN_ReelL-1-1	10,355,686	9,097,648	3,191,761	35.08	320,517	6.2
TN_CumbR-2-1	19,354,296	17,207,756	5,195,263	30.19	462,483	8.94
TN_CumbR-3-1	36,494,782	32,818,124	14,319,135	43.63	515,083	9.96
TN_CumbR-4-1	17,966,530	15,923,956	6,234,885	39.15	367,110	7.1
TN_CumbR-4-2	12,655,016	11,533,036	7,934,194	68.8	345,515	6.68
TN_CumbR-4-3	8,054,986	7,484,796	1,447,877	19.34	204,691	3.96
TN_CumbR-4-4	14,534,358	13,248,786	2,363,868	17.84	318,561	6.16
TN_CumbR-4-5	10,450,598	9,688,994	6,605,257	68.17	265,356	5.13
TN_CumbR-4-6	33,244	30,322	14,954	49.32	6,371	0.12
TN_CumbR-4-7	21,390,236	19,646,820	15,556,871	79.18	386,075	7.47
TN_CumbR-4-8	7,556,972	6,762,500	1,935,606	28.62	265,970	5.14
TN_CumbR-4-9	7,450,142	6,805,320	2,660,364	39.09	276,604	5.35
TN_CumbR-4-10	8,686,358	7,697,082	2,547,807	33.1	285,403	5.52
WV_SandS-1-1	45,266	41,264	26,125	63.31	11,416	0.22

Inference of Population Structure

Population structure was analysed using FASTSTRUCTURE implemented in the R package LEA.v.3.6.0, a faster version of STRUCTURE which uses a Bayesian method of inferring population ancestry (Raj et al., 2014). Group number was estimated for K values 1 to 100 using a sparse Non-Negative Matrix Factorization algorithm. The results were summarised in the Figure 4-2a by plotting the resulting admixture coefficients estimated for K value. This approach relies on assumptions of HWE and linkage equilibrium between loci within populations which may not be reliable for a clonally reproducing species like pawpaw, as mentioned in chapter 2. It was no surprise to see that FASTSTRUCTURE had difficulty in identifying distinct ancestral clusters. The FASTSTRUCTURE estimated ancestry coefficients were lowest between $K = 60$ and $K = 70$. This is representative of the number of sample sites (62) or even the sample patches (90). As in the VApop only analysis using the previous reference genome, this method only showed that there was enough local variation to distinguish sample sites but gave no indications on other levels of population structure.

Consistent with Wyatt et al (2020) who had identified a two group structure using nine nuclear microsatellite loci, the principal components analysis revealed two distinct clusters approximating an east to west divide of the Appalachian Mountain (AM) range, Figure 4-3b. The first three principal components (PCs) only accounted for 10.13% of the total variance. The eigenvalues indicated there was a significant drop off in contribution to variation from the first to the second and approximately 20 PCs preceding explained only 1% or less of the variation. This shows that even over the entire native range sampled there is little variation. This was demonstrated by the VApop only analysis, where the furthest distance between any sample was ~550 km. However, in the merged Tpop data set many samples were 1,000 to 2,000 km apart, separated by multiple mountains, rivers, and temperature gradients.

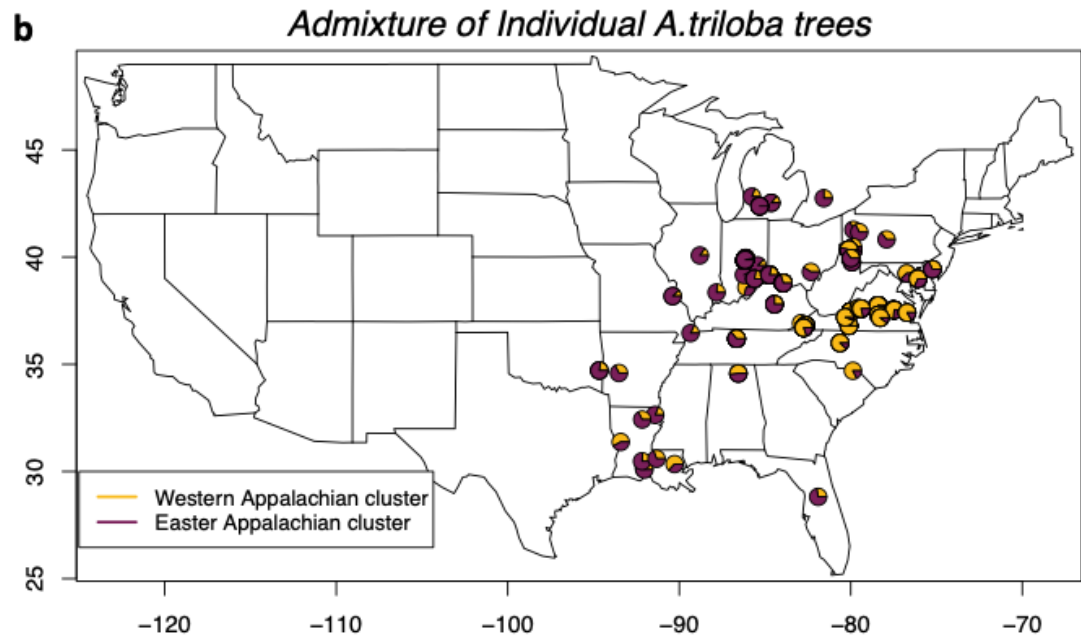
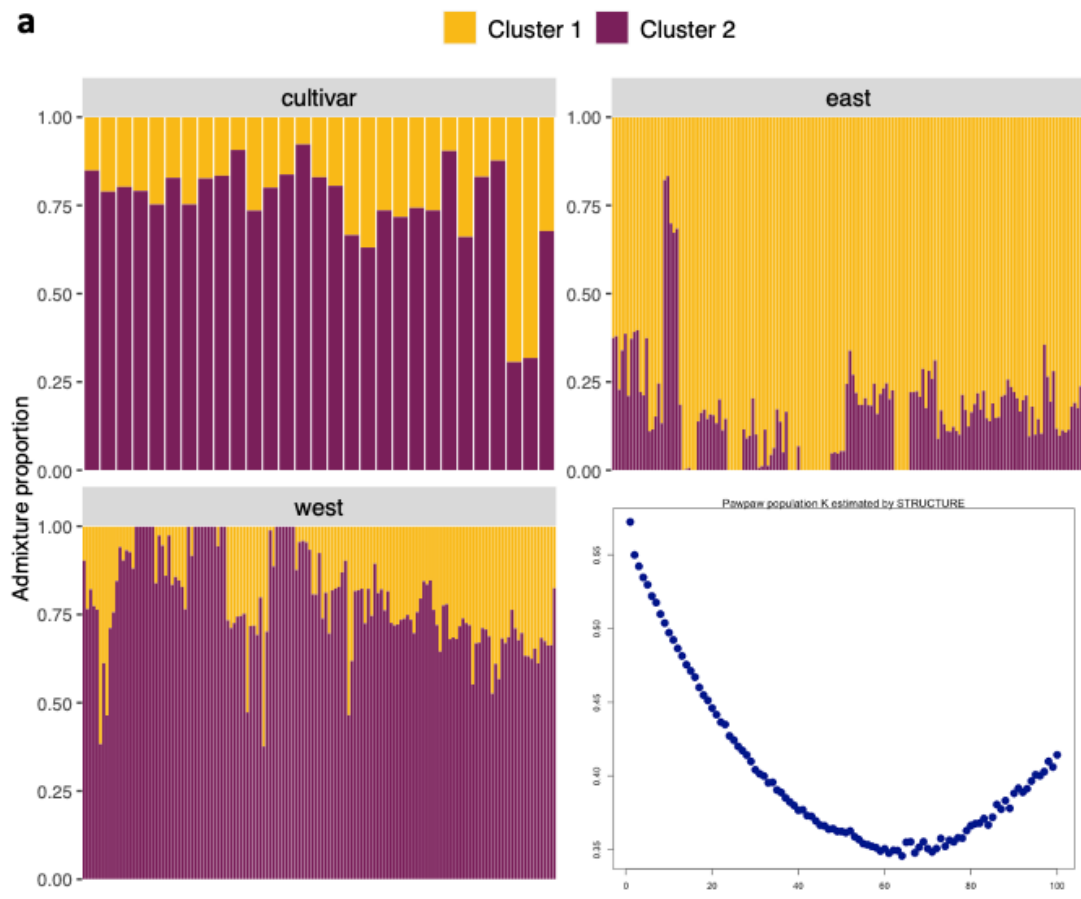


Figure 4-2 Population Admixture

Panel showing the admixture of each sample location. a) The top row and left-hand plots show admixture in each cluster at K=2. Cluster 1 is primarily East, and cluster 2 is primarily west of the Appalachian Mountains. Lower right plot shows FASTSTRUCTURE estimation population structure, x-axis shows number of populations, ancestry coefficient on is the y-axis. b) Map of sample locations with k=2 admixture at each site.

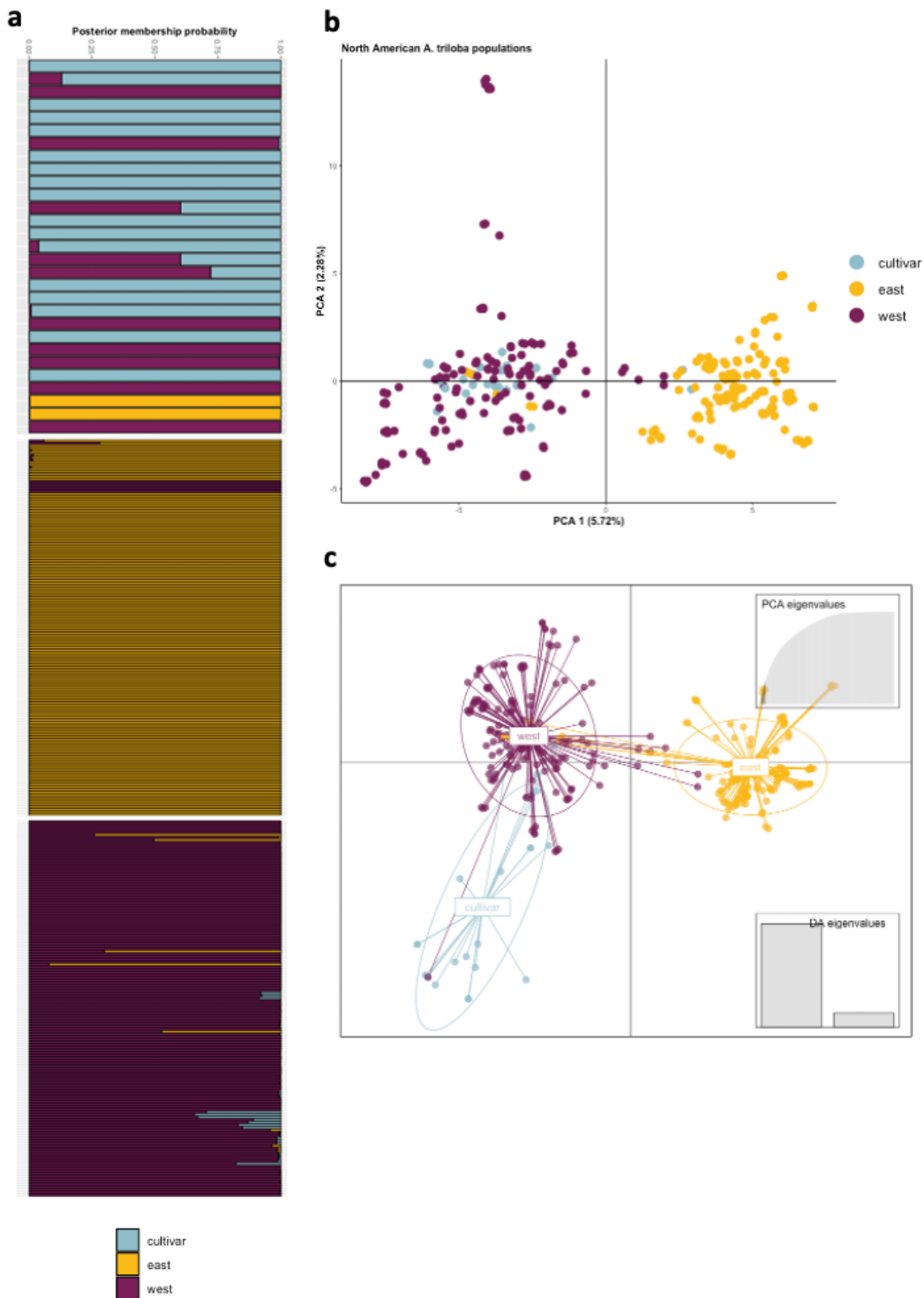


Figure 4-3 Population clustering and membership

Panel showing PCA and DAPC clustering a) DAPC posterior membership probability showing percentage membership of each individual to its assigned cluster. b) PCA clustering with samples coloured by assigned groups. The variance explained by PC1 is 5.72% and by PC2 is 2.28%. c) DAPC clustering of samples in pre-assigned groups.

Regarding quality control, we were concerned that combining the two GBS libraries might result in “batch effects”. Combining data from separate high throughput whole genome sequencing projects has been shown to introduce such issues (Friel et al., 2021; Tom et al., 2017). These effects can come from any number of biological or non-biological sources, including who performed each of the library preparations, laboratory conditions on the day of preparation, reagent lots, preservation of material during shipping etc., and can lead to incorrect conclusions from introduced bias in the variants calling, allele frequencies, or haplotype representation (Chen et al., 2011; Leek et al., 2010). The PCA allowed us to assess potential library biases as we could expect to see clustering by data type or sequencing library. Initially this appeared to explain the two clusters in the PCA, with VApop representing much of the eastern cluster the majority of western cluster comprised the NRpop library. However, we could clearly see that there were a number of samples from either library within each of the two clusters, indicating that the clusters represent true biological variation and not technical variations.

In the Virginia population we identified river basins and large rivers as a factor is the flow of genetic diversity in pawpaw. The sampling of our Tpop relied on volunteers finding pawpaws from any location they had access to and was not targeted to assay geographic factors influencing genetic diversity. Our samples were spread over 10 different major water resource regions (Erie Basin, Great Lakes, Ohio, Mid Atlantic Gulf, South Atlantic Gulf, Tennessee, Lower Mississippi, Upper Mississippi, Texas Gulf and Arkansas-White-Red) (Henderson et al., 2015) but no pattern of effect was seen in group clustering with any combination of PCs.

To identify the optimal number of clusters for the DAPC, a k-means clustering was performed using the Bayesian information criterion (BIC) for $K = 1:100$, for comparison with FASTSTRUCTURE. The optimal clustering solution is inferred from the lowest BIC value (Sun et al., 2014). The BIC values decreased with increasing K , levelling out around $K = 80$

[Figure S4-2]. This is essentially the same as with FASTSTRUCTURE; the optimal number of clusters to explain the genetic variation between individuals matches closely the number of sample patches, pointing to increased local variation with little variation on a population level. Considering this outcome we performed the DAPC comparing the two groups identified in the PC and in Wyatt et al (2020). These were those labelled simply as ‘East’ and ‘West’ representing the east to west split around the AM range, with a third group which included all of the named varieties and curated trees collected from the BTI experimental orchard and the Harvard arboretum, this group was labelled as ‘Cultivars’. The optimal number of PCs to include for the DAPC was estimated by running the ADEGENET function *optim.a.score()* resulting in 11 PCs being retained for the discriminant analysis [Figure S4-2b]. Each groups genetic representation using 11 PCs was cross validated using the ADEGENET function *dapc.summary()* and appeared to support the groups well [Figure S4-2c]. The DAPC scatterplot [Figure 4-2c] showed the three assigned groups as distinct clusters with some individuals sourced from east of the AM in the West cluster and vice versa. Many of the cultivar samples were clustered with West group, which agrees with the recorded information on the origin of cultivated genotypes (Lu et al., 2011; Peterson, 2003, 1991). DAPCs posterior membership probability estimates at $K = 3$ shows the three distinct groups with many cultivars having higher probability of belonging to one of the other two groups than to a third cultivated group [Figure 4-2a]. Pawpaw cultivation is in its early stages with many named cultivars not being developed since they were first selected from wild populations around 100 years ago (Lu et al., 2011; Peterson, 2003, 1991).

