8-2022

Genome Evolution in the Salicaceae: Genetic Novelty, Horizontal Gene Transfer, and Comparative Genomics

Timothy Yates
tyates6@vols.utk.edu

Recommended Citation

Yates, Timothy, "Genome Evolution in the Salicaceae: Genetic Novelty, Horizontal Gene Transfer, and Comparative Genomics. " PhD diss., University of Tennessee, 2022.
https://trace.tennessee.edu/utk_graddiss/7325

To the Graduate Council:

I am submitting herewith a dissertation written by Timothy Yates entitled "Genome Evolution in the Salicaceae: Genetic Novelty, Horizontal Gene Transfer, and Comparative Genomics." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Energy Science and Engineering.

Wellington Muchero, Major Professor

We have read this dissertation and recommend its acceptance:

Jin-Gui Chen, Margaret Staton, Bode Olukolu

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Genome Evolution in the *Salicaceae*: Genetic Novelty, Horizontal Gene Transfer, and Comparative Genomics

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

Timothy B. Yates
August 2022

DEDICATION

I dedicate this work to my parents, Brad Yates and Susan Lagasse, and my fiancé Sonja Feck, whose support during graduate school made this work possible.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Wellington Muchero for his support and mentorship throughout graduate school. I also would like to thank members of my committee, Dr. Jay Chen, Dr. Margaret Staton, and Dr. Bode Olukulu for their support, guidance, and helpful feedback.

I also want to thank the plant systems biology group at ORNL. To Lee Gunter, Mindy Clark, Sara Jawdy, and Mac McLennan without whom this work would not be possible. To Priya Ranjan, Kai Feng, Jin Zhang, and Cai John for their discussion, friendship, and help which enabled this work.

I also would like to thank the Bredesen Center staff, Dr. Lee Reidinger, Wanda Davis, Dr. Suresh Babu, and Allie Burns for their support.

I also want to thank my friends and family in Tennessee and California who supported me on this journey and made grad school much more enjoyable. Lastly, I want to thank Sonja Feck, my fiancé, for her unwavering support throughout this process.

ABSTRACT

Genome evolution is a powerful force which shapes genomes over time through processes like mutation, horizontal transfer, and sexual reproduction. Although questions which aim to explore genome evolution are broad, they are all understood through the discovery and comparison of genetic variation. For example, genetic diversity may explain differences in phenotypes, etiology of disease, and is essential for phylogenomic analysis. Recently, the democratization of next generation and third generation DNA sequencing technologies have allowed for genomics to produce large amounts of sequence data. This has facilitated the capture of genetic variation at species and population scales.

Populus and *Salix* are members of the *Salicaceae* family and are ecologically and economically important woody plants. Currently, there are multiple high-quality reference genomes available for these two genera. Two important sources of genome evolution that will be explored here are genetic novelty in the form of new genes and horizontal gene transfer from the organelle genomes. In the context of genome evolution, both processes have been shown to contribute to beneficial phenotypes as well as disease. The primary contributions of this dissertation research are to identify and assign putative functions to orphan and *de novo* genes in *P. trichocarpa*, identify and compare horizontal transfer from the organelle genomes to the nuclear genomes of *P. trichocarpa* and *P. deltoides*, and generate new organelle genome resources for 6 different *Salix* species.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ATTACHMENTS

Chapter I
Introduction

INTRODUCTION

Genetic Novelty and *de novo* gene evolution

De novo genes are genes whose origins lie in ancestrally non-coding sequence (J. Schmitz, Ullrich, & Bornberg-bauer, 2017). They may be either protein-coding or RNA genes (J. F. Schmitz & Bornberg-Bauer, 2017). They are part of class of genes known as orphan genes which can be defined as genes which do not have homology to other genes. In most studies, genetic novelty is explored along a lineage or within a species or taxa and are referred to as lineage specific genes and taxonomically restricted genes (TRGs), respectively.

Prior to widespread genome sequencing, gene duplication had long been thought to be the primary source of novel protein coding genes. At the time, there was no empirical evidence for new genes originating from ancestrally non-coding sequence. Consequently, the duplication-divergence model was both supported by Susumu Ohno and François Jacob, who claimed de novo gene origination from non-coding sequence was highly unlikely if not impossible (Jacob, 1977; Ohno, 1970). The completion of the yeast genome in 1996 was the primary driver for the reexamination of the concept of de novo genes when it was determined that nearly 30-35% of the total open reading frames (ORFs) had no known homologs (Dujon, 1996). Although this percentage would diminish with the continued sequencing of genomes closely related to yeast, Bernard Dujon commented on this relatively large percentage and proposed the term “orphan genes” (Dujon, 1996). In the mid 1990s, orphan genes were defined as genes with unknown function and genes that were structurally dissimilar to any other genes in any database, the latter being the basis of the current operating definition (Khalturin, Hemmrich, Fraune, Augustin, & Bosch, 2009). However, it was not until the early-2000s when empirical evidence for de novo evolved genes was available. The first example being five genes that evolved de novo from non-coding sequence in Drosophila (Levine, Jones, Kern, Lindfors, & Begun, 2006). Following this study, there were also several additional examples of genes evolving from non-coding sequence in primates, humans, and plants (Knowles, Mclysaght, Knowles, & Mclysaght, 2009; Toll-Riera et al., 2009; Xiao et al., 2009). For some time, it was thought orphan genes were a product of poor annotation (and to an extent this is true, many orphans are poorly annotated), and lacking genomic data in other species (Casari, De Daruvar, Sander, & Schneider, 1996). The general sentiment was that as sequence data became better the number of orphan genes would decrease

(Casari et al., 1996). However, this was not the case, and because of next generation sequencing, orphan genes have been found in all domains of life. Every taxonomic group studied so far contains 5-15% of genes that are lineage specific (Khalturin et al., 2009).

Many models have put forth possible mechanisms of *de novo* gene birth (Bornberg-bauer et al. 2015; Schmitz & Bornberg-Bauer 2017; Carvunis et al. 2012; Wilson et al. 2017). Here, I will describe the two of the best studied models: 'ORF first vs. transcription first' and the 'Out of Testis' hypothesis. The 'ORF first' vs. 'transcription first' model relies on the fact that as genes, *de novo* genes must acquire an open reading frame (ORF) and be transcribed (Schlotterer, 2015). The order of these events is unimportant, as both need to occur for a functional RNA or protein product to be produced and subjected to selection. Additionally, there is evidence of *de novo* gene origination from both 'ORF-first' and 'transcription first' paths. (Guerzoni & McLysaght, 2016). The 'transcription first' model requires transcription of the locus and assumes no ORF is present. Next, a random DNA mutation or mutations occur which create an ORF. Two additional sources of evidence which support the 'transcription first' model include pervasive transcription of the genome and evidence of expression at the orthologous region in a closely related lineage (Neme & Tautz, 2016; Reinhardt et al., 2013). Alternatively, the 'ORF-first' model proposes that random mutations which occur in cis-regulatory may allow for a previously established ORF to be transcribed. The 'Out of Testis' hypothesis offers another model by which *de novo* genes may arise. Early papers in orphan gene research identified that new genes were preferentially expressed in *Drosophila* male reproductive organs (Begun, Lindfors, Kern, & Jones, 2007; Betrán, Thornton, & Long, 2002). From these observations, more formal hypotheses were formed which stated that young genes might initially be expressed in the testis and later may be gain expression in more tissues (Vinckenbosch, Dupanloup, & Kaessmann, 2006). In mammals, transcription in the testes is complex and thought to be promiscuous due to an open chromatin state (Soumillon et al., 2013). Similarly, in monocot and dicot plants pollen-biased expression of young genes is apparent and pollen shares a similar open chromatin environment (Cui et al., 2015). Therefore, male reproductive tissues may serve as a testing ground for new genes which may become fixed if they offer some advantage.

Identification and curation of *de novo* and orphan genes are impacted by a range of factors. Two of the most salient are the genome annotation pipeline employed and the methodology used for identification. The lack of standardization in gene annotation pipelines is significant problem for the accurate identification of orphan and *de novo* genes and in some cases may be biased against new genes (Klasberg, Bitard-Feildel, & Mallet, 2016). Most common genome annotation methods employ a combination of homology-based and *ab initio* derived prediction to annotate the genome. Because orphan and *de* novo genes lack homology to any other genes these methods are unsuitable for their identification. Furthermore, *ab initio* techniques, which are models trained on known proteins or genes may be biased against young genes which have atypical structures such as short length and are often overlapping another gene on the antisense strand. (Ardern, Neuhaus, & Scherer, 2020; Seetharam et al., 2019). One solution to this is to include additional evidence such as RNA-Seq, Ribo-Seq, or proteomic data to correct missing annotations. However, due to the high tissue specificity and low expression levels of young *de novo* and orphan genes multiple tissues and developmental stages should be sampled at high sequencing depth. Additionally, some *de novo* gene studies do not limit their analyses to annotated genes which adds further to variation in curation methodology (Blevins et al., 2021; Ruiz-Orera et al., 2015). The methodology employed to identify *de novo* genes also varies considerably across studies. Two of the most common methods are phylostratigraphy and synteny-based identification of *de novo* genes. Phylostratigraphy employs BLAST to stratify genes by their age and relies on sequence similarity (Arendsee et al. 2019). Consequently, phylostratigraphy identifies *de novo* genes that originate from non-coding sequence and short rapidly evolving genes which may be too diverged from their closest homolog (Domazet-Lošo et al., 2017). Additionally, phylostratigraphy relies on negative evidence (the lack of an identifiable homolog) to assign an age to a target feature is therefore heavily influenced by the genomes that are available, and the quality of the genome assemblies and gene annotations (Arendsee et al. 2019). To improve phylostratigraphy, more sensitive sequence similarity based methods like PSI-BLAST may be employed to detect ancestrally coding sequence and reduce the identification of false positive *de novo* genes (Van Oss & Carvunis, 2019). Synteny-based detection is the preferred method for the identification of young *de novo* genes. It relies on conserved sequence order between the target species and outgroups and can distinguish true *de novo* genes which have originated from non-coding sequence from those who have rapidly evolved from ancestrally coding sequence (Knowles et al., 2009). Most commonly, genes are used as markers, which can then be assembled in syntenic blocks, which represent genomic segments where gene order is conserved (Soderlund, Bomhoff, & Nelson, 2011; Tang et al., 2008; Y. Wang et al., 2012). Recently, whole genome alignment of the query and target genomes is becoming more popular as a much more dense syntenic map is produced (Grabherr et al. 2010; Marçais et al. 2018; Arendsee et al. 2019). However, regardless of the methodology used to generate syntenic information between a target and query species, inspection of syntenic sequence of the target outgroup species for coding potential is of utmost importance (McLysaght & Hurst, 2016). The target sequence should clearly be non-coding and ideally shared disabling mutations should be conserved in the outgroup sequence (Vakirlis & McLysaght, 2019). Synteny-based methods also have limitations. For synteny to be used there must be a sufficient number of closely related high-quality genome assemblies. Generally, synteny becomes less conserved as species relatedness to the focal genome decreases. Methodology to automate and integrate phylostratigraphy and synteny-based identification of *de novo* genes is now available and represents a reproducible and simpler approach to what is generally a cumbersome and non-trivial task (Arendsee et al. 2019).

Orphan and *de novo* genes have many features which distinguish them from older genes. For example, some of the most common include shorter length, narrow expression breadth across tissues, lower expression levels, and low sequence complexity due to their non-coding origins (Khalturin et al., 2009). Additionally, their codon bias and GC contents differ from older genes (Basile, Sachenkova, Light, & Elofsson, 2016). From a protein structural perspective they tend to have high intrinsic disorder, high β-sheet preference, and overall low complexity (Ahrens, Dos Santos, & Siltberg-Liberles, 2016; Light, Sagit, Sachenkova, Ekman, & Elofsson, 2013). They also appear to be associated with common biological themes. In humans, they are shown to be expressed in the brain and testes and have been linked to cancer (*CLLU1*) and other diseases (Knowles et al., 2009; McLysaght & Hurst, 2016). In plants, they have been linked to stress response and plant pathogen interaction. One of the best studied examples in plants is *QQS,* an *Arabidopsis* orphan gene, and starch biosynthesis regulator, is likely a product of retroposition which was determined by its unique gene and its regulatory control sequences (L. Li et al., 2009). This same gene has also been shown to reduce susceptibility to cyst nematodes when heterologously expressed in soybean (Qi et al., 2019). Another orphan gene which confers resistance to a pathogen is the Bassicaceae-specific *EWR1* gene which when heterologously expressed in *N. benthamiana* results in reduced susceptibility to *V. dahlia*e (Yadeta, Valkenburg, Hanemian, Marco, & Thomma, 2014). Beyond known examples, it has also been shown that functional proteins can be generated from random protein libraries. In *Arabidopsis,* after screening over 2,000 transgenic plants that were transformed with small random peptides a subset showed phenotypes that affected photosynthesis, flowering, and the red light response (Bao, Clancy, Carvalho, Elliott, & Folta, 2017). It is evident that orphan genes largely have roles in response to stress but also are involved in various functions.

A largely unexplored area in *de novo* gene evolution is regarding how their regulation and integration into existing regulatory networks (Z. W. Arendsee, Li, & Wurtele, 2014). In order to be integrated into existing networks *de novo* genes must acquire functional *cis*-elements or enhancers (X. Wu & Sharp, 2013). Additionally, orphan genes have been shown to not be constitutively expressed. One indicator of functionality is change in expression in response to functional needs and several studies have described these patterns (Colbourne et al., 2011; Heinen, Staubach, Häming, & Tautz, 2009). One paper showed differential translation of young genes in stress conditions and an enrichment of young genes adjacent to transcription factor binding sites that are known to regulate response to stress (Carvunis et al., 2012). Furthermore, several studies have explored the *cis*-elements of *de novo* genes but primarily through homology-based detection of known transcription factor binding sites (TFBS) (Li et al. 2016; Carvunis et al. 2012). More recently, orphan gene regulation was explored in *P. pacificus* and it was found orphan genes are located in open chromatin and have different enhancers compared to conserved genes (Werner et al., 2018). These findings were explored further with single-cell RNA-seq and ATAC-seq data and showed that open regions of chromatin facilitated *de novo* evolution and enhancers are important for edge creation and rewiring of regulatory networks (Majic & Payne, 2019). Studies which explore regulatory networks and incorporate functional studies in additional systems will allow further understanding of how orphan genes are integrated in networks and the functional niches they are present in.

Organelle derived transfer to the nuclear genome

DNA transfer from organelle genomes to the nuclear genome is a ubiquitous process and is present in most eukaryotic nuclear genomes (Zhang et al., 2020). Beyond organelle DNA transfer to the nuclear genome, transfers also occur from plastid to mitochondrion, mitochondrion to plastid, and nucleus to mitochondrion (Kleine, Maier, & Leister, 2009).

Organelle derived transfer to the nuclear genome is referred to as nuclear mitochondrial DNAs (NUMTs) and nuclear plastid DNAs (NUPTs).

Mitochondria and chloroplasts were once free-living prokaryotes, their closest relatives being α-proteobacteria and cyanobacteria, respectively. Chloroplast genomes contain between 20 to 200 protein-coding genes and mitogenomes encode a comparatively smaller number ranging from 3 to 67 protein-coding genes (Richly & Leister, 2004a, 2004b). Extant α-proteobacteria and cyanobacteria have comparatively larger genomes and numbers of protein-coding genes, highlighting the extensive loss in size of chloroplast and mitogenomes. For example, the α-proteobacteria *M. loti* has a 7 Mb genome size and 6,700 protein-coding genes (Timmis, Ayliff, Huang, & Martin, 2004). Similarly, the cyanobacteria *Nostoc punctiforme*, has an 8.2 Mb genome size and close to 7,000 genes. This drastic reduction in organelle size can be explained by endosymbiotic theory where over time large numbers of genes were transferred from organelle genomes to the nuclear genome. This process is commonly referred to as endosymbiotic gene transfer (EGT) (Timmis et al., 2004). After integration into the nuclear genome, the nuclear copy has several possible fates. It can retain its original function and be retargeted to the organelle, be retargeted to a different compartment in the cell, or may evolve different or additional functions (Hazkani-Covo & Martin, 2017). For example in *Arabidopsis,* 4500 (18%) nuclear genes are estimated to be derived from cyanobacteria and about 50% of these diverge in function from their original plastid roles (Martin et al., 2002).

This extensive transfer to the nuclear genome also raises questions about why organelle genomes are necessary and are still present in the cell. To address this, two prominent theories have been put forth. First, the ‘hydrophobicity hypothesis’ states that due to the hydrophobic nature of proteins encoded in organelle genomes, their transport from the nucleus back to the organelle through the cytosol would not be possible (Daley & Whelan, 2005). Second, large numbers of proteins in organelles play roles in redox reactions and are responsive to the redox state in organelle. It has been proposed by the "co-location for redox regulation" (CoRR) hypothesis that organelle genes that remain in the organelle were essential for controlling redox state in α-proteobacteria and cyanobacteria and the co-location of these proteins are essential for maintaining that balance (Allen, 2017).

EGT in the form of NUPT and NUMT transfer is an active process for most genomes which have organelles. The rate of transfer to the nucleus is high. The rate of transfer to the nucleus was determined by transplastomic *N. benthamiana* lines which had a nucleus specific neomycin gene integrated into the chrloroplast genome (Huang, Ayliffe, & Timmis, 2003). Seedlings were then screened for kanamycin resistance, and one transposition event to the nucleus was observed per 16,000 pollen grains. Additionally, abiotic stress such as heat stress may increase the frequency of transfer to the nuclear genome (D. Wang, Lloyd, & Timmis, 2012) Mechanisms of insertion into the nuclear genome were initially thought to be RNA-mediated (Nugent & Palmer, 1991). However, as additional genomes became available, this has been shown to be false and any DNA segment of organelle origin can be transferred. Additionally, studies have also shown there is no evidence of post-transcriptional modification like polyadenylation or RNA splicing prior to insertion (Woischnik & Moraes, 2002). The primary mechanism is thought to be double-strand break repair followed by non-homologous end joining (NHEJ) (Hazkani-Covo & Covo, 2008). There are many possible processes by which DNA could escape from the organelle as a result of damage to the organelle membrane, some of which include autophagy, cell division, and cell stress (Kleine et al., 2009). The escaped organelle DNA can then be accessed by nuclear import machinery and be inserted into nuclear DNA (Thorsness & Fox, 1990).

Based on the identification and analysis of NUMTs and NUMTs in 6 plant genomes patterns of integration and variation in size and numbers were apparent (Michalovova, Vyskot, & Kejnovsky, 2013). Additionally, a larger analysis of NUPTs in 199 plant nuclear genomes identified considerable differences in size as the smallest and largest cumulative lengths were 1,038 bp and 9.83 Mb in *P. tricomutum* and *T. urartu*, respectively (Zhang et al., 2020). The largest integrations of NUPT and NUMT content were the longest and had high identity to their organelle sequence from which they were derived. This suggests that longer stretches of NUPTs and NUMTs are younger. These large NUPTs and NUMTs were also preferentially located in centromeres and peri-centromeric regions. NUPTs and NUMTs have also been shown to be organized as clusters (S. F. Li et al., 2019). In general, NUPTs and NUMTs are organized in several different ways in the nuclear genome (Leister, 2005). These arrangements include continuous insertions which are collinear with organelle DNA, rearranged stretches which diverge from the original orientation in the nuclear genome, and concatemers or mosaic fragments consisting of both NUPTs and NUMTs. Number and total length of NUPTs and NUMTs present within the nuclear genome were also shown to be correlated with nuclear genome size and the number of organelles in the cell (Hazkani-Covo, Zeller, & Martin, 2010; Smith, Crosby, & Lee, 2011).

Besides differences at taxonomic levels, genome assembly and methodology used for curation can have considerable impacts on the numbers and total size of NUPTs and NUMTs are identified (Hazkani-Covo & Martin, 2017). Some of these features are also edited out incorrectly during the genome assembly process (Dayama, Emery, Kidd, & Mills, 2014). The most common alignment tool to identify NUPTs and NUMTs in nuclear genomes is BLAST (Altschul, 1977). There is a large amount of heterogeneity in how BLAST output is filtered between studies. The most significant variations are present in differences in alignment length thresholds, e-value thresholds, identity thresholds, merging thresholds, and how organelle repeats are handled (Ma et al., 2020; Michalovova et al., 2013; Pinard, Myburg, & Mizrachi, 2019). A more unified framework on the best practices for NUPT and NUMT identification and downstream filtering would be beneficial not only for those studying these features but also for studies which assemble genomes and would like to accurately assess NUPT and NUMT content.

Once NUPT and NUMT features are integrated into the nuclear genome they undergo various outcomes as evolutionary time progresses. As described by Zhang et al. 2020, there are five possible fates for organelle DNA present in the nuclear genome: elimination, mutation, fragmentation, rearrangement, and proliferation. For elimination, it was determined in rice that 80% percent of NUPTs were eliminated within 1 million years after they were inserted (Matsuo, Ito, Yamauchi, & Obokata, 2005). This study suggested that the integration and elimination of organelle DNA are in equilibrium, and this is essential for the maintenance of genome size. For mutation, the majority of NUPTs and NUMTs will not be under selective pressure and will therefore decay (Huang, Grünheit, Ahmadinejad, Timmis, & Martin, 2005). Specifically, new NUPTs and NUMTs are thought to be neutral or deleterious when integrated as they may cause genome instability (Yoshida, Furihata, To, Kakutani, & Kawabe, 2019). Similarly, transposable elements are also known to cause instability in the genome and are in part regulated by epigenetic modification in the forms of DNA and histone methylation (Du, Johnson, Jacobsen, & Patel, 2015). A corollary of NUPT and NUMT methylation as a source of repression is mutation as methylation can cause the deamination of cytosine. This results in biased cytosine to thymine (C to T) and guanine to adenine (G to A) transition mutations. This is also a useful characteristic for assessing the relative age of NUPTs and NUMTs as new integration events should be highly methylated and over time become less methylated. This epigenetic modification is likely essential for the suppression of spurious transcription of NUPTs and NUMTs and therefore maintenance of regulatory networks that exist between nuclear and organelle genomes (Yoshida et al., 2019). In terms of fragmentation and rearrangement, new organelle derived features tend to be longer, have higher identity, and are in general collinear to their sister organelle sequences. The insertion of TEs or other NUMTs/NUMTs into existing NUPT features can have an important impact on the arrangement of these features in the nuclear genome as was observed in rice and higher rates of rearrangement should be present in older features (Matsuo et al., 2005). Lastly, proliferation of NUPTs and NUMTs can occur and is likely mediated through retrotransposon activity (Vanburen & Ming, 2013).

Most NUMT and NUPT transfer is non-coding in nature. In order to be functional, the putative NUPT or NUMT must have accompanying regulatory sequences such as promoter, terminator, and transit peptide sequences. Some examples in the context of mitochondrial gene transfer have shown that most events insert into pre-existing nuclear promoter sequences and transit-peptide sequences (Timmis et al., 2004). Furthermore, it has been shown that organelle DNA preferentially inserts into open chromatin (D. Wang & Timmis, 2013). Additionally, it is also now evident that organelle derived regulatory sequences can drive nuclear transcription, albeit at lower levels than nuclear promoters (D. Wang, Qu, Adelson, Zhu, & Timmis, 2014). The transfer of whole functional genes may be unlikely but, NUPT or NUMT transfer has been shown to be implicated in remodeling existing genes. For example one study identified novel exons contributed by NUMTS and NUPTs in *Arabidopsis, S. cerevisiae, humans,* and rice (Noutsos, Kleine, Armbruster, DalCorso, & Leister, 2007). Based on an analysis of these four species, 45 integration events contributed sequences to 49 protein-coding exons. Although NUPT and NUMT insertion may provide a source of evolutionary innovation, they also have been linked to disease in humans (Hazkani-Covo et al., 2010). One study described five different NUMT insertions which cause disease in humans (Chen, Chuzhanova, Stenson, Férec, & Cooper, 2005). Overall, NUPTs and NUMTs are important forces which shape genomes and adaptive evolution, are sources of genetic innovation, are causes of disease in humans.

Organelle genomics and phylogenomics in plants

In the cell, chloroplasts and mitochondria are membrane bound organelles that perform photosynthesis and generate energy in the form of ATP, respectively (Mustárdy, Buttle, Steinbach, & Garab, 2008; Siekevitz, 1957). They both have haploid genomes and are of prokaryotic origin as proposed by endosymbiotic theory. Their utility in phylogenomics and comparative genomics will be explored here.

N. benthamiana (tobacco) and *M. polymorpha* (liverwort) were the first chloroplast genomes to be sequenced and were released in 1986 (Ohyama et al., 1986; Shinozaki et al., 1986). At present, land plant chloroplast genomes are available for more than 2,500 genera NCBI and are far more numerous compared to mitogenome and nuclear genome assemblies (accessed 01/2022). Land plant chloroplast genomes have a highly conserved quadripartite structure which consists of a large single-copy (LSC), small single-copy (SSC), and two inverted repeat (IR) regions. Additionally, almost all land plant chloroplast genomes range in size from 120 kb to 160 kb (Fan, Wu, Yang, Shahzad, & Li, 2018). Although most chloroplast genomes map as circular molecules, there is also evidence for branched-linear forms. For example, the *A. thaliana* chloroplast genome has been observed in both circular and linear conformations (Oldenburg & Bendich, 2015). Chloroplast gene arrangement reflects a bacterial origin, as most genes are organized into structures similar to operons (Zoschke & Bock, 2018). However, unlike bacteria, chloroplasts have highly complex expression regulation and posttranscriptional RNA modification mechanisms. Some of these include RNA splicing, RNA editing, and processing of RNA transcripts (Börner, Aleynikova, Zubo, & Kusnetsov, 2015).

The highly conserved structure of chloroplasts among land plants has allowed for the widespread use of reference-based assembly (J. Wu et al., 2012). Furthermore, high copy number of organelle DNA in the cell has resulted in high-quality DNA extraction and allowed for genome-skimming approaches (Golczyk et al., 2014). Before next-generation sequencing technologies were widely available much more intensive procedures were required to prepare organelle DNA for sequencing (Saski et al., 2005). These procedures often included the creation of bacterial artificial chromosomes (BAC) libraries, sugar gradients or ultra-centrifugation, and long-range PCR to enrich for organelle genomes (Jansen et al., 2005). Recently, dedicated genome assembly software for plastid genomes has also been developed which further simplifies the chloroplast genome assembly process (Dierckxsens, Mardulyn, & Smits, 2017; Jin et al., 2020). These assembly methods require a partial assembly or starting sequence such as a conserved chloroplast gene and therefore don't rely on a closely related reference genomes. Both of these assembly software use baiting and iterative mapping approach for assembly. Additionally, a large number of automated gene annotation methods are available for chloroplast genomes, most of which rely on the transfer of existing annotations to the new assembly (Cheng, Zeng, Ren, & Liu, 2013; Shi et al., 2019; Tillich et al., 2017). These tools greatly simplify the annotation and assembly process of chloroplast genomes.

The first two sequenced plant mitogenomes were *M. polymorpha* and *A. thaliana* (Oda et al., 1992; Unseld, Marienfeld, Brandt, & Brennicke, 1997). The total numbers of land plant mitochondrial genomes which are publicly available at NCBI is quite low at 290, which is representative of 204 genera (accessed 01/2022). However, these mitogenomes do represent all seven major lineages of land plants which represents a promising opportunity for comparative analyses (Mower, 2020). One of the main reasons for low numbers are primarily due to the challenges of mitogenome assembly with next-generation sequencing (NGS) data as mitogenomes generally have large repeats which may exceed the short read length and the heterogenous nature of mitogenomes even among closely related species. For example, in the genus *Silene, S. latifola* and *S. conica* have 250 kb and 11.3 Mb mitogenomes, respectively (Sloan et al., 2012; Sloan, Alverson, Štorchová, Palmer, & Taylor, 2010). Beyond differences in size, variation in structure is also apparent (Sloan, 2013). Plant mitogenomes are typically thought to be circular and that the 'master circle' which is the entire sequence content derived from the recombination of subgenomic circles is the primary conformation. However, this is not always the case as multipartite genomes, subgenomic circles, and linear molecules have been observed (Mower, 2020). Variation in land plant protein-coding gene content is also apparent with numbers as low as 19 in the genus *Viscum* and more than 50 in *M. polymorpha* (Oda et al., 1992; Skippingtona, Barkmanb, Ricea, & Palmera, 2015). Mitogenomes also have two types of introns (group I and group II) which have different secondary structures and splicing mechanisms and must be removed from precursor transcripts (Bonen, 2012). Intron content and numbers are highly variable across plants with group II introns being more common in land plants compared to group I introns (Mukhopadhyay & Hausner, 2021). Splicing mechanisms in plant mitochondria can occur in both *cis* and *trans* and vary between land plants. The most common mechanism is *cis*, where introns are removed from same primary transcript.

Alternatively, as a result of fragmentation during evolution, some genes may be split apart and *trans*-splicing of introns occurs over long distances (Guo, Zhu, Fan, Adams, & Mower, 2020).

Because mitogenomes are highly variable, reference-based assembly is a challenge. This results in a variety of non-standard methods used for assembly most of which rely on iterative baiting and assembly or *de novo* assembly followed by gap filling (Choi et al., 2017; Jackman et al., 2020; Kersten et al., 2016). Dedicated plant mitogenome assembly software such as GetOrganelle and NOVOPlasty are developed for the use of next-generation sequencing data and therefore may not be as useful for mitogenomes which have large repeats (Dierckxsens et al., 2017; Jin et al., 2020). The increased adoption of third-generation sequencing data such as PacBio or Oxford Nanopore should simplify plant mitogenome assembly considerably. Studies which use third-generation sequencing data often use hybrid assembly methods with like those employed in SPAdes or Unicycler or a dedicated long-read assembler like Canu followed by polishing with next-generation sequencing data (Koren et al., 2017; Prjibelski, Antipov, Meleshko, Lapidus, & Korobeynikov, 2020; Wick, Judd, Gorrie, & Holt, 2017). Plant mitogenome annotation software is less mature compared to chloroplast genome annotation software and is limited to mitofy and MFannot (https://github.com/BFL-lab/Mfannot) (Alverson et al., 2010). In summary, as the numbers of assembled plant mitogenomes continues to increase, additional plant mitogenome specific assembly and annotation methods would be beneficial.

