

**Developments in Hardwood Genomics: The Generation and Application
of Molecular Resources for Species Preservation**

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Matthew L. Huff
May 2023**

Copyright © 2023 by Matthew L. Huff
All rights reserved.

DEDICATION

This work is dedicated to the memory of Charles Leonard, Sr.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my boss and mentor, Dr. Margaret Staton, for all of her guidance and support throughout my work to obtain my Doctorate and for continuing to help my growth as a scientist. I would like to thank the current members of the Staton lab, in particular Zane Smith for his help in building Chapters IV and V of my dissertation. I would like to thank Beant Kapoor and Trinity Hamm for their help in implementing specific packages. I would like to thank Dr. Lav Yadav, Allyson Dekovich, Meher Ony, and Holly Brabazon for their general help throughout the dissertation process. I would like to thank Abdullah Almasaeed for his help in general coding advice early on in the research period. I would also like to thank the numerous Staton Lab Alumni who have since graduated. I would like to thank my committee members - Dr. Bode Olukolu, Dr. Denita Hadziabdic-Guerry, Dr. DeWayne Shoemaker, and Dr. Jeanne Romero-Severson for their support, suggestions, and greater conversations about my work in order to improve it. I would like to thank my collaborators for their efforts in the green ash genome paper and the muscadine chapter. I would like to thank Dr. Bill Klingeman, Evan Staton, Grace Pietsch, Josh Granger, Jason Granger, Ellen Crocker, and Megan Buland for their help in collecting green ash samples in Tennessee and beyond. And finally, I would like to thank my friends and family for their continued support.

ABSTRACT

The field of plant genomics aims to better understand plants, and the threats they face, through analysis of their genomes. Plant species, particularly woody tree species, offer special challenges when compared to other genomes. Recent advances in sequencing technologies, and continually decreasing sequencing costs, have allowed for more plant genomes to be produced. This project utilized bioinformatics technologies to characterize the genomes of economically important tree species by generating new genomes. My projects have a particular focus on green ash (*Fraxinus pennsylvanica*). Once a prominent forest species with economic and ecological importance, green ash is now threatened by the invasive emerald ash borer (EAB). In my first project, I report on a chromosome-level assembly for green ash and describe its annotation. This genome is used to characterize a whole genome duplication event within the genome, and is also used as a template for chromosome-scale scaffolds for 27 other ash species. In the following project, I explore the mechanisms of sex determination in green ash using this genome. I used a genotyping-by-sequencing (GBS) approach to identify a region of interest that was more likely to be present in female ash than in male ash. My final project focuses on the muscadine grape (*Vitis rotundifolia*), where I report a long-read assembly for the muscadine cultivar ‘Carlos.’ I used this assembly to generate a pan-genome using existing *Vitis* species genomes - including two other muscadine cultivars - and to identify unique genetic responses to inoculation with the bacterial pathogen *Xylella fastidiosa* subsp. *fastidiosa* compared to those of common grapevine (*Vitis vinifera*). This project demonstrates the utility of bioinformatics approaches in providing genomic resources to economically important or threatened species.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION TO TREE GENOMES AND SEQUENCING	
APPROACHES.....	1
Overview of Plant Genomics	1
Importance for Hardwoods, Ornamentals, and Woody Fruiting Plants.....	1
Challenges in Tree Genomics	3
Pan-genomes and the future standard of genomics.....	4
Overview of Sequencing Technologies	4
Short-Read Sequencing: Illumina	5
Long-Read Sequencing: Nanopore and Pac-Bio HiFi	6
Applications of Sequencing Technologies.....	7
Whole Genome Sequencing.....	8
Reduced-Representation Sequencing.....	9
RNA Sequencing	10
REFERENCES	12
CHAPTER II: INTRODUCTION TO GREEN ASH (<i>FRAXINUS PENNSYLVANICA</i>). 21	
<i>Fraxinus pennsylvanica</i> (green ash)	21
Distribution, usage, and importance	21
Emerald ash borer and other threats.....	23
EAB resistance in green ash and breeding efforts	23
Green ash dioecy	25
Overview of sex-determining systems in ash species.....	25
Self-incompatibility systems in ash	26
Currently known systems of plant dioecy.....	27
Two-gene systems.....	27
Single-gene systems.....	28
The overall model and its applicability to green ash	29
REFERENCES	31
CHAPTER III: A HIGH-QUALITY REFERENCE GENOME FOR <i>FRAXINUS</i>	
<i>PENNSYLVANICA</i> FOR ASH SPECIES RESTORATION AND RESEARCH 35	
Abstract	36
Introduction.....	36
Materials and Methods.....	38
Genome annotation and quality assessment	38
Characterizing the whole genome duplication event	38
Fraxinus spp. reference-guided genome scaffolding	39
Comparative genomics with Asterids	39
Results.....	39
Genetic linkage map	39
Assembly and scaffolding.....	39
Annotation.....	42
rRNA characterization	42
Characterization of whole genome duplications	43

Comparison of <i>F. pennsylvanica</i> to <i>F. excelsior</i>	43
Examination of EAB resistance candidate genes.....	45
Reference-guided assembly of worldwide <i>Fraxinus</i> genomes.....	45
The <i>F. pennsylvanica</i> genome shares conserved blocks of synteny with other asterids	47
Discussion	47
REFERENCES	53
APPENDIX.....	59
Supplemental Methods.....	59
CHAPTER IV: IDENTIFICATION OF MARKERS OF SEX DETERMINATION IN <i>FRAXINUS PENNSYLVANICA</i>	60
Abstract	61
Introduction.....	61
Materials and Methods.....	63
Sample collection.....	63
Genotyping-by-sequencing read processing and variant calling	64
Results.....	65
Sample population	65
Genotyping-by-sequencing SNPs were not associated with sex	65
A candidate presence/absence variation region identified on Chromosome 8	65
Annotation of genes located in close proximity to the region of interest	65
Discussion	73
REFERENCES	76
CHAPTER V: LONG-READ, CHROMOSOME-SCALE ASSEMBLY OF <i>VITIS ROTUNDIFOLIA</i> CV. CARLOS AND ITS UNIQUE RESISTANCE TO <i>XYLELLA FASTIDIOSA</i> SUBSP. <i>FASTIDIOSA</i>	80
Abstract	81
Introduction.....	81
Materials and Methods.....	83
Plant sampling - DNA sequencing.....	83
Plant sampling - RNA Sequencing	83
Genome scaffolding and assembly	84
Gene annotation	84
Comparative genomics of 'Carlos' to other genomes.....	85
Differential expression analysis.....	85
Results.....	86
Genome assembly information and statistics.....	86
Gene annotation report.....	86
Comparison of assembly to existing cultivars and common grape genome	88
Differential expression in response to Pierce's Disease	92
Discussion	100
REFERENCES	104
CHAPTER VI: CONCLUSION	112
VITA.....	114

LIST OF TABLES

Table 1. Summary of the <i>Fraxinus pennsylvanica</i> genome assembly.....	41
Table 2. Comparison of <i>Fraxinus excelsior</i> genome statistics before (version 0.5) and after (version 0.7) reference-guided assembly and reannotation.	46
Table 3. Orthogroup statistics of genes from six Asterids.....	49
Table 4. Ash to olive chromosomal synteny defined by single-copy orthologs.	50
Table 5. Significant SNPs from FreeBayes lack expected genotype segregation for a sex- linked marker.	68
Table 6. Statistics of variants of interest in the region on Chromosome 8.	71
Table 7. Genes from <i>Fraxinus pennsylvanica</i> located in proximity to the region of interest on Chromosome 8.	72
Table 8. Summary of the <i>Vitis rotundifolia</i> 'Carlos' genome assembly.....	87
Table 9. Quantitative statistics of collinearity between <i>Vitis rotundifolia</i> 'Carlos', <i>V.</i> <i>rotundifolia</i> 'Trayshed', and <i>V. rotundifolia</i> 'Noble' assemblies.	90
Table 10. Muscadine pan-genome information and statistics.....	93
Table 11. Differential expression results in <i>Vitis rotundifolia</i> 'Carlos' and <i>Vitis vinifera</i>	95
Table 12. Placement of candidate R-genes in Orthogroups.....	97
Table 13. Functional annotation of orthogroups of interest.....	99

LIST OF FIGURES

Figure 1. Phenotypic characteristics of green ash (<i>Fraxinus pennsylvanica</i>).....	22
Figure 2. The effects of emerald ash borer (EAB; <i>Agrilus planipennis</i>) on <i>Fraxinus pennsylvanica</i>	24
Figure 3. High-density, consensus genetic linkage map of <i>Fraxinus pennsylvanica</i> (Huff et al., 2022).	40
Figure 4. Characterization of Genome Duplications in <i>Fraxinus pennsylvanica</i>	44
Figure 5. Reference-guided assembly of <i>Fraxinus</i> taxa genomes.	48
Figure 6. Map of <i>Fraxinus</i> sp. Samples Collected for Study.	66
Figure 7. Flower structure of male (A) and female (B) <i>Fraxinus pennsylvanica</i> flowers.	67
Figure 8. Manhattan plot of significantly present/absent variants obtained from FreeBayes.	69
Figure 9. A region along Chromosome 8 with preferential presence in females of <i>Fraxinus</i> section <i>Melioides</i> females and absence in males of <i>Fraxinus</i> sect. <i>Melioides</i>	70
Figure 10. Plot of collinearity between <i>Vitis rotundifolia</i> 'Carlos', <i>V. rotundifolia</i> 'Trayshed', and <i>V. rotundifolia</i> 'Noble' assemblies.	89
Figure 11. Plot of collinearity between <i>Vitis rotundifolia</i> 'Carlos' and <i>Vitis vinifera</i> v12.1 assemblies.	91
Figure 12. Gene Counts of R-Gene Classes for Three Muscadine Cultivars and Common Grape.	94
Figure 13. Overlap of orthogroups associated with differentially expressed genes in <i>Vitis rotundifolia</i> 'Carlos' and <i>Vitis vinifera</i>	98

LIST OF ATTACHMENTS

Supplemental Tables (MHuffDissertation_SupplementalTables.xlsx)

Supplemental Figures (MHuffDissertation_SupplementalFigures.docx)

CHAPTER I: INTRODUCTION TO TREE GENOMES AND SEQUENCING APPROACHES

Overview of Plant Genomics

Genomics is the study of the genome, particularly focusing on the structure, function, and evolution of genomes, and how they compare to those of other species (Del Giacco & Cattaneo, 2012). Genomes are assembled using bioinformatics pipelines, and they are then used to answer questions about the organism of interest and its relationship to other species. For instance, genomes can be used to identify traits of interest in an economically important species, or they can be used to compare gene expression in response to a stressor of interest. The first full reference plant genome became available in 2000 with the *Arabidopsis thaliana* assembly (Initiative & The Arabidopsis Genome Initiative, 2000), and since then the number of available plant genomes has only continued to increase with advancements in sequencing technology and reduced sequencing costs. While only a small fraction of existing land plants have been sequenced, and while there is a bias towards primarily herbaceous species of economic importance, current sequencing efforts explore a wide range of plants across taxonomies (Marks et al., 2021). Of particular interest to some researchers are non-herbaceous plants with woody tissues, including forest tree species, ornamental tree species, and woody vine plant species.

Importance for Hardwoods, Ornamentals, and Woody Fruiting Plants

Hardwood trees provide a multitude of products and labor for the lumber industry and beyond, with an estimated value of \$135.6 billion for the United States' hardwood products (*Hardwood Federation - Hardwood Industry Economic Impact*, n.d.). Many native hardwood species have become threatened due to a combination of climate change, human disturbance, and the introduction of invasive pests and pathogens. Plant genomics is important to understand underlying mechanisms and genetic components of tree stress response, disease resistance, and other traits of interest in forest health and conservation. The growing field of tree genomics has given researchers the tools needed to identify genes of interest and their relation to traits that can improve hardwood survival (Neale & Kremer, 2011; Plomion et al., 2016). As a prominent example, the American chestnut (*Castanea dentata*) was once a dominant species in American forests and a source of food and lumber, but the introduction of the chestnut blight in the 1920s rendered it functionally extinct. Plant researchers have worked to restore the American chestnut to the environment through breeding, biotechnology, and biocontrol (Anagnostakis, 2012; Bolgiano & Novak, 2007; Hebard, 1994). While attempts at breeding with resistant species, such as the Chinese chestnut (*Castanea mollissima*), have brought some level of success, a truly resistant American chestnut has yet to be produced. Genomes for both species have been published online (Sandercock et al., 2022; Staton et al., 2020), which will allow researchers to identify markers of resistance in Chinese chestnuts. This offers two opportunities to improve attempts at conferring resistance to American chestnuts - either by tracking these markers through a breeding program, or by using transgenic technology to introduce resistant genes into individual trees (Powell et al., 2019; Westbrook et al., 2020).

A number of annotated hardwood tree genomes have been generated in recent years, and their effects on the research surrounding their respective species cannot be understated. The

release of the first chromosome scale oak genome, the pedunculate oak (*Quercus robur*) (Plomion et al., 2018) spurred new intra- and interspecific *Quercus* genomic research. In the realm of population genetics, a recent paper made use of the *Q. robur* genome in identifying the introgression of genes from *Q. dentata* in populations of *Q. crispula*. *Q. dentata* is adapted to coastal stress, while *Q. crispula* is an inland species, but a population of *Q. crispula* with the traits of *dentata* has emerged in high-altitude coastal areas of Northern Japan (Nagamitsu et al., 2020). For *Q. robur* itself, it is the target of two exotic fungal pathogens: *Erysiphe alphitoides* and *Phytophthora cinnamomi*. Native species of *Q. robur* do not have natural resistance to either fungus, and usually succumb to infection. However, some of these “naive” trees display what is known as “exapted” resistance, or resistance that evolved for some function outside of resistance to a particular pathogen (Newcombe, 1998). Utilizing the recent assembly, researchers have identified two genomic regions containing six quantitative trait loci (QTLs) that highly contribute to phenotypic variation in response to these two pathogens (Bartholomé et al., 2020). These studies indicate an ability for plants to adapt to the exotic pathogens that threaten them, and knowledge of their genomic structure allowed them to identify QTLs associated with these mechanisms.

For ornamental trees, or trees with value based on their flowers and plant architectures, genomics can be used to identify the basis of a trait of interest and how to spread that trait to other populations. Some flower colors are more desirable than others, and an understanding of what genes drive the development of these valuable phenotypes will allow breeders to better select crosses to produce progeny with desirable traits. The first ornamental plant genome sequence was completed in 2012, with the ornamental tree *Prunus mume* (Q. Zhang et al., 2012). Since then, more ornamental trees have undergone genome sequencing (Zheng et al., 2021). In Tennessee, the dogwood (*Cornus* sp.) is a species of key economic importance with an estimated market value of \$7.4 million in Tennessee alone in 2019 according to the USDA (2019 *Census of Horticultural Specialties*, n.d.). Several cultivars of dogwood are susceptible to powdery mildew, an infection caused by fungal pathogens *Erysiphe pulchra* and *Phyllactinia guttata* (Mmbaga et al., 2016). While it is unlikely to kill dogwoods, powdery mildew harms the value of trees due to the unattractive appearance of infected leaves. Certain cultivars of dogwood have been bred to have resistance to powdery mildew, but their bracts are purely white. Here, the development of dogwood genomes will allow researchers to identify the genetic bases of powdery mildew resistance and bract color to select for resistant cultivars with red or pink bracts. Recent QTL analyses have identified candidate traits for powdery mildew resistance (Pfarr Moreau et al., 2022), and expanded genomic resources will allow researchers to identify specific genes controlling bract color.

Finally, many commercial fruits grow on woody trees or vines. Here, genomics can be used to improve fruit trees and their fruit products by identifying desirable traits such as fruit quality, phenolic content, and pathogen response (Licciardello et al., 2021). Once a phenotype has been linked to a genotype, breeders will be able to make more informed decisions about how to breed a plant to select for traits of interest. This can be made complex by some cultivated fruits having complex polyploid genomes, such as the tetraploid blueberry (Cappai et al., 2020) and octoploid strawberry (Edger et al., 2019). The development of new techniques for genome sequencing and their resulting genomic resources is important in fruit species with complex, polyploid genomes. Furthermore, many fruits are affected by diseases that result in reduced

yields, such as the gram-negative bacteria *Xylella fastidiosa* subsp. *fastidiosa* (*Xfsf*), which causes disease in species such as citrus trees, grapevines, and olive trees (Sicard et al., 2018). As with diseases in hardwood and ornamental trees, genome sequencing can be used to identify markers of inheritable disease resistance for breeding efforts to propagate. In apple (*Malus* sp.), a recent QTL experiment identified reduced fire blight (*Erwinia amylovora*) susceptibility alleles in the ‘Honeycrisp’ cultivar (Kostick et al., 2021). Similar QTL mapping efforts have identified *PdR1*, a gene driving *Xfsf*-resistance, in grapevine (*Vitis* sp.) (Krivanek et al., 2006). These studies provide a foundation for breeders to identify cultivars with disease resistance, allowing for introgression of resistant alleles into currently susceptible cultivars.

Challenges in Tree Genomics

Plant genomes vary widely in size, structure, and complexity, requiring significant resources to fully sequence. The first plant genome, *Arabidopsis thaliana*, was selected for its small size, around 135 Mb in five haploid chromosomes, and its tractability as a laboratory model for plant growth and development (Initiative & The Arabidopsis Genome Initiative, 2000). By comparison, the black cottonwood (*Populus trichocarpa*) genome - the first sequenced tree genome - is around ~500 Mb with 19 haploid chromosomes (Tuskan et al., 2006). Gymnosperms, such as conifers and ginkgos, have particularly large and complex genomes (Morse et al., 2009), with sugar pine (*Pinus lambertiana*) having a genome size of around 31 Gb (Stevens et al., 2016). In addition to size considerations, many angiosperms have undergone whole genome duplication events and/or hybridization events, leading to polyploidization, i.e., having more than two sets of chromosomes. Angiosperms may have 4x (tetraploid), 6x (hexaploid), 8x (octoploid) or even higher copy numbers of their chromosomes (Kyriakidou et al., 2018). Angiosperms also often have historical whole genome duplication events that are only partially purged, leading to complex chromosomal structure with residual homology despite diploid-like chromosomal pairing. While these internal duplications are important for the development of genes with novel functions (Crow, 2006), it becomes more difficult for researchers to tie phenotypes to genotypes. In polyploid genomes, it is important to correctly identify different haplotypes - known as haplotype phasing - for more accurate genome assemblies (Browning & Browning, 2011). Finally, plant genomes often contain a large percentage of repetitive regions, which creates issues in genome assembly strategies (Mehrotra & Goyal, 2014). Overall, the process of creating a new genome in tree species is often complex and has significant barriers for producing accurate, chromosome-scale assemblies.

In woody species in particular, associating the genome with traits of interest is challenging. One major approach to determining regions of the genome that control phenotypes of interest is through controlled breeding. In comparison to breeding programs in large scale agricultural systems or with laboratory model species like *Arabidopsis thaliana*, similar programs in forest tree species face significant obstacles. Model and agricultural plants are often herbaceous, annual plants that require a relatively small amount of space to grow and have short generation times. Forest trees are perennial, long-lived, large and have longer generation times than herbaceous species. Tree breeding programs take much longer and are more expensive than other programs to account for these differences. Certain traits of interest, such as wood development and related phenotypes, do not manifest for years or even decades. In some tree

species, flowering first occurs anywhere from 10 to 25 years after planting (Hausman et al., 2014; Schopmeyer & Others, 1974).

Breeding programs may face additional restrictions based on the reproductive physiology of the target plant species. Many plants are dioecious, where male and female traits are separated between individuals. Often, the sex-determining mechanisms are unknown, and individuals can only be sexed by observing their first flowers. These factors can delay breeding efforts due to uncertainty that fully grown will be compatible with one another. Similarly, many hermaphrodite plants, or plants with both sex traits on the same flower, have a self-incompatibility (SI) system, which ensures that the plants are unable to self-fertilize (Watanabe et al., 2012). In some systems, researchers have worked to overcome SI systems to promote self-compatibility (SC). This results in improved crop productivity but runs the risk of producing overly in-bred populations by allowing plants to self-cross.

Pan-genomes and the future standard of genomics

While the standard model of a single genome representing each organism has worked for many years, it is unable to capture all of the phenotypic and genotypic variation between cultivars in a species (Bayer et al., 2020; Hufford et al., 2021). With the recent advances in sequencing technology and lower sequencing costs, a new concept, the “pan-genome,” aims to overcome these limitations (Bayer et al., 2020; Della Coletta et al., 2021). A pan-genome is a collection of genomes representing multiple genotypes of a species, or multiple species within a genus. Genes in a pan-genome are marked as “core” if they are present in every member of the pan-genome or “accessory”/“variable” if they are absent in at least one member of the pan-genome. Pan-genomes may be “open” or “closed,” depending on how many gene families are captured in the pan-genome. An open pan-genome will see more gene families added with additional genomes, while a closed pan-genome will have a comprehensive collection of gene families. As such, an open pan-genome is able to be updated with new gene families with the addition of new genomes. In the context of plant genomics, pan-genomes are quickly becoming indispensable due to this ability to capture within-species variation (Bayer et al., 2020; Della Coletta et al., 2021). Pan-genomes have been created for multiple crops of interest, including maize (Woodhouse et al., 2021), tomato (Gao et al., 2019), and rice (Zhao et al., 2018; Zhou et al., 2020). For woody plants, pan-genomes have been developed for poplar (B. Zhang et al., 2019) and pecan (Lovell et al., 2021).

Overview of Sequencing Technologies

Since the introduction of Sanger sequencing in 1977, the field of genome sequencing has grown from a combination of higher-throughput and lower costs. This growth was led by next-generation sequencing, short-read technology that sacrificed read-length in favor of lower cost. As next-generation sequencing nears its 20th year, it continues to dominate the sequencing market (Research & Markets ltd, n.d.). However, recent advances in long-read sequencing, including PacBio’s HiFi sequencing technology, offer a promising alternative that overcomes the limitations of short-read sequencing.

Short-Read Sequencing: Illumina

Next-generation sequencing (NGS), also known as second-generation sequencing (SGS), has led to significant and continuous growth in genomic sequencing since its introduction in 2005 (Metzker, 2010). NGS technologies, most notably Illumina sequencing, sacrificed the longer-read sequencing of first-generation technologies in favor of high accuracy, high throughput, and parallel sequencing capabilities at a low cost in a short period of time. Illumina reads may be used for a variety of purposes, from sequencing the entire genome to produce an assembly to sequencing the transcriptome of an organism. The core mechanism of Illumina sequencing is a sequencing by synthesis (SBS) approach, in which nucleotides are called as the DNA is replicated (Maxam & Gilbert, 1977). After fragmentation of the original sequence, short adapter sequences are attached to the fragments so that the fragments are able to attach to the Illumina flow cell. While on the flow cell, the DNA is replicated and broken down into single-stranded DNA for a second round of sequencing. This second round utilizes four fluorescently-tagged terminators to identify each base added to the sequence. A camera captures each fluorescent label, and the computer records them. The terminator is cut from the sequence, and sequencing continues until the next terminator is added. The resulting sequences are saved as “reads” on the computer, and these reads are what are used in genomic analyses. The SBS approach utilizes the natural process of DNA replication to generate highly accurate reads, and it also produces a large number of sequence data, often referred to as “high-throughput.” Additionally, sequences from different samples may be multiplexed, or pooled into a single Illumina plate, to reduce costs.

Despite the high accuracy of individual reads, Illumina faces greater challenges in large-scale applications, specifically in the realm of genome sequencing. Eukaryotic genomes contain a large proportion of repetitive elements, including tandem repeats, transposable elements (TEs) and telomeric/centromeric regions. Plant genomes are rich in repetitive regions, with up to 90% of the genome size considered repetitive (Mehrotra & Goyal, 2014). Illumina reads are short, only spanning 30-400 bp in length, and these highly repetitive regions of the genome result in multiple reads containing near-identical sequences. Most short-read assemblers end up collapsing TE-rich repetitive regions into a “consensus” sequence shorter than the true sequence, which can result in assembly errors (Tørresen et al., 2019). Another possible sequencing error stems from chimeric contigs, in which the assembler incorrectly placed reads flanking repetitive regions (Treangen & Salzberg, 2011). Repetitive regions can also result in unique sequences being left out of a contig that are unable to be integrated into the larger assembly. If short reads are used to assemble a genome, researchers will use certain strategies to reduce the issues caused by repetitive regions. First, chromatin mapping using Hi-C and other chromosome conformation capture (3C) techniques can be used to more accurately place contigs based on their 3d structure, as captured by Hi-C sequencing (Harewood et al., 2017). Second, researchers have generated linkage maps, associated with linkage markers found in the genome, to place contigs based on their alignment to the linkage map (Nürnberg et al., 2021). Finally, recent advances in long-read sequences allow them to be used in tandem with short-read sequencing to create hybrid genome assemblies (Panthee et al., 2022).

Another weakness of short-read sequencing is the ability to capture large-scale structural variants (SVs), or heritable rearrangements in a region of a genome (Mills et al., 2011). SVs range from small-scale insertion/deletion events (InDels) all the way up to large scale inversions,

duplications, and translocation of sequences between chromosomes. In the context of plant genomics, SVs are responsible for important phenotypic changes in crop species (Gabur et al., 2019), including improvements to fruit yield and quality (Butelli et al., 2012) and resistance to pests and diseases (Jobson & Roberts, 2022). Illumina short reads are capable of detecting small-scale InDels, but larger SVs are more difficult to identify because they can span beyond the maximum read-length and may have similar alignment patterns to other SVs (Mahmoud et al., 2019; Sedlazeck et al., 2018). In the realm of genome assemblies, the use of short reads can either cause a true SV to be lost - i.e., a tandem duplication event that is collapsed - or can cause a “false” SV to form due to misassembly of short reads. (Mascher et al., 2021; Vicente-Salvador et al., 2017). As with repetitive regions, SVs can be detected in short-read assemblies using Hi-C, linkage maps, or hybrid assemblies.

Long-Read Sequencing: Nanopore and Pac-Bio HiFi

Considered the “third generation” of sequencing technology (TGS), current long-read sequencing approaches focus on overcoming the limitations of short reads to generate more contiguous genome assemblies and detect repetitive regions (Rice & Green, 2019). One of the first companies to offer long-read sequencing was Oxford Nanopore Technologies (ONT), which achieves real-time, long-read sequencing through a nanopore placed in an electrical resistant membrane (Bowden et al., 2019; Jain et al., 2016). The main ONT sequencer, the MiniION, accepts a DNA read through the nanopore after the addition of adaptor sequences to both ends of the DNA and threads the read through the nanopore. DNA can be single stranded (1D) or double-stranded (2D), with 2D reads having higher accuracy. As the read is threaded through the nanopore, the sequencer measures changes in ionic current, which corresponds to a change in nucleotide sequence. The sequencer interprets these changes as a sequence of 3-6 nucleotide long k-mers of length k , and each k-mer event is aligned and chained together to produce a high-quality read of up to 100 kb in length. ONT reads have been reported to generate reads as long as 2 mb. Compared to the SBS method of Illumina, which must continually terminate sequencing, ONT sequencing is continuous. MiniION sequencing can be used to detect direct base modifications, such as cytosine methylation, without the need for bisulfite sequencing. In addition to providing long reads, the MiniION sequencer is portable and can be used on the field to obtain sequencing data in real time. However, MiniION reads are error-prone, with an error rate of about 14% (Workman et al., 2019), making them difficult to use for genome assemblies on their own. Most genome assemblers that accept Nanopore offer error correction, usually using Illumina short reads to guide correction or by comparing the existing reads to one another (Koren et al., 2017). Additionally, ONT has developed a neural network base calling algorithm called Guppy, a mixed-species basecalling model which is reported to improve read accuracy up to 96% without sacrificing speed (Wick et al., 2019). This varies by organism, and a separate study indicated that, without training for a specific organism of interest, Guppy’s default model may reduce base-calling quality and read accuracy (Ferguson et al., 2022).

Another company that offers long-read sequencing technology is Pacific Biosciences (PacBio). PacBio offers two strategies for obtaining long reads: continuous long read (CLR) sequencing and high fidelity (HiFi) sequencing, also known as circular consensus sequencing (CCS). The older technology, CLR, utilizes single-molecule real-time (SMRT) sequencing to obtain reads without the need to pause between read steps (Schadt et al., 2010). A target

molecule, called a SMRTbell template, is created by ligating hairpin adapters to a target, double-stranded DNA (dsDNA) molecule (Rhoads & Au, 2015). When the SMRTbell is loaded into a SMRT cell, reads are collected through a process similar to SBS. Each nucleotide is marked with a distinct fluorescent label, and the cell collects this information as the SMRTbell is sequenced. Unlike Illumina, which continually pauses and restarts sequencing through the terminator nucleotide, CLR sequencing does not pause after the addition of each nucleotide. The CLR sequencing method achieves longer read lengths, with up to 200 kb reported in CLR (Wenger et al., 2019). Nevertheless, like ONT, PacBio CLR reads are marred by lower base quality, lower throughput, and higher cost when compared to Illumina reads. A common use of PacBio CLR sequencing, like that of ONT, is by pairing this data with Illumina reads to produce hybrid assemblies. This methodology combines the accuracy of short read sequencing with the longer read length of ONT and PacBio CLR, allowing for assemblies that more accurately span repetitive regions than those using only short reads (Panthee et al., 2022). Though the long reads offered by Nanopore and PacBio have strong benefits in sequencing experiments, their low accuracy and high cost represent barriers to widespread adoption for researchers.

In 2019, PacBio introduced HiFi sequencing technology, a method of long-read sequencing that aims to achieve a higher accuracy than the ONT or CLR techniques (Wenger et al., 2019). This technique uses SMRTbell templates to circularize high-weight, double-stranded DNA after annealing primers and DNA polymerase to the sequence. The circularized DNA is then sequenced in multiple passes; this allows the sequencer to make multiple attempts at sequencing a read. After trimming adapters, the subreads are used to call a consensus read sequence - the HiFi read. It is expected that the sequencer will make errors during the sequencing process, and the HiFi circular consensus sequencing (CCS) technique addresses these errors through calling a consensus through the subreads. The more subreads are produced, the more accurate the final read. While the HiFi method sacrifices read length - only up to 24 kb compared to 200 kb for PacBio CLR and 4 Mb for ONT reads - the resulting bases are of higher quality and are more accurate than previous methods. Genomes produced using HiFi reads have high contiguity, even among organisms with complex, polyploid genomes (Hon et al., 2020). Recent comparisons of PacBio HiFi assemblies to those of short read and hybrid assemblies found that assemblies using PacBio HiFi alone were the most accurate and contiguous (Rayamajhi et al., 2022). The main barrier to using PacBio HiFi reads is cost, as they are more expensive than both Illumina and Nanopore reads. However, recent price drops promise to make this option more accessible for labs.

Applications of Sequencing Technologies

With the advancements NGS brought to sequencing technology, researchers have found a variety of applications for their collected reads. These applications depend on if DNA or RNA were collected and sequenced, as these alter the types of questions that can be answered with the sequencing data. DNA sequencing is primarily used in whole genome sequencing and restriction site associated sequencing analyses, which aim to compare genomic content between different samples against the reference genome. RNA sequencing is primarily used in transcriptome sequencing to compare gene expression between different samples by measuring transcript abundance.

Whole Genome Sequencing

Whole genome sequencing (WGS) is a widely-used application of DNA sequencing (DNASeq) technology that aims to collect sequencing data from the entire genome of an organism (Metzker, 2010; Ng & Kirkness, 2010). WGS is a direct product of the innovations driven by NGS, as the first-generation sequencing methods were unfeasible due to their high costs and long runtimes. In WGS, high-quality DNA collected from an organism is fragmented into smaller sequences that are placed into the sequencer. There are two considerations in obtaining this type of data: breadth of coverage and depth of coverage (Sims et al., 2014). The breadth of coverage refers to the percent of the target genome that is sequenced. The depth of coverage refers to the redundancy of coverage - how many times a nucleotide is read during sequencing. A high depth of coverage is important to confirm results and to parse out sequencing errors from true variation, but a high breadth of coverage is also needed to cover as much of the genome as possible. Labs performing WGS experiments must balance these two factors to get accurate results, especially if the lab is looking for variants in the sample against an existing reference genome. Additionally, DNASeq may be run with single-end sequencing - in which only the forward or reverse strands are sequenced - or with paired-end sequencing - in which both strands are sequenced (Risca & Greenleaf, 2015). While single-end sequencing is cheaper, paired-end sequencing produces more informative alignment results.

A common usage of WGS data is for a Variant Calling (VC) pipeline. This analysis identifies single nucleotide polymorphisms (SNPs), or differences in base content compared to the reference genome, in samples. While many SNPs result in synonymous or “silent” mutations that result in the same protein product, there are nonsynonymous SNPs that alter the protein product, ranging from a change in an amino acid altering function to an early termination producing a truncated protein. Both types of SNPs are valuable in plant genomics for different reasons. While synonymous SNPs are less likely to impact gene function, the rate of substitutions per synonymous site (K_s) is used to confirm family-specific whole-genome duplication events (Sollars et al., 2017). SNPs are the focus in trait-association studies, in which researchers work to tie the effects of a SNP to a physical trait of interest. There are two strategies for trait-association: genome-wide association studies (GWAS) and quantitative trait locus (QTL) analyses. QTL analyses require a pedigree to tie the heritability of a trait of interest to a point in the genome (Myles & Wayne, 2008), whereas GWAS analyses rely on linkage disequilibrium (LD) between unrelated members of a species (Uffelmann et al., 2021). These approaches complement one another, with GWAS having a broader range of individuals to work with while QTLs are better able to identify the region of interest. In soybean, a combination of QTL and GWAS analyses identified candidate genes associated with protein and oil content (T. Zhang et al., 2019). Current efforts to track dogwood powdery mildew resistance involves the use of QTLs (Pfarr Moreau et al., 2022).

Another common use for WGS reads is to construct novel reference genomes for organisms without an existing reference, commonly known as *de novo* genome assembly (Baker, 2012). The assembly process requires programs known as assemblers to join individual reads into contigs - a combination of reads that are found to have overlapping sequences (Sohn & Nam, 2016). Contigs are then placed together into scaffolds - which are commonly filled with uncalled gap regions joining individual contigs - which may be further placed together into chromosome-scale super-scaffolds. The goal of an assembly is to achieve as high a contiguity as

possible. The more contiguous the genome assembly is, the more accurate the genome. Contiguity in genomes is tested by searching for key single copy genes - also referred to as Benchmarking Universal Single-Copy Orthologs (BUSCOs) (Seppey et al., 2019) - and by determining the N50 of the assembly - the length of the shortest contig that covers 50% of the genome. As discussed earlier, short reads have greater difficulty in providing a contiguous genome due to complex repetitive regions and structural variants (Tørresen et al., 2019). In both short-read and long read assemblies, scaffolding can be improved using Hi-C contact maps (Harewood et al., 2017) and linkage maps (Nürnberg et al., 2021) to more accurately order and orient contigs. Long-read sequencing often has higher contiguity before being placed into scaffold-level assemblies, and Hi-C can recover full chromosome scale assemblies. For trees that are in significant danger of becoming extinct, Nanopore offers free sequencing for any critically endangered species through a pilot stage project (<https://ORG.one>).

The main drawback to using WGS is the cost of sequencing. Despite substantial price drops in Illumina sequencing in the years following its popularization (Kris A. Wetterstrand, 2019), the costs of sequences are still substantial for tree genomics. Given that many tree species have large genomes, even Illumina WGS experiments can be a costly expenditure for a lab. These concerns drive the need to balance between breadth of coverage and depth of coverage. In a perfect experiment, both will be prioritized, but that can lead to increased costs, so compromises must be made to keep costs low. These cost concerns especially apply to a long-read approach, and before the recent attempts to improve accuracy, there were concerns that the money used would not provide accurate results. While WGS is a strong tool for analyzing DNA sequencing, the focus on the entire genome can lead to a more expensive experiment.

Reduced-Representation Sequencing

There is an alternative DNaseq approach that attempts to minimize costs. In contrast to WGS, reduced-representation sequencing (RRS) does not aim to sequence the entirety of a genome, but instead focus coverage on select regions (Davey & Blaxter, 2010). The breadth of sequencing coverage is sacrificed so that these select regions can have a greater depth of coverage. These regions are commonly selected using restriction enzymes (REs) - enzymes that cut at a predetermined sequence in the genome - after the DNA isolation step. While having a genome assembly available for an organism of interest is ideal, RADSeq sequencing can be used in GWAS and QTL analyses without the need for a complete genome assembly. While all RRS technologies rely on fragmented DNA collected from RE digested DNA, the exact protocols vary. This is the difference between reduced-association DNA sequencing (RADSeq) and genotyping-by-sequencing (GBS) RRS approaches (Campbell et al., 2018). Additionally, RRS can be run with single-digest (one RE) or double-digest (two REs) approaches.