Clustering by Neighbor-joining tree

A NJ tree based on the fraction of different sites between samples and rooted with an *Asimina parviflora* sample as the out-group was run using POPPR.v. 2.9.3 (Kamvar et al., 2014), with 100x bootstrapping support. In comparison to the previous cluster analysis methods, a NJ tree uses a distance matrix to construct a tree by determining which terminal nodes are “neighbours” via an iterative clustering process (Saitou and Nei, 1987) and unlike the previous methods, suggests only one possible clustering assignment. A benefit of the NJ tree is that it uses a greedy algorithm, meaning that the applied problem-solving heuristic can make locally optimal choices at each stage (Zhang et al., 2000), and with bootstrapping support it can assign confidence measures to each node of the tree. The tree produced [Figure 4-4] again supports the previous East and West clustering with individuals from the Cultivar group correctly assigned to the reported state they came from. Some exceptions were found in the Wells (Origin: Indiana) and Wilson (Origin: Kentucky) cultivars which clustered most closely with samples from Maryland. This is a possibly a mistake or mislabelling during sampling, some historical confusion about the origin of the genotypes, or, if the sample and history are correct, it is possible evidence of some genotypes being moved vast distances across the country. In the NJ tree [Figure 4-4] we see that in almost all cases individuals from a particular state cluster with others from the same state, and that states close to each other also cluster. This is not surprising given that in our previous study, Chapter 2, we showed that, on a state level, geneflow was slightly increased among individuals in close proximity.

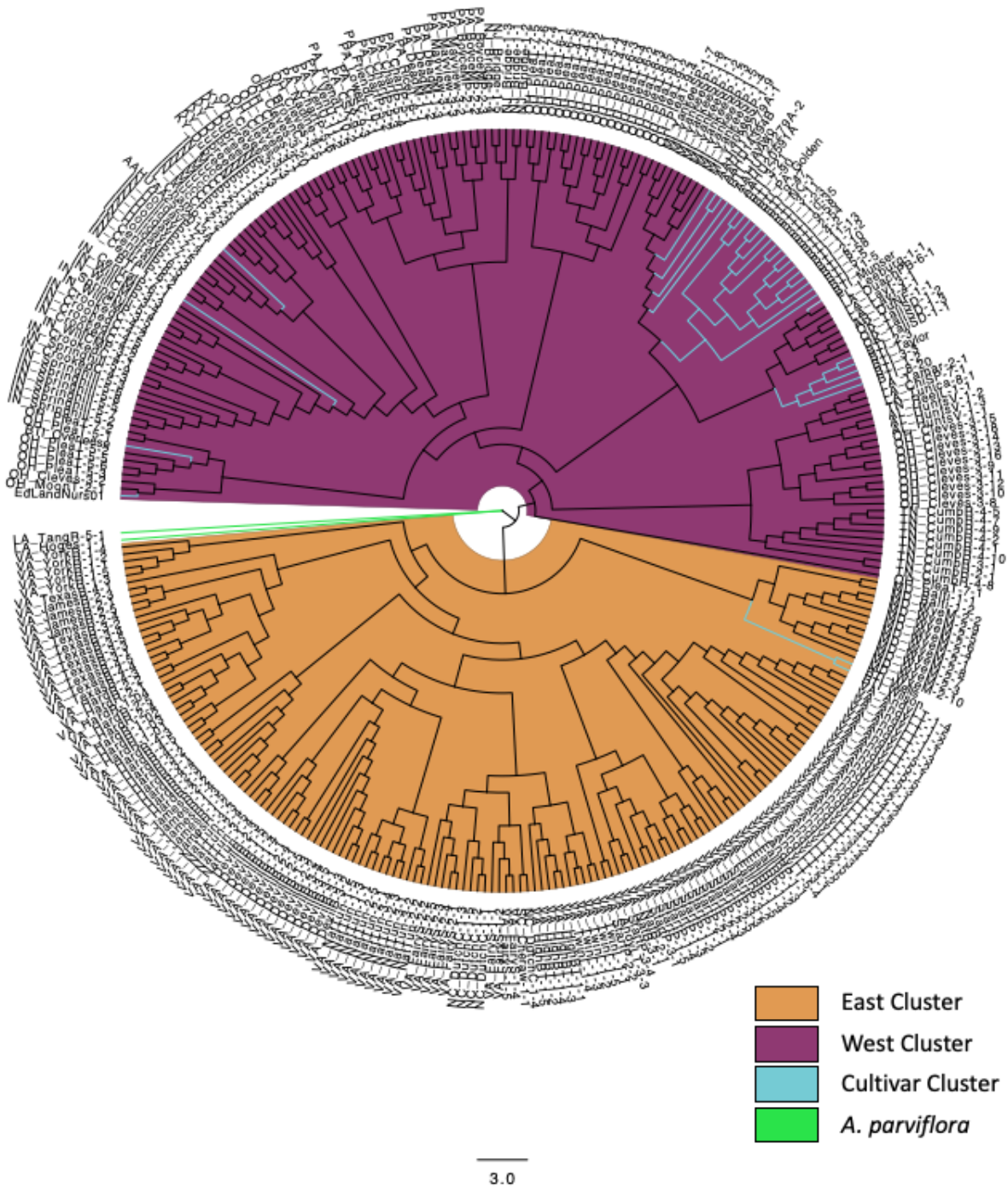


Figure 4-4 Neighbor-joining tree

Cladogram from neighbor-joining tree. Tree representing estimation of the dissimilarity and Euclidian distances distance matrix comparison. The tree was generated using the R package POPPR and run with 100 bootstrapping replicates. Taxon coloured by Appalachian Mountain (AM) clusters. *A. parviflora* used as the outgroup (green) to root the tree.

Isolation-by-distance (IBD)

However, we previously observed a small effect of geographic separation in the Virginia pawpaw population whereby the genetic differences between patches increased with distance. Thus, we ran the same IBD test to see if this effect was maintained over larger distances across the entire native range. We performed a Mantel's correlation test on a matrix of Nei's distances (Nei, 1972) and a matrix of the Euclidean geographic distances between individuals. Results revealed no evidence of IBD in the population [Figure 4-5a], which could indicate that geneflow among pawpaw populations over long distances is maintained at a higher level than genetic drift (Slatkin, 1993) but not a smaller scale, as we saw in the Virginia population.

During the Holocene warming, many of the extant North American species generally moved north from warmer refugia (Delcourt and Delcourt, 1988; Iverson et al., 1999), including pawpaw (Wyatt et al., 2020). Pollen records show that the rate is variable, but that short swift movements can sometimes occur (Davis, 1981; Davis and Shaw, 2001), even as fast as 100m/year (Clark, 1998). But such shifts not only involve migration but also local adaptation (Davis and Shaw, 2001) which may include the need to adapt to new environmental conditions, biotic agents, new species interactions or new photoperiods (Mimura and Aitken, 2007). In migrating populations, adaptations may result from a combination of gene flow and local selection, leading to the stepping stone migration model of IBD within relatively few generations, and resulting in adaptive population divergence across the range (Mimura and Aitken, 2007). If a population is in equilibrium between genetic drift and gene flow, IBD can be expected, making it more likely in long-established populations, as high gene flow tends to homogenize populations (Mimura and Aitken, 2007; Sharbel et al., 2000). In pawpaw, we could then expect to see IBD across the range if geneflow from neighbouring populations is higher than the long-term geneflow over long distances. However, we are not seeing this, the Virginia population showed a steppingstone pattern, but over the entire range we no longer see

any effect. Perhaps geneflow and genetic drift are not in equilibrium across the population, and are acting to homogenize the population, possibly as a result of a rapid northward expansion. In this case it would make it harder to observe genetic variation and effect of IBD over a large range but may still be visible in smaller sample populations.

AMOVA

To evaluate the genetic variation across the AM range, a hierarchical partitioning and analysis of genetic diversity was carried out by running an Analysis of Molecular Variance (AMOVA) (Meirmans, 2006), results are shown in table S4-3. AMOVA assumes a HWE that is not appropriate in a population which likely contains clones. Thus, we performed a Monte-Carlo significance testing with clone correction [Figure 4-5b]. Following previously identified population structure, the AMOVA was performed comparing the two groups identified by PCA. The results indicated that 21% of the variation is between subpopulations, 110% was within samples, and -35% was between samples within subpopulations. The negative number can be explained as an absence of population structure between samples, or as a greater variation within groups than among groups (Meirmans, 2006).

Table 4-3 AMOVA results

Summary of AMOVA conducted on the three groups (East, West and Cultivar)

<i>Variation</i>	Df	Sum Sq	Mean Sq
Between Pop	3	1,042.17	347.40
Between Subpop within pop	112	9,572.57	85.47
Between samples within subpop	213	4,465.08	20.96
Within samples	329	19,399.50	58.96
Total	657	34,479.31	52.48

components of covariance			
<i>Variation</i>		Sigma	%
Between Pop		2.12	3.95
Between Subpop within pop		11.41	21.33
Between samples within subpop		-19.00	-35.52
Within samples		58.97	110.23
Total		53.49	100.00

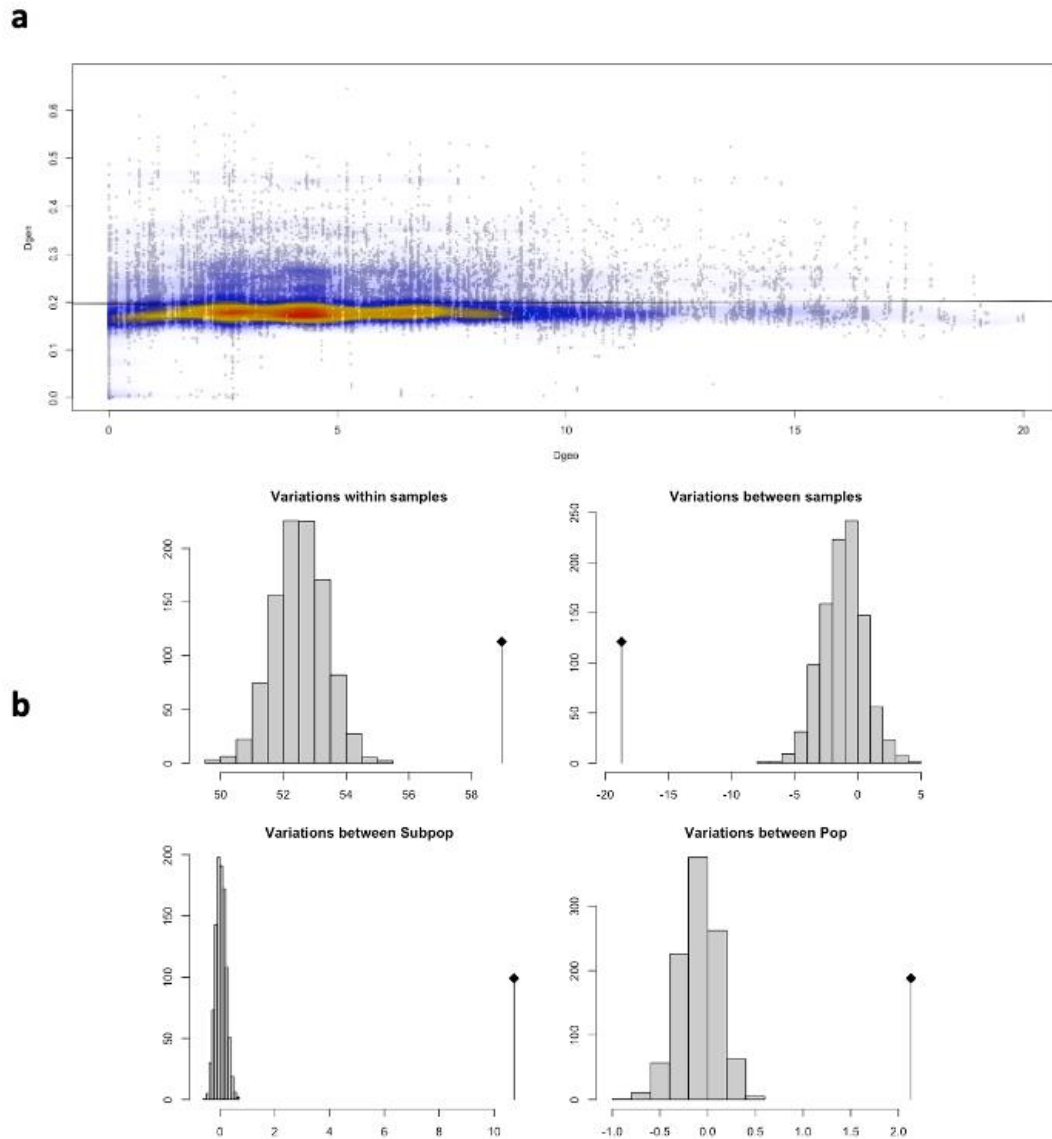


Figure 4-5 IBD and AMOVA

a) Isolation-by-distance plot. x-axis shows geographic distance and y-axis shows the genetic distance between samples. b) Histogram plots showing the results of AMOVA significance testing at each strata. Histogram bars represent the expected variation, the black line and dot represent the measured variation.

Genetic diversity

Table 4-4 Genetic diversity

Summary of genetic variation. H_o : observed heterozygosity within group; H_s : genetic diversity within group; H_T : overall genetic diversity; H_{TP} : corrected H_T ; D_{ST} : diversity among samples; D_{STP} : corrected D_{ST} ; F_{ST} : fixation index; Inbreeding coefficient: F_{IS}

AM Clusters	Fsts	west	cultivar	east
	west	°	°	°
	cultivar	0.03	°	°
	east	0.05	0.06	°
	Mean (H_e)	0.27	0.25	0.27
	Mean (H_o)	0.28	0.25	0.28
	Mean (F_{IS})	0.00	0.04	0.00
	segregating sites	178.00	326.00	333.00
	nucleotide diversity	80.40	97.92	100.99
	Tajima's D	0.20	1.29	2.51
	Watterson's Theta	77.96	72.14	56.51
Global population	H_o	0.27		
	H_s	0.27		
	H_T	0.28		
	D_{ST}	0.01		
	H_{TP}	0.28		
	D_{STP}	0.01		
	F_{ST}	0.03		
	F_{STP}	0.05		
	F_{IS}	-0.01		
	Dest	0.02		
(ANGSD) H_o	0.13			

Heterozygosity and F-statistics

Heterozygosity across the range was estimated using three different approaches for comparison, and three different results were received, see table 4-4. Because DARTR and POPGENOME make use of only a filtered set of loci is shared across all individuals, many heterozygous sites may be removed from the estimation. ANGSD on the other hand uses as input the BAM files containing all unfiltered loci to estimate the total population's observed heterozygosity directly from BAM files. ANGSD estimated the observed heterozygosity to be 0.13, slightly higher than previously estimated (0.07) in the Virginia only population but still lower than other species in the same range, *Cercis Canadensis* (Ony et al., 2021) *Quercus robra* (Götz et al., 2022).

DARTR estimated the total populations observed heterozygosity to be 0.27, which is the same as the expected (0.27). This is lower than the Virginia only H_e reported in Chapter 2 as 0.32. There was no difference in the H_o reported in the two east (0.28) and west (0.28) clusters, and only a slight decrease in the cultivar cluster (0.25). No difference between the two clusters could mean, as discussed above in relation to IBD, that geneflow is still acting to homogenize the level of genetic diversity. Unlike our Virginia only results, heterozygosity is not much higher than the HWE expected, pointing to a slowly evolving species (Edwards, 2008). It seems that across the range the heterozygosity levels are maintained around the expected levels while on a local level heterozygosity between may be higher. It may also indicate little time has passed since the populations split. While a limited selection of the wild diversity selected for cultivation can lead to a founder effect or an inbreeding effect altering heterozygosity (Leberg, 1992; Ony et al., 2021), the slight decrease seen in the cultivated population is not significant according to our pairwise Games-Howell test.

The result of a pairwise F_{st} comparison between the East and West clusters was 0.046 indicating that there exists very little genetic differentiation between the two clusters. (Holsinger and

Weir, 2009; Weir and Cockerham, 1984). Pairwise comparisons of the Cultivar group with West (0.03) and Cultivar with East (0.61) showed that there is little differentiation between any of the groups. Additionally, the cultivars and the wild individuals forming the western cluster are more similar to each other than both of the wild populations (East and West of the AM) are to each other. A likely reason for this is that most the cultivars originate from the West cluster. Comparing inbreeding coefficients positive estimates of Wright's F_{IS} inbreeding coefficient indicates fewer observed heterozygotes compared with the expected by HWE, while a negative F_{IS} indicates an excess (Wright, 1922). F_{IS} estimated by DARTR, the two wild clusters were -0.001 (West) and -0.005 (East). The group of cultivars was the only positive number (0.045). The actual differences are small and do not support inbreeding deficiencies in any of the clusters.