Plastid sequences are the most common choice for the reconstruction of land plant phylogenies (Gitzendanner et al. 2018). Due to their small size, homogeneity in structure, and uniparental inheritance the chloroplast genome is an optimal choice for phylogenomic studies (Gao, Su, & Wang, 2010). Additionally, an early paper identified that plastid genes had appropriate rates of evolution for phylogenetic studies (Palmer, Jansen, Michaels, Chase, & Manhart, 1988). Prior to widespread whole chloroplast genome sequencing, single genes (*rbcL*) were common targets for the reconstruction of phylogenies in land plants (Chase et al., 1993; Davis, Xi, & Mathews, 2014). More recently, additional single copy regions which have higher rates of variation have been proposed for barcoding all land plants (Chase et al., 2007). These universal barcodes have also allowed for phylogenetic inference at more granular levels. Currently, phylogenomics leverages widely available whole chloroplast genomes to infer phylogenies (Ruhfel, Gitzendanner, Soltis, Soltis, & Burleigh, 2014). Phylogenies inferred from whole chloroplast genome information are commonly derived from all single copy orthologs or whole genome alignments (Gitzendanner et al. 2018).

Due to its complex structure, poor conservation across land plants, and a slower evolutionary rate compared to the nuclear (12 times slower) and chloroplast genomes (3 times slower) the plant mitogenome is much less suited for phylogenetic analyses (Drouin, Daoud, & Xia, 2008; Duminil & Besnard, 2021; Van de Paer, Bouchez, & Besnard, 2018). This contrasts animal and fungal mitogenomes which are much smaller and more conserved in structure and are commonly used in phylogenetic studies (Boore, 1999; Salavirta et al., 2014). Because of structural divergence even among closely related species plant mitogenome derived phylogenies should be constructed from genes (X. Wang et al., 2018). However, some studies have inferred phylogenies from mitogenome data which agree with plastome data (Van de Paer et al., 2018; S. Wang et al., 2018). Additional land plant mitogenomes should become available as third generation sequencing becomes more widespread. This should allow for a more comprehensive understanding of their use in phylogenetic inference and how they compare to plastome based phylogenies.

Outline and scientific contribution of dissertation

The genera *Populus* and *Salix* are ecologically and economically important trees and shrubs. Their natural range is expansive and extends across North America, Europe, and Asia. Their uses are broad, some of which include; fiber, fuel, and medicine (Erichsen-Brown, 1979). The development of genetic resources for these important species has and will benefit the domestication and breeding process (Stanton et al., 2014; Zhou et al., 2020). This is particularly important as they are excellent feedstocks for the creation of lignocellulosic biofuels which aim to replace fossil fuels primarily in ocean shipping and aviation industries (Fulton et. al., 2015). As demand for liquid fuels remains likely until 2075, development of cleaner alternatives to fossil fuels are of paramount importance (Fulton et al., 2015; Hannon et al., 2020).

The primary contribution of this thesis will be the creation of new genomic resources and the exploration of genome evolution from various facets in the *Salicaceae* family. The first chapter provides a comprehensive introduction to chapters two through five. The second chapter explores orphan and *de novo* gene evolution in *P. trichocarpa* through the lens of a whole genome duplication. The third chapter identifies and compares organelle derived nuclear DNA in the nuclear genomes of *P. trichocarpa* and *P. deltoides*. The fourth chapter generates new organelle genome resources from a pedigree of 11 F1 *Salix* genomes and further characterizes them through comparative genomics. The following paragraphs will expand further on the specific contributions of each chapter.

Chapter two is focused on the identification, curation, and regulation of *de novo* genes in *P. trichocarpa*. All members of the *Salicaceae* share a common whole genome duplication event called the Salicoid whole genome duplication which resulted in highly syntenic sister chromosomes (Dai et al., 2014; Tuskan et al., 2007). This genomic feature allows for inter-genera, inter-species, and intra-genome comparisons of the steps of *de novo* gene evolution. Additionally, *P. trichocarpa* is a good platform for characterizing orphan and *de novo* gene evolution because of the genome wide association study (GWAS) population which includes genomic and transcriptomic resources for thousands of *P. trichocarpa* genotypes. This population allowed for determination of variation in *de novo* genes along the natural range of *P. trichocarpa.* Furthermore, it also allowed for the assignment of *de novo* and orphan genes to regulatory networks. Through functional enrichment of genes in these networks, *de novo* and orphan genes could be assigned a putative function by guilt by association.

Chapter three identifies and compares putative coding and non-coding organelle derived integration events in the nuclear genomes of *P. trichocarpa* and *P. deltoides*. To assess variation in these features between *P. trichcoarpa* and *P. deltoides* developed we developed a presence and absence variation (PAV) curation process. To validate PAV and further characterize these organelle derived features we determined their methylation levels and their genetic distances to their original organelle feature. To determine expression evidence of putative protein-coding organelle derived features we used the JGI Gene Atlas expression datasets in both *P. trichocarpa* and *P. deltoides*. Additionally, we further characterized putative protein-coding features with expression datasets from the *P. trichocarpa* GWAS population. Lastly, we examined the whole chloroplast integration event reported in Huang et al., 2017 and identified structural differences at this locus in newer *P. trichocarpa* assemblies.

Chapter four generates new sequence data and organelle assemblies for 11 F1 *Salix* genomes which consisted of six different species. These assemblies were then annotated, and gene presence absence was assessed. Furthermore, mitogenome structural variation was identified and large amounts of heterogeneity even at the species level was identified. We also assessed structural variation at inverted repeat (IR) regions in chloroplast genomes. To determine if there was evidence of adaptive evolution, we calculated pairwise Ka/Ks between species and identified both purifying and positive selection. Finally, we constructed a species phylogeny from publicly available *Salix* organelle genomes and the 11 *Salix* organelle genomes assembled here.

References

Ahrens, J., Dos Santos, H. G., & Siltberg-Liberles, J. (2016). The Nuanced Interplay of Intrinsic Disorder and Other Structural Properties Driving Protein Evolution. *Molecular Biology and Evolution*, *33*(9), 2248–2256. https://doi.org/10.1093/molbev/msw092

Allen, J. F. (2017). The CoRR hypothesis for genes in organelles. *Journal of Theoretical Biology*, *434*, 50–57. https://doi.org/10.1016/j.jtbi.2017.04.008

Altschul, S. (1977). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.*, *25*(17), 3389–3402.

Alverson, A. J., Wei, X., Rice, D. W., Stern, D. B., Barry, K., & Palmer, J. D. (2010). Insights into the evolution of mitochondrial genome size from complete sequences of citrullus lanatus and cucurbita pepo (Cucurbitaceae). *Molecular Biology and Evolution*, *27*(6), 1436–1448. https://doi.org/10.1093/molbev/msq029

Ardern, Z., Neuhaus, K., & Scherer, S. (2020). Are Antisense Proteins in Prokaryotes Functional? *Frontiers in Molecular Biosciences*, *7*(August), 1–12. https://doi.org/10.3389/fmolb.2020.00187

Arendsee, Z., Li, J., Singh, U., Bhandary, P., Seetharam, A., & Wurtele, E. S. (2019). Fagin: Synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics*, *20*(1), 1–14. https://doi.org/10.1186/s12859-019-3023-y

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K., & Wurtele, E. S. (2019). Phylostratr : a Framework for Phylostratigraphy . *Bioinformatics*, (March), 1–11. https://doi.org/10.1093/bioinformatics/btz171

Arendsee, Z. W., Li, L., & Wurtele, E. S. (2014). Coming of age: Orphan genes in plants. *Trends in Plant Science*, *19*(11), 698–708. https://doi.org/10.1016/j.tplants.2014.07.003

Arendsee, Z., Wilkey, A., Singh, U., Li, J., Hur, M., & Wurtele, E. (2019). Synder: Inferring Genomic Orthologs From Synteny Maps. *BioRxiv*, 554501. https://doi.org/10.1101/554501

Bao, Z., Clancy, M. A., Carvalho, R. F., Elliott, K., & Folta, K. M. (2017). Identification of novel growth regulators in plant populations expressing random peptides. *Plant Physiology*, *175*(2), 619–627. https://doi.org/10.1104/pp.17.00577

Basile, W., Sachenkova, O., Light, S., & Elofsson, A. (2016). High GC Content Causes De Novo Created Proteins to be Intrinsically Disordered. *BioRxiv*, 070003. https://doi.org/10.1101/070003

Begun, D. J., Lindfors, H. A., Kern, A. D., & Jones, C. D. (2007). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics*, *176*(2), 1131–1137. https://doi.org/10.1534/genetics.106.069245

Betrán, E., Thornton, K., & Long, M. (2002). Retroposed new genes out of the X in Drosophila. *Genome Research*, *12*(12), 1854–1859. https://doi.org/10.1101/gr.6049

Blevins, W. R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J. L., Espinar, L., … Albà, M. M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature Communications*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-20911-3

Bonen, L. (2012). *Evolution of Mitochondrial Introns in Plants and Photosynthetic Microbes*. *Advances in Botanical Research* (Vol. 63). Elsevier. https://doi.org/10.1016/B978-0-12-394279-1.00007-7

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, *27*(8), 1767–1780. https://doi.org/10.1093/nar/27.8.1767

Bornberg-bauer, E., Schmitz, J., & Heberlein, M. (2015). Emergence of de novo proteins from ‘

dark genomic matter ' by ' grow slow and moult ,' 867–873. https://doi.org/10.1042/BST20150089

Börner, T., Aleynikova, A. Y., Zubo, Y. O., & Kusnetsov, V. V. (2015). Chloroplast RNA polymerases: Role in chloroplast biogenesis. *Biochimica et Biophysica Acta - Bioenergetics*, *1847*(9), 761–769. https://doi.org/10.1016/j.bbabio.2015.02.004

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–374. https://doi.org/10.1038/nature11184

Casari, G., De Daruvar, A., Sander, C., & Schneider, R. (1996). Bioinformatics and the discovery of gene function. *Trends in Genetics*, *12*(7), 244–245. https://doi.org/10.1016/0168-9525(96)30057-7

Chase, M. W., Cowan, R. S., Hollingsworth, P. M., Van Den Berg, C., Madriñán, S., Petersen, G., … Wilkinson, M. (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon*, *56*(2), 295–299. https://doi.org/10.1002/tax.562004

Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H., Mishler, B. D., … Albert, V. A. (1993). Phylogenetics of Seed Plants: An Analysis of Nucleotide Sequences from the Plastid Gene rbcL. *Annals of the Missouri Botanical Garden*, *80*(3), 528. https://doi.org/10.2307/2399846

Chen, J. M., Chuzhanova, N., Stenson, P. D., Férec, C., & Cooper, D. N. (2005). Meta-analysis of gross insertions causing human genetic disease: Novel mutational mechanisms and the role of replication slippage. *Human Mutation*, *25*(2), 207–221. https://doi.org/10.1002/humu.20133

Cheng, J., Zeng, X., Ren, G., & Liu, Z. (2013). CGAP: A new comprehensive platform for the comparative analysis of chloroplast genomes. *BMC Bioinformatics*, *14*. https://doi.org/10.1186/1471-2105-14-95

Choi, M. N., Han, M., Lee, H., Park, H. S., Kim, M. Y., Kim, J. S., … Park, E. J. (2017). The complete mitochondrial genome sequence of Populus davidiana Dode. *Mitochondrial DNA Part B: Resources*, *2*(1), 113–114. https://doi.org/10.1080/23802359.2017.1289346

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., … Boore, J. L. (2011). The ecoresponsive genome of Daphnia pulex. *Science*, *331*(6017), 555–561. https://doi.org/10.1126/science.1197761

Cui, X., Lv, Y., Chen, M., Nikoloski, Z., Twell, D., & Zhang, D. (2015). Young genes out of the male: An insight from evolutionary age analysis of the pollen transcriptome. *Molecular Plant*, *8*(6), 935–945. https://doi.org/10.1016/j.molp.2014.12.008

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G. A., … Yin, T. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research*, *24*(10), 1274–1277. https://doi.org/10.1038/cr.2014.83

Daley, D. O., & Whelan, J. (2005). Why genes persist in organelle genomes. *Genome Biology*, *6*(5). https://doi.org/10.1186/gb-2005-6-5-110

Davis, C. C., Xi, Z., & Mathews, S. (2014). Plastid phylogenomics and green plant phylogeny: Almost full circle but not quite there. *BMC Biology*, *12*, 2–5. https://doi.org/10.1186/1741-7007-12-11

Dayama, G., Emery, S. B., Kidd, J. M., & Mills, R. E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research*, *42*(20), 12640–12649. https://doi.org/10.1093/nar/gku1038

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4). https://doi.org/10.1093/nar/gkw955

Domazet-Lošo, T., Carvunis, A.-R., Mar Albà, M., Sebastijan Šestak, M., Bakarić, R., Neme, R., & Tautz, D. (2017). No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Molecular Biology and Evolution, 34*(4), msw284. https://doi.org/10.1093/molbev/msw284

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., & Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evolutionary Biology*, *11*(1), 47. https://doi.org/10.1186/1471-2148-11-47

Drouin, G., Daoud, H., & Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular Phylogenetics and Evolution*, *49*(3), 827–831. https://doi.org/10.1016/j.ympev.2008.09.009

Du, J., Johnson, L. M., Jacobsen, S. E., & Patel, D. J. (2015). DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews Molecular Cell Biology*, *16*(9), 519–532. https://doi.org/10.1038/nrm4043

Dujon, B. (1996). The yeast genome project: what did we learn? *Trends in Genetics : TIG*, *12*(7), 263–270. https://doi.org/10.1016/0168-9525(96)10027-5

Duminil, J., & Besnard, G. (2021). Utility of the Mitochondrial Genome in Plant Taxonomic Studies. *Methods in Molecular Biology*, *2222*, 107–118. https://doi.org/10.1007/978-1-0716-0997-2_6

Erichsen-Brown, C. (1979). Use of plants for the past 500 years.

Fan, W. B., Wu, Y., Yang, J., Shahzad, K., & Li, Z. H. (2018). Comparative chloroplast genomics of dipsacales species: Insights into sequence variation, adaptive evolution, and phylogenetic relationships. *Frontiers in Plant Science*, *9*(May), 1–13. https://doi.org/10.3389/fpls.2018.00689

Fulton, L., Lynd, L. R., Körner, A., Greene, N., & Tonachel, L. R. (2015). The need for biofuels as part of a low carbon energy future Lewis. *Biofuels, Bioproducts and Biorefining*, *9*, 476–483. https://doi.org/10.1002/bbb

Gao, L., Su, Y. J., & Wang, T. (2010). Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. *Journal of Systematics and Evolution*, *48*(2), 77–93. https://doi.org/10.1111/j.1759-6831.2010.00071.x

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., & Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany*, *105*(3), 291–301. https://doi.org/10.1002/ajb2.1048

Gitzendanner, M. A., Soltis, P. S., Yi, T. S., Li, D. Z., & Soltis, D. E. (2018). *Plastome Phylogenetics: 30 Years of Inferences Into Plant Evolution*. *Advances in Botanical Research* (1st ed., Vol. 85). Elsevier Ltd. https://doi.org/10.1016/bs.abr.2017.11.016

Golczyk, H., Greiner, S., Wanner, G., Weihe, A., Bock, R., Börner, T., & Herrmann, R. G. (2014). Chloroplast DNA in mature and senescing leaves: A reappraisal. *Plant Cell*, *26*(3), 847–854. https://doi.org/10.1105/tpc.113.117465

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., di Palma, F., & Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, *26*(9), 1145–1151. https://doi.org/10.1093/bioinformatics/btq102

Guerzoni, D., & McLysaght, A. (2016). De Novo Genes Arise at a Slow but Steady Rate along the Primate Lineage and Have Been Subject to Incomplete Lineage Sorting. *Genome Biology and Evolution*, *8*(4), 1222–1232. https://doi.org/10.1093/gbe/evw074

Guo, W., Zhu, A., Fan, W., Adams, R. P., & Mower, J. P. (2020). Extensive shifts from cis-to trans-splicing of gymnosperm mitochondrial introns. *Molecular Biology and Evolution*, *37*(6), 1615–1620. https://doi.org/10.1093/molbev/msaa029

Hannon, J. R., Lynd, L. R., Andrade, O., Benavides, P. T., Beckham, G. T., Biddy, M. J., … Wyman, C. E. (2020). Technoeconomic and life-cycle analysis of single-step catalytic conversion of wet ethanol into fungible fuel blendstocks. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(23), 12576–12583. https://doi.org/10.1073/pnas.1821684116

Hazkani-Covo, E., & Covo, S. (2008). Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genetics*, *4*(10). https://doi.org/10.1371/journal.pgen.1000237

Hazkani-Covo, E., & Martin, W. F. (2017). Quantifying the number of independent organelle DNA insertions in genome evolution and human health. *Genome Biology and Evolution*, *9*(5), 1190–1203. https://doi.org/10.1093/gbe/evx078

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics*, *6*(2). https://doi.org/10.1371/journal.pgen.1000834

Heinen, T. J. A. J., Staubach, F., Häming, D., & Tautz, D. (2009). Emergence of a New Gene from an Intergenic Region. *Current Biology*, *19*(18), 1527–1531. https://doi.org/10.1016/j.cub.2009.07.049

Huang, C. Y., Ayliffe, M. A., & Timmis, J. N. (2003). Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature*, *422*(6927), 72–76. https://doi.org/10.1038/nature01435

Huang, C. Y., Grünheit, N., Ahmadinejad, N., Timmis, J. N., & Martin, W. (2005). Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology*, *138*(3), 1723–1733. https://doi.org/10.1104/pp.105.060327

Jackman, S. D., Coombe, L., Warren, R. L., Kirk, H., Trinh, E., MacLeod, T., … Birol, I. (2020). Complete Mitochondrial Genome of a Gymnosperm, Sitka Spruce (Picea sitchensis), Indicates a Complex Physical Structure. *Genome Biology and Evolution*, *12*(7), 1174–1179. https://doi.org/10.1093/gbe/evaa108

Jacob, F. (1977). Evolution and Tinkering. *Science*, *196*(4295), 1161–1166. https://doi.org/10.1210/jcem-10-10-1361

Jansen, R. K., Raubeson, L. A., Boore, J. L., DePamphilis, C. W., Chumley, T. W., Haberle, R. C., … Cui, L. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology*, *395*, 348–384. https://doi.org/10.1016/S0076-6879(05)95020-9

Jin, J. J., Yu, W. Bin, Yang, J. B., Song, Y., Depamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 1–31. https://doi.org/10.1186/s13059-020-02154-5

Kersten, B., Rampant, P. F., Mader, M., Le Paslier, M. C., Bounon, R., Berard, A., … Fladung, M. (2016). Genome sequences of Populus tremula chloroplast and mitochondrion: Implications for holistic poplar breeding. *PLoS ONE*, *11*(1), 1–21. https://doi.org/10.1371/journal.pone.0147209

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., & Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*, *25*(9), 404–413. https://doi.org/10.1016/j.tig.2009.07.006

Klasberg, S., Bitard-Feildel, T., & Mallet, L. (2016). Computational identification of novel genes: Current and future perspectives. *Bioinformatics and Biology Insights*, *10*, 121–131. https://doi.org/10.4137/BBI.S39950

Kleine, T., Maier, U. G., & Leister, D. (2009). DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. *Annual Review of Plant Biology*, *60*(1), 115–138. https://doi.org/10.1146/annurev.arplant.043008.092119

Knowles, D. G., Mclysaght, A., Knowles, D. G., & Mclysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, *19*, 1–9. https://doi.org/10.1101/gr.095026.109

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 κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Leister, D. (2005). Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends in Genetics*, *21*(12), 655–663. https://doi.org/10.1016/j.tig.2005.09.004

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., & Begun, D. J. (2006). Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9935–9939. https://doi.org/10.1073/pnas.0509809103

Li, L., Foster, C. M., Gan, Q., Nettleton, D., James, M. G., Myers, A. M., & Wurtele, E. S. (2009). Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant Journal*, *58*(3), 485–498. https://doi.org/10.1111/j.1365-313X.2009.03793.x

Li, S. F., Li, J. R., Wang, J., Dong, R., Jia, K. L., Zhu, H. W., … Gao, W. J. (2019). Cytogenetic and genomic organization analyses of chloroplast DNA invasions in the nuclear genome of Asparagus officinalis L. provides signatures of evolutionary complexity and informativity in sex chromosome evolution. *BMC Plant Biology*, *19*(1), 1–13. https://doi.org/10.1186/s12870-019-1975-8

Li, Z.-W., Chen, X., Wu, Q., Hagmann, J., Han, T.-S., Zou, Y.-P., … Guo, Y.-L. (2016). On the Origin of De Novo Genes in Arabidopsis thaliana Populations. *Genome Biology and Evolution*, *8*(7), 2190–2202. https://doi.org/10.1093/gbe/evw164

Light, S., Sagit, R., Sachenkova, O., Ekman, D., & Elofsson, A. (2013). Protein expansion is primarily due to indels in intrinsically disordered regions. *Molecular Biology and Evolution*, *30*(12), 2645–2653. https://doi.org/10.1093/molbev/mst157

Ma, X., Fan, J., Wu, Y., Zhao, S., Zheng, X., Sun, C., & Tan, L. (2020). Whole-genome de novo assemblies reveal extensive structural variations and dynamic organelle-to-nucleus DNA transfers in African and Asian rice. *Plant Journal*, *104*(3), 596–612. https://doi.org/10.1111/tpj.14946

Majic, P., & Payne, J. L. (2019). Enhancers facilitate the birth of de novo genes and their integration into regulatory networks. *BioRxiv*, 616581. https://doi.org/10.1101/616581

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), 1–14. https://doi.org/10.1371/journal.pcbi.1005944

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., … Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the*

National Academy of Sciences of the United States of America, *99*(19), 12246–12251. https://doi.org/10.1073/pnas.182432999

Matsuo, M., Ito, Y., Yamauchi, R., & Obokata, J. (2005). The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell*, *17*(3), 665–675. https://doi.org/10.1105/tpc.104.027706

McLysaght, A., & Hurst, L. D. (2016). Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, *17*(9), 567–578. https://doi.org/10.1038/nrg.2016.78

Michalovova, M., Vyskot, B., & Kejnovsky, E. (2013). Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: Size, relative age and chromosomal localization. *Heredity*, *111*(4), 314–320. https://doi.org/10.1038/hdy.2013.51

Mower, J. P. (2020). Variation in protein gene and intron content among land plant mitogenomes. *Mitochondrion*, *53*(May), 203–213. https://doi.org/10.1016/j.mito.2020.06.002

Mukhopadhyay, J., & Hausner, G. (2021). Organellar introns in fungi, algae, and plants. *Cells*, *10*(8). https://doi.org/10.3390/cells10082001

Mustárdy, L., Buttle, K., Steinbach, G., & Garab, G. (2008). The three-dimensional network of the thylakoid membranes in plants: Quasihelical model of the granum-stroma assembly. *Plant Cell*, *20*(10), 2552–2557. https://doi.org/10.1105/tpc.108.059147

Neme, R., & Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *ELife*, *5*(FEBRUARY2016), 1–20. https://doi.org/10.7554/eLife.09977

Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., & Leister, D. (2007). Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. *Trends in Genetics*, *23*(12), 597–601. https://doi.org/10.1016/j.tig.2007.08.016

Nugent, J. M., & Palmer, J. D. (1991). RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. *Cell*, *66*(3), 473–481. https://doi.org/10.1016/0092-8674(81)90011-8

Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Takemura, M., Nozato, N., … Ohyama, K. (1992). Gene organization deduced from the complete sequence of liverwort Marchantia polymorpha mitochondrial DNA. A primitive form of plant mitochondrial genome. *Journal of Molecular Biology*, *223*(1), 1–7. https://doi.org/10.1016/0022-2836(92)90708-R

Ohno, S. (1970). *Evolution by gene duplication*. Springer Science \& Business Media.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., … Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort marchantia polymorpha chloroplast DNA. *Nature*, *322*(6079), 572–574. https://doi.org/10.1038/322572a0

Oldenburg, D. J., & Bendich, A. J. (2015). DNA maintenance in plastids and mitochondria of plants. *Frontiers in Plant Science*, *6*(OCTOBER), 1–15. https://doi.org/10.3389/fpls.2015.00883

Palmer, J. D., Jansen, R. K., Michaels, H. J., Chase, M. W., & Manhart, J. R. (1988). Chloroplast DNA Variation and Plant Phylogeny. *Annals of the Missouri Botanical Garden*, *75*(4), 1180–1206.

Pinard, D., Myburg, A. A., & Mizrachi, E. (2019). The plastid and mitochondrial genomes of Eucalyptus grandis. *BMC Genomics*, *20*(1), 1–14. https://doi.org/10.1186/s12864-019-5444-4

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1), 1–29. https://doi.org/10.1002/cpbi.102

Qi, M., Zheng, W., Zhao, X., Hohenstein, J. D., Kandel, Y., O'Conner, S., … Li, L. (2019). QQS orphan gene and its interactor NF-YC4 reduce susceptibility to pathogens and pests. *Plant Biotechnology Journal*, *17*(1), 252–263. https://doi.org/10.1111/pbi.12961

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., & Jones, C. D. (2013). De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. *PLoS Genetics*, *9*(10). https://doi.org/10.1371/journal.pgen.1003860

Richly, E., & Leister, D. (2004a). NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution*, *21*(6), 1081–1084. https://doi.org/10.1093/molbev/msh110

Richly, E., & Leister, D. (2004b). NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Molecular Biology and Evolution*, *21*(10), 1972–1980. https://doi.org/10.1093/molbev/msh210

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., & Burleigh, J. G. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes, *14*(23).

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., … Albà, M. M. (2015). Origins of De Novo Genes in Human and Chimpanzee. *PLoS Genetics*, *11*(12), 1–24. https://doi.org/10.1371/journal.pgen.1005721

Salavirta, H., Oksanen, I., Kuuskeri, J., Mak̈elä, M., Laine, P., Paulin, L., & Lundell, T. (2014). Mitochondrial genome of Phlebia radiata is the second largest (156 kbp) among fungi and features signs of genome flexibility and recent recombination events. *PLoS ONE*, *9*(5). https://doi.org/10.1371/journal.pone.0097141

Saski, C., Lee, S. B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H. G., & Jansen, R. K. (2005). Complete chloroplast genome sequence of glycine max and comparative analyses with other legume genomes. *Plant Molecular Biology*, *59*(2), 309–322. https://doi.org/10.1007/s11103-005-8882-0

Schlotterer, C. (2015). Genes from scratch - the evolutionary fate of de novo genes. *Trends in Genetics*, *31*(4), 215–219. https://doi.org/10.1016/j.tig.2015.02.007

Schmitz, J. F., & Bornberg-Bauer, E. (2017). Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Research*, *6*(0), 57. https://doi.org/10.12688/f1000research.10079.1

Schmitz, J., Ullrich, K., & Bornberg-bauer, E. (2017). De Novo Genes are " Frozen Accidents " which Escaped Rapid Turnover of Pervasively Transcribed ORFs, (Saibil 2013).

Seetharam, A., Singh, U., Li, J., Bhandary, P., Arendsee, Z., & Wurtele, E. S. (2019). Maximizing prediction of orphan genes in assembled genomes. *BioRxiv*, 37–39. https://doi.org/10.1101/2019.12.17.880294

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., & Liu, C. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Research*, *47*(W1), W65–W73. https://doi.org/10.1093/nar/gkz345

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., … Sugiura, M. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. *EMBO Journal*, *5*(9), 2043–2049. https://doi.org/10.1002/j.1460-2075.1986.tb04464.x

Siekevitz, P. (1957). Powerhouse of the Cell. *Scientific American*, *197*(1), 131–144. https://doi.org/10.1038/scientificamerican0757-131

Skippingtona, E., Barkmanb, T. J., Ricea, D. W., & Palmera, J. D. (2015). Miniaturized mitogenome of the parasitic plant viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(27), E3515–E3524. https://doi.org/10.1073/pnas.1504491112

Sloan, D. B. (2013). One ring to rule them all? Genome sequencing provides new insights into the “master circle” model of plant mitochondrial DNA structure. *New Phytologist*, *200*(4), 978–985. https://doi.org/10.1111/nph.12395

Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D., & Taylor, D. R. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biology*, *10*(1). https://doi.org/10.1371/journal.pbio.1001241

Sloan, D. B., Alverson, A. J., Štorchová, H., Palmer, J. D., & Taylor, D. R. (2010). Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm Silene latifolia. *BMC Evolutionary Biology*, *10*(1). https://doi.org/10.1186/1471-2148-10-274

Smith, D. R., Crosby, K., & Lee, R. W. (2011). Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. *Genome Biology and Evolution*, *3*(1), 365–371. https://doi.org/10.1093/gbe/evr001

Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3.4: A turnkey synteny system with application to plant genomes. *Nucleic Acids Research*, *39*(10). https://doi.org/10.1093/nar/gkr123

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., … Kaessmann, H. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*, *3*(6), 2179–2190. https://doi.org/10.1016/j.celrep.2013.05.031

Stanton, B. J., Serapiglia, M. J., Smart, L. B., & others. (2014). The domestication and conservation of Populus and Salix genetic resources. *Poplars and Willows: Trees for Society and the Environment. Wallingford, UK: CAB International*, 124–199.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and Collinearity in Plant Genomes. *Science*, *320*(5875), 486–488. https://doi.org/10.1126/science.1153917

Thorsness, P., & Fox, T. (1990). Escape of DNA from mitochondria to the nucleus in Saccharomyces cerevisiae. *Nature*, *346*(4), 376–379. https://doi.org/10.1134/s032097251904002x

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11. https://doi.org/10.1093/nar/gkx391

Timmis, J. N., Ayliff, M. A., Huang, C. Y., & Martin, W. (2004). Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, *5*(2), 123–135. https://doi.org/10.1038/nrg1271

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., & Mar Albà, M. (2009). Origin of primate orphan genes: A comparative genomics approach. *Molecular Biology and Evolution*, *26*(3), 603–612. https://doi.org/10.1093/molbev/msn281

Tuskan, G. A., & Torr, P. (2007). The Genome of Black Cottonwood ,. *Science*, *1596*(2006), 1596–1605. https://doi.org/10.1126/science.1128691

Unseld, M., Marienfeld, R. J., Brandt, P., & Brennicke, A. (1997). The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides, *15*, 57–61.