The cheaper sequencing costs of RRS compared to WGS do not come without drawbacks. While a lower breadth of coverage is expected in a RRS experiment, RRS technologies can only capture around 5% of the total genome content of an organism (Hoban et al., 2016; Lowry et al., 2017). In the context of identifying novel SNPs associated with a trait of interest, there is a low likelihood that the RRS approach will capture the true variant of interest. Further, RRS suffers from a high amount of missing data in each individual sample, due to uneven coverage. Some RADSeq pipelines use a process known as imputation to overcome missing genotypes by using linkage disequilibrium (LD) to fill in missing genotype data (Swarts

et al., 2014). This usually relies on having a chromosome-scale assembly of the species of interest, though recent imputation strategies have proven to be effective without a genome of this scale (Munyengwa et al., 2021). Imputation is also limited by the sampled cohort's overall LD. Furthermore, RRS is only as informative as the reference genome, and if the region containing the true SNP of interest is missing from the assembly, then the imputation strategy will be unable to pick up the SNP in the reference (Hoban et al., 2016). However, programs like STACKS offer *de novo* assembly of RRS data, allowing researchers to overcome issues stemming from the reference itself (Heller et al., 2021). Overall, while RRS represents a cheaper option to a WGS approach, the drawbacks of decreased coverage and missing data require careful consideration and often additional analysis steps.

RRS approaches have been used in the context of plant genomics to answer similar questions to WGS sequencing at a lower cost. RRS has been used in QTL mapping and GWAS analyses across a broad range of plant species, including apple (Lee et al., 2017), citrus (Imai et al., 2018), and sunflowers (Ma et al., 2018). In *Olea europaea*, a GBS approach was found to be just as effective in comparing genetic diversity between populations as a WGS approach (Friel et al., 2021). RADSeq approaches, particularly GBS with imputation, were also confirmed to be an effective SNP discovery approach in maize (Wang et al., 2020). In bread wheat, the imputation strategy has been shown to be effective in accurately filling in missing genotypes, improving the utility of a GBS approach (Alipour et al., 2019). While imputing generally relies on a chromosome-scale assembly, a GBS experiment in rubber trees has shown that imputation is possible without a full genome (Munyengwa et al., 2021).

RNA Sequencing

DNA sequencing is useful in identifying SNPs associated with a known phenotype, but it will not always answer how this SNP results in an altered phenotype. Furthermore, gene expression can be altered independently of SNPs in the reference genome, and DNaseq is ill-equipped for this type of analysis. Introduced in 2007, RNA sequencing (RNASeq) is a complementary sequencing technology that captures the range of RNA present at the time of collection (the transcriptome) (Emrich et al., 2007; Lister et al., 2008). RNASeq is primarily used in differential gene expression (DGE) experiments to compare levels of messenger RNA (mRNA) between experimental groups, such as different times of collection and different treatments of samples (Stark et al., 2019). Like RADseq, RNASeq is a cheaper alternative to WGS, as only a small percentage of the genome is transcribed into mRNA. The RNA of an organism is isolated, converted to more stable clone DNA (cDNA), and sequenced. In most plant RNASeq applications, only the mRNA is selected using an oligo(dT) adapter that anneals to the poly(A) tail prior to reverse transcription. While SNPs can be identified using mRNA data, the main point of interest in an RNASeq experiment is measuring transcript abundance and identifying significantly different transcript levels between experimental groups.

Experimental design in an RNASeq analysis is of the utmost importance, as all sources of gene expression variation must be considered to obtain accurate results. A sufficient number of replicates, or samples with the same experimental factors, is needed to tease out natural biological variation and random variation caused by the protocol itself (Lamarre et al., 2018). Two types of replicates are needed to confirm this information: biological replicates, or biologically distinct samples undergoing the same treatment, and technical replicates, or repeated

measurements of a single sample. Replicate number has been shown to have a larger impact on the DE analysis than library size, though many publications use a low number of biological replicates (Lamarre et al., 2018). A 48-replicate study in yeast determined that a minimum of six replicates is ideal for DGE analysis, though most studies limit themselves to three or four replicates (Schurch et al., 2016). This is because more biological replicates mean more samples to sequence, which results in increased sequencing costs. Additional considerations for RNASeq match those of DNASeq, including depth of coverage and use of single- or paired-end reads.

RNASeq is most commonly performed using Illumina short read sequencing. This strategy measures changes in transcript abundance by aligning short reads to the genome and counting the number of reads associated with genes in its annotation. Studies across multiple plant species, including blueberry (Yu et al., 2021), grapevine (Rapicavoli et al., 2018), and poplar (Leite Montalvão et al., 2022), have used this RNASeq strategy to answer questions about gene function and the effect of pathogens on plants. However, short reads are not as well-equipped to measure differential alternative splicing, or changes in isoform abundance in response to a treatment, because they are heavily fragmented. The introduction of long read sequencing technologies has expanded the possibilities for RNASeq experiments. First, long-reads can capture transcript-level reads instead of fragments, allowing for a direct comparison between isoforms of the same genes (Rhoads & Au, 2015; Soneson et al., n.d.). Additionally, direct RNASeq (dRNASeq) applications, which directly measure RNA abundance without a cDNA intermediate, offer reads representing full-length mRNAs. While these advancements allow for direct comparison of isoforms, the higher throughput and lower cost of short-read sequencing allows the older method to still be in use. This is coupled with an updated variation of the classic differential expression pipeline where, instead of mapping genes back to the genome, programs compare transcripts within to quantify transcript abundance (Bray et al., 2016; Patro et al., 2017).

REFERENCES

- 2019 census of horticultural specialties. (n.d.). Retrieved March 10, 2023, from https://www.nass.usda.gov/Publications/AgCensus/2017/Online_Resources/Census_of_Horticulture_Specialties/index.php
- Alipour, H., Bai, G., Zhang, G., Bihamta, M. R., Mohammadi, V., & Peyghambari, S. A. (2019). Imputation accuracy of wheat genotyping-by-sequencing (GBS) data using barley and wheat genome references. In *PLOS ONE* (Vol. 14, Issue 1, p. e0208614). <https://doi.org/10.1371/journal.pone.0208614>
- Anagnostakis, S. L. (2012). Chestnut Breeding in the United States for Disease and Insect Resistance. *Plant Disease*, 96(10), 1392–1403.
- Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, 9(4), 333–337.
- Bartholomé, J., Brachi, B., Marçais, B., Mougou-Hamdane, A., Bodénès, C., Plomion, C., Robin, C., & Desprez-Loustau, M.-L. (2020). The genetics of exapted resistance to two exotic pathogens in pedunculate oak. *The New Phytologist*, 226(4), 1088–1103.
- Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6(8), 914–920.
- Bolgiano, C., & Novak, G. (2007). *Mighty Giants: An American Chestnut Anthology*. American Chestnut Foundation.
- Bowden, R., Davies, R. W., Heger, A., Pagnamenta, A. T., de Cesare, M., Oikkonen, L. E., Parkes, D., Freeman, C., Dhalla, F., Patel, S. Y., Popitsch, N., Ip, C. L. C., Roberts, H. E., Salatino, S., Lockstone, H., Lunter, G., Taylor, J. C., Buck, D., Simpson, M. A., & Donnelly, P. (2019). Sequencing of human genomes with nanopore technology. *Nature Communications*, 10(1), 1869.
- Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525–527.
- Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. In *Nature Reviews Genetics* (Vol. 12, Issue 10, pp. 703–714). <https://doi.org/10.1038/nrg3054>
- Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., & Martin, C. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *The Plant Cell*, 24(3), 1242–1255.
- Campbell, E. O., Brunet, B. M. T., Dupuis, J. R., & Sperling, F. A. H. (2018). Would an RRS by any other name sound as RAD ? *Methods in Ecology and Evolution / British Ecological Society*, 9(9), 1920–1927.
- Cappai, F., Amadeu, R. R., Benevenuto, J., Cullen, R., Garcia, A., Grossman, A., Ferrão, L. F. V., & Munoz, P. (2020). High-Resolution Linkage Map and QTL Analyses of Fruit Firmness in Autotetraploid Blueberry. *Frontiers in Plant Science*, 11, 562171.
- Crow, K. D. (2006). What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? In *Molecular Biology and Evolution* (Vol. 23, Issue 5, pp. 887–892). <https://doi.org/10.1093/molbev/msj083>
- Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. In *Briefings in Functional Genomics* (Vol. 9, Issues 5-6, pp. 416–423). <https://doi.org/10.1093/bfgp/elq031>

- Del Giacco, L., & Cattaneo, C. (2012). Introduction to genomics. *Methods in Molecular Biology*, 823, 79–88.
- Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biology*, 22(1), 3.
- Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., Smith, R. D., Teresi, S. J., Nelson, A. D. L., Wai, C. M., Alger, E. I., Bird, K. A., Yocca, A. E., Pumplin, N., Ou, S., Ben-Zvi, G., Brodt, A., Baruch, K., Swale, T., ... Knapp, S. J. (2019). Origin and evolution of the octoploid strawberry genome. *Nature Genetics*, 51(3), 541–547.
- Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, 17(1), 69–73.
- Ferguson, S., McLay, T., Andrew, R. L., Bruhl, J. J., Schwessinger, B., Borevitz, J., & Jones, A. (2022). Species-specific basecallers improve actual accuracy of nanopore sequencing in plants. *Plant Methods*, 18(1), 137.
- 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(11). <https://doi.org/10.3390/plants10112514>
- Gabur, I., Chawla, H. S., & Snowdon, R. J. (2019). Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied*. <https://link.springer.com/article/10.1007/s00122-018-3233-0>
- Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., ... Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, 51(6), 1044–1051.
- Hardwood Federation - Hardwood Industry Economic Impact. (n.d.). Retrieved March 14, 2023, from <http://www.hardwoodfederation.com/Hardwood-Industry-Economic-Impact>
- Harewood, L., Kishore, K., Eldridge, M. D., Wingett, S., Pearson, D., Schoenfelder, S., Collins, V. P., & Fraser, P. (2017). Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biology*, 18(1), 125.
- Hausman, C. E., Bertke, M. M., Jaeger, J. F., & Rocha, O. J. (2014). Genetic structure of green ash (*Fraxinus pennsylvanica*): implications for the establishment of ex situ conservation protocols in light of the invasion of the emerald ash borer. *Plant Genetic Resources: Characterization and Utilization*, 12(3), 286–297.
- Hebard, F. V. (1994). The American Chestnut Foundation breeding plan: beginning and intermediate steps. *J. American Chestnut*. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.172.9371&rep=rep1&type=pdf#page=16>
- 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. *Molecular Ecology Resources*, 21(4), 1085–1097.
- 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. *The American Naturalist*, 188(4), 379–397.
- Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, 7(1), 399.
- Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., ... Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, 373(6555), 655–662.
- Imai, A., Nonaka, K., Kuniga, T., Yoshioka, T., & Hayashi, T. (2018). Genome-wide association mapping of fruit-quality traits using genotyping-by-sequencing approach in citrus landraces, modern cultivars, and breeding lines in Japan. In *Tree Genetics & Genomes* (Vol. 14, Issue 2). <https://doi.org/10.1007/s11295-018-1238-0>
- Initiative, T. A. G., & The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. In *Nature* (Vol. 408, Issue 6814, pp. 796–815). <https://doi.org/10.1038/35048692>
- Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1), 256.
- Jobson, E., & Roberts, R. (2022). Genomic structural variation in tomato and its role in plant immunity. *Molecular Horticulture*, 2(1), 1–10.
- 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 Research*, 27(5), 722–736.
- Kostick, S. A., Teh, S. L., Norelli, J. L., Vanderzande, S., Peace, C., & Evans, K. M. (2021). Fire blight QTL analysis in a multi-family apple population identifies a reduced-susceptibility allele in “Honeycrisp.” *Horticulture Research*, 8(1), 28.
- Kris A. Wetterstrand, M. S. (2019, March 13). *DNA sequencing costs: Data*. Genome.gov; NHGRI. <https://www.genome.gov/sequencingcostsdata/>
- Krivanek, A. F., Riaz, S., & Walker, M. A. (2006). Identification and molecular mapping of *PdR1*, a primary resistance gene to Pierce’s disease in *Vitis*. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 112(6), 1125–1131.
- Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., & Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, 9, 1660.
- Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M., & Maza, E. (2018). Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number

- and Library Size. *Frontiers in Plant Science*, 9, 108.
- Lee, S. J., Ban, S. H., Kim, G. H., Kwon, S. I., Kim, J. H., & Choi, C. (2017). Identification of potential gene-associated major traits using GBS-GWAS for Korean apple germplasm collections. *Plant Breeding = Zeitschrift Fur Pflanzenzuchtung*, 136(6), 977–986.
- Leite Montalvão, A. P., Kersten, B., Kim, G., Fladung, M., & Müller, N. A. (2022). ARR17 controls dioecy in *Populus* by repressing B-class MADS-box gene expression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 377(1850), 20210217.
- Licciardello, C., Perrone, I., Gambino, G., Talon, M., & Velasco, R. (2021). Editorial: Functional Genomics in Fruit Trees: From 'Omics to Sustainable Biotechnologies. *Frontiers in Plant Science*, 12, 729714.
- Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 133(3), 523–536.
- Lovell, J. T., Bentley, N. B., Bhattarai, G., Jenkins, J. W., Sreedasyam, A., Alarcon, Y., Bock, C., Boston, L. B., Carlson, J., Cervantes, K., Clermont, K., Duke, S., Krom, N., Kubenka, K., Mamidi, S., Mattison, C. P., Monteros, M. J., Pisani, C., Plott, C., ... Randall, J. J. (2021). Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. *Nature Communications*, 12(1), 4125.
- 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. *Molecular Ecology Resources*, 17(2), 142–152.
- Ma, G. J., Song, Q. J., Markell, S. G., & Qi, L. L. (2018). High-throughput genotyping-by-sequencing facilitates molecular tagging of a novel rust resistance gene, R15, in sunflower (*Helianthus annuus* L.). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 131(7), 1423–1432.
- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, 20(1), 246.
- Marks, R. A., Hotaling, S., Frandsen, P. B., & VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nature Plants*, 7(12), 1571–1578.
- Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C. S., Ens, J., Gundlach, H., Boston, L. B., Tulpová, Z., Holden, S., Hernández-Pinzón, I., Scholz, U., Mayer, K. F. X., Spannagl, M., Pozniak, C. J., Sharpe, A. G., Šimková, H., Moscou, M. J., ... Stein, N. (2021). Long-read sequence assembly: a technical evaluation in barley. *The Plant Cell*, 33(6), 1888–1906.
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), 560–564.
- Mehrotra, S., & Goyal, V. (2014). Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics, Proteomics & Bioinformatics*, 12(4), 164–171.

- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, 11(1), 31–46.
- Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., ... Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332), 59–65.
- Mmbaga, M. T., Mrema, F. A., Mackasmiel, L., & Rotich, E. (2016). Effect of bacteria isolates in powdery mildew control in flowering dogwoods (*Cornus florida* L.). In *Crop Protection* (Vol. 89, pp. 51–57). <https://doi.org/10.1016/j.cropro.2016.06.011>
- Morse, A. M., Peterson, D. G., Islam-Faridi, M. N., Smith, K. E., Magbanua, Z., Garcia, S. A., Kubisiak, T. L., Amerson, H. V., Carlson, J. E., Nelson, C. D., & Davis, J. M. (2009). Evolution of genome size and complexity in *Pinus*. *PloS One*, 4(2), e4332.
- Munyengwa, N., Le Guen, V., Bille, H. N., Souza, L. M., Clément-Demange, A., Mournet, P., Masson, A., Soumahoro, M., Kouassi, D., & Cros, D. (2021). Optimizing imputation of marker data from genotyping-by-sequencing (GBS) for genomic selection in non-model species: Rubber tree (*Hevea brasiliensis*) as a case study. In *Genomics* (Vol. 113, Issue 2, pp. 655–668). <https://doi.org/10.1016/j.ygeno.2021.01.012>
- Myles, C., & Wayne, M. (2008). Quantitative trait locus (QTL) analysis. *Nature Education* 1 (1), 208. https://www.researchgate.net/profile/Seied-Mehdi-Miri/post/How_can_I_conduct_a_substitution_mapping/attachment/5b7c68b93843b0e6f2ab1896/AS%3A662148569112588%401534879929564/download/B+90.pdf
- Nagamitsu, T., Uchiyama, K., Izuno, A., Shimizu, H., & Nakanishi, A. (2020). Environment-dependent introgression from *Quercus dentata* to a coastal ecotype of *Quercus mongolica* var. *crispula* in northern Japan. *The New Phytologist*, 226(4), 1018–1028.
- Neale, D. B., & Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nature Reviews. Genetics*, 12(2), 111–122.
- Newcombe, G. (1998). A review of exapted resistance to diseases of *Populus*. *Forest Pathology*, 28(4), 209–216.
- Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. *Methods in Molecular Biology*, 628, 215–226.
- Nürnberg, B., Baird, S. J. E., Čížková, D., Bryjová, A., Mudd, A. B., Blaxter, M. L., & Szymura, J. M. (2021). A dense linkage map for a large repetitive genome: discovery of the sex-determining region in hybridizing fire-bellied toads (*Bombina bombina* and *Bombina variegata*). In *G3 Genes|Genomes|Genetics* (Vol. 11, Issue 12). <https://doi.org/10.1093/g3journal/jkab286>
- Panthee, S., Hamamoto, H., Paudel, A., Kaito, C., Suzuki, Y., & Sekimizu, K. (2022). Hybrid assembly using long reads resolves repeats and completes the genome sequence of a laboratory strain of *Staphylococcus aureus* subsp. *aureus* RN4220. In *Heliyon* (Vol. 8, Issue 11, p. e11376). <https://doi.org/10.1016/j.heliyon.2022.e11376>
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4),

- 417–419.
- Pfarr Moreau, E., Honig, J. A., & Molnar, T. J. (2022). High-Density Linkage Mapping and Identification of Quantitative Trait Loci Associated with Powdery Mildew Resistance in Flowering Dogwood (*Cornus florida*). *Horticulturae*, 8(5), 405.
- Plomion, C., Aury, J.-M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., Faye, S., Francillon, N., Labadie, K., Le Provost, G., Lesur, I., Bartholomé, J., Faivre-Rampant, P., Kohler, A., Leplé, J.-C., Chantret, N., Chen, J., Diévert, A., Alaeitabar, T., ... Salse, J. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, 4(7), 440–452.
- Plomion, C., Bastien, C., Bogeat-Triboulot, M.-B., Bouffier, L., Déjardin, A., Duplessis, S., Fady, B., Heuertz, M., Le Gac, A.-L., Le Provost, G., Legué, V., Lelu-Walter, M.-A., Leplé, J.-C., Maury, S., Morel, A., Oddou-Muratorio, S., Pilate, G., Sanchez, L., Scotti, I., ... Vacher, C. (2016). Forest tree genomics: 10 achievements from the past 10 years and future prospects. *Annals of Forest Science*, 73(1), 77–103.
- Powell, W. A., Newhouse, A. E., & Coffey, V. (2019). Developing Blight-Tolerant American Chestnut Trees. In *Cold Spring Harbor Perspectives in Biology* (Vol. 11, Issue 7, p. a034587). <https://doi.org/10.1101/cshperspect.a034587>
- Rapicavoli, J. N., Blanco-Ulate, B., Muszyński, A., Figueroa-Balderas, R., Morales-Cruz, A., Azadi, P., Dobruchowska, J. M., Castro, C., Cantu, D., & Roper, M. C. (2018). Lipopolysaccharide O-antigen delays plant innate immune recognition of *Xylella fastidiosa*. *Nature Communications*, 9(1), 390.
- Rayamajhi, N., Cheng, C.-H. C., & Catchen, J. M. (2022). Evaluating Illumina-, Nanopore-, and PacBio-based genome assembly strategies with the bald notothen, *Trematomus borchgrevinki*. *G3*, 12(11). <https://doi.org/10.1093/g3journal/jkac192>
- Research, & Markets ltd. (n.d.). *Next-generation Sequencing Data Analysis Market Size, Share & Trends Analysis Report by Product (NGS Commercial Software), by Workflow, by Mode, by End-use, by Read Length, by Region, and Segment Forecasts, 2022-2030*. Retrieved March 16, 2023, from https://www.researchandmarkets.com/reports/5415633/next-generation-sequencing-data-analysis-market?utm_source=GNOM&utm_medium=PressRelease&utm_code=kxkfhf&utm_campaign=1784471+-+Global+Next-Generation+Sequencing+Data+Analysis+Market+Report+2022%3a+Rising+Academic+Research+Based+On+NGS+Protocols+Boosts+Growth&utm_exec=como322prd
- Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5), 278–289.
- Rice, E. S., & Green, R. E. (2019). New Approaches for Genome Assembly and Scaffolding. *Annual Review of Animal Biosciences*, 7, 17–40.
- Risca, V. I., & Greenleaf, W. J. (2015). Beyond the Linear Genome: Paired-End Sequencing as a Biophysical Tool. *Trends in Cell Biology*, 25(12), 716–719.
- Sandercock, A. M., Westbrook, J. W., Zhang, Q., Johnson, H. A., Saielli, T. M., Scrivani, J. A., Fitzsimmons, S. F., Collins, K., Perkins, M. T., Craddock, J. H., Schmutz, J., Grimwood, J., & Holliday, J. A. (2022). Frozen in time: Rangewide genomic

- diversity, structure, and demographic history of relict American chestnut populations. *Molecular Ecology*, 31(18), 4640–4655.
- Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2), R227–R240.
- Schopmeyer, C. S., & Others. (1974). Seeds of woody plants in the United States. *Seeds of Woody Plants in the United States.*, 450.
<https://www.cabdirect.org/cabdirect/abstract/19750622387>
- Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? In *RNA* (Vol. 22, Issue 6, pp. 839–851).
<https://doi.org/10.1261/rna.053959.115>
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, 15(6), 461–468.
- Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology*, 1962, 227–245.
- Sicard, A., Zeilinger, A. R., Vanhove, M., Scharrel, T. E., Beal, D. J., Daugherty, M. P., & Almeida, R. P. P. (2018). *Xylella fastidiosa*: Insights into an Emerging Plant Pathogen. *Annual Review of Phytopathology*, 56, 181–202.
- Sims, D., Sudbery, I., Illott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics*, 15(2), 121–132.
- Sohn, J.-I., & Nam, J.-W. (2016). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, 19(1), 23–40.
- Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., Kaithakottil, G., Cooper, E. D., Uauy, C., Havlickova, L., Worswick, G., Studholme, D. J., Zohren, J., Salmon, D. L., Clavijo, B. J., Li, Y., He, Z., Fellgett, A., McKinney, L. V., ... Buggs, R. J. A. (2017). Genome sequence and genetic diversity of European ash trees. *Nature*, 541(7636), 212–216.
- Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M. D., & Hussain, S. (n.d.). *A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes*.
<https://doi.org/10.1101/574525>
- Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews. Genetics*, 20(11), 631–656.
- Staton, M., Addo-Quaye, C., Cannon, N., Yu, J., Zhebentyayeva, T., Huff, M., Islam-Faridi, N., Fan, S., Georgi, L. L., Dana Nelson, C., Bellis, E., Fitzsimmons, S., Henry, N., Drautz-Moses, D., Noorai, R. E., Ficklin, S., Saski, C., Mandal, M., Wagner, T. K., ... Carlson, J. E. (2020). “A reference genome assembly and adaptive trait analysis of *Castanea mollissima* ‘Vanuxem,’ a source of resistance to chestnut blight in restoration breeding.” In *Tree Genetics & Genomes* (Vol. 16, Issue 4).
<https://doi.org/10.1007/s11295-020-01454-y>
- Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R.,

- Gonzalez-Ibeas, D., Koriabine, M., Holtz-Morris, A. E., Martínez-García, P. J., Sezen, U. U., Marçais, G., Jermstad, K., McGuire, P. E., Loopstra, C. A., Davis, J. M., Eckert, A., de Jong, P., ... Langley, C. H. (2016). Sequence of the Sugar Pine Megagenome. *Genetics*, 204(4), 1613–1626.
- Swarts, K., Li, H., Alberto Romero Navarro, J., An, D., Romay, M. C., Hearne, S., Acharya, C., Glaubitz, J. C., Mitchell, S., Elshire, R. J., Buckler, E. S., & Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. In *The Plant Genome* (Vol. 7, Issue 3). <https://doi.org/10.3835/plantgenome2014.05.0023>
- Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A. V., Promponas, V. J., Anisimova, M., Jakobsen, K. S., & Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 47(21), 10994–11006.
- Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1), 36–46.
- Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhallerao, R. R., Bhallerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., ... Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 313(5793), 1596–1604.
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., & Lappalainen, T. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 1–21.
- Vicente-Salvador, D., Puig, M., Gayà-Vidal, M., Pacheco, S., Giner-Delgado, C., Noguera, I., Izquierdo, D., Martínez-Fundichely, A., Ruiz-Herrera, A., Estivill, X., Aguado, C., Lucas-Lledó, J. I., & Cáceres, M. (2017). Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Human Molecular Genetics*, 26(3), 567–581.
- Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., Gowda, M., Nair, S. K., Hao, Z., Lu, Y., San Vicente, F., Prasanna, B. M., Li, X., & Zhang, X. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Scientific Reports*, 10(1), 16308.
- Watanabe, M., Suwabe, K., & Suzuki, G. (2012). Molecular genetics, physiology and biology of self-incompatibility in Brassicaceae. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, 88(10), 519–535.
- Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Functammasan, 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., ... Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10), 1155–1162.
- Westbrook, J. W., Holliday, J. A., Newhouse, A. E., & Powell, W. A. (2020). A plan to diversify a transgenic blight-tolerant American chestnut population using citizen

- science. In *PLANTS, PEOPLE, PLANET* (Vol. 2, Issue 1, pp. 84–95).
<https://doi.org/10.1002/ppp3.10061>
- Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. In *Genome Biology* (Vol. 20, Issue 1).
<https://doi.org/10.1186/s13059-019-1727-y>
- Woodhouse, M. R., Cannon, E. K., Portwood, J. L., 2nd, Harper, L. C., Gardiner, J. M., Schaeffer, M. L., & Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, 21(1), 385.
- Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Soulette, C. M., Snutch, T. P., Loman, N., Paten, B., Loose, M., ... Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods*, 16(12), 1297–1305.
- Yu, J., Hulse-Kemp, A. M., Babiker, E., & Staton, M. (2021). High-quality reference genome and annotation aids understanding of berry development for evergreen blueberry (*Vaccinium darrowii*). *Horticulture Research*, 8(1), 228.
- Zhang, B., Zhu, W., Diao, S., Wu, X., Lu, J., Ding, C., & Su, X. (2019). The poplar pangenome provides insights into the evolutionary history of the genus. *Communications Biology*, 2, 215.
- Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W., Tao, Y., Wang, J., Yuan, Z., Fan, G., Xing, Z., Han, C., Pan, H., Zhong, X., Shi, W., Liang, X., Du, D., Sun, F., Xu, Z., ... Wang, J. (2012). The genome of *Prunus mume*. *Nature Communications*, 3, 1318.
- Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., Hou, W., Wu, C., Han, T., & Sun, S. (2019). A Combined Linkage and GWAS Analysis Identifies QTLs Linked to Soybean Seed Protein and Oil Content. *International Journal of Molecular Sciences*, 20(23). <https://doi.org/10.3390/ijms20235915>
- Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., ... Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, 50(2), 278–284.
- Zheng, T., Li, P., Li, L., & Zhang, Q. (2021). Research advances in and prospects of ornamental plant genomics. In *Horticulture Research* (Vol. 8, Issue 1).
<https://doi.org/10.1038/s41438-021-00499-x>
- Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., Parakkal, P., Arbelaez, L. J., Franco, N., Alexandrov, N., Hamilton, N. R. S., Leung, H., Mauleon, R., Lorieux, M., Zuccolo, A., McNally, K., ... Wing, R. A. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. *Scientific Data*, 7(1), 113.

CHAPTER II: INTRODUCTION TO GREEN ASH (*FRAXINUS PENNSYLVANICA*).

Fraxinus pennsylvanica (green ash)

Distribution, usage, and importance

Fraxinus pennsylvanica Marsh., commonly known as green ash (also as red ash, swamp ash, or water ash), is a deciduous tree species within the Ash (*Fraxinus*) genus native to the eastern North American coast. It is one of over forty *Fraxinus* species, distributed throughout the temperate forests of North America, Europe, and Asia. Green ash is the most widely distributed *Fraxinus* species found in North America (Kennedy, 1949). It is part of the Oleaceae family, also known as the Olive family, alongside the economically important olive trees (*Olea* sp.). Species in the genus are characterized by “winged” seeds - also called samara - and leaves arranged opposite of one another on the stem. In green ash, the samara are pointed (Figure 1). Green ash is dioecious, meaning that sex is distinct between male and female individuals (Figure 1). Male flowers consist solely of stamen structures, and female flowers consist solely of carpel structures. In the wild, green ash is found primarily near streams, rivers, and bottomlands, though it is adaptable to other environments as well.

Green ash is a species of economic importance in the United States. Its adaptability applies to urban environments, allowing it to function as an ornamental tree. It became a popular ornamental following the large-scale loss of American elm trees due to Dutch Elm Disease (Burns & Honkala, 1990), though this practice arguably exacerbated the later damage due to establishing monocultures. Outside of ornamental use, green ash wood is commonly used to make a number of products. Its lumber is popular for woodworking, and has been used to create wooden baseball bats, pool cues, and tool handles, among a variety of other uses. The guitar company Fender traditionally used green ash in building the bodies of its electric guitars, including its Stratocaster and Telecaster models. Comparison of solid-body guitars made from ash wood to those of other trees found that ash wood improved their acoustic properties by reducing vibration damping in the neck of the guitar (Ray et al., 2021). Due to its abundance in North America, its popularity as an ornamental, and the products made from its lumber, green ash represents an important species of hardwood tree in North America’s economy, having contributed close to a billion dollars (Kovacs et al., 2010).

Outside of its economic value, green ash is also an ecologically important species. Its leaves serve as food for a variety of insects, and its seeds serve as food for birds and small mammals (Hilty, n.d.). Due to the low tannin content of their fruit, they serve as an excellent food source for tadpoles (Stephens et al., 2013). Tadpoles grown in habitats containing ash trees were found to be healthier than those in habitats primarily containing tannin-rich trees. Foliage and twigs of green ash are palatable to herbivorous mammals, including hoofed mammals such as deer and horses, as well as small mammals such as rabbits (Hilty, n.d.). Ash trees are used in soil conservation and have widely been used in revegetation efforts in the aftermath of strip mining (Wright, n.d.). In general, green ash has been a useful species for ecosystem restoration efforts in riparian areas. In recent years, however, the introduction of an invasive insect species has caused severe population declines in both natural and urban environments.

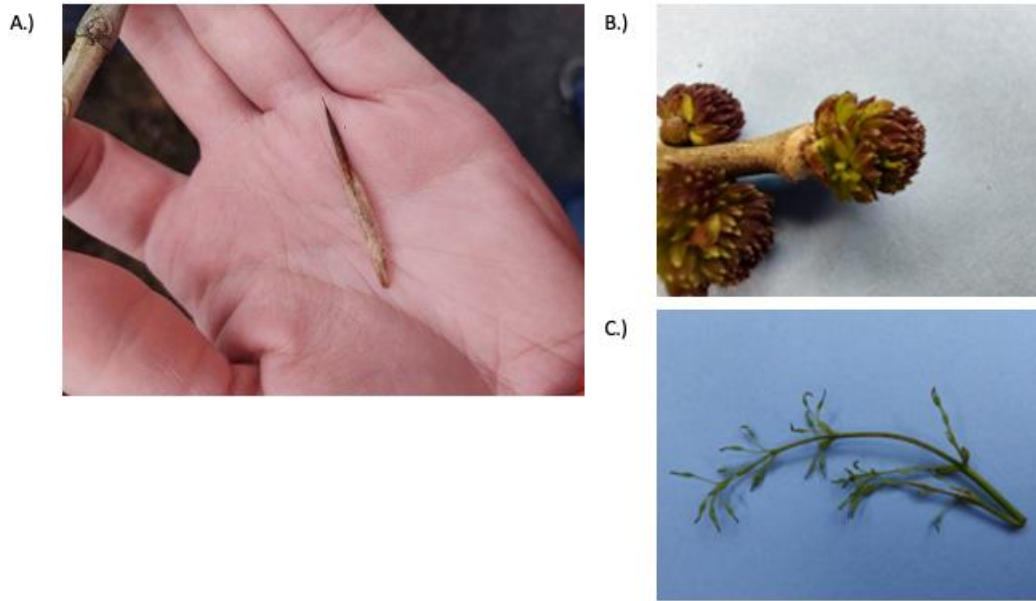


Figure 1. Phenotypic characteristics of green ash (*Fraxinus pennsylvanica*). (A). Samara of green ash. (B). Flowers of male green ash. (C.) Flowers of female green ash. Samara and flower samples were all collected from Turkey Creek, Knoxville, TN, USA.

Emerald ash borer and other threats

The emerald ash borer (EAB; *Agrilus planipennis*) is an invasive species of jewel beetle that feeds on and targets the bark of ash trees, and it is a critical threat to green ash and all of the other native ash populations of North America. Native to northeast Asia, the EAB is believed to have arrived in America through wood packing materials. While EAB was first confirmed in the United States in 2002, evidence suggests the invasive population can be traced back to southeast Michigan in the early 1990s (Siegert et al., 2014). Females lay their eggs on the bark of the tree, and after hatching, the larvae feed on phloem and vascular cambium. This feeding pattern results in characteristic galleries that damage the wood (Figure 2), which disrupts the transport of sugars and water from the tree's roots to its canopy. Symptoms of EAB infestation include crown dieback and characteristic, D-shaped exit holes. Within 6 years after introduction, the presence of EAB can reduce a healthy population of North American ash trees to near complete mortality (Knight et al., 2013).

While some ash trees native to Asia have resistance to EAB, species outside of this range - including those of North America and Europe - are highly susceptible to the beetle's effects (Rebek et al., 2008). Projected damages caused by the pest have been estimated to be as much as \$10.7 billion, including the replacement of 17 million ash trees (Kovacs et al., 2010). Management of the pest has proven difficult; the EAB flies from tree to tree, complicating quarantine efforts. To minimize the spread of EAB from state to state, the USDA recommends that no firewood should be transported between states. Most treatment options for ash trees are preventative in nature, focusing on preserving healthy trees instead of trying to rescue trees that have already been infested. The loss of green ash has resulted in a decline in use of its lumber. Fender no longer uses green ash wood for electric guitar bodies due to the losses caused by the EAB. Ecologically, the shift in forest populations from trees lacking tannin to trees with high tannin levels - representing the replacement of susceptible ash species with other species - is associated with reduced fitness in tadpoles (Stephens et al., 2013).

Outside of the United States, the ash trees of Europe face an additional threat in the form of an Ascomycete fungus, *Hymenoscyphus fraxineus*. *H. fraxineus* causes a fatal disease in ash known as ash dieback, named for the crown dieback resulting from the disease (Bakys et al., 2009). Symptoms of ash dieback include necrotic spots on stems that enlarge into cankers and wilting of leaves. It has primarily impacted growth of European ash (*F. excelsior*, section *Fraxinus*) populations, but other species have been affected as well. European ash trees infected with *H. fraxineus* have an estimated mortality rate of 85% in plantations and 69% in woodland populations (Coker et al., 2019). While it has not been detected in the United States at this time, experiments indicate that American species, including green ash, are mildly susceptible to *H. fraxineus* (Nielsen et al., 2017). The introduction of another fatal invasive to its native range would likely exacerbate the current issues facing green ash.

EAB resistance in green ash and breeding efforts

Despite the overall high mortality in the native green ash population, a small number (0.1-1%) of ash trees remain healthy and survive in otherwise highly infested forests after all other green ash have died (Knight et al., 2012, 2013). While some surviving trees are "escaped" susceptible trees that die after further exposure to EAB, other trees have reproducible resistance to the EAB. Controlled screening trials have confirmed some ash genotypes, hereafter referred to

A.)



B.)



Figure 2. The effects of emerald ash borer (EAB; *Agrilus planipennis*) on *Fraxinus pennsylvanica*. (A). An EAB gallery on an ash tree. (B). Close-up of a D-shaped exit hole on a damaged tree. Pictures taken in Founder's Park at Campbell Station, Farragut, TN.

as “lingering ash,” have reproducibly higher resistance to the pest (Koch et al., 2015). Resistance to EAB, in this trial, was defined as the ability of the plant to kill or slow the development of EAB larvae during the feeding stage, preventing the beetle from completing its life cycle. Ash species native to the natural range of EAB kill a high percentage of larvae when they feed on the inner bark, which prevents major structural damage. Evolutionarily, these EAB resistant ash trees are distant from green ash and cannot be used to breed resistance into the susceptible species. The exact mechanisms driving the lingering phenotype remain unclear, and current efforts focus on exploring this trait further through a combination of breeding, genetics, and metabolomics. Current evidence indicates that more successful lingering ash trees (high larval kill; HLK) produce lower levels of select secoiridoids than low larval kill (LLK) trees but higher levels of aromatic alkaloids (Stanley et al., n.d.). These results suggest shared mechanisms to respond and overcome EAB infestation in many lingering green ash. If the few trees that mount an effective defense were to breed, their progeny may have a strong defense response to EAB infestation.

The US Forest Service has developed an ongoing breeding program that focuses on identifying the genetic bases for the lingering trait and propagating EAB resistant individuals to restore the natural ash population (Koch et al., 2012). The breeding program takes two strategies to propagate resistance to the EAB in North American ash species. First, the program looks for Asian ash species with sympatric resistance that are capable of crossing with North American species. While this strategy has seen success when crossing black ash (*F. nigra*, sect. *Fraxinus*) with Manchurian ash (*F. mandshurica*, sect. *Fraxinus*), it has seen limited results with *Fraxinus* section *Melioides* species due to their incompatibility with Asian species (Romero-Severson & Koch, 2016). The second strategy identifies native trees with allopatric resistance or tolerance to EAB and breeds them to produce progeny carrying traits of interest. Samples of lingering ash from the wild are collected, preserved, and replicated through grafting of lingering ash material to other ash trees. Following this, EAB-resistant lingering ash are distinguished from EAB-susceptible ash by screening the grafts for larval kill, reduced larval development, and host defense responses (i.e., callous formation). This strategy shows greater promise for green ash and other native ash species. Of note, in 2010, the nursery growing area in the Forest Service Laboratory at Delaware, OH, underwent a naturally occurring EAB infestation (Koch et al., 2012). While the susceptible grafted selections were infested, the immediately adjacent lingering ash grafted selections remained uninfested.