Tajima's D assess the level of genetic variation within a population and can detect the presence of selection pressures (Tajima, 1989). We used POPGENOME to evaluate genetic variation and possible selection process influences on the genetic variation in each cluster, see table 4-4. Estimates for each cluster were 0.2 (West), the lowest of all the three estimations, this value may indicate that the population west of the AM may be evolving at the expected rate for neutral selection. It seems that, as the population north from the Gulf of Mexico, the migrating western population experienced few selection pressures. A Tajima's D of 2.5 was estimated for the eastern population, and finally, the Tajima's D score for the cultivar group was 1.29. Any positive value over zero indicates a relatively high level of genetic variation and could point to a recent population expansion or the presence of weak or balancing selection which might be consequence of the recent migration from refugia in the Gulf of Mexico (Wyatt et al., 2020). Our analysis of the Virginia population revealed that the heterozygosity and genetic variation was high across the state while the genetic variation among populations was low, potentially due to high levels of geneflow, clonal reproduction, and low mutation rates. In comparison

with the entire native range, both genetic variation among groups, and the overall diversity was low. This discrepancy might be an artifact of the different reference genomes used in either study, or even allele dropout rates from the inclusion of poorly sequenced individuals (Heller et al., 2021). In both studies, we were looking at genetic variation on different levels of the population and it is possible that signatures of allele frequencies are dilute over the wider population leading to lower diversity estimates. It is also important to consider that, low levels of genetic diversity have been attributed to species having had extended periods of restricted population sizes, known as genetic bottlenecks, or to founder effects during recolonization during post-glacial (Leberg, 1992) and may be a true representation of the population.

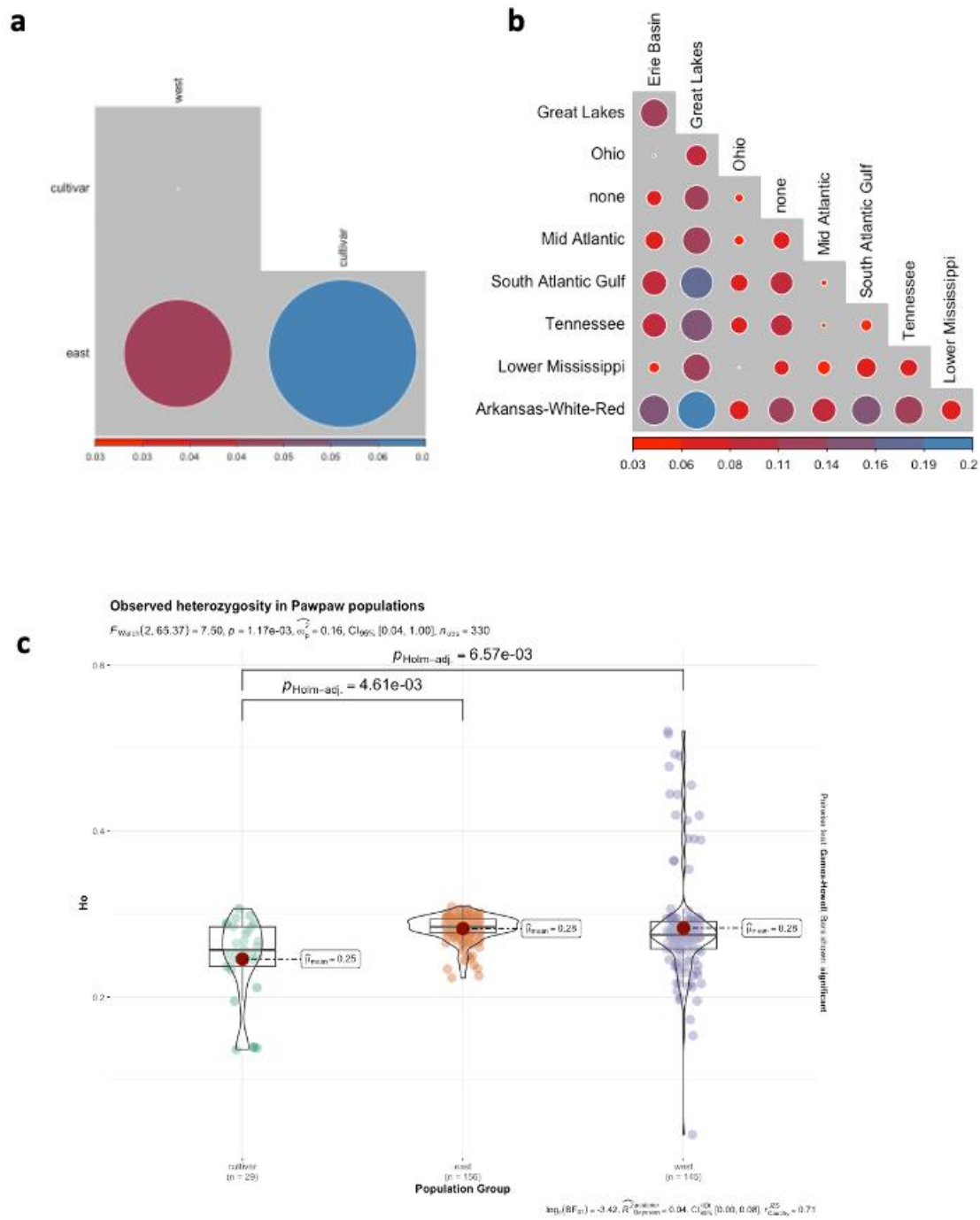


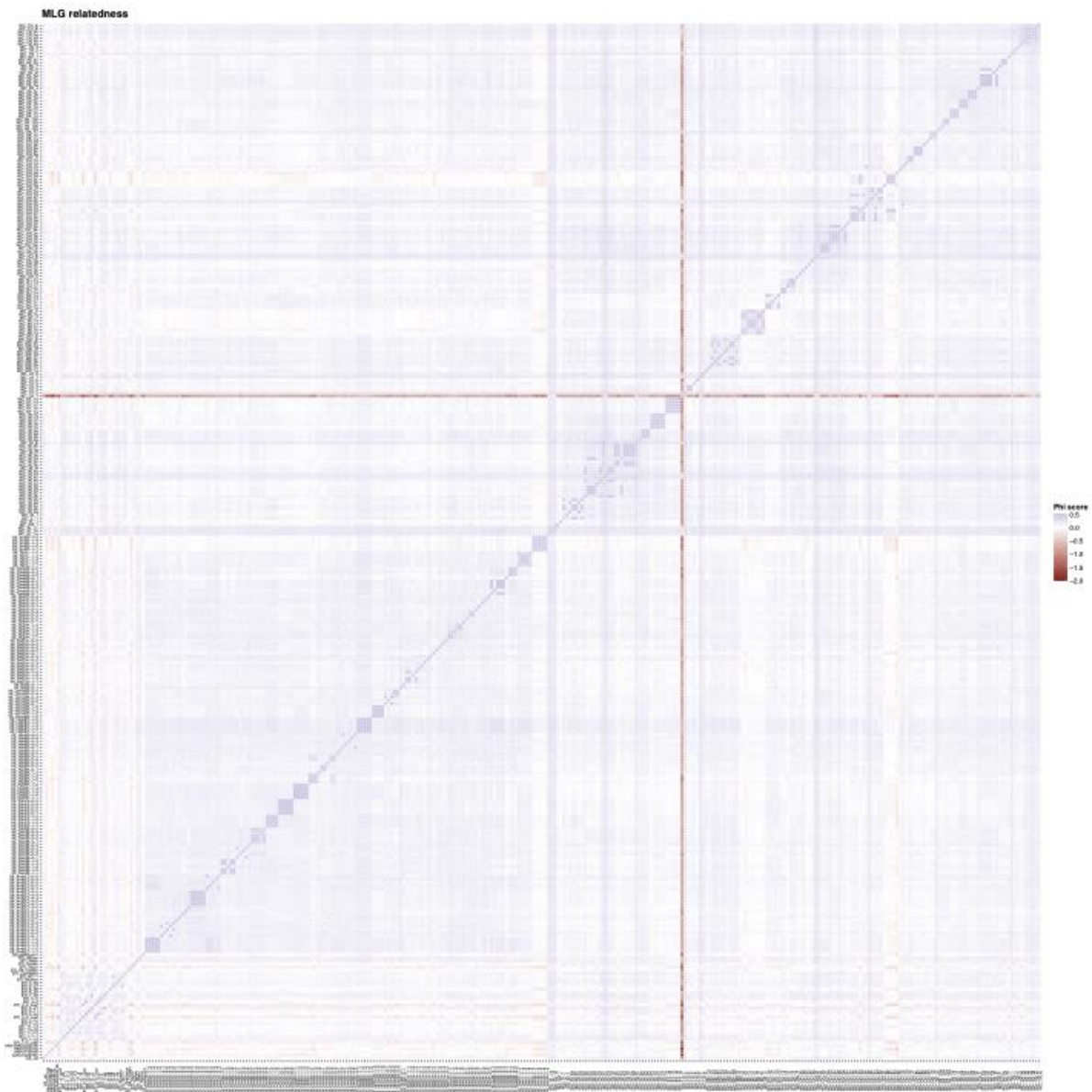
Figure 4-6 Diversity among clusters

a) Corr plot of F_{st} values for pairwise comparisons of East, West and Cultivar groups. b) Corr plot of F_{st} value comparison samples from different major water sources. Colour represents changes from no difference (Red) to low/moderate difference (Blue). Circle size corresponds to F_{st} values. c) Plot compares East, West and Cultivar heterozygosity values of each individual. Pairwise significance testing was done with Games-Howell analysis.

Conclusions

In conclusion, our study supports the findings of Wyatt et al. (2020) who modelled the ecological niches of pawpaw and migration after the last Glacial maxima. Their study made use of microsatellites and identified two large clusters separated by the AM and overall high genetic diversity. Our study using SNPs confirmed these findings, but we also found some interesting discrepancies. Specifically, we found that heterozygosity was low, coming close to the HWE expectations. This could be due to differences in the marker type used between the studies. However, it might also indicate that the species is evolving slowly (Edwards, 2008), perhaps as a result of low mutation rates, high clonal reproduction and infrequent sexual reproduction in a species still emerging from a genetic bottleneck. Whatever the cause, the discrepancy highlights the problems of potential variability introduced from user choices of marker type, reference-free or reference-based approaches, and even the reference version, all of which can make comparison between related work difficult and obscure biological significance. We demonstrated that the genetic diversity over the species distribution range, and variation among the two major populations clusters, is low.

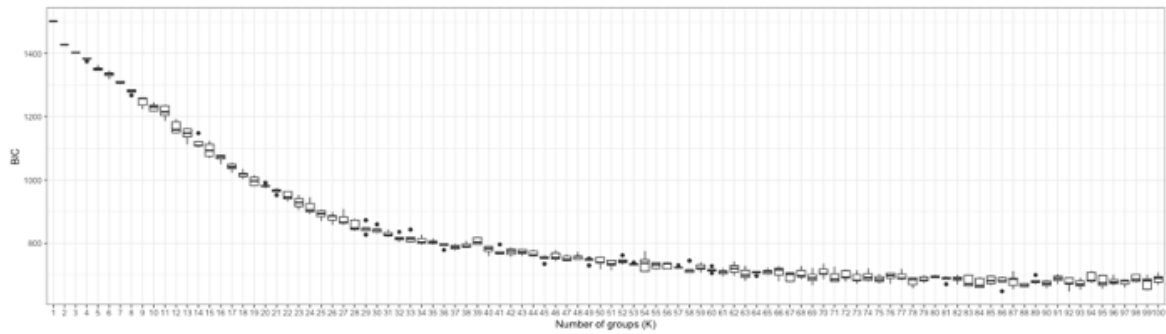
Supplemental Figures



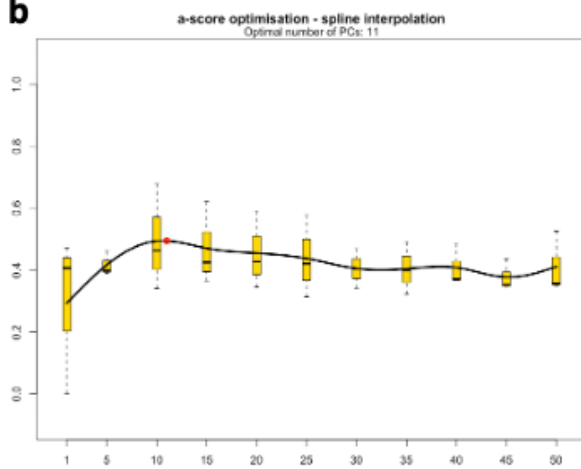
Supplemental Figure 4-1 Relatedness heatmap

Heatmap showing the KING estimated phi values for each pairwise genotype comparison. Values close to 0.5 indicate high genetic similarity. Red line shown here for the *A. parviflora* sample used as an outgroup.

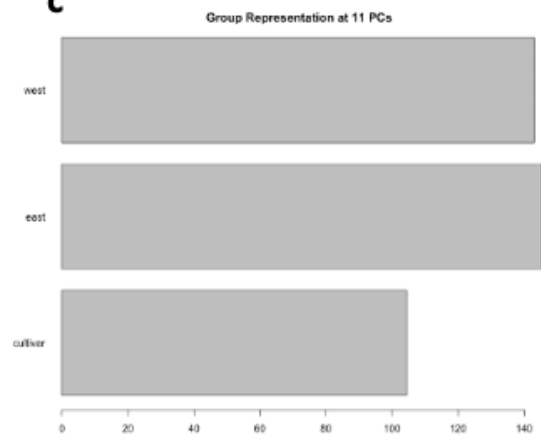
a



b

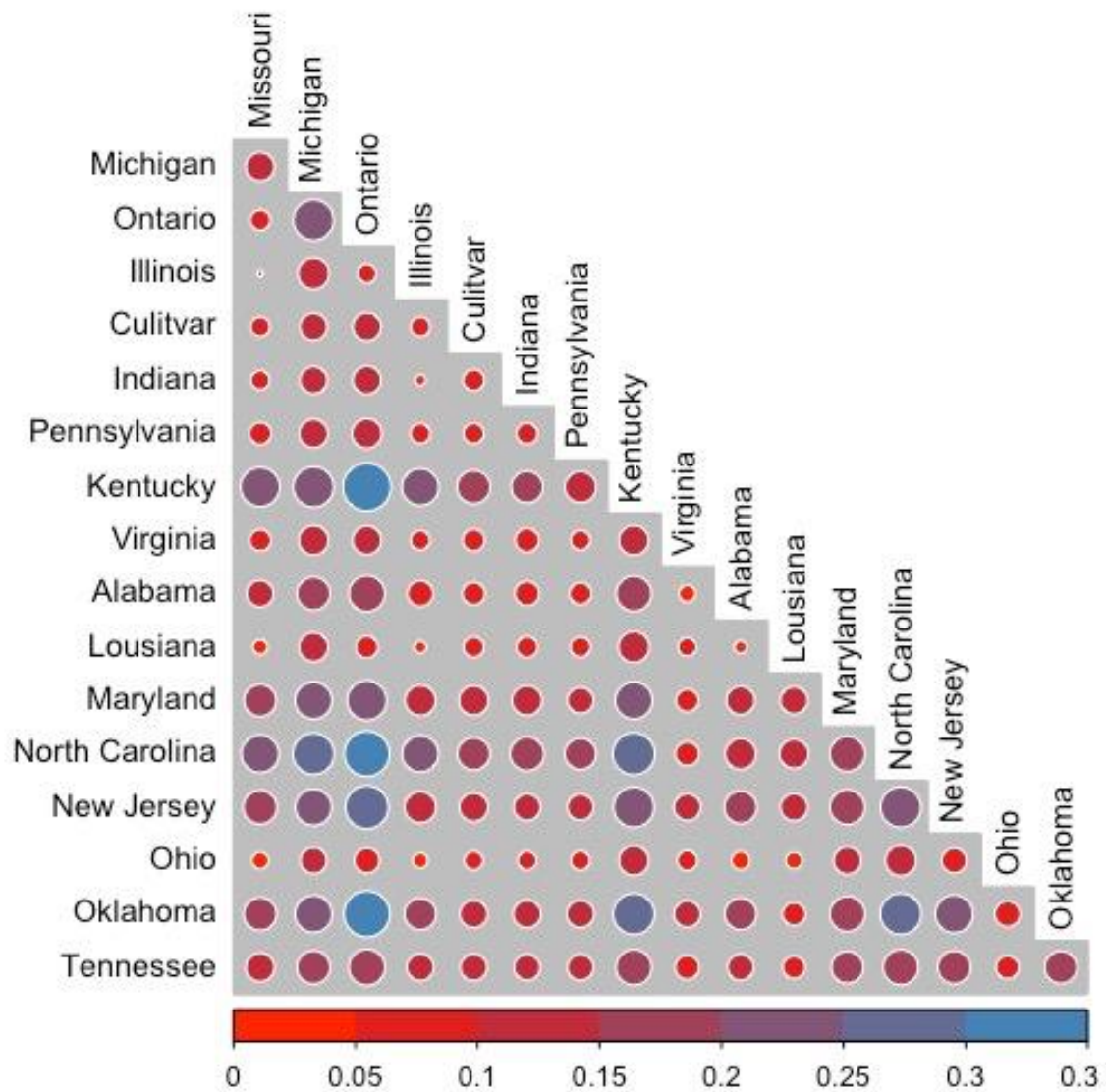


c



Supplemental Figure 4-2 DAPC group optimization

a) Bayesian information criterion (BIC) for values of discriminant analysis of principal components (DAPC) $K = 1:100$. b) DAPC PC optimization testing. c) Barplot showing the group diversity representation estimates when keeping the recommended PCs by DAPC optimization.