Vakirlis, N., & McLysaght, A. (2019). Computational prediction of De novo emerged protein-coding genes. In *Computational Methods in Protein Evolution* (pp. 63–81). Springer.

Van de Paer, C., Bouchez, O., & Besnard, G. (2018). Prospects on the evolutionary mitogenomics of plants: A case study on the olive family (Oleaceae). *Molecular Ecology Resources*, *18*(3), 407–423. https://doi.org/10.1111/1755-0998.12742

Van Oss, S. B., & Carvunis, A. R. (2019). De novo gene birth. *PLoS Genetics*, *15*(5), 1–23. https://doi.org/10.1371/journal.pgen.1008160

Vanburen, R., & Ming, R. (2013). Organelle DNA accumulation in the recently evolved papaya sex chromosomes. *Molecular Genetics and Genomics*, *288*(5–6), 277–284. https://doi.org/10.1007/s00438-013-0747-7

Vinckenbosch, N., Dupanloup, I., & Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(9), 3220–3225. https://doi.org/10.1073/pnas.0511307103

Wang, D., Lloyd, A. H., & Timmis, J. N. (2012). Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(7), 2444–2448. https://doi.org/10.1073/pnas.1117890109

Wang, D., Qu, Z., Adelson, D. L., Zhu, J. K., & Timmis, J. N. (2014). Transcription of nuclear organellar DNA in a model plant system. *Genome Biology and Evolution*, *6*(6), 1327–1334. https://doi.org/10.1093/gbe/evu111

Wang, D., & Timmis, J. N. (2013). Cytoplasmic organelle DNA preferentially inserts into open chromatin. *Genome Biology and Evolution*, *5*(6), 1060–1064. https://doi.org/10.1093/gbe/evt070

Wang, S., Song, Q., Li, S., Hu, Z., Dong, G., Song, C., … Liu, Y. (2018). Assembly of a complete mitogenome of chrysanthemum nankingense using oxford nanopore long reads and the diversity and evolution of asteraceae mitogenomes. *Genes*, *9*(11). https://doi.org/10.3390/genes9110547

Wang, X., Cheng, F., Rohlsen, D., Bi, C., Wang, C., Xu, Y., … Ye, N. (2018). Organellar genome assembly methods and comparative analysis of horticultural plants. *Horticulture Research*, *5*(1), 1–13. https://doi.org/10.1038/s41438-017-0002-1

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., … Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7). https://doi.org/10.1093/nar/gkr1293

Werner, M. S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C., & Sommer, R. J. (2018). Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Research*, *28*(11), 1675–1687. https://doi.org/10.1101/gr.234872.118

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6), 1–22. https://doi.org/10.1371/journal.pcbi.1005595

Wilson, B. A., Foy, S. G., Neme, R., & Masel, J. (n.d.). Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecology & Evolution*, *1*(6), 0146. https://doi.org/10.1038/s41559-017-0146

Woischnik, M., & Moraes, C. T. (2002). Pattern of Organization of Human Mitochondrial

Pseudogenes in the Nuclear Genome. *Genome Research*, *12*(6), 885–893. https://doi.org/10.1101/gr.227202

Wu, J., Liu, B., Cheng, F., Ramchiary, N., Choi, S. R., Lim, Y. P., & Wang, X. W. (2012). Sequencing of chloroplast genome using whole cellular DNA and Solexa sequencing technology. *Frontiers in Plant Science*, *3*(NOV), 1–7. https://doi.org/10.3389/fpls.2012.00243

Wu, X., & Sharp, P. A. (2013). Divergent transcription: A driving force for new gene origination? *Cell*, *155*(5), 990–996. https://doi.org/10.1016/j.cell.2013.10.048

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., & Wang, S. (2009). A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE*, *4*(2), 1–12. https://doi.org/10.1371/journal.pone.0004603

Yadeta, K. A., Valkenburg, D. J., Hanemian, M., Marco, Y., & Thomma, B. P. H. J. (2014). The Brassicaceae-specific EWR1 gene provides resistance to vascular wilt pathogens. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0088230

Yoshida, T., Furihata, H. Y., To, T. K., Kakutani, T., & Kawabe, A. (2019). Genome defense against integrated organellar DNA fragments from plastids into plant nuclear genomes through DNA methylation. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-38607-6

Zhang, G. J., Dong, R., Lan, L. N., Li, S. F., Gao, W. J., & Niu, H. X. (2020). Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. *International Journal of Molecular Sciences*, *21*(3). https://doi.org/10.3390/ijms21030707

Zhou, R., Macaya-Sanz, D., Carlson, C. H., Schmutz, J., Jenkins, J. W., Kudrna, D., … Difazio, S. P. (2020). A willow sex chromosome reveals convergent evolution of complex palindromic repeats. *Genome Biology*, *21*(1), 1–19. https://doi.org/10.1186/s13059-020-1952-4

Zoschke, R., & Bock, R. (2018). Chloroplast translation: Structural and functional organization, operational control, and regulation. *Plant Cell*, *30*(4), 745–770. https://doi.org/10.1105/tpc.18.00016

CHAPTER II

The ancient Salicoid genome duplication event: A platform for reconstruction of *de novo* gene evolution in *Populus trichocarpa*

This chapter has been previously published and contains contributions from multiple authors:

Timothy B Yates, Kai Feng, Jin Zhang, Vasanth Singan, Sara S Jawdy, Priya Ranjan, Paul E Abraham, Kerrie Barry, Anna Lipzen, Chongle Pan, Jeremy Schmutz, Jin-Gui Chen, Gerald A Tuskan, Wellington Muchero, The Ancient Salicoid Genome Duplication Event: A Platform for Reconstruction of De Novo Gene Evolution in Populus trichocarpa, Genome Biology and Evolution, Volume 13, Issue 9, September 2021, evab198, https://doi.org/10.1093/gbe/evab198

Author contributions: T.B.Y. and W.M. designed this study. T.B.Y conducted data analysis and wrote the manuscript. V.S., K.B., and A.L. analyzed and curated the RNA-Seq and WGS data. S.J.S. and L.G. generated data. K.F. performed variant calling and performed data analysis. T.B.Y and J.Z. performed eQTL analysis. C.P. and P.R. contributed to data analysis. P.E.A. generated and analyzed proteomic data. T.B.Y, P.E.A., J.S., J.-G.C, G.A.T., and W.M. revised the manuscript.

Abstract

Orphan genes are characteristic genomic features that have no detectable homology to genes in any other species and represent an important attribute of genome evolution as sources of novel genetic functions. Here, we identified 445 genes specific to *Populus trichocarpa.* Of these, we performed deeper reconstruction of 13 orphan genes to provide evidence of *de novo* gene evolution. *Populus* and its sister genera *Salix* are particularly well suited for the study of orphan gene evolution because of the Salicoid whole-genome duplication event (WGD) which resulted in highly syntenic sister chromosomal segments across the Salicaceae. We leveraged this genomic feature to reconstruct *de novo* gene evolution from inter-genera, inter-species, and intra-genomic perspectives by comparing the syntenic regions within the *P. trichocarpa* reference, then *P. deltoides,* and finally *Salix purpurea.* Furthermore, we demonstrated that 86.5% of the putative orphan genes had evidence of transcription. Additionally, we also utilized the *Populus* genome-wide association mapping panel (GWAS), a collection of 1,084 undomesticated *P. trichocarpa* genotypes to further determine putative regulatory networks of orphan genes using expression quantitative trait loci (eQTL) mapping. Functional enrichment of these eQTL subnetworks identified common biological themes associated with orphan genes such as response to stress and defense response. We also identify a putative *cis*-element for a *de novo* gene and leverage conserved synteny to describe evolution of a putative transcription factor binding site. Overall, 45% of orphan genes were captured in *trans*-eQTL networks.

Introduction

To date, each new species sequenced has contained a cadre of orphan genes (Tautz & Domazet-Lošo, 2011). Although detection of orphan genes is highly dependent on the group of species being searched against, they can be defined as any gene that lacks identifiable homologs in any other species. Additionally, some orphan genes when examined in the context of closely related species may show evidence of *de novo* evolution. *De novo* genes must arise from noncoding ancestrally noncoding sequence (Arendsee et al., 2014).

Initially, there was considerable speculation surrounding the possibility of *de novo* gene evolution. Both Susumu Ohno and François Jacob supported the duplication and divergence model and thought *de novo* gene origination from noncoding sequence was highly unlikely, if not impossible (Jacob, 1977; Ohno, 1970). It was not until the early-2000s when empirical evidence for *de novo* evolved genes was available. The first example being five genes that evolved from noncoding sequence in *Drosophila* (Levine et al., 2006). Following this study, several additional examples of *de novo* genes were described in humans, plants, and primates (Knowles, & Mclysaght, 2009; Toll-Riera et al., 2009; Xiao et al., 2009).

Here, we provide evidence of *de novo*-evolved orphan genes in *Populus trichocarpa* via intra-genomic, inter-specific, and inter-genera syntenic analyses. This novel analysis is possible as a result of the Salicoid whole-genome duplication (WGD) event that all members of the Salicaceae family share (Dai et al., 2014; Tuskan et al., 2007). The Salicoid WGD occurred 58 million years ago in the ancestor of *Populus* and *Salix*, which was followed by the divergence of *Populus* and *Salix* 6 million years after this WGD event (Dai et al., 2014). Additionally, we used transcriptomic and proteomic analysis across multiple experiments and tissue types to provide evidence of functionality. We also analyzed polymorphism of orphan genes throughout the *Populus* GWAS mapping population and provide range-wide features of *P. trichocarpa* orphan genes. Lastly, we performed eQTL mapping to identify putative regulatory networks surrounding orphan genes. From these analyses, we propose novel insights into the mechanisms of *de novo* gene evolution, evidence of functionality, and characterization at the population scale.

Results

Identification and curation of *de novo* genes in *P. trichocarpa*

Orphan genes represent an important aspect of genome evolution and their accurate identification allows for insights into adaptive processes. We first identified orphan genes in *P. trichocarpa* v3.1 reference genome assembly using several filtering processes, primarily utilizing tools such as BLAST (Altschul 1997) to exclude genes with homologs in the NCBI database. Following the steps described in Figure II-S1, we identified 445 putative orphan genes in *P. trichocarpa* which had no detectable homology to any known genes based on the following analyses. For the initial BLASTP step, we used a total of 63 plant genomes available on the Phytozome database (Table II-S1). Next, we compared the remaining genes against the conserved domain database and the non-redundant protein database (nr database) (Marchler-Bauer et al., 2015). Finally, we removed candidates with missing open reading frames (ORFs) and then removed genes that had a copy number greater than one in *P. trichocarpa* to simplify downstream syntenic searches. Next, genes that had assigned gene models or transcriptomic evidence in *P. deltoides* or *S. purpurea* genomes were removed (Table II-S11). The remaining 445 genes were classified as specific to *P. trichocarpa* (Table II-S2). Within this set of 445, 386 (86.5%) met our threshold for expression based on five RNA-Seq datasets. These included 533 xylem, 529 root and 470 leaf transcriptomes from the *Populus* Genomewide Association Mapping Studies (GWAS) panel, 438 xylem transcriptomes from a *P. trichocarpa* x *P. deltoides* pseudobackcross mapping population and 37 transcriptomes from the Joint Genome Institute (JGI) Plant Gene Atlas database representing various tissue types (Table II-S3). We also found proteomic evidence for 16 putative orphan genes based on a limited scan of leaf tissue from six *P. trichocarpa* genotypes from the GWAS mapping panel (J. Zhang et al., 2018). Next, for each gene, we attempted to locate the non-genic syntenic sequence in *P. trichocarpa, P. deltoides, and S. purpurea* and only retained genes that had alignments to all syntenic regions. From this analysis, we selected 13 genes that had tell-tale evidence of *de novo* gene evolution as illustrative cases. All 13 candidates had expression evidence, and two of them had proteomic evidence (Figure II-S2).

Reconstruction of *de novo* gene evolution in *P. trichocarpa*

We used the common Salicoid WGD and resulting conserved macrosynteny to explore *de novo* gene evolution in the ancestral and extant states in *Populus* and *Salix* (Figure II-1A, II-1B). *De novo* gene reconstruction relies on identifying enabling mutations in noncoding sequence leading to the formation of the functional ORF. For example, Figure II-1C profiles a putative *de novo* gene on chromosome 1 (Chr01) of the *P. trichocarpa* reference genome. Syntenic blocks of genes exist between chromosome 1 and 3 and these syntenic regions were extracted and aligned. Two hypotheses arose from this alignment. The first is that after the WGD there were functional open reading frames, and over time the ORFs on the syntenic chromosomes accumulated mutations and eroded away, rendering them non-functional pseudogenes. The alternative, and more probable hypothesis, is that intergenic sequence was duplicated, and as such, the multiple sister chromosomes or secondary syntenic sequences represent noncoding ancestral states and outgroups that allow for a stepwise analysis of the enabling mutations leading to the evolution of a functional *de novo* gene. That is, the primary syntenic regions to Potri.001G124250 (*P. deltoides* Chr01 and *S. purpurea* Chr01) both have mutations resulting in a non-functional start codon, and all five syntenic regions, including *P. trichocarpa* Chr03, *P. deltoides* Chr03, and *S. purpurea* Chr03, have a non-functional stop codon (Figure II-1C). This hypothesis is further supported by the high number of shared disabling mutations near residue 25, residue 62, and residue 89 as well as the high level of sequence conservation shared between the sister syntenic chromosomes of *P. trichocarpa*, *P. deltoides*, and *S. purpurea* (Figure II-1C).

Figure II-1D profiles another example, Potri.001G400201, and depicts an alternative example where the gain of a start and stop codon can be observed along with a deletion at base 38 that places a potential stop codon out of frame therefore leading to a *de novo* gene in *P. trichocarpa*. We identified similar trends across most of the putative *de novo* gene alignments. Specifically, percent identity to the *de novo* gene of interest was highest for genes on the primary noncoding syntenic region (Figure II-S3A) and noncoding secondary syntenic sequences were most similar to each other (Figure II-S3B). The high similarity in the sister syntenic sequences serves as a useful comparator when examining the steps of *de novo* gene evolution. Finally, alignments to *S. purpurea* consistently had even lower identity values compared with *P. deltoides*, which can be explained by the earlier divergence of *Salix* and *Populus* (Dai et al., 2014). For example, average identity across all *P. trichocarpa de novo* genes compared to the *P. deltoides* primary syntenic chromosome alignments was 93.1% compared to 59.4% for *S. purpurea* alignments. Overall, it is evident the Salicoid WGD provides a valuable resource through highly conserved sister syntenic sequence which allows for a clear understanding of ancestral sequence level changes essential to the evolution of a *de novo* gene. Alignments and genomic coordinates for other putative *de novo* genes against their non-coding syntenic regions are available in Figure II-S4-S14, Table II-S12.

Direction of selection of the 13 *de novo genes* in the GWAS mapping panel was assessed with the ratio of piN to piS to infer molecular function (Table II-S4). Two *de novo* genes showed evidence of purifying selection (piN/piS < 1) and one showed evidence of neutral selection (piN=piS). The remaining 10 had piN/piS values greater than one, or where piN was greater than piS which may indicate positive or balancing selection.

***De novo*-evolved genes are polymorphic within the *P. trichocarpa* genome-wide association study population**

Genomic variant profiles were generated for all 42,950 *P. trichocarpa* genes and frequency of mutations were compared between non-orphan and orphan genes (Figure II-2A). Non-orphan genes are defined as genes excluded as orphan genes as a result of our curation pipeline. For all mutation impact classes, orphan genes had higher frequencies of mutations when compared to non-orphan genes. Orphan genes were not significantly correlated with regions of high or low nucleotide diversity across 1 Mb windows (Figure II-S20). Additionally, the subset of 13 *de novo* genes were inspected in more detail for variants affecting their coding potential. This analysis was performed with 917 *P. trichocarpa* individuals (Zhang et al., 2018). Although all 13 genes had high impact mutations, 7 of 13 did not have mutations impacting their ORF suggesting that they were fixed in the population (Table II-S5). Two examples that deviated substantially from the Nisqually-1 reference genome based on variant profile and had deleterious variants in their coding region, were Potri.002G127150 and Potri.005G061300 (Figure II-2B). For example, Potri.002G127150 had three nonsynonymous coding single nucleotide polymorphisms (SNPs) that were homozygous for the alternate allele in greater than 75% of individuals. This same gene also had a small proportion of individuals (0.5%) where the stop codon was lost. Additionally, Potri.005G061300 had a nonsynonymous coding SNP that was homozygous for the alternate allele in 86% of individuals and a homozygous nonsense mutation that resulted in a gained stop codon in 90% of the individuals. On the other hand, *de novo* genes that more closely matched Nisqually-1's variant profile included Potri.001G124250, Potri.001G257200, and Potri.009G129850 (Figure II-2B). Specifically, nearly all variants in the coding region of Potri.001G124250 were homozygous for the reference allele, with the exception of 5% of individuals that had a mutation resulting in the loss of a stop codon. Additionally, Potri.001G257200, and Potri.009G129850, closely matched the Nisqually-1 variants and at the population level had small numbers of individuals that were homozygous for nonsynonymous or other deleterious mutations. In summary, based on the high similarity of genotype profiles in the GWAS mapping panel to the Nisqually-1 reference genotype profile (the *P. trichocarpa* genotype where *de novo* genes were identified) and low frequency of high impact or other deleterious variants, these three genes are relatively homogenous in the GWAS mapping panel.

P. trichocarpa **orphan genes exhibit a narrow expression breadth, lower expression levels, and a subset show evidence of translation**

Orphan genes often exhibit lower expression levels and tend to have expression patterns that are relegated to specific tissues, commonly reported in male-biased and/or reproductive tissues (Cui et al., 2015; Levine et al., 2006). Thirteen expression libraries from various tissues (Table II-S5) were selected from the *Populus* Gene Atlas project and expression breadth (number of tissues in which transcription evidence is available) was determined for each of the 445 orphan genes. Expression breadth was assessed for both orphan and non-orphan genes. Results indicated that orphan genes were expressed in a smaller number of tissues and had narrower expression breadths (Figure II-3A, B). Additionally, as shown in Figure II-3A, more than 40% of orphan genes were not expressed in the tissues analyzed, compared to less than 20% for non-orphan genes. Nearly 20% of orphan genes were expressed in all 13 tissues, in comparison to nearly 30% for non-orphan genes. To further complement the expression breadth analysis, the tissue specificity index (tau) was calculated using the same 13 tissues above and confirmed that orphan gene expression was more specific to particular tissues when compared to non-orphan genes (Figure II-S15). Additionally, expression analysis across the five RNAseq datasets described above determined that orphan genes exhibited lower expression levels across all datasets when compared to non-orphan genes (Figure II-S16). In addition to transcriptome data, proteomics data was generated from leaf tissue for six *P. trichocarpa* genotypes from the GWAS mapping panel (Figure II-3C). It was determined that 16 orphan genes, including two *de novo* genes, showed evidence of translation.

cis-eQTL analysis of orphan genes provides insight into proximal regulatory features

Even though orphan genes had relatively low expression, we observed significant expression variance across the GWAS population sufficient to facilitate expression quantitative trait loci (eQTL) mapping (Figure II-3D, Table II-S3). To that end, eQTL mapping was performed using leaf and xylem transcriptomes to predict putative *cis*- and *trans*-regulatory elements underlying orphan gene expression. We identified putative *cis* elements for 88 orphan genes that exhibited expression variation in the GWAS mapping panel (Figure II-S17A, Table II-S6). Of these 88, 19 had predicted *cis* elements that could be aligned to all primary and secondary syntenic regions in *P. trichocarpa, P. deltoides,* and *S. purpurea.* Among these 19 was a *de novo* gene, Potri.001G400201 previously described above. The predicted *cis* element of Potri.001G400201 occurred in a 3-kb interval upstream of the transcription start site. The flanking sequence of three SNPs were annotated as putative transcription factor binding site (TFBS) (Chr01:42192609, Chr01:42193303, and Chr01:42193318) based on motif searches. The most significant SNP in the *cis*-eQTL region of Potri.001G400201 was predicted to fall within a homeobox TFBS (Figure II-4A, B). Additionally, after synteny-based reconstruction of this potential TFBS, the *P. deltoides* primary syntenic chromosome was the only syntenic sequence that had the same predicted homeobox TFBS. The other syntenic sequences either did not have any predicted TFBS (secondary *P. deltoides*, and primary/secondary *S. purpurea)* or had an entirely different predicted TFBS (secondary *P. trichocarpa*).

Another *de novo* gene identified with a *cis*-eQTL signal was Potri.002G037600. There were three nonsynonymous SNPs within the gene body that were associated with its expression and are likely in linkage disequilibrium with the causal variant (Figure II-4C). It is evident there was a SNP-effect on gene expression as the reference genotypes in the first two SNPs (Chr02:2428758 and Chr02:2428788) had lower expression when compared to the homozygous alternate. The opposite was true for the third SNP located in the gene body (Chr02:2428932), where the homozygous reference genotype had higher expression when compared to the homozygous alternate genotype (Figure II-4D). Collectively, we identified *cis*-acting elements for a subset of orphan genes and profiled the proximal regulatory elements of two *de novo* evolved orphan genes.

Orphan genes are found in leaf and xylem *trans*-eQTL networks

After eQTL mapping, 128 and 136 orphan genes in leaf and xylem, respectively, were shown to be putatively regulated by one or more *trans*-eQTLs (Table II-S7-S9). In these orphan gene sets, four *de* novo genes, used as illustrative cases above, could be associated with one or more *trans*-eQTLs. Interestingly, some orphan genes were found to be associated with *trans*-eQTLs that putatively regulate multiple genes (Figure II-5A). Overall, 231 and 221 SNPs within leaf and xylem *trans*-eQTLs, respectively, had 2 or more orphan genes in their putative regulatory networks (Tables II-S8-S9). Some orphan genes had *trans*-eQTLs that were exclusively found in either leaf (14.5%) or xylem (16.5%) (Fig II-S17B). Lastly, a small cohort (2.6%) had overlapping eQTL regions in xylem and leaf (Figure II-S17B, Table II-S10). From this small cohort of orphan genes, Potri.003G199150 was an example where the same *trans*-eQTL on chromosome 14 was predicted in both xylem and leaf transcriptomes (Fig II-5B). This example is also particularly interesting because of Potri.014G135300, an auxin response factor (ARF), transcription factor was present within the *trans*-eQTL interval. The closest homolog to Potri.014G135300 is ARF2 in *Arabidopsis* and has been shown to be a suppressor of auxin signaling and regulate many important developmental processes (Figure II-5D) (Lim et al., 2010).

37 and 47 *trans*-eQTL intervals that were associated with expression of more than 15 genes with at least 1 orphan gene in leaf and xylem, respectively, allowing for functional enrichment analyses using both the GO biological process and molecular function ontology (Figure II-S18-S19). Of those co-regulated networks with significant enrichment, some shared biological processes across leaf and xylem, were apparent. Most notably, those included homeostasis, cell wall related processes, nucleoside and glycosyl metabolic process, response to wounding and stress, protein modification, ubiquitination, methylation, and alkylation functions. The corresponding GO molecular function processes include calcium ion binding, pectinesterase activity, endoribonuclease activity, enzyme inhibitor activity, serine-type endopeptidase activity, and ubiquitin-protein transferase activity.

It was evident from a comparative GO enrichment in two tissues, that orphan genes were present in *trans*-eQTL-regulated networks that appear responsive to environmental stresses, although they were also present in networks representing diverse functions. One example of response to biotic stress in leaf was evident in three different co-regulated networks targeted by *trans*-eQTLs (Chr17:5204928-5223212, Chr18:6692837-6751087, and scaffold_3123:90-159). These *trans*-eQTL intervals shared response to wounding as one of their enriched biological processes (Figure II-5C). The same co-regulated networks also had serine-type endopeptidase inhibitor activity as their most significantly enriched molecular function. The overrepresentation of serine-type endopeptidases likely represents an induced defense mechanism against pathogen proteases (Gottwald et al., 2012). The eQTL interval on Chr17 had a potential regulator, (Potri.017G057500) that is an alkaline ceramidase and has been shown to play an important role in defense response in *Arabidopsis* (Wu et al., 2015). The remaining two intervals were located in gene deserts and therefore putative regulator assignment was not possible during this study.

Discussion and Conclusion

In this work, we identified 13 orphan genes which showed evidence of *de novo* evolution. All 13 of these putative *de novo* genes were expressed and two showed evidence of translation. Seven of the *de novo* genes were located on the antisense strand of an existing gene which is consistent with other studies where antisense *de novo* transcripts were identified and found to be functional (Ardern, Neuhaus, & Scherer, 2020; Blevins et al., 2021). Additionally, 10 may be under positive or balancing selection, two were under purifying selection, and one was under neutral selection. Based on a recent study of 13 genomes of rice, gene age is positively correlated with the degree of purifying selection, which is also apparent in our results (Stein et al., 2018). In total, 445 orphan genes were identified representing 1% of genes in the *P. trichocarpa* genome that can be classified as species-specific, which is lower than common estimates of 5-15%. However, these estimates vary considerably by species and methodology used for identification (Arendsee et al., 2014). One reason for this lower percentage may be that our curation pipeline was conservative (Vakirlis & McLysaght, 2019). Specifically, two primary aspects contributing to this lower percentage include locating missing gene models in *P. deltoides* and *S. purpurea,* two close relatives to *P. trichocarpa,* as well as requiring there be no expression evidence after re-mapping *P. deltoides* and *S. purpurea* RNA-Seq data to the *P. trichocarpa* genome.

Expression of orphan genes is often relegated to particular tissues when compared to more broadly expressed non-orphan genes. Most commonly, these include male reproductive tissues such as the testis, as well as the brain in humans (Begun et al., 2007; Cui et al., 2015; Li et al., 2010; Zhao et al., 2014). Consistent with previous findings we also observe significantly lower expression, expression breadth, and higher tissue specificity compared to non-orphan genes. We also examine population-level expression variation of orphan genes in leaf and xylem. Although, orphan genes have considerably lower expression variation when compared to non-orphan genes, we were still able to identify strong associations with putative *cis*-elements and *trans*-regulatory factors. Future work could further examine the genomic context underlying variation in the *P. trichocarpa* GWAS population, possibly through epigenetic and pan-genome approaches. In summary, through our extensive use of multi-tissue and population-level transcriptome datasets, we are able to confirm previously accepted orphan gene expression trends in the context of *P. trichocarpa* and show there was sufficient expression variation for association studies. We also provide evidence of translation for 16 orphan genes (including two orphan genes which showed evidence of *de novo* evolution) based on a subset of six *P. trichocarpa* genotypes. Although this is a much lower percentage compared to recent studies, a more thorough sampling of the GWAS mapping panel for proteomics analysis and ribosome profiling would allow for further discovery of orphan genes with evidence of translation (L. Zhang et al., 2019).

Regulatory networks play essential roles in the control of transcription, signaling and development (Prud'homme et al., 2007). Recently, the process of *de novo* gene integration into existing networks has been explored in more depth (Majic & Payne, 2020). Several studies have examined the integration of *de novo* genes into existing regulatory networks and rely on sequence homology of known *cis*-elements (Carvunis et al., 2012; Li et al., 2016). In our study, we expand upon homology-based detection of known regulatory elements through the use *cis*-eQTL analysis. Through the identification of probable causal SNPs that were associated with gene expression, we identified potential TFBS and provided evidence of network rewiring for the addition of *de novo* genes. The use of *cis*-eQTL analysis adds an additional layer of evidence in addition to homology guided detection of binding sites that the TFBS is likely functional and provides a foundation for future molecular validations.

In addition to gene reconstruction, synteny can also be leveraged to reconstruct the evolution of *cis* elements of *de* novo genes. *Cis* elements have been shown to be essential to *de* *novo* gene evolution by creating an appropriate context for transcription and integrating new genes into existing networks (Majic & Payne, 2020; Werner et al., 2018). Our findings provide clear examples of the evolution of a TFBS in the context of a *cis*-eQTL study. Although, we do not provide experimental evidence of transcription factor binding, future work could validate these predictions through molecular assays.

Thus far, insight into orphan gene functional repertoire has primarily been limited to examples of molecular validation. To expand and confirm upon known functions provided by molecular studies, we used functional enrichment of putative *trans*-eQTL regulatory networks that contain orphan genes to assign putative functional roles to orphan genes found in those networks. Our functional enrichment results aligned with well-known functional niches of orphan genes such as response to environmental stress and host-pathogen interaction and uncovered additional functional diversity as well. Collectively, this study captured 45% of *P. trichocarpa* orphan genes in *trans*-eQTL networks.

Forest trees are keystone species and have significant environmental importance. *P. trichocarpa* has an extensive species range which spans considerable abiotic and biotic diversity (Evans et al., 2014). From previously described functional validation in other systems, orphan genes have been shown play essential roles in adaptation and phenotypic novelty (Qi et al., 2019; Xiao et al., 2009). Moreover, the identification of orphan genes in *P. trichocarpa* may provide a platform for future studies interested in their roles in adaptive processes. By exploring *de novo* gene evolution through the lens of a WGD event which serves as an outgroup that has extensive syntenic conservation, we were able to concretely provide evidence of a noncoding ancestral state. Furthermore, the methodology developed here may enable future *de novo* gene studies in species which lack closely related outgroups but retain highly conserved whole genome duplications. Additionally, through the use of multi-omics data available to *P. trichocarpa* we could accurately describe orphans' primary functional niches and regulatory origins. Future work will place an emphasis on empirically validating their function and regulation.