Green ash dioecy

Overview of sex-determining systems in ash species

Breeding programs in green ash are dependent on the ability to produce seeds in as efficient a manner as possible. As a dioecious species, green ash trees can only produce seed when a male tree produces pollen that is carried by wind to a female tree’s stigma and germinated (Wallander, 2008, 2012). At present, it is impossible to sex green ash at the seedling stage, and sex is currently determined when a tree sexually matures and first produces flowers. Green ash reaches reproductive maturity at a range between 10 and 25 years after germination (Hausman et al., 2014; Schopmeyer & Others, 1974), meaning that it could potentially take many years for planters to realize they planted an inefficient seed orchard. Breeding efforts, therefore,

are delayed due to the time it takes ash trees to reach sexual maturity. A better understanding of how sex is determined in green ash can help identify markers of sex determination, which would allow researchers to sex ash trees at the seedling stage.

Sex determination mechanisms in *Fraxinus* species vary widely and include hermaphroditism, polygamy, dioecy, and androdioecy (Wallander, 2012). All fifteen species in *Fraxinus* section *Melioides*, including green ash, are dioecious with unisexual, apetalous, and wind-pollinated flowers. *Fraxinus* phylogenetic studies suggest that section *Melioides* branched off from a polygamous ancestor (Wallander, 2008, 2012). Outside of the *Melioides* ancestor, dioecy has evolved independently in two other species of ash: Chinese ash (*F. chinensis*, section *Ornus*) and Manchurian ash (*F. mandshurica*, section *Fraxinus*). Unlike the female flowers of *Fraxinus* section *Melioides* species, which contain no trace of male features, Manchurian ash female flowers contain rudimentary anthers (Wallander, 2012). Recent work has indicated a mechanism of subdioecy driven by self-incompatibility in the polygamous species *F. excelsior* (Saumitou-Laprade et al., 2018), suggesting a possible route to the evolution of dioecy. In plants, polygamy refers to the presence of both unisexual and bisexual flowers on a single tree. In extant species, European ash was commonly described as polygamous due having a wide range of sexual phenotypes - female flowers, hermaphrodite flowers, and male flowers can all be found in this species. However, European ash was also found to have a stable, genetically determined sex expression that is uncharacteristic of most polygamous plants (Albert et al., 2013). This observation, combined with confirmation of a diallelic SI (DSI) system in the closely related manna ash (*F. ornus*, section *Ornus*) (Vernet et al., 2016), allowed researchers to identify the role of DSI in European ash (Saumitou-Laprade et al., 2018). Their results indicated that, due to the DSI system being present in European ash, they are more accurately described as functionally subdioecious. While technically hermaphroditic, European ash are placed into two compatibility groups - G1 and G2 - that, despite sharing the same DSI system with other Oleaceae trees (Mariotti et al., 2019; Vernet et al., 2016), its compatibility groups are more strongly associated with functional sex than in other species. Ash trees in G2 function as males, and they can only breed with functionally female trees in G1. Conversely, while G1 trees can produce stamen, they are only able to sire seeds using pollen from G2 trees. This represents a form of subdioecy called androdioecy, where males co-occur with hermaphroditic trees. Due to these findings, it is unclear if the ancestral state of the *Melioides* ancestor, prior to the evolution of dioecy, was polygamy, as previously assumed, or this form of subdioecy.

Self-incompatibility systems in ash

It is unclear if the DSI system described in European ash, or another system of incompatibility groups, exists in green ash. In hermaphroditic plants, SI systems prevent the plant from being able to self-fertilize (Watanabe et al., 2012). Hermaphroditic plants with perfect flowers (i.e., both sex organs are in one flower) are more likely to result in self-pollination, which would produce offspring with a low genetic diversity. An SI system places a plant into a compatibility group, ensuring that the plant can only pollinate a plant outside of its own group. While pollen from a plant can make contact with a stigma from its own compatibility group, the SI system arrests pollen tube development and inhibits germination.

Traditionally, it has been understood that dioecy and self-incompatibility are negatively correlated and unlikely to occur in the same family (Baker, 1959; Charlesworth, 1985). In this

case, however, there is an overlap in dioecy and self-incompatibility systems in the Oleaceae family. As described previously, European ash has a DSI system that is directly driving a subdioecious system (Saumitou-Laprade et al., 2018). Olive (*Olea europaea*) shares the same DSI system, though it is hermaphroditic and thus not associated with functional sex as found in European ash (Mariotti et al., 2019). In the past, green ash has been shown to be capable of crossing with European ash, though some crosses are not successful (Ireland. Dept. of Agriculture and Fisheries, 1914). While the SI system has been overcome in other species by removing the tissues driving SI, most notably in *Ramosmania rodriguesii* (Magdalena, 2010), these steps have not been reported in these ash crosses. Furthermore, breeders have reported that some green ash trees appear to be sexually incompatible with one another (Koch et al., 2010). These observations would indicate that, despite evolving full dioecy, green ash may have remnants of its ancestral DSI system.

Currently known systems of plant dioecy

Two-gene systems

While little is known about the mechanisms of dioecy in ash species, research has determined the genomic mechanisms in other plant species. Dioecy has evolved independently in approximately 5-6% of angiosperms, spanning around 15,600 species, representing 987 genera from more than 175 families. Dioecious angiosperms can have a male heterogametic system (XY/XX) or a female heterogametic system (ZW/ZZ), with the majority of angiosperms with known sex-determining systems having the XY system (Leite Montalvão et al., 2020). Unlike most animal genomes, plant “sex chromosomes” are rarely heteromorphic. The male and female haplotypes are largely homologous, except at a sex-determining region (SDR) that contains one or more genes that differentiate the haplotype. The classic model of the evolution of dioecy involves two critical mutations with a gynodioecious intermediate (Charlesworth & Charlesworth, 1978), but more recent findings indicate alternative models are possible (Golenberg & West, 2013; Müller et al., 2020).

The classic model of evolution from hermaphroditism to dioecy involves two linked mutations on the same chromosome (Charlesworth & Charlesworth, 1978). The first is a recessive, male-sterility mutation that results in a gynodioecious population (where females co-exist with hermaphrodites). Following this, a dominant, female-sterility mutation must occur and must be linked with the male-sterility mutation to create a region of reduced recombination. This creates a non-recombining region in the chromosome - the SDR - and the plant has become dioecious. These mutations, most likely, will affect different genes on the same chromosome, though they may occur in a single gene. While European ash has not evolved to full dioecy, its current subdioecious system represents an exception as it is androdioecious (males co-exist with hermaphrodites), and, should the *Melioides* ancestor have evolved to dioecy from this current system, represents an alternative form of the model (Saumitou-Laprade et al., 2018). Nevertheless, the underlying mechanisms would likely be similar: one gene would regulate female flower development and the other would regulate male flower development.

This model of dioecy has been observed in garden asparagus (Harkess et al., 2020), kiwifruit (Akagi et al., 2018), and grape (Badouin et al., 2020; Massonnet et al., 2020). In asparagus (*Asparagus officinalis*), the Y-specific *SUPPRESSOR OF FEMALE FUNCTION* (*SOFF*) gene acts as the suppressor of female floral organ development. Knockout of *SOFF* in

asparagus through radiation knockout resulted in the conversion of male flowers to hermaphrodites (Harkess et al., 2017). A closely linked gene, *DEFECTIVE IN TAPETUM DEVELOPMENT AND FUNCTION 1* (*TDF1*), has been confirmed as promoting male function through knockout experiments (Harkess et al., 2020). Knockout of *TDF1* through ethyl methanesulfonate (EMS) mutagenesis converted male flowers to asexual neuters, and knockout of both *SOFF* and *TDF1* converted males to females. In kiwifruit (*Actinidia deliciosa*), the suppressor of female function is a type-C cytokinin response regulator (“*SHY GIRL*”; *SyGI*) (Akagi et al., 2018), and the promoter of male function is a *MICROSPORE AND TAPETUM REGULATOR 1* (*MTR1*) protein (“*FRIENDLY BOY*”; *FrBy*) (Akagi et al., 2019). In grapevines, the candidate male-sterility mutation is *INAPERTURATE POLLEN 1* (*INP1*), mutations in which result in sterile pollen, and the candidate female-sterility mutation is a *YABBY3* transcription factor, mutations in which result in a sterile pistil (Massonnet et al., 2020). Human cultivation has resulted in a hermaphroditic subspecies of grape (*V. vinifera* subsp. *vinifera*), and the current model of grape dioecy indicates that this is the result of a rare recombination event in a male during domestication that removed the *YABBY3* mutation and restored pistil function. These are just a few of the dioecious plants observed to follow the Charlesworth model.

Single-gene systems

In recent years, more exceptions to the Charlesworth evolutionary model of plant dioecy have been observed, commonly in relation to cases where dioecy evolved from monoecy (the presence of both unisexual flowers on a single plant at different times) (Renner, 2016). In most cases, these exceptions are driven by a single gene that affects unlinked genes driving male or female function (Golenberg & West, 2013). Rather than being connected by physical linkage, the factors that determine sex are connected through epistatic interactions, and the gene controlling dioecy alters these interactions to produce a different sex. Turnovers between XY and ZW systems are common in these plants due to the single gene requirement (Müller et al., 2020). Monoecious plants can be artificially induced to become dioecious by altering genes affecting hormones, as has been performed in maize (*Zea mays*) (Jones, 1934) and melon (*Cucumis melo*) (Boualem et al., 2015). In other plants, floral MADS-box genes are excellent candidate genes in single gene models. If the required genes are missing, then the features of one sex will not form. Two B-class MADS-box genes - *PISTILLATA* (*PI*) and *APETALA3* (*AP3*) - are required for the formation of stamen and prevent the expression of carpels wherever they are expressed (Cronk & Müller, 2020). If *PI/AP3* are over-expressed in a flower, it is expected that they will form flowers lacking any female organs. Similarly, if these genes were to be silenced, then only carpels would form. These examples indicate alternative paths to the evolution of dioecy that do not rely on the two-gene model (Renner, 2016).

One such single-gene model, involving MADS-box genes, has become well-characterized in poplar species (*Populus* sp.). Some species of poplars, including *P. alba*, have a ZW sex-determining system, while most others have an XY system (Müller et al., 2020). The same gene, an ortholog of *ARABIDOPSIS RESPONSE REGULATOR 17* (*ARR17*), is responsible for sex-determination in both systems, despite having different SDRs. The SDR in *Populus* was first identified on chromosome 19 in a variety of species, and later genome-wide association studies (GWAS) indicated that *ARR17* was strongly correlated with sex determination in poplar (Geraldes et al., 2015; McKown et al., 2017). Knockout experiments have confirmed that the

loss of *ARR17* in female poplar results in the growth of male flowers (Müller et al., 2020). Furthermore, this study characterized the difference between sex determination in XY and ZW poplars. In ZW poplars, the male (Z) haplotype lacks the *ARR17* gene, and the female (W) haplotype contains a single copy. The location of the *ARR17* gene serves as the SDR in these poplars. In XY poplars, both the female (X) and male (Y) haplotypes contain *ARR17*, but the male haplotype also contains an *ARR17* pseudogene inverted repeat (*ΨARR17-IR*). This pseudogene produces small RNA that targets *ARR17* transcripts for silencing, and the region containing the pseudogene represents the SDR in poplars with this system. Gene expression analysis of male and female poplar identified up-regulation of *PI/AP3* exclusively in males (Cronk et al., 2020), indicating that *ARR17* functions by targeting these MADS-box genes for silencing. Further expression analyses comparing genetically identical aspen (*P. tremula*) with or without an *arr17* knockout confirmed that *PI* expression is inhibited by *ARR17* (Leite Montalvão et al., 2022). Overexpression of *PI/AP3* produces stamen-only flowers, as observed in male poplars, and blocking the expression of *PI/AP3* inhibits stamen growth and results in carpal-only flowers, as observed in female poplars.

The overall model and its applicability to green ash

An updated, unified model of plant dioecy has been proposed to take these findings into account (Leite Montalvão et al., 2020). Despite the surface-level differences between the two-gene and single-gene models, the genes involved in all currently observed systems are connected in a single genetic network of promoters and suppressors of male and female traits. In monoecious plants and dioecious plants with a single-gene system, the “sex-switch” gene acts on a high level by repressing a female suppressor. Without the female suppressor, male-promoting genes are inhibited. This results in flowers lacking vestigial remnants of the opposite sex. In monoecious plants, this is localized based on hormones, while in dioecious plants, this is determined by either presence/absence of the sex-switch gene or silencing of the sex-switch gene by small RNA. Two-gene systems, by contrast, do not have genes that interact with one another, but are linked within the sex-determining genomic region. This linkage between a female suppressor and male suppressor is essential to control flower development, as the genes are controlled independently. Proper linkage ensures that hermaphrodites and asexual neuters do not happen, though crossover and reversion to hermaphroditic flowers was observed in the case of domesticated grapes (Massonnet et al., 2020). The unified model builds off of the classic model (Charlesworth & Charlesworth, 1978), while incorporating what appeared to be exceptions to this rule of plant dioecy (Renner, 2016).

Based on these observations of dioecy in other plant species, we have a guideline for what systems are most likely driving sex determination in green ash. The current evolutionary hypothesis in European ash - subdioecy with functional males pollinating functional females - would point to a form of the two gene pathway, with one gene acting as a male floral organ inhibitor and the other acting as a female floral organ inhibitor. However, whereas the most common form of this two-gene system sees gynodioecy as the halfway point between hermaphroditism and full dioecy, the system observed in ash is closer to androdioecy. Furthermore, the similarities between poplar and green ash floral structures strongly indicates that MADS-box genes play a role in sex-determination. Most sex-determining systems in dioecious plants, such as grapevine, result in flowers containing vestigial organs of the opposite

sex (Massonnet et al., 2020). Using grapevine as an example, the mutations in *INP1* do not prevent the growth of stamen in female flowers, but they result in stunted stamen that produce sterile pollen. By contrast, in poplar, there is a clean deletion of opposite sex structures - male flowers consist solely of stamen, and female flowers consist solely of carpels (Müller et al., 2020). This is the result of *ARR17* targeting *PI/AP3*, MADS-box genes responsible for stamen formation, for silencing, and the absence (ZW) or silencing (XY) of *ARR17* allows them to be overexpressed. Rather than resulting in stunted sexual organs of the opposite sex, the silencing and overexpression of MADS-box genes prevents their development entirely. The structure of male and female green ash flowers is similar to poplar in that they do not contain any features from the opposite sex (Figure 1) (Wallander, 2012). For this reason, we expect that one or more genes in green ash drive dioecy by targeting MADS-box genes such as *PI/AP3*.

REFERENCES

- Akagi, T., Henry, I. M., Ohtani, H., Morimoto, T., Beppu, K., Kataoka, I., & Tao, R. (2018). A Y-Encoded Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response Regulator in Kiwifruit. *The Plant Cell*, 30(4), 780–795.
- Akagi, T., Pilkington, S. M., Varkonyi-Gasic, E., Henry, I. M., Sugano, S. S., Sonoda, M., Firl, A., McNeilage, M. A., Douglas, M. J., Wang, T., Rebstock, R., Voogd, C., Datson, P., Allan, A. C., Beppu, K., Kataoka, I., & Tao, R. (2019). Two Y-chromosome-encoded genes determine sex in kiwifruit. *Nature Plants*, 5(8), 801–809.
- Albert, B., Morand-Prieur, M.-É., Brachet, S., Gouyon, P.-H., Frascaria-Lacoste, N., & Raquin, C. (2013). Sex expression and reproductive biology in a tree species, *Fraxinus excelsior* L. In *Comptes Rendus Biologies* (Vol. 336, Issue 10, pp. 479–485). <https://doi.org/10.1016/j.crv.2013.08.004>
- Badouin, H., Velt, A., Gindraud, F., Flutre, T., Dumas, V., Vautrin, S., Marande, W., Corbi, J., Sallet, E., Ganofsky, J., Santoni, S., Guyot, D., Ricciardelli, E., Jepsen, K., Käfer, J., Berges, H., Duchêne, E., Picard, F., Hugueney, P., ... Marais, G. A. B. (2020). The wild grape genome sequence provides insights into the transition from dioecy to hermaphroditism during grape domestication. *Genome Biology*, 21(1), 223.
- Baker, H. G. (1959). Reproductive methods as factors in speciation in flowering plants. *Cold Spring Harbor Symposia on Quantitative Biology*, 24, 177–191.
- Bakys, R., Vasaitis, R., Barklund, P., Ihrmark, K., & Stenlid, J. (2009). Investigations concerning the role of *Chalara fraxinea* in declining *Fraxinus excelsior*. *Plant Pathology*, 58(2), 284–292.
- Boualem, A., Troadec, C., Camps, C., Lemhemdi, A., Morin, H., Sari, M.-A., Fraenkel-Zagouri, R., Kovalski, I., Dogimont, C., Perl-Treves, R., & Bendahmane, A. (2015). A cucurbit androecy gene reveals how unisexual flowers develop and dioecy emerges. *Science*, 350(6261), 688–691.
- Burns, R. M., & Honkala, B. M. (1990). Silvics of North America. Vol. 1. Conifers. US Dep. Agric Agric. Handb, 654.
- Charlesworth, D. (1985). Distribution of dioecy and selfincompatibility in angiosperms. *Evolution-Essays in Honour of John Maynard Smith*, 237–267.
- Charlesworth, D., & Charlesworth, B. (1978). Population genetics of partial male-sterility and the evolution of monoecy and dioecy. *Heredity*, 41(2), 137–153.
- Coker, T. L. R., Rozsypálek, J., Edwards, A., Harwood, T. P., Butfoy, L., & Buggs, R. J. A. (2019). Estimating mortality rates of European ash (*Fraxinus excelsior*) under the ash dieback (*Hymenoscyphus fraxineus*) epidemic. *Plants, People, Planet*, 1(1), 48–58.
- Cronk, Q., & Müller, N. A. (2020). Default Sex and Single Gene Sex Determination in Dioecious Plants. *Frontiers in Plant Science*, 11, 1162.
- Cronk, Q., Soolanayakanahally, R., & Bräutigam, K. (2020). Gene expression trajectories during male and female reproductive development in balsam poplar (*Populus balsamifera* L.). *Scientific Reports*, 10(1), 8413.
- Geraldes, A., Hefer, C. A., Capron, A., Kolosova, N., Martinez-Nuñez, F., Soolanayakanahally, R. Y., Stanton, B., Guy, R. D., Mansfield, S. D., Douglas, C. J.,

- & Cronk, Q. C. B. (2015). Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). In *Molecular Ecology* (Vol. 24, Issue 13, pp. 3243–3256). <https://doi.org/10.1111/mec.13126>
- Golenberg, E. M., & West, N. W. (2013). Hormonal interactions and gene regulation can link monoecy and environmental plasticity to the evolution of dioecy in plants. *American Journal of Botany*, 100(6), 1022–1037.
- Harkess, A., Huang, K., van der Hulst, R., Tissen, B., Caplan, J. L., Koppula, A., Batish, M., Meyers, B. C., & Leebens-Mack, J. (2020). Sex Determination by Two Y-Linked Genes in Garden Asparagus. *The Plant Cell*, 32(6), 1790–1796.
- Harkess, A., Zhou, J., Xu, C., Bowers, J. E., Van der Hulst, R., Ayyampalayam, S., Mercati, F., Riccardi, P., McKain, M. R., Kakrana, A., Tang, H., Ray, J., Groenendijk, J., Arikiti, S., Mathioni, S. M., Nakano, M., Shan, H., Telgmann-Rauber, A., Kanno, A., ... Chen, G. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature Communications*, 8(1), 1279.
- Hausman, C. E., Bertke, M. M., Jaeger, J. F., & Rocha, O. J. (2014). Genetic structure of green ash (*Fraxinus pennsylvanica*): implications for the establishment of ex situ conservation protocols in light of the invasion of the emerald ash borer. *Plant Genetic Resources: Characterization and Utilization*, 12(3), 286–297.
- Hilty, J. (n.d.). *Green Ash (Fraxinus pennsylvanica lanceolata)*. Illinois Wildflowers. Retrieved March 3, 2023, from https://www.illinoiswildflowers.info/trees/plants/green_ash.htm
- Ireland. Dept. of Agriculture and Fisheries. (1914). *Journal*.
- Jones, D. F. (1934). Unisexual Maize Plants and Their Bearing on Sex Differentiation in Other Plants and in Animals. *Genetics*, 19(6), 552–567.
- Kennedy, H. E., Jr. (1949). *Fraxinus pennsylvanica* Marsh. green ash. *Agriculture Handbook*, 2(654), 348.
- Knight, K. S., Brown, J. P., & Long, R. P. (2013). Factors affecting the survival of ash (*Fraxinus* spp.) trees infested by emerald ash borer (*Agrilus planipennis*). *Biological Invasions*, 15(2), 371–383.
- Knight, K. S., Herms, D., Plumb, R., Sawyer, E., Spalink, D., Pisarczyk, E., Wiggins, B., Kappler, R., Ziegler, E., & Menard, K. (2012). Dynamics of surviving ash (*Fraxinus* spp.) populations in areas long infested by emerald ash borer (*Agrilus planipennis*). *Proceedings of the 4th International Workshop on Genetics Host-Parasite Interactions in Forestry, USDA Forest Service Pacific Southwest Research Station, PSW-GTR-240*, 143–152.
- Koch, J. L., Carey, D. W., Knight, K. S., Poland, T., Herms, D. A., & Mason, M. E. (2012). Breeding strategies for the development of emerald ash borer-resistant North American ash. In: *Snieszko, Richard A.; Yanchuk, Alvin D.; Kliejunas, John T.; Palmieri, Katharine M.; Alexander, Janice M.; Frankel, Susan J., Tech. Coords. Proceedings of the Fourth International Workshop on the Genetics of Host-Parasite Interactions in Forestry: Disease and Insect Resistance in Forest Trees. Gen. Tech. Rep. PSW-GTR-240. Albany, CA: Pacific Southwest Research Station, Forest Service, US Department of Agriculture. Pp. 235-239., 240, 235–239.*
- Koch, J. L., Carey, D. W., Mason, M. E., & Nurul Islam-Faridi, M. (2010, March 9).

- Overcoming obstacles to interspecies hybridization of ash.*
<https://www.nrs.fs.usda.gov/pubs/gtr/gtr-p-72papers/26-koch-p-72.pdf>
- Koch, J. L., Carey, D. W., Mason, M. E., Poland, T. M., & Knight, K. S. (2015). Intraspecific variation in *Fraxinus pennsylvanica* responses to emerald ash borer (*Agrilus planipennis*). *New Forests*, 46(5), 995–1011.
- Kovacs, K. F., Haight, R. G., McCullough, D. G., Mercader, R. J., Siegert, N. W., & Liebhold, A. M. (2010). Cost of potential emerald ash borer damage in U.S. communities, 2009–2019. In *Ecological Economics* (Vol. 69, Issue 3, pp. 569–578). <https://doi.org/10.1016/j.ecolecon.2009.09.004>
- Leite Montalvão, A. P., Kersten, B., Fladung, M., & Müller, N. A. (2020). The Diversity and Dynamics of Sex Determination in Dioecious Plants. *Frontiers in Plant Science*, 11, 580488.
- Leite Montalvão, A. P., Kersten, B., Kim, G., Fladung, M., & Müller, N. A. (2022). ARR17 controls dioecy in *Populus* by repressing B-class MADS-box gene expression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 377(1850), 20210217.
- Magdalena, C. (2010). Raising the living dead: *Ramosmania rodriguesii*. *Sibbaldia: The International Journal of Botanic Garden Horticulture*, 8, 63–73.
- Mariotti, R., Fornasiero, A., Mousavi, S., Cultrera, N. G. M., Brizioli, F., Pandolfi, S., Passeri, V., Rossi, M., Magris, G., Scalabrin, S., Scaglione, D., Di Gaspero, G., Saumitou-Laprade, P., Vernet, P., Alagna, F., Morgante, M., & Baldoni, L. (2019). Genetic Mapping of the Incompatibility Locus in Olive and Development of a Linked Sequence-Tagged Site Marker. *Frontiers in Plant Science*, 10, 1760.
- Massonnet, M., Cochetel, N., Minio, A., Vondras, A. M., Lin, J., Muyle, A., Garcia, J. F., Zhou, Y., Delledonne, M., Riaz, S., Figueroa-Balderas, R., Gaut, B. S., & Cantu, D. (2020). The genetic basis of sex determination in grapes. *Nature Communications*, 11(1), 2902.
- McKown, A. D., Klápště, J., Guy, R. D., Soolanayakanahally, R. Y., La Mantia, J., Porth, I., Skyba, O., Unda, F., Douglas, C. J., El-Kassaby, Y. A., Hamelin, R. C., Mansfield, S. D., & Cronk, Q. C. B. (2017). Sexual homomorphism in dioecious trees: extensive tests fail to detect sexual dimorphism in *Populus* †. *Scientific Reports*, 7(1), 1831.
- Müller, N. A., Kersten, B., Leite Montalvão, A. P., Mähler, N., Bernhardsson, C., Bräutigam, K., Carracedo Lorenzo, Z., Hoenicka, H., Kumar, V., Mader, M., Pakull, B., Robinson, K. M., Sabatti, M., Vettori, C., Ingvarsson, P. K., Cronk, Q., Street, N. R., & Fladung, M. (2020). A single gene underlies the dynamic evolution of poplar sex determination. *Nature Plants*, 6(6), 630–637.
- Nielsen, L. R., McKinney, L. V., Hietala, A. M., & Kjær, E. D. (2017). The susceptibility of Asian, European and North American *Fraxinus* species to the ash dieback pathogen *Hymenoscyphus fraxineus* reflects their phylogenetic history. In *European Journal of Forest Research* (Vol. 136, Issue 1, pp. 59–73). <https://doi.org/10.1007/s10342-016-1009-0>
- Ray, T., Kaljun, J., & Straže, A. (2021). Comparison of the Vibration Damping of the Wood Species Used for the Body of an Electric Guitar on the Vibration Response of Open-Strings. *Materials*, 14(18). <https://doi.org/10.3390/ma14185281>

- Rebek, E. J., Herms, D. A., & Smitley, D. R. (2008). Interspecific variation in resistance to emerald ash borer (Coleoptera: Buprestidae) among North American and Asian ash (*Fraxinus* spp.). *Environmental Entomology*, 37(1), 242–246.
- Renner, S. S. (2016). Pathways for making unisexual flowers and unisexual plants: Moving beyond the “two mutations linked on one chromosome” model. In *American Journal of Botany* (Vol. 103, Issue 4, pp. 587–589). <https://doi.org/10.3732/ajb.1600029>
- Romero-Severson, J., & Koch, J. L. (2016). Saving green ash. *Proceedings of Workshop on Gene Conservation of Tree Species--Banking on the Future*, 102.
- Saumitou-Laprade, P., Vernet, P., Dowkiw, A., Bertrand, S., Billiard, S., Albert, B., Gouyon, P.-H., & Dufay, M. (2018). Polygamy or subdioecy? The impact of diallelic self-incompatibility on the sexual system in *Fraxinus excelsior* (Oleaceae). *Proceedings. Biological Sciences / The Royal Society*, 285(1873). <https://doi.org/10.1098/rspb.2018.0004>
- Schopmeyer, C. S., & Others. (1974). Seeds of woody plants in the United States. *Seeds of Woody Plants in the United States.*, 450. <https://www.cabdirect.org/cabdirect/abstract/19750622387>
- Siebert, N. W., McCullough, D. G., Liebhold, A. M., & Telewski, F. W. (2014). Dendrochronological reconstruction of the epicentre and early spread of emerald ash borer in North America. In *Diversity and Distributions* (Vol. 20, Issue 7, pp. 847–858). <https://doi.org/10.1111/ddi.12212>
- Stanley, R. K., Carey, D. W., Mason, M. E., Poland, T. M., Koch, J. L., Daniel Jones, A., & Romero-Severson, J. (n.d.). *Profiles of secoiridoids and alkaloids in tissue of susceptible and resistant green ash progeny reveal patterns of induced responses to emerald ash borer in Fraxinus pennsylvanica*. <https://doi.org/10.1101/2022.05.18.492370>
- Stephens, J. P., Berven, K. A., & Tiegs, S. D. (2013). Anthropogenic changes to leaf litter input affect the fitness of a larval amphibian. *Freshwater Biology*, 58(8), 1631–1646.
- Vernet, P., Lepercq, P., Billiard, S., Bourceaux, A., Lepart, J., Dommée, B., & Saumitou-Laprade, P. (2016). Evidence for the long-term maintenance of a rare self-incompatibility system in Oleaceae. *The New Phytologist*, 210(4), 1408–1417.
- Wallander, E. (2008). Systematics of *Fraxinus* (Oleaceae) and evolution of dioecy. *Osterreichische Botanische Zeitschrift*, 273(1-2), 25–49.
- Wallander, E. (2012). Systematics and floral evolution in *Fraxinus* (Oleaceae). *Belgische Dendrologie Belge*, 2012, 39–58.
- Watanabe, M., Suwabe, K., & Suzuki, G. (2012). Molecular genetics, physiology and biology of self-incompatibility in Brassicaceae. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, 88(10), 519–535.
- Wright, J. W. (n.d.). Green ash (*Fraxinus pennsylvanica* Marsh.). *Silvics of Forest Trees of the United States*. HA Fowells.

CHAPTER III:
A HIGH-QUALITY REFERENCE GENOME FOR *FRAXINUS*
***PENNSYLVANICA* FOR ASH SPECIES RESTORATION AND RESEARCH**

A version of this chapter was originally published by Huff et al.:

Huff, M., Seaman, J., Wu, D., Zhebentyayeva, T., Kelly, L. J., Faridi, N., Nelson, C. D., Cooper, E., Best, T., Steiner, K., Koch, J., Romero Severson, J., Carlson, J. E., Buggs, R., & Staton, M. (2022). A high-quality reference genome for *Fraxinus pennsylvanica* for ash species restoration and research. *Molecular ecology resources*, 22(4), 1284–1302.
<https://doi.org/10.1111/1755-0998.13545>

The sections of this publication have been modified for uniformity of the present work and the figure and table numbers have been changed to reflect chronological numbering within this dissertation. Work from this publication that is irrelevant to the dissertation work has been excluded, and any results that mention irrelevant work cite the original publication.

Abstract

Green ash (*Fraxinus pennsylvanica*) is the most widely distributed ash tree in North America. Once common, it has experienced high mortality from the non-native invasive emerald ash borer (EAB; *Agrilus planipennis*). A small percentage of native green ash trees that remain healthy in long-infested areas, termed “lingering ash,” display partial resistance to the insect, indicating that breeding and propagating populations with higher resistance to EAB may be possible. To assist in ash breeding, ecology and evolution studies, we report the first chromosome-level assembly from the *Fraxinus* genus for *F. pennsylvanica* with over 99% of bases anchored to 23 haploid chromosomes, spanning 757Mb in total, composed of 49.43% repetitive DNA, and containing 35,470 high confidence gene models assigned to 22,976 Asterid orthogroups. We also present results of range-wide genetic variation studies, the identification of candidate genes for important traits including potential EAB-resistance genes, and an investigation of comparative genome organization among Asterids based on this reference genome platform. Residual duplicated regions within the genome likely resulting from a recent whole genome duplication event in Oleaceae were visualized in relation to wild olive (*Olea europaea* var. *sylvestris*). We used our *F. pennsylvanica* chromosome assembly to construct reference-guided assemblies of 27 previously sequenced *Fraxinus* taxa, including *F. excelsior*. Thus we present a significant step forward in genomic resources for research and protection of *Fraxinus* species.

Introduction

Fraxinus pennsylvanica Marsh., commonly known as green ash (also as red ash, swamp ash, or water ash), is the most widely distributed ash species in North America with a broad range across the east and midwest of the continent (Kennedy 2021). Green ash has economic importance (Kovacs et al., 2010) particularly as a landscaping tree. Due to green ash’s adaptability to urban environments, it became a popular ornamental tree following the large-scale loss of American elm trees due to Dutch Elm Disease (Burns & Honkala, 1990). Its lumber is popular for woodworking, and it is used for many specialty products. Ecologically, green ash

is adapted to a wide variety of environmental conditions and is considered an important source of food and/or cover for wildlife across much of its range (Gucker, 2005).

The emerald ash borer (EAB; *Agrilus planipennis*) is an invasive species of jewel beetle that feeds on ash trees and is a critical threat to the native ash populations of North America, including green ash. Native to northeast Asia, the EAB is believed to have arrived in America through wood packing materials. While it was first found in the United States in 2002, evidence suggests the invasive population can be traced back to southeast Michigan in the early 1990s (Siegert, McCullough, Liebhold, & Telewski, 2014). Females lay their eggs on the bark of an ash tree and after hatching, the larvae chew through the bark and feed on the phloem and vascular cambium, disrupting the transport of sugars and water (Poland & McCullough, 2006). Within six years, the presence of the EAB can reduce a healthy population of North American ash trees to near complete mortality (Knight, Brown, & Long, 2013).

Worldwide there are over forty *Fraxinus* species, distributed throughout the temperate forests of North America, Europe and Asia. While some ash species native to Asia have resistance to EAB, species outside of the beetle's native range - including those of North America and Europe - are largely highly susceptible to the beetle's effects (Kelly et al., 2020; Rebek, Herms, & Smitley, 2008). Economic impact from EAB in the United States has been estimated to be as much as \$10.7 billion including the replacement of 17 million ornamental ash trees (Kovacs et al., 2010). Despite the overall high mortality, a small percentage of ash trees remain healthy. Further controlled EAB screening trials have confirmed some ash genotypes as having reproducibly higher resistance to the pest. Identifying the genetic basis for this trait could benefit an ongoing US Forest Service breeding program to develop ash with enhanced resistance to EAB (Koch et al., 2012).

The *Fraxinus* species of Europe face an additional threat in the form of the Ascomycete fungus *Hymenoscyphus fraxineus*. *H. fraxineus* is the causative agent of a fatal disease in ash known as ash dieback, named for the crown dieback caused by the disease (Bakys, Vasaitis, Barklund, Ihrmark, & Stenlid, 2009). Symptoms of the infection include necrotic spots on stems that eventually enlarge into cankers and wilting of leaves. European ash trees infected with *H. fraxineus* have an estimated mortality rate of 85% in plantations and 69% in woodland populations (Coker et al., 2019). It has primarily impacted European ash (*Fraxinus excelsior*) populations, though other species have been affected as well. While not yet detected in North America, experiments indicate that American species, including green ash, are mildly susceptible (Nielsen, McKinney, Hietala, & Kjær, 2017).

Genomics can be an important tool to combat threats posed by invasive pests and pathogens. Annotated genomes have contributed to resistance breeding programs in a number of crop species by allowing researchers to identify genes for resistance (Babu et al., 2020; Pérez-de-Castro et al., 2012) and recent studies show promising results for tree species (Dario Grattapaglia et al., 2018; Neale & Kremer, 2011; Plomion et al., 2016). European ash (*F. excelsior*) has the most contiguous and well annotated ash genome to date, with 89,514 nuclear scaffolds and 38,852 protein-coding genes (Sollars et al., 2017). Less contiguous scaffold- or contig-level genomes are also available for 26 other *Fraxinus* taxa, including three green ash individuals, though at present these do not have *de novo* gene annotations available (Kelly et al., 2020). A green ash transcriptome was published in 2016, exploring the effects of EAB feeding and other stresses on gene expression in green ash trees (Lane et al., 2016), and a genetic linkage map for

green ash was reported in 2019 (Wu, Koch, Coggeshall, & Carlson, 2019). This linkage map contained a total of 23 linkage groups spanning about 2009 cM, with a total of 1,201 markers and an average inter-marker distance of 1.67 cM. (Wu et al., 2019).

To support efforts to combat EAB and other threats to ash species, we produced a chromosome-scale reference genome for green ash scaffolded with an improved genetic linkage map, which provides a foundation for exploring population diversity across the native range, discovery of trait-associated loci including for survival after EAB infestation, and comparative genomics across the Asterids. Genome-guided scaffolding of the draft genomes from an additional 27 *Fraxinus* taxa representing 23 species extends these resources to support research and breeding efforts in other threatened ash species.

Materials and Methods

Genome annotation and quality assessment

Following identification and softmasking of repetitive elements with RepeatModeler version 1.0.11 and RepeatMasker version 4.0.9, gene annotations were predicted using BRAKER version 2.1.5 (Hoff et al., [2016](#), [2019](#); Smit et al., [2015a](#), [2015b](#)). Annotations were filtered by structure and function using gFACs version 1.1.2 and EnTAP version 0.9.1, respectively (Caballero & Wegrzyn, [2019](#); Hart et al., [2020](#)). Benchmarking Universal Single-Copy Orthologs (BUSCO) version 4.0 was run to assess the completeness of the *Fraxinus pennsylvanica* assembly (Seppey et al., [2019](#)). Statistical analysis of the pseudomolecules and scaffolds of the input genome was performed using QAST version 5.0.2 (Gurevich et al., [2013](#)).