Supplemental Figure 4-3 F_{ST} Pairwise comparison by state

a) Correlation plot of F_{ST} values for pairwise comparisons of grouping samples by state. Colours represent changes from no difference (Red) to low/moderate difference (Blue). Circle size corresponds to F_{ST} values.

Chapter 5 : Final Conclusions

Conclusions

The objective of this PhD thesis was to understand the factors influencing the evolutionary history of the species *A. triloba* and the distribution of genetic variations over its native range. To better understand the species, we have produced a high-quality reference genome, used a GBS approach to sequence over 300 wild individuals from across North America, and performed analysis to infer population structure.

Firstly, we assembled the draft genome of *A. triloba* using PacBio's Sequel II's long reads and then polished with Illumina short reads, which resulted in a genome with an N90 of 0.07 Mb and L90 of 2334. This scored well in quality assessment with a BUSCO completeness over 90% and provided a valuable resource for the study of *A. triloba* and, more broadly, to study the evolutionary history of Annonaceae and Magnoliids. However, we aimed to further improve the genome assembly by producing a new version incorporating HiFi, Hi-C and RNA-seq data. The use of these technologies allowed us to generate a highly contiguous and accurate genome assembly with a contig N90 of 92.55 Mb and an L90 of 8. This improved genome assembly provided a more comprehensive view of the genome structure and organization of *A. triloba*, allowed us to construct pseudochromosomes, and to annotate the genome, including the identification of repeat elements, transposable elements and gene models. While the quality of the improved reference genome is excellent, the annotation still has much room for improvement. Genome annotation is a highly complex process that is complicated further when working with poorly studied species. There are many reasons why it might not be successful, including highly complex genomes with poor coverage, poorly resolved repeat regions, a lack

of comparative information from closely related species, or limited computational resources. In our case, there is limited information on closely related species for homology-based annotation but another more likely issue is in the quantity of RNA-seq data we had access to during annotation. By performing more sequencing on other tissue types, we would be able to identify more coding and non-coding regions which would have helped in training the gene prediction software.

Furthermore, we conducted a population genetics study of *A. triloba* using a GBS approach in a sample population from the state of Virginia (VA), sampling from multiple sites along rivers and waterways. In the VA study we mapped our GBS reads to our first draft assembly to identify SNPs. Our analysis using those SNPs allowed us to show the influence of geographic feature on the structure of the population. Specifically, we showed the influence of rivers on gene flow and the presence of low total genetic diversity at moderate level among river basin populations. The results of this study have important implications for the mechanisms underpinning genetic diversity in the species and we showed, for the first-time, rivers as important vehicles for long distance gene dispersal in pawpaw trees.

We followed up this study with samples collected from all across the natural range of the species. We generated GBS libraries and sequenced all the collected samples and mapped the sequencing reads to our improved HiFi and Hi-C assembly. Our analysis provided insight into the migration of the species after the retreat of the Laurentide Ice Sheet. In our study using SNPs, the findings supported those of a previous study by Wyatt et al. (2020) who modelled the ecological niches of pawpaw and migration during the Holocene warming using SSR markers. Both their study and ours identified two large clusters which appear to be split by the Appalachian Mountains. Interestingly, we had some discrepancies; the levels of genetic

diversity heterozygosity was lower than expected which might be due to differences in the marker type and approaches used in either study. Overall, our study indicated that the species may have experienced a recent bottleneck and genetic diversity remains low. Pawpaw is an obligate out-crossing species but can reproduce clonally. Low mutation rates and a bias towards asexual reproduction might explain our findings.

The Pawpaw Network

It is important to note that one of the most essential aspects of this thesis was the contribution made by the Pawpaw Network, a group of volunteers who agreed to go out into the pawpaw trees natural habitat to collect and ship samples of wild individuals. We established this network by engaging with the general public, primarily on Facebook, contacting groups with an interest in foraging native North American species, and groups that had a specific interest in pawpaw. We found multiple groups dedicated to just pawpaw, the largest group was over 9,000 members at the beginning of the Ph.D in 2019 but has since grown to over 11,000. Engagement involved posting the goals of the project, requesting help sampling and answering botanical questions, engagement with user posts, and posing interesting and fun questions about pawpaw to the community. A post was made containing collecting and shipping instructions [Figure 5-1] on multiple groups with an explanation of the project and our goals. Over the course of the project, we received 189 pawpaw samples, 71 patches from 17 states [Figure 5-2], as well as seeds from *Asimina obovata*, and leaf tissue from 8 other members of the genus. In response to the incredible generosity of the Pawpaw Network, we designed and printed pawpaw genetic diversity project t-shirts and stickers, sending one to every contributor [Figure 5-3].


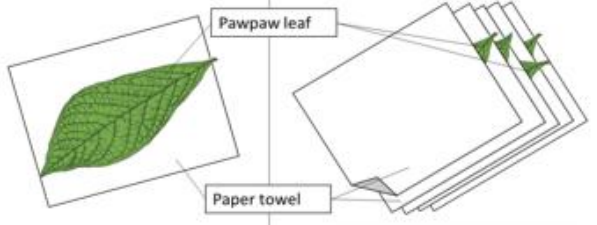
<p>Leaf samples:</p> <p>What Locations? All native Pawpaw states</p> <p>What type of material? <i>A.triloba</i> (Pawpaw) All other <i>Asimina</i> species</p> <p>How much material? 5 leaves from each tree sampled If you are in an area with a patch of pawpaws then up to 5 trees from the patch would be incredibly helpful.</p>	<p>Seed samples:</p> <p>What locations? Only Northern and Southern States. We need seeds from locations that have long cold winters and locations that rarely go below 26°F</p> <p>What type of material? <i>A.triloba</i> (Pawpaw) <i>A.obovata</i> or similar southern species</p> <p>How much material? As many as possible! I am aiming for a total of 60 seeds from each group: 60 northern pawpaw seeds 60 southern pawpaw seeds 60 <i>A.obovata</i> seeds Anything you can contribute to this total figure will help a lot.</p> <p>How to transport? Seeds will need to be clean and dried before shipping. The postal service might be a bit rough on the seeds so they will need some kind of protection. I would recommend a padded envelope but anything that provides protection will be fine.</p>	
 <p>Step 1: Harvest leaves and record location</p> <p>Ideally 5 leaves from 5 different trees in the same patch. (But anything is a big help)</p> <p>It is important record your location the best you can. Something like map coordinates from a GPS but town and State is great too.</p>	 <p>Step 2: Remove excess moisture</p> <p>Gently pat the leaves dry the leaves using a paper towel. Excess water could damage the envelope in transport or lead to the plant material rotting due to fungal infection</p> <p>Step 3: Shipping</p> <p>Prepare for shipping by placing in dry sheet of paper towel in-between each leaf.</p> <p>Then any envelope that you can fit that in will be just fine.</p>	
<p>Please send all mail to the Virginia tech lab run by my collaborator Prof. Haak</p> <p>Postal address:</p> <p>David Haak, 411 Latham hall, 220 Ag Quad, Blacksburg, VA 24061</p> <p>My contact details:</p> <p>James Friel james.friel@unimi.it https://bombarelylab.com/ - This is the new lab website from my supervisor Prof. Bombarely. I will likely post updates on here unless I managed to make a dedicated site.</p> <p>I'm on linkedin and researchgate if you want to find me on either of those.</p>		

Figure 5-1 Sampling instructions

Example of the shipping instruction created and posted to online groups to provide background in the project and details on sampling and shipping.

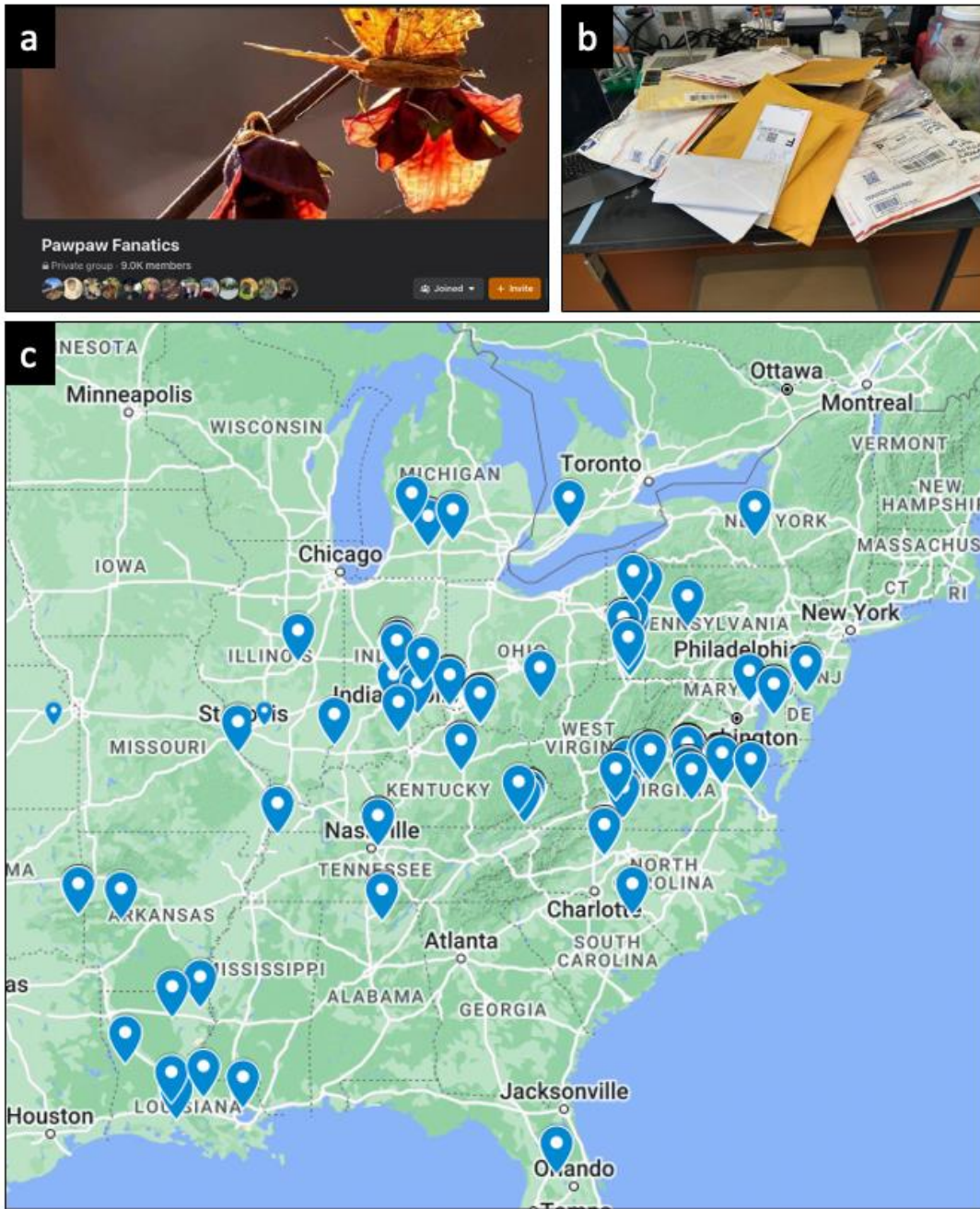


Figure 5-2 Pawpaw Network

Panel showing the effectiveness of the Pawpaw Network. a) Screenshot of one of the group pages used to source material. b) Example of the packages send in by volunteers. c) Google maps image showing sampling locations reached used the Pawpaw Network, figure includes the Virginia population.

Reaching out to people allowed us to collect a large number of samples from diverse geographical locations despite the logistical challenges posed by the pandemic. It is difficult to imagine how we could have collected samples from so many locations even without the imposed travel restrictions. The time and financial costs required to travel such vast distances would be well beyond the means of our research group and made a project like this impossible in a 3-year time frame. Moreover, without any prior information or the local knowledge of the Pawpaw Network members, we would have had to search in forested locations all over the US hoping to find pawpaw trees and would certainly have been unsuccessful on many occasions. People very kindly collected and shipped samples from near their homes when out on walks, along mountain trails when hiking, or even kayaking along rivers in difficult to reach locations, meaning we received samples from areas we could never have reached ourselves.

Thanks to the Pawpaw Network, we were able to analyse the genetic diversity of pawpaw trees across the length of its native range, from the Gulf of Mexico to Ontario, and Canada approximately 2,000 km away. The successful use of citizen science in this study highlights the potential of this approach to overcome some logistical challenges that can occur when studying other widespread wild populations.

In summary, this thesis has made significant contributions to the understanding of the genome structure and organization of *A. triloba* and its population genetics. The improved genome assembly and population genetics data generated in this study will serve as a valuable resource for further functional and comparative genomics studies, as well as for conservation and breeding programs. The use of citizen science in this study has shown its potential to overcome significant logistical challenges by engaging the general public in scientific research.

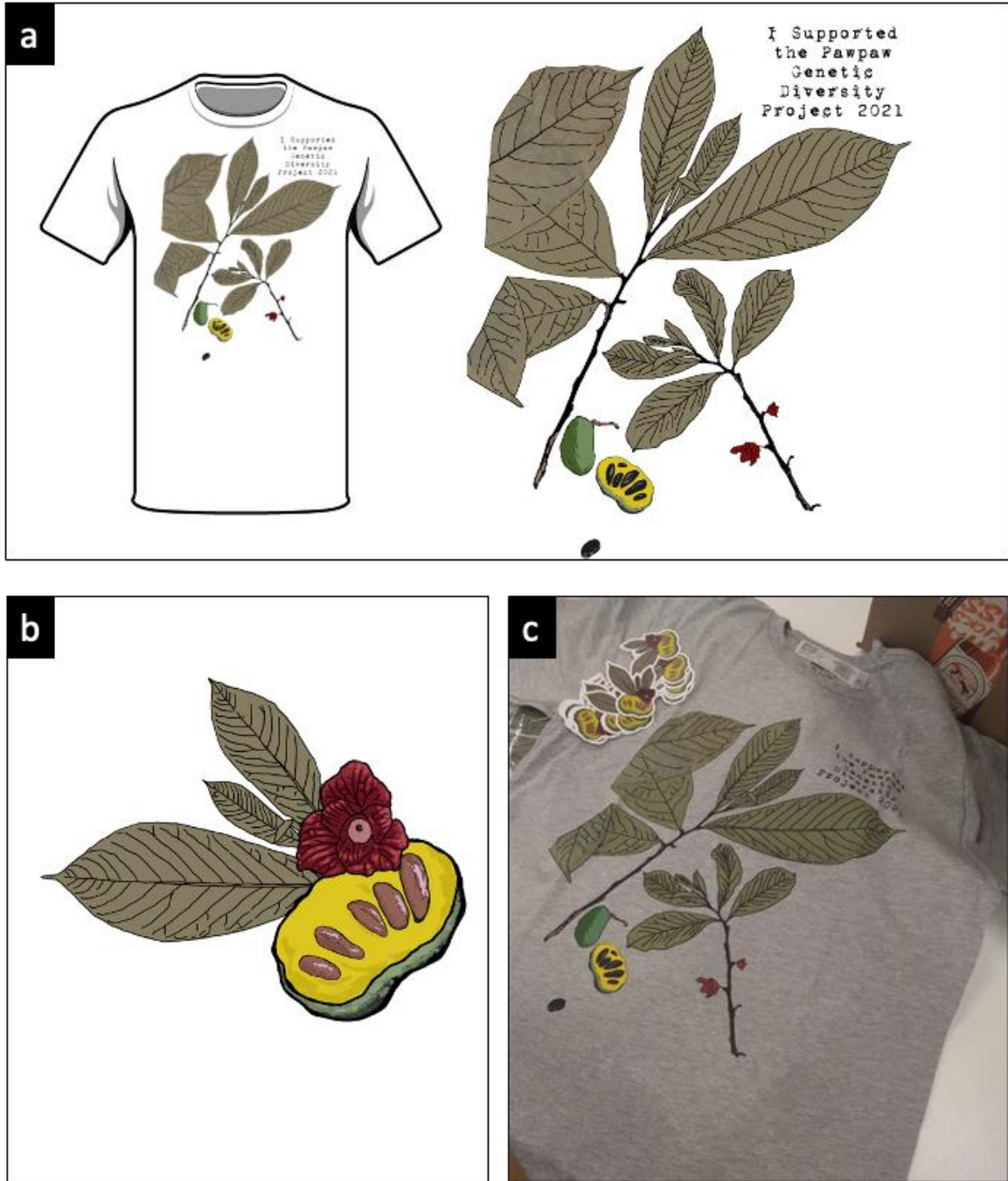


Figure 5-3 Pawpaw T-shirt design

Panel showing t-shirt we created to thank the Pawpaw Network. a) The final T-shirt design. b) Sticker design we also created. c) Picture of the printed t-shirt and stickers that were sent to all volunteers.