Materials and Methods

Phylogenetics, synteny, and selection analysis: The species tree was constructed with Orthofinder v2.2.6 with default parameters using the primary protein isoform sequences of *P. trichocarpa* v3.1, *P. deltoides* v2.1, and *S. purpurea* v1.1 (Emms & Kelly, 2019). Macrosynteny relationships between *P. trichocarpa*, *P. detloides*, and *S. purpurea* were constructed with MCScan using default parameters (JCVI utility libraries v0.8.12) with the primary transcripts of *P. trichocarpa* v3.1, *P. deltoides* v2.1, and *S. purpurea* v1.1 (Tang et al., 2008). Selection analysis (piN/piS) was performed with SNPGenie with 917 individuals, using bi-allelic SNPs found in the CDS of the *de novo* gene with a minimum allele frequency of 0.01 (Nelson et al., 2015).

Orphan gene curation: The *P. trichocarpa* v.3.1 genome using the primary transcript from all genes was searched against 63 proteomes (Table II-S1) available in Phytozome 12 using BLASTP 2.6.0+ with an e-value cutoff of 0.001 (Altschul, 1997). This resulted in 1,079 genes that were found to be exclusive to *P. trichocarpa*. These 1,079 genes were then analyzed with BLASTP 2.6.0+ with an e-value threshold of 0.001 against the NCBI nr database excluding *Populus spp*., and 32 genes had hits within the NCBI nr database, reducing the count to 1,047. The remaining genes were then analyzed within the Conserved Domain Database which contains 50,369 PSSM, and resulted in 4 additional hits, and reduced the gene count to 1,043. Genes were then analyzed for their coding intactness, which resulted in 68 orphan genes that had missing start or stop codons which resulted in 977 total genes. These genes were further analyzed for missing homologous gene models in *S. purpurea* v.1.0 and *P. deltoides* v.2.1 genomes with genblastG v1.0.138 (She et al., 2011). An e-value cutoff of 1e-5, coverage threshold of 90%, and identity threshold of 50% was used. This excluded 329 candidate orphan genes that had missing homologous gene models, and resulted in 648 genes. A script was also used to verify the validity of the genblastG gene models and predicted models that did not have start or stop codons, had internal stop codons, or were not divisible by 3 were removed from consideration. Next, 55 genes that were duplicated in clusters of 2 or more removed, which resulted in 593 genes. The remaining 593 genes were then analyzed for expression evidence in *P. deltoides* D124 and across six tissues in *S. purpurea* 94006 (Table II-S11). These RNA-Seq datasets were aligned to *P. trichocarpa* v3.1 reference genome following the methods described below. Genes that had expression evidence of counts per million (CPM) greater than one in one replicate or greater than zero in two replicates in *P. deltoides* D124 or across six tissues in *S. purpurea* were removed, which resulted in 445 genes.

The current *P. trichocarpa* v4.1 Nisqually-1 reference genome assembly used a homology-based annotation method which excluded the majority of orphan genes since our curation pipeline specifically eliminated genes that had any detectable homology with genic features in all existing genomic databases. The exception was 24 orphan genes which were annotated based on homology to gene models in another *P. trichocarpa* genome assembly, Stettler 14 v1.1 (https://phytozome-next.jgi.doe.gov/info/PtrichocarpaStettler14_v1_1). These 24 did not have homology to any other genes outside of the 2 *P. trichocarpa* genome assemblies, further supporting our conclusions of species-specificity. Regardless of exclusion by annotation methodology, we used genblastG and were able to identify 437 of 445 orphan genes in the v4.1 reference genome suggesting a 98% transfer rate across assemblies.

***De novo* gene identification**: First, MCScanX with default parameters (Wang et al., 2012), was used to generate a collinearity map in the following way, *P. trichocarpa* vs *P. trichocarpa, P. trichocarpa* vs. *P. deltoides* and *P. trichocarpa* vs *S. purpurea* (Wang et al., 2012). In order to identify the primary and secondary syntenic chromosomes for the 445 orphan genes, five genes flanking the orphan gene were used to search each collinearity map. Some genes were missing their primary and secondary syntenic regions and following this step only 250 genes had syntenic regions in *P. trichocarpa*, *P. deltoides* and *S. purpurea.* Next, whole-genome alignments were generated with nucmer from the Mummer4 package with the same species comparisons as above (Marçais et al., 2018). The boundary coordinates of the five flanking genes were then used to extract the orphan gene region and using the identified primary and secondary chromosomes the synteny map was split into primary and secondary syntenic maps. Next, Synder 0.28.0 was used to identify the expected syntenic region of the candidate *de novo* gene within the primary and secondary syntenic map (Arendsee et al., 2019). The resulting region was then filtered to include the highest scoring interval that was at least the size of the candidate *de novo* gene and less than 125 kb, which resulted in 202 genes which had intervals in the expected syntenic regions. The regions were extracted and compiled with the candidate *de novo* gene and aligned with MAFFT linsi v7.407 with default parameter settings (Katoh & Standley, 2013). To ensure that the primary and secondary syntenic regions were non-genic we utilized multiple lines of evidence: the syntenic region was required to not overlap a gene model (annotated by JGI) on the same strand second, the use of genblastG did not result in a reasonable gene model based on the thresholds above, the absence of expression of the query *de novo* gene after remapping RNA-Seq data from both *P. deltoides* and *S. purpurea* after applying the thresholds above, and identification of shared disabling mutations in the syntenic alignments. This analysis resulted in 13 *de novo* genes that had high- quality alignments. Upon further inspection of the non-genic syntenic sequence with BLASTN 2.6.0 + (Altschul, 1997), 7 *de novo* genes could be classified as having their origins in overlapping existing gene features on the antisense strand while the remaining 6 are from intergenic regions.

SNP effect and polymorphism analysis in 917 *P. trichocarpa* individuals: A vcf, with the variant calls from 917 *P. trichocarpa* individuals was annotated with SnpEff 4.3t using the *P. trichocarpa* v.3.1 GFF3 file (available on https://phytozome.jgi.doe.gov/pz/portal.html) (Cingolani et al., 2012). To determine the variant frequency by mutation class, gene regions were extracted with bcftools v1.9, each mutation class count was divided by the gene length, and then multiplied by 1000. For similarity to Nisqually-1, gene regions were extracted with bcftools v1.9 (H. Li, 2011). Then, Plink v1.90 was used to calculate pairwise identity by state with default parameter settings, multiallelic SNPs were excluded (Purcell et al., 2007). Identity by state was calculated for each of the 13 *de novo* genes. Nucleotide diversity was calculated with the same vcf as above with VCFtools v0.1.16 with a window size of 1 Mb (Danecek et al., 2011).

Mass Spectrometry of 6 *P. trichocarpa* genotypes

Protein extraction and digestion: Six genotypes, BESC-377, BESC-907, BESC-901, BESC-886, BESC-900 and Nisqually-1, from the *P. trichocarpa* GWAS population were selected for analysis by Mass Spectrometry. Each genotype had three technical replicates. 100 mg of leaf tissue was ground with two 5 mm stainless steel grinding beads in the Qiagen TissueLyser twice for 30 s at 30 Hz. Ground tissue pellets were suspended in sodium dodecyl sulfate (SDS) lysis buffer (2% in 100 mM of NH4HCO3, 10 mM DTT). Samples were physically disrupted by bead beating (0.15 mm) at 8000 rpm for 5 min. Crude lysates were boiled for 5 min at 90 °C and then samples were adjusted to 30 mM IAA and incubated in the dark for 15 min at room temperature to avoid disulfide bridge reformation. Proteins were precipitated using a chloroform/methanol/water extraction. Dried protein pellets were resolubilized in 2% (w/v) sodium deoxycholate (SDC) (100 mM NH4HCO3) and protein amounts were estimated by performing a BCA assay (Pierce Biotechnology). For each sample, an aliquot of approximately 500 µg of protein was digested via two aliquots of sequencing-grade trypsin (Promega, 1:75 [w:w]) at two different sample dilutions, (overnight) followed by incubating 3 hours at 37°C. The peptide mixture was adjusted to 0.5%

formaldehyde (FA) to precipitate SDC. Hydrated ethyl acetate was added to each sample at a 1:1 [v:v] ratio three times to effectively remove SDC. Samples were then placed in a SpeedVac Concentrator (Thermo Fischer Scientific) to remove ethyl acetate and further concentrate the sample. The peptide-enriched flow through was quantified using the BCA assay, desalted on RP-C18 stage tips (Pierce Biotechnology) and then stored at −80°C prior to prior to LC-MS/MS analysis.

LC-MS/MS analysis: All samples were analyzed on a Q Exactive Plus mass spectrometer (Thermo Fisher Scientific) coupled with a Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific). Peptides were separated on a 75 μm inner diameter microcapillary column packed with 25 cm of Kinetex C18 resin (1.7 μm, 100 Å, Phenomenex). For each sample, a 2 μg aliquot was loaded in buffer A (0.1% formic acid, 2% acetonitrile) and eluted with a linear 150 min gradient of 2 – 20% of buffer B (0.1% formic acid, 80% acetonitrile), followed by an increase in buffer B to 30% for 10 minutes, another increase to 50% buffer for 10 minutes and concluding with a 10 min wash at 98% buffer A. The flow rate was kept at 200 nL/min. MS data was acquired with the Thermo Xcalibur software v2.2, a topN method where N could be up to 15. Target values for the full scan MS spectra were 1 x 106 charges in the 300–1,500 m/z range with a maximum injection time of 25 ms. Transient times corresponding to a resolution of 70,000 at m/z 200 were chosen. A 1.6 m/z isolation window and fragmentation of precursor ions was performed by higher-energy C-trap dissociation with a normalized collision energy of 27 eV. MS/MS sans were performed at a resolution of 17,500 at m/z 200 with an ion target-value of 1 x 106 and a maximum injection time of 50 ms. Dynamic exclusion was set to 30 s to avoid repeated sequencing of peptides.

Peptide identification and protein inference: MS raw data files were searched against the *P. trichocarpa* v3.0 reference FASTA database to which mitochondrial and chloroplast-encoded proteins had been added. A list of common protein contaminants (e.g., keratin) were appended to the reference database. A decoy database, consisting of the reversed sequences of the target database, was appended to discern the false-discovery rate (FDR) at the spectral level. For standard database searching, the peptide fragmentation spectra (MS/MS) were analyzed by the Crux pipeline v3.0 (McIlwain et al., 2014). The MS/MS were searched using the Tide algorithm and was configured to derive fully-tryptic peptides using default settings except for the following parameters: allowed clip nterm-methionine, a precursor mass tolerance of 10 ppm, a static modification on cysteines (iodoacetamide; +57.0214 Da), and dynamic modifications on methionine (oxidation; 15.9949). The results were processed by Percolator to estimate q values. Peptide spectrum matches and peptides were considered identified at a q-value <0.01. Across the entire experimental dataset, proteins were required to have at least 2 distinct peptide sequences and 2 minimum spectra per protein. All proteomics spectral data in this study was deposited at ProteomeXchange Consortium via the MASSIVE repository (https://massive.ucsd.edu/). The data can be reviewed under the username "MSV000087050_reviewer" and password "muchero".

Protein quantification: For label-free quantification, MS1-level precursor intensities were derived from MOFF (Argentini et al., 2016) using the following parameters: 10 ppm mass tolerance, retention time window for extracted ion chromatogram was 3 min, time window to get the apex for MS/MS precursor was 30 s. Protein intensity-based values, which were calculated by summing together quantified peptides, normalized by dividing by protein length and then LOESS and median central tendency procedures were performed on $\log_2$-transformed data using the freely available software Perseus (http://www.perseus-framework.org). Missing values were replaced by random numbers drawn from a normal distribution (width=0.3 and downshift = 2.8).

RNA-Seq and data analysis: For RNA extraction procedures, refer to (J. Zhang et al., 2018). Raw RNA-Seq reads were filtered and trimmed using the JGI QC pipeline. BBDuk (https://sourceforge.net/projects/bbmap/) was used to evaluate raw reads were for sequence artifacts by kmer matching (kmer=25) allowing 1 mismatch and detected artifacts were trimmed from the 3' end of the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Following trimming, reads under the length threshold were removed (minimum length 25 bases or 1/3 of the original read length; whichever was longer). Raw reads from each library were aligned to the *P. trichocarpa* v3.1 reference genome using STAR v2.6.1b (Dobin et al., 2013). FeatureCounts 1.6.3 in stranded mode was used to generate raw gene counts, excluding multi-mapping reads (Liao, et al., 2014). For the QTL mapping pedigree, eQTL xylem, root, and leaf samples, and the *P. trichocarpa* GeneAtlas dataset, EdgeR 3.24.3 was used to generate scaling factors and was subsequently converted to CPM to account for RNA composition and library size between samples (Robinson et al., 2009). Orphan genes were considered to be expressed if CPM was greater than one in one dataset or greater than zero in two datasets. The JGI Plant Gene Atlas can be found at https://phytozome.jgi.doe.gov/phytomine/aspect.do?name=Expression. The *Salix* dataset was processed with the same JGI QC pipeline, the raw reads were then mapped against the *P. trichocarpa* v3.1 reference with STAR v2.6.1b, and raw counts were generated with featureCounts in unstranded mode, excluding multi-mapping reads, and converted to CPM. For population level expression variation, the statistic used is variance. For the tissue specificity index (tau), at least one tissue was required to have a CPM value of greater than 1, the script used for this analysis is available in (Le Béguec et al., 2018). Refer to Table II-S11 for all SRA identifiers for RNA-Seq data used.

eQTL analysis: Whole-genome sequencing, short variant discovery, and functional annotation of 545 *P. trichocarpa* individuals is described in (Evans et al., 2014). The exact same analysis pipeline was used in this study, with the exception being that 917 individuals were used. This SNP dataset is available at http://bioenergycenter.org/besc/gwas/. 390 and 444 RNA-Seq samples in leaf and xylem, respectively, were used to perform eQTL analysis with EMMAX v20120210 with default parameters (Kang et al., 2010). All genes with expression evidence were used in the eQTL analysis, which is 40,301 in leaf and 39,380 in xylem. Association results were then filtered with a threshold of p-value less than or equal to 1e-10, followed by bedtools merge with a distance of 100kb to extract the eQTL interval (Quinlan et al., 2010). Additionally, in order for the eQTL interval to be considered significant, at least five SNPs needed to be present in the peak. eQTLs on different chromosomes than the target gene were considered to be *trans*-eQTL, and eQTLs on the same chromosome within 1 Mb of the target gene were classified as c*is*-eQTLs. TFBS were reconstructed with synteny via same methods described above. TFBS were identified by adding a flanking sequence (5 bp) to target SNPs of interest, which was then searched with FIMO v4.11.2 using default parameters (Grant et al., 2011) against the PlantPAN v3.0 position weight matrix (Chow et al., 2019). Functional enrichment of the networks with orphan genes was performed with ClusterProfiler v3.14.3 enrichGO function, p-values were adjusted with Benjamini-Hochberg correction (Yu et al., 2012).

References

Altschul, S. (1977). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.*, *25*(17), 3389–3402.

Ardern, Z., Neuhaus, K., & Scherer, S. (2020). Are Antisense Proteins in Prokaryotes Functional? *Frontiers in Molecular Biosciences*, *7*(August), 1–12. https://doi.org/10.3389/fmolb.2020.00187

Arendsee, Z. W., Li, L., & Wurtele, E. S. (2014). Coming of age: Orphan genes in plants. *Trends in Plant Science*, *19*(11), 698–708. https://doi.org/10.1016/j.tplants.2014.07.003

Arendsee, Z., Wilkey, A., Singh, U., Li, J., Hur, M., & Wurtele, E. (2019). Synder: Inferring Genomic Orthologs From Synteny Maps. *BioRxiv*, 554501. https://doi.org/10.1101/554501

Argentini, A., Goeminne, L. J. E., Verheggen, K., Hulstaert, N., Staes, A., Clement, L., & Martens, L. (2016). MoFF: A robust and automated approach to extract peptide ion intensities. *Nature Methods*, *13*(12), 964–966. https://doi.org/10.1038/nmeth.4075

Begun, D. J., Lindfors, H. A., Kern, A. D., & Jones, C. D. (2007). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics*, *176*(2), 1131–1137. https://doi.org/10.1534/genetics.106.069245

Blevins, W. R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J. L., Espinar, L., … Albà, M. M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature Communications*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-20911-3

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–374. https://doi.org/10.1038/nature11184

Chow, C. N., Lee, T. Y., Hung, Y. C., Li, G. Z., Tseng, K. C., Liu, Y. H., … Chang, W. C. (2019). Plantpan3.0: A new and updated resource for reconstructing transcriptional regulatory networks from chip-seq experiments in plants. *Nucleic Acids Research*, *47*(D1), D1155–D1163. https://doi.org/10.1093/nar/gky1081

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cui, X., Lv, Y., Chen, M., Nikoloski, Z., Twell, D., & Zhang, D. (2015). Young genes out of the male: An insight from evolutionary age analysis of the pollen transcriptome. *Molecular Plant*, *8*(6), 935–945. https://doi.org/10.1016/j.molp.2014.12.008

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G. A., … Yin, T. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research*, *24*(10), 1274–1277. https://doi.org/10.1038/cr.2014.83

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 1–14. https://doi.org/10.1101/466201

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., …

DiFazio, S. P. (2014). Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nature Genetics*, *46*(10), 1089–1096. https://doi.org/10.1038/ng.3075

Gottwald, S., Samans, B., Lück, S., & Friedt, W. (2012). Jasmonate and ethylene dependent defence gene expression and suppression of fungal virulence factors: Two essential mechanisms of Fusarium head blight resistance in wheat? *BMC Genomics*, *13*(1). https://doi.org/10.1186/1471-2164-13-369

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Jacob, F. (1977). Evolution and Tinkering. *Science*, *196*(4295), 1161–1166. https://doi.org/10.1210/jcem-10-10-1361

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

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

Knowles, D. G., Mclysaght, A., Knowles, D. G., & Mclysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, *19*, 1–9. https://doi.org/10.1101/gr.095026.109

Le Béguec, C., Wucher, V., Lagoutte, L., Cadieu, E., Botherel, N., Hédan, B., … Hitte, C. (2018). Characterisation and functional predictions of canine long non-coding RNAs. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-31770-2

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., & Begun, D. J. (2006). Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9935–9939. https://doi.org/10.1073/pnas.0509809103

Li, C. Y., Zhang, Y., Wang, Z., Zhang, Y., Cao, C., Zhang, P. W., … Wei, L. (2010). A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Computational Biology*, *6*(3). https://doi.org/10.1371/journal.pcbi.1000734

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, Z.-W., Chen, X., Wu, Q., Hagmann, J., Han, T.-S., Zou, Y.-P., … Guo, Y.-L. (2016). On the Origin of De Novo Genes in Arabidopsis thaliana Populations. *Genome Biology and Evolution*, *8*(7), 2190–2202. https://doi.org/10.1093/gbe/evw164

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lim, P. O., Lee, I. C., Kim, J., Kim, H. J., Ryu, J. S., Woo, H. R., & Nam, H. G. (2010). Auxin response factor 2 (ARF2) plays a major role in regulating auxin-mediated leaf longevity. *Journal of Experimental Botany*, *61*(5), 1419–1430. https://doi.org/10.1093/jxb/erq010

Majic, P., & Payne, J. L. (2020). Enhancers Facilitate the Birth of De Novo Genes and Gene Integration into Regulatory Networks. *Molecular Biology and Evolution*, *37*(4), 1165–1178. https://doi.org/10.1093/molbev/msz300

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018).

MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology, 14*(1), 1–14. https://doi.org/10.1371/journal.pcbi.1005944

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., … Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research, 43*(D1), D222–D226. https://doi.org/10.1093/nar/gku1221

McIlwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C. E., Diament, B., Frewen, B., … Noble, W. S. (2014). Crux: Rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research*, *13*(10), 4488–4491. https://doi.org/10.1021/pr500741y

Nelson, C. W., Moncla, L. H., & Hughes, A. L. (2015). SNPGenie: Estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics, 31*(22), 3709–3711. https://doi.org/10.1093/bioinformatics/btv449

Ohno, S. (1970). The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*, 120–130.

Prud'homme, B., Gompel, N., & Carroll, S. B. (2007). Emerging principles of regulatory evolution. *In the Light of Evolution, 1*, 109–127. https://doi.org/10.17226/11790

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. https://doi.org/10.1086/519795

Qi, M., Zheng, W., Zhao, X., Hohenstein, J. D., Kandel, Y., O'Conner, S., … Li, L. (2019). QQS orphan gene and its interactor NF-YC4 reduce susceptibility to pathogens and pests. *Plant Biotechnology Journal*, *17*(1), 252–263. https://doi.org/10.1111/pbi.12961

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

She, R., Chu, J. S., Uyar, B., Wang, J., Wang, K., & Chen, N. (2011). genBlastG : using BLAST searches to build homologous gene models, *27*(15), 2141–2143. https://doi.org/10.1093/bioinformatics/btr342

Stein, J. C., Yu, Y., Copetti, D., Zwickl, D. J., Zhang, L., Zhang, C., … Wing, R. A. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nature Genetics*, *50*(2), 285–296. https://doi.org/10.1038/s41588-018-0040-0

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and Collinearity in Plant Genomes. *Science, 320*(5875), 486–488. https://doi.org/10.1126/science.1153917

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, *12*(10), 692–702. https://doi.org/10.1038/nrg3053

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., & Mar Albà, M. (2009). Origin of primate orphan genes: A comparative genomics approach. *Molecular Biology and Evolution*, *26*(3), 603–612. https://doi.org/10.1093/molbev/msn281

Tuskan, G. A., & Torr, P. (2007). The Genome of Black Cottonwood ,. *Science, 1596*(2006), 1596–1605. https://doi.org/10.1126/science.1128691

Vakirlis, N., & McLysaght, A. (2019). Computational prediction of De novo emerged protein-coding genes. In *Computational Methods in Protein Evolution* (pp. 63–81). Springer.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., … Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7). https://doi.org/10.1093/nar/gkr1293

Werner, M. S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C., & Sommer, R. J. (2018). Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Research*, *28*(11), 1675–1687. https://doi.org/10.1101/gr.234872.118

Wu, J. X., Li, J., Liu, Z., Yin, J., Chang, Z. Y., Rong, C., … Yao, N. (2015). The Arabidopsis ceramidase AtACER functions in disease resistance and salt tolerance. *Plant Journal*, *81*(5), 767–780. https://doi.org/10.1111/tpj.12769

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., & Wang, S. (2009). A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE*, *4*(2), 1–12. https://doi.org/10.1371/journal.pone.0004603

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang, J., Yang, Y., Zheng, K., Xie, M., Feng, K., Jawdy, S. S., … Muchero, W. (2018). Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in Populus. *New Phytologist*, *220*(2), 502–516. https://doi.org/10.1111/nph.15297

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A. R., … Long, M. (2019). Rapid evolution of protein diversity by de novo origination in Oryza. *Nature Ecology and Evolution*, *3*(4), 679–690. https://doi.org/10.1038/s41559-019-0822-5

Zhao, L., Saelao, P., Jones, C. D., & Begun, D. J. (2014). Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science*, *343*(6172), 769–772. https://doi.org/10.1126/science.1248286

Appendix

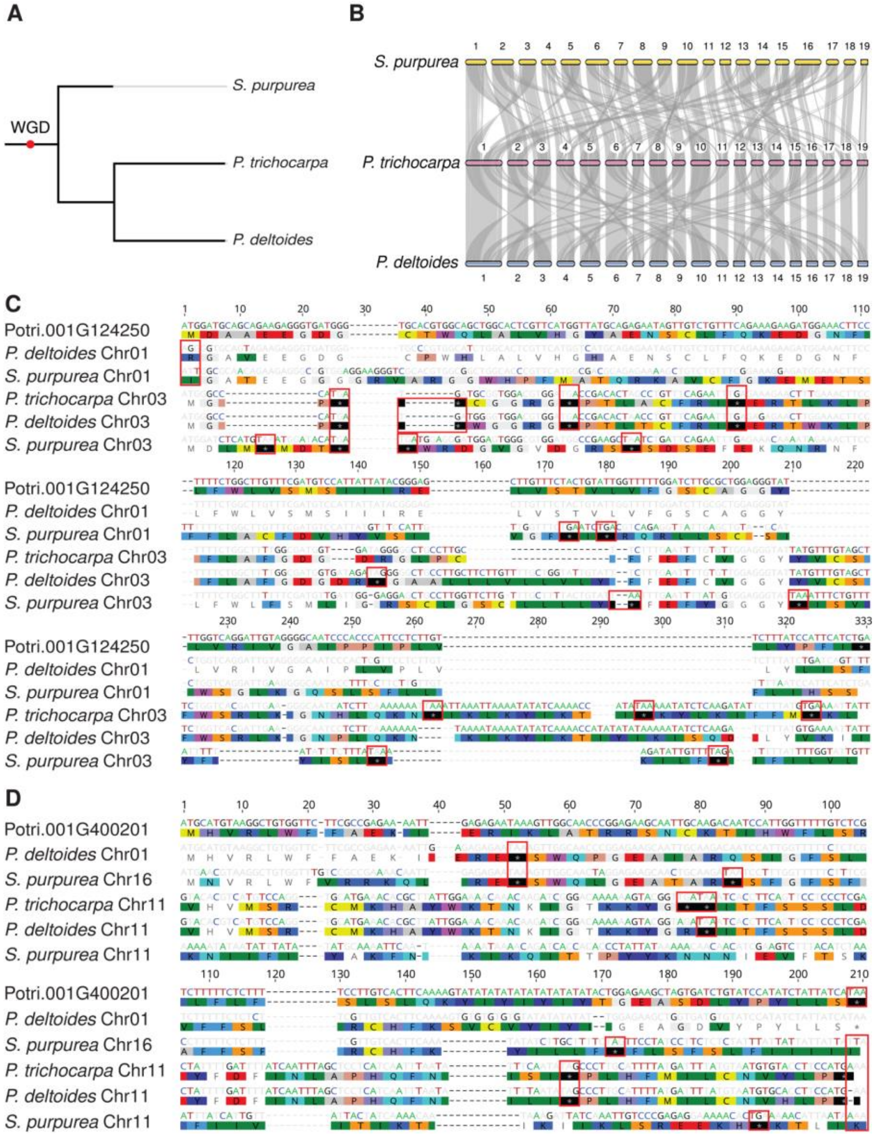

Figure II-1. A species tree and macrosynteny relationships between *P. trichocarpa*, *P. deltoides* and *S. purpurea* and alignments of *de novo* genes Potri.001G124250 and Potri.001G400201 against their respective syntenic intergenic regions in *Populus deltoides* and *Salix purpurea*. Highlighted bases and residues indicate disagreements to the *de novo* gene in *Populus trichocarpa*. Red boxes indicate disabling mutations in the intergenic syntenic sequence. Both alignments follow the sequential structure of *P. trichocarpa* (*de novo* gene), *P. deltoides* (primary syntenic region), *S. purpurea* (primary syntenic region), *P. trichocarpa* (secondary syntenic region), *P. deltoides* (secondary syntenic region), and *S. purpurea* (secondary syntenic region) (A) A species tree of the phylogeny of *P. trichocarpa*, *P. deltoides*, and *S. purpurea*, the red point on the root of the tree is indicative of the WGD which occurred in the ancestor of *Populus* and *Salix* (B) Macrosynteny relationships between *P. trichocarpa* and *S. purpurea*, and *P. trichocarpa* and *P. deltoides*, (C) Alignment of Potri.001G124250 against intergenic syntenic regions, (D) Alignment of Potri.001G400201 against intergenic syntenic regions.

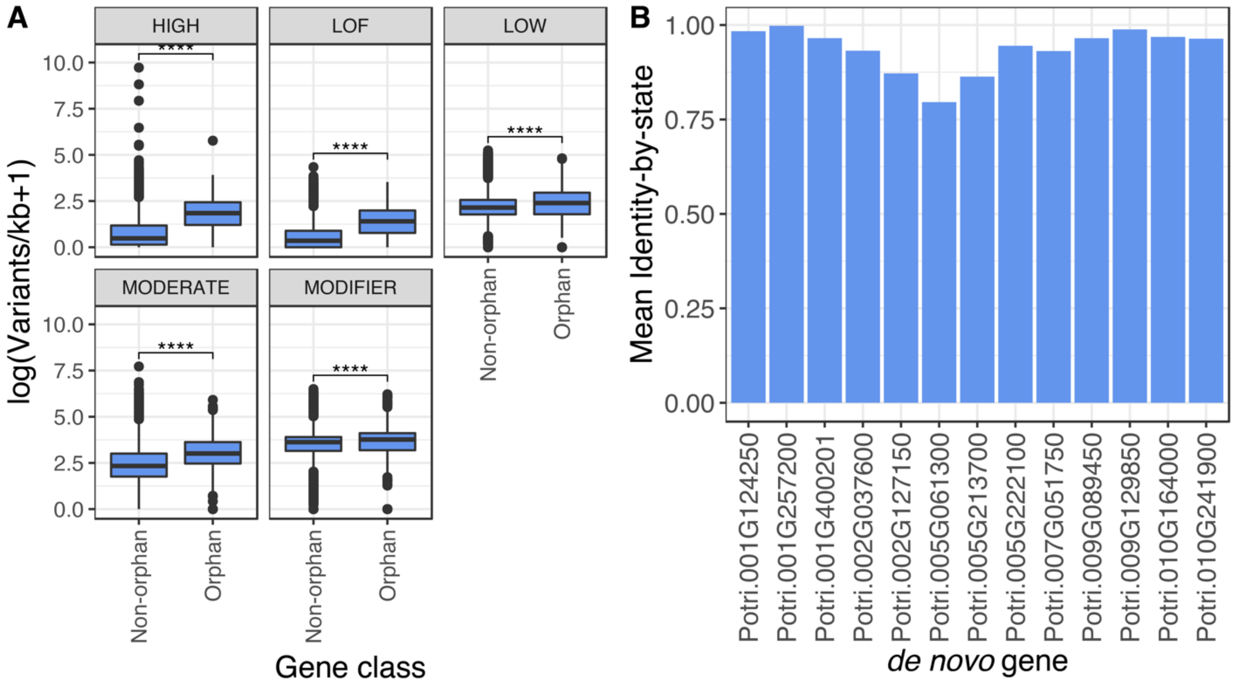

Figure II-2. Orphan gene impact class frequency and variant profile similarity to Nisqually-1 (reference genome), (A) Frequency of variant impacts by mutation class (high, LOF=predicted loss of function, low, moderate, and modifier) in 445 orphan and 42,505 non-orphan genes. Non-orphan genes are defined as genes excluded as orphan genes as a result of our curation pipeline. (****=*p*<1e-4, Wilcoxon signed rank test), (B) The 13 *de novo* genes and similarity to Nisqually-1's (reference genome) genotype profile, represented as mean identity by state.