Characterizing the whole genome duplication event

Four species from within the Lamiids subclade of the Asterid clade were selected for comparative genomics: wild olive (*Olea europaea* var. *sylvestris*) (Unver, Wu, Sterck, & Turktas, 2017), common yellow monkeyflower (*Erythranthe guttata*, formerly known as *Mimulus guttatus*) (Hellsten et al., 2013), tomato (*Solanum lycopersicum*) (Tomato Genome Consortium, 2012), and coffee (*Coffea canephora*) (Denoëud et al., 2014). Carrot (*Daucus carota*) was selected as an outgroup (Iorizzo et al., 2016). The rate of synonymous mutations (K_s) in each species was determined using reciprocal BLASTP analyses and inputting these results into the custom KsPlotter python script as previously described (Sollars et al., 2017). After generating the K_s plot, we isolated gene clusters in *F. pennsylvanica* that represent the most recent WGD event predicted in the plot, indicated with a K_s of 0.25 or less. A Circos plot to visualize these gene clusters was created with a custom python script (https://github.com/MattHuff/Green_Ash_Annotation/blob/master/get_coordinates_ash_ash.py) to match genes in a cluster with their coordinates along the genome (Krzywinski et al., 2009).

***Fraxinus* spp. reference-guided genome scaffolding**

BUSCO and RepeatMasker were run on the BAT0.5 *F. excelsior* assembly (Sollars et al., 2017) using the same parameters previously described (Seppey et al., 2019; Smit et al., 2015a). We downloaded 28 other publicly available *Fraxinus* scaffold- and contig-level assemblies (Kelly et al., 2020). We used RagTag version 1.0.1 to align the scaffolds of each assembly to the chromosomes of *F. pennsylvanica* and join them to produce chromosome-scale assemblies (Alonge et al., 2019). Following the same methods as above, we masked repetitive elements from the genomes with RepeatMasker (Smit et al., 2015a) and gene annotations with BRAKER2 (Hoff et al., 2016, 2019).

Comparative genomics with Asterids

OrthoFinder v2.3.12 was used to identify orthologs in *F. pennsylvanica* and the five other species of Asterids used for WGD analysis. Multiple sequence alignment (MSA) mode was selected, which used MAFFT v7.467 to infer gene trees and obtain sequence alignments (Emms & Kelly, 2019; Katoh & Standley, 2013). The output of OrthoFinder was used to identify blocks of synteny between *F. pennsylvanica* and the other asterid species. Orthologs among *F. pennsylvanica*, *O. europaea*, and *C. canephora* were selected by identifying orthogroups with a single gene member from each species (Staton et al., 2020). The command line version of Circos, version 0.69-6, used these orthologous links to visualize synteny between the genomes (Krzywinski et al., 2009).

Results

Genetic linkage map

The *Fraxinus pennsylvanica* genetic linkage map based on 95 progeny (Wu et al., 2019) from the PE0028 x PE0248 cross was expanded by genotyping a further 160 F₂ individuals (Huff et al., 2022). This resulted in a consensus linkage map representing both parents composed of 4,193 SNPs organized in 23 linkage groups representing the 23 haploid chromosomes (Figure 3A, Supplementary Table S1) (Löve, 1982). The total map length was 1675.9 cM with linkage groups ranging from 49.6 cM (LG16) to 104.5cM (LG2) yielding an average marker density of 0.4 cM per SNP.

Assembly and scaffolding

The reference genome was generated from PE00248, a male tree with EAB resistance and a parent of the cross used for genetic mapping (Kelly et al., 2020; Wu et al., 2019). The assembly utilized Illumina paired end and mate pair reads, then underwent additional scaffolding with Hi-C data and the new genetic linkage map (Huff et al., 2022). This yielded a high quality, chromosome-scale reference genome (version 1.4) with 23 primary sequences corresponding to the expected 23 haploid chromosomes and spanning 755 Mb (Table 1). An additional 87 scaffolds of 10 kb or more in length remain unplaced (totaling 2 Mb). The genome assembly has a total length of 757 Mb, representing 81.4% to 85.1% of the total predicted length of 890-930

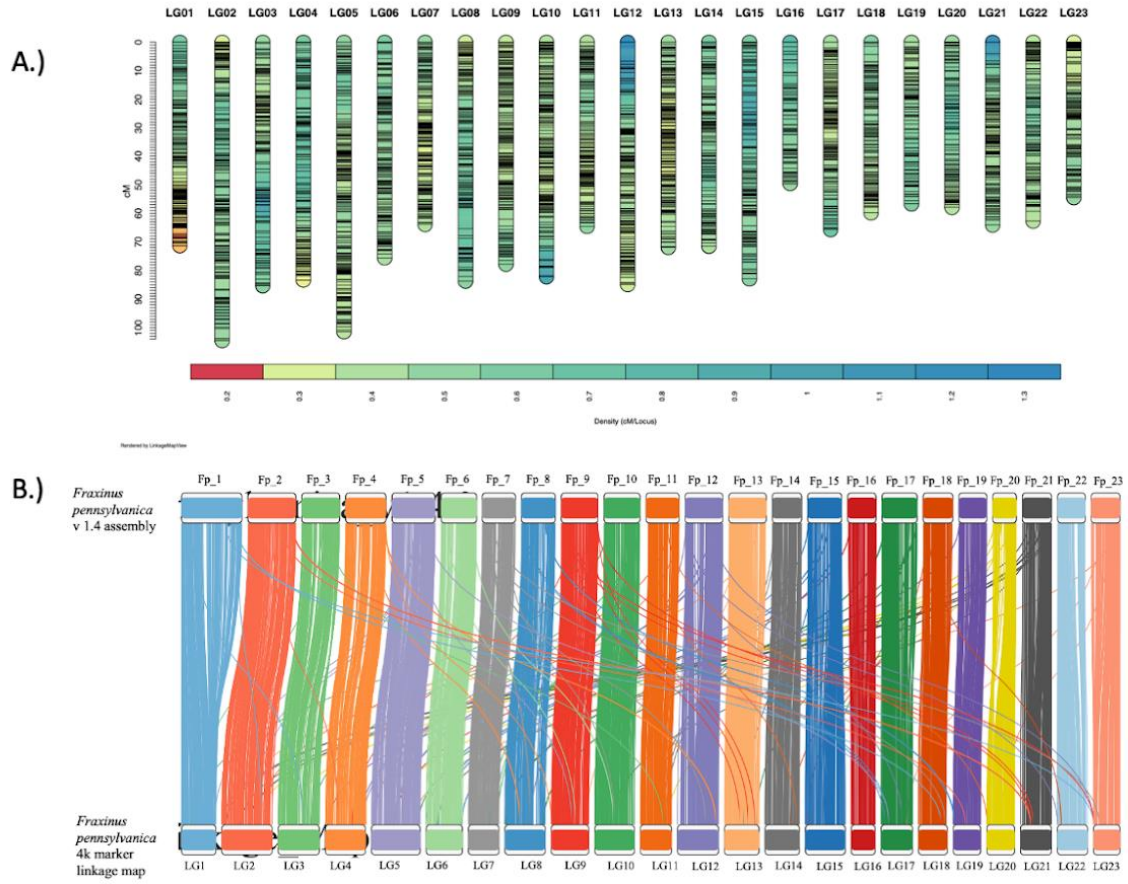


Figure 3. High-density, consensus genetic linkage map of *Fraxinus pennsylvanica* (Huff et al., 2022). (a) Consensus genetic map of the PE0048 × PE0248 *F. pennsylvanica* cross composed of 4193 SNPs. Black tickmarks represent markers segregating either in female or male parent. Vertical scale on the left reflects map genetic distances in centimorgans (cM). Coloured scale in the bottom shows variation in marker density (cM per locus) across the linkage groups. (b) Alignment of sequence-based genetic markers from the high-resolution linkage map (bottom) to the chromosomes of the *F. pennsylvanica* genome assembly (top)

Table 1. Summary of the *Fraxinus pennsylvanica* genome assembly.

Total Length	756,791,288
# of Scaffolds	110
Average Length	6,879,920.8
Largest Scaffold	56,547,140
Smallest Scaffold	10,000
Number of Ns	91,729,614
Chromosome Length	755,065,760
Chromosome Scaffolds (%)	99.77%
GC (%)	34.40%
Repetitive Elements (%)	48.80%
Protein Coding Gene Models	35,470
Reads Mapped (%)	88.97%
Collinear Markers (%)	97.40%
BUSCO Description	Number in Genome
Complete BUSCOs (C)	1566 (97.03%)
Complete and single-copy BUSCOs (S)	1345 (83.33%)
Complete and duplicated BUSCOs (D)	221 (13.69%)
Fragmented BUSCOs (F)	25 (1.55 %)
Missing BUSCOs (M)	23 (1.43%)
Total BUSCO groups searched	1614 (100%)

Mb estimated from flow cytometry (Kelly et al., 2020). The 23 chromosome sequences range from 22.2 Mb to 56.5 Mb in length and make up 99.7% of the total sequence content. Genome quality was assessed with multiple methods. First, to assess the accuracy of the assembly of the original contigs and scaffolding into chromosome order, sequences for each marker from the high-resolution linkage map were aligned to the current assembly (Figure 3B). Of the 4,117 markers that aligned to the assembly, 4,010 (97.4%) aligned to their expected linkage group on the map. Next, the proportion of the genome captured in the assembly was evaluated by mapping the original paired end short reads to the final assembly. While the reference genome has fewer bases than the estimated genome length and a significant proportion of uncalled bases (12.1%), we found that over 89% of short read pairs map to the genome sequence. This indicates that a large majority of the genome is represented in the assembly, and the reduced length and number of Ns may be due to collapsed assembly of repetitive areas. Finally, completeness was evaluated with BUSCO (Benchmarking Universal Single-Copy Orthologs) to confirm the presence of expected orthologs (Table 1). Of the 1,614 BUSCO groups searched, 97% were complete and present at least once in the genome. 83.3% of these complete BUSCOs were single copy while 13.7% were duplicated. All BUSCOs were located on the 23 chromosomes. Overall, all evaluation metrics support that the assembly is largely complete and correctly scaffolded.

Annotation

Using known plant repetitive element sequences and *de novo* repeat discovery, 49.43% of the genome was identified as repetitive and masked prior to gene annotation (Supplemental Table S1). Like most characterized plant genomes, long terminal repeats (LTRs) were most commonly identified (24.53%), primarily from the Ty1-Copia (13.53%) and Ty3-Gypsy (10.26%) families (Baucom et al., 2009).

Gene annotation yielded an initial set of 53,977 gene predictions that were further filtered by structural and functional annotation to a set of 35,470 high confidence gene models (Hart et al., 2020). All high confidence genes are located on the 23 chromosome sequences. The majority of high-quality genes were annotated with a sequence similarity match to a protein database (29,501) and assignment to an eggNOG gene family (35,085). 29,408 genes were assigned at least one gene ontology (GO) term and 8,495 have at least one pathway assignment from KEGG. Annotation of tRNA identified 723 candidate loci located on all 23 pseudomolecules as well as in unplaced scaffolds 26, 41, 74, and 97.

rRNA characterization

The rRNA genes were annotated to identify the nucleolus organizer regions and compare the patterns to previously analyzed *Fraxinus* species. In the green ash chromosomal assembly, rDNA sequences were found only on chromosome 1 at 17.59Mb. The region includes the 5S, 5.8S, and 25S genes and part of the internal transcribed spacer but is missing the 18S subunit. An rDNA sequence with all eukaryotic subunits (5S, 18S, 5.8S, and 25S) was also identified on Scaffold_24 of the green ash assembly, which has not been placed in a chromosomal location. Scaffold_24 includes two markers from the genetic linkage map: “15188_21” from linkage group 18 and “15186_147” from linkage group 20. In both chromosome 1 and scaffold_24, the rDNA sequences are present in a single copy, indicating the tandemly repeated array of rDNA sequences was collapsed during assembly to a single copy. The tandem repeat nature of rDNA

arrays make them particularly difficult to assemble from short reads and accurately place along chromosomes (Tørresen et al., 2019).

Characterization of whole genome duplications

Previous studies have suggested the occurrence of two recent whole genome duplication (WGD) events shared across the Oleaceae family (Sollars et al., 2017; Unver et al., 2017). A recent asterid-wide phylogeny and WGD analysis placed an Oleaceae-specific WGD at around 35 million years ago and an additional older WGD shared by the Oleaceae and Carlemanniaceae families at around 78Mya (Zhang et al., 2020). Evidence of these events were assessed in *F. pennsylvanica* through pairwise synonymous site divergence (K_s) plotting (Blanc & Wolfe, 2004). *F. pennsylvanica* shares a peak at $K_s = 0.25$ with *F. excelsior* and *O. europaea* (wild olive), corroborating the presence of a recent, Oleaceae-specific whole genome duplication event (Figure 4A). Consistent with the results of Sollars et al., *F. pennsylvanica* and *F. excelsior* also share a peak at $K_s = 0.6$ with *O. europaea*; this is present along with peaks corresponding with previously reported WGD events predicted for *Daucus carota* (carrot) and *Solanum lycopersicum* (tomato) (Iorizzo et al., 2016; Song, Guo, Sun, & Wang, 2012; Tomato Genome Consortium, 2012).

To identify duplicated genomic regions likely to result from the most recent WGD in *F. pennsylvanica*, we identified all gene pairs with a K_s value of 0.25 or less (Figure 4B, Supplemental Figures S1 and S2). These are mainly found in large collinear blocks in the green ash genome (Figure 4C). All chromosomes share syntenic blocks with at least one other chromosome. Five chromosome pairs appear to originate from a single ancestral chromosome with zero to three internal rearrangements: 6 and 15, 7 and 10, 8 and 17, 9 and 13, and 11 and 14. The remaining chromosomes have a more complex synteny pattern encompassing multiple chromosomes but generally only a few large detectable rearrangements. The one exception is the distal end of chromosomes 4 and 12 which are a mosaic of small blocks. The internal synteny pattern of the WGD is consistent with the locations of 182 of the duplicated BUSCOs, while 13 are locally duplicated - meaning genes located on the same chromosome within 10 genes apart - and 33 did not appear to have been the result of WGD (Supplemental Table S2).

Comparison of *F. pennsylvanica* to *F. excelsior*

A scaffold-level assembly of the European ash (*Fraxinus excelsior*) genome (BATG v0.5) (Sollars et al., 2017) is available with 89,514 scaffolds and an N50 of 104 kbp. We ran BUSCO on the *F. excelsior* assembly and identified duplicated BUSCOs common between both assemblies. BUSCO analysis of the *F. excelsior* assembly indicated that 1436 (88.9%) complete BUSCOs were present, of which 188 (11.6%) are duplicated. We identified 127 duplicated BUSCOs shared between the two species assemblies. These shared BUSCOs consist of 57.4% of the duplicated BUSCOs identified in *F. pennsylvanica* and 67.5% of those found in *F. excelsior*. Five of the 12 locally duplicated BUSCOs in *F. pennsylvanica* were found in the list of duplicated BUSCOs in *F. excelsior*. Supplemental Table S3 provides the names of all 127 BUSCOs shared between *F. pennsylvanica* and *F. excelsior*, along with their functions.

The same repeat annotation pipeline for *F. pennsylvanica* was performed for *F. excelsior* (version BATG0.5). The two genome assemblies were largely similar in repetitive content, with

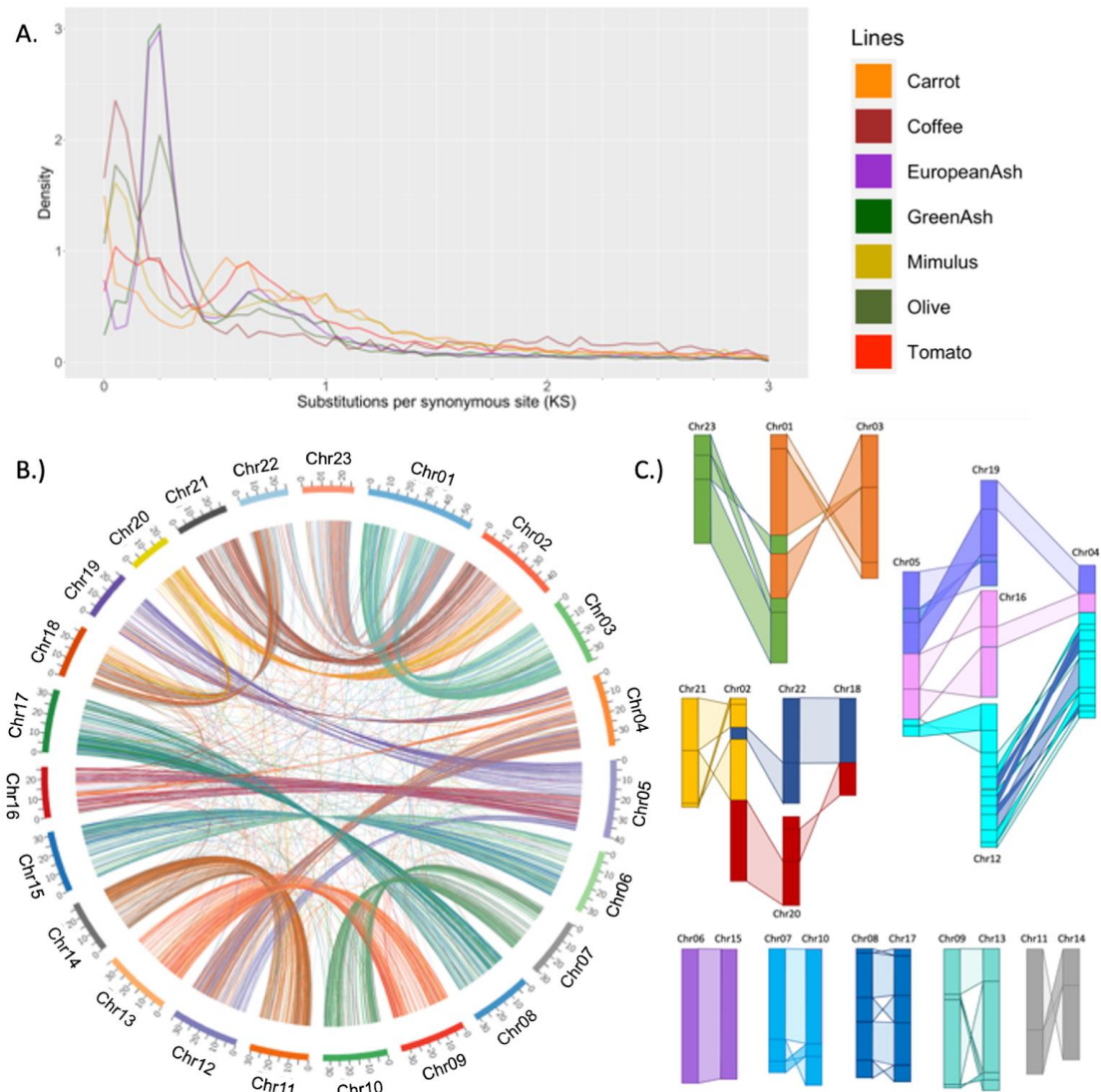


Figure 4. Characterization of Genome Duplications in *Fraxinus pennsylvanica*. (A.) Distribution of K_s values of *F. pennsylvanica* and other Asterid species. (B.) Visualization of blocks of syntenies within the green ash genome likely due to an Oleaceae specific WGD. Lines link paralogs with a K_s value of 0.25 or less. (C.) Block diagram of internal syntenies between ash chromosomes.

F. excelsior containing a total of 49.6% repetitive bases versus *F. pennsylvanica* with 49.43% (Supplementary Table S3). The two genomes share LTR elements as the most common repeat class, but *F. excelsior* has a higher percentage (29.6%) in comparison with *F. pennsylvanica* (24.53%). More of the repetitive elements in *F. pennsylvanica* were unclassified (18.5%) than in *F. excelsior* (11.61%).

While a genetic map was recently constructed for *F. pennsylvanica*, supporting the assembly used in this study, *F. excelsior* does not currently have such a map. To provide additional resources for the European ash research community, we constructed a chromosome-level assembly of *F. excelsior* using the *F. pennsylvanica* genome as a guide. Of the 89,514 input scaffolds in the *F. excelsior* assembly, 17,586 were placed on a chromosome, comprising a total of 728.3 Mbp or 84% of the BATG v0.5 assembly (Table 2). We refer to this reference-guided assembly as BATG v0.7. Of the 38,949 annotated genes identified in *F. excelsior*, 35,752 (91.8%) were placed in BATG v0.7. To build new gene models, BATG v0.7 underwent the same pipeline of repeat masking and gene annotation as *F. pennsylvanica*, yielding 41,824 gene models. BUSCO results for BATG v0.7 improved from 88.9% to 94.9% complete BUSCOs, suggesting the guided scaffolding improved the gene contiguity.

Examination of EAB resistance candidate genes

Previous evaluation of EAB response across *Fraxinus* species found that resistance arose independently within three separate phylogenetic lineages. To study signatures of convergent evolution, Kelly et al. produced a set of contig and scaffold-level assemblies for 26 taxa representing 22 species sampled from across the phylogenetic breadth of the genus (Kelly et al., 2020). A comparative analysis of these draft genomes between the resistant and susceptible species identified 53 candidate genes containing evidence of convergent evolution correlated to EAB resistance (Kelly et al., 2020). Based on sequence similarity and the reference-guided scaffolding of *F. excelsior*, 51 of the 53 candidate genes were located in the *F. pennsylvanica* annotation (Supplemental Table S4); four of these were removed from the final annotation during filtering. Based on the results described in Kelly et al., 2020, OG11720 underwent a start codon loss mutation in another *F. pennsylvanica* individual, which might account for its absence from the annotation. Functionally, all seven candidate genes associated with the phenylpropanoid biosynthesis pathway have at least one ortholog in *F. pennsylvanica* and thirteen of the fifteen candidate genes associated with herbivorous insect defence response were annotated as well. Candidate OG27080 - involved in the phenylpropanoid pathway - was predicted to be nonfunctional in *F. ornus*, but the associated mutation is not present in the associated gene in *F. pennsylvanica*. OG11720, absent from this annotation, is predicted to play a role in defense response against herbivores. Two candidates - OG32176 and OG47560 - appear to contain two copies in green ash located within the same chromosome; in the case of OG47560, the genes in green ash are within 1000 bp of each other.

Reference-guided assembly of worldwide *Fraxinus* genomes

N50 scores for the assemblies from the other *Fraxinus* species ranged from 2,665 to 103,995 bases and their BUSCO scores varied (Supplemental Table S5). To provide a new community resource, we performed *F. pennsylvanica* guided-scaffolding of 27 contig and

Table 2. Comparison of *Fraxinus excelsior* genome statistics before (version 0.5) and after (version 0.7) reference-guided assembly and reannotation.

	BATG v0.5	BATG v0.7
Total Scaffolds	89,514	71,971
Assembly Size (Mbp)	867.5	869.2
N50	103,995	30,774,430
Ns	149,164,818 (17.19%)	150,919,118 (17.36%)
Complete BUSCOs (All)	1436 (88.97%)	1532 (94.92%)
Complete BUSCOs (Single Copy)	1248 (77.32%)	1343 (83.21%)
Complete BUSCOs (Duplicated)	188 (11.65%)	189 (11.71%)
Fragmented BUSCOs	79 (4.90%)	38 (2.35%)
Missing BUSCOs	99 (6.13%)	44 (2.73%)
Total BUSCOs Searched	1614	1614

scaffold-level assemblies derived from these taxa (excluding *F. excelsior*) (Kelly et al., 2020), 10 of which we first improved with additional Illumina sequence data. This strategy varied in success, with a range of 44% to 86% of bases being placed on the pseudomolecule-level assembly (Figure 5, Supplemental Table S5). While this strategy is unable to fully anchor all bases or to identify structural variations among the genomes, similarly to *F. excelsior*, it could provide much higher gene model quality by joining neighbor scaffolds. To test this, following assembly, we masked repeats and produced new gene annotations for each new reference-guided genome version (Supplemental Table S6).

The F. pennsylvanica genome shares conserved blocks of synteny with other asterids

Protein sequences from *F. pennsylvanica*, *O. europaea*, *M. guttatus*, *C. canephora*, *S. lycopersicum* and *D. carota* were compared to each other using orthofinder. A total of 184,340 (88.7%) genes from the full set of 207,754 were placed into orthogroups based upon degree of sequence similarity. A total of 22,976 orthogroups were identified by orthofinder; of these orthogroups, 9882 had a member from all six query species, and 1376 of these were “single-copy,” meaning that all six species had exactly one copy of the gene (Supplemental Table S7). All species had 88%–91% of their genes assigned to an orthogroup, except for *S. lycopersicum* with only 84% (Table 3). The 5085 species-specific orthogroups—defined as any orthogroup containing only genes from a single species—varied in number between species, as did the total number of genes within these species-specific orthogroups (Table 3).

The links among single-copy orthologs enable structural synteny to be examined at a macroscale. Using only the orthogroups containing one gene from each of the species, we identified strong regions of synteny between the 23 chromosomes of green ash and the 23 chromosomes of wild olive (Figure S7) with most chromosomes showing one-to-one synteny. Only ash chromosomes 12 and 22 appear to have syntenic regions to two chromosomes in wild olive (Table 4). In contrast, green ash chromosomes tended to have syntenic blocks in more than one of the 11 chromosomes that comprise the genome of coffee, and these often occurred in 2-to-1 patterns (Figure S8). These patterns fit well with previous results suggesting a shared Oleaceae-specific WGD in green ash and wild olive (Unver et al., 2017) but a lack of recent WGD in coffee (Denoeud et al., 2014).

Discussion

Five ash tree species native to North America including green ash are now listed on the IUCN (The International Union for Conservation of Nature) Red List of Threatened Species as “Critically Endangered” due to the ongoing devastation of EAB (IUCN, 2017). Rare individual green ash trees with moderate EAB resistance have been documented, and these individuals are being used as a critical foundation for conservation and restoration work (Koch et al., 2012). However, the genetic or other mechanisms of this trait remain largely unknown with research efforts only just beginning. We have developed a high quality, annotated reference genome from one of these “lingering” green ash to act as a valuable research tool for understanding and leveraging the genetic component of resistance against EAB. Our chromosome-level assembly spans 757Mb with over 99% of bases anchored to the 23 haploid chromosomes. The scaffolding was based on an expanded *F. pennsylvanica* genetic map with 4,193 SNPs markers and Hi-C sequencing, a proximity ligation approach that yields highly accurate plant genomes (Michael &

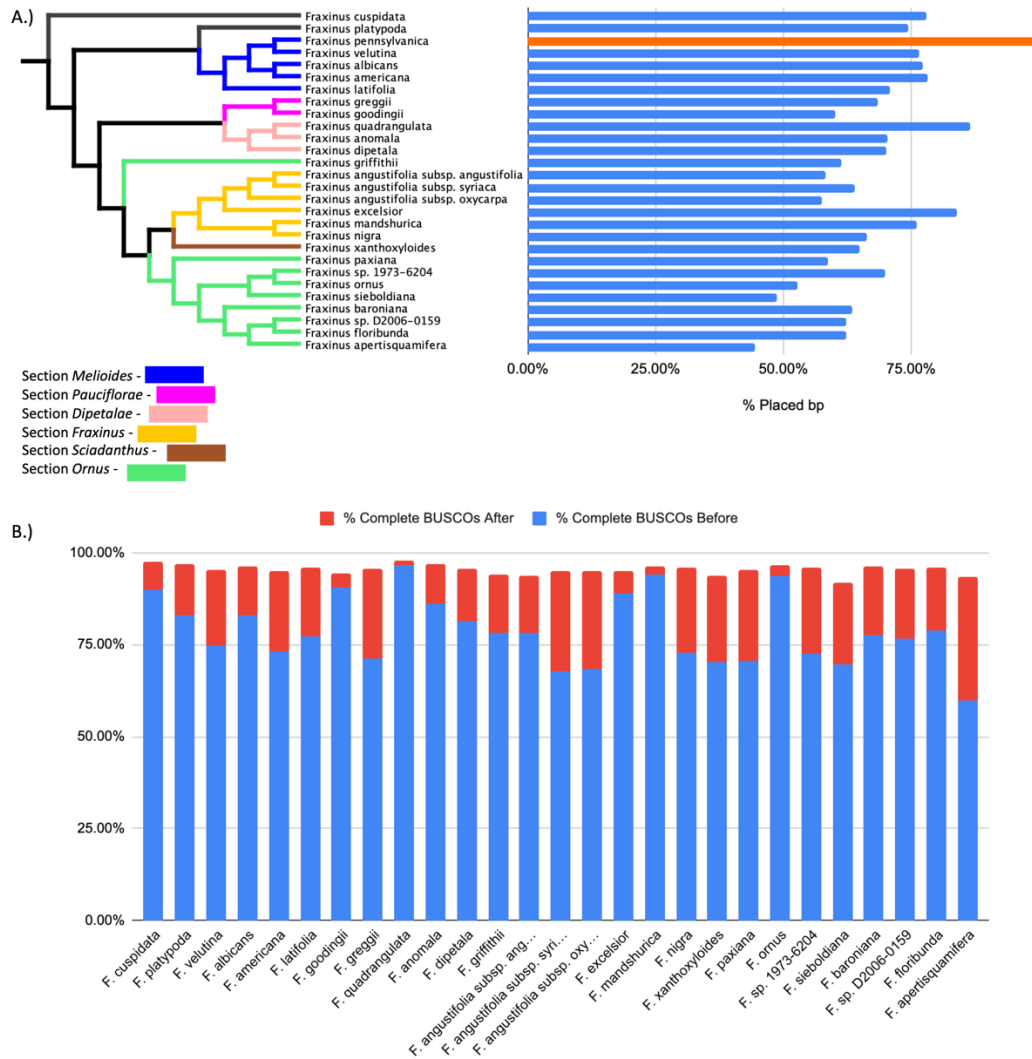


Figure 5. Reference-guided assembly of *Fraxinus* taxa genomes. A.) Pairing of a phylogenetic tree of available genomes in the *Fraxinus* genus and a bar chart illustrating percent placement of base pairs of the original genomes to the green ash genome. *Fraxinus pennsylvanica*'s placement is in orange to denote phylogenetic location relative to other species. B.) Illustration of the total complete BUSCOs identified before and after RagTag scaffolding. Abbreviations: "Subsp. ang..." = subspecies *angustifolia*, "Subsp. syri..." = subspecies *syriaca*, and "Subsp. oxy..." = subspecies *oxycarpa*. The red bars indicate additional, complete BUSCOs detected after reference-guided scaffolding.

Table 3. Orthogroup statistics of genes from six Asterids.

	<i>F. pennsylvanica</i>	<i>O. europaea</i>	<i>M. guttatus</i>	<i>C. canephora</i>	<i>S. lycopersicum</i>	<i>D. carota</i>	TOTAL
Total Genes	35,470	50,684	28,140	25,574	35,768	32,118	207,754
Number of Genes in Orthogroups	32,239 (90.90%)	45,008 (88.80%)	25,534 (90.70%)	23,062 (90.20%)	30,009 (83.90%)	28,488 (88.70%)	184,340 (88.70%)
Number of Species-Specific Orthogroups	675	939	580	530	1,071	1,290	5,085
Number of Genes in Species- Specific Orthogroups	1,766 (4.98%)	8,194 (16.17%)	2,658 (9.45%)	2,361 (9.23%)	5,083 (14.21%)	6,130 (19.09%)	26,192 (12.61%)
Number of Unassigned Genes	3,231 (9.1%)	5,676 (11.2%)	2,606 (9.3%)	2,512 (9.8%)	5,759 (16.1%)	3,630 (11.3%)	23,414 (11.3%)
Predicted Gene Duplication Events	7,292	19,050	7,904	5,761	11,282	11,239	62,528

Table 4. Ash to olive chromosomal syteny defined by single-copy orthologs. RC indicates the chromosome is in the reverse complemented orientation in the wild olive genome (Unver et al., 2017).

Ash Chromosome	Olive Chromosome
Chr01	Chr10
Chr02	Chr06 (RC)
Chr03	Chr18
Chr04	Chr13 (RC)
Chr05	Chr11 (RC)
Chr06	Chr07
Chr07	Chr03 (RC)
Chr08	Chr02 (RC)
Chr09	Chr01
Chr10	Chr15 (RC)
Chr11	Chr04
Chr12	Chr18, Chr19
Chr13	Chr12
Chr14	Chr22
Chr15	Chr14 (RC)
Chr16	Chr20 (RC)
Chr17	Chr17
Chr18	Chr16
Chr19	Chr09
Chr20	Chr21
Chr21	Chr08
Chr22	Chr05 (RC), Chr16
Chr23	Chr23

VanBuren, 2020). Assessment of the assembly and annotation with BUSCO, read alignment, and comparisons to other sequenced plant genomes indicates the genome sequence is largely complete and accurately scaffolded and annotated.

Despite the overall accuracy of the assembly, the placement of the major 35S rDNA array is still uncertain. FISH analysis identified one major rRNA locus with colocalized 35S and 5S arrays and another minor 35S locus on a different chromosome pair (Huff et al., 2022). Our assembly also has two rRNA loci; one on chromosome 1 and one on unplaced Scaffold_24, which contains markers for both Chromosome 18 and 20. Future research will be needed to confirm its exact location. Islam-Faridi *et al.* performed FISH analysis of Manchurian ash (*F. mandshurica*) and, similarly to green ash, identified two 35S (18S-5.8S-25S) rDNA and one 5S rDNA sites. They also assessed blue ash (*F. quadrangulata*) and found three 35S rDNA and one 5S rDNA site (N. Islam-Faridi, Mason, Koch, & Nelson, 2020). In both species, the 5S was colocalized with one of the major 35S sites. Fully characterizing the location of rDNA arrays across ash species could help to predict successful cross species hybridizations, which could support efforts to introgress EAB resistance and other traits across ash species.

Supporting previous findings from European ash and wild olive genomes (Sollars et al., 2017; Unver et al., 2017), we found evidence for a WGD event shared by species in the Oleaceae family and confirmed through a K_s plot and internal synteny analysis. The two copies of each original chromosome are largely syntenic within the green ash genome with some major rearrangements detected. By delineating these conserved, internal blocks of synteny within the green ash genome, we provide a strong foundation for designing genetic markers unique between the syntenic regions and future studies of gene loss and diversification across *Fraxinus* species after the WGD. In comparing the structure of the *F. pennsylvanica* genome to *O. europaea* by collinear order of orthologous genes, we observed a surprisingly high amount of structural conservation, with most chromosomes between the species having one-to-one synteny (Supplemental Figure S3). In examining a more phylogenetically distant asterid, coffee (*Coffea canephora*), more extensive structural rearrangements were common, however, large blocks of synteny were still identifiable (Supplemental Figure S4). Both coffee and olive are interesting comparators to green ash, as both have ongoing international agricultural and genetic research focused on disease and insect resistance.

With EAB threatening North American ashes, ash dieback threatening European ash species, and resistance to both found in Asian ash species, there is a strong need to develop genomic and genetic resources for the entire *Fraxinus* genus. We have used the green ash genome to enhance 28 currently available ash genomes, spanning 23 species and 6 sections (Wallander, 2012). Guided scaffolding of contig and scaffold-level genome assemblies, using the green ash genome as a reference, yielded partial chromosome-level assemblies. We have also provided updated repeat and gene annotations for the scaffolded genomes. These reference-scaffolded genomes have major limitations: they anchor only a portion of the contigs (ranging from 44% to 86%) and are unable to detect differences in genome architecture between species. However, with gene regions having a higher percent identity across species and thus gene-rich regions being preferentially anchored, we were able to show significantly improved gene annotation after scaffolding. Until independently scaffolded genomes are available for these species, this new resource could improve analysis of transcriptome experiments and improve

studies utilizing genetic markers by better contextualizing the marker's location relative to known genes and other genomic features.