Final thoughts and Future perspectives

Overall, this thesis has provided an important step forward in understanding the genomic basis of adaptation and diversity in *A. triloba* and proved an important resources for conservation and breeding programs. The research conducted in this thesis has provided valuable insights into *A. triloba*'s genetic diversity through application of population genetic analysis, and in producing a reference genome we revealed information on the structure, repeat content, orthologous genes, and clarification on chromosome number. However, there are several avenues for future research that can build upon this work and aspects of the project we were unable to achieve because of the pandemic.

Areas of application

Functional Genomics: The improved genome assembly generated in this thesis provides an excellent resource for functional genomics studies. Future research can focus on identifying functional elements in the genome, such as gene expression patterns, regulatory elements, and cis- and trans-acting factors that are involved in the adaptation and diversification of *A. triloba*. In particular, genes involved in the acetogenin pathways are of particular interest to pawpaw breeders because they have potential anti-cancer and pesticidal applications but can also accumulate in bruised overripe fruit changing flavour profiles and even act as an emetic.

Comparative Genomics: The improved genome assembly of *A. triloba* can be used to conduct comparative genomics studies with other closely related species in the genus *Asimina*. This will help to understand the genetic basis of adaptation and diversification in the genus and provide insight into the evolution of the genus and in the cold adaptation found only in *A. triloba* and, to a much lesser extent, *A. parviflora*.

Population Genetics: The population genetics study conducted in this thesis provides valuable insights into the genetic structure and diversity of *A. triloba*. However, further sampling from wild populations in different regions and an increased number of markers will give a more comprehensive understanding of the population genetics of the species and its evolutionary history. The role of Indigenous America people in the distribution of the species remains unresolved. One approach might be to sample pawpaw trees around known anthropogenic sites, this would allow for comparison of SNPs from pawpaw trees at human and wild sites. Comparing the frequency of specific SNPs at these sites, which could indicate specific movement of individuals, indicating the presence or absence of human intervention.

Breeding Programs: The improved genome assembly and population genetics data generated in this thesis can be used to inform breeding programs for *A. triloba*. Identifying regions of the genome that are under selection and regions of high genetic diversity can help to identify desirable traits and create more efficient breeding strategies.

Citizen Science: The use of citizen science in this study has shown its potential to overcome logistical challenges through involvement of the public in scientific research. Future research on pawpaw could continue to explore the use of citizen science in collecting samples and build on the already established Pawpaw Network.

Appendix

Additional publication from work done during the course of the PhD

Friel, J., Bombarely, A., Dorca Fornell, C., Luque, F., Fernández-Ocaña, AM. (2021) Comparative Analysis of Genotyping by Sequencing and Whole-Genome Sequencing Methods in Diversity Studies of *Olea europaea* L. *Plants* 2021, 10(11), 2514; <https://doi.org/10.3390/plants10112514>

Impact factor: 4.67

Comparative Analysis of Genotyping By Sequencing and Whole-Genome Sequencing Methods in Diversity Studies of *Olea europaea* L.

James Friel¹, **Aureliano Bombarely**^{1,2}, **Carmen Dorca Fornell**³, **Francisco Luque**⁴ and **Ana Maria Fernández-Ocaña**^{5,*}

1. ¹Dipartimento di Bioscienze, Università degli Studi di Milano, 20122 Milan, Italy; james.friel@unimi.it (J.F.); abombarely@ibmcp.upv.es (A.B.)
2. ²Instituto de Biología Molecular y Celular de Plantas (IBMCP), CSIC, Universitat Politecnica de Valencia, 46011 Valencia, Spain
3. ³Universidad Internacional de la Rioja (UNIR)Facultad de Educación. Departamento de Didáctica de las Matemáticas y las Ciencias Experimentales. 26006. Logroño Spain; mariadelcarmen.dorcafornell@unir.net
4. ⁴Instituto Universitario de Investigación en Olivares y Aceites de Oliva (INUO), Universidad de Jaén, 23071 Jaén, Spain; ffluque@ujaen.es
5. ⁵Departamento de Biología Animal, Biología Vegetal y Ecología, Facultad de Ciencias Experimentales, Campus de Las Lagunillas s/n, Universidad de Jaén UJA, 23071 Jaén, Spain

* Correspondence: amocana@ujaen.es

Author Contributions:

Conceptualisation, A.M.F.-O., A.B. and F.L.;

Methodology, A.M.F.-O., A.B., and **James Friel.**;

Formal analysis, **James Friel.**;

writing—original draft preparation, **James Friel.**;

writing—review and editing, **James Friel**, C.D.F., A.B., F.L. and A.M.F.-O.

Bibliography

- Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. <https://doi.org/10.1101/gr.094052.109>
- Alkofahi, A., Rupprecht, J.K., Anderson, J.E., McLaughlin, J.L., Mikolajczak, K.L., Scott, B.A., 1989. Search for New Pesticides from Higher Plants, in: *Insecticides of Plant Origin*, ACS Symposium Series. American Chemical Society, pp. 25–43. <https://doi.org/10.1021/bk-1989-0387.ch003>
- Altenhoff, A.M., Gil, M., Gonnet, G.H., Dessimoz, C., 2013. Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. *PLoS ONE* 8, e53786. <https://doi.org/10.1371/journal.pone.0053786>
- Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92. <https://doi.org/10.1038/nrg.2015.28>
- Andrews, S., 2010. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.
- Angel, V.D.D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O.V., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B.L., Soler, L., Binzer-Panchal, M., Lantz, H., 2018. Ten steps to get started in Genome Assembly and Annotation. <https://doi.org/10.12688/f1000research.13598.1>
- Angiosperm Phylogeny Group, 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* 181, 1–20. <https://doi.org/10.1111/boj.12385>

- Arnaud-Haond, S., Alberto, F., Teixeira, S., Procaccini, G., Serrão, E.A., Duarte, C.M., 2005. Assessing genetic diversity in clonal organisms: low diversity or low resolution? Combining power and cost efficiency in selecting markers. *J. Hered.* 96, 434–440. <https://doi.org/10.1093/jhered/esi043>
- Arnold, B., Corbett-Detig, R.B., Hartl, D., Bomblies, K., 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* 22, 3179–3190. <https://doi.org/10.1111/mec.12276>
- Aronesty, E., 2013. Comparison of Sequencing Utility Programs. *Open Bioinforma. J.* 7.
- Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A., 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE* 3, e3376. <https://doi.org/10.1371/journal.pone.0003376>
- Balloux, F., Lehmann, L., de Meeûs, T., 2003. The population genetics of clonal and partially clonal diploids. *Genetics* 164, 1635–1644.
- Bankevich, A., Bzikadze, A.V., Kolmogorov, M., Antipov, D., Pevzner, P.A., 2022. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat. Biotechnol.* 40, 1075–1081. <https://doi.org/10.1038/s41587-022-01220-6>
- Barlow, C., 2001. *The Ghosts Of Evolution: Nonsensical Fruit, Missing Partners, and Other Ecological Anachronisms.*
- Belletti, P., Monteleone, I., Ferrazzini, D., 2008. A population genetic study in a scattered forest species, wild service tree [*Sorbus torminalis* (L.) Crantz], using RAPD markers. *Eur. J. For. Res.* 127, 103–114. <https://doi.org/10.1007/s10342-007-0187-1>
- Bellini, E., Nin, S., Cocchi, M., 2003. The Pawpaw Research Program at the Horticulture Department of the University of Florence. *HortTechnology* 13, 455–457. <https://doi.org/10.21273/HORTTECH.13.3.0455>

Belton, J.-M., McCord, R.P., Gibcus, J., Naumova, N., Zhan, Y., Dekker, J., 2012. Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods San Diego Calif* 58, 10.1016/j.ymeth.2012.05.001. <https://doi.org/10.1016/j.ymeth.2012.05.001>

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoshler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.-D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo,

- S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, Andrew C., Pike, Alger C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, John, Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, Jane, Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. <https://doi.org/10.1038/nature07517>
- Berković, B., Coelho, N., Gouveia, L., Serrão, E.A., Alberto, F., 2018. Individual-based genetic analyses support asexual hydrochory dispersal in *Zostera noltei*. *PLOS ONE* 13, e0199275. <https://doi.org/10.1371/journal.pone.0199275>
- Berry, E.W., 1916. The Lower Eocene Floras of Southeastern North America. *Berry E W 1916 Low. Eocene FLoras Southeast. N. Am.* 91, 90.
- Blanchet, S., Prunier, J.G., Paz-Vinas, I., Saint-Pé, K., Rey, O., Raffard, A., Mathieu-Bégné, E., Loot, G., Fournet, L., Dubut, V., 2020. A river runs through it: The causes, consequences, and management of intraspecific diversity in river networks. *Evol. Appl.* 13, 1195–1213. <https://doi.org/10.1111/eva.12941>
- Bohling, J., 2020. Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecol. Evol.* 10, 7585–7601. <https://doi.org/10.1002/ece3.6483>

- Bombarely, A., 2022. aubombarely/GenoToolBox.
- Botkins, J., Pomper, K.W., Lowe, J.D., Crabtree, S.B., 2012. Pawpaw Patch Genetic Diversity, and Clonality, and its Impact on the Establishment of Invasive Species in the Forest Understory. *J. Ky. Acad. Sci.* 73, 113–121. <https://doi.org/10.3101/1098-7096-73.2.113>
- Bowden, W.M., 1949. Triploid Mutants among Diploid Seedling Populations of *Asimina triloba*. *Bull. Torrey Bot. Club* 76, 1–6. <https://doi.org/10.2307/2481881>
- Bowden, W.M., 1940. Diploidy, Polyploidy, and Winter Hardiness Relationships in the Flowering Plants. *Am. J. Bot.* 27, 357–371. <https://doi.org/10.2307/2436450>
- Brannan, R.G., Peters, T., Talcott, S.T., 2015. Phytochemical analysis of ten varieties of pawpaw (*Asimina triloba* [L.] Dunal) fruit pulp. *Food Chem.* 168, 656–661. <https://doi.org/10.1016/j.foodchem.2014.07.018>
- Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., Borodovsky, M., 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* 3, lqaa108. <https://doi.org/10.1093/nargab/lqaa108>
- Brunke, J., Radespiel, U., Russo, I.-R., Bruford, M.W., Goossens, B., 2019. Messing about on the river: the role of geographic barriers in shaping the genetic structure of Bornean small mammals in a fragmented landscape. *Conserv. Genet.* 20, 691–704. <https://doi.org/10.1007/s10592-019-01159-3>
- Callway, M.B., 1992. Current research for the commercial development of pawpaw *Asimina triloba* (L.) dunal. *HortScience* 27, 90.
- Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., Yandell, M., 2008. MAKER: An easy-to-use annotation pipeline designed for emerging

- model organism genomes. *Genome Res.* 18, 188–196.
<https://doi.org/10.1101/gr.6743907>
- Carneiro, M.O., Russ, C., Ross, M.G., Gabriel, S.B., Nusbaum, C., DePristo, M.A., 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* 13, 375. <https://doi.org/10.1186/1471-2164-13-375>
- Carson, A.R., Smith, E.N., Matsui, H., Brækkan, S.K., Jepsen, K., Hansen, J.-B., Frazer, K.A., 2014. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* 15, 125. <https://doi.org/10.1186/1471-2105-15-125>
- Chaisson, M.J., Pevzner, P.A., 2008. Short read fragment assembly of bacterial genomes. *Genome Res.* 18, 324–330. <https://doi.org/10.1101/gr.7088808>
- Chaput-Bardy, A., Fleurant, C., Lemaire, C., Secondi, J., 2009. Modelling the effect of in-stream and overland dispersal on gene flow in river networks. *Ecol. Model., Selected Papers on Spatially Explicit Landscape Modelling: Current practices and challenges* 220, 3589–3598. <https://doi.org/10.1016/j.ecolmodel.2009.06.027>
- Chaw, S.-M., Liu, Y.-C., Wu, Y.-W., Wang, H.-Y., Lin, C.-Y.I., Wu, C.-S., Ke, H.-M., Chang, L.-Y., Hsu, C.-Y., Yang, H.-T., Sudianto, E., Hsu, M.-H., Wu, K.-P., Wang, L.-N., Leebens-Mack, J.H., Tsai, I.J., 2019. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* 5, 63–73. <https://doi.org/10.1038/s41477-018-0337-0>
- Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., Liu, C., 2011. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLOS ONE* 6, e17238. <https://doi.org/10.1371/journal.pone.0017238>

- Chen, J., Hao, Z., Guang, X., Zhao, C., Wang, P., Xue, L., Zhu, Qihui, Yang, Linfeng, Sheng, Y., Zhou, Y., Xu, H., Xie, H., Long, X., Zhang, J., Wang, Z., Shi, M., Lu, Y., Liu, S., Guan, L., Zhu, Qianhua, Yang, Liming, Ge, S., Cheng, T., Laux, T., Gao, Q., Peng, Y., Liu, N., Yang, S., Shi, J., 2019. *Liriodendron* genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nat. Plants* 5, 18–25. <https://doi.org/10.1038/s41477-018-0323-6>
- Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., Hwang, C.-C., 2013. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLOS ONE* 8, e62856. <https://doi.org/10.1371/journal.pone.0062856>
- Chen, Z., Erickson, D.L., Meng, J., 2020. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics* 21, 631. <https://doi.org/10.1186/s12864-020-07041-8>
- Christiansen, H., Heindler, F.M., Hellemans, B., Jossart, Q., Pasotti, F., Robert, H., Verheyde, M., Danis, B., Kochzius, M., Leliaert, F., Moreau, C., Patel, T., Van de Putte, A.P., Vanreusel, A., Volckaert, F.A.M., Schön, I., 2021. Facilitating population genomics of non-model organisms through optimized experimental design for reduced representation sequencing. *BMC Genomics* 22, 625. <https://doi.org/10.1186/s12864-021-07917-3>
- Chu, J., Mohamadi, H., Warren, R.L., Yang, C., Birol, I., 2017. Innovations and challenges in detecting long read overlaps: An evaluation of the state-of-the-art. *Bioinformatics* 33, 1261–1270. <https://doi.org/10.1093/bioinformatics/btw811>
- Clark, J.S., 1998. Why Trees Migrate So Fast: Confronting Theory with Dispersal Biology and the Paleorecord. *Am. Nat.* 152, 204–224. <https://doi.org/10.1086/286162>
- Coleman, A.P., 1901. Glacial and Interglacial Beds near Toronto. *J. Geol.* 9, 285–310.