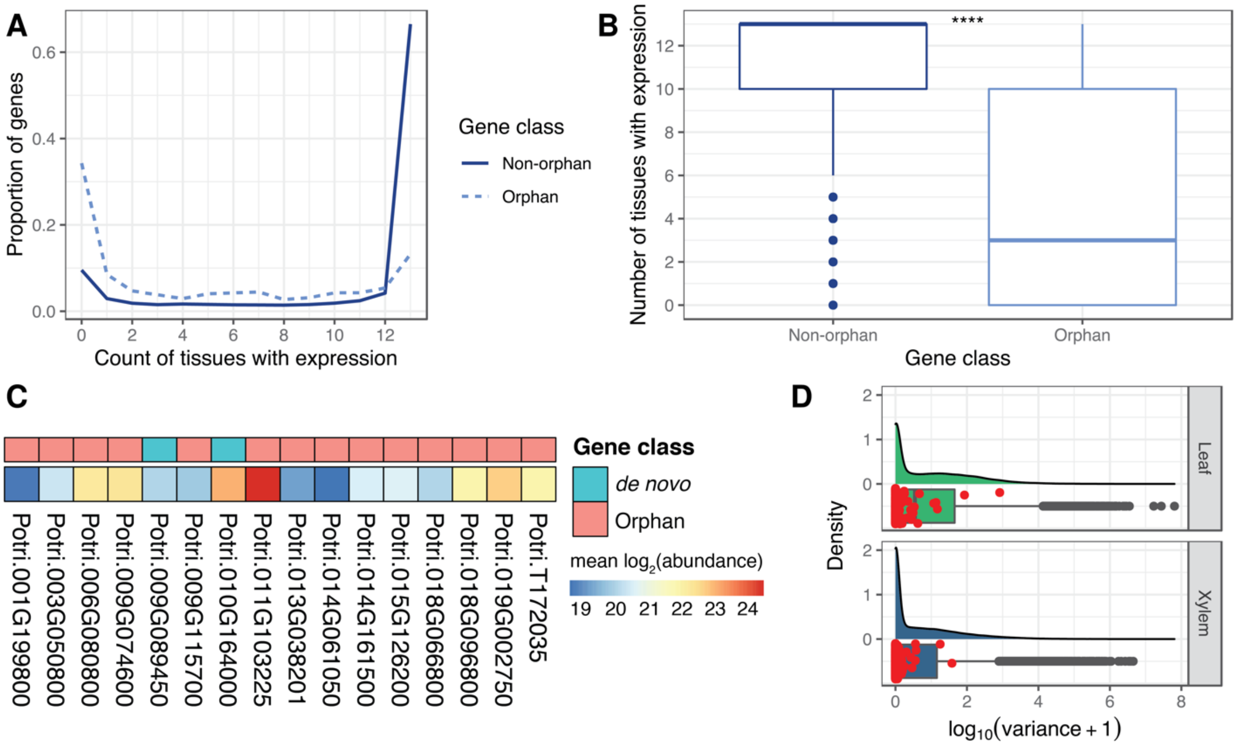

Figure II-3. Orphan gene expression breadth, translation evidence, and expression variation in the GWAS mapping panel, (A) Expression breadth across 13 *Populus trichocarpa* tissues, (B) Expression breadth distribution across 13 tissues derived from the JGI Gene Atlas project, which is publicly available RNA-Seq data (Table II-S11), (**** = *p*<1e-4, Wilcoxon signed rank test), (C) 16 orphan and *de novo* genes with proteomic data (MS/MS), shown is the mean of the $\log_2$ transformed abundance values across six *P. trichocarpa* genotypes (D) Expression variation distribution in xylem (533 genotypes) and leaf (470 genotypes) from publicly available RNA-Seq data (Table II-S11) for all *P. trichocarpa* genes with expression evidence, orphan genes are highlighted as red points.

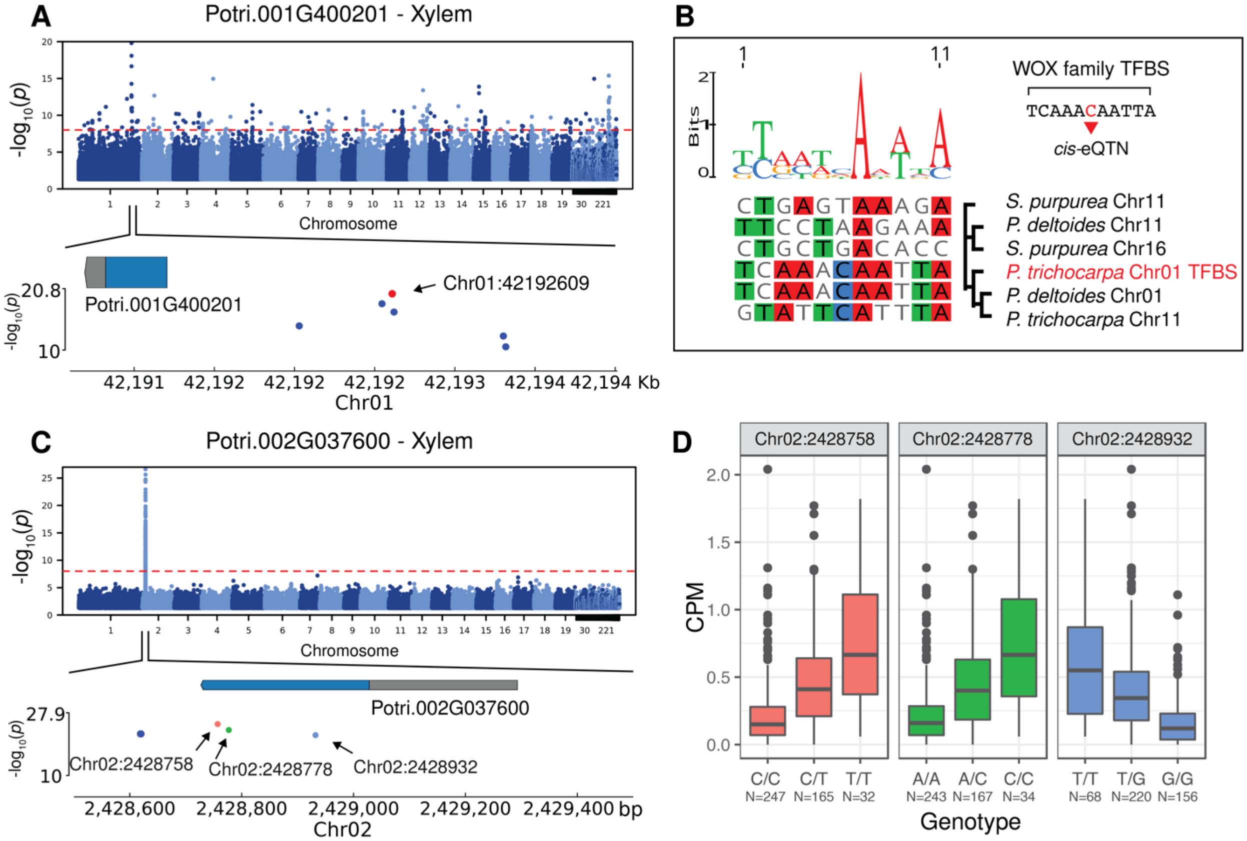

Figure II-4. *cis*-eQTL analysis facilitates assignment and reconstruction of probable proximal regulatory elements, (A) Potri.001G400201 Manhattan plot depicting the *cis*-eQTL interval, the red dot is the most significant SNP, with a significant match to a homeobox TFBS, (B) The reconstructed homeobox TFBS from the most significant SNP in Potri.001G400201's *cis*-eQTL interval (C) Potri.002G037600 Manhattan plot depicting the *cis*-eQTL interval, the three points within the exon are all non-synonymous SNPs, (D) SNP effect on gene expression for Potri.002G037600, for three exonic SNPs, Chr02:2428758, Chr02:2428788, and Chr02:2428932.

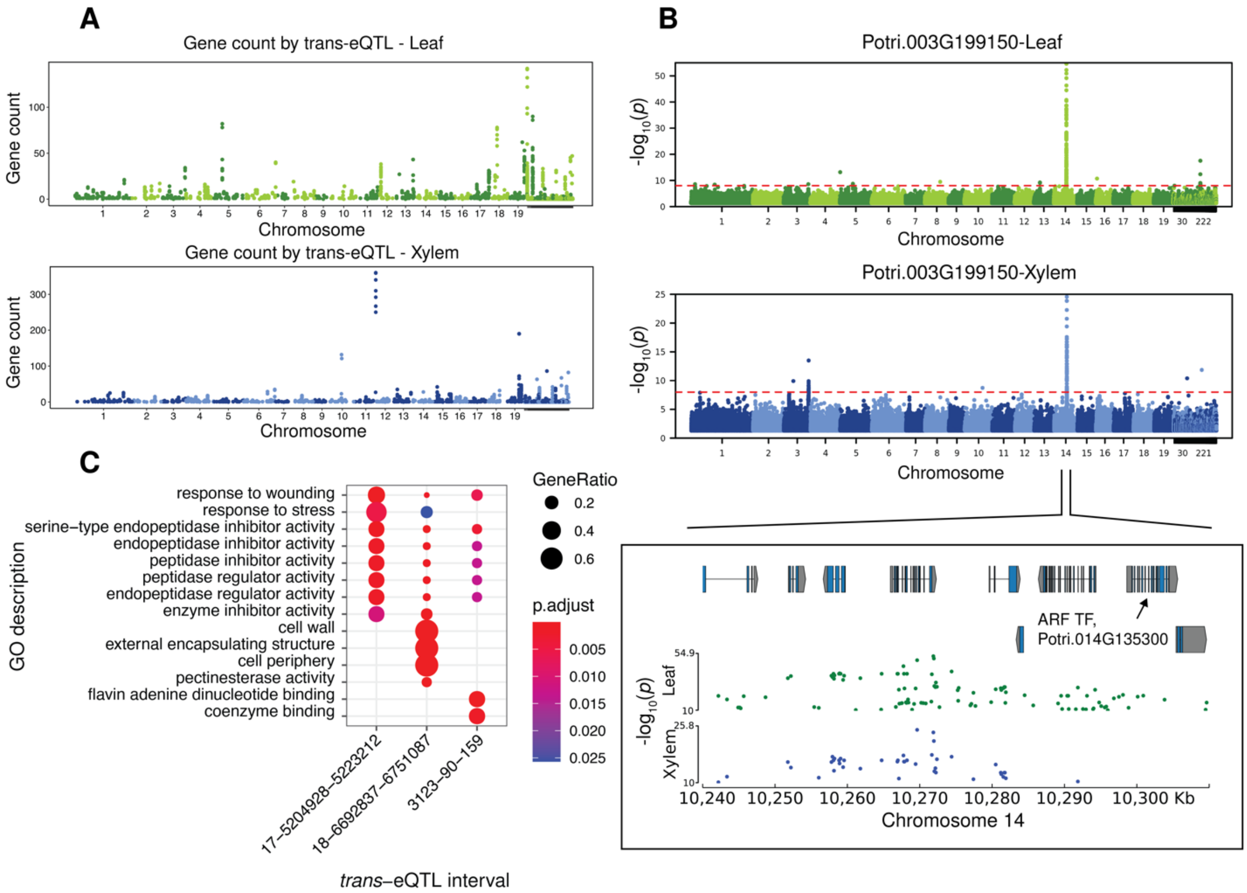

Figure II-5. *trans*-eQTL analysis of orphan genes in the GWAS mapping panel (A) Count of genes regulated by putative *trans*-eQTL (SNPs) in leaf and xylem, all trans-eQTL shown putatively regulate one or more orphan gene, (B) *trans*-eQTL analysis in leaf and xylem for Potri.003G199150, (C) Functional enrichment of three *trans*-eQTL intervals with similar functional roles, all three GO ontologies are represented in this figure.

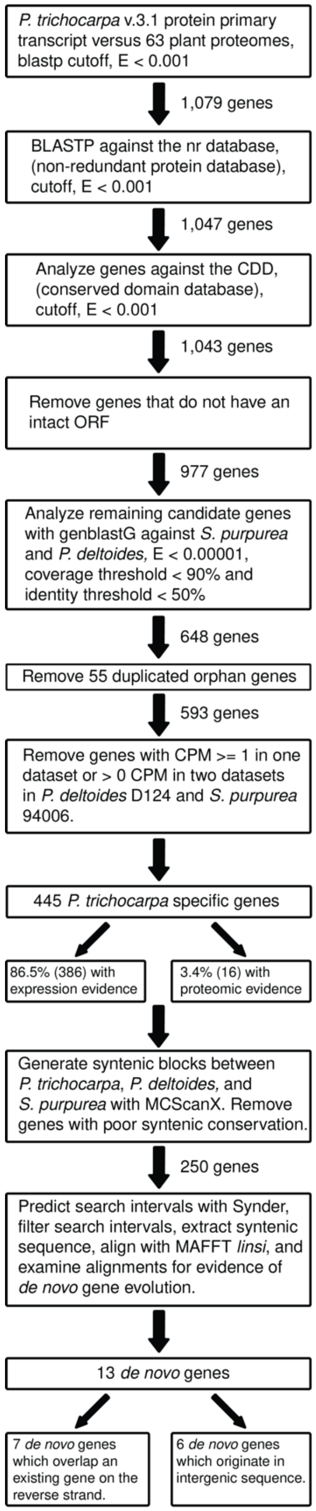
P. trichocarpa v.3.1 protein primary transcript versus 63 plant proteomes, blastp cutoff, E < 0.001
1,079 genes
BLASTP against the nr database, (non-redundant protein database), cutoff, E < 0.001
1,047 genes
Analyze genes against the CDD, (conserved domain database), cutoff, E < 0.001
1,043 genes
Remove genes that do not have an intact ORF
977 genes
Analyze remaining candidate genes with genblastG against S. purpurea and P. deltoides, E < 0.00001, coverage threshold < 90% and identity threshold < 50%
648 genes
Remove 55 duplicated orphan genes
593 genes
Remove genes with CPM >= 1 in one dataset or > 0 CPM in two datasets in P. deltoides D124 and S. purpurea 94006.
445 P. trichocarpa specific genes
86.5% (386) with expression evidence
3.4% (16) with proteomic evidence
Generate syntenic blocks between P. trichocarpa, P. deltoides, and S. purpurea with MCScanX. Remove genes with poor syntenic conservation.
250 genes
Predict search intervals with Synder, filter search intervals, extract syntenic sequence, align with MAFFT linsi, and examine alignments for evidence of de novo gene evolution.
13 de novo genes
7 de novo genes which overlap an existing gene on the reverse strand.
6 de novo genes which originate in intergenic sequence.

Figure II-S1. Orphan and *de novo* gene curation pipeline.

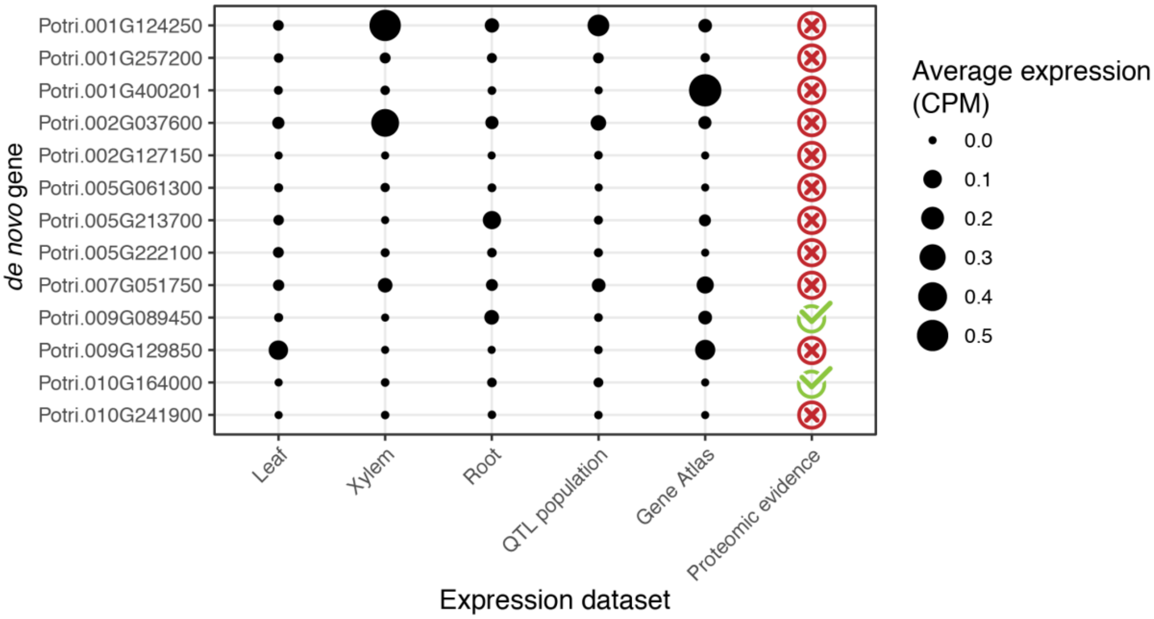

Figure II-S2. Expression and proteomic evidence for the 13 *P. trichocarpa de novo* genes. All RNA-Seq data displayed here is publicly available (Table II-S11).

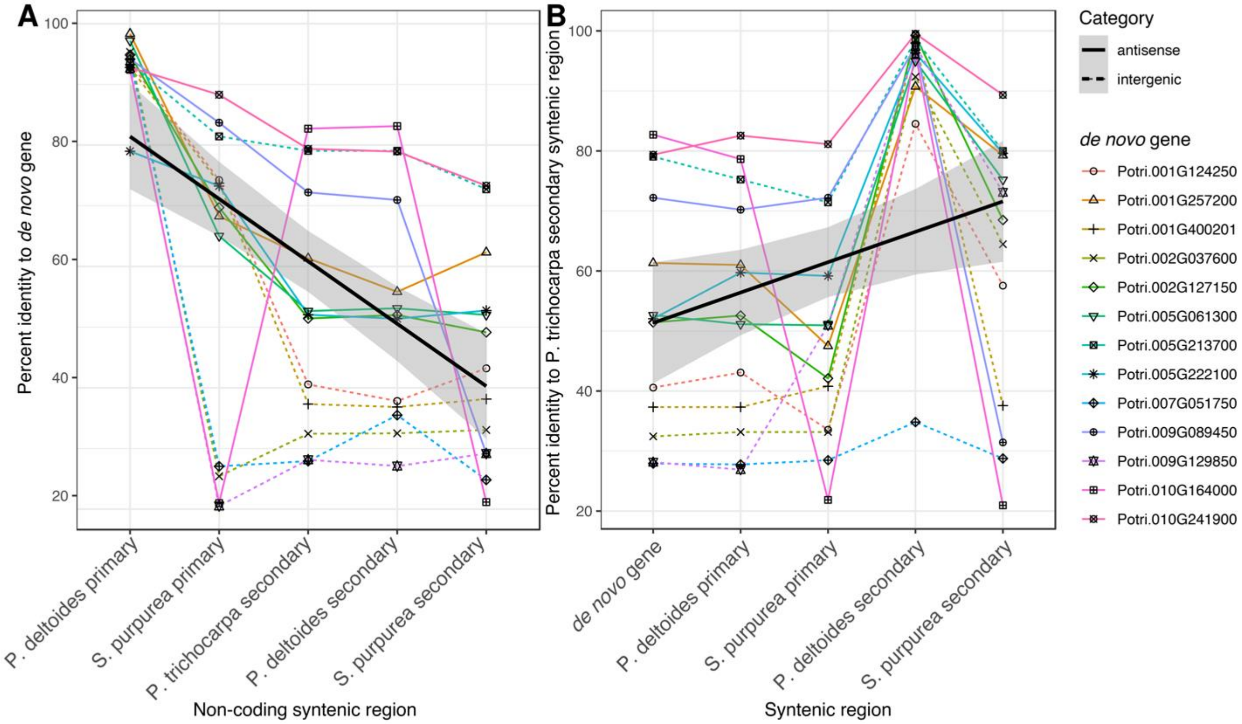

Figure II-S3. Identity plots to syntenic regions, (A) Identity of the noncoding syntenic region to the *de novo gene*, (B) identity of the noncoding *P. trichocarpa* secondary syntenic region to the other syntenic positions.

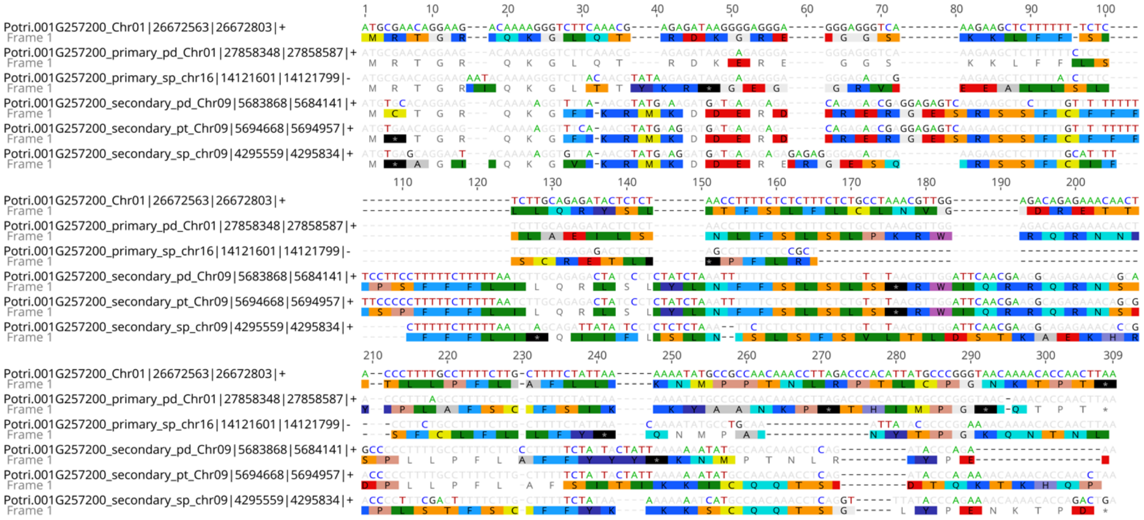
Potri.001G257200_Chr01|26672563|26672803|+
Potri.001G257200_primary_pd_Chr01|27858348|27858587|+
Potri.001G257200_primary_sp_chr16|14121601|14121799|-
Potri.001G257200_secondary_pd_Chr09|5683868|5684141|+
Potri.001G257200_secondary_pt_Chr09|5694668|5694957|+
Potri.001G257200_secondary_sp_chr09|4295559|4295834|+
Frame 1

Figure II-S4. Alignment of *de novo* gene Potri.001G257200 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

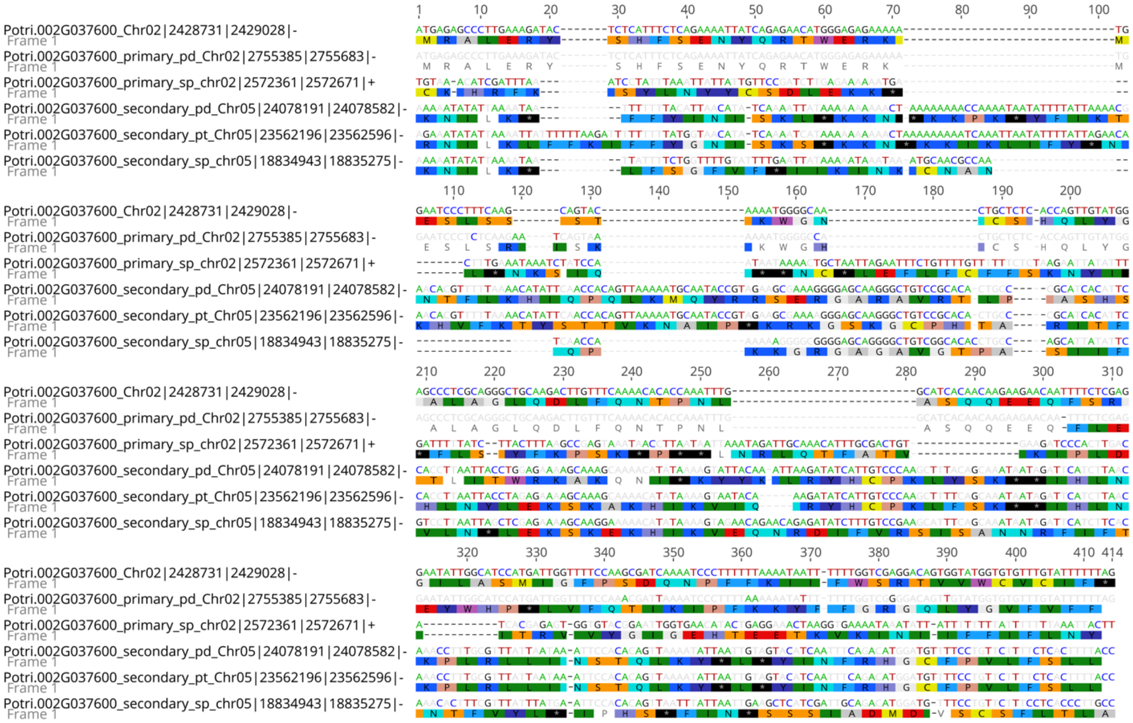
Potri.002G037600_Chr02|2428731|2429028|-
Potri.002G037600_primary_pd_Chr02|2755385|2755683|-
Potri.002G037600_primary_sp_chr02|2572361|2572671|+
Potri.002G037600_secondary_pd_Chr05|24078191|24078582|-
Potri.002G037600_secondary_pt_Chr05|23562196|23562596|-
Potri.002G037600_secondary_sp_chr05|18834943|18835275|-
Frame 1

Figure II-S5. Alignment of *de novo* gene Potri.002G037600 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

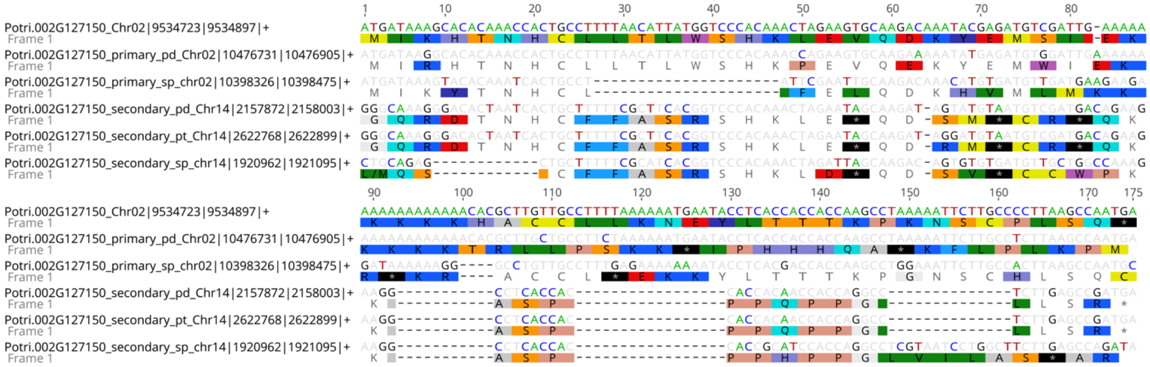
Potri.002G127150_Chr02|9534723|9534897|+
Potri.002G127150_primary_pd_Chr02|10476731|10476905|+
Potri.002G127150_primary_sp_chr02|10398326|10398475|+
Potri.002G127150_secondary_pd_Chr14|2157872|2158003|+
Potri.002G127150_secondary_pt_Chr14|2622768|2622899|+
Potri.002G127150_secondary_sp_chr14|1920962|1921095|+
Frame 1

Figure II-S6. Alignment of *de novo* gene Potri.002G127150 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