REFERENCES

- Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., ... Schatz, M. C. (2019). RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, 20(1), 224.
- Babu, P., Baranwal, D. K., Harikrishna, Pal, D., Bharti, H., Joshi, P., ... Singh, A. (2020). Application of Genomics Tools in Wheat Breeding to Attain Durable Rust Resistance. *Frontiers in Plant Science*, 11, 567147.
- Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, 3(10), e3376.
- Bakys, R., Vasaitis, R., Barklund, P., Ihrmark, K., & Stenlid, J. (2009). Investigations concerning the role of *Chalara fraxinea* in declining *Fraxinus excelsior*. *Plant Pathology*, 58(2), 284–292.
- Baucom, R. S., Estill, J. C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.-M., ... Bennetzen, J. L. (2009). Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *PLoS Genetics*, 5(11), e1000732.
- Berlin, S., Trybush, S. O., Fogelqvist, J., Gyllenstrand, N., Hallingbäck, H. R., Åhman, I., ... Hanley, S. J. (2014). Genetic diversity, population structure and phenotypic variation in European *Salix viminalis* L. (Salicaceae). *Tree Genetics & Genomes*, 10(6), 1595–1610.
- Blanc, G., & Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell*, 16(7), 1667–1678.
- Borkowski, D. S., Hoban, S. M., Chatwin, W., & Romero-Severson, J. (2017). Rangewide population differentiation and population substructure in *Quercus rubra* L. *Tree Genetics & Genomes*, 13(3), 67.
- Burns, R. M., & Honkala, B. M. (1990). Silvics of North America. Vol. 1. Conifers. US Dep. Agric Agric. Handb, (654).
- Caballero, M., & Wegrzyn, J. (2019). gFACs: Gene Filtering, Analysis, and Conversion to Unify Genome Annotations Across Alignment and Gene Prediction Frameworks. *Genomics, Proteomics & Bioinformatics*, 17(3), 305–310.
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140.
- Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in Molecular Biology*, 1962, 1–14.
- Clarke, J. D. (2009). Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harbor Protocols*, 2009(3), db.prot5177.
- Coker, T. L. R., Rozsypálek, J., Edwards, A., Harwood, T. P., Butfoy, L., & Buggs, R. J. A. (2019). Estimating mortality rates of European ash (*Fraxinus excelsior*) under the ash dieback (*Hymenoscyphus fraxineus*) epidemic. *Plants, People, Planet*, 1(1), 48–58.
- Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., ... Lashermes, P. (2014). The coffee genome provides insight into the convergent

- evolution of caffeine biosynthesis. *Science*, 345(6201), 1181–1184.
- Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4(2), 359–361.
- 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(5), e19379.
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238.
- Endelman, J. B., & Plomion, C. (2014). LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics*, 30(11), 1623–1624.
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14(8), 2611–2620.
- Gaur, R., Azam, S., Jeena, G., Khan, A. W., Choudhary, S., Jain, M., ... Bhatia, S. (2012). High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (*Cicer arietinum* L.). *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 19(5), 357–373.
- Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS One*, 9(2), e90346.
- Grattapaglia, D., & Sederoff, R. (1994). Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics*, 137(4), 1121–1137.
- Grattapaglia, D., Silva-Junior, O. B., Resende, R. T., Cappa, E. P., Müller, B. S. F., Tan, B., ... El-Kassaby, Y. A. (2018). Quantitative Genetics and Genomics Converge to Accelerate Forest Tree Breeding. *Frontiers in Plant Science*, 9, 1693.
- Gucker, C. L. (2005). *Fraxinus pennsylvanica*. USDA Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory.
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075.
- Hart, A. J., Ginzburg, S., Xu, M. (sam), Fisher, C. R., Rahmatpour, N., Mitton, J. B., ... Wegrzyn, J. L. (2020). EnTAP : Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Molecular Ecology Resources*, 20(2), 591–604.
- Hellsten, U., Wright, K. M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S. R., ... Rokhsar, D. S. (2013). Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48), 19478–19482.
- Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, 33(1), 54–78.
- Hoban, S. M., Borkowski, D. S., Brosi, S. L., McCLEARY, T. S., Thompson, L. M., McLACHLAN, J. S., ... Romero-Severson, J. (2010). Range-wide distribution of genetic diversity in the North American tree *Juglans cinerea*: a product of range

- shifts, not ecological marginality or recent population decline. *Molecular Ecology*, 19(22), 4876–4891.
- Hoff, K. J., Lange, S., Lomsadze, A., & Borodovsky, M. (2016). BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 32(5), 767–769.
- Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER. *Methods in Molecular Biology*, 1962, 65–95.
- Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., ... Simon, P. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics*, 48(6), 657–666.
- Islam-Faridi, M. N., Nelson, C. D., DiFazio, S. P., Gunter, L. E., & Tuskan, G. A. (2009). Cytogenetic analysis of *Populus trichocarpa*--ribosomal DNA, telomere repeat sequence, and marker-selected BACs. *Cytogenetic and Genome Research*, 125(1), 74–80.
- Islam-Faridi, N., Mason, M. E., Koch, J. L., & Nelson, C. D. (2020). Cytogenetics of *Fraxinus mandshurica* and *F. quadrangulata*: ploidy determination and rDNA analysis. *Tree Genetics & Genomes*, 16(1), 26.
- Islam-Faridi, N., Sakhanokho, H. F., & Dana Nelson, C. (2020). New chromosome number and cyto-molecular characterization of the African Baobab (*Adansonia digitata* L.) - ‘The Tree of Life’. *Scientific Reports*, 10(1), 13174.
- IUCN. (2017). *Fraxinus nigra*: Jerome, D., Westwood, M., Oldfield, S. & Romero-Severson, J [Data set]. *IUCN Red List of Threatened Species*. IUCN. Retrieved from <https://doi.org/10.2305/iucn.uk.2017-2.rlts.t61918683a61918721.en>
- Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14), 1801–1806.
- Jewell, D. C., & Islam-Faridi, N. (1994). A Technique for Somatic Chromosome Preparation and C-banding of Maize. In W. V. Freeling M. (Ed.), *The Maize Handbook* (pp. 484–493). Springer, New York, NY: Springer Lab Manuals.
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780.
- Kelly, L. J., Plumb, W. J., Carey, D. W., Mason, M. E., Cooper, E. D., Crowther, W., ... Buggs, R. J. A. (2020). Convergent molecular evolution among ash species resistant to the emerald ash borer. *Nature Ecology & Evolution*, 4(8), 1116–1128.
- Knight, K. S., Brown, J. P., & Long, R. P. (2013). Factors affecting the survival of ash (*Fraxinus* spp.) trees infested by emerald ash borer (*Agrilus planipennis*). *Biological Invasions*, 15(2), 371–383.
- Koch, J. L., Carey, D. W., Knight, K. S., Poland, T., Herms, D. A., & Mason, M. E. (2012). Breeding strategies for the development of emerald ash borer-resistant North American ash. In In: *Snieszko, Richard A.; Yanchuk, Alvin D.; Kliejunas, John T.; Palmieri, Katharine M.; Alexander, Janice M.; Frankel, Susan J., tech. coords. Proceedings of the fourth international workshop on the genetics of host-parasite*

- interactions in forestry: Disease and insect resistance in forest trees. Gen. Tech. Rep. PSW-GTR-240. Albany, CA: Pacific Southwest Research Station, Forest Service, US Department of Agriculture. pp. 235-239. (Vol. 240, pp. 235–239).*
- Kovacs, K. F., Haight, R. G., McCullough, D. G., Mercader, R. J., Siegert, N. W., & Liebhold, A. M. (2010). Cost of potential emerald ash borer damage in U.S. communities, 2009–2019. *Ecological Economics*, 69(3), 569–578.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639–1645.
- Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9), 3100–3108.
- Lane, T., Best, T., Zembower, N., Davitt, J., Henry, N., Xu, Y., ... Staton, M. E. (2016). The green ash transcriptome and identification of genes responding to abiotic and biotic stresses. *BMC Genomics*, 17(1), 702.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5), 589–595.
- Lind, J. F., & Gailing, O. (2013). Genetic structure of *Quercus rubra* L. and *Quercus ellipsoidalis* E. J. Hill populations at gene-based EST-SSR and nuclear SSR markers. *Tree Genetics & Genomes*, 9(3), 707–722.
- Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., ... Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18), 2397–2399.
- Löve, Á. (1982). IOPB Chromosome Number Reports LXXIV. *Taxon*, 31(1), 119–128.
- Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5), 955–964.
- Marroni, F., Pinosio, S., Di Centa, E., Jurman, I., Boerjan, W., Felice, N., ... Morgante, M. (2011). Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: towards next-generation Ecotilling. *The Plant Journal: For Cell and Molecular Biology*, 67(4), 736–745.
- Michael, T. P., & VanBuren, R. (2020). Building near-complete plant genomes. *Current Opinion in Plant Biology*, 54, 26–33.
- Neale, D. B., & Kremer, A. (2011). Forest tree genomics: growing resources and applications. *Nature Reviews. Genetics*, 12(2), 111–122.
- Nielsen, L. R., McKinney, L. V., Hietala, A. M., & Kjær, E. D. (2017). The susceptibility of Asian, European and North American *Fraxinus* species to the ash dieback pathogen *Hymenoscyphus fraxineus* reflects their phylogenetic history. *European Journal of Forest Research*, 136(1), 59–73.
- Ouellette, L. A., Reid, R. W., Blanchard, S. G., & Brouwer, C. R. (2018). LinkageMapView-rendering high-resolution linkage and QTL maps. *Bioinformatics*, 34(2), 306–307.
- Pérez-de-Castro, A. M., Vilanova, S., Cañizares, J., Pascual, L., Blanca, J. M., Díez, M. J.,

- ... Picó, B. (2012). Application of genomic tools in plant breeding. *Current Genomics*, 13(3), 179–195.
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*. Retrieved from <https://doi.org/10.1371/journal.pone.0037135>
- Plomion, C., Bastien, C., Bogeat-Triboulot, M.-B., Bouffier, L., Déjardin, A., Duplessis, S., ... Vacher, C. (2016). Forest tree genomics: 10 achievements from the past 10 years and future prospects. *Annals of Forest Science*, 73(1), 77–103.
- Poland, T. M., & McCullough, D. G. (2006). Emerald ash borer: invasion of the urban forest and the threat to North America's ash resource. *Journal of Forestry*, 104(3), 118–124.
- Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., ... Tangphatsornruang, S. (2015). Construction of a high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). *Frontiers in Plant Science*, 6, 367.
- Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in Genetics*, 4, 98.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575.
- Rebek, E. J., Herms, D. A., & Smitley, D. R. (2008). Interspecific variation in resistance to emerald ash borer (Coleoptera: Buprestidae) among North American and Asian ash (*Fraxinus* spp.). *Environmental Entomology*, 37(1), 242–246.
- Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology*, 1962, 227–245.
- Siebert, N. W., McCullough, D. G., Liebhold, A. M., & Telewski, F. W. (2014). Dendrochronological reconstruction of the epicentre and early spread of emerald ash borer in North America. *Diversity and Distributions*, 20(7), 847–858.
- Smit, A. F. A., Hubley, R., & Green, P. (2015a). RepeatMasker Open-4.0. 2013--2015.
- Smit, A. F. A., Hubley, R., & Green, P. (2015b). RepeatModeler Open-1.0. 2008--2015. *Seattle, USA: Institute for Systems Biology*. Available from: [Httpwww.Repeatmasker.Org](http://www.Repeatmasker.Org), Last Accessed May, 1, 2018.
- Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., ... Buggs, R. J. A. (2017). Genome sequence and genetic diversity of European ash trees. *Nature*, 541(7636), 212–216.
- Song, C., Guo, J., Sun, W., & Wang, Y. (2012). Whole genome duplication of intra- and inter-chromosomes in the tomato genome. *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, 39(7), 361–368.
- Staton, M., Addo-Quaye, C., Cannon, N., Yu, J., Zhebentyayeva, T., Huff, M., ... Carlson, J. E. (2020). 'A reference genome assembly and adaptive trait analysis of *Castanea*

- mollissima “Vanuxem,” a source of resistance to chestnut blight in restoration breeding’. *Tree Genetics & Genomes*, 16(4), 57.
- Steiner, K. C., Graboski, L. E., Knight, K. S., Koch, J. L., & Mason, M. E. (2019). Genetic, spatial, and temporal aspects of decline and mortality in a *Fraxinus* provenance test following invasion by the emerald ash borer. *Biological Invasions*, 21(11), 3439–3450.
- Steiner, K. C., Williams, M. W., DeHayes, D. H., & Hall, R. B. (1988). Juvenile performance in a range-wide provenance test of *Fraxinus pennsylvanica* Marsh. *Silvae Genetica*, 37(3-4), 104–111.
- Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., ... Zhang, Z. (2016). GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *The Plant Genome*, 9(2). Retrieved from <https://doi.org/10.3835/plantgenome2015.11.0120>
- Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400), 635–641.
- Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., ... Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 47(21), 10994–11006.
- Unver, T., Wu, Z., Sterck, L., & Turktas, M. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 114(44), E9413–E9422.
- Van Ooijen, J. W. (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. *Kyazma BV, Wageningen*, 33(10.1371).
- Wallander, E. (2012). Systematics and floral evolution in *Fraxinus* (Oleaceae). *Belgische Dendrologie Belge*, 2012, 39–58.
- Wang, Q., Tian, F., Pan, Y., Buckler, E. S., & Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PloS One*, 9(9), e107684.
- Wu, D., Koch, J., Coggeshall, M., & Carlson, J. (2019). The first genetic linkage map for *Fraxinus pennsylvanica* and syntenic relationships with four related species. *Plant Molecular Biology*, 99(3), 251–264.
- Zhang, C., Zhang, T., Luebert, F., Xiang, Y., Huang, C.-H., Hu, Y., ... Ma, H. (2020). Asterid Phylogenomics/Phylotranscriptomics Uncover Morphological Evolutionary Histories and Support Phylogenetic Placement for Numerous Whole-Genome Duplications. *Molecular Biology and Evolution*, 37(11), 3188–3210.

APPENDIX

Supplemental Methods

Genome Annotation:

Repetitive elements in the *Fraxinus pennsylvanica* genome were identified with RepeatModeler v1.0.11 (Smit, Hubley, & Green, 2015b). These repeat sequences were combined with RepBase and repetitive elements previously identified in *Fraxinus excelsior* (Sollars et al., 2017), then provided to RepeatMasker v4.0.9 (species “eudicotyledon”) (Smit, Hubley, & Green, 2015a). Gene annotations were predicted using BRAKER v2.1.5 using RNA sequencing data from multiple tissues and stress conditions (Lane et al., 2016) (25.2 Gb) and 50,841 gene models from *Fraxinus excelsior* (Sollars et al., 2017) (Katharina J. Hoff, Lomsadze, Borodovsky, & Stanke, 2019; K. J. Hoff, Lange, Lomsadze, & Borodovsky, 2016). Genes were filtered by gFACs v1.1.2 for gene structure (Caballero & Wegrzyn, 2019). Genes were filtered if they met these criteria: CDS size less than 300 bp, exon and intron sizes of less than 9 bp, in-frame stop codons, and/or lack of start or stop codons. Only the longest unique gene models were kept to represent each locus. Genes were next filtered by EnTAP v0.9.1 based on predicted function (Hart et al., 2020). For EnTAP, genes were compared against the Swissprot and TrEMBL plant protein databases and the EggNOG gene family database. Genes kept by EnTAP were again filtered by gFACs, yielding the final list of gene models. Transfer RNA (tRNA) were predicted using tRNAscan version 2.0 on default settings (Chan & Lowe, 2019; Lowe & Eddy, 1997). Ribosomal DNA (rDNA) locations were predicted using the program RNAmmer v1.2 and aligning known rRNA sequences from *Olea europaea* and *F. excelsior* to the genome assembly (Lagesen et al., 2007).

***Fraxinus* spp. Reference-Guided Genome Scaffolding:**

BUSCO and RepeatMasker were run on the BAT0.5 *F. excelsior* assembly (Sollars et al., 2017) using the same parameters previously described (Seppey et al., 2019; Smit et al., 2015a). We downloaded 27 other publicly available *Fraxinus* scaffold and contig level assemblies (Kelly et al., 2020). For *F. angustifolia* subsp. *angustifolia*, *F. dipetala*, *F. latifolia*, *F. mandshurica*, *F. ornus*, *F. paxiana*, *F. quadrangulata*, and *F. sieboldiana*, new 800bp insert size library data was sequenced with the Illumina HiSeq 4000 platform by the Centre for Genomic Research at the University of Liverpool. These ten species were all de novo assembled using the same methods as Kelly et al (2020). An additional species, *F. apertisquamifera*, was also sequenced and assembled following the same methods of Kelly et al., 2020. Plant material for this assembly was collected from accession 20071420-L, provided by the Royal Botanic Garden Edinburgh. For all 28 assemblies, we used RagTag v1.0.1 to align the scaffolds of each assembly to the chromosomes of *F. pennsylvanica* and join them to produce chromosome-scale assemblies (Alonge et al., 2019). Following the same methods as above, we masked repetitive elements from the genomes with RepeatMasker with the repetitive elements that were identified in *F. pennsylvanica* as the repeat database (Smit et al., 2015a). Gene annotations were predicted utilizing BRAKER2; this time, gene models from *F. pennsylvanica*, *F. excelsior*, and *Olea europaea* were used as input (Katharina J. Hoff et al., 2019; K. J. Hoff et al., 2016). The output of BRAKER2 was filtered with a single round of gFACs using identical parameters to the methods above (Caballero & Wegrzyn, 2019).

CHAPTER IV:
IDENTIFICATION OF MARKERS OF SEX DETERMINATION IN
FRAXINUS PENNSYLVANICA

Abstract

Green ash (*Fraxinus pennsylvanica*) is dioecious, with distinct male and female individuals. At present, the mechanisms of dioecy in green ash and other species in the *Melioides* section of *Fraxinus* are unknown, and trees cannot be sexed until after they first flower. Current efforts to breed green ash with enhanced resistance to the invasive emerald ash borer (EAB; *Agrilus planipennis*) would be improved by finding markers of sex determination, allowing researchers to sex trees at the seedling stage. In this study, we performed a double-digest genotyping-by-sequencing (ddGBS) experiment to identify these markers in the green ash genome. Samples were collected from ash species in section *Melioides* found throughout their native range. Using a large sample size of 192 trees, the ddGBS data yielded 4,993,034 high quality SNPs. While no SNPs were perfectly associated with male versus female samples, we found candidate SNPs of interest in a region of chromosome 8 with preferential detection in female samples over male samples. These results are consistent with an XY sex-determining system, where reads in this region are more often detected in female samples because females have two copies of this haplotype. Annotation of nearby genes included a phospholipase A2-like protein, which has been shown to be critical for pollen formation in other plants. However, the structure of ash flowers (the complete absence of stamens in females and carpels in males) also suggests an unidentified role of MADS-box genes in ash sexual dimorphism. Our results demonstrate the limits of using ddGBS in exploring sex-determination in green ash but also suggest a potential region of interest for future studies.

Introduction

Fraxinus pennsylvanica Marsh., or green ash (also called red ash, swamp ash, or water ash), is the most widely distributed ash species in North America and a species of high economic importance (Kennedy, 1949). In recent years, it has come under threat from an invasive, wood-boring beetle known as Emerald Ash Borer (EAB; *Agrilus planipennis*), which has an estimated economic impact of \$10.7 billion, including the replacement of 17 million ornamental ash (Kovacs et al., 2010). Native to eastern Asia, EAB was purportedly introduced in North America in the early 1990s, but was first identified near Detroit, Michigan in 2002 (Siegert et al., 2014). To colonize new hosts, EAB lays its eggs on the bark of ash trees. After hatching, nascent larvae chew through the bark and feed on the phloem and cambium, disrupting the transport of water and nutrients (Poland & McCullough, 2006). While most ash trees die within 6 years of EAB infestation, a small percentage of individuals remain healthy in highly infested forests. In an effort to combat the loss of green ash across its native range, the US Forest Service has established an ongoing breeding program to develop ash with improved EAB resistance by propagating and breeding so-called “lingering ash” with heritable tolerance to EAB (Koch et al., 2012).

However, breeding efforts in ash are delayed by the substantial time required to reach sexual maturity. Thus, efforts to make breeding more efficient are of particular interest. One way to ensure that seedling plots will produce viable offspring is by identifying markers of sex determination. Sex determination mechanisms in ash species range from hermaphroditic systems to completely dioecious systems (Wallander, 2012). As a member of the *Melioides* section of the *Fraxinus* genus, green ash has a dioecious system with complete segregation of male and female

flowers between individuals. In contrast, European ash (*F. excelsior*; section *Fraxinus*) is subdioecious with a range of sexual phenotypes from all female to hermaphrodite to all male flowers. However, individuals function primarily as male or female due to a self-incompatibility (SI) system (Saumitou-Laprade et al., 2018). The same diallelic SI system is shared with other Oleaceae species, including *F. ornus*, *Phillyrea angustifolia*, and *Olea europaea* (Mariotti et al., 2019; Vernet et al., 2016). It is currently accepted that the *Melioides* section of *Fraxinus* acquired the dioecy trait from a common ancestor after diverging from another, non-dioecious species (Wallander, 2008, 2012). Nevertheless, the exact mechanism that drives sex determination in green ash is currently unknown. Moreover, the heterogametic sex is also unknown, making it impossible to sex trees prior to flowering, which can occur anywhere from 10 to 25 years after germination (Hausman et al., 2014; Schopmeyer & Others, 1974). Breeders report that some individuals appear to be sexually incompatible (Koch et al., 2010), suggesting green ash may have an SI system, but if the SI system is the same as that previously described for other species or its interaction with dioecy have yet to be determined. Thus, efforts to breed EAB resistant individuals would be made more efficient by identifying markers of sex determination and SI group at the seedling stage.

While information on dioecy in green ash is currently limited, studies in other dioecious plant species provide insight into the overall trends of dioecy that will help identify the mechanisms in green ash. Dioecy has evolved independently across numerous distinct plant lineages (Charlesworth, 2002), but a generalized mechanism of dioecy becomes clear when comparing species with well-characterized sex-determination mechanisms (Leite Montalvão et al., 2020). In plants, sex-determining regions (SDRs) of chromosomes contain one or more genes that differ between the male and female genotypes. Unlike animals, where sex chromosomes are often heteromorphic, the chromosomes share broad scale homology except at the SDR. This SDR varies in size, but is always a region lacking recombination (Renner & Müller, 2021). Due to this lack of recombination, the SDR evolves into two haplotypes that share little homology. For example, sex determination may be driven by two genes [e.g., asparagus (Harkess et al., 2020), grape (Massonnet et al., 2020), and kiwifruit (Akagi et al., 2018)] or by a single gene [e.g., poplar (Leite Montalvão et al., 2022; Müller et al., 2020)].

In the traditional model of the evolution of dioecy, it occurs as a two-gene system, with an initial, recessive male-sterility mutation followed by a dominant female-sterility mutation (Charlesworth, 2002; Charlesworth & Charlesworth, 1978). These mutations affect a pair of genes with strong linkage, meaning that recombination of the SDR is extremely rare or impossible. Genomics studies of dioecious plants have found support for the two gene model but also exceptions, including single gene systems. The most commonly observed form of a two gene-system sees one gene act as a promoter of one sex, and the other acts as a suppressor of the opposite sex. In the example of asparagus, the Y-specific *SUPPRESSOR OF FEMALE FUNCTION* (*SOFF*) gene prevents development of female organs, and *DEFECTIVE IN TAPETUM DEVELOPMENT AND FUNCTION 1* (*TDF1*) promotes development of male organs (Harkess et al., 2020). Knockout of only *SOFF* converts males to hermaphrodites, knockout of only *TDF1* converts males to asexual neuter flowers, and knockout of both genes converts males to females.

In a single gene system, sex is determined by the presence or absence of a gene, which may result in transitions between XY and ZW systems. In poplar, an ortholog of *ARABIDOPSIS*

RESPONSE REGULATOR 17 (*ARR17*) serves as the sex-determining gene. Poplar species with a ZW sex-determining system (*Populus alba*) lack *ARR17* in the Z haplotype, and females carry a single copy of *ARR17* in the W haplotype. Poplar species with an XY sex-determining system (*P. tremula* and *P. trichocarpa*) contain *ARR17* in both X and Y haplotypes, but the Y haplotype contains partial duplicates of *ARR17*, which produce small RNAs that silence *ARR17* expression. As evidence, knockouts of *ARR17* with CRISPR-Cas9-induced mutations converted female *P. tremula* into males (Müller et al., 2020). Gene expression analysis of *ARR17* knockouts determined that *ARR17* represses the expression of B-class MADS-box genes responsible for stamen development (Leite Montalvão et al., 2022). We do not expect green ash to share these exact mechanisms, but we expect the mechanism of sex-determination in green ash to be similarly driven by alterations in a MADS-box-associated pathway.

The goal of this project was to identify markers of sex determination in the green ash genome. We performed a double-digest genotyping-by-sequencing (ddGBS) analysis of 192 ash individuals to identify regions of the genome that significantly correlated with the male and female phenotypes. Compared to whole genome sequencing (WGS), which aims to cover as much of the genome as possible, GBS is a reduced-representation sequencing (RRS) strategy that focuses on capturing higher coverage in a lower percentage of the genome. In olive, this analysis has been shown to be effective for comparing genetic diversity between samples as a cheaper alternative to WGS (Friel et al., 2021). In amphibians, a clade of animals with homomorphic sex chromosomes, GBS experiments have been used to successfully identify SDRs (Luo et al., 2020; Yang et al., 2022). Thus, due to the homomorphic architecture of most plant chromosomes, we deemed GBS to be a reasonable first approach to identifying genomic regions associated with the sex of an organism. Consequently, our results identified a region of interest on Chromosome 8 in the recently published green ash genome (Huff et al., 2022). Further study with more comprehensive genomic techniques is needed for confirmation of this locus.

Materials and Methods

Sample collection

Samples were collected for DNA extraction from the leaves of green ash throughout its native range, though the majority of samples were collected in East Tennessee due to proximity. To confirm the sex of collected material, collecting was performed during the spring, when male and female flowers were present. No samples were collected from trees that did not flower or did not display remnant inflorescences from previous flowerings. Samples were confirmed as green ash by assessing bud scar and samara morphology (of female trees). Additionally, samples of white ash (*F. americana*) and Biltmore ash (*F. biltmoreana*) were collected as we expect a shared sex determining system due to recent evolutionary divergence. Samples were initially dried following collection and intermittently frozen using liquid nitrogen and homogenized four times using 5 silica beads in a Fisherbrand Bead Mill 24 Homogenizer (Thermo Fisher Scientific, Pittsburgh, PA, USA) before DNA extraction. Our DNA extraction protocol followed the E.Z.N.A.® Plant DNA DS Mini Kit protocols for dried samples (Omega Bio-Tek Inc., Norcross, GA, USA) with the following modifications: centrifugations speeds were universally set to 14,400 rpm, centrifugation time was increased by 1 minute per cycle, and elution buffer was heated to 65 °C and allowed to incubate at room temperature on the HiBind® DNA Mini Column for 5 minutes before each centrifugation cycle. After testing multiple combinations of

REs on test data collected from green ash leaves, PstI was selected as the first enzyme for this double-digest experiment, followed by MseI as the second enzyme, as this combination provided the best fragment size and coverage. GBS libraries for Illumina short-read sequencing were prepared and paired-end sequencing (150 bp) was performed on the Illumina NovaSeq 6000 sequencing platform (Illumina, San Diego, CA, USA) at the University of Wisconsin Madison Genomics Core (Madison, WI, USA).

Genotyping-by-sequencing read processing and variant calling

Reads were demultiplexed using SABRE, version 1.0 (<https://github.com/najoshi/sabre>), and trimmed of adapter content using Skewer, version 0.2.2 (Jiang et al., 2014). We allowed a mean quality score of 30 or higher before trimming, and a minimum read length of 30 bp after trimming. Following read trimming, we aligned reads to the chromosome-scale genome assembly (Huff et al., 2022) using the Burrows-Wheeler Aligner tool (BWA), version 0.7.17 (Li & Durbin, 2010). The SAM output of BWA was converted to BAM format using samtools, version 1.10 (Danecek et al., 2021; Li et al., 2009), and which was also used throughout the project to obtain index files.

Read groups were created using Sambamba (Tarasov et al., 2015), and variant calling was performed using FreeBayes, version 1.3.1, on the individual chromosomes in the green ash genome for each sample (Garrison & Marth, 2012). We did not filter out InDels in this analysis, choosing instead to include every type of variant in an attempt to capture larger structural changes. We created a combined VCF file using BCFtools, version 1.10.2, to concatenate the individual VCF files into a combined file for each chromosome, followed by merging these chromosome files (Danecek et al., 2021). We filtered our samples using VCFtools (Danecek et al., 2011), selecting only variants with a minor allele frequency (MAF) of 0.05 or more. We reran this filtering step with different “max missing” values to identify SNPs found in 50%, 70%, and 90% of our samples.

Variants strongly associated with male and female phenotypes were called using PLINK, version v1.90b6.21 (Purcell et al., 2007). PLINK was run without parental information. We also allowed PLINK to include excess chromosomes, in this case referring to unplaced scaffolds beyond the 23 chromosome-scale scaffolds (Chapter III). PLINK association files were visualized using qqman version 0.1.8 (Turner, 2014).

We next created a file to test for regions with significant presence/absence variation correlated to sex. We altered the vcf file to reflect locus presence or absence. At each locus, samples with genotype information were encoded as “present” by the use of 0/0 and samples without genotype information were encoded as “missing” by 1/1. PLINK was run on this new file with the same approach previously described, without filtering for “max missing” values.

Regions of interest were visualized using the Integrative Genomics Viewer (IGV) web viewer, version 1.12.9 (Robinson et al., 2011), and annotations for genes of interest were obtained from our green ash genome assembly (Huff et al., 2022).

Results

Sample population

A total of 263 ash floral samples were collected, and 230 were considered viable for sequencing. Samples were collected from Kentucky, Mississippi, Nebraska, North Carolina, Ohio, South Carolina, and Tennessee (Figure 6). We sequenced 192 of these samples, consisting of 170 green ash samples, 20 white ash (*F. americana*) samples, and 2 Biltmore ash (*F. biltmoreana*) samples, which represented the most comprehensive geographic distribution of individuals possible with our dataset. We collected 99 samples from female trees and 92 samples from male trees, confirmed by identification and collection of floral material (Figure 7).

Genotyping-by-sequencing SNPs were not associated with sex

We first opted to use the GBS data to identify SNPs and test for association with sex. While SDRs are usually large regions with significant structural variation, we hypothesized that neighboring variants may be in linkage disequilibrium with the SDR region. A total of 4,993,034 variants were called using FreeBayes, 60,359 of which had an MAF of at least 0.05. This number was reduced to 5,837 when filtering out genotypes called in 50% or more samples, 1,953 for variants called in 70% or more samples, and 466 for variants in 90% or more samples. Each variant set was tested for association with sex using PLINK, which identified only 1 significant variant ($p \leq 0.05$) before any genotype filtering, and no significant variants after filtering for variants called in 50% or more of samples. For the significant SNP, despite the general association of homozygous reference allele with females and homozygous alternate allele with males, multiple samples did not fit this pattern (Table 5).

A candidate presence/absence variation region identified on Chromosome 8

By altering the VCF file to test for presence/absence of genotype calls instead of the specific allelic variants, we identified a region of interest along Chromosome 8 spanning from 17,168,000 to 17,170,000 bp, with additional SNPs of interest past this point (Figures 8 and 9, Table 6). Of the 20 variants in this region, 8 were considered significant, and most of these variants were considered the most significant across all variants. Outside of two outliers, variants were more likely to be present in female samples and absent in male samples (Table 6), and even in these outliers the variants were observed most often in female samples. While we observed variants that were called in male samples, they were far less abundant in comparison to the overall number of male samples compared to the female samples. The haplotype phasing does not support a heterozygous SNP in these called male SNPs, but the bias towards its absence in males indicates this is a region of interest.

Annotation of genes located in close proximity to the region of interest

We confirmed the presence of two genes in the current green ash annotation surrounding this region of interest (Figure 9A), Fp_g19699 and Fp_g19700. While functional annotation of Fp_g19699 indicated it did not have any known protein functions (Table 7), Fp_g19700 was annotated as a homolog of phospholipase A2-alpha (PLA2). On a broad scale, PLA family enzymes hydrolyze phospholipids to release acyl groups, which are then used to produce free

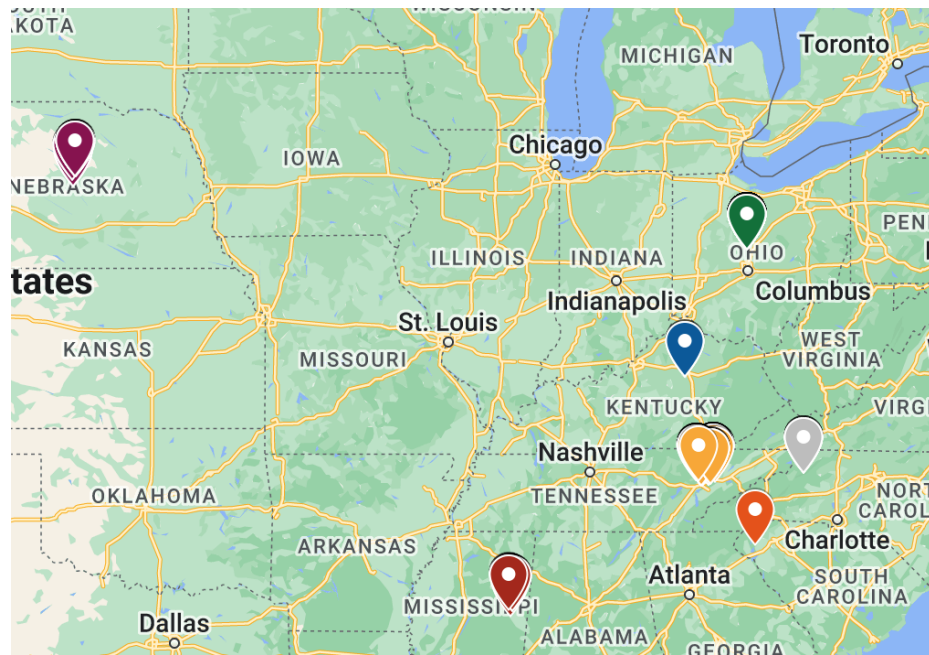


Figure 6. Map of *Fraxinus* sp. Samples Collected for Study. Coordinates obtained from TreeSnap collections information.

A.)



B.)



Figure 7. Flower structure of male (A) and female (B) *Fraxinus pennsylvanica* flowers.
Samples collected from Turkey Creek, Knoxville, TN, USA.

Table 5. Significant SNPs from FreeBayes lack expected genotype segregation for a sex-linked marker. SNPs were obtained from FreeBayes, version 1.3.1, and called as significant by PLINK, version 1.90b6.21.

Chromosome	Position	<i>p</i> -value	Adj. <i>p</i> -value	Called (%)	Uncalled (%)	Male Samples	Female Samples	0/0 (M,F)	0/1 (M,F)	1/1 (M,F)
Chr09	12,880,363	3.901e-08	0.002352	30 (15.6%)	162 (84.4%)	21	9	17,0	2,2	2,7

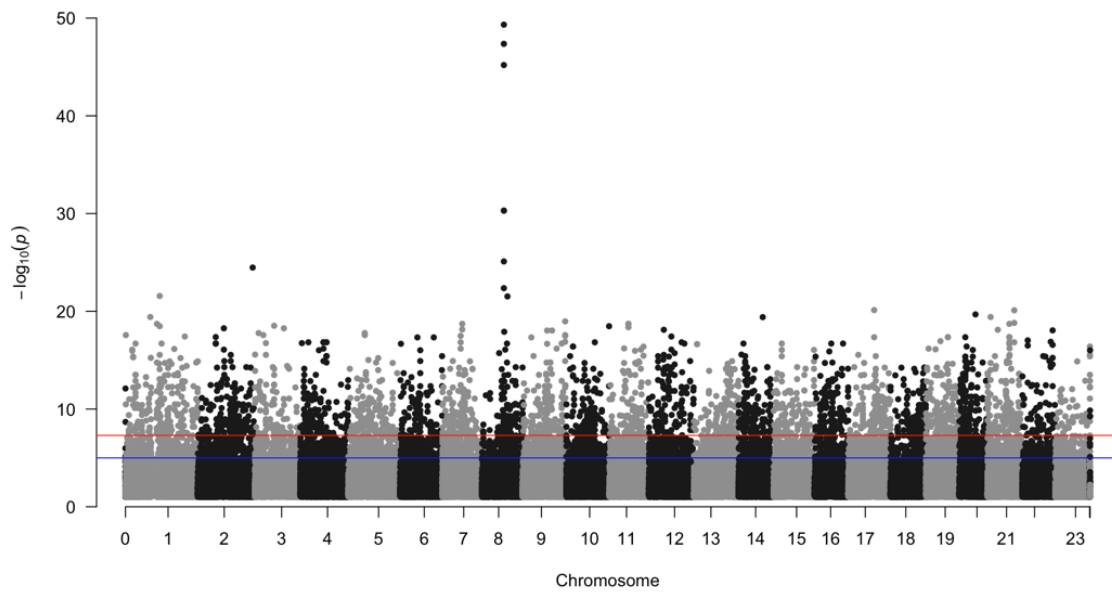
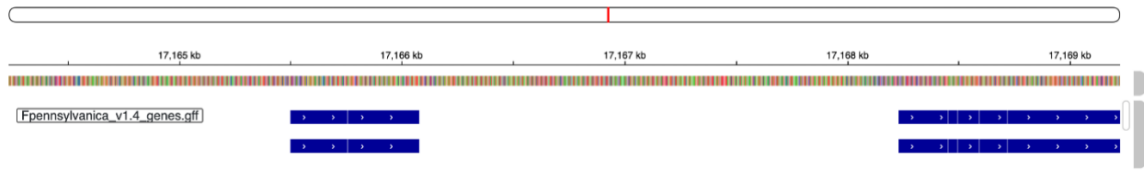


Figure 8. Manhattan plot of significantly present/absent variants obtained from FreeBayes. Blue lines represent the suggestive cut-off for each plot, and red lines represent the genome-wide significance cut-off for each plot. Plot only includes variants where $p < 0.10$.

A.)



B.)