- Cushman, S.A., Max, T., Meneses, N., Evans, L.M., Ferrier, S., Honchak, B., Whitham, T.G., Allan, G.J., 2014. Landscape genetic connectivity in a riparian foundation tree is jointly driven by climatic gradients and river networks. *Ecol. Appl.* 24, 1000–1014. <https://doi.org/10.1890/13-1612.1>
- Cypher, B.L., Cypher, E.A., 1999. Germination Rates of Tree Seeds Ingested by Coyotes and Raccoons. *Am. Midl. Nat.* 142, 71–76. [https://doi.org/10.1674/0003-0031\(1999\)142\[0071:GROTSI\]2.0.CO;2](https://doi.org/10.1674/0003-0031(1999)142[0071:GROTSI]2.0.CO;2)
- Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 2011. The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Darrow, G.M., 1975. Minor temperate fruits. J Janick JN Moore Eds *Adv. Fruit Breed.* Purdue Univ Press West Lafayette IN 276–277.
- Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. <https://doi.org/10.1038/nrg3012>
- Davis, M.B., 1981. Quaternary History and the Stability of Forest Communities, in: West, D.C., Shugart, H.H., Botkin, D.B. (Eds.), *Forest Succession: Concepts and Application*, Springer Advanced Texts in Life Sciences. Springer, New York, NY, pp. 132–153. https://doi.org/10.1007/978-1-4612-5950-3_10
- Davis, M.B., Shaw, R.G., 2001. Range shifts and adaptive responses to Quaternary climate change. *Science* 292, 673–679. <https://doi.org/10.1126/science.292.5517.673>
- de la Vega, E., Chalk, T.B., Wilson, P.A., Bysani, R.P., Foster, G.L., 2020. Atmospheric CO₂ during the Mid-Piacenzian Warm Period and the M2 glaciation. *Sci. Rep.* 10, 11002. <https://doi.org/10.1038/s41598-020-67154-8>

- Delahaye, C., Nicolas, J., 2021. Sequencing DNA with nanopores: Troubles and biases. *PLOS ONE* 16, e0257521. <https://doi.org/10.1371/journal.pone.0257521>
- Delcourt, H.R., Delcourt, P.A., 1988. Quaternary landscape ecology: Relevant scales in space and time. *Landsc. Ecol.* 2, 23–44. <https://doi.org/10.1007/BF00138906>
- Díaz-Arce, N., Rodríguez-Ezpeleta, N., 2019. Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better? *Front. Genet.* 10, 533. <https://doi.org/10.3389/fgene.2019.00533>
- Dobin, A., 2023. STAR 2.7.10b.
- Dong, S., Liu, M., 2020. The genome assembly and annotation of *Magnolia biondii* Pamp., a phylogenetically, economically, and medicinally important ornamental tree species. <https://doi.org/10.5061/DRYAD.S4MW6M947>
- Dowsett, H.J., Foley, K.M., Stoll, D.K., Chandler, M.A., Sohl, L.E., Bentsen, M., Otto-Bliesner, B.L., Bragg, F.J., Chan, W.-L., Contoux, C., Dolan, A.M., Haywood, A.M., Jonas, J.A., Jost, A., Kamae, Y., Lohmann, G., Lunt, D.J., Nisancioglu, K.H., Abe-Ouchi, A., Ramstein, G., Riesselman, C.R., Robinson, M.M., Rosenbloom, N.A., Salzmann, U., Stepanek, C., Strother, S.L., Ueda, H., Yan, Q., Zhang, Z., 2013. Sea Surface Temperature of the mid-Piacenzian Ocean: A Data-Model Comparison. *Sci. Rep.* 3, 2013. <https://doi.org/10.1038/srep02013>
- Dray, S., Dufour, A.-B., 2007. The ade4 Package: Implementing the Duality Diagram for Ecologists. *J. Stat. Softw.* 22, 1–20. <https://doi.org/10.18637/jss.v022.i04>
- Duarte, G.T., Volkova, P.Yu., Geras'kin, S.A., 2021. A Pipeline for Non-model Organisms for de novo Transcriptome Assembly, Annotation, and Gene Ontology Analysis Using Open Tools: Case Study with Scots Pine. *Bio-Protoc.* 11, e3912. <https://doi.org/10.21769/BioProtoc.3912>

- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., Aiden, E.L., 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95. <https://doi.org/10.1126/science.aal3327>
- Duffrin, M.W., Pomper, K.W., 2006. Development of Flavor Descriptors for Pawpaw Fruit Puree: A Step Toward the Establishment of a Native Tree Fruit Industry. *Fam. Consum. Sci. Res. J.* 35, 118–130. <https://doi.org/10.1177/1077727X06292931>
- Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3, 95–98. <https://doi.org/10.1016/j.cels.2016.07.002>
- Edwards, A.W.F., 2008. G. H. Hardy (1908) and Hardy–Weinberg Equilibrium. *Genetics* 179, 1143–1150. <https://doi.org/10.1534/genetics.104.92940>
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 323, 133–138. <https://doi.org/10.1126/science.1162986>
- Ekblom, R., Wolf, J.B.W., 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* 7, 1026–1042. <https://doi.org/10.1111/eva.12178>
- Ellegren, H., 2014. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 29, 51–63. <https://doi.org/10.1016/j.tree.2013.09.008>

- Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE* 6, e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E., Holzapfel, C.M., 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proc. Natl. Acad. Sci.* 107, 16196–16200. <https://doi.org/10.1073/pnas.1006538107>
- Englund, P.T., 1972. The 3'-terminal nucleotide sequences of T7 DNA. *J. Mol. Biol.* 66, 209–224. [https://doi.org/10.1016/0022-2836\(72\)90474-3](https://doi.org/10.1016/0022-2836(72)90474-3)
- Englund, P.T., 1971. Analysis of nucleotide sequences at 3' termini of duplex deoxyribonucleic acid with the use of the T4 deoxyribonucleic acid polymerase. *J. Biol. Chem.* 246, 3269–3276.
- Erkens, R.H.J., Blanpain, L.M.P., Carrascosa Jara, I., Runge, K., Verspagen, N., Cosiaux, A., Couvreur, T.L.P., 2022. Spatial distribution of Annonaceae across biomes and anthromes: Knowledge gaps in spatial and ecological data. *PLANTS PEOPLE PLANET* n/a. <https://doi.org/10.1002/ppp3.10321>
- Etherington, G.J., Heavens, D., Baker, D., Lister, A., McNelly, R., Garcia, G., Clavijo, B., Macaulay, I., Haerty, W., Di Palma, F., 2020. Sequencing smart: De novo sequencing and assembly approaches for a non-model mammal. *GigaScience* 9, giaa045. <https://doi.org/10.1093/gigascience/giaa045>
- Excoffier, L., Smouse, P.E., Quattro, J.M., 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491. <https://doi.org/10.1093/genetics/131.2.479>

- Ferrer-Blanco, C., Hormaza, J.I., Lora, J., 2022. Phenological growth stages of “pawpaw” [*Asimina triloba* (L.) Dunal, Annonaceae] according to the BBCH scale. *Sci. Hortic.* 295, 110853. <https://doi.org/10.1016/j.scienta.2021.110853>
- Fischer, M.C., Rellstab, C., Leuzinger, M., Roumet, M., Gugerli, F., Shimizu, K.K., Holderegger, R., Widmer, A., 2017. Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics* 18, 69. <https://doi.org/10.1186/s12864-016-3459-7>
- Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* 117, 9451–9457. <https://doi.org/10.1073/pnas.1921046117>
- Fox, S., 2012. Picking Up the Pawpaws: The Rare Woody Plants of Ontario Program at the University of Guelph Arboretum. *Arnoldia* 69 2–13 12.
- Frichot, E., François, O., 2014. LEA: an R package for Landscape and Ecological Association studies.
- Friel, J., Bombarely, A., Fornell, C.D., Luque, F., Fernández-Ocaña, A.M., 2021. Comparative Analysis of Genotyping by Sequencing and Whole-Genome Sequencing Methods in Diversity Studies of *Olea europaea* L. *Plants* 10, 2514. <https://doi.org/10.3390/plants10112514>
- Garrison, E., 2022. ekg/bamaddrg.
- Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio*.
- Geng, Q., Yao, Z., Yang, J., He, J., Wang, D., Wang, Z., Liu, H., 2015. Effect of Yangtze River on population genetic structure of the relict plant *Parrotia subaequalis* in eastern China. *Ecol. Evol.* 5, 4617–4627. <https://doi.org/10.1002/ece3.1734>

- Giani, A.M., Gallo, G.R., Gianfranceschi, L., Formenti, G., 2020. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* 18, 9–19. <https://doi.org/10.1016/j.csbj.2019.11.002>
- Gilbert, W., Maxam, A., 1973. The Nucleotide Sequence of the lac Operator. *Proc. Natl. Acad. Sci.* 70, 3581–3584. <https://doi.org/10.1073/pnas.70.12.3581>
- Gonzalez de la Rosa, P.M., Thomson, M., Trivedi, U., Tracey, A., Tandonnet, S., Blaxter, M., 2021. A telomere-to-telomere assembly of *Oscheius tipulae* and the evolution of rhabditid nematode chromosomes. *G3 GenesGenomesGenetics* 11, jkaa020. <https://doi.org/10.1093/g3journal/jkaa020>
- Goodrich, K.R., Zjhra, M.L., Ley, C.A., Raguso, R.A., 2006. When Flowers Smell Fermented: The Chemistry and Ontogeny of Yeasty Floral Scent in Pawpaw (*Asimina triloba*: Annonaceae). *Int. J. Plant Sci.* 167, 33–46. <https://doi.org/10.1086/498351>
- Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. <https://doi.org/10.1038/nrg.2016.49>
- Gottsberger, G., 1999. Pollination and evolution in neotropical Annonaceae. *Plant Species Biol.* 14, 143–152. <https://doi.org/10.1046/j.1442-1984.1999.00018.x>
- Götz, J., Rajora, O.P., Gailing, O., 2022. Genetic Structure of Natural Northern Range-Margin Mainland, Peninsular, and Island Populations of Northern Red Oak (*Quercus rubra* L.). *Front. Ecol. Evol.* 10.
- Grady, K.C., Ferrier, S.M., Kolb, T.E., Hart, S.C., Allan, G.J., Whitham, T.G., 2011. Genetic variation in productivity of foundation riparian species at the edge of their distribution: implications for restoration and assisted migration in a warming climate. *Glob. Change Biol.* 17, 3724–3735. <https://doi.org/10.1111/j.1365-2486.2011.02524.x>

- Groenendael, J.M. van, Klimes, L., Klimesova, J., Hendriks, R.J.J., 1996. Comparative ecology of clonal plants. 1339.
- Gruber, B., Unmack, P.J., Berry, O.F., Georges, A., 2018. dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Mol. Ecol. Resour.* 18, 691–699. <https://doi.org/10.1111/1755-0998.12745>
- Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., Durbin, R., 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinforma. Oxf. Engl.* 36, 2896–2898. <https://doi.org/10.1093/bioinformatics/btaa025>
- Hamrick, J.L., Godt, M.J.W., 1997. Effects of life history traits on genetic diversity in plant species. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 351, 1291–1298. <https://doi.org/10.1098/rstb.1996.0112>
- Hamrick, J.L., Godt, M.J.W., 1990. Allozyme diversity in plant species. *Plant Popul. Genet. Breed. Genet. Resour.* 43–63.
- Hamrick, J.L., Godt, M.J.W., Sherman-Broyles, S.L., 1992. Factors influencing levels of genetic diversity in woody plant species, in: Adams, W.T., Strauss, S.H., Copes, D.L., Griffin, A.R. (Eds.), *Population Genetics of Forest Trees: Proceedings of the International Symposium on Population Genetics of Forest Trees Corvallis, Oregon, U.S.A., July 31–August 2, 1990, Forestry Sciences*. Springer Netherlands, Dordrecht, pp. 95–124. https://doi.org/10.1007/978-94-011-2815-5_7
- Hansen, M., Kraft, T., Christiansson, M., Nilsson, N.-O., 1999. Evaluation of AFLP in Beta. *Theor. Appl. Genet.* 98, 845–852. <https://doi.org/10.1007/s001220051143>
- Hare, E.E., Johnston, J.S., 2011. Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol. Biol. Clifton NJ* 772, 3–12. https://doi.org/10.1007/978-1-61779-228-1_1

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., Gene Ontology Consortium, 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258-261. <https://doi.org/10.1093/nar/gkh036>

Hasing, T., Rinaldi, E., Manrique, S., Colombo, L., Haak, D.C., Zaitlin, D., Bombarely, A., 2019. Extensive phenotypic diversity in the cultivated Florist's Gloxinia, *Sinningia speciosa* (Lodd.) Hiern, is derived from the domestication of a single founder population. *PLANTS PEOPLE PLANET* 1, 363–374. <https://doi.org/10.1002/ppp3.10065>

Hass, B., Papanicolaou, A., 2016. TransDecoder.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., Ordoukhanian, P., 2014. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques* 56, 61-passim. <https://doi.org/10.2144/000114133>

Heller, R., Nursyifa, C., Garcia-Erill, G., Salmona, J., Chikhi, L., Meisner, J., Korneliussen, T.S., Albrechtsen, A., 2021. A reference-free approach to analyse RADseq data using standard next generation sequencing toolkits. *Mol. Ecol. Resour.* 21, 1085–1097. <https://doi.org/10.1111/1755-0998.13324>

- Henderson, J., Rodgers, C., Jones, R., Smith, J., Strzepek, K., Martinich, J., 2015. Economic impacts of climate change on water resources in the coterminous United States. *Mitig. Adapt. Strateg. Glob. Change* 20, 135–157. <https://doi.org/10.1007/s11027-013-9483-x>
- Herten, K., Hestand, M.S., Vermeesch, J.R., Van Houdt, J.K., 2015. GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics* 16, 73. <https://doi.org/10.1186/s12859-015-0514-3>
- Hewitt, G.M., 2004. Genetic consequences of climatic oscillations in the Quaternary. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 359, 183–195. <https://doi.org/10.1098/rstb.2003.1388>
- Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y., Fostier, J., 2017. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* 18. <https://doi.org/10.1186/s12859-017-1784-8>
- Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B., Poss, M.L., Reed, L.K., Storfer, A., Whitlock, M.C., 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *Am. Nat.* 188, 379–397. <https://doi.org/10.1086/688018>
- Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R., Zamir, A., 1965. STRUCTURE OF A RIBONUCLEIC ACID. *Science* 147, 1462–1465. <https://doi.org/10.1126/science.147.3664.1462>
- Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nat. Rev. Genet.* 10, 639–650. <https://doi.org/10.1038/nrg2611>
- Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491. <https://doi.org/10.1186/1471-2105-12-491>

- Hormaza, J.I., 2014. The Pawpaw, a Forgotten North American Fruit Tree. *Arnoldia Mag. Arnold Arbor.* 72, 11.
- Horn, C.N., 2015. A New Hybrid of *Asimina* (Annonaceae) Based on Morphological and Ecological Data. *Castanea* 80, 262–272. <https://doi.org/10.2179/15-067>
- Hou, X., Wang, D., Cheng, Z., Wang, Y., Jiao, Y., 2022. A near-complete assembly of an *Arabidopsis thaliana* genome. *Mol. Plant* 15, 1247–1250. <https://doi.org/10.1016/j.molp.2022.05.014>
- Hrabovetska, O.A., Derevyanko, V.M., Khokhlov, S., 2006. *Asimina triloba* (*Asimina triloba* (L.) Dun.): state and prospects of culture, bioecological features in growing conditions in the south of Ukraine Introduction of plants. *Introd. Plants* 3, 21–25.
- Huang, H., Knowles, L.L., 2016. Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Syst. Biol.* 65, 357–365. <https://doi.org/10.1093/sysbio/syu046>
- Huang, H., Layne, D.R., Kubisiak, T.L., 2003. Molecular Characterization of Cultivated Pawpaw (*Asimina triloba*) Using RAPD Markers. *J. Am. Soc. Hortic. Sci.* 128, 85–93. <https://doi.org/10.21273/JASHS.128.1.0085>
- Huang, H., Layne, D.R., Kubisiak, T.L., 2000. RAPD Inheritance and Diversity in Pawpaw (*Asimina triloba*). *J. Am. Soc. Hortic. Sci.* 125, 454–459. <https://doi.org/10.21273/JASHS.125.4.454>
- Huang, H., Layne, D.R., Peterson, R.N., 1997. Using Isozyme Polymorphisms for Identifying and Assessing Genetic Variation in Cultivated Pawpaw [*Asimina triloba* (L.) Dunal]. *J. Am. Soc. Hortic. Sci.* 122, 504–511. <https://doi.org/10.21273/JASHS.122.4.504>
- Huang, H., Layne, D.R., Riemenschneider, D.E., 1998. Genetic Diversity and Geographic Differentiation in Pawpaw [*Asimina triloba* (L.) Dunal] Populations from Nine States

- as Revealed by Allozyme Analysis. *J. Am. Soc. Hortic. Sci.* 123, 635–641.
<https://doi.org/10.21273/JASHS.123.4.635>
- Hufton, A.L., Panopoulou, G., 2009. Polyploidy and genome restructuring: a variety of outcomes. *Curr. Opin. Genet. Dev., Genomes and evolution* 19, 600–606.
<https://doi.org/10.1016/j.gde.2009.10.005>
- Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., Lemainque, A., Engelen, S., Wincker, P., Schacherer, J., Aury, J.-M., 2017. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *GigaScience* 6, giw018.
<https://doi.org/10.1093/gigascience/giw018>
- Ito I., Mutsuura O., 1956. Studies on the induced polyploids of the North American Pawpaws (I): Chromosome number and mixoploids in *Asimina triloba* DUNAL. *Sci. Rep. Saikyo Univ. Agric.* 8, 57-"60-2".
- Iverson, L.R., Prasad, A., Schwartz, M.W., 1999. Modeling potential future individual tree-species distributions in the eastern United States under a climate change scenario: a case study with *Pinus virginiana*. *Ecol. Model.* 115, 77–93.
[https://doi.org/10.1016/S0304-3800\(98\)00200-2](https://doi.org/10.1016/S0304-3800(98)00200-2)
- Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., Akeson, M., 2015. Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* 12, 351–356.
<https://doi.org/10.1038/nmeth.3290>
- Janzen, D.H., Martin, P.S., 1982. Neotropical Anachronisms: The Fruits the Gomphotheres Ate. *Science* 215, 19–27. <https://doi.org/10.1126/science.215.4528.19>
- Jiménez-Ruiz, J., Ramírez-Tejero, J.A., Fernández-Pozo, N., Leyva-Pérez, M. de la O., Yan, H., Rosa, R. de la, Belaj, A., Montes, E., Rodríguez-Ariza, M.O., Navarro, F., Barroso, J.B., Beuzón, C.R., Valpuesta, V., Bombarely, A., Luque, F., 2020. Transposon