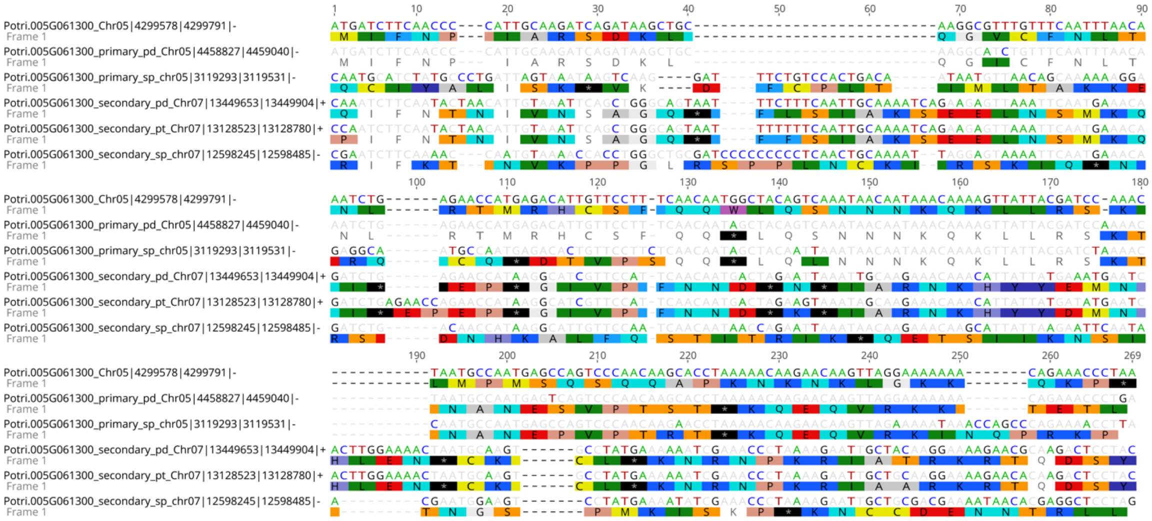
Potri.005G061300_Chr05|4299578|4299791|-
Potri.005G061300_primary_pd_Chr05|4458827|4459040|-
Potri.005G061300_primary_sp_chr05|3119293|3119531|-
Potri.005G061300_secondary_pd_Chr07|13449653|13449904|+
Potri.005G061300_secondary_pt_Chr07|13128523|13128780|+
Potri.005G061300_secondary_sp_chr07|12598245|12598485|-
Frame 1

Figure II-S7. Alignment of *de novo* gene Potri.005G061300 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

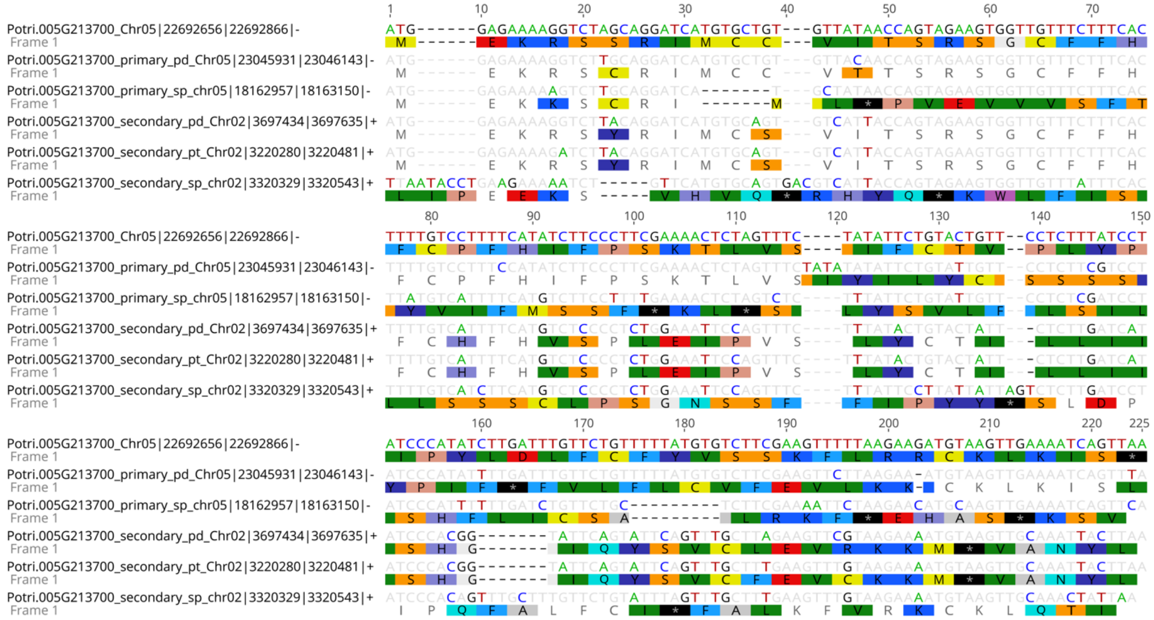
Potri.005G213700_Chr05|22692656|22692866|-
Potri.005G213700_primary_pd_Chr05|23045931|23046143|-
Potri.005G213700_primary_sp_chr05|18162957|18163150|-
Potri.005G213700_secondary_pd_Chr02|3697434|3697635|+
Potri.005G213700_secondary_pt_Chr02|3220280|3220481|+
Potri.005G213700_secondary_sp_chr02|3320329|3320543|+
Frame 1

Figure II-S8. Alignment of *de novo* gene Potri.005G213700 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

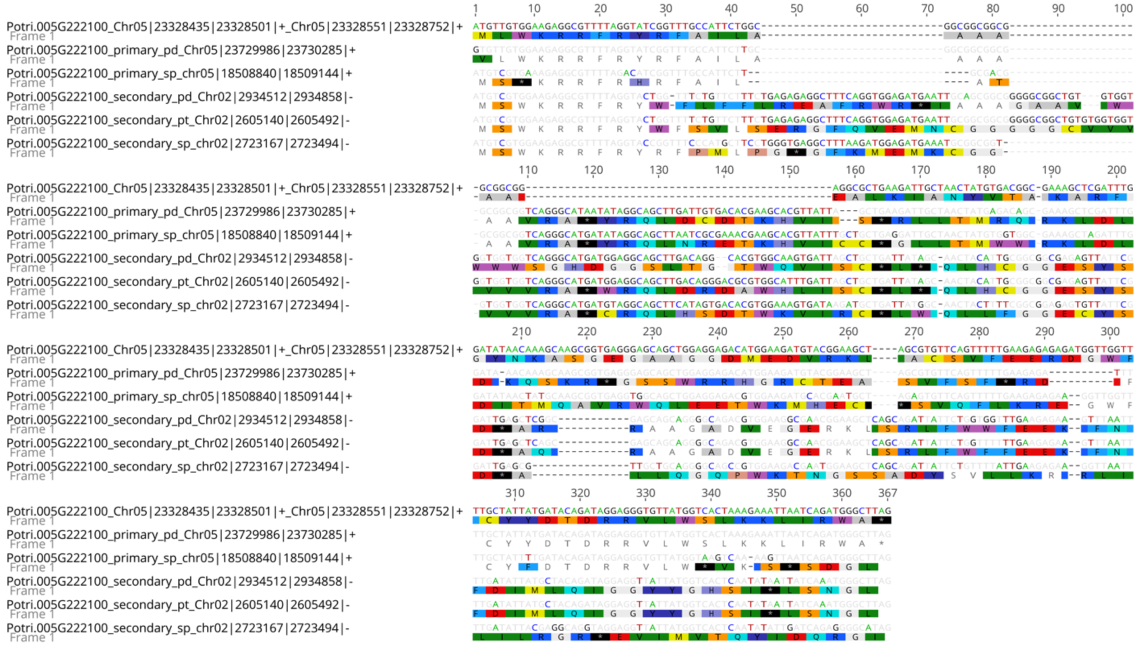
Potri.005G222100_Chr05|23328435|23328501|+_Chr05|23328551|23328752|+
Potri.005G222100_primary_pd_Chr05|23729986|23730285|+
Potri.005G222100_primary_sp_chr05|18508840|18509144|+
Potri.005G222100_secondary_pd_Chr02|2934512|2934858|-
Potri.005G222100_secondary_pt_Chr02|2605140|2605492|-
Potri.005G222100_secondary_sp_chr02|2723167|2723494|-
Frame 1

Figure II-S9. Alignment of *de novo* gene Potri.005G222100 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

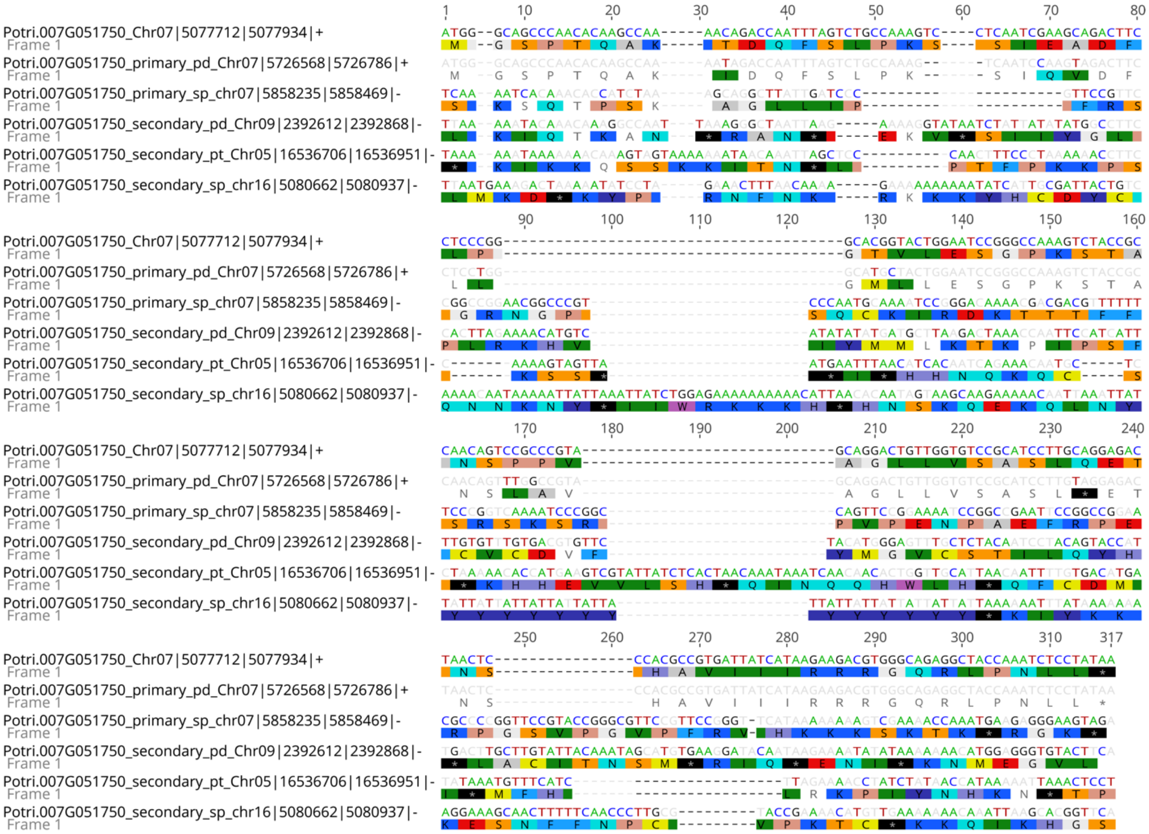
Potri.007G051750_Chr07|5077712|5077934|+
Potri.007G051750_primary_pd_Chr07|5726568|5726786|+
Potri.007G051750_primary_sp_chr07|5858235|5858469|-
Potri.007G051750_secondary_pd_Chr09|2392612|2392868|-
Potri.007G051750_secondary_pt_Chr05|16536706|16536951|-
Potri.007G051750_secondary_sp_chr16|5080662|5080937|-
Frame 1

Figure II-S10. Alignment of *de novo* gene Potri.007G051750 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

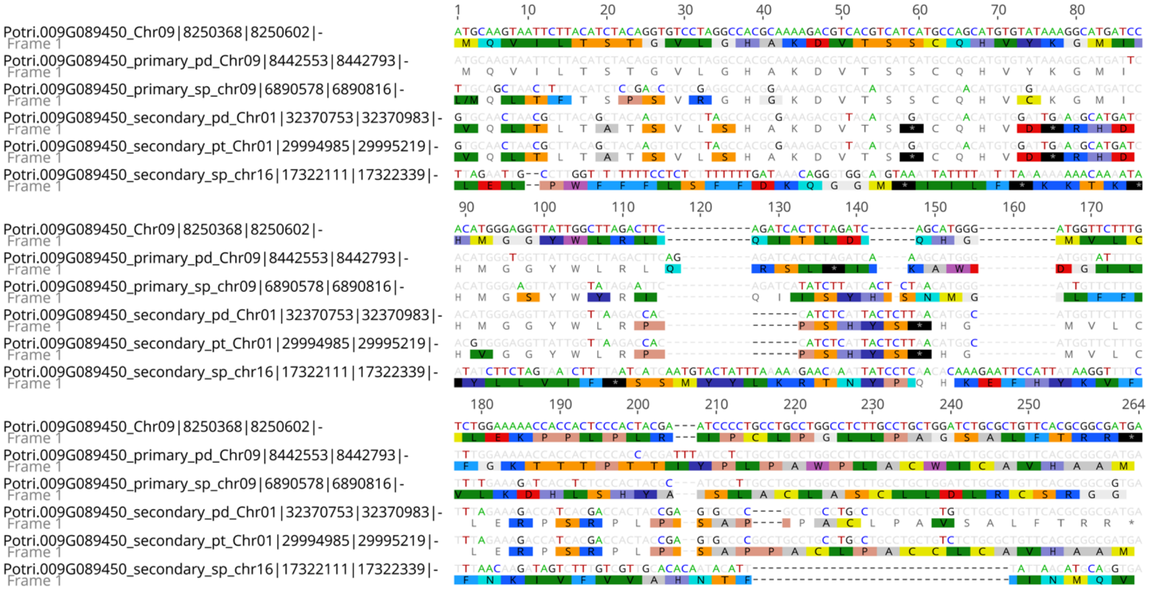
1
10
20
30
40
50
60
70
80
Potri.009G089450_Chr09|8250368|8250602|-
Frame 1
Potri.009G089450_primary_pd_Chr09|8442553|8442793|-
Frame 1
Potri.009G089450_primary_sp_chr09|6890578|6890816|-
Frame 1
Potri.009G089450_secondary_pd_Chr01|32370753|32370983|-
Frame 1
Potri.009G089450_secondary_pt_Chr01|29994985|29995219|-
Frame 1
Potri.009G089450_secondary_sp_chr16|17322111|17322339|-
Frame 1
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
264

Figure II-S11 Alignment of *de novo* gene Potri.009G089450 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

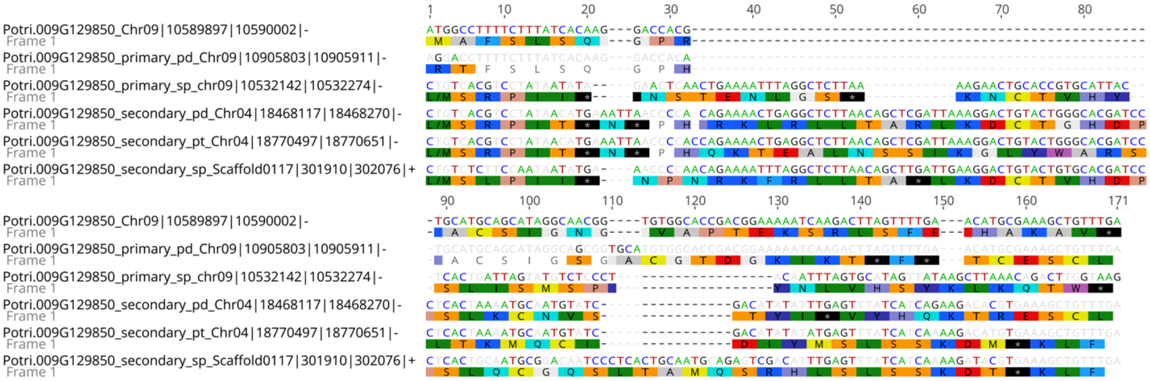
1
10
20
30
40
50
60
70
80
Potri.009G129850_Chr09|10589897|10590002|-
Frame 1
Potri.009G129850_primary_pd_Chr09|10905803|10905911|-
Frame 1
Potri.009G129850_primary_sp_chr09|10532142|10532274|-
Frame 1
Potri.009G129850_secondary_pd_Chr04|18468117|18468270|-
Frame 1
Potri.009G129850_secondary_pt_Chr04|18770497|18770651|-
Frame 1
Potri.009G129850_secondary_sp_Scaffold0117|301910|302076|+
Frame 1
90
100
110
120
130
140
150
160
171

Figure II-S12. Alignment of *de novo* gene Potri.009G129850 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

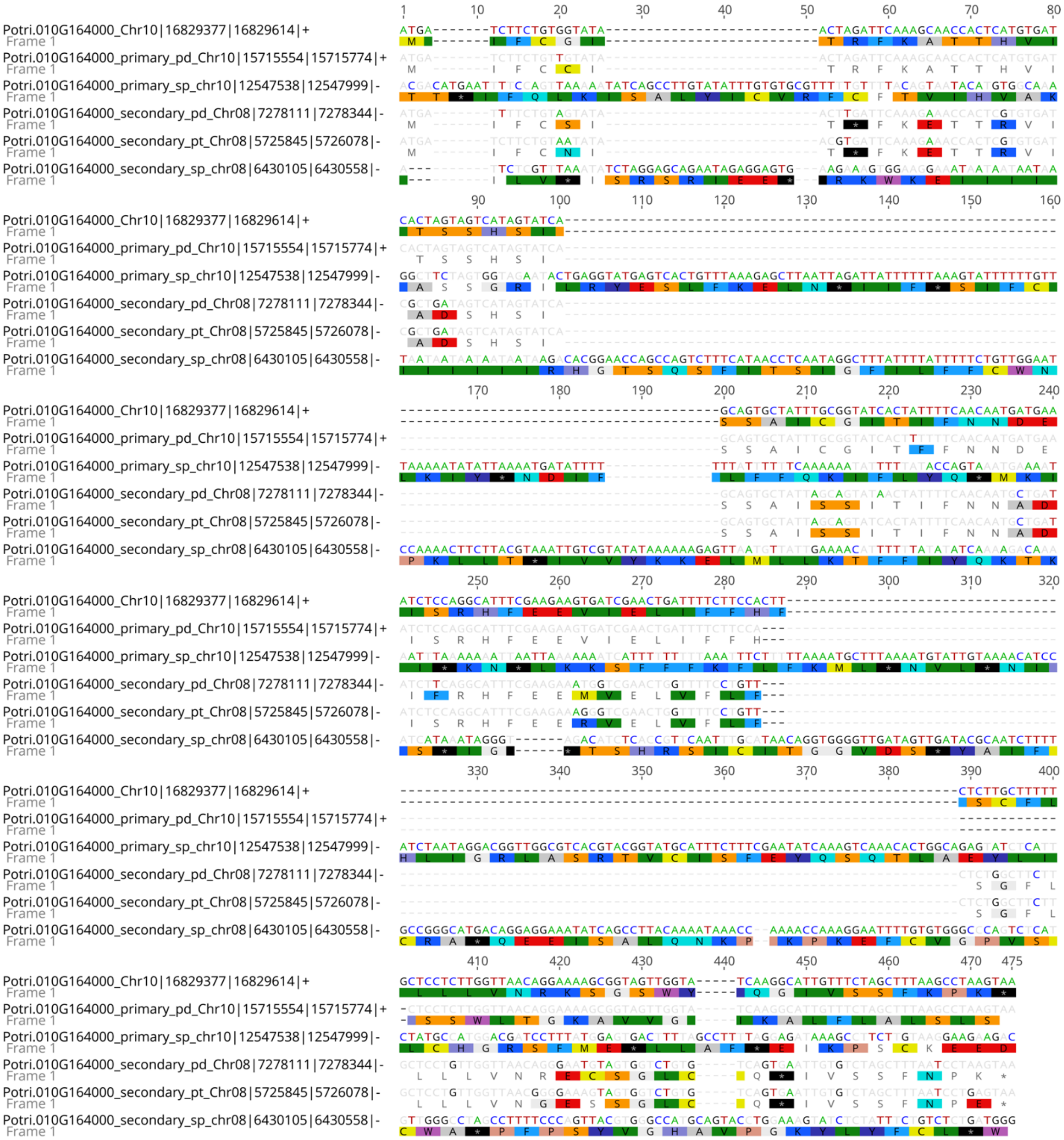
Potri.010G164000_Chr10|16829377|16829614|+
Potri.010G164000_primary_pd_Chr10|15715554|15715774|+
Potri.010G164000_primary_sp_chr10|12547538|12547999|-
Potri.010G164000_secondary_pd_Chr08|7278111|7278344|-
Potri.010G164000_secondary_pt_Chr08|5725845|5726078|-
Potri.010G164000_secondary_sp_chr08|6430105|6430558|-
Frame 1

Figure II-S13. Alignment of *de novo* gene Potri.010G164000 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

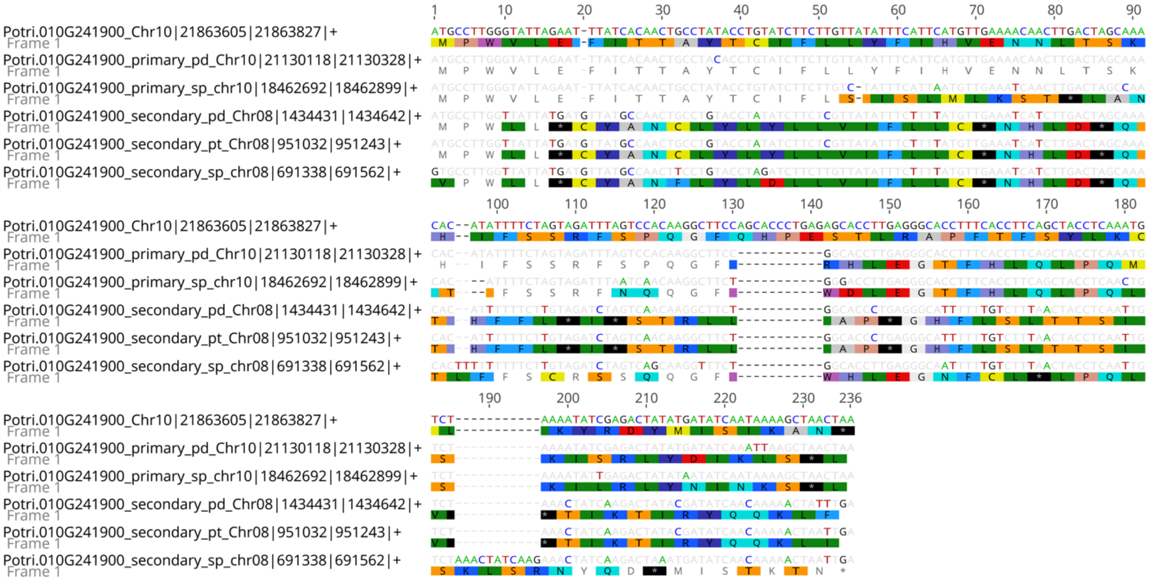
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
236
Potri.010G241900_Chr10|21863605|21863827|+
Frame 1
Potri.010G241900_primary_pd_Chr10|21130118|21130328|+
Frame 1
Potri.010G241900_primary_sp_chr10|18462692|18462899|+
Frame 1
Potri.010G241900_secondary_pd_Chr08|1434431|1434642|+
Frame 1
Potri.010G241900_secondary_pt_Chr08|951032|951243|+
Frame 1
Potri.010G241900_secondary_sp_chr08|691338|691562|+
Frame 1

Figure II-S14. Alignment of *de novo* gene Potri.010G241900 against its non-coding syntenic regions. Highlighting shows disagreement to the *de novo* gene. The sequence label includes information about species and coordinates for the region; pd = *P. deltoides,* sp = *S. purpurea*, pt = *P. trichocarpa*.

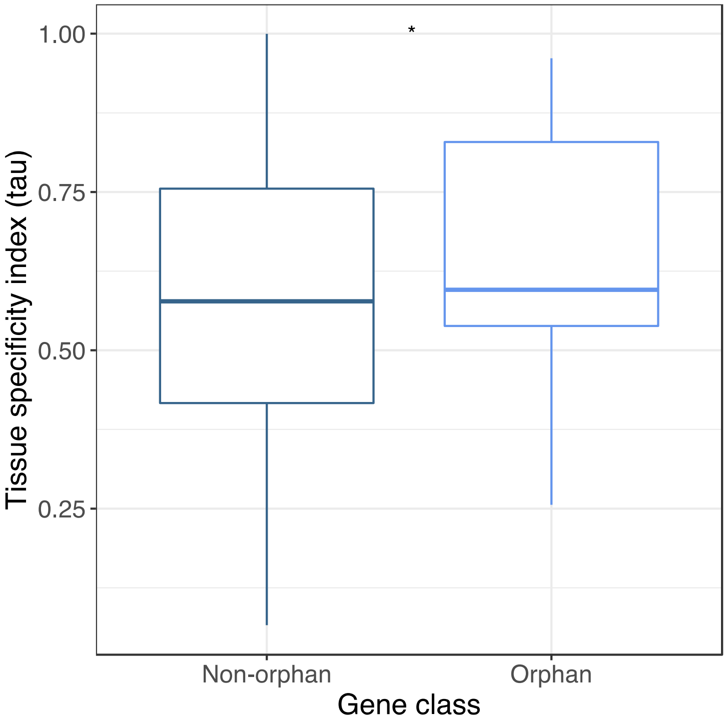

Figure II-S15. Tissue specificity index (tau) comparison between non-orphan and orphan genes. *: $p < 0.05$, Wilcoxon signed rank test.

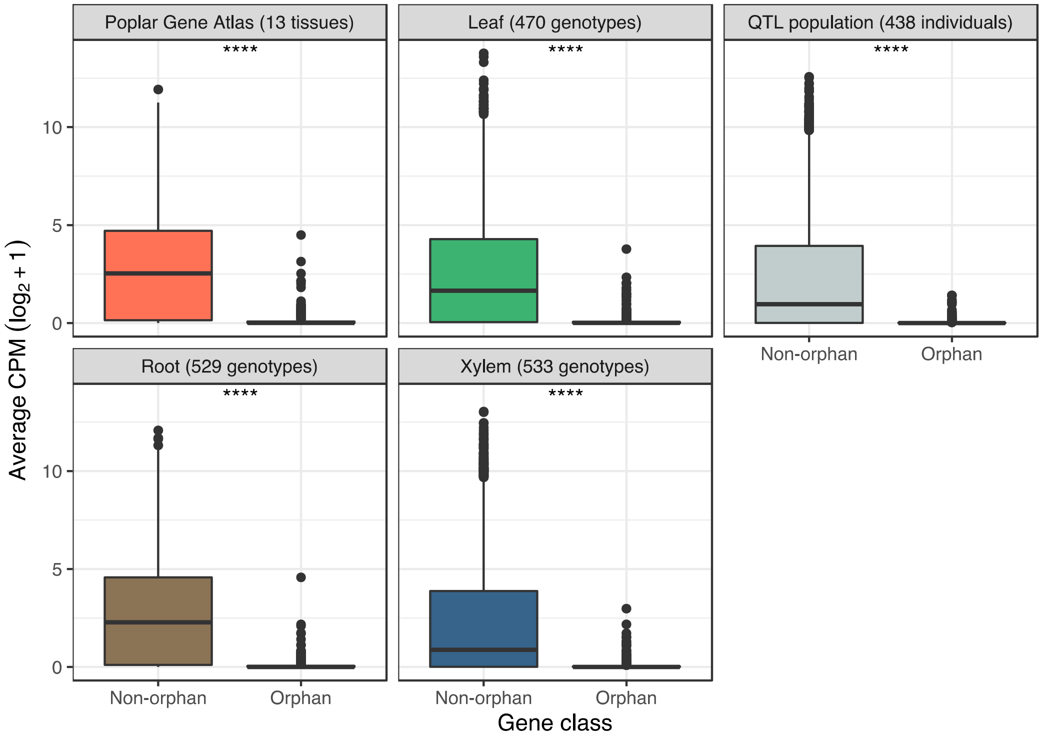

Figure II-S16. Non-orphan gene and orphan gene expression across five data sets. ****: p <= 0.0001, Wilcoxon signed rank test. Non-orphan genes are defined as genes excluded as orphan genes as a result of our curation pipeline. All RNA-Seq data displayed here is publicly available (Table II-S11).

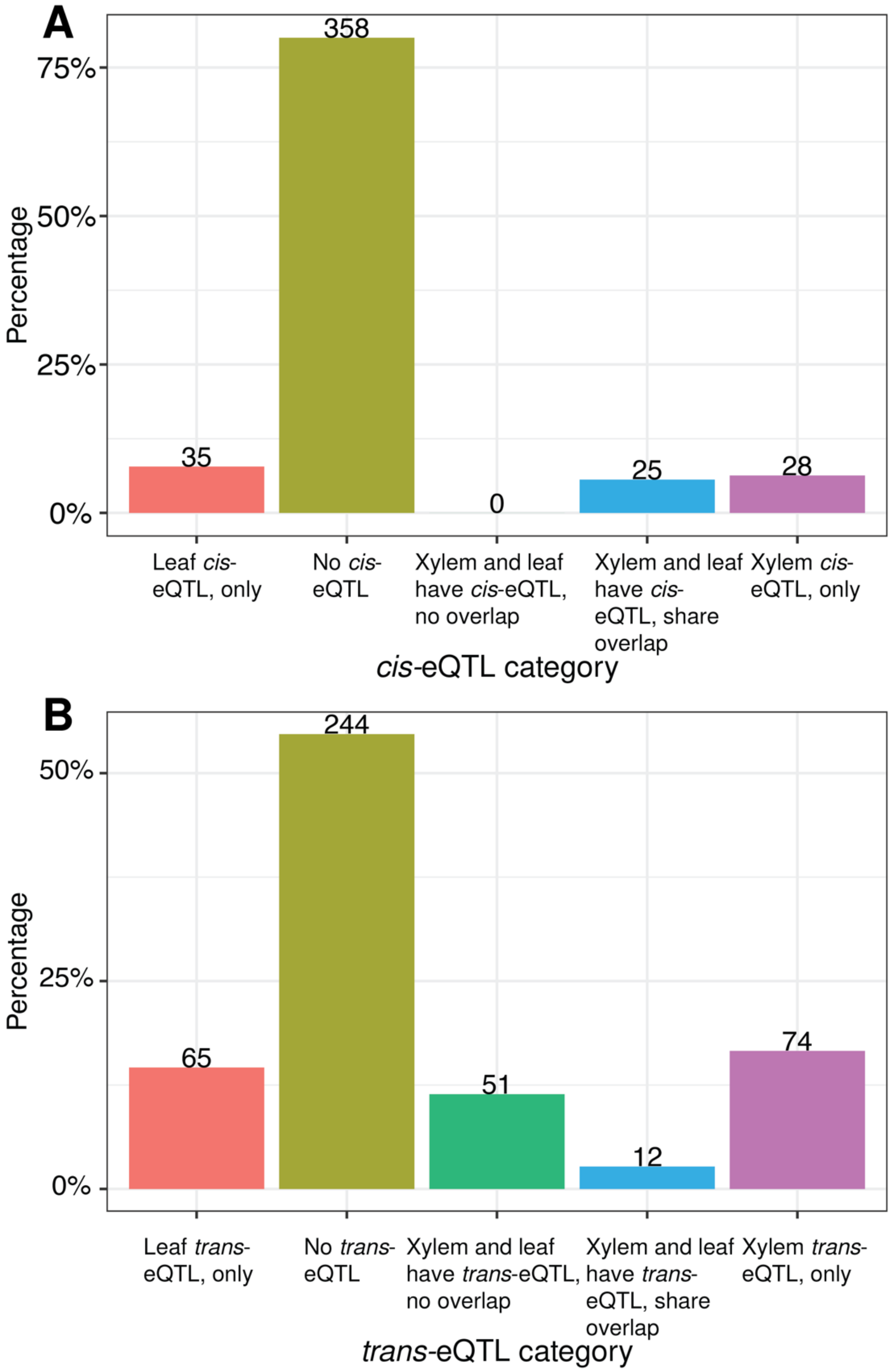

Figure II-S17. Percentage and count of orphan genes by eQTL category, (A) *cis*-eQTL and (B) *trans*-eQTL.