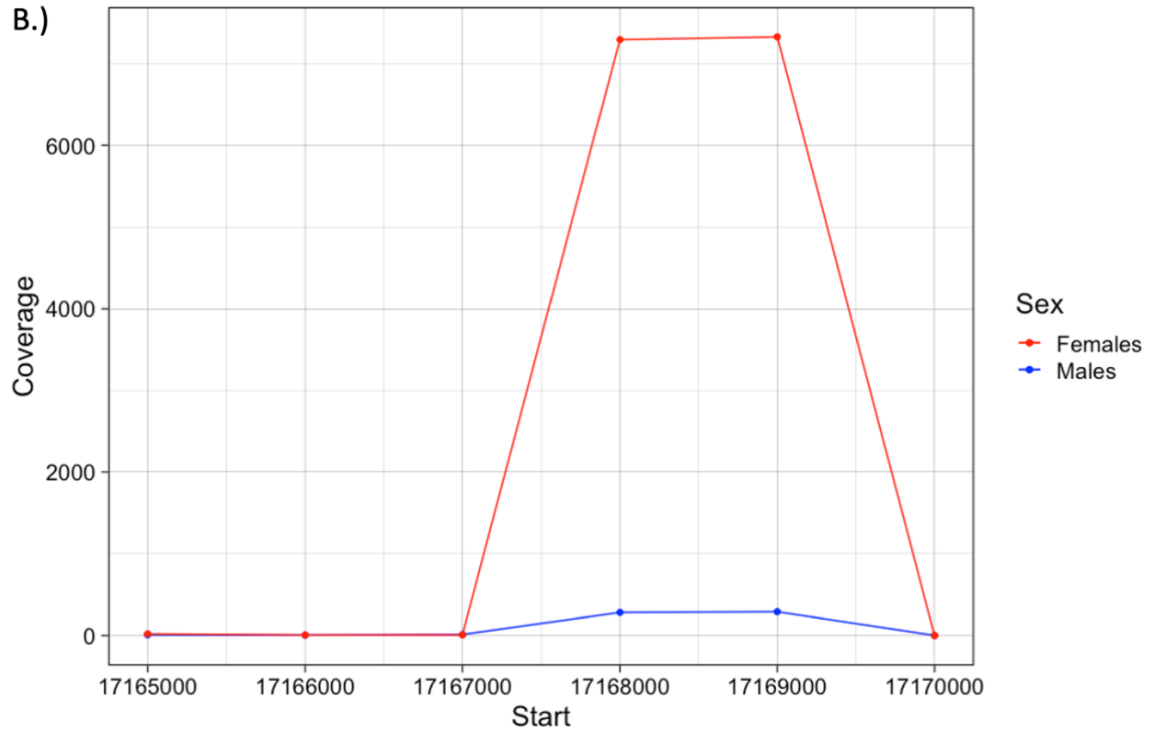


Figure 9. A region along Chromosome 8 with preferential presence in females of *Fraxinus* section *Melioides* females and absence in males of *Fraxinus* sect. *Melioides*. (A) An overview of the region, obtained from IGV, web version 1.12.9. (B) A plot of read coverage around this region. Read coverage was obtained from BEDTools, version 2.25.0.

Table 6. Statistics of variants of interest in the region on Chromosome 8. Adjusted *p*-values used Bonferroni correction and were determined using PLINK, version 1.9.

SNP Location	Called (%)	Uncalled (%)	Male Samples	Female Samples	<i>p</i> -value	Adj. <i>p</i> -value
17168052	41 (21%)	151 (79%)	0	41	4.30E-23	1.97E-16
17168074	69 (36%)	123 (64%)	0	69	6.38E-46	2.92E-39
17168095	73 (38%)	119 (62%)	0	73	4.74E-50	2.17E-43
17168123	81 (42%)	111 (58%)	4	77	4.34E-48	1.99E-41
17168128	45 (23%)	147 (77%)	0	45	7.90E-26	3.62E-19
17168330	55 (29%)	137 (71%)	0	55	n/a	n/a
17168345	55 (29%)	137 (71%)	0	55	n/a	n/a
17178050	90 (47%)	102 (53%)	2	88	5.02E-31	1.26E-08
17178055	117 (61%)	75 (39%)	29	88	5.02E-31	1.26E-08
17178069	32 (17%)	160 (83%)	0	32	0.1694	1
17178084	46 (24%)	146 (76%)	1	45	0.1435	1
17178088	26 (14%)	166 (86%)	0	26	0.1694	1
17178089	29 (15%)	163 (85%)	0	29	0.1694	1
17178097	32 (17%)	160 (83%)	2	30	1.38E-13	6.34E-07
17178103	30 (16%)	162 (84%)	1	29	0.1435	1
17178106	62 (32%)	130 (68%)	1	61	0.246	1
17178118	29 (15%)	163 (85%)	13	16	n/a	n/a
17178124	63 (33%)	129 (67%)	3	60	0.01086	1
17178129	82 (43%)	110 (57%)	39	43	n/a	n/a
17178152	50 (26%)	142 (74%)	7	43	n/a	n/a

Table 7. Genes from *Fraxinus pennsylvanica* located in proximity to the region of interest on Chromosome 8. “Upstream” means the gene is found before the SNP, and “downstream” means the gene is found after the SNP.

Gene Name	Start	End	Location	Function
Fp_g19699	17165500	17166078	Upstream	uncharacterized protein LOC111376422 isoform X1 [<i>Olea europaea</i> var. <i>sylvestris</i>]
Fp_g19700	17168230	17170215	Downstream	phospholipase A2-alpha-like [<i>Olea europaea</i> var. <i>sylvestris</i>]

fatty acids and lysophospholipids (Chen et al., 2013). In plants, PLAs play a variety of roles in plant defense, stress response, and developmental processes (Takáč et al., 2019). Of note to this study, mutated PLAs in *Arabidopsis* have been shown to result in defective pollen development, reduced germination after contact with the stigma, and issues in jasmonic acid (JA) biosynthesis (Ishiguro et al., 2001; Kim et al., 2011).

Discussion

Green ash populations have sustained significant losses following the introduction of the invasive emerald ash borer (EAB) (Kovacs et al., 2010). Certain trees have been shown to survive otherwise infested woods, leading to interest in propagating these “lingering” trees to breed EAB-resistant progeny. Breeding ash trees takes many years, and the dioecious trees cannot be sexed until about 10-25 years after germination (Hausman et al., 2014; Schopmeyer & Others, 1974). Breeding programs can benefit from an earlier molecular-based sex detection system, for example in planning crosses and planting seed orchards. By uncovering a marker of sex determination in green ash, breeders will be able to identify the sex of individual trees prior to reaching maturity, improving the efficiency of designing and planting long-term seed orchards. In this study, we searched for SNPs with significant association to green ash sex determination through a double-digest genotyping-by-sequencing (ddGBS) approach. While we have found a novel SNP of interest in our analysis, our overall results suggest a weakness in using GBS experiments in the search for markers of dioecy.

In the closely related *Olea europaea*, GBS (a cheaper alternative to WGS) was found to be a viable alternative for comparing the genetic relationships of multiple accessions (Friel et al., 2021). In theory, a ddGBS approach would allow us to identify a SNP that is in linkage disequilibrium with the SDR, but our analysis did not find any SNPs matching this pattern. The single significant SNP did not match the expected genotype segregation for a sex-linked marker, with no full segregation of haplotypes between male and female samples. Most plants with well-understood mechanisms of dioecy are not characterized by a single SNP determining sex as male or female but feature some structural differences between the two haplotypes. This can be observed as a missing gene, such as the *ARR17* gene in *Populus alba* (Müller et al., 2020), or as a whole region of the chromosome, such as the two-gene system of asparagus (Harkess et al., 2020). This is a major drawback to identifying a sex-determining region, where a presence-absence analysis is rarely employed. Restriction-associated digest sequencing (RADseq) technologies such as GBS can only capture a fragment of the entire genome, with some estimates pointing to only 5% of the genome (Hoban et al., 2016).

To this point, altering our VCF file for a presence/absence approach identified a region of interest for our study goal along Chromosome 8 of the green ash genome (Huff et al., 2022). Variants called in this region were found to strongly skew towards being called in female samples while remaining uncalled or called in much lower numbers in male samples. These results indicate that this region of Chromosome 8 is a candidate location for the sex-determining region (SDR) due to the strong association between presence of the genotype in females and its absence in males. Based on these analyses, we predict that green ash has an XY sex determining system with males as the heteromorphic sex. Whereas in a ZW system—with heteromorphic females—we would be unlikely to observe any alignment of female-exclusive SNPs due to the male assembly lacking the female haplotype. This could also explain why a few male samples

contained SNPs along this region of interest as males have a copy of the female haplotype in addition to the male haplotype. Overall, these results indicate that the current green ash genome, which was not haplotype phased, has either captured the female haplotype or consists of a chimera of both haplotypes. Functional annotation of nearby genes identified a phospholipase A2-alpha-like (PLA2) downstream of the region. In *Arabidopsis*, PLAs play a critical role in the development of pollen and in germination once pollen reaches a stigma, and mutations in the *PLA* genes can result in sterile pollen (Ishiguro et al., 2001; Kim et al., 2011). While the gene itself does not appear to be mutated—the gene is located downstream of the region of interest—SNPs in the region are sufficiently proximal to affect expression if it were to occur in a promoter region. It is important to note that the case of *ARR17* in poplar has shown that predicted gene function *in silico* does not always indicate gene function *in vivo* (Leite Montalvão et al., 2022; Müller et al., 2020).

A comparison of our organism to dioecious plants with known gene determinants strongly indicates that the genes responsible for sex determination function by altering expression of floral MADS-box genes, which are required for the development of sex-linked floral organs. In plants with sex-determining systems that do not rely on MADS-box genes, such as grape (*Vitis vinifera*), there are vestigial remnants of the opposite sex in the flower structure, such as female flowers producing stunted, infertile stamen (Massonnet et al., 2020). By contrast, in plants with sex-determining systems that rely on MADS-box genes, such as poplar, there are no vestigial remnants of the opposite sex in the flower structure (Müller et al., 2020). Expression of a combination of specific MADS-box genes is required to form specific organs. Thus, if these genes are not present, the organs will not develop (Cronk & Müller, 2020). In the case of poplar, males consistently express the MADS-box genes *PISTILLATA* (*PI*) and *APETALA3* (*AP3*), genes involved in the development of stamen (Cronk et al., 2020). *ARR17* represses expression of *PI*, resulting in female poplars lacking vestigial stamen (Leite Montalvão et al., 2022). Without *ARR17*, *PI* is over-expressed, and the pistil is unable to develop, resulting in male poplars lacking any form of pistils. The flower structure of green ash is similar to that of poplar, i.e., no features of the opposite sex are present in either male or female flowers (Figure 7). For this reason, we believe that one or more genes determine sex by targeting a MADS-box gene, such as *PI* or *AP3*. This would also indicate that mutations associated with the *PLA2* gene are not the main driver of dioecy, as there would likely be vestigial remnants of stamen in female flowers that are not present in green ash.

Future research to assess this region along Chromosome 8 will utilize WGS to further confirm the SDR. One approach is to produce k-mers, or subsequences of length *k*, exclusive to male and female ash trees. Identifying sex-exclusive k-mers and comparing their frequencies is a well-established method to identify SDRs in dioecious plants, and this method has been used to identify SDRs in nightshade (Wu et al., 2021), date palm tree (Torres et al., 2018), kiwifruit (Akagi et al., 2018), and jojoba (Al-Dossary et al., 2021). Based on the findings of these papers, we would expect to find an enrichment of male-specific k-mers that strongly align to a region in the genome if green ash displays an XY sex-determining system. However, given that the current genome is not haplotype-phased (Huff et al., 2022), it is also possible that we will see an enrichment of female-specific k-mers in our region of interest. If male is the heterogametic sex, our genome will possibly include both haplotypes. If a GBS approach were employed again in

the future, a follow-up approach would utilize REs that flank the putative SDR on Chromosome 8 to better confirm reduced presence in male samples.

Another analysis that could help characterize the expression pathways involved in dioecy is an RNA sequencing experiment. This experiment would compare gene expression between male and female flowers collected at multiple time points. This would identify genes with significant differential expression between male and female flowers, indicating one or more pathways that drive the sex-determining pathway. Gene expression was utilized in poplar to identify differential expression of *PI* and *AP3* in male flowers (Cronk et al., 2020) and was later used to confirm the role of *ARR17* in inhibiting MADS-box genes (Leite Montalvão et al., 2022). While our results found a gene of interest potentially associated with pollen tube development, comparisons with flower structure indicate differences in a MADS-box-gene-associated pathway. Comparing gene expression between male and female flowers could help clarify the pathway of sex determination, either by confirming sex-specific expression of key MADS-box genes or by indicating a different mechanism of sex determination. As the SDR in green ash becomes better characterized, future RNASeq experiments could characterize the effects of gene knockouts in female flowers using synthetic biology.

While our results have identified a candidate SDR in the green ash genome, we have yet to confirm it as a genetic marker of sex determination. While a GBS approach identified this region, we found that this technique was not sufficiently informative to characterize the SDR in green ash. Future research is warranted to evaluate this region of interest as the potential SDR in green ash. Confirming this region will provide a reliable marker of sex determination and improve current breeding efforts for restoration.

REFERENCES

- Akagi, T., Henry, I. M., Ohtani, H., Morimoto, T., Beppu, K., Kataoka, I., & Tao, R. (2018). A Y-Encoded Suppressor of Feminization Arose via Lineage-Specific Duplication of a Cytokinin Response Regulator in Kiwifruit. *The Plant Cell*, 30(4), 780–795.
- Al-Dossary, O., Alsubaie, B., Kharabian-Masouleh, A., Al-Mssallem, I., Furtado, A., & Henry, R. J. (2021). The jojoba genome reveals wide divergence of the sex chromosomes in a dioecious plant. In *The Plant Journal* (Vol. 108, Issue 5, pp. 1283–1294). <https://doi.org/10.1111/tpj.15509>
- Charlesworth, D. (2002). Plant sex determination and sex chromosomes. *Heredity*, 88(2), 94–101.
- Charlesworth, D., & Charlesworth, B. (1978). Population genetics of partial male-sterility and the evolution of monoecy and dioecy. *Heredity*, 41(2), 137–153.
- Chen, G., Greer, M. S., & Weselake, R. J. (2013). Plant phospholipase A: advances in molecular biology, biochemistry, and cellular function. *Biomolecular Concepts*, 4(5), 527–532.
- Cronk, Q., & Müller, N. A. (2020). Default Sex and Single Gene Sex Determination in Dioecious Plants. *Frontiers in Plant Science*, 11, 1162.
- Cronk, Q., Soolanayakanahally, R., & Bräutigam, K. (2020). Gene expression trajectories during male and female reproductive development in balsam poplar (*Populus balsamifera* L.). *Scientific Reports*, 10(1), 8413.
- 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., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158.
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). <https://doi.org/10.1093/gigascience/giab008>
- 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(11). <https://doi.org/10.3390/plants10112514>
- Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. In *arXiv [q-bio.GN]*. arXiv. <http://arxiv.org/abs/1207.3907>
- Harkess, A., Huang, K., van der Hulst, R., Tissen, B., Caplan, J. L., Koppula, A., Batish, M., Meyers, B. C., & Leebens-Mack, J. (2020). Sex Determination by Two Y-Linked Genes in Garden Asparagus. *The Plant Cell*, 32(6), 1790–1796.
- Hausman, C. E., Bertke, M. M., Jaeger, J. F., & Rocha, O. J. (2014). Genetic structure of green ash (*Fraxinus pennsylvanica*): implications for the establishment of ex situ conservation protocols in light of the invasion of the emerald ash borer. *Plant Genetic Resources: Characterization and Utilization*, 12(3), 286–297.
- 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. *The*

- American Naturalist*, 188(4), 379–397.
- Huff, M., Seaman, J., Wu, D., Zhebentyayeva, T., Kelly, L. J., Faridi, N., Nelson, C. D., Cooper, E., Best, T., Steiner, K., Koch, J., Romero Severson, J., Carlson, J. E., Buggs, R., & Staton, M. (2022). A high-quality reference genome for *Fraxinus pennsylvanica* for ash species restoration and research. *Molecular Ecology Resources*, 22(4), 1284–1302.
- Ishiguro, S., Kawai-Oda, A., Ueda, J., Nishida, I., & Okada, K. (2001). The DEFECTIVE IN ANTHR DEHISCENCE1 Gene Encodes a Novel Phospholipase A1 Catalyzing the Initial Step of Jasmonic Acid Biosynthesis, Which Synchronizes Pollen Maturation, Anther Dehiscence, and Flower Opening in Arabidopsis. In *The Plant Cell* (Vol. 13, Issue 10, p. 2191). <https://doi.org/10.2307/3871502>
- Jiang, H., Lei, R., Ding, S.-W., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15, 182.
- Kennedy, H. E., Jr. (1949). *Fraxinus pennsylvanica* Marsh. green ash. *Agriculture Handbook*, 2(654), 348.
- Kim, H. J., Ok, S. H., Bahn, S. C., Jang, J., Oh, S. A., Park, S. K., Twell, D., Ryu, S. B., & Shin, J. S. (2011). Endoplasmic reticulum- and Golgi-localized phospholipase A2 plays critical roles in Arabidopsis pollen development and germination. *The Plant Cell*, 23(1), 94–110.
- Koch, J. L., Carey, D. W., Knight, K. S., Poland, T., Herms, D. A., & Mason, M. E. (2012). Breeding strategies for the development of emerald ash borer-resistant North American ash. In: *Snieszko, Richard A.; Yanchuk, Alvin D.; Kliejunas, John T.; Palmieri, Katharine M.; Alexander, Janice M.; Frankel, Susan J., Tech. Coords. Proceedings of the Fourth International Workshop on the Genetics of Host-Parasite Interactions in Forestry: Disease and Insect Resistance in Forest Trees. Gen. Tech. Rep. PSW-GTR-240. Albany, CA: Pacific Southwest Research Station, Forest Service, US Department of Agriculture. Pp. 235-239., 240, 235–239.*
- Koch, J. L., Carey, D. W., Mason, M. E., & Nurul Islam-Faridi, M. (2010, March 9). *Overcoming obstacles to interspecies hybridization of ash.* <https://www.nrs.fs.usda.gov/pubs/gtr/gtr-p-72papers/26-koch-p-72.pdf>
- Kovacs, K. F., Haight, R. G., McCullough, D. G., Mercader, R. J., Siegert, N. W., & Liebhold, A. M. (2010). Cost of potential emerald ash borer damage in U.S. communities, 2009–2019. In *Ecological Economics* (Vol. 69, Issue 3, pp. 569–578). <https://doi.org/10.1016/j.ecolecon.2009.09.004>
- Leite Montalvão, A. P., Kersten, B., Fladung, M., & Müller, N. A. (2020). The Diversity and Dynamics of Sex Determination in Dioecious Plants. *Frontiers in Plant Science*, 11, 580488.
- Leite Montalvão, A. P., Kersten, B., Kim, G., Fladung, M., & Müller, N. A. (2022). ARR17 controls dioecy in *Populus* by repressing B-class MADS-box gene expression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 377(1850), 20210217.
- Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5), 589–595.

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.
- Luo, W., Xia, Y., Yue, B., & Zeng, X. (2020). Assigning the Sex-Specific Markers via Genotyping-by-Sequencing onto the Y Chromosome for a Torrent Frog *Amolops mantzorum*. *Genes*, 11(7). <https://doi.org/10.3390/genes11070727>
- Mariotti, R., Fornasiero, A., Mousavi, S., Cultrera, N. G. M., Brizioli, F., Pandolfi, S., Passeri, V., Rossi, M., Magris, G., Scalabrin, S., Scaglione, D., Di Gaspero, G., Saumitou-Laprade, P., Vernet, P., Alagna, F., Morgante, M., & Baldoni, L. (2019). Genetic Mapping of the Incompatibility Locus in Olive and Development of a Linked Sequence-Tagged Site Marker. *Frontiers in Plant Science*, 10, 1760.
- Massonnet, M., Cochetel, N., Minio, A., Vondras, A. M., Lin, J., Muyle, A., Garcia, J. F., Zhou, Y., Delledonne, M., Riaz, S., Figueroa-Balderas, R., Gaut, B. S., & Cantu, D. (2020). The genetic basis of sex determination in grapes. *Nature Communications*, 11(1), 2902.
- Müller, N. A., Kersten, B., Leite Montalvão, A. P., Mähler, N., Bernhardsson, C., Bräutigam, K., Carracedo Lorenzo, Z., Hoenicka, H., Kumar, V., Mader, M., Pakull, B., Robinson, K. M., Sabatti, M., Vettori, C., Ingvarsson, P. K., Cronk, Q., Street, N. R., & Fladung, M. (2020). A single gene underlies the dynamic evolution of poplar sex determination. *Nature Plants*, 6(6), 630–637.
- Poland, T. M., & McCullough, D. G. (2006). Emerald ash borer: invasion of the urban forest and the threat to North America's ash resource. *Journal of Forestry*, 104(3), 118–124.
- 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. *American Journal of Human Genetics*, 81(3), 559–575.
- Renner, S. S., & Müller, N. A. (2021). Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. *Nature Plants*, 7(4), 392–402.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26.
- Saumitou-Laprade, P., Vernet, P., Dowkiw, A., Bertrand, S., Billiard, S., Albert, B., Gouyon, P.-H., & Dufay, M. (2018). Polygamy or subdioecy? The impact of diallelic self-incompatibility on the sexual system in *Fraxinus excelsior* (Oleaceae). *Proceedings. Biological Sciences / The Royal Society*, 285(1873). <https://doi.org/10.1098/rspb.2018.0004>
- Schopmeyer, C. S., & Others. (1974). Seeds of woody plants in the United States. *Seeds of Woody Plants in the United States.*, 450. <https://www.cabdirect.org/cabdirect/abstract/19750622387>
- Siebert, N. W., McCullough, D. G., Liebhold, A. M., & Telewski, F. W. (2014). Dendrochronological reconstruction of the epicentre and early spread of emerald ash

- borer in North America. In *Diversity and Distributions* (Vol. 20, Issue 7, pp. 847–858). <https://doi.org/10.1111/ddi.12212>
- Takáč, T., Novák, D., & Šamaj, J. (2019). Recent Advances in the Cellular and Developmental Biology of Phospholipases in Plants. *Frontiers in Plant Science*, 10, 362.
- Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12), 2032–2034.
- Torres, M. F., Mathew, L. S., Ahmed, I., Al-Azwani, I. K., Krueger, R., Rivera-Núñez, D., Mohamoud, Y. A., Clark, A. G., Suhre, K., & Malek, J. A. (2018). Genus-wide sequencing supports a two-locus model for sex-determination in Phoenix. *Nature Communications*, 9(1), 3969.
- Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. In *bioRxiv* (p. 005165). <https://doi.org/10.1101/005165>
- Vernet, P., Lepercq, P., Billiard, S., Bourceaux, A., Lepart, J., Dommée, B., & Saumitou-Laprade, P. (2016). Evidence for the long-term maintenance of a rare self-incompatibility system in Oleaceae. *The New Phytologist*, 210(4), 1408–1417.
- Wallander, E. (2008). Systematics of Fraxinus (Oleaceae) and evolution of dioecy. *Osterreichische Botanische Zeitschrift*, 273(1-2), 25–49.
- Wallander, E. (2012). Systematics and floral evolution in Fraxinus (Oleaceae). *Belgische Dendrologie Belge*, 2012, 39–58.
- Wu, M., Haak, D. C., Anderson, G. J., Hahn, M. W., Moyle, L. C., & Guerrero, R. F. (2021). Inferring the Genetic Basis of Sex Determination from the Genome of a Dioecious Nightshade. *Molecular Biology and Evolution*, 38(7), 2946–2957.
- Yang, X., Luo, W., Xia, Y., & Zeng, X. (2022). Using Sex-Linked Markers via Genotyping-by-Sequencing to Identify XX/XY Sex Chromosomes in the Spiny Frog (*Quasipaa boulengeri*). *Genes*, 13(4). <https://doi.org/10.3390/genes13040575>

CHAPTER V:
LONG-READ, CHROMOSOME-SCALE ASSEMBLY OF *VITIS*
***ROTUNDIFOLIA* CV. CARLOS AND ITS UNIQUE RESISTANCE TO**
XYLELLA FASTIDIOSA* SUBSP. *FASTIDIOSA

Abstract

Muscadine grape (*Vitis rotundifolia*) is resistant to many of the pathogens that negatively impact the production of common grape (*V. vinifera*), including the bacterial pathogen *Xylella fastidiosa*, subsp. *Fastidiosa* (*Xfsf*), which causes Pierce's Disease (PD). Previous studies in common grape have indicated *Xfsf* delays host immune response with a complex O-chain antigen produced by the *wzy* gene. Muscadine cultivars range from tolerant to completely resistant to *Xfsf*, but the mechanism is unknown. We assembled and annotated a new, long-read genome assembly for 'Carlos,' a cultivar of muscadine that exhibits tolerance, to build upon the existing genetic resources available for muscadine. We used these resources to construct an initial pan-genome for three cultivars of muscadine and one cultivar of common grape. This pan-genome contains a total of 34,970 synteny-constrained entries containing genes of similar structure. Comparison of resistance gene content between the 'Carlos' and common grape genomes indicates an expansion of resistance (R) genes in 'Carlos.' We further identified genes involved in *Xfsf* response by transcriptome sequencing 'Carlos' plants inoculated with *Xfsf*. We observed 234 differentially expressed genes with functions related to lipid catabolism, oxidation-reduction signaling, and abscisic acid (ABA) signaling as well as seven R genes. Leveraging public data from previous experiments of common grape inoculated with *Xfsf*, we determined that most differentially expressed genes in the muscadine response were not found in common grape, and three of the R genes identified as differentially expressed in muscadine do not have an ortholog in the common grape genome. Our results support the utility of a pan-genome approach to identify candidate genes for traits of interest, particularly disease resistance to *Xfsf*, within and between muscadine and common grape.

Introduction

Muscadine (*Vitis rotundifolia* Michx., syn. *Muscadinia rotundifolia* $2n = 2x = 40$) is a species of grape native to the southeastern United States. It is part of the *Muscadinia* subgenus of *Vitis*, which contains only three species (Hickey et al., 2019). In comparison, the subgenus *Euvitis* (bunch grapes) exhibits greater diversity and includes about 60 species. Compared to the common table grape (*V. vinifera*; a member of the *Euvitis* subgenus), muscadine grapes are a small-market crop (Hoffmann, 2020). Native Americans used the muscadine fruit both as a food source and a dye prior to European colonization, and wineries in the Southeast have produced special red and white wines from muscadine berries since the 16th century (Reimer et al., 1906). Muscadine grapes have increased in commercial interest outside of culinary applications, including medicinal use and nutritional supplements, due to higher antioxidant content and an increased diversity of bioactive compounds relative to other grape species. Several studies demonstrate treatment of cancer cells *in vitro* with muscadine grape skin extracts (MGSEs) has been found to increase cytotoxicity and trigger apoptosis in breast cancer cells (Luo et al., 2017; Mendonca et al., 2019) and prostate cancer cells (Ignacio et al., 2019). In the context of grape breeding, muscadines have attracted interest in providing a resource for beneficial traits to be introduced into bunch grape species. However, breeding efforts are complicated by the different number of chromosomes caused by fusion of two muscadine chromosomes in the *Vitis vinifera* genome.

Pierce's Disease (PD) has become a major limiting factor in the production of wine and table grapes in southern California, the center of winemaking in the continental United States (US). Symptoms include leaf necrosis, leaf scorch, and dieback, leading to vine death within three to five years. PD is caused by the xylem-limited bacterial pathogen *Xylella fastidiosa* subsp. *fastidiosa* (*Xfsf*) and is spread by xylem-feeding sharpshooter family (Cicadelladea) insect vectors (Krivanek & Walker, 2005). Disease spread was limited by short flying distance and low feeding rate of the insect vector, however, the introduction of a more efficient vector in the 1990s, the glassy-winged sharpshooter (*Homalodisca vitripennis*), exacerbated the impact of PD in California (Blua et al., 1999). *Xfsf* is a gram-negative bacterium that is indigenous to the Gulf Coast region of the US, but it has since spread to other regions of both American continents as an invasive disease (Hopkins & Purcell, 2002). The bacteria clog the xylem vessels of susceptible grape genotypes, leading to water and nutrient stress. PD symptoms usually appear in late summer during high temperatures or when the plants are exposed to drought conditions. In most muscadine grapes, PD is of little concern, with disease symptoms typically limited to occasional necrosis or marginal leaf burn on tolerant cultivars.

Previous studies have found clues to the physiology and molecular basis of resistance to PD in both muscadines and common grapes. In comparison to susceptible bunch grapevines, tolerant muscadine cultivars infected with *Xfsf* produced higher frequencies of tyloses and gums, which occlude xylem vessels to encapsulate and prevent internal spread of the bacteria (Mollenhauer & Hopkins, 1976). In addition, gums and tannins occur more frequently in the PD-resistant muscadine grape cultivar 'Noble' after PD inoculation than in the PD-tolerant cultivar 'Carlos' (Fry et al., 1990; Fry & Milholland, 1990). Furthermore, higher amounts of xylem sap β -1,3-glucanase and peroxidase was observed in PD-tolerant muscadine grapes compared to PD-susceptible bunch grapes (Basha et al., 2010). In PD, knockout of the *wzy* gene in *Xfsf* was observed to result in a tolerant host response (Rapicavoli et al., 2018). The *wzy* mutants were unable to produce a complex lipopolysaccharide O-antigen that, in wild-type *Xfsf*, delays host immune response. Comparison of infected and uninfected grape vines (var. Thompson Seedless) found an up-regulation of antioxidant strategies, cell wall modification enzymes, and pathogenesis-related proteins (Zaini et al., 2018). Although structural barriers provide some degree of resistance/tolerance against pathogens, little is known about the relationship between muscadine grape anatomy and its tolerance to *Xfsf*.

At present, two muscadine genomes have been published. The first genome was a haplotype-phased, chromosome-level assembly of the male cultivar 'Trayshed,' representing about 86.2% of the estimated 484.3 Mbp genome size (Cochetel et al., 2021). The genome was used to verify the fusion of muscadine chromosome 7 and 20 in common grape as chromosome 7, previously only supported by muscadine linkage group evidence (Blanc et al., 2012; Lewter et al., 2019). Compared to the assembly of Cabernet Sauvignon, a disease-susceptible cultivar of bunch grape, the 'Trayshed' assembly had an expansion of Toll/Interleukin-1 Receptor-like Nucleotide-Binding Site Leucine-Rich Repeat (TIR-NBS-LRR) proteins, a class of disease resistance genes. A second assembly was recently published for the 'Noble' cultivar and represents 81.5% of the estimated total genome size (Park et al., 2022). Additionally, this study included a genome-wide association study (GWAS) which identified 52 quantitative trait nucleotides (QTNs) associated with 12 berry-related traits. Due to the similarity between *Muscadinia* and *Euvitis* genomes, these QTN markers can be used to study comparative genetics

between species. As the dynamic of genomics shifts towards a pan-genome model, which covers genetic content across multiple cultivars of a species, these two assemblies can be used together to provide a more accurate understanding of the relationship between different cultivars (Bayer et al., 2020; Della Coletta et al., 2021).

To examine gene-expression changes in response to inoculation with *Xfsf*, RNA-Seq was performed on the muscadine ‘Carlos,’ a bronze-skinned, commercial cultivar with tolerance to PD infection. A ‘Carlos’ reference genome was constructed to improve the accuracy of the RNA-Seq data analysis and expand existing muscadine genomic resources. To establish a genomic framework for this comparison within and across species, we developed a pan-genome resource incorporating three muscadine genome assemblies and a common grape genome assembly. To further understand the differences in response between muscadine and common grapes, we compared our gene expression data to grapes inoculated with both wild-type *Xfsf* and *wzy* knock-out mutants (Rapicavoli et al., 2018).

Materials and Methods

Plant sampling - DNA sequencing

The muscadine grape clone ‘US19-33’, a self-pollinated accession of *V. rotundifolia* Michx., ($2n=2x=40$) cultivar ‘Carlos’ (Fry & Milholland, 1990), maintained at the USDA ARS Southern Horticultural Research Laboratory in Poplarville, MS was selected for genomic DNA (gDNA) extraction (30.8402° N, 89.5342° W). Young leaves were collected from a single plant of clone ‘US19-33’ and used for gDNA extraction using a modified hexadecyltrimethylammonium bromide (CTAB) protocol (Nishiyama et al. 2021). Isolated gDNA was quantified with a Nanodrop 2000 spectrophotometer and a Qubit dsDNA HS assay kit (ThermoFisher Scientific, Waltham, MA), and gDNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Tech, Santa Clara, CA, USA). Ultimately, gDNA libraries for Illumina short-read sequencing were prepared, and paired-end sequencing (2×150 bp, 30× coverage) was performed on the Illumina NovaSeq 6000 sequencing platform (Illumina, San Diego, CA, USA).

To isolate high molecular weight (HMW) DNA suitable for PacBio long-read sequencing, young expanding leaves were dark-treated for 48 h prior to harvest and flash frozen in liquid nitrogen prior to gDNA extraction. HMW DNA was used to construct libraries following the manufacturer’s protocol and sequenced using the CLR (continuous long read) protocol on the Pacific Biosciences (PacBio) sequencing platform (Pacific Bioscience, Menlo Park, CA, USA).

Plant sampling - RNA Sequencing

A total of 8 plantlets from tissue culture propagated muscadine grape cultivar ‘Carlos’ were transferred from Magenta boxes into 1-gallon plastic pots containing a mixture of pine bark mulch and sand (1:1, v/v) with a pH of 5.6. Stems were needle-inoculated approximately 7 cm above the media surface with 10 µl of succinate-citrate buffer containing 10^6 cells of *X. fastidiosa* subsp. *fastidiosa* (*Xfsf*) strain NOB1 (Mavrodi et al., 2020). Control plants were mock inoculated with 10 µl of succinate-citrate buffer. Plants were stored inside an incubator

maintained at the following conditions: 25° C temperatures during 16 h photoperiods of 154 $\mu\text{mol m}^{-2}\text{s}^{-1}$ photon flux density, 22° C temperatures during 8 h skotoperiods, and 65% consistent relative humidity.

Approximately 0.1g of leaf tissues were collected 28 days post-inoculation and used for DNA and RNA extraction. Total genomic DNA was extracted using a CTAB method and used for the detection of *Xfsf* using RST31/RST33 primer pairs as described by Minsavage et al. 1994. Total RNA was isolated from leaf tissues collected from control and PD-inoculated plants using the Spectrum plant total RNA kit protocol (Sigma-Aldrich, MO, USA). RNA quantity and purity were assessed with a Nanodrop 2000 spectrophotometer and a Qubit 2.0 Fluorometer (ThermoFisher, Waltham, MA, USA), respectively, and RNA quality was evaluated using an Agilent 2100 Bioanalyzer (Agilent Tech, Santa Clara, CA, USA). RNA sequencing libraries were constructed using the NEBNextTM II Directional RNA library prep kit, and sequencing was performed on the Illumina NovoSeq 6000 platform (Illumina, San Diego, CA, USA).

Genome scaffolding and assembly

CLR reads were initially assembled into contigs using MECAT2 version 2019.03.04 (Xiao et al., 2017), which implemented CANU version 1.9 (Koren et al., 2017). Resulting contigs underwent one round of polishing with Arrow (Pacific BioSciences SMRT Tools Reference Guide, 2019), followed by additional polishing with Pilon (Walker et al., 2014). Circular contigs, which were labeled in the output of CANU (Koren et al., 2017)v. 1.9, were removed from the assembly.

Following initial contig-joining, RagTag version 1.0.1 (Alonge et al., 2019) was used to align the 'Carlos' contigs to the previously published 'Trayshed' assembly (focusing on haplotype 1). This produced a chromosome-level assembly with gap regions. These gaps were resolved with raw PacBio reads by running TGS-GapCloser, version 1.1.1 (Xu et al., 2020). Benchmarking Universal Single-Copy Orthologs (BUSCO), version 5.2.2, was run to assess the completeness of the 'Carlos' cultivar assembly (Seppey et al., 2019). Statistical analysis of the scaffolds was performed using the BBtools “stats” command (Bushnell, 2018). Repetitive regions were identified using RepeatModeler (Smit et al., 2015b) and softmasked using RepeatMasker (Smit et al., 2015a). RNA-associated repetitive sequences were left unmasked.

Gene annotation

Gene Model Maker (GeMoMa) version 1.8 was used to predict the “Carlos” gene annotation using the existing annotation for the 'Trayshed' haplotype 1 genome assembly (Keilwagen et al., 2019). GeMoMa includes an option to enhance the lift-over by incorporating RNASeq read alignment files, which we used to include the PD-inoculation transcriptome data in this analysis. Thereafter, input reads were trimmed with Skewer version 0.2.2 (Jiang et al., 2014), and trimmed reads were aligned to the 'Carlos' assembly with STAR version 2.7.9a (Dobin et al., 2013). All alignment files were merged into a single file using the Samtools version 1.10 “merge” command (H. Li et al., 2009). In addition to including the alignment data, GeMoMa was run with options to include untranslated regions (UTRs) and other features from the 'Trayshed' annotation. Identified genes were renamed, and gFACs version 1.1.2 was run without any filtering parameters to produce updated coding sequence (CDS) and amino acid (AA) FASTA files (Caballero & Wegrzyn, 2019). Non-coding RNA features such as rRNA and

tRNA were annotated using RNAmmer version 1.2 (Lagesen et al., 2007) and tRNAScan version 2.0.9. (Chan & Lowe, 2019; Lowe & Eddy, 1997).

Comparative genomics of 'Carlos' to other genomes

The 'Carlos' chromosomes were aligned to the 'Trayshed' haplotype 1 chromosomes and the 'Noble' chromosomes using MiniMap2 version 2.24 with default settings (H. Li, 2018). The resulting alignment files were run through synteny and rearrangement identifier (SyRI), version 1.6.3, to identify any structural rearrangements (Goel et al., 2019). The resulting output was placed into SyRI's sister program PlotSR, version 0.5.4, to visualize all structural rearrangements between 'Carlos' and the aforementioned 'Noble' and 'Trayshed' genomes (Goel & Schneeberger, 2022). In-depth gene annotation comparison and initial pan-genome construction was performed using GENESPACE, version 0.9.4 (Lovell et al., 2022). The common grape (*Vitis vinifera*), version 12.1, annotation was included in the GENESPACE analysis (Jaillon et al., 2007). A custom script was developed to obtain statistics from the GENESPACE results

(https://github.com/MattHuff/GENESPACE-Companion-Script/blob/main/parse_pangenome.py). Plant disease resistance-related genes (R-genes) were identified in these four genomes using the Disease Resistance Analysis and Gene Orthology (DRAGO2) pipeline (Osuna-Cruz et al., 2018).