- activation is a major driver in the genome evolution of cultivated olive trees (*Olea europaea* L.). *Plant Genome* 13, e20010. <https://doi.org/10.1002/tpg2.20010>
- Johnson, H.A., Gordon, J., McLaughlin, J.L., 1996. Monthly variations in biological activity of *Asimina triloba*. ASHS Press.
- Jombart, T., 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>
- Jombart, T., Ahmed, I., 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. <https://doi.org/10.1093/bioinformatics/btr521>
- Jombart, T., Devillard, S., Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. <https://doi.org/10.1186/1471-2156-11-94>
- Jung, H., Ventura, T., Chung, J.S., Kim, W.-J., Nam, B.-H., Kong, H.J., Kim, Y.-O., Jeon, M.-S., Eyun, S., 2020. Twelve quick steps for genome assembly and annotation in the classroom. *PLOS Comput. Biol.* 16, e1008325. <https://doi.org/10.1371/journal.pcbi.1008325>
- Jung, H., Winefield, C., Bombarely, A., Prentis, P., Waterhouse, P., 2019a. Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. *Trends Plant Sci.* <https://doi.org/10.1016/j.tplants.2019.05.003>
- Jung, H., Winefield, C., Bombarely, A., Prentis, P., Waterhouse, P., 2019b. Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. *Trends Plant Sci.* 24, 700–724. <https://doi.org/10.1016/j.tplants.2019.05.003>
- Kamvar, Z.N., Tabima, J.F., Grünwald, N.J., 2014. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2, e281. <https://doi.org/10.7717/peerj.281>

- Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Karrow, P.F., 1990. Interglacial Beds at Toronto, Ontario. *Géographie Phys. Quat.* 44, 289–297. <https://doi.org/10.7202/032830ar>
- Keener, C., Kuhns, E., 1997. The Impact of Iroquoian Populations on the Northern Distribution of Pawpaws in the Northeast. *North Am. Archaeol.* 18, 327–342. <https://doi.org/10.2190/AT0W-VEDT-0E0P-W21V>
- Kille, B., Balaji, A., Sedlazeck, F.J., Nute, M., Treangen, T.J., 2022. Multiple genome alignment in the telomere-to-telomere assembly era. *Genome Biol.* 23, 182. <https://doi.org/10.1186/s13059-022-02735-6>
- Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* gr.215087.116. <https://doi.org/10.1101/gr.215087.116>
- Korf, I., 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5, 59. <https://doi.org/10.1186/1471-2105-5-59>
- Korneliussen, T.S., Albrechtsen, A., Nielsen, R., 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15, 356. <https://doi.org/10.1186/s12859-014-0356-4>
- Kral, R., 1960. A revision of *Asimina* and *Deeringothamnus* (Annonaceae). *Brittonia* 12, 233–278. <https://doi.org/10.2307/2805119>
- Laetsch, D.R., Blaxter, M.L., 2017. BlobTools: Interrogation of genome assemblies. <https://doi.org/10.12688/f1000research.12232.1>
- Lagrange, R.L., Tramer, E.J., 1985. Geographic Variation in Size and Reproductive Success in the Paw Paw (*Asimina Triloba*).

- Lamichhaney, S., Barrio, A.M., Rafati, N., Sundström, G., Rubin, C.-J., Gilbert, E.R., Berglund, J., Wetterbom, A., Laikre, L., Webster, M.T., Grabherr, M., Ryman, N., Andersson, L., 2012. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proc. Natl. Acad. Sci. U. S. A.* 109, 19345–19350. <https://doi.org/10.1073/pnas.1216128109>
- Lampton, R.K., 1957. Floral Morphology in *Asimina triloba* Dunal. I. Development of Ovule and Embryo Sac. *Bull. Torrey Bot. Club* 84, 151–156. <https://doi.org/10.2307/2482886>
- Lawson, D.J., Dorp, L. van, Falush, D., 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* 9, 1–11. <https://doi.org/10.1038/s41467-018-05257-7>
- Layne, D.R., 1996. The Pawpaw [*Asimina triloba* (L.) Dunal]: A New Fruit Crop for Kentucky and the United States. *HortScience* 31, 777–784. <https://doi.org/10.21273/HORTSCI.31.5.777>
- Leberg, P.L., 1992. Effects of Population Bottlenecks on Genetic Diversity as Measured by Allozyme Electrophoresis. *Evolution* 46, 477–494. <https://doi.org/10.2307/2409866>
- Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. <https://doi.org/10.1038/nrg2825>
- Lehner, C., Roth, T., Kaul, H.-P., Neugschwandtner, R.W., 2022. (L.) Dunal: Langzeitstudie über 13 Sorten in Österreich. *Bodenkult. J. Land Manag. Food Environ.* 73, 41–54. <https://doi.org/10.2478/boku-2022-0003>
- Levin, S.A., Muller-Landau, *Helene C., Nathan, *Ran, Chave, *Jérôme, 2003. The Ecology and Evolution of Seed Dispersal: A Theoretical Perspective. *Annu. Rev. Ecol. Evol. Syst.* 34, 575–604. <https://doi.org/10.1146/annurev.ecolsys.34.011802.132428>

- Li, F.-W., Harkess, A., 2018. A guide to sequence your favorite plant genomes. *Appl. Plant Sci.* 6, e1030. <https://doi.org/10.1002/aps3.1030>
- Li, H., 2022. *lh3/seqtk*.
- Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio*.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, Y.-C., Korol, A.B., Fahima, T., Beiles, A., Nevo, E., 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11, 2453–2465. <https://doi.org/10.1046/j.1365-294x.2002.01643.x>
- Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., Fan, W., 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief. Funct. Genomics* 11, 25–37. <https://doi.org/10.1093/bfgp/elr035>
- Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. <https://doi.org/10.1126/science.1181369>
- Limasset, A., Cazaux, B., Rivals, E., Peterlongo, P., 2016. Read mapping on de Bruijn graphs. *BMC Bioinformatics* 17, 237. <https://doi.org/10.1186/s12859-016-1103-9>

- Locke, J.F., 1936. Microsporogenesis and Cytokinesis in *Asimina triloba*. *Bot. Gaz.* 98, 159–168. <https://doi.org/10.1086/334624>
- Lolletti, D., Principio, L., Ciorba, R., Mitrano, F., Ceccarelli, D., Antonucci, F., Manganiello, R., Ciccoritti, R., 2021. *Asimina triloba*: Crop years, cultivars and ripening time influence on qualitative parameters. *Sci. Hortic.* 289, 110481. <https://doi.org/10.1016/j.scienta.2021.110481>
- Lomsadze, A., Burns, P.D., Borodovsky, M., 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* 42, e119. <https://doi.org/10.1093/nar/gku557>
- Looy, K. van, Jacquemyn, H., Breyne, P., Honnay, O., 2009. Effects of flood events on the genetic structure of riparian populations of the grassland plant *Origanum vulgare*. *Biol. Conserv.* 142, 870–878.
- Lopez, O.R., 2001. Seed flotation and postflooding germination in tropical terra firme and seasonally flooded forest species. *Funct. Ecol.* 15, 763–771. <https://doi.org/10.1046/j.0269-8463.2001.00586.x>
- Losada, J.M., Hormaza, J.I., Lora, J., 2017. Pollen-pistil interaction in pawpaw (*Asimina triloba*), the northernmost species of the mainly tropical family Annonaceae. *Am. J. Bot.* 104, 1891–1903. <https://doi.org/10.3732/ajb.1700319>
- Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F., Storfer, A., 2017. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* 17, 142–152. <https://doi.org/10.1111/1755-0998.12635>
- Lu, L., Pomper, K.W., Lowe, J.D., Crabtree, S.B., 2011. Genetic Variation in Pawpaw Cultivars Using Microsatellite Analysis. *J. Am. Soc. Hortic. Sci.* 136, 415–421. <https://doi.org/10.21273/JASHS.136.6.415>

- Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., Chen, W.-M., 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. <https://doi.org/10.1093/bioinformatics/btq559>
- Marine, R., Polson, S.W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas, M., Wommack, K.E., 2011. Evaluation of a Transposase Protocol for Rapid Generation of Shotgun High-Throughput Sequencing Libraries from Nanogram Quantities of DNA. *Appl. Environ. Microbiol.* 77, 8071–8079. <https://doi.org/10.1128/AEM.05610-11>
- Mathebula, T.P., 2016. ProHint, A Static Code Analysis Tool for OpenEdge.
- McCage, C.M., Ward, S.M., Paling, C.A., Fisher, D.A., Flynn, P.J., McLaughlin, J.L., 2002. Development of a paw paw herbal shampoo for the removal of head lice. *Phytomedicine* 9, 743–748. <https://doi.org/10.1078/094471102321621377>
- McGrath, M.J., Karahadian, C., 1994. Evaluation of physical, chemical, and sensory properties of pawpaw fruit (*Asimina triloba*) as indicators of ripeness. *J. Agric. Food Chem. USA.*
- McLaughlin, J.L., 2008. Paw Paw and Cancer: Annonaceous Acetogenins from Discovery to Commercial Products. *J. Nat. Prod.* 71, 1311–1321. <https://doi.org/10.1021/np800191t>
- Meger, J., Ulaszewski, B., Vendramin, G.G., Burczyk, J., 2019. Using reduced representation libraries sequencing methods to identify cpDNA polymorphisms in European beech (*Fagus sylvatica* L). *Tree Genet. Genomes* 15, 7. <https://doi.org/10.1007/s11295-018-1313-6>
- Meirmans, P.G., 2006. Using the Amova Framework to Estimate a Standardized Genetic Differentiation Measure. *Evolution* 60, 2399–2402. <https://doi.org/10.1111/j.0014-3820.2006.tb01874.x>
- Meloni, M., Reid, A., Caujapé-Castells, J., Marrero, Á., Fernández-Palacios, J.M., Mesa-Coelo, R.A., Conti, E., 2013. Effects of clonality on the genetic variability of rare,

- insular species: the case of *Ruta microcarpa* from the Canary Islands. *Ecol. Evol.* 3, 1569–1579. <https://doi.org/10.1002/ece3.571>
- Messing, J., Crea, R., Seeburg, P.H., 1981. A system for shotgun DNA sequencing. *Nucleic Acids Res.* 9, 309–321. <https://doi.org/10.1093/nar/9.2.309>
- Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., Schneider, V.A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., Markovic, C., Maduro, V., Dutra, A., Bouffard, G.G., Chang, A.M., Hansen, N.F., Wilfert, A.B., Thibaud-Nissen, F., Schmitt, A.D., Belton, J.-M., Selvaraj, S., Dennis, M.Y., Soto, D.C., Sahasrabudhe, R., Kaya, G., Quick, J., Loman, N.J., Holmes, N., Loose, M., Surti, U., Risques, R. ana, Graves Lindsay, T.A., Fulton, R., Hall, I., Paten, B., Howe, K., Timp, W., Young, A., Mullikin, J.C., Pevzner, P.A., Gerton, J.L., Sullivan, B.A., Eichler, E.E., Phillippy, A.M., 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585, 79–84. <https://doi.org/10.1038/s41586-020-2547-7>
- Mimura, M., Aitken, S.N., 2007. Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. *Heredity* 99, 224–232. <https://doi.org/10.1038/sj.hdy.6800987>
- Montero-Mendieta, S., Grabherr, M., Lantz, H., De la Riva, I., Leonard, J.A., Webster, M.T., Vilà, C., 2017. A practical guide to build de-novo assemblies for single tissues of non-model organisms: the example of a Neotropical frog. *PeerJ* 5, e3702. <https://doi.org/10.7717/peerj.3702>
- Moore, A., 2015. *Pawpaw: In Search of America's Forgotten Fruit*. Chelsea Green Publishing.
- Muneepeerakul, R., Bertuzzo, E., Rinaldo, A., Rodriguez-Iturbe, I., 2008. Patterns of vegetation biodiversity: the roles of dispersal directionality and river network structure. *J. Theor. Biol.* 252, 221–229. <https://doi.org/10.1016/j.jtbi.2008.02.001>

- Murphy, J.L., 2001. Pawpaws, Persimmons, and 'Possums: On the Natural Distribution of Pawpaws in the Northeast. *North Am. Archaeol.* 22, 93–115. <https://doi.org/10.2190/TT5Y-YAXJ-0KAU-GUFK>
- Myers, E.W., 2005. The fragment assembly string graph. *Bioinformatics* 21, ii79–ii85. <https://doi.org/10.1093/bioinformatics/bti1114>
- Nam, J.-S., Park, S.-Y., Lee, S.-O., Lee, H.-J., Jang, H.-L., Rhee, Y.H., 2021. The growth-inhibitory effects of pawpaw (*Asimina triloba* [L.] Dunal) roots, twigs, leaves, and fruit against human gastric (AGS) and cervical (HeLa) cancer cells and their anti-inflammatory activities. *Mol. Biol. Rep.* <https://doi.org/10.1007/s11033-021-06226-y>
- Nam, J.-S., Park, S.-Y., Oh, H.-J., Jang, H.-L., Rhee, Y.H., 2019. Phenolic Profiles, Antioxidant and Antimicrobial Activities of Pawpaw Pulp (*Asimina triloba* [L.] Dunal) at Different Ripening Stages. *J. Food Sci.* 84, 174–182. <https://doi.org/10.1111/1750-3841.14414>
- Nei, M., 1972. Genetic Distance between Populations. *Am. Nat.* 106, 283–292. <https://doi.org/10.1086/282771>
- Nilsson, C., Brown, R.L., Jansson, R., Merritt, D.M., 2010. The role of hydrochory in structuring riparian and wetland vegetation. *Biol. Rev. Camb. Philos. Soc.* 85, 837–858. <https://doi.org/10.1111/j.1469-185X.2010.00129.x>
- Nilsson, C., Gardfjell, M., Grelsson, G., 1991. Importance of hydrochory in structuring plant communities along rivers. *Can. J. Bot.* 69, 2631–2633. <https://doi.org/10.1139/b91-328>
- Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Logsdon, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., Koren, S., 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30, 1291–1305. <https://doi.org/10.1101/gr.263566.120>

- Okada, H., Ueda, K., 1984. Cytotaxonomical Studies on Asian Annonaceae. *Plant Syst. Evol.* 144, 165–177.
- Ony, M., Klingeman, W.E., Zobel, J., Trigiano, R.N., Ginzl, M., Nowicki, M., Boggess, S.L., Everhart, S., Hadziabdic, D., 2021. Genetic diversity in North American *Cercis Canadensis* reveals an ancient population bottleneck that originated after the last glacial maximum. *Sci. Rep.* 11, 21803. <https://doi.org/10.1038/s41598-021-01020-z>
- Ortiz, D.A., Lima, A.P., Werneck, F.P., 2018. Environmental transition zone and rivers shape intraspecific population structure and genetic diversity of an Amazonian rain forest tree frog. *Evol. Ecol.* 32, 359–378. <https://doi.org/10.1007/s10682-018-9939-2>
- Ou, S., Chen, J., Jiang, N., 2018. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46, e126. <https://doi.org/10.1093/nar/gky730>
- Ou, S., Jiang, N., 2018. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* 176, 1410–1422. <https://doi.org/10.1104/pp.17.01310>
- Paradis, E., Schliep, K., 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinforma. Oxf. Engl.* 35, 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- Pellicer, J., Hidalgo, O., Dodsworth, S., Leitch, I.J., 2018. Genome Size Diversity and Its Impact on the Evolution of Land Plants. *Genes* 9, 88. <https://doi.org/10.3390/genes9020088>
- Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. <https://doi.org/10.1038/nbt.3122>
- Peterson, R.N., 2003. Pawpaw Variety Development: A History and Future Prospects. *HortTechnology* 13, 449–454. <https://doi.org/10.21273/HORTTECH.13.3.0449>

- Peterson, R.N., 1991. PAWPAW (ASIMINA), in: *Acta Horticulturae*. International Society for Horticultural Science (ISHS), Leuven, Belgium, pp. 569–602.
<https://doi.org/10.17660/ActaHortic.1991.290.13>
- Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E., Lercher, M.J., 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* 31, 1929–1936. <https://doi.org/10.1093/molbev/msu136>
- Pomper, K.W., Crabtree, S.B., Brown, S.P., Jones, S.C., Bonney, T.M., Layne, D.R., 2003. Assessment of Genetic Diversity of Pawpaw (*Asimina triloba*) Cultivars with Intersimple Sequence Repeat Markers. *J. Am. Soc. Hortic. Sci.* 128, 521–525.
<https://doi.org/10.21273/JASHS.128.4.0521>
- Pomper, K.W., Layne, D., 2003. *The North American pawpaw: botany and horticulture*, Horticultural Reviews. John Wiley & Sons.
- Pomper, K.W., Lowe, J.D., Lu, L., Crabtree, S.B., Collins, L.A., 2009. Clonality of Pawpaw (*Asimina triloba*) Patches in Kentucky. *J. Ky. Acad. Sci.* 70, 3–11.
<https://doi.org/10.3101/1098-7096-70.1.3>
- Pomper, K.W., Lowe, J.D., Lu, L., Crabtree, S.B., Dutta, S., Schneider, K., Tidwell, J., 2010. Characterization and Identification of Pawpaw Cultivars and Advanced Selections by Simple Sequence Repeat Markers. *J. Am. Soc. Hortic. Sci.* 135, 143–149.
<https://doi.org/10.21273/JASHS.135.2.143>
- Pongpanich, M., Sullivan, P.F., Tzeng, J.-Y., 2010. A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics* 26, 1731–1737.
<https://doi.org/10.1093/bioinformatics/btq272>
- Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–959.