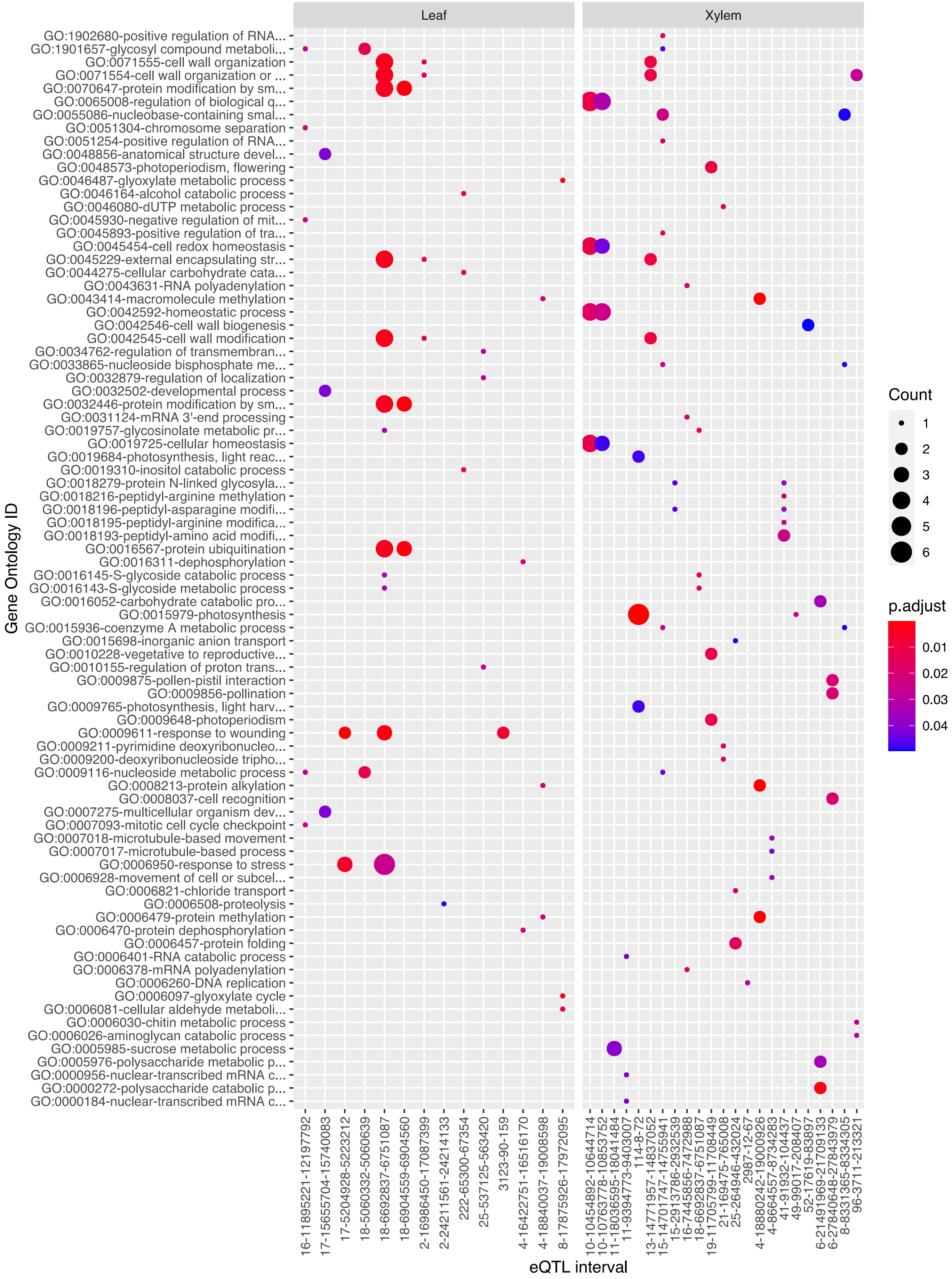

Figure II-S18. Comparative functional analysis of genes in trans-eQTL intervals in xylem and leaf based on the ‘BP’ GO ontology.

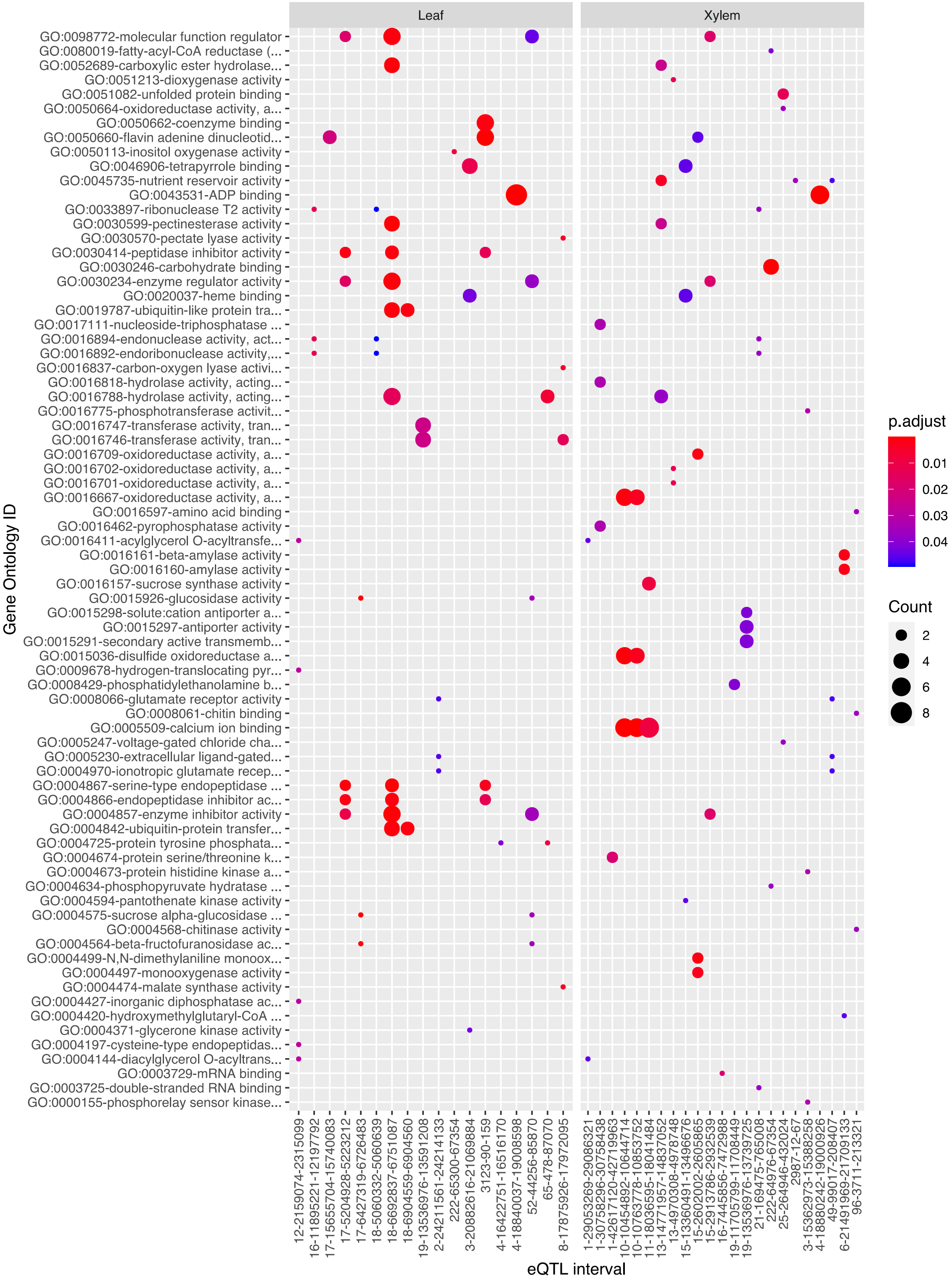

Figure II-S19. Comparative functional analysis of genes in trans-eQTL intervals in xylem and leaf based on the 'MF' GO ontology.

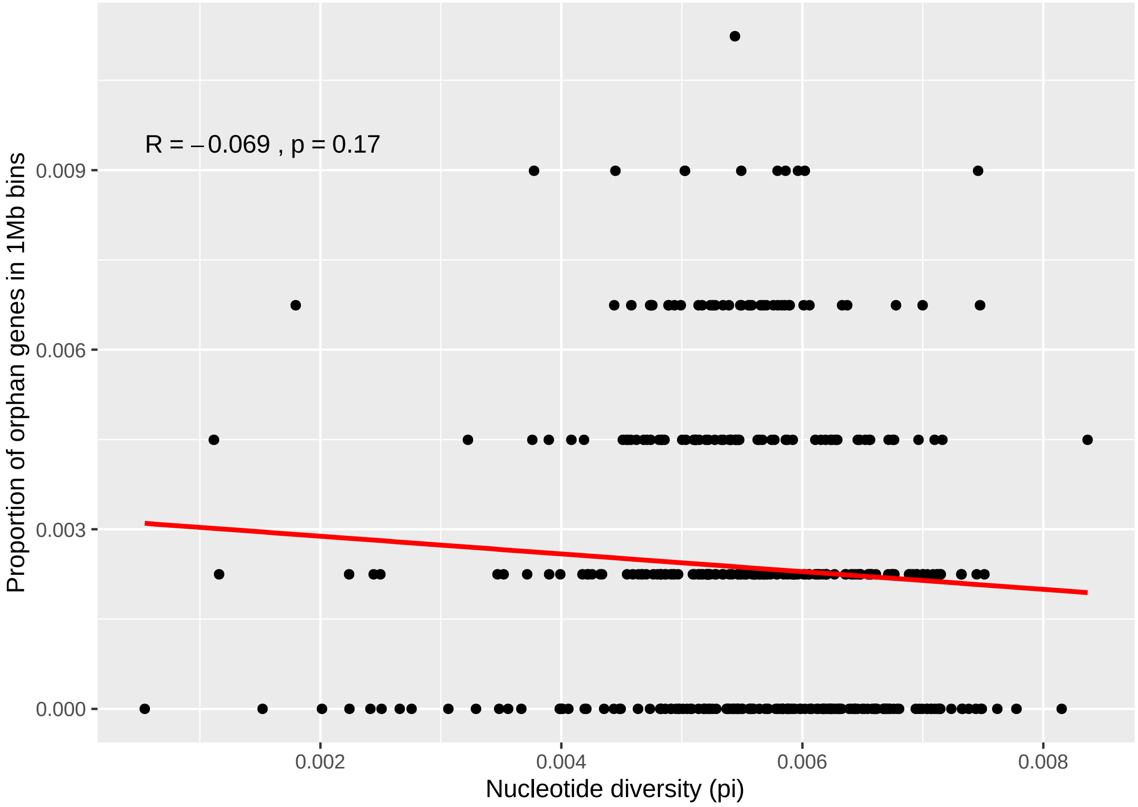

Figure II-S20. Correlation of the proportion orphan genes in 1 Mb bins with nucleotide diversity.

CHAPTER III
A comparative analyses of organelle DNA transfer to nuclear genomes in *Populus trichocarpa* and *Populus deltoides*

A version of this chapter is in preparation for publication by:

Timothy B. Yates, Kai Feng, Paul E. Abraham, Vasanth Singan, Sara S. Jawdy, Kerrie Barry, Anna Lipzen, Chongle Pan, Jeremy Schmutz, Jin-Gui Chen, Gerald A. Tuskan, and Wellington Muchero

Author contributions: T.B.Y. and W.M. designed this study. T.B.Y conducted data analysis and wrote the manuscript. V.S., K.B., and A.L. analyzed and curated the RNA-Seq data. S.J.S. and L.G. generated data. K.F. performed performed data analysis. C.P. contributed to data analysis. P.E.A. generated and analyzed proteomic data. T.B.Y, P.E.A., J.S., J.-G.C, G.A.T., and W.M. revised the manuscript.

Abstract

Organelle-derived features which have integrated into the nuclear genome from the chloroplast and mitochondria are known as NUPTs (“new plastid”) and NUMTs (“new mito”), respectively. Assembled organelle genomes allow for phylogenomic analyses as well as identification of NUPTs and NUMTs in *Populus trichocarpa* and *Populus deltoides* to facilitate comparisons at the species level. We identified 1,076 (0.62 Mb) and 1,409 (0.41 Mb) NUPTs and NUMTs in *P. trichocarpa* while the *P. deltoides* analyses revealed significantly more NUPT and NUMT features with 1,201 (0.99 Mb) and 1,448 (0.76 Mb), respectively. Of these 586 (0.35 Mb) and 667 (0.74 Mb) NUPTs and 732 (0.2 Mb) and 802 (0.26 Mb) NUMTs were shared between *P. trichocarpa* and *P. deltoides*, respectively. Based on transcriptome analyses of protein-coding NUPTs and NUMTs, we provide expression evidence based on uniquely mapping transcripts for 64% in *P. trichocarpa*. Additionally, 28% of protein-coding NUPTs in *P. trichocarpa* had unique peptide evidence based on a proteomics dataset from 6 genotypes. In general, expression levels were lower compared to non-organelle derived features. Aging of these features based on AT transversion-based genetic distance showed that features that were shared between *P. trichocarpa* and *P. deltoides* had higher AT transversion rates and had lower methylation levels in their gene body and promoter regions. Lastly, we characterized the whole chloroplast integration event on chromosome 13 in *P. trichocarpa* and *P. deltoides* and revealed marked expansion in *P. deltoides* resulting in a 370 Kb region with 8 NUPTs compared to *P. trichocarpa*’s 165 Kb and 4 NUPTs.

Introduction

Organelle DNA transfer to the nuclear genome is a ubiquitous process in eukaryotes and has played a substantial role in genome evolution (Timmis et al., 2004). The evolutionary impacts of organelle DNA transfer are extensive. Broadly, endosymbiotic transfer allowed for the establishment of organelle genomes. Mitochondria and chloroplasts were once free-living prokaryotes, their closest relatives being α-proteobacteria and cyanobacteria, respectively (Deusch et al., 2008; Ku et al., 2015). Comparison of organelle genomes with their progenitors has revealed extensive gene loss. Typical gene numbers in plant mitochondria range from 3 to 67 and 15 to 209 in chloroplasts. Comparatively, the α-proteobacterium *Mesorhizobium loti* genome (7 Mb) encodes 6,700 proteins and the cyanobacteria Nostoc sp. PCC 7,120 genome is 6.4 Mb in size and encodes 5,400 proteins (Keeling & Palmer, 2008; Timmis et al., 2004). This drastic reduction in genetic content in chloroplasts and mitochondria is a product of redundancy as entire organelle genomes, genes, and biochemical pathways have transferred to the nuclear genome (Kleine et al., 2009). A large proportion of these nuclear mitochondrial DNAs (NUMTs) and nuclear plastid DNAs (NUPTs) are noncoding sequences (Kleine et al., 2009). However, some are functional and are retargeted to their respective organelle, while others acquire novel functions following horizontal transfer and are involved in extraorganellar functions from all functional categories (Deusch et al., 2008). The general mechanism of organelle DNA transfer to the nuclear genome involves organelle DNA escape, uptake by nuclear import machinery, which is then followed by double-strand break repair through non-homologous end joining (NHEJ) (Hazkani-Covo & Covo, 2008; Thorsness & Fox, 1990).

NUPTs and NUMTs can transfer to the nuclear genome as entire protein-coding genes and can alter existing nuclear gene structure. Although the majority of organelle DNA transfers to the nuclear genome are likely to be non-functional as the transfer must acquire the appropriate regulatory sequences, there is evidence that organelle DNA transfer has contributed functional genes which retained their original function as well as new genes with diverged function (Sheppard & Timmis, 2009). For example, in *Arabidopsis thaliana*, of the 3,500 to 4,500 genes of cyanobacterial origin, approximately half do not localize to the chloroplast (Martin et al., 2002). Additionally, organelle transfers may be responsible for genetic novelty in the form of new exons in existing genes and have also been associated with diseases in Humans (Noutsos et al., 2007; Turner et al., 2003). NUPTs and NUMTs have been shown to preferentially insert into open chromatin and can be transcribed by a chloroplast promoter, albeit at lower levels compared to nuclear genes (Wang et al., 2014; Wang & Timmis, 2013).

Previously, Huang et al., (2017) identified NUPTs in *P. trichocarpa* and identified the transfer of the entire chloroplast genome to chromosome 13. However, due to the lack of a *P. trichocarpa* mitochondria genome assembly and organelle genome assemblies in *P. deltoides* these organelle derived features could neither be identified nor compared between these two closely related *Populus* species. Comparative analysis of NUPT and NUMT presence and absence variation (PAV) at the species level has only been explored in humans as well as domesticated and wild rice (Dayama et al., 2014; Lang et al., 2012; Ma et al., 2020; Wang & Timmis, 2013). It was determined in both rice and humans that the presence and absence of NUPTs and NUMTs varied between populations. To address this, we compared presence and absence polymorphism of NUPTs and NUMTs in *P. trichocarpa* and *P. deltoides* which have species ranges along the west coast and in eastern and central parts of the United States, respectively. Additionally, we also explored DNA methylation and AT-transversion levels of NUPTs and NUMTs in the context of PAV.

To that end, we generated new organelle assemblies for *P. trichocarpa* and *P. deltoides* and identified organelle derived DNA transfers to the nuclear genome. Additionally, we examined PAV for NUPTs and NUMTs between *P. trichocarpa* and *P. deltoides* and further characterized methylation and genetic distances between features which were present or absent. We also explored proteomic data in *P. trichocarpa* and transcriptome data across multiple tissues and experiments in *P. trichocarpa* and *P. deltoides* to assess expression evidence and variation of putative functional transfers from the organelle genome. Lastly, we investigated the whole chloroplast integration event which is shared in *P. trichocarpa* and *P. deltoides* from comparative genomics and epigenomics perspectives.

Results

Assembly and annotation of the *P. trichocarpa* mitochondrial genome and the *P. deltoides* chloroplast and mitochondrial genomes

The assembly of the mitochondrial and chloroplast genomes was performed with Unicyler. The *P. deltoides* WV94 chloroplast assembled into a single circular contig of 156,967 bp which was slightly smaller than the publicly available *P. trichocarpa* chloroplast genome which is 157,033 bp (Figure III-1A). Nucleotide identity between the *P. deltoides* WV94 and the publicly available *P. trichocarpa* chloroplast genomes was 99.2%. Protein-coding gene numbers totaled 97 and 100 in the *P. deltoides* WV94 and publicly available *P. trichocarpa* chloroplast genomes, respectively. Additionally, the numbers of RNA genes were identical as both genomes had 8 rRNA and 37 tRNA. We explored the phylogenetic relationship of the *P. deltoides* WV94 chloroplast genome to 38 other *Populus* species (Table III-S1) based on the alignment of single copy orthologs (Figure III-1B). Our phylogenomic analysis was consistent with previous analyses (Y. Huang et al., 2017; Wang et al., 2015). We observed clades specific to the new world (*P. trichocarpa, P. fremontii, P. balsamifera,* and *P. deltoides*) and old world (*P. tremula, P. alba, P. euphratica,* and *P. davidiana*) *Populus* species.

The *P. trichocarpa* and *P. deltoides* mitochondria genomes assembled into three circular contigs. However, in both assemblies, the full 'master circle' which contains the complete mitochondrial sequence content and is generated through intermolecular recombination of subgenomic circles could not be recovered. It is clear through whole genome self-alignment that repeat regions are shared between the three subgenomic circles in both *P. trichocarpa* and *P. deltoides* (Figure III-1C, III-1D, III-S1, III-S2) which further supports the 'master circle' model. However, if there was a 'master circle', it should have been recovered as a result of the hybrid assembly method which used long and short reads. Multipartite mitochondria structures have been documented previously. Notably, the genus *Silene* has very large mitogenome sizes made up of large numbers of subgenomic circles, the largest being *S. conica* with a 11.3 Mb mitogenome comprised of 128 circular subgenomes (Sloan et al., 2012). The total size of the assemblies for *P. trichocarpa* and *P. deltoides* was 804,157 bp and 802,312 bp, respectively. Nucleotide identity between the two mitochondrial genomes was 99.84%. Gene content was highly similar between the two assemblies, with both having 3 rRNA and 21 tRNA genes. *P. deltoides* had 51 protein-coding genes while *P. trichocarpa* had 52. This difference in protein-coding gene number occurred at a locus on subgenome 2 for a hypothetical protein-coding gene. Although gene numbers are very similar, some variation was present in the presence and absence of hypothetical protein-coding genes. On subgenomes 2 in *P. trichocarpa* and *P. deltoides* there were two hypothetical protein-coding genes which were absent at the expected position in each respective species. Similarly, on subgenome 3, there was one hypothetical protein-coding gene absent in *P. deltoides* which is present in *P. trichocarpa.* The phylogeny inferred from single copy orthologs of mitochondrial protein-coding genes agreed with the analysis performed with the whole chloroplast genome sequences (Figure III-1E).

Organelle derived transfer to the nuclear genome in *P. trichocarpa* and *P. deltoides*

We used BLAST to search for NUPTs and NUMTs in the *P. trichocarpa* and *P. deltoides* nuclear genomes. In total, 1,076 (0.62 Mb) and 1,201 (0.99 Mb) NUPTs were identified in *P. trichocarpa* and *P. deltoides*, respectively (Table III-S2). For NUMTs, 1,409 (0.41 Mb) and 1,448 (0.76 Mb) were identified in *P. trichocarpa* and *P. deltoides*, respectively. Average size of NUPTs and NUMTs were considerably lower in *P. trichocarpa* compared to *P. deltoides* and the maximum length of NUPTs and NUMTs was much larger in *P. deltoides* (Figure III-2A). The largest NUPT in *P. deltoides* was 91 Kb versus 40 Kb in *P. trichocarpa* (Figure III-2A). Furthermore, the same trend was apparent with NUMTs where *P. deltoides* longest NUMT was 41 Kb compared to 4.4 Kb in *P. trichocarpa*.

Three large mitochondrial derived features in *P. deltoides* were removed as NUMTs and considered possible sources of misassembly (Table III-S7). One of these features on Chr02 was 336 kb in size, was located at Chr02:20,020,773-20,357,166 which is in the centromeric region (Weighill et al., 2019). This large feature aligned entirely with the *P. deltoides* mitogenome subgenome 1 with 99.82 % identity. Additionally, methylation frequency in this region was considerably lower when compared to its flanking sequence which is unexpected as NUPTs and NUMTs are rapidly methylated following integration suggesting that this interval may be a product of misassembly (Figure III-S3) (C. Y. Huang et al., 2005). The other two features located on scaffold_22 are derived from subgenome 2. Together, these two features cover 99.6% of subgenome 2 and average 99.9% identity. Scaffold_22 also has very low levels of methylation, further supporting that it is a mitochondrial subgenome which was not removed from the nuclear genome assembly (Figure III-S4).

Putative protein-coding NUPT and NUMT transfer to the nuclear genome was also identified (Table III-S4). We observed larger numbers of protein-coding NUPTs compared to protein-coding NUMTs in both *P. trichocarpa* and *P. deltoides*. *P. deltoides* had larger numbers of protein-coding NUPTs (353) and NUMTs (18) compared to *P. trichocarpa*'s protein-coding

NUPTs (200) and NUMTs (10). Functional enrichment of the putative protein-coding NUPTs and NUMTs was also performed (Figure III-2B). Photosynthesis and translation were two most significantly enriched GO terms in the biological process ontology for protein-coding NUPTs in both *P. trichocarpa* and *P. deltoides*. For protein-coding NUMTs in *P. deltoides* the most significant enrichments in the biological process ontology were for ATP synthesis coupled proton transport and cytochrome complex assembly. Due to the small number of protein-coding NUMTs in *P. trichocarpa* there was no significant functional enrichment.

Presence and absence of NUPTs and NUMTs between *P. deltoides* and *P. trichocarpa*

Differences in organelle-derived content between the nuclear genomes of *P. trichocarpa* and *P. deltoides* was significant. Some chromosomes were found to have considerable differences in transferred organellar DNA (Figure III-3A). For NUPTs, Chr11 and Chr13 in *P. deltoides* had 22kb and 341kb more NUPT content, respectively, compared to their sister chromosomes in *P. trichocarpa*. Furthermore, NUMTs in *P. deltoides* on Chr18 had 65kb more NUMT content compared to the same chromosomes in *P. trichocarpa.*

We explored differences in organelle DNA transfer between *P. trichocarpa* and *P. deltoides* through a combination of BLAST and syntenic analysis to infer presence and absence of NUPTs and NUMTs (Figure III-3B). For NUPTs, 586 (349 kb) features in *P. trichocarpa* had sequence and syntenic correspondence with 667 features in *P. deltoides* (743 kb), while 490 (274 kb) and 534 (246 kb) NUPTs were uniquely found in *P. trichocarpa* and in *P. deltoides,* respectively. For NUMTs, 732 (202 kb) *P. trichocarpa* features corresponded with 802 (261 kb) features in *P. deltoides* while 677 (205 kb) and 644 (209 kb) were uniquely found in *P. trichocarpa* in *P. deltoides,* respectively.

Methylation levels were explored in NUPTs and NUMTs in both shared and species-unique categories (Figure III-S5). Except for *P. trichocarpa* NUPTs in the CG methylation context, all NUPTs and NUMTs that were classified as shared had lower methylation frequencies compared to those determined to be unique to one species. Lower methylation levels of shared features can be indicative of age, as older NUPTs and NUMTs have been shown to have lower methylation levels compared to recently integrated features.

Additionally, AT transversion-based genetic distance of each NUPT and NUMT to its organelle sequence was also examined in the lens of presence and absence variation (Figure III-S6). Shared NUPT and NUMT features, had a higher AT transversion-based genetic distance compared to features found uniquely in one species. This finding corroborates the methylation analysis and further supports the effectiveness of the presence and absence variation analysis as a reliable way to age these features. Additional work will be needed to address why *P. trichocarpa* NUPTs in present and absent classes had similar methylation levels in the CG methylation context.

Expression and proteomic evidence for putative protein-coding NUPTs and NUPTs

Expression evidence for protein-coding NUPTs and NUMTs was first assessed with the Joint Genome Institute (JGI) Plant Gene Atlas database which included 13 and 9 tissues in *P. trichocarpa* and *P. deltoides,* respectively (Table III-S5). For *P. trichocarpa,* 4/10 (40%) putative protein-coding NUMTs and 40/200 (20%) NUPTs had expression evidence based on uniquely mapping transcripts. In *P. deltoides,* 5/18 (20%) putative protein-coding NUMTs and 30/353 (8.4%) had evidence of expression based on uniquely mapping transcripts. Although *P. deltoides* had lower percentages of protein-coding NUPTs and NUMTs compared to *P. trichocarpa,* this could be due to the larger numbers of total features which may be duplicated and therefore lack expression evidence due to multimapping reads.

To further assess expression evidence of putative protein-coding NUPTs and NUMTs in *P. trichocarpa* at the population level we utilized expanded expression datasets (Table III-S5). These transcriptome datasets included leaf, root, and developing xylem tissue which was sampled from hundreds of individual *P. trichocarpa* genotypes well as a QTL pedigree which included 438 individuals. When the expanded datasets were included, 127/200 putative protein-coding NUPTs (63%) had expression and of the 10 putative protein-coding NUMTs in *P. trichocarpa*, all but two (80%) had expression evidence. In leaf, root, and xylem tissues in *P. trichocarpa* where RNA-Seq data was gathered across large numbers of individual genotypes, protein-coding NUPTs had the highest mean expression levels in leaf tissue compared to root and xylem tissues (Figure III-2C). Additionally, gene expression variance in protein-coding NUPTs and NUMTs was highest across individuals in leaf and xylem tissue, respectively. Putative protein-coding NUPTs also had proteomic evidence based on the analysis of unique peptides from mass spectrometry data generated from leaf tissue of six *P. trichocarpa* genotypes. Specifically, 57/200 (28%) protein-coding NUPTs had proteomic evidence and none of the protein-coding NUMTs had proteomic evidence.