Due to grape and muscadine having a different total number of chromosomes, SyRI could not be run to compare assembly structure. Instead, Orthofinder version 2.3.12 was used to identify orthologs between 'Carlos' muscadine and common grape (D. M. Emms & Kelly, 2019). Protein sets from a total of 8 other species, including the 'Trayshed' haplotype 1 annotation, were included as outgroups to enhance OrthoFinder phylogenetic results based on the suggested best practices in the OrthoFinder documentation (D. Emms, 2019). Multiple sequence alignment (MSA) mode was selected using MAFFT version 7.467 to obtain MSA files (Katoh & Standley, 2013). Orthogroups with a single gene member from both 'Carlos' and common grape were selected (Huff et al., 2022; Staton et al., 2020), and the related orthologous links were plotted using RIdeogram version 0.2.2 to visualize synteny between the genomes (Hao et al., 2020).

Differential expression analysis

Our 'Carlos' RNASeq data were trimmed using Skewer, version 0.2.2, with identical parameters to those used in the gene annotation section (Jiang et al., 2014). STAR, version 2.7.9a (Dobin et al., 2013), was rerun for the muscadine samples against the 'Carlos' genome with gene annotation included. Following alignment, gene features were associated with read counts using HTSeq-Count version 0.13.5 (Anders et al., 2010). Differential expression was determined using DESeq2 version 1.26.0 analysis of the read count data (Love et al., 2014). Expression was compared between the control samples and the inoculated samples. Significantly differentially expressed genes were called with an adjusted *p*-value of 0.05. DRAGO2 was rerun on the differentially expressed genes to identify candidate R-genes.

Common grapevine RNASeq data were obtained from a previous study that confirmed the effect of *wzy* knockout in *Xfsf*, strain *Temecula1*, on host response (NCBI accession: PRJNA345471) (Rapicavoli et al., 2018). Due to the difference in strains, we compared the existing annotations of both strains to confirm that no significant structural variation occurred in *Wzy* between the strains. To match the conditions of our 'Carlos' RNASeq data, samples were

only included if collected 4 weeks after inoculation from petioles. A total of 9 samples were analyzed: 3 representing control, 3 representing infection with wild-type *Temecula1* strain *Xfsf*, and 3 representing infection with mutant *wzy* knockout *Temecula1* strain *Xfsf*. For both sets of samples, Skewer, version 0.2.2, was run using identical parameters to those used in the gene annotation section. The same version and parameters of STAR were run to align the reads with the *V. vinifera* genome, version 2.1. For the common grape data, expression was compared between the control and inoculated samples, the inoculated and mutant samples, and the control and mutant samples (SRR4345376-SRR4345378, SRR4345409-SRR4345411, and SRR4345433-SRR4345435) using the same adjusted p-value to call differentially expressed genes.

To compare the expression of genes in 'Carlos' muscadine to common grape, the results of the earlier OrthoFinder run were used. Orthogroups containing significantly differentially expressed genes from analyses were identified. Functional gene annotation of 'Carlos' was determined using EnTAP version 0.10.8 (Hart et al., 2020), which assigned functions based on sequence similarity, Gene Ontology (GO) term assignment, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotation. The Muscadine proteins of interest were also functionally annotated using KEGG's BlastKOALA and GhostKOALA functions (Kanehisa et al., 2016). Impacted pathways were determined by running KEGG separately on up- and down-regulated genes.

Results

Genome assembly information and statistics

Our 'Carlos' chromosome-scale assembly utilized PacBio CLR long-read sequencing to generate DNA scaffolds. The initial assembly consisted of 384 contigs arranged into 355 scaffolds with a total of 413 Mb sequences and a contig N50 of 2.3 Mb (Supplemental Table S8). Reference-guided scaffolding of the initial 'Carlos' assembly to the 'Trayshed' haplotype 1 assembly (Cochetel et al., 2021) arranged the contigs into 76 scaffolds, including 20 chromosome-level scaffolds that contained 99% of the sequenced bases. After resolving gaps, the assembly consisted of 112 contigs in 76 scaffolds with a low percentage of uncalled bases (Table 8). The final assembly has a total sequence length of 413 Mb with the 20 assembled chromosomes representing 97.7% of the total length. Assessment of our assembly by benchmarking universal single-copy orthologs (BUSCOs) indicated the sequence was largely complete (Seppey et al., 2019). This assembly is slightly longer than the two other sequenced cultivars (394Mb for 'Noble' and 400Mb for 'Trayshed'), but all three are notably lower than the flow cytometry estimate of 483Mb, suggesting the flow cytometry is likely an overestimate (Cochetel et al., 2021; Park et al., 2022).

Gene annotation report

Using a combination of *de novo* repeat calling and known plant repeats, 51.12% of the 'Carlos' assembly was identified as repetitive (Supplemental Table S9). Long-terminal repeats (LTRs) were the most common repeats (15.96%), primarily consisting of Gypsy/DIRS1 (8.06%) and Ty1-Copia (7.33%) retrotransposon families. A total of 27,923 genes were identified in the 'Carlos' assembly, of which 27,747 (99.4%) of these genes were located on the 20 assembled

Table 8. Summary of the *Vitis rotundifolia* 'Carlos' genome assembly. BUSCO scores were identical when considering all scaffolds and only chromosomes.

	All Scaffolds	Chromosomes Only
Total Scaffold Length	413,913,702	404,563,012
Total Contig Length	413,908,995	404,559,443
No. of Scaffolds	76	20
No. of Contigs	112	55
Largest Scaffold	31,234,097	31,234,097
Smallest Scaffold	2,758	13,845,986
# of Ns	3,707	3,569
GC (%)	34.11	34.01
Protein Coding Genes	27,923	27,747
Total Transcripts (w/Isoforms)	40,416	40,219
BUSCO category	Number in genome	
Complete BUSCOs (C)	1583 (97.9%)	
Complete, Single-Copy (S)	1556 (96.4%)	
Complete, Duplicated (D)	25 (1.5%)	
Fragmented BUSCOs (F)	21 (1.3%)	
Missing BUSCOs (M)	12 (0.8%)	
Total	1614	

chromosomes. The inclusion of RNA-Seq alignment data allowed GeMoMa to identify 40,219 gene isoforms (Keilwagen et al., 2019), and tRNAscan identified 611 tRNAs across all 20 chromosomes (Chan & Lowe, 2019). RNAmmer identified two nucleolus organizing regions (NORs) with 18S-5.8S-26S (35S) rDNA arrays on chromosomes 15 and 17, as well as two distinct 5S rDNA regions on chromosome 17 (Figure 10) (Lagesen et al., 2007).

Comparison of assembly to existing cultivars and common grape genome

Overall chromosome size was consistent among all three assemblies, though chromosome 18 was notably longer in the 'Trayshed' assembly than in the other cultivar assemblies. Alignment of the 'Carlos' cultivar assembly to the existing 'Trayshed' and 'Noble' cultivar assemblies demonstrated an overall similar structure among cultivars (Figure 10) (Cochetel et al., 2021; Park et al., 2022). Most regions in all three genomes were syntenic, with 83.8% of the 'Carlos' genome syntenic with the 'Trayshed' assembly and 73.4% syntenic with the 'Noble' assembly (Table 9). In these syntenic comparisons with 'Carlos', 'Trayshed' had a higher number of sites with translocations compared to 'Noble'. In contrast, the 'Noble' genome assembly had a larger number of inversions despite containing more syntenic regions with 'Carlos'. Chromosome 8 in the 'Carlos' assembly had a gap in synteny in its alignments to both 'Trayshed' and 'Noble' at approximately 5 Mb.

Alignment of the 'Carlos' cultivar to the common grape genome, *V. vinifera* version 12.1 (Jaillon et al., 2007), showed strong collinearity between the two assemblies, and chromosome 7 in the common grape assembly was shown to be split between chromosomes 7 and 20 in the 'Carlos' assembly, consistent with previous muscadine cultivar assemblies and linkage maps (Figure 11) (Cochetel et al., 2021; Delame et al., 2019; Lewter et al., 2019; Park et al., 2022).

Comparison of the three muscadine and the common grape genome annotation sets with GENESPACE, version 0.9.4, produced a synteny-based pan-genome (Table 10, Supplemental Table S10) (Lovell et al., 2022). GENESPACE predicted a total of 34,970 synteny-constrained orthogroups. 17,457 orthogroups contain at least one gene from each genome, representing the core *Vitis* gene set. Of these genes, 10,750 match exactly one gene from each genome, representing single copy orthologs. The remaining 17,513 orthogroups are missing genes from at least one genome. 4,919 orthogroups are represented by only genes from one genome, and 4,321 of these have a single gene from a single genome, hereafter referred to as singleton genes. Of the singletons, 426 were observed in 'Carlos', 607 were observed in 'Noble', 455 were observed in 'Trayshed', and 2,833 were observed in common grapes. Annotation of the 'Carlos' singletons indicated the presence of RPV1 and RUN1, two muscadine disease-resistance genes (Feechan et al., 2013), among other plant defense proteins (Supplemental Table S11). The GENESPACE riparian synteny plot is consistent with the synteny plots produced by SyRi and RIdeogram (Supplemental Figure S5).

We ran RNAmmer, version 1.2 (Lagesen et al., 2007), on the assemblies for 'Trayshed', 'Noble', and common grape to compare predicted rDNA regions between other muscadine and grape assemblies (Figure 5, Supplemental Tables S12-15). In the 'Carlos' assembly, two nucleolus organizing regions (NORs) were found on chromosomes 15 and 17, and two 5S arrays were found on chromosome 17. In 'Trayshed', an NOR was only detectable on chromosome 17. In 'Noble', a single NOR on chromosome 15 was detected. We did not detect any complete

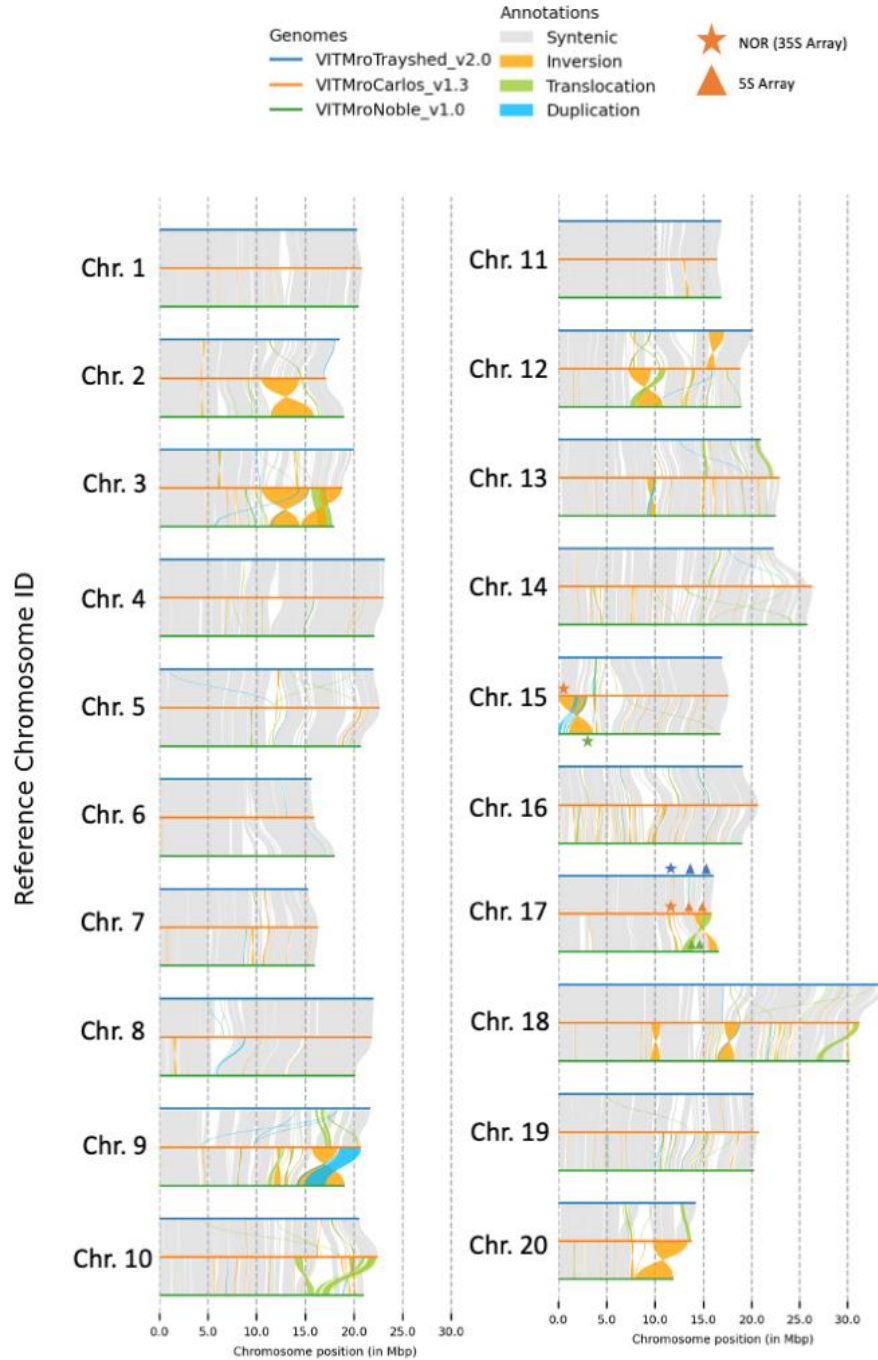


Figure 10. Plot of collinearity between *Vitis rotundifolia* 'Carlos', *V. rotundifolia* 'Trayshed', and *V. rotundifolia* 'Noble' assemblies. The location of rDNA containing a nucleolus organizing region (NOR) is denoted by a star, while rDNA containing a 5S array is denoted by a triangle. Symbols are correlated by color to the associated cultivar (see figure legend).

Table 9. Quantitative statistics of collinearity between *Vitis rotundifolia* 'Carlos', *V. rotundifolia* 'Trayshed', and *V. rotundifolia* 'Noble' assemblies.

'Carlos' Aligned to 'Trayshed'			
Variation_type	Count	Length ('Trayshed')	Length ('Carlos')
Syntenic regions	551	337,011,859	334,838,553
Inversions	43	2,924,794	2,055,345
Translocations	562	7,624,712	7,764,274
Duplications ('Trayshed')	211	1,719,701	-
Duplications ('Carlos')	401	-	1,477,822
Not aligned ('Trayshed')	1272	52,190,426	-
Not aligned ('Carlos')	1504	-	58,644,910
'Carlos' Aligned to 'Noble'			
Variation_type	Count	Length ('Noble')	Length ('Carlos')
Syntenic regions	563	301,890,417	297,062,683
Inversions	278	37,080,995	35,655,433
Translocations	427	11,422,729	11,474,627
Duplications ('Noble')	353	1,807,614	-
Duplications ('Carlos')	229	-	3,255,480
Not aligned ('Noble')	1580	43,540,564	-
Not aligned ('Carlos')	1345	-	59,119,165
'Trayshed' Aligned to 'Noble'			
Variation_type	Count	Length ('Noble')	Length ('Trayshed')
Syntenic regions	624	294,764,948	294,233,443
Inversions	241	30,059,987	34,873,805
Translocations	608	16,742,580	16,733,441
Duplications ('Noble')	229	3,805,281	-
Duplications ('Trayshed')	715	-	4,612,284
Not aligned ('Noble')	1482	54,144,482	-
Not aligned ('Trayshed')	2001	-	51,185,359

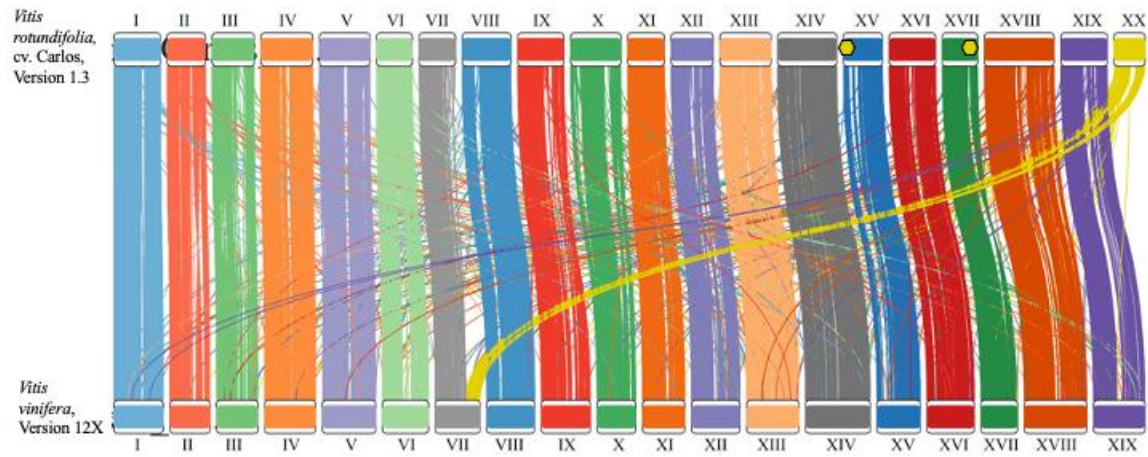


Figure 11. Plot of collinearity between *Vitis rotundifolia* 'Carlos' and *Vitis vinifera* v12.1 assemblies. Figure generated using RIdeogram version 0.2.2. Yellow hexagons indicate the location of 35S rDNA arrays.

NORs in the Version 2.1 assembly of the common grape genome (Figure 11). Two 5S arrays along chromosome 17 were detected in all assemblies.

We identified and classified predicted plant disease resistance-related genes (R-genes) in 'Carlos', 'Trayshed', 'Noble', and common grape to compare between tolerant, resistant, and susceptible species (Figure 12). We first identified genes with R-gene relevant domains, including coiled-coil (CC), kinase, leucine rich repeat (LRR), nucleotide binding site (NBS), Toll/interleukin-1 receptor (TIR), and transmembrane helix domains. 'Carlos' contained a total of 3,522 genes containing domains associated with R-genes, a comparable number to 'Trayshed' (3,274), but higher than the number observed in 'Noble' (2,394) and common grape (2,332). These sets of genes were further categorized by domain combinations into common R-gene classes: receptor-like kinases (RLK), receptor-like proteins (RLP), TIR-NBS-LRRs (TNL), CC-NBS-LRRs (CNLs), and other NBS-LRRs (NLs). 'Carlos' and 'Trayshed' have similar gene counts in most classes, and are consistently higher than in 'Noble', except the TNL class. Common grape has the lowest number in all R-gene categories except NLs, where it has one gene more than 'Noble', yet still less than the other two muscadine cultivars. This reveals an overall trend of expansion of R-genes in muscadine grape species that is not present in common grape species. All four gene sets had a higher number of RLK and RLP class genes than other NBS-related classes (Supplemental Table S16).

We then identified orthogroups from GENESPACE that contained one or more R-gene candidates from 'Carlos'. The 3,522 candidate R-genes were placed into 2,061 synteny-constrained orthogroups. Of these orthogroups, 250 were predicted to function as RLKs, 303 as RLPs, 84 as TNLs, 132 as CNLs, and 89 as other NLs. In comparing the number of genes in these orthogroups from 'Carlos' to those of common grape, we found that 518 of these orthogroups contain genes novel to 'Carlos', and 392 contain more genes in 'Carlos' compared to grape. The orthogroups unique to 'Carlos' contained 64 genes predicted to function as RLKs, 109 as RLPs, 50 as TNLs, 54 as CNLs, and 37 as other NLs. Of the 392 orthogroups with more 'Carlos' genes, 65 are predicted to function as RLKs, 166 as RLPs, 36 as TNLs, 87 as CNLs, and 57 as other NLs. For the remaining orthogroups, 244 contained more genes in grape than in 'Carlos', and 907 contained an equal number of genes between 'Carlos' and grape. These numbers support that a number of R-genes have either undergone expansions in muscadine or are unique to muscadine.

Differential expression in response to Pierce's Disease

Four weeks after stem inoculation of *Xfsf* in 'Carlos' cultivar plants, RNA sequencing of leaf tissue revealed 234 differentially expressed (DE) genes compared to control inoculated plants (Table 11, Supplemental Table S17). 89 of these genes were up-regulated in infected 'Carlos' vines, and 145 were down-regulated. Annotation of these differentially expressed genes indicated a down-regulation of genes associated with lipid catabolism (mostly consisting of GDSL lipases), both up and down regulation of genes associated with oxidation-reduction, and an up-regulation of genes associated with ubiquitin, a proteolysis mediator associated with plant defense (Sharma et al., 2016). Genes predicted to function as lignin-degrading laccases were down-regulated, possibly indicating a strengthening of the cell wall through lignin fortification (Christopher et al., 2014). Genes associated with abscisic acid (ABA) biosynthesis, a pathway involved in drought resistance and stress response in plants (Muhammad Aslam et al., 2022),

Table 10. Muscadine pan-genome information and statistics. Four gene sets, three from muscadine genome assemblies ‘Carlos’, ‘Trayshed’, ‘Noble’ and one from *V. vinifera* were utilized. “Syntenic links” are regions of the genomes that share gene order. “Core” orthogroups refer to orthogroups with at least one gene member from each of the four gene sets. “Single-copy” orthogroups refer to core orthogroups with exactly one gene from each of the four gene sets. “Accessory” orthogroups refer to orthogroups where at least one gene set does not have an included gene. “Single-individual” orthogroups refer to orthogroups where genes from only a single gene set are present. “Singletons” refer to orthogroups containing only a single gene from a single gene set.

Number of Initial Orthogroups	22,864
Number of Syntenic Links	1,084
Number of Synteny-Constrained Orthogroups	34,970
Number of Core <i>Vitis</i> Orthogroups	17,457 (49.9%)
Number of Single-Copy Orthogroups	10,750 (30.7%)
Number of Accessory <i>Vitis</i> Orthogroups	17,513 (50.2%)
Number of Single-Individual Orthogroups	4,919 (14.1%)
Number of Singletons	4,321 (12.4%)
Number of 'Carlos' Singletons	426 (1.2%)
Number of 'Noble' Singletons	607 (1.7%)
Number of 'Trayshed' Singletons	455 (1.3%)
Number of Common Grape Singletons	2,833 (8.1%)
Number of Non-Singleton Single-Species Orthogroups	598 (1.7%)

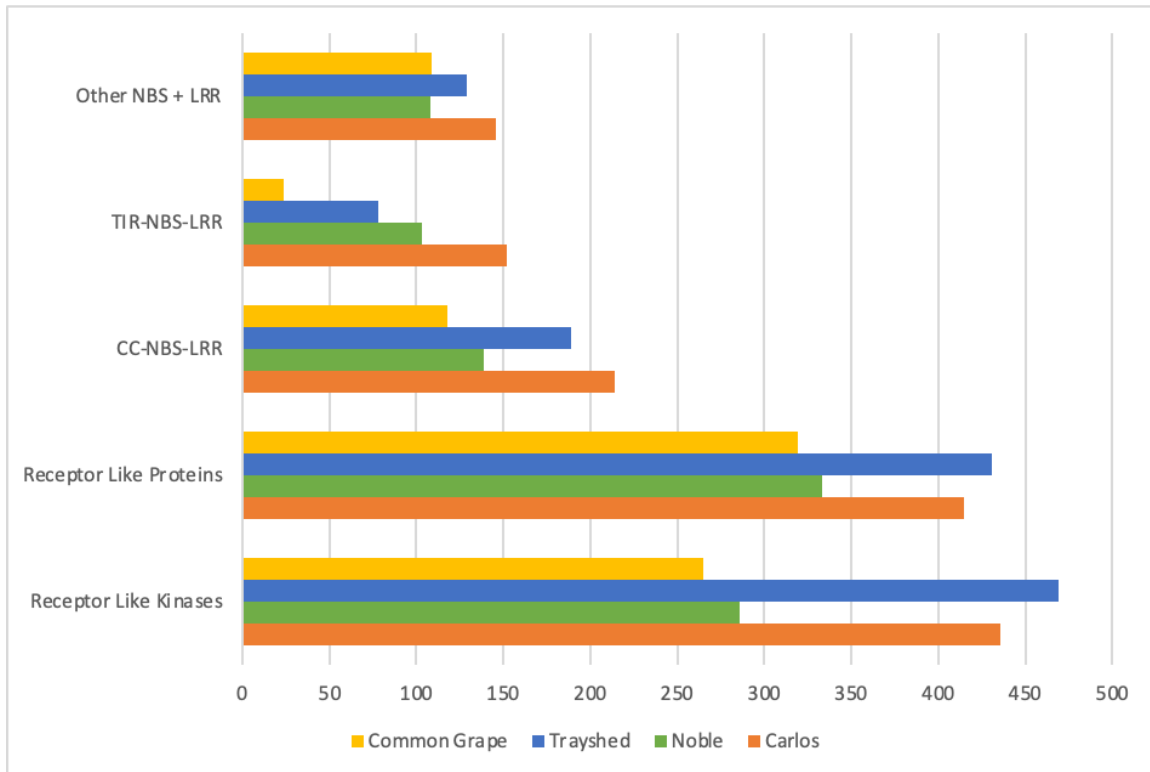


Figure 12. Gene Counts of R-Gene Classes for Three Muscadine Cultivars and Common Grape. Gene class was assigned based on the set of domains found in the gene (Supplemental Table S16).

Table 11. Differential expression results in *Vitis rotundifolia* 'Carlos' and *Vitis vinifera*.

	Up-Regulated	Down-Regulated	Total	Representative Orthogroups
Muscadine (Control vs. Inoculated)	89	145	234	208
Grape (Control vs Inoculated)	66	31	97	78
Grape (Inoculated vs wzy KO)	206	12	218	165
Grape (Control vs wzy KO)	124	11	135	102

were mostly up-regulated, though some genes activated by ABA were down-regulated. Sixteen of the differentially expressed genes were found to have R-gene domains, 5 of which were assigned to the Kinase R-gene class, indicating a plant defense response after inoculation with *Xfsf* (Table 12). Of these candidate R-genes, 9 were down-regulated and 7 were up-regulated. VITMroCarlos_v1.3.g17552, an up-regulated gene with RLP domains, is predicted to function as Eix2, a receptor driving disease resistance in plants (Ron & Avni, 2004). GhostKOALA, version 2.2 (Kanehisa et al., 2016), analysis of the differentially expressed genes indicated an enrichment of pathways associated with metabolism, cell signaling, and genetic processing. Up-regulated genes were associated with the Toll and immune deficiency (IMD) signaling pathways, ubiquitin-mediated proteolysis, and terpenoid biosynthesis. The Toll and IMD pathways are consistent with up-regulation of R-genes containing Toll domains, and plant terpenoids have well-characterized roles in plant defense (Singh & Sharma, 2015). Down-regulated genes were associated with biosynthesis of phenylpropanoids and other plant secondary metabolites, plant-pathogen interactions, and cell cycle processes.

In common grape, tolerance can be induced in otherwise susceptible cultivars by knockout of the *wzy* gene in *Xfsf*, which results in truncation of the complex, O-chain antigen that delays PAMP-triggered immunity (PTI) (Clifford et al., 2013; Rapicavoli et al., 2018). It is currently unclear how muscadine grapes, which have greater tolerance to wild type *Xfsf*, are able to bypass the O-chain antigen, or if it has an alternative method of tolerance. To gain insight into the possible differences or overlaps of these two grape species to *Xsfs*, we examined the overlap of the gene orthogroups between our dataset and data from Rapicavoli et al., 2018. Rapicavoli et al. 2018 reported an RNASeq experiment with inoculation of 'Cabernet Sauvignon' grapes with wild type *WZY Xfsf*, mutant *wzy Xfsf*, or control buffer. RNA sequencing in tissue both local and systemic to the site of inoculation spanned collection times ranging from hours to 4 weeks after inoculation (Rapicavoli et al., 2018). We selected data that mirrored our experimental design, i.e. systemic tissue collected four weeks after inoculation, for comparison.

Genes were compared between species by orthogroups, with each orthogroup representing a common ancestral grape gene for all orthogroup members. The 25,556 orthogroups identified by OrthoFinder analysis were filtered for orthogroups containing the genes of interest - including orthogroups containing two or more differentially expressed genes - and an overlap between all four conditions was determined (Figure 13). Most orthogroups containing muscadine DE genes were unique to the muscadine response, with any grape gene members not reported as differentially expressed. Only 1 orthogroup, OG0000266, shared genes that were differentially expressed (all up-regulated) in all four comparisons (Table 13). Functional annotation of the muscadine DE gene defined it as a Dirigent protein, a plant defense gene associated with lignin biosynthesis (N. Li et al., 2017; Paniagua et al., 2017). Four orthogroups contained genes that were differentially expressed in the three comparisons associated with greater tolerance to PD. Functional annotation of these orthogroups of interest identified more genes associated with plant pathogen response, including Snakin-1 and an auxin-binding protein (Table 13). Half of these genes were up-regulated in muscadine, and the other half were down-regulated. A full list of orthogroups, including those containing DE muscadine genes, is available in Supplemental Table S18.

The 16 differentially expressed R-gene candidates from the 'Carlos' DE gene set were placed into 13 orthogroups in this analysis. While most of these orthogroups contained genes

Table 12. Placement of candidate R-genes in Orthogroups. RLP = Receptor-like protein; KIN = Kinase; CLK = Coiled-coil, leucine-rich repeat, kinase; RLK = Receptor-like kinase; CK = Coiled-coil, kinase; TNL = Toll/Interleukin receptor, nucleotide binding site, leucine-rich repeat. Log fold change from 'Carlos' RNASeq experiment, with positive values indicating up regulation in inoculated samples and vice versa. Grape and 'Carlos' Gene Count are the total number of genes from each found in the orthogroup. Differentially expressed genes that are not present in common grape are shown in bold.

Class	Orthogroup	Candidate R-Gene	LogFoldChange	Grape Gene Count	'Carlos' Gene Count
CK	OG0002568	VITMroCarlos_v1.3.g25887	-2.1	2	2
	OG0009426	VITMroCarlos_v1.3.g15699	-1.8	1	1
CLK	OG0008270	VITMroCarlos_v1.3.g12650	0.9	5	5
KIN	OG0000032	VITMroCarlos_v1.3.g16025	3.6	29	32
	OG0000484	VITMroCarlos_v1.3.g10347	-1.3	2	3
	OG0005040	VITMroCarlos_v1.3.g19102	-2.2	1	2
	OG0013253	VITMroCarlos_v1.3.g17633	-6.8	0	9
		VITMroCarlos_v1.3.g17647	-7.0	0	9
	OG0015829	VITMroCarlos_v1.3.g20953	6.5	0	5
RLK	OG0000068	VITMroCarlos_v1.3.g21507	3.2	14	19
	OG0002084	VITMroCarlos_v1.3.g15393	-2.2	2	2
		VITMroCarlos_v1.3.g24514	-5.2	2	2
RLP	OG0000075	VITMroCarlos_v1.3.g01514	-1.1	9	6
	OG0005036	VITMroCarlos_v1.3.g17550	3.2	2	9
		VITMroCarlos_v1.3.g17552	3.3	2	9
TNL	OG0008267	VITMroCarlos_v1.3.g24911	3.2	1	8

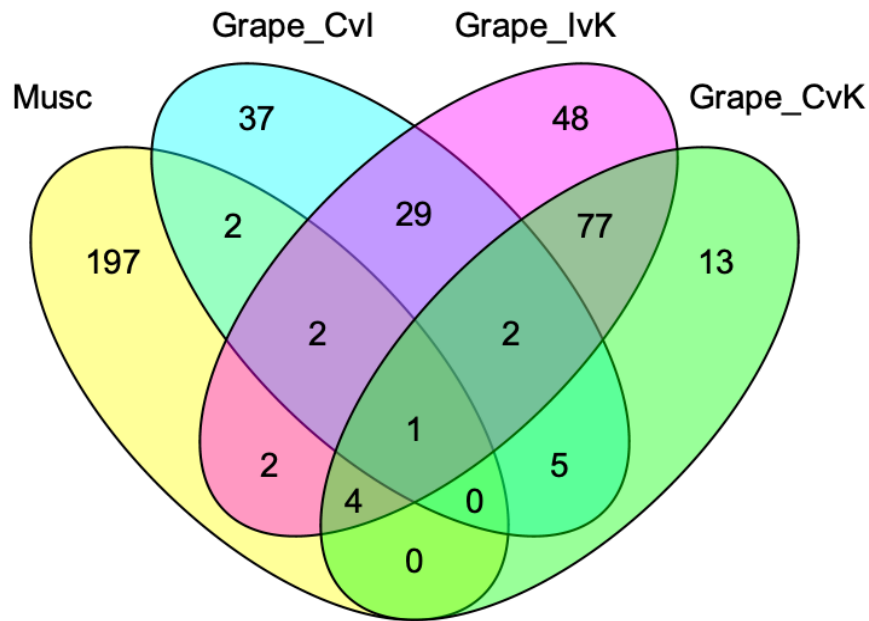


Figure 13. Overlap of orthogroups associated with differentially expressed genes in *Vitis rotundifolia* 'Carlos' and *Vitis vinifera*. Orthogroups contained one or more genes that were differentially expressed in the comparison. “Musc” = Muscadine Control vs. Inoculated. “Grape_CvI” = Grape Control vs. Inoculated (Wild-type). “Grape_IvK” = Grape Inoculated (Wild-type) vs. Inoculated (wzy Knockout). “Grape_CvK” = Grape Control vs. Inoculated (wzy Knockout).

Table 13. Functional annotation of orthogroups of interest. Functional annotation was performed using EnTAP version 0.10.8. The listed Gene IDs include only the genes that are differentially expressed. “Musc” indicates the Muscadine control vs. inoculated comparison, “Grape IvK” indicates the wild-type *X. fastidiosa* subsp. *fastidiosa* inoculated Grape vs. wzy knockout strain inoculated Grape comparison, and “Grape CvK” indicates the Grape control vs. wzy knockout strain inoculated Grape comparison.

Orthogroup ID	<i>V. vinifera</i> DE Gene IDs	<i>V. rotundifolia</i> (Carlos) DE Gene IDs	DE in Comparisons	Protein Function	DE in Infected Muscadine
OG0000266	VIT_206s0004g01000.1, VIT_206s0004g01015.1	VITMroCarlos _v1.3.g23332	All	Dirigent Protein, Disease resistance response protein	Up
OG0000539	VIT_218s0001g14810.1	VITMroCarlos _v1.3.g24572	Musc, Grape IvK, and Grape CvK	Lipase_GDSL domain- containing proteinGDSL esterase lipase	Down
OG0000585	VIT_207s0129g00580.1	VITMroCarlos _v1.3.g18058	Musc, Grape IvK, and Grape CvK	Snakin-1, GASA Family protein	Down
OG0001153	VIT_212s0034g01890.1, VIT_212s0034g01930.1, VIT_212s0034g01950.1, VIT_212s0034g01970.1	VITMroCarlos _v1.3.g00557	Musc, Grape IvK, and Grape CvK	Globulin seed storage protein 2	Up
OG0011935	VIT_209s0002g01320.1	VITMroCarlos _v1.3.g13207	Musc, Grape IvK, and Grape CvK	Germin-like protein, auxin- binding protein ABP19a-like	Up

from the common grape assembly, only the muscadine genes were differentially expressed in response to infection with *Xfsf* (Table 12). One orthogroup contained more genes in common grapes compared to 'Carlos', and 5 contained an equal number of genes between the two species. The remaining orthogroups contained more genes from 'Carlos' than in common grapes. Of these, 2 orthogroups - OG0013253 and OG0015829 - contained no genes from the common grape annotation, indicating proteins novel to the muscadine annotation (Table 12, bold). These R-genes are predicted to function as kinases - from both the annotation results and the R-gene analysis - with the R-genes of OG0013253 being down-regulated in response to *Xfsf* and the R-gene of OG0015829 being up-regulated. Another orthogroup, OG0005036, contained VITMroCarlos_v1.3.g17552, previously established as a predicted plant defense gene. Orthogroup analysis indicates an expansion in this family in 'Carlos' compared to common grapes, supporting an R-gene expansion.

Discussion

Pierce's Disease, caused by the gram-negative bacteria *Xylella fastidiosa subsp. fastidiosa*, represents a limiting factor in the production of Californian common grapes, yet is inconsequential to wild and commercial muscadine grapes grown as a native food crop in the Southeast. While current understanding of muscadine's mechanisms of tolerance and resistance is limited, current research suggests that muscadine grapes produce novel proteins that improve host tolerance (Katam et al., 2015) or produce tyloses to prevent spread of the bacteria (Mollenhauer & Hopkins, 1976). A better understanding of genetic and molecular resistance mechanisms will allow researchers to improve tolerance in common grape varieties through either gene introgression or selection of traits through breeding. To that end, we have worked to expand our existing knowledge of the muscadine genome and PD tolerance by providing a genome assembly and *Xfsf* response transcriptome for the tolerant cultivar 'Carlos'.