- Pryszcz, L.P., Gabaldón, T., 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44, e113. <https://doi.org/10.1093/nar/gkw294>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575.
- Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13, 341. <https://doi.org/10.1186/1471-2164-13-341>
- Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J.H.J., Becker-Ziaja, B., Boettcher, J.P., Cabeza-Cabrerizo, M., Camino-Sánchez, Á., Carter, L.L., Doerrbecker, J., Enkirch, T., Dorival, I.G., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L.E., Koropogui, M., Kosgey, A., Kuisma, E., Logue, C.H., Mazzei, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallasch, E., Patrono, L.V., Portmann, J., Repits, J.G., Rickett, N.Y., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R.L., Zekeng, E.G., Racine, T., Bello, A., Sall, A.A., Faye, Ousmane, Faye, Oumar, Magassouba, N., Williams, C.V., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, Alimou, Somlare, H., Camara, Abdoulaye, Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K.Y., Diarra, A., Savane, Y., Pallawo, R.B., Gutierrez, G.J., Milhano, N., Roger, I., Williams, C.J., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., J. Turner, D., Pollakis, G., Hiscox, J.A., Matthews, D.A., Shea, M.K.O., Johnston, A.M., Wilson, D., Hutley, E., Smit, E., Di Caro, A., Wölfel, R., Stoecker, K.,

- Fleischmann, E., Gabriel, M., Weller, S.A., Koivogui, L., Diallo, B., Keïta, S., Rambaut, A., Formenty, P., Günther, S., Carroll, M.W., 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. <https://doi.org/10.1038/nature16996>
- Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Raj, A., Stephens, M., Pritchard, J.K., 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics* 197, 573–589. <https://doi.org/10.1534/genetics.114.164350>
- Ramaut, A., 2018. FigTree.
- Ratnayake, S., Rupprecht, K.J., Potter, W.M., McLaughlin, J.L., 1992. Evaluation of Various Parts of the Paw Paw Tree, *Asimina triloba* (Annonaceae), as Commercial Sources of the Pesticidal Annonaceous Acetogenins. *J. Econ. Entomol.* 85, 2353–2356. <https://doi.org/10.1093/jee/85.6.2353>
- Ravinet, M., Westram, A., Johannesson, K., Butlin, R., André, C., Panova, M., 2016. Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Mol. Ecol.* 25, 287–305. <https://doi.org/10.1111/mec.13332>
- Rhie, A., Walenz, B.P., Koren, S., Phillippy, A.M., 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21, 245. <https://doi.org/10.1186/s13059-020-02134-9>
- Rhoads, A., Au, K.F., 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics, SI: Metagenomics of Marine Environments* 13, 278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>

- Rivera-Colón, A.G., Rochette, N.C., Catchen, J.M., 2021. Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. *Mol. Ecol. Resour.* 21, 363–378. <https://doi.org/10.1111/1755-0998.13163>
- Roberts, D.R., Hamann, A., 2015. Glacial refugia and modern genetic diversity of 22 western North American tree species. *Proc. R. Soc. B Biol. Sci.* 282, 20142903. <https://doi.org/10.1098/rspb.2014.2903>
- Rogstad, S.H., Wolff, K., Schaal, B.A., 1991. Geographical Variation in *Asimina triloba* Dunal (Annonaceae) Revealed by the M13 “DNA Fingerprinting” Probe. *Am. J. Bot.* 78, 1391–1396. <https://doi.org/10.2307/2445277>
- Rubin, B.E.R., Ree, R.H., Moreau, C.S., 2012. Inferring Phylogenies from RAD Sequence Data. *PLOS ONE* 7, e33394. <https://doi.org/10.1371/journal.pone.0033394>
- Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>
- Salzberg, S.L., 2019. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* 20, 92. <https://doi.org/10.1186/s13059-019-1715-2>
- Salzberg, S.L., Yorke, J.A., 2005. Beware of mis-assembled genomes. *Bioinformatics* 21, 4320–4321. <https://doi.org/10.1093/bioinformatics/bti769>
- Sanger, F., Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94, 441–448. [https://doi.org/10.1016/0022-2836\(75\)90213-2](https://doi.org/10.1016/0022-2836(75)90213-2)
- Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., Petersen, G.B., 1982. Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* 162, 729–773. [https://doi.org/10.1016/0022-2836\(82\)90546-0](https://doi.org/10.1016/0022-2836(82)90546-0)

- Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>
- Sanger, F., Thompson, E.O.P., 1953a. The amino-acid sequence in the glyceryl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.* 53, 353–366. <https://doi.org/10.1042/bj0530353>
- Sanger, F., Thompson, E.O.P., 1953b. The amino-acid sequence in the glyceryl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem. J.* 53, 366–374. <https://doi.org/10.1042/bj0530366>
- Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240. <https://doi.org/10.1093/hmg/ddq416>
- Schleuning, M., Huamán, V., Matthies, D., 2008. Flooding and canopy dynamics shape the demography of a clonal Amazon understorey herb. *J. Ecol.* 96, 1045–1055. <https://doi.org/10.1111/j.1365-2745.2008.01416.x>
- Schuster, S.C., 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18. <https://doi.org/10.1038/nmeth1156>
- Seo, T.S., Bai, X., Kim, D.H., Meng, Q., Shi, S., Ruparel, H., Li, Z., Turro, N.J., Ju, J., 2005. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl. Acad. Sci.* 102, 5926–5931. <https://doi.org/10.1073/pnas.0501965102>
- Seppy, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness, in: Kollmar, M. (Ed.), *Gene Prediction: Methods and Protocols, Methods in Molecular Biology*. Springer, New York, NY, pp. 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

- Sharbel, T.F., Haubold, B., Mitchell-Olds, T., 2000. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* 9, 2109–2118. <https://doi.org/10.1046/j.1365-294X.2000.01122.x>
- Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H., 2017. DNA sequencing at 40: past, present and future. *Nature* 550, 345–353. <https://doi.org/10.1038/nature24286>
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Smit, A., Hubley, R., Green, P., 2015. RepeatMasker.
- Smith, L.M., Fung, S., Hunkapiller, M.W., Hunkapiller, T.J., Hood, L.E., 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* 13, 2399–2412. <https://doi.org/10.1093/nar/13.7.2399>
- Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., Hood, L.E., 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* 321, 674–679. <https://doi.org/10.1038/321674a0>
- Staden, R., 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* 6, 2601–2610. <https://doi.org/10.1093/nar/6.7.2601>
- Stanke, M., Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19, ii215–ii225. <https://doi.org/10.1093/bioinformatics/btg1080>
- Strijk, J.S., Hinsinger, D.D., Roeder, M.M., Chatrou, L.W., Couvreur, T.L.P., Erkens, R.H.J., Sauquet, H., Pirie, M.D., Thomas, D.C., Cao, K., 2021. Chromosome-level reference

- genome of the soursop (*Annona muricata*): A new resource for Magnoliid research and tropical pomology. *Mol. Ecol. Resour.* 21, 1608–1619. <https://doi.org/10.1111/1755-0998.13353>
- Sun, J.-T., Jiang, X.-Y., Wang, M.-M., Hong, X.-Y., 2014. Development of microsatellite markers for, and a preliminary population genetic analysis of, the white-backed planthopper. *Bull. Entomol. Res.* 104, 765–773. <https://doi.org/10.1017/S0007485314000613>
- Tabacu, A.F., Butcaru, A.C., Stan, A., Mihai, C.A., Stănică, F., 2020. Pawpaw Hybrid Genotypes (*Asimina triloba* (L.) Dunal) Cultivated in the Bucharest Area. *Bull. Hortic.* 77, 14. <https://doi.org/DOI:10.15835/buasvmcn-hort:2020.0014>
- Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. <https://doi.org/10.1093/genetics/123.3.585>
- Tanaka, R., Okada, H., 1972. Karyological studies in four species of Annonaceae, a primitive angiosperm. *J. Sci. Hiroshima Univ. Ser. B Div 2*, 85–105. *J Sci Hiroshima Univ, Ser. B Div 2*, 85–105.
- Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* 25, 4.10.1-4.10.14. <https://doi.org/10.1002/0471250953.bi0410s25>
- Teo, Y.Y., Fry, A.E., Clark, T.G., Tai, E.S., Seielstad, M., 2007. On the usage of HWE for identifying genotyping errors. *Ann. Hum. Genet.* 71, 701–703; author reply 704. <https://doi.org/10.1111/j.1469-1809.2007.00356.x>
- Tom, J.A., Reeder, J., Forrest, W.F., Graham, R.R., Hunkapiller, J., Behrens, T.W., Bhangale, T.R., 2017. Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics* 18, 351. <https://doi.org/10.1186/s12859-017-1756-z>

- Tomé, L.M.R., da Silva, F.F., Fonseca, P.L.C., Mendes-Pereira, T., Azevedo, V.A. de C., Brenig, B., Badotti, F., Góes-Neto, A., 2022. Hybrid Assembly Improves Genome Quality and Completeness of *Trametes villosa* CCMB561 and Reveals a Huge Potential for Lignocellulose Breakdown. *J. Fungi* 8, 142. <https://doi.org/10.3390/jof8020142>
- Torkamaneh, D., Laroche, J., Belzile, F., 2016. Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. *PLOS ONE* 11, e0161333. <https://doi.org/10.1371/journal.pone.0161333>
- Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38, e159. <https://doi.org/10.1093/nar/gkq543>
- Tsykun, T., Rellstab, C., Dutech, C., Sipos, G., Prospero, S., 2017. Comparative assessment of SSR and SNP markers for inferring the population genetic structure of the common fungus *Armillaria cepistipes*. *Heredity* 119, 371–380. <https://doi.org/10.1038/hdy.2017.48>
- Tulowiecki, S.J., 2021. Modeling the geographic distribution of pawpaw (*Asimina triloba* [L.] Dunal) in a portion of its northern range limits, western New York State, USA. *Plant Ecol.* 222, 193–208. <https://doi.org/10.1007/s11258-020-01098-x>
- Turakulov, R., Easteal, S., 2003. Number of SNPS loci needed to detect population structure. *Hum. Hered.* 55, 37–45. <https://doi.org/10.1159/000071808>
- van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., Thermes, C., 2018. The Third Revolution in Sequencing Technology. *Trends Genet.* 34, 666–681. <https://doi.org/10.1016/j.tig.2018.05.008>

- Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., Schatz, M.C., 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. <https://doi.org/10.1093/bioinformatics/btx153>
- Wang, Y., Reighard, G.L., Layne, D.R., Abbott, A.G., Huang, H., 2005. Inheritance of AFLP Markers and Their Use for Genetic Diversity Analysis in Wild and Domesticated Pawpaw [*Asimina triloba* (L.) Dunal]. *J. Am. Soc. Hortic. Sci.* 130, 561–568. <https://doi.org/10.21273/JASHS.130.4.561>
- Wei, X., Meng, H., Jiang, M., 2013. Landscape Genetic Structure of a Streamside Tree Species *Euptelea pleiospermum* (Eupteleaceae): Contrasting Roles of River Valley and Mountain Ridge. *PLOS ONE* 8, e66928. <https://doi.org/10.1371/journal.pone.0066928>
- Weir, B.S., Cockerham, C.C., 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38, 1358–1370. <https://doi.org/10.2307/2408641>
- Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37, 1155–1162. <https://doi.org/10.1038/s41587-019-0217-9>
- Whitlock, M.C., Lotterhos, K.E., 2015. Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F_{ST} . *Am. Nat.* 186, S24–S36. <https://doi.org/10.1086/682949>
- Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York.

- Wilkinson, M., 1995. Coping with Abundant Missing Entries in Phylogenetic Inference Using Parsimony. *Syst. Biol.* 44, 501–514. <https://doi.org/10.1093/sysbio/44.4.501>
- Willson, M.F., Schemske, D.W., 1980. Pollinator Limitation, Fruit Production, and Floral Display in Pawpaw (*Asimina triloba*). *Bull. Torrey Bot. Club* 107, 401–408. <https://doi.org/10.2307/2484160>
- Winter, D., 2022. repeatR.
- Wright, B.R., Grueber, C.E., Lott, M.J., Belov, K., Johnson, R.N., Hogg, C.J., 2019. Impact of reduced-representation sequencing protocols on detecting population structure in a threatened marsupial. *Mol. Biol. Rep.* 46, 5575–5580. <https://doi.org/10.1007/s11033-019-04966-6>
- Wright, S., 1943. Isolation by Distance. *Genetics* 28, 114–138.
- Wright, S., 1922. Coefficients of Inbreeding and Relationship. *Am. Nat.* 56, 330–338.
- Wu, R., Kaiser, A.D., 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* 35, 523–537. [https://doi.org/10.1016/S0022-2836\(68\)80012-9](https://doi.org/10.1016/S0022-2836(68)80012-9)
- Wyatt, G.E., Hamrick, J.L., Trapnell, D.W., 2021. The role of anthropogenic dispersal in shaping the distribution and genetic composition of a widespread North American tree species. *Ecol. Evol.* 11, 11515–11532. <https://doi.org/10.1002/ece3.7944>
- Wyatt, G.E., Hamrick, J.L., Trapnell, D.W., 2020. Ecological niche modelling and phylogeography reveal range shifts of pawpaw, a North American understory tree. *J. Biogeogr.* n/a. <https://doi.org/10.1111/jbi.14054>
- Wykoff, W., 2009. On the Natural Distribution of Pawpaw in the Northeast.
- Xinkun, H., Lijuan, W., Jundong, H., Shian, S., 2021. Research Progress of Introduced Pawpaw Tree from North America. *J. Sichuan For. Sci. Technol.*

- Xu, F., Li, X., Ren, H., Zeng, R., Wang, Z., Hu, H., Bao, J., Que, Y., 2022. The First Telomere-to-Telomere Chromosome-Level Genome Assembly of *Stagonospora tainanensis* Causing Sugarcane Leaf Blight. *J. Fungi Basel Switz.* 8, 1088. <https://doi.org/10.3390/jof8101088>
- Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., Zhang, G., Gu, Y.Q., Coleman-Derr, D., Xia, Q., Wang, Y., 2019. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 47, W52–W58. <https://doi.org/10.1093/nar/gkz333>
- Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342. <https://doi.org/10.1038/nrg3174>
- Yi, X., Latch, E.K., 2022. Nonrandom missing data can bias Principal Component Analysis inference of population genetic structure. *Mol. Ecol. Resour.* 22, 602–611. <https://doi.org/10.1111/1755-0998.13498>
- Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. <https://doi.org/10.1101/gr.074492.107>
- Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A Greedy Algorithm for Aligning DNA Sequences. *J. Comput. Biol.* 7, 203–214. <https://doi.org/10.1089/10665270050081478>
- Zhang, Z.-Y., Zheng, X.-M., Ge, S., 2007. Population genetic structure of *Vitex negundo* (Verbenaceae) in Three-Gorge Area of the Yangtze River: The riverine barrier to seed dispersal in plants. *Biochem. Syst. Ecol.* 35, 506–516. <https://doi.org/10.1016/j.bse.2007.01.014>
- Zhao, G.-X., Miesbauer, L.R., Smith, D.L., McLaughlin, J.L., 1994. Asimin, Asiminacin, and Asiminecin: Novel Highly Cytotoxic Asimicin Isomers from *Asimina triloba*. *J. Med. Chem.* 37, 1971–1976. <https://doi.org/10.1021/jm00039a009>

Zimmerman, G.A., 1941. HYBRIDS OF THE AMERICAN PAPAWE. *J. Hered.* 32, 83–93.

<https://doi.org/10.1093/oxfordjournals.jhered.a105006>

Zomlefer, W.B., Comer, J.R., Lucardi, R.D., Hamrick, J.L., Allison, J.R., 2018. Distribution and genetic diversity of the rare plant *Veratrum woodii* (Liliales: Melanthiaceae) in Georgia: A preliminary study with AFLP fingerprint data. *Syst. Bot.* 43, 858–869.

<https://doi.org/10.1600/036364418X697779>