The 'Carlos' muscadine genome assembly is the third public reference genome from muscadine, following the 'Trayshed' and 'Noble' reference genomes (Cochetel et al., 2021; Park et al., 2022). Using the 'Trayshed' assembly as a guide, our assembly anchored over 99% of sequenced bases into 20 chromosome-level scaffolds. By default, this method resulted in chromosome-level scaffolds, but still containing a number of gap regions. We were able to resolve a number of the gaps using raw PacBio reads, resulting in an assembly with a low percentage of uncalled bases. Structural comparison of the 'Carlos' assembly to the 'Trayshed' assembly showed that, because of the high-quality contig assembly, we were able to identify structural variants between the two assemblies (Cochetel et al., 2021). Comparison of 'Carlos' with the 'Noble' assembly is consistent with the collinear analysis from the 'Noble' paper (Park et al., 2022), finding most structural variants previously observed between 'Noble' and 'Trayshed' were also identified when comparing 'Carlos' and 'Noble'. Due to the 'Noble' assembly utilizing a combination Illumina and PacBio CLR reads - as opposed to our PacBio CLR-only approach - it is likely that some of these structural variations may be artifacts of a lower contiguity assembly in 'Noble'. Supporting this, the contig N50 of the 'Noble' assembly is 105.8 KB, while the contig N50 for the 'Carlos' assembly is 11.3 MB. Recent developments in long-read sequencing, in particular the PacBio HiFi technology, have resulted in more accurate assemblies using long-reads alone than either hybrid- or short-read assemblies only (Hon et al., 2020; Rayamajhi et al., 2022). As accurate long-read sequencing becomes more common, future genome assemblies will

provide more accurate insights into true structural variation, as genome assembly quality is a key factor in these pan-genome type comparisons (Mascher et al., 2021).

Ribosomal gene arrays (rDNA arrays) are fundamental parts of genome architecture with high tandem copy numbers. In plants, the two array types - 5S and 35S (18S-5.8S-26S) - can occur in one or a few locations in the genome. In *Vitis* species, rDNA arrays have been used to profile chromosome architecture and to infer phylogenies, and results of the available assemblies indicate the number and length of both array types varies within *V. vinifera* cultivars and across *Vitis* species (Castro et al., 2018; Pierozzi, 2011). These arrays represent a challenge to most assembly algorithms, as their total length may exceed even long read sequencing reads. To build on the findings of the existing muscadine genome assemblies, we compared placement of rDNA arrays along the 'Carlos' assembly to those of existing muscadine and grape cultivars. We found some overlap in rDNA locations between our 'Carlos' assembly and the 'Trayshed' and 'Noble' muscadine assemblies, particularly in regard to the placement of 5S sequences, but found that each of the existing genomes missed one of the two 35S arrays we identified in 'Carlos'. Often, rDNA array presence was associated with large structural variations (Figure 10); of note, the NOR along Chromosome 15 in 'Carlos' is placed in a different region of the same chromosome in the 'Noble' assembly. In the case of 'Trayshed', the region containing the array on Chromosome 15 was not scaffolded to the current assembly. These differences in rDNA array locations between otherwise similar genomes highlight the difficulty in capturing these repetitive regions in genome assemblies. Similarly, we found that NORs could not be reliably found in the common grapevine version 12.1 genome. Further cytogenetic work in these specific cultivars is needed to confirm the actual 5S and 35S loci locations, confirm nearby structural variants (or indicate errors in assembly), and tie the reference genomes back to the chromosomal karyotypes.

There are over 100 cultivars of muscadine available in and beyond its native range (Andersen et al., 2020). Recent studies have found substantial genetic diversity within muscadine cultivars (Buck & Worthington, 2022; Cao et al., 2020). Attempting to use a single one of the three available genomes to characterize every cultivar will result in gaps in knowledge due to the phenotypic and genotypic differences between each cultivar (Bayer et al., 2020; Hufford et al., 2021). The solution is to further develop resources for these three cultivars - as well as additional cultivars - to develop a pan-genome. Pan-genomes cover the complete gene content of multiple cultivars of a single species, or multiple species of a closely-related genus. The pan-genome approach has been applied with great success to common crops such as maize, tomato, and rice (L. Gao et al., 2019; Woodhouse et al., 2021; Zhao et al., 2018; Zhou et al., 2020). In the context of *Vitis* species, pan-genome data can be used to better understand the mechanisms driving disease resistance, drought tolerance, and other traits of interest. For this study, we produced a foundational muscadine pan-genome consisting of the three muscadine cultivar annotations, as well as the common grape annotation. Of the 34,970 syntenic-constrained orthogroups, almost half contain at least one gene from all four annotations, representing the "core" gene set across all four genomes. The other orthogroups represent "accessory" genes, including 4,919 orthogroups containing only genes from a single genome. As more grape, muscadine, and other *Vitis* genomes continue to be produced, we will add new *Vitis* genomes to our pan-genome, and our understanding of what genes are currently considered "core" and "accessory" to evolve.

In common grapes, pattern-associated molecular pattern (PAMP)-triggered immunity (PTI) is delayed by a long O-chain antigen associated with lipopolysaccharides (LPSs) in *Xfsf*.

Knock-out of the *wzy* gene results in grapevines mounting a faster immune response and having greater tolerance to infection (Rapicavoli et al., 2018). By contrast, muscadine cultivars range from tolerant to resistant against the bacteria, indicating that it is successfully detecting the pathogen by overcoming the O-chain antigen or utilizing another mechanism of detection. Plant resistance (R) genes are of particular importance to pathogen detection and response signaling (Pandolfi et al., 2017), and the immunity observed in muscadine cultivars could be attributed to an expanded suite of R-genes compared to the R-genes of common grapes (Feechan et al., 2013). The ‘Trayshed’ assembly paper reported an expansion of R genes in muscadine grapes compared to common grapes (Cochetel et al., 2021). Analysis of R-genes in the ‘Carlos’ assembly supports this expansion, with more R-genes identified in the ‘Carlos’ assembly than in ‘Trayshed’. The Trayshed genome paper described an expansion of Toll/Interleukin-1 Receptor-like Nucleotide-Binding Site Leucine-Rich Repeat (TIR-NBS-LRR) genes compared to the common grape. R-gene analysis of our gene annotation supported this expansion and indicated a nearly two-fold expansion of TIR-NBS-LRR genes in ‘Carlos’ compared to ‘Trayshed’. The ‘Noble’ assembly appears to have a contraction in most R-genes compared to other muscadine cultivars, though this may be the result of a different annotation strategy and deserves further exploration. Despite the overall lower number of R-genes compared to the other cultivars, ‘Noble’ had more proteins containing TIR-NBS-LRR domains compared to the ‘Trayshed’ assembly. The disparity in R-genes identified in Noble may be attributed to a lower quality assembly. An enriched set of R-genes in muscadine grape genomes supports the notion of improved disease resistance mechanisms in muscadine compared to the common grape. Further work will be needed to tease out candidates associated with improved response to infection with *Xfsf*.

The results of our gene expression analysis in ‘Carlos’, reveal that abiotic stress response, particularly drought tolerance, were up-regulated in ‘Carlos’ inoculated with *Xfsf*, with a few up-regulated genes associated with defense. Of these defense genes, we observed a candidate gene, VITMroCarlos_v1.3.g17552, that showed similarity to EIX2, a receptor-like protein that confers resistance to a fungal pathogen of tomato after detecting an elicitor protein (Ron & Avni, 2004). While LPSs are well-characterized elicitors in common grapes, they are not the only PAMP that can induce PTI (Burbank & Ochoa, 2022). Orthogroup placement of this gene indicates an expansion in this family compared to grape, which could point to an expanded suite of PAMP response elements in muscadine compared to common grape (Table 12). It is important to note that, instead of samples collected from the site of inoculation, our sequencing data measured systemic responses to *Xfsf* inoculation. These results, therefore, offer no conclusions on parenchymal cell response in the xylem, where initial detection of *Xfsf* occurs. Further research will be needed to fully characterize the muscadine response to inoculation with *Xfsf*. In particular, it will need to be determined if the tolerance observed in ‘Carlos’, as opposed to resistance in other cultivars, can be attributed to a different immune response, or if it is the result of greater drought tolerance compared to common grapes.

To contextualize how these observed genes overlap with the common grape response to *Xfsf*, we compared the gene expression of ‘Carlos’ muscadines inoculated with *Xfsf* to grapes inoculated with two strains: a wild-type strain and a *wzy*-knockout strain (Rapicavoli et al., 2018). Rather than attempting to map reads from one organism to the genome of another, we utilized OrthoFinder to identify genes with a common ancestor of both species. Of the 208 orthogroups containing differentially expressed genes in muscadine, 197 were unique to the

muscadine response to PD. Of note, 19 of these orthogroups contained no genes from the common grape annotation, suggesting them as interesting candidates for follow up study. For the remaining orthogroups with both muscadine and common grape members, unique changes in gene sequence or expression patterns may have emerged after the organisms diverged evolutionarily to drive a unique host response to PD.

Despite the variation in host susceptibility, some overlaps in DE genes between muscadine and common grape were found. One orthogroup contained genes that were differentially expressed across all four conditions, and four contained genes that were differentially expressed among all PD-tolerant responses (Figure 13). The single orthogroup representing all four conditions, OG0000266, was annotated as a Dirigent (DIR) protein. Among other important functions, plant (DIR) proteins play a role in plant defense response by modulating cell wall metabolism through lignin accumulation (N. Li et al., 2017). In muscadine, this gene was up-regulated in response to inoculation with *Xfsf*. The four orthogroups associated with PD-tolerant responses were also functionally annotated as genes associated with plant pathogen responses (Almasia et al., 2020; Cândido et al., 2011; M. Gao et al., 2017; Pei et al., 2019). Among these, the muscadine gene coding for a GDSL lipase was down-regulated, which is consistent with the observed DEGs. While GDSL lipases can play a role in plant defenses (Lee et al., 2009), some plant systems see improved host resistance when GDSL lipases are silenced (M. Gao et al., 2017). While there is overlap in genes differentially expressed in muscadines inoculated with *X. fastidiosa* subsp. *fastidiosa*, our main conclusion is that most differentially expressed genes were uniquely expressed in muscadine, supporting a unique host response not found in common grapes.

While hybrids between common grapes and muscadine grapes have been produced, these efforts are made complicated by a chromosome-fusion event in common grapes. By identifying specific genes with high association to PD-resistance in muscadine, researchers will be better able to focus their efforts on candidate resistance genes. Our efforts in this study have provided three major outcomes to advance this goal. First, our 'Carlos' assembly adds to the existing genome reference sequences and gene annotation sets available for muscadine grapes. Second, our muscadine pan-genome provides a new resource for comparative genomics between muscadine cultivars and those of the common grape, while allowing room for expansion. Third, we provide a comparison of differentially expressed genes that are informative in comparing the immune response between muscadine and common grapes. There is more work to be done to fully characterize the muscadine response to Pierce's disease, from an expanded time course at different times and locations after infection to host and pathogen knockout experiments. Our results have provided new resources to expand this understanding and provided a framework for future sequencing efforts.

REFERENCES

- Almasia, N. I., Nahirñak, V., Hopp, H. E., & Vazquez-Rovere, C. (2020). Potato Snakin-1: an antimicrobial player of the trade-off between host defense and development. *Plant Cell Reports*, 39(7), 839–849. <https://doi.org/10.1007/s00299-020-02557-5>
- Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, 20(1), 224. <https://doi.org/10.1186/s13059-019-1829-6>
- Andersen, P. C., Sarkhosh, A., Huff, D., & Breman, J. (2020). The Muscadine Grape (*Vitis rotundifolia* Michx): HS763/HS100, rev. 10/2020. *EDIS*. <https://journals.flvc.org/edis/article/download/123942/126076>
- Anders, S., Pyl, P. T., & Huber, W. (2010). *HTSeq: Analysing high-throughput sequencing data with Python*.
- Basha, S. M., Mazhar, H., & Vasanthaiah, H. K. N. (2010). Proteomics approach to identify unique xylem sap proteins in Pierce's disease-tolerant *Vitis* species. *Applied Biochemistry and Biotechnology*, 160(3), 932–944. <https://doi.org/10.1007/s12010-009-8620-1>
- Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6(8), 914–920. <https://doi.org/10.1038/s41477-020-0733-0>
- Blanc, S., Wiedemann-Merdinoglu, S., Dumas, V., Mestre, P., & Merdinoglu, D. (2012). A reference genetic map of *Muscadinia rotundifolia* and identification of Ren5, a new major locus for resistance to grapevine powdery mildew. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 125(8), 1663–1675. <https://doi.org/10.1007/s00122-012-1942-3>
- Blua, M. J., Phillips, P. A., & Redak, R. A. (1999). A new sharpshooter threatens both crops and ornamentals. In *California Agriculture* (Vol. 53, Issue 2, pp. 22–25). <https://doi.org/10.3733/ca.v053n02p22>
- Buck, K., & Worthington, M. (2022). Genetic Diversity of Wild and Cultivated Muscadine Grapes (*Vitis rotundifolia* Michx.). *Frontiers in Plant Science*, 13, 852130. <https://doi.org/10.3389/fpls.2022.852130>
- Burbank, L. P., & Ochoa, J. (2022). Evidence for Elicitation of an Oxidative Burst in *Vitis vinifera* by *Xylella fastidiosa* Cold Shock Protein Peptide csp20. *PhytoFrontiers™*, 2(4), 339–341. <https://doi.org/10.1094/PHYTOFR-05-22-0048-SC>
- Bushnell, B. (2018). BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. *Joint Genome Institute*.
- Caballero, M., & Wegrzyn, J. (2019). gFACs: Gene Filtering, Analysis, and Conversion to Unify Genome Annotations Across Alignment and Gene Prediction Frameworks. *Genomics, Proteomics & Bioinformatics*, 17(3), 305–310. <https://doi.org/10.1016/j.gpb.2019.04.002>
- Cândido, E. de S., Pinto, M. F. S., Pelegrini, P. B., Lima, T. B., Silva, O. N., Pogue, R., Grossi-de-Sá, M. F., & Franco, O. L. (2011). Plant storage proteins with antimicrobial activity: novel insights into plant defense mechanisms. *FASEB Journal: Official Publication of the Federation of American Societies for*

- Experimental Biology*, 25(10), 3290–3305. <https://doi.org/10.1096/fj.11-184291>
- Cao, S., Stringer, S., Gunawan, G., McGregor, C., & Conner, P. J. (2020). Genetic Diversity and Pedigree Analysis of Muscadine Grape Using SSR Markers. In *Journal of the American Society for Horticultural Science* (Vol. 145, Issue 3, pp. 143–151). <https://doi.org/10.21273/jashs04856-20>
- Castro, C., Carvalho, A., Pavia, I., Leal, F., Moutinho-Pereira, J., & Lima-Brito, J. (2018). Nucleolar activity and physical location of ribosomal DNA loci in *Vitis vinifera* L. by silver staining and sequential FISH. *Scientia Horticulturae*, 232, 57–62. <https://doi.org/10.1016/j.scienta.2017.12.064>
- Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in Molecular Biology*, 1962, 1–14. https://doi.org/10.1007/978-1-4939-9173-0_1
- Christopher, L. P., Yao, B., & Ji, Y. (2014). Lignin Biodegradation with Laccase-Mediator Systems. *Frontiers in Energy Research*, 2. <https://doi.org/10.3389/fenrg.2014.00012>
- Clifford, J. C., Rapicavoli, J. N., & Caroline Roper, M. (2013). A Rhamnose-Rich O-Antigen Mediates Adhesion, Virulence, and Host Colonization for the Xylem-Limited Phytopathogen *Xylella fastidiosa*. In *Molecular Plant-Microbe Interactions®* (Vol. 26, Issue 6, pp. 676–685). <https://doi.org/10.1094/mpmi-12-12-0283-r>
- Cochetel, N., Minio, A., Massonnet, M., & Vondras, A. M. (2021). Diploid chromosome-scale assembly of the *Muscadinia rotundifolia* genome supports chromosome fusion and disease resistance gene expansion during *Vitis* and G3 . <https://academic.oup.com/g3journal/article-abstract/11/4/jkab033/6129119>
- Delame, M., Prado, E., Blanc, S., Robert-Siegwald, G., Schneider, C., Mestre, P., Rustenholz, C., & Merdinoglu, D. (2019). Introgression reshapes recombination distribution in grapevine interspecific hybrids. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 132(4), 1073–1087. <https://doi.org/10.1007/s00122-018-3260-x>
- Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B., & Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement. *Genome Biology*, 22(1), 3. <https://doi.org/10.1186/s13059-020-02224-8>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Emms, D. (2019, September 21). *OrthoFinder best practices*. OrthoFinder Tutorials. https://davidemms.github.io/orthofinder_tutorials/orthofinder-best-practices.html
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238. <https://doi.org/10.1186/s13059-019-1832-y>
- Feechan, A., Anderson, C., Torregrosa, L., Jermakow, A., Mestre, P., Wiedemann-Merdinoglu, S., Merdinoglu, D., Walker, A. R., Cadle-Davidson, L., Reisch, B., Aubourg, S., Bentahar, N., Shrestha, B., Bouquet, A., Adam-Blondon, A.-F., Thomas, M. R., & Dry, I. B. (2013). Genetic dissection of a TIR-NB-LRR locus from the wild North American grapevine species *Muscadinia rotundifolia* identifies

- paralogous genes conferring resistance to major fungal and oomycete pathogens in cultivated grapevine. *The Plant Journal: For Cell and Molecular Biology*, 76(4), 661–674. <https://doi.org/10.1111/tpj.12327>
- Fry, S. M., & Milholland, R. D. (1990). Response of resistant, tolerant, and susceptible grapevine tissues to invasion by the Pierce's disease bacterium, *Xylella fastidiosa*. *Phytopathology*, 80(1), 66–69. https://www.apsnet.org/publications/phytopathology/backissues/Documents/1990Articles/Phyto80n01_66.pdf
- Fry, S. M., Milholland, R. D., & Others. (1990). Multiplication and translocation of *Xylella fastidiosa* in petioles and stems of grapevine resistant, tolerant, and susceptible to Pierce's disease. *Phytopathology*, 80(1), 61–65. https://www.apsnet.org/publications/phytopathology/backissues/Documents/1990Articles/Phyto80n01_61.pdf
- Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L., Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., van der Knaap, E., Huang, S., Klee, H. J., ... Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, 51(6), 1044–1051. <https://doi.org/10.1038/s41588-019-0410-2>
- Gao, M., Yin, X., Yang, W., Lam, S. M., Tong, X., Liu, J., Wang, X., Li, Q., Shui, G., & He, Z. (2017). GDLS lipases modulate immunity through lipid homeostasis in rice. *PLoS Pathogens*, 13(11), e1006724. <https://doi.org/10.1371/journal.ppat.1006724>
- Goel, M., & Schneeberger, K. (2022). plotsr: Visualising structural similarities and rearrangements between multiple genomes. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btac196>
- Goel, M., Sun, H., Jiao, W.-B., & Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, 20(1), 277. <https://doi.org/10.1186/s13059-019-1911-0>
- Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., & Chen, J. (2020). RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Computer Science*, 6, e251. <https://doi.org/10.7717/peerj-cs.251>
- Hart, A. J., Ginzburg, S., Xu, M. (sam), Fisher, C. R., Rahmatpour, N., Mitton, J. B., Paul, R., & Wegrzyn, J. L. (2020). EnTAP : Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Molecular Ecology Resources*, 20(2), 591–604. <https://doi.org/10.1111/1755-0998.13106>
- Hoffmann, M. (2020). *Muscadine grape production guide for the Southeast*.
- Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D. R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, 7(1), 399. <https://doi.org/10.1038/s41597-020-00743-4>
- Hopkins, D. L., & Purcell, A. H. (2002). *Xylella fastidiosa*: Cause of Pierce's Disease of Grapevine and Other Emergent Diseases. *Plant Disease*, 86(10), 1056–1066. <https://doi.org/10.1094/PDIS.2002.86.10.1056>

- Huff, M., Seaman, J., Wu, D., Zhebentyayeva, T., Kelly, L. J., Faridi, N., Nelson, C. D., Cooper, E., Best, T., Steiner, K., Koch, J., Romero Severson, J., Carlson, J. E., Buggs, R., & Staton, M. (2022). A high-quality reference genome for *Fraxinus pennsylvanica* for ash species restoration and research. *Molecular Ecology Resources*, 22(4), 1284–1302. <https://doi.org/10.1111/1755-0998.13545>
- Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., ... Dawe, R. K. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, 373(6555), 655–662. <https://doi.org/10.1126/science.abg5289>
- Ignacio, D. N., Mason, K. D., Hackett-Morton, E. C., Albanese, C., Ringer, L., Wagner, W. D., Wang, P. C., Carducci, M. A., Kachhap, S. K., Paller, C. J., Mendonca, J., Li-Ying Chan, L., Lin, B., Hartle, D. K., Green, J. E., Brown, C. A., & Hudson, T. S. (2019). Muscadine grape skin extract inhibits prostate cancer cells by inducing cell-cycle arrest, and decreasing migration through heat shock protein 40. *Heliyon*, 5(1), e01128. <https://doi.org/10.1016/j.heliyon.2019.e01128>
- Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., ... French-Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161), 463–467. <https://doi.org/10.1038/nature06148>
- Jiang, H., Lei, R., Ding, S.-W., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15, 182. <https://doi.org/10.1186/1471-2105-15-182>
- Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, 428(4), 726–731. <https://doi.org/10.1016/j.jmb.2015.11.006>
- Katam, R., Chibanguza, K., Latinwo, L. M., & Smith, D. (2015). Proteome Biomarkers in Xylem Reveal Pierce's Disease Tolerance in Grape. *Journal of Proteomics & Bioinformatics*, 8(9), 217–224. <https://doi.org/10.4172/jpb.1000372>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods in Molecular Biology*, 1962, 161–177. https://doi.org/10.1007/978-1-4939-9173-0_9
- 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 Research*, 27(5), 722–736. <https://doi.org/10.1101/gr.215087.116>
- Krivanek, A. F., & Walker, M. A. (2005). Vitis Resistance to Pierce's Disease Is

- Characterized by Differential *Xylella fastidiosa* Populations in Stems and Leaves. *Phytopathology*, 95(1), 44–52. <https://doi.org/10.1094/PHYTO-95-0044>
- Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9), 3100–3108. <https://doi.org/10.1093/nar/gkm160>
- Lee, D. S., Kim, B. K., Kwon, S. J., Jin, H. C., & Park, O. K. (2009). Arabidopsis GDGL lipase 2 plays a role in pathogen defense via negative regulation of auxin signaling. *Biochemical and Biophysical Research Communications*, 379(4), 1038–1042. <https://doi.org/10.1016/j.bbrc.2009.01.006>
- Lewter, J., Worthington, M. L., Clark, J. R., Varanasi, A. V., Nelson, L., Owens, C. L., Conner, P., & Gunawan, G. (2019). High-density linkage maps and loci for berry color and flower sex in muscadine grape (*Vitis rotundifolia*). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 132(5), 1571–1585. <https://doi.org/10.1007/s00122-019-03302-7>
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, N., Zhao, M., Liu, T., Dong, L., Cheng, Q., Wu, J., Wang, L., Chen, X., Zhang, C., Lu, W., Xu, P., & Zhang, S. (2017). A Novel Soybean Dirigent Gene GmDIR22 Contributes to Promotion of Lignan Biosynthesis and Enhances Resistance to *Phytophthora sojae*. *Frontiers in Plant Science*, 8, 1185. <https://doi.org/10.3389/fpls.2017.01185>
- Lovell, J. T., Sreedasyam, A., Eric Schranz, M., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. In *eLife* (Vol. 11). <https://doi.org/10.7554/elife.78526>
- Love, M., Anders, S., & Huber, W. (2014). Differential analysis of count data--the DESeq2 package. *Genome Biology*, 15(550), 10–1186. <https://www.nanjxs.com/www/packages/3.3/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf>
- Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5), 955–964. <https://doi.org/10.1093/nar/25.5.955>
- Luo, J., Song, S., Wei, Z., Huang, Y., Zhang, Y., & Lu, J. (2017). The comparative study among different fractions of muscadine grape “Noble” pomace extracts regarding anti-oxidative activities, cell cycle arrest and apoptosis in breast cancer. *Food & Nutrition Research*, 61(1), 1412795. <https://doi.org/10.1080/16546628.2017.1412795>
- Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C. S., Ens, J., Gundlach, H., Boston, L. B., Tulpová, Z., Holden, S., Hernández-Pinzón, I., Scholz, U., Mayer, K. F. X., Spannagl, M., Pozniak, C. J., Sharpe, A. G., Šimková, H., Moscou, M. J., ...

- Stein, N. (2021). Long-read sequence assembly: a technical evaluation in barley. *The Plant Cell*, 33(6), 1888–1906. <https://doi.org/10.1093/plcell/koab077>
- Mavrodi, O. V., Mavrodi, D. V., Stafne, E. T., Adamczyk, J. J., & Babiker, E. M. (2020). Draft Genome Sequences of *Xylella fastidiosa* subsp. *fastidiosa* Strains OK3, VB11, and NOB1, Isolated from Bunch and Muscadine Grapes Grown in Southern Mississippi. *Microbiology Resource Announcements*, 9(25). <https://doi.org/10.1128/MRA.00562-20>
- Mendonca, P., Darwish, A. G., Tsoleva, V., El-Sharkawy, I., & Soliman, K. F. A. (2019). The Anticancer and Antioxidant Effects of Muscadine Grape Extracts on Racially Different Triple-negative Breast Cancer Cells. *Anticancer Research*, 39(8), 4043–4053. <https://doi.org/10.21873/anticancer.13560>
- Minsavage, G.V., Thompson, C.M., Hopkins, D.L., Leite, R., and Stall, R.E. (1994). Development of a Polymerase Chain Reaction Protocol for Detection of *Xylella fastidiosa* in Plant Tissue. *Phytopathology*, 84(5), 456–461. <https://doi.org/10.1094/Phyto-84-456>
- Mollenhauer, H. H., & Hopkins, D. L. (1976). Xylem morphology of Pierce's disease-infected grapevines with different levels of tolerance. *Physiological Plant Pathology*, 9(1), 95–100. [https://doi.org/10.1016/0048-4059\(76\)90079-5](https://doi.org/10.1016/0048-4059(76)90079-5)
- Muhammad Aslam, M., Waseem, M., Jakada, B. H., Okal, E. J., Lei, Z., Saqib, H. S. A., Yuan, W., Xu, W., & Zhang, Q. (2022). Mechanisms of Absciscic Acid-Mediated Drought Stress Responses in Plants. *International Journal of Molecular Sciences*, 23(3). <https://doi.org/10.3390/ijms23031084>
- Osuna-Cruz, C. M., Paytuví-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Aiese Cigliano, R., Sanseverino, W., & Ercolano, M. R. (2018). PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Research*, 46(D1), D1197–D1201. <https://doi.org/10.1093/nar/gkx1119>
- Pandolfi, V., Neto, J. R. C. F., da Silva, M. D., Amorim, L. L. B., Wanderley-Nogueira, A. C., de Oliveira Silva, R. L., Kido, E. A., Crovella, S., & Iseppon, A. M. B. (2017). Resistance (R) Genes: Applications and Prospects for Plant Biotechnology and Breeding. *Current Protein & Peptide Science*, 18(4), 323–334. <https://doi.org/10.2174/1389203717666160724195248>
- Paniagua, C., Bilkova, A., Jackson, P., Dabravolski, S., Riber, W., Didi, V., Houser, J., Gigli-Bisceglia, N., Wimmerova, M., Budínská, E., Hamann, T., & Hejatko, J. (2017). Dirigent proteins in plants: modulating cell wall metabolism during abiotic and biotic stress exposure. *Journal of Experimental Botany*, 68(13), 3287–3301. <https://doi.org/10.1093/jxb/erx141>
- Park, M., Vera, D., Kambrianda, D., Gajjar, P., Cadle-Davidson, L., Tsoleva, V., & El-Sharkawy, I. (2022). Chromosome-level genome sequence assembly and genome-wide association study of *Muscadinia rotundifolia* reveal the genetics of 12 berry-related traits. *Horticulture Research*. <https://doi.org/10.1093/hr/uhab011>
- Pei, Y., Li, X., Zhu, Y., Ge, X., Sun, Y., Liu, N., Jia, Y., Li, F., & Hou, Y. (2019). GhABP19, a Novel Germin-Like Protein From *Gossypium hirsutum*, Plays an Important Role in the Regulation of Resistance to *Verticillium* and *Fusarium* Wilt

- Pathogens. *Frontiers in Plant Science*, 10. <https://doi.org/10.3389/fpls.2019.00583>
- Pierozzi, N. I. (2011). Karyotype and NOR-banding of mitotic chromosomes of some *Vitis* L. species. In *Revista Brasileira de Fruticultura* (Vol. 33, Issue spe1, pp. 564–570). <https://doi.org/10.1590/s0100-29452011000500077>
- Rapicavoli, J. N., Blanco-Ulate, B., Muszyński, A., Figueroa-Balderas, R., Morales-Cruz, A., Azadi, P., Dobruchowska, J. M., Castro, C., Cantu, D., & Roper, M. C. (2018). Lipopolysaccharide O-antigen delays plant innate immune recognition of *Xylella fastidiosa*. *Nature Communications*, 9(1), 390. <https://doi.org/10.1038/s41467-018-02861-5>
- Rayamajhi, N., Cheng, C.-H. C., & Catchen, J. M. (2022). Evaluating Illumina-, Nanopore-, and PacBio-based genome assembly strategies with the bald notothen, *Trematomus borchgrevinki*. *G3*, 12(11). <https://doi.org/10.1093/g3journal/jkac192>
- Reimer, F. C., Hume, H. H., Michels, J., Smith, R. I., & Curtis, R. S. (1906). *Scuppernong and Other Muscadine Grapes: Origin and Importance*. North Carolina Agricultural Experiment Station of the College of Agriculture and Mechanic Arts. <https://play.google.com/store/books/details?id=fsdQAQAAMAAJ>
- Ron, M., & Avni, A. (2004). The receptor for the fungal elicitor ethylene-inducing xylanase is a member of a resistance-like gene family in tomato. *The Plant Cell*, 16(6), 1604–1615. <https://doi.org/10.1105/tpc.022475>
- Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology*, 1962, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14
- Sharma, B., Joshi, D., Yadav, P. K., Gupta, A. K., & Bhatt, T. K. (2016). Role of Ubiquitin-Mediated Degradation System in Plant Biology. *Frontiers in Plant Science*, 7, 806. <https://doi.org/10.3389/fpls.2016.00806>
- Singh, B., & Sharma, R. A. (2015). Plant terpenes: defense responses, phylogenetic analysis, regulation and clinical applications. In 3 *Biotech* (Vol. 5, Issue 2, pp. 129–151). <https://doi.org/10.1007/s13205-014-0220-2>
- Smit, A. F. A., Hubley, R., & Green, P. (2015a). *RepeatMasker Open-4.0. 2013--2015*.
- Smit, A. F. A., Hubley, R., & Green, P. (2015b). *RepeatModeler Open-1.0. 2008--2015*. Seattle, USA: Institute for Systems Biology. Available from: [Httpwww.Repeatmasker.Org](http://www.Repeatmasker.Org), Last Accessed May, 1, 2018.
- Staton, M., Addo-Quaye, C., Cannon, N., Yu, J., Zhebentyayeva, T., Huff, M., Islam-Faridi, N., Fan, S., Georgi, L. L., Dana Nelson, C., Bellis, E., Fitzsimmons, S., Henry, N., Drautz-Moses, D., Noorai, R. E., Ficklin, S., Saski, C., Mandal, M., Wagner, T. K., ... Carlson, J. E. (2020). “A reference genome assembly and adaptive trait analysis of *Castanea mollissima* ‘Vanuxem,’ a source of resistance to chestnut blight in restoration breeding.” In *Tree Genetics & Genomes* (Vol. 16, Issue 4). <https://doi.org/10.1007/s11295-020-01454-y>
- Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, 9(11), e112963. <https://doi.org/10.1371/journal.pone.0112963>

- Woodhouse, M. R., Cannon, E. K., Portwood, J. L., 2nd, Harper, L. C., Gardiner, J. M., Schaeffer, M. L., & Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, 21(1), 385. <https://doi.org/10.1186/s12870-021-03173-5>
- Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., & Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature Methods*, 14(11), 1072–1074. <https://doi.org/10.1038/nmeth.4432>
- Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B. A., Fan, G., Liu, X., Xu, X., Deng, L., & Zhang, Y. (2020). TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*, 9(9). <https://doi.org/10.1093/gigascience/giaa094>
- Zaini, P. A., Nascimento, R., Gouran, H., Cantu, D., Chakraborty, S., Phu, M., Goulart, L. R., & Dandekar, A. M. (2018). Molecular Profiling of Pierce's Disease Outlines the Response Circuitry of *Vitis vinifera* to *Xylella fastidiosa* Infection. In *Frontiers in Plant Science* (Vol. 9). <https://doi.org/10.3389/fpls.2018.00771>
- Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., ... Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, 50(2), 278–284. <https://doi.org/10.1038/s41588-018-0041-z>
- Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., Parakkal, P., Arbelaez, L. J., Franco, N., Alexandrov, N., Hamilton, N. R. S., Leung, H., Mauleon, R., Lorieux, M., Zuccolo, A., McNally, K., ... Wing, R. A. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. *Scientific Data*, 7(1), 113. <https://doi.org/10.1038/s41597-020-0438-2>

CHAPTER VI: CONCLUSION

The field of plant genomics has expanded in the past few years with the introduction of new sequencing technologies and pipelines to analyze them. In these three chapters, I have demonstrated the use of these methods to produce and annotate genomes for the purpose of answering meaningful, biological questions relating to our species of interest. In the first chapter of my research, I utilized a chromosome-scale genome assembly to generate a gene annotation for green ash (*Fraxinus pennsylvanica*), a species that is threatened by the invasive emerald ash borer (EAB; *Agrilus planipennis*) (Chapter III). As part of this research, I also generated chromosome-level scaffolds for other ash species, using our genome as a reference, to provide new resources for these species. In the following chapter, I used this reference genome to search for regions in the genome associated with dioecy (Chapter IV). While I have found a region of interest on Chromosome 8, I am interested in doing more work to confirm this region of interest. In the final chapter, I performed a long-read assembly of muscadine grape (*Vitis rotundifolia*, cv Carlos) and laid the foundation for a *Vitis* species pan-genome (Chapter V). Complementing this work, I performed a functional comparison of 'Carlos' and common grapevine (*Vitis vinifera*) in response to inoculation with the bacterial pathogen *Xylella fastidiosa* subsp. *fastidiosa*.

One aspect of this research that I want to continue developing further is the concept of a pan-genome. Rather than taking one specific reference as the guideline for an entire species, pan-genomes use genomes from multiple cultivars in a species to identify genes shared by all members of the species ("core" genes) versus those that are not ("accessory/variable" genes). The pan-genome approach is also applicable at the genus-level, comparing gene content between different species in a closely-related genus. In Chapter V, I utilized this approach in muscadine to compare the genetic content of three cultivars of muscadine ('Carlos', 'Noble', and 'Trayshed') as well as that of common grapevine. Completion of this pan-genome allowed me to confirm an expansion of resistance-genes in most muscadine genomes compared to the grape genome assembly. This pan-genome is far from complete - there are multiple cultivars of muscadine and grape available, and the internal diversity of muscadine alone is robust - and I aim to add new *Vitis* sp. genomes as they become available. As more *Fraxinus* sp. genomes become available, I aim to generate a similar ash pan-genome to compare gene content between and within the different sections of the genus. In the context of EAB research, most American ash species are incompatible with Asian species with resistance to EAB. A pan-genome would be a useful tool to identify expanded gene families in EAB-resistant species compared to EAB-susceptible species.

I also predict a continued focus on assemblies that are built with long-read sequences. Comparing the 'Carlos' muscadine assembly - which used PacBio CLR sequencing - to the Illumina short-read assembly of green ash, I found that the muscadine assembly had a higher contiguity and fewer gap sequences compared to the green ash assembly. Short-read assemblies have been observed to have difficulties placing repetitive regions, and I observed this in the green ash assembly and its inability to place the rDNA array in a chromosome-level scaffold. By contrast, the muscadine genome had no difficulty placing two separate nucleolar organizing regions (NORs) into their respective chromosomes. Further development of genomic resources for green ash will be necessary to characterize sex determination at the seedling stage, as Chapter IV demonstrated the limited ability to distinguish markers of sex determination with an unphased

genome. Due to the nature of sex-determining regions (SDRs) in plants, I believe a long-read approach would have a stronger chance of capturing the difference between the male and female haplotypes. While the use of long-read sequencing has been limited by cost, recent price drops for both Oxford Nanopore Technologies (ONT) and PacBio HiFi reads promise to make long-read assemblies a more attainable option for research laboratories.

VITA

Originally from South Carolina, Matthew Huff was born in Myrtle Beach, South Carolina, and grew up in Lexington, South Carolina. After high school, he attended Clemson University and received a Bachelor of Science degree in Genetics. He attended the Medical University of South Carolina a year after graduation and pursued a Master of Science in Biomedical Science with a focus on Bioinformatics. Following graduation, he began a job as a Bioinformatics Research Associate for the University of Tennessee, Knoxville. Two years after being hired, he chose to pursue a Doctor of Philosophy degree in Entomology and Plant Pathology with a concentration in Bioinformatics. His research interests include green ash, with a particular focus on understanding sex determination, and genome assembly and associated analyses. In his position as a Research Associate, he has performed, or assisted in, analyses on a variety of organisms with students and researchers from other departments in the University. After graduation, he will continue his Research Associate duties at the University of Tennessee while searching for new positions. He is grateful for the support from his family as he looks towards new career opportunities.