Additionally, the putative protein-coding NUPTs and NUMTs were not distributed evenly across chromosomes. In *P. trichocarpa*, putative protein-coding NUPTs were overrepresented (Fisher's exact test, $p < 0.05$) on Chr13 and Chr19 while in *P. deltoides* they are overrepresented on Chr11, Chr13, and Chr16. For putative protein-coding NUMTs, no chromosomes were overrepresented in either *P. trichocarpa* or *P. deltoides.*

Although putative protein-coding NUPTs and NUMTs are generally thought to be non-functional we identified 47 putative protein-coding NUPTs that have expression evidence and proteomic evidence based on uniquely mapping transcripts and peptides, respectively (Table III-S4). Future molecular validation studies will be necessary to assess their function.

The *P. trichocarpa* and *P. deltoides* whole chloroplast integration events

Previously, Huang et al., (2017), identified a whole chloroplast integration event with an alignment length of 165kb near the end of Chr13 in *P. trichocarpa* (Chr13:14,730,412-14,896,306) (Figure III-4A). This event is also shared in *P. deltoides* (Chr13:15,455,935-15,826,865) as the *P. deltoides* chloroplast genome aligned fully to the nuclear genome, albeit with a larger total alignment length of 370 kb. Syntenic alignments in this region totaled 157 kb or 95% of the query length and 160 kb or 43% of the target length in *P. trichocarpa* and *P. deltoides,* respectively. Regions which could not be aligned between *P. trichocarpa* and *P. deltoides* were significantly larger in *P. deltoides* with a total size of 73 kb compared to 149 bp in *P. trichocarpa.*

In *P. deltoides,* this region accounted for 481/497kb (96%) of the NUPT content on chromosome 13 and 48% of all NUPT content genomewide. Similarly, this region in *P. trichocarpa* contained 137/156kb (87%) of the NUPT content on Chr13, but overall represented a lower proportion of total NUPT content at 21% genomewide. Based on the presence and absence analysis described above, NUPTs that were specific to *P. deltoides* and *P. trichocarpa* were relatively similar in size and number, but it was evident *P. deltoides* has a significantly larger total size of shared NUPTs. Two hypotheses arise from this finding, either a one-to-many or many-to-many relationship exists from *P. trichocarpa* to *P. deltoides* where a large amount of duplication has occurred on Chr13 in *P. deltoides* or there is a misassembly in the genome of *P. deltoides*. To address this, we further examined the Chr13 whole chloroplast integration event region with a focus on the syntenic relationships of putative protein-coding genes within an expanded region that includes the whole chloroplast integration events in *P. trichocarpa* and *P. deltoides* (Figure III-4B). The expanded region in *P. deltoides* is 564kb (Chr13:15,308,379-15,872,913) and the corresponding region (Chr13:14,660,555-14,907,618) in *P. trichocarpa* is 247kb is less than half the size. Putative protein-coding NUPTs which were collinear in *P. trichocarpa* and *P. deltoides* were used as borders to define the expanded in this region. Of the 200 putative protein-coding NUPTs in *P. trichocarpa,* 43 or 20% were located in this region in syntenic blocks. Similarly, of the 353 putative protein-coding NUPTs in *P. deltoides,* 121 or 34% are found in this interval. We identified several large expansions of protein-coding NUPT gene content from *P. trichocarpa* to *P. deltoides* (Table III-S6). For example, the largest one-to-many collinear relationships from *P. trichocarpa* to *P. deltoides* in this interval were present for three genes (Potri.013G138612, Potri.013G138300, Potri.013G136730) which have six homologous copies in *P. deltoides*. Although methylation was present in this region, it was lower compared to flanking regions which may suggest misassembly (Figure III-S7). To further characterize this region we aligned the *P. trichocarpa* chloroplast genome to two additional *P. trichocarpa* assemblies, *P. trichocarpa* v4.0 and *P. trichocarpa* Stettler v1.0. The chloroplast genome aligned to similar regions in Chr13 in *P. trichocarpa* v4.0 and *P. trichocarpa* Stettler v1.0 as it did in *P. trichocarpa* v3.0, however, both alignments were shorter with lengths of 61 kb and 76 kb, respectively. Although conserved synteny in this region between *P. trichocarpa* and *P. deltoides* may indicate a legitimate whole chloroplast integration event that was shared in the ancestor of *P. trichocarpa* and *P. deltoides* this is not supported by methylation data and only partially supported by more contiguous *P. trichocarpa* v4.0 and *P. trichocarpa* Stettler v1.0 assemblies. Future genome assemblies of *P. trichocarpa* and *P. deltoides* should evaluate this region for whole or partial chloroplast integration events.

Discussion and Conclusion

In this study we provided new organelle genome resources, identified, compared, and characterized organelle-derived putative coding and non-coding NUPTs and NUMTs in the *P. trichocarpa* and *P. deltoides* nuclear genomes. Phylogenomic analyses agreed with previous work and correctly assigned new and old world *Populus* species to correct clades (Y. Huang et al., 2017; Wang et al., 2015). However, based on the low bootstrap values observed in the mitogenome derived phylogeny, a more comprehensive sampling of species from *Populus* and *Salix* genera would be beneficial. The widespread use of long-read sequencing should enable future studies exploring the phylogenomics of the Salicaceae. Genome assembly of plant mitogenomes is challenging compared to the assembly of chloroplast genomes. Specifically, the high repeat content, rearrangement, and heterogeneity between closely related mitogenomes makes assembly difficult (Sloan, 2013). Although we did not recover a ‘master circle’ which is the format of other publicly available *Populus* mitogenomes, we did capture the expected gene content for plant mitogenomes, and gene content was highly similar between the mitogenome assemblies generated here (Choi et al., 2017; Kersten et al., 2016).

We identified NUPTs and NUMTs in *P. trichocarpa* and *P. deltoides* with BLAST which has been commonly used in previous studies. To generate a more conservative group of BLAST hits, additional methodological improvements were performed. These included the use of doubled organelle genomes, only retaining features which are identical or fully contained by a larger feature, and merging features which were in close proximity to each other (Hazkani-Covo & Martin, 2017). *P. trichocarpa* NUPTs have been identified using BLAST and with LASTZ in previous studies and the cumulative size is comparable to what were reported in this study (Y. Huang et al., 2017; Zhang et al., 2020). We also investigated putative protein-coding gene transfer to the nuclear genome and characterized these features with transcriptomic and proteomic data. Based on unique peptides and uniquely mapping transcripts, we provided evidence of features that may be functional in the nuclear genome. Through the use of population-scale expression datasets from the *P. trichocarpa* GWAS mapping panel, we provided expression evidence for 63% and 80% of putative protein-coding NUPTs and NUMTs, respectively. Unsurprisingly, as a result of their recent origin, these features do have lower expression overall and expression variation compared to genes which do not originate from organelle transfer events. Although the functionality of these features may be dubious we identified putative protein-coding NUPTs in *P. trichocarpa* which have expression and proteomic evidence and may be good candidates for functional validation.

P. deltoides had significantly higher numbers and total cumulative size of NUPTs and NUMTs compared to *P. trichocarpa*. To characterize this variation further, we used BLAST and conserved synteny between *P. trichocarpa* and *P. deltoides* to classify NUPTs and NUMTs as either present or absent across genomes of the two species. We identified differences such as Chr18, which had a large amount of NUMT content specific to *P. deltoides* and shared regions like the end of Chr13 which harbors the whole chloroplast integration event. This analysis highlights the high variation of organellar transfer rates to the nuclear genome even between very closely related species. Additionally, future sequencing projects where large numbers of related individuals or species are assembled may find the 'absent' class of NUPTs and NUMTs useful to assess possible misassembly in the nuclear genome. To further support our presence and absence variation pipeline we leveraged AT transversion-based genetic distance and methylation frequencies of present and absent NUPTs and NUMTs and were able to confirm that present and absent features fall into the appropriate age classes, which aligns with previous work which describes methylation decay over evolutionary time (Yoshida, et al., 2019). AT-transversion based genetic distance and methylation levels complement each other as metrics for assigning age to NUPTs and NUMTs. Based on AT-transversion based genetic distance we observed NUPTs in *P. trichocarpa* and *P. deltoides* are more divergent from their progenitor organelle sequence compared to NUMTS. This suggests that NUPTs may experience lower rates of mutation compared to NUMTs in the nuclear genome.

NUPTs and NUMTs primarily consist of short insertion events with a smaller number of longer more recent integration events. The whole chloroplast integration event was shared between *P. trichocarpa* and *P. deltoides*, indicating that this occurred prior to speciation. Notably, the fates of the two insertion events have differed considerably after integration. This region in *P. deltoides* has undergone considerable duplication and fragmentation compared to *P. trichocarpa.* Methylation frequency in this region also appears to be lower compared to flanking regions in both *P. trichocarpa* and *P. deltoides*. Furthermore, more contiguous *P. trichocarpa* assemblies only contain portions of the whole chloroplast genome at this locus, suggesting it could be an assembly artifact.

Populus as a model tree species is an important biological resource and spans large environmental variation. Previously, comparative genomics has shown that NUPTs and NUMTs vary considerably in populations and contribute significantly to differences in genome evolution even among closely related species (Dayama et al., 2014; Ma et al., 2020). Here, we observe similar trends, where NUPT and NUMT were variable across two closely related species.

However, additional high quality *Populus* genomes or long read sequencing will be necessary to assess segregation of these features accurately along the natural species ranges of *P. trichocarpa* and *P. deltoides.* Additionally, the methodology described here for comparing NUPT and NUMT content from comparative genomics or genome quality perspective may benefit future studies which seek to compare closely related genomes. As a result of the population-scale multi-omics data available to *P. trichocarpa* we were able to identify a cohort of putative protein-coding *P. trichcocarpa* NUPTs and NUMTs which had expression and proteomic evidence and exhibited expression variation in the GWAS mapping panel. Future work will focus on empirically validating their function.

Materials and Methods

Organelle genome assembly and annotation: PacBio continuous long read (CLR) and illumina whole genome sequencing (WGS) data was generated for both *P. trichocarpa* Nisqually-1 and *P. deltoides* WV94. These data are available at PRJNA333102 and PRJNA791651. To enrich for organelle derived reads from PacBio data, long reads were mapped with nucmer v4.0.0rc1 with an alignment length threshold of 1kb and the 'maxmatch' flag against doubled organelle genomes (Table III-S1) (Marçais et al., 2018). Doubled genomes were used to ensure reads mapped to the entire circular molecule. To enrich for organelle derived reads in illumina data, short reads were mapped with bowtie2 v2.4.2 against the same doubled organelle genomes (Langmead & Salzberg, 2012). Prior to assembly, both the long and short organelle enriched reads were randomly down sampled to 500X coverage. The organelle genomes were then assembled with Unicycler v0.4.8 with default parameters (Wick et al., 2017). The *P. trichocarpa* Nisqually-1 and *P. deltoides* WV94 mitochondria genomes were annotated with Geneious Prime 2021.1.1 by transferring the mitochondria annotations from *P. alba* (NC_041085), *P. davidiana* (KY216145), and *P. tremula* (KT337313) with a sequence similarity threshold of 85%. The *P. deltoides* WV94 chloroplast genome was annotated with GeSeq (Tillich et al., 2017). The previously annotated chloroplast genomes of *P. trichocarpa* Nisqually-1 (EF489041) and *P. deltoides* (MT789695) were used as references for the annotation, and the BLAT search identity thresholds were set at 85% for protein, rRNA, tRNA, and DNA. The mitochondria and chloroplast genomes generated in this study have been deposited at NCBI (OM461348-OM461354), refer to Table III-S1 for all accession numbers.

Phylogenomic analysis of *Populus* chloroplast and mitochondrial genomes: Publicly available organelle genomes were downloaded from NCBI (Table III-S1). Protein sequences from these genomes were used as input to Orthofinder v2.5.4 to identify single copy orthologs (Emms & Kelly, 2019). The concatenated alignment generated by Orthofinder v2.5.4 with MAFFT v7.487 was used for tree construction (Katoh & Standley, 2013). The Maximum Likelihood tree was constructed with IQ-Tree v2.1.2 with 1000 bootstraps and visualized with Geneious Prime 2021.1.1 (Kearse et al., 2012; Minh et al., 2020).

Identification of organelle derived transfers to the nuclear genome: Regions of organelle DNA that were integrated into the nuclear genome were identified with BLASTN 2.11.0+. Parameters used are a word size of 11, e-value less than 1E-5, percent identity greater than or equal to 75%, and minimum alignment length of 100 bp. The organelle genome was doubled to account for transfer from a circular molecule. NUPTs and NUMTs were then merged with bedtools merge v2.29.2 with a distance of 400 bp to account for overlapping and closely adjacent features. Some NUPT and NUMT features were identified at identical locations in the nuclear genome, these features were classified as ambiguous and were not included in any downstream analysis (Table III-S3). Similarly, some NUPTs and NUMTs were fully contained in a larger feature, the larger feature was retained while the smaller features were excluded from downstream analysis (Table III-S3). Protein-coding transfer from the organelle to the nuclear genomes was determined with the nucleotide CDS sequences with BLASTN 2.11.0+, a coverage requirement of 75% in either the subject or the query was required as well alignment identity greater than or equal to 80%. Duplicate protein-coding transfer as a result of similarity between mitochondrial and chloroplast genes was removed by retaining the nuclear copy with the highest coverage to the organelle gene. Expression data processed as described in (Yates et al., 2021) with the exception that the organelle genomes and annotations for *P. trichocarpa* and *P. deltoides* were added at the mapping step. Similarly, proteomic data from six *P. trichocarpa* genotypes was processed as described previously, with the exception that organelle proteins were added to the *P. trichocarpa* v3.0 reference database. The thresholds used to determine proteomic evidence included at least one unique peptide in two datasets or two or more unique peptides in a single dataset.

AT transversion-based genetic distance: First, organelle sequences and the corresponding NUPT or NUMT were aligned with MAFFT v7.487 *linsi*. For the calculation of AT transversion-based genetic distance, refer to (Yoshida et al., 2019).

Presence and absence of NUPTs and NUMTs in *P. trichocarpa* and *P. deltoides*: Presence and absence of NUPTs and NUMTs was determined through a combination of BLAST and syntenic searches. First, reciprocal BLASTN v2.11.0+ searches with a word size of 11, e-value less than 1E-5, and at least 75% coverage in the subject or query were required. Next, syntenic intervals were predicted for each NUPT and NUMT with Synder 0.28.0 from a syntenic map derived from pairwise whole genome alignments between *P. trichocarpa* and *P. deltoides* with nucmer v4.0.0rc1 with the 'maxmatch' flag. NUPTs or NUMTs which met the blast requirements and could be identified in predicted syntenic intervals were classified as present while those which failed either step were considered absent. Total content and presence and absence of NUPT and NUMT content was visualized with Circos v0.69 (Krzywinski et al., 2009).

Synteny and structural variation analysis between *P. trichocarpa* and *P. deltoides*: Gene level synteny analyses were performed with MCScanX (Y. Wang et al., 2012), with default parameters, and the resulting collinearity file was analyzed to determine duplication from *P. trichocarpa* to *P. deltoides.* Whole genome level synteny and structural variation was determined through whole genome alignment between *P. trichocarpa* and *P. deltoides* with nucmer v4.0.0rc1 with the 'maxmatch' flag. From these alignments, structural rearrangement and synteny was identified with SyRI, default parameters were used (Goel et al., 2019).

Comparative methylation analysis of NUPTs and NUMTs in *P. trichocarpa* and *P. deltoides:* Fresh young leaf tissue (approximately 100 mg) was collected and ground in liquid nitrogen using the Qiagen TissueLyser II with one 5mm stainless steel bead. DNA extraction was performed using a modified CTAB based protocol (Doyle & Doyle, 1987). Briefly, the organic and aqueous phase were extracted using chloroform:isoamyl alcohol 24:1. After separation, a SPRI bead solution was used to select for reads greater than 1kb (Mayjonade et al., 2016). Following extraction, 1 ug of DNA was used as input to Oxford Nanopore's genomic DNA by ligation sequencing kit (SQK-LSK109) and the subsequent library was sequenced on a R.9.4.1 flow cell. Coverage generated was 86X and 69X in *P. trichocarpa* and *P. deltoides,* respectively. Guppy v5.0.11 was used to basecall the raw fast5 files with the 'dna_r9.4.1_450bps_hac_prom.cfg' config file. Tombo v1.5.1 was used to align the fast5 files to the *P. trichocarpa* v3.0 and *P. deltoides* v2.0 reference genomes which included the complete organelle genomes. DeepSignal plant was then used to call modifications with the 'model.dp2.CNN.arabnrice2-1_120m_R9.4plus_tem.bn13_sn16.both_bilstm.epoch6.ckpt' model and calculate modification frequencies across the genome (Ni et al., 2021). Modification sites with greater than or equal to five mapped reads were retained for all subsequent analysis.

Alignment, synteny, and orthogroup analysis of the whole chloroplast genome insertion event in *P. trichocarpa* and *P. deltoides*: Whole genome alignment was performed as described by (Y. Huang et al., 2017). Briefly, the chloroplast sequence was aligned to the nuclear genome with lastz v1.0.2 with the 'chain' flag. Next, the 'lav' output was converted to 'axt' format with lavToAxt which was then converted to chain format with axtChain. Orthogroups were identified from with Orthofinder v2.5.4 with default parameters, primary protein isoforms were used (Emms & Kelly, 2019). The figure was generated with gggenomes (Hackl et al., 2021).

References

Choi, M. N., Han, M., Lee, H., Park, H. S., Kim, M. Y., Kim, J. S., … Park, E. J. (2017). The complete mitochondrial genome sequence of Populus davidiana Dode. *Mitochondrial DNA Part B: Resources*, *2*(1), 113–114. https://doi.org/10.1080/23802359.2017.1289346

Dayama, G., Emery, S. B., Kidd, J. M., & Mills, R. E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research*, *42*(20), 12640–12649. https://doi.org/10.1093/nar/gku1038

Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K. V., Allen, J. F., … Dagan, T. (2008). Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Molecular Biology and Evolution*, *25*(4), 748–761. https://doi.org/10.1093/molbev/msn022

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *PHYTOCHEMICAL BULLETIN*, (RESEARCH). Retrieved from http://worldveg.tind.io/record/33886

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 1–14. https://doi.org/10.1101/466201

Goel, M., Sun, H., Jiao, W. B., & Schneeberger, K. (2019). SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *BioRxiv*, 1–13. https://doi.org/10.1101/546622

Hackl, T., Duponchel, S., Barenhoff, K., Weinmann, A., & Fischer, M. G. (2021). Virophages and retrotransposons colonize the genomes of a heterotrophic flagellate. *ELife*, *10*. https://doi.org/10.7554/eLife.72674

Hazkani-Covo, E., & Covo, S. (2008). Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genetics*, *4*(10). https://doi.org/10.1371/journal.pgen.1000237

Hazkani-Covo, E., & Martin, W. F. (2017). Quantifying the number of independent organelle DNA insertions in genome evolution and human health. *Genome Biology and Evolution*, *9*(5), 1190–1203. https://doi.org/10.1093/gbe/evx078

Huang, C. Y., Grünheit, N., Ahmadinejad, N., Timmis, J. N., & Martin, W. (2005). Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology*, *138*(3), 1723–1733. https://doi.org/10.1104/pp.105.060327

Huang, Y., Wang, J., Yang, Y., Fan, C., & Chen, J. (2017). Phylogenomic analysis and dynamic evolution of chloroplast genomes in salicaceae. *Frontiers in Plant Science*, *8*(June), 1–13. https://doi.org/10.3389/fpls.2017.01050

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, *9*(8), 605–618. https://doi.org/10.1038/nrg2386

Kersten, B., Rampant, P. F., Mader, M., Le Paslier, M. C., Bounon, R., Berard, A., … Fladung, M. (2016). Genome sequences of Populus tremula chloroplast and mitochondrion:

Implications for holistic poplar breeding. *PLoS ONE*, *11*(1), 1–21. https://doi.org/10.1371/journal.pone.0147209

Kleine, T., Maier, U. G., & Leister, D. (2009). DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. *Annual Review of Plant Biology*, *60*(1), 115–138. https://doi.org/10.1146/annurev.arplant.043008.092119

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. https://doi.org/10.1101/gr.092759.109

Ku, C., Nelson-Sathi, S., Roettger, M., Garg, S., Hazkani-Covo, E., & Martin, W. F. (2015). Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(33), 10139–10146. https://doi.org/10.1073/pnas.1421385112

Lang, M., Sazzini, M., Calabrese, F. M., Simone, D., Boattini, A., Romeo, G., … Gasparre, G. (2012). Polymorphic NumtS trace human population relationships. *Human Genetics*, *131*(5), 757–771. https://doi.org/10.1007/s00439-011-1125-3

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Ma, X., Fan, J., Wu, Y., Zhao, S., Zheng, X., Sun, C., & Tan, L. (2020). Whole-genome de novo assemblies reveal extensive structural variations and dynamic organelle-to-nucleus DNA transfers in African and Asian rice. *Plant Journal*, *104*(3), 596–612. https://doi.org/10.1111/tpj.14946

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), 1–14. https://doi.org/10.1371/journal.pcbi.1005944

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., … Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(19), 12246–12251. https://doi.org/10.1073/pnas.182432999

Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., … Muños, S. (2016). Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *BioTechniques*, *61*(4), 203–205. https://doi.org/10.2144/000114460

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., … Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Ni, P., Huang, N., Nie, F., Zhang, J., Zhang, Z., Wu, B., … Wang, J. (2021). Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nature Communications*, *12*(1), 1–11. https://doi.org/10.1038/s41467-021-26278-9

Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., & Leister, D. (2007). Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. *Trends in Genetics*, *23*(12), 597–601. https://doi.org/10.1016/j.tig.2007.08.016

Sheppard, A. E., & Timmis, J. N. (2009). Instability of plastid DNA in the nuclear genome. *PLoS Genetics*, *5*(1), 1–8. https://doi.org/10.1371/journal.pgen.1000323

Sloan, D. B. (2013). One ring to rule them all? Genome sequencing provides new insights into the “master circle” model of plant mitochondrial DNA structure. *New Phytologist*, *200*(4), 978–985. https://doi.org/10.1111/nph.12395

Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D., & Taylor, D. R. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biology*, *10*(1). https://doi.org/10.1371/journal.pbio.1001241

Thorsness, P., & Fox, T. (1990). Escape of DNA from mitochondria to the nucleus in Saccharomyces cerevisiae. *Nature*, *346*(4), 376–379. https://doi.org/10.1134/s032097251904002x

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11. https://doi.org/10.1093/nar/gkx391

Timmis, J. N., Ayliff, M. A., Huang, C. Y., & Martin, W. (2004). Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, *5*(2), 123–135. https://doi.org/10.1038/nrg1271

Turner, C., Killoran, C., Thomas, N. S. T., Rosenberg, M., Chuzhanova, N. A., Johnston, J., … Biesecker, L. G. (2003). Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Human Genetics*, *112*(3), 303–309. https://doi.org/10.1007/s00439-002-0892-2

Wang, Dong, Qu, Z., Adelson, D. L., Zhu, J. K., & Timmis, J. N. (2014). Transcription of nuclear organellar DNA in a model plant system. *Genome Biology and Evolution*, *6*(6), 1327–1334. https://doi.org/10.1093/gbe/evu111

Wang, Dong, & Timmis, J. N. (2013). Cytoplasmic organelle DNA preferentially inserts into open chromatin. *Genome Biology and Evolution*, *5*(6), 1060–1064. https://doi.org/10.1093/gbe/evt070

Wang, Dongsheng, Wang, Z., Du, S., & Zhang, J. (2015). Phylogeny of section Leuce (Populus, Salicaceae) inferred from 34 chloroplast DNA fragments. *Biochemical Systematics and Ecology*, *63*, 212–217. https://doi.org/10.1016/j.bse.2015.09.020

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., … Paterson, A. H. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7). https://doi.org/10.1093/nar/gkr1293

Weighill, D., Macaya-Sanz, D., DiFazio, S. P., Joubert, W., Shah, M., Schmutz, J., … Jacobson, D. (2019). Wavelet-based genomic signal processing for centromere identification and hypothesis generation. *Frontiers in Genetics*, *10*(MAY). https://doi.org/10.3389/fgene.2019.00487

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6), 1–22. https://doi.org/10.1371/journal.pcbi.1005595

Yates, T. B., Feng, K., Zhang, J., Singan, V., Jawdy, S. S., Ranjan, P., … Muchero, W. (2021). The Ancient Salicoid Genome Duplication Event: A Platform for Reconstruction of De Novo Gene Evolution in Populus trichocarpa . *Genome Biology and Evolution*, *13*(9), 1–14. https://doi.org/10.1093/gbe/evab198

Yoshida, T., Furihata, H. Y., To, T. K., Kakutani, T., & Kawabe, A. (2019). Genome defense against integrated organellar DNA fragments from plastids into plant nuclear genomes through DNA methylation. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-38607-6

Zhang, G. J., Dong, R., Lan, L. N., Li, S. F., Gao, W. J., & Niu, H. X. (2020). Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. *International Journal of Molecular Sciences*, *21*(3). https://doi.org/10.3390/ijms21030707

Appendix

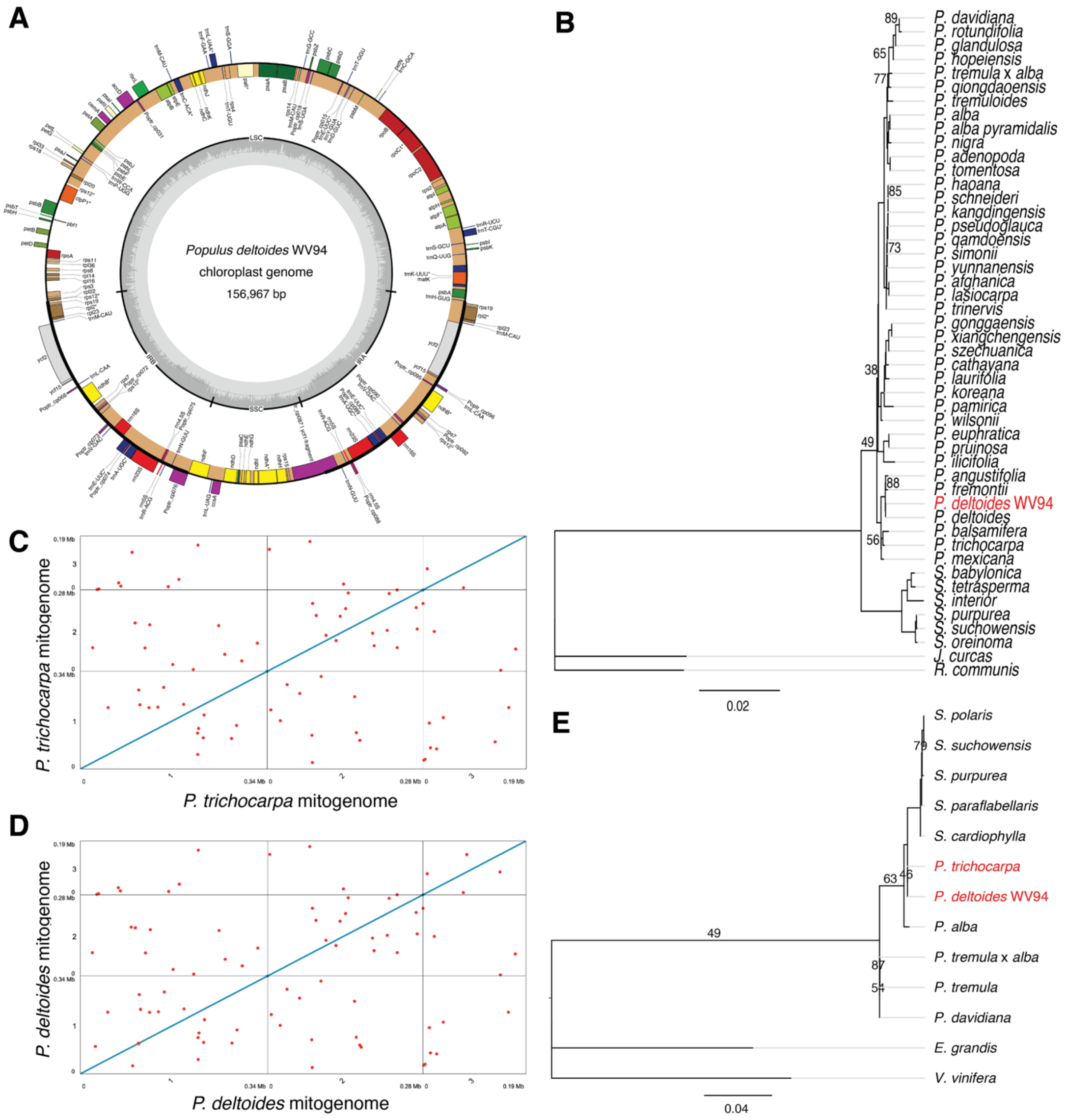

Figure III-1. New organelle assemblies in *P. trichocarpa* and *P. deltoides* facilitate phylogenomic analysis with existing publicly available chloroplast and mitochondrial genomes. (A) The *P. deltoides* WV94 chloroplast genome assembly. (B) Maximum Likelihood tree derived from alignment of single copy orthologs which depict the phylogenomic relationship of the newly assembled *P. deltoides* WV94 genome (highlighted in red) with 38 other publicly available *Populus* chloroplast genomes. Bootstrap support values less than 90% are shown. (C) Whole genome self-alignment of the *P. trichocarpa* Nisqually-1 mitogenome assembly, red points indicate repetitive alignments. (D) Whole genome self-alignment of the *P. deltoides* WV94 mitogenome assembly, red points indicated repetitive alignments. (E) Maximum Likelihood tree derived from alignment of single copy orthologs depicting the phylogenomic relationship of the newly assembled *P. deltoides* WV94 and *P. trichocarpa* Nisqually-1 mitogenomes (highlighted in red) with existing publicly available *Populus* and *Salix* mitogenome assemblies. Bootstrap support values less than 90% are shown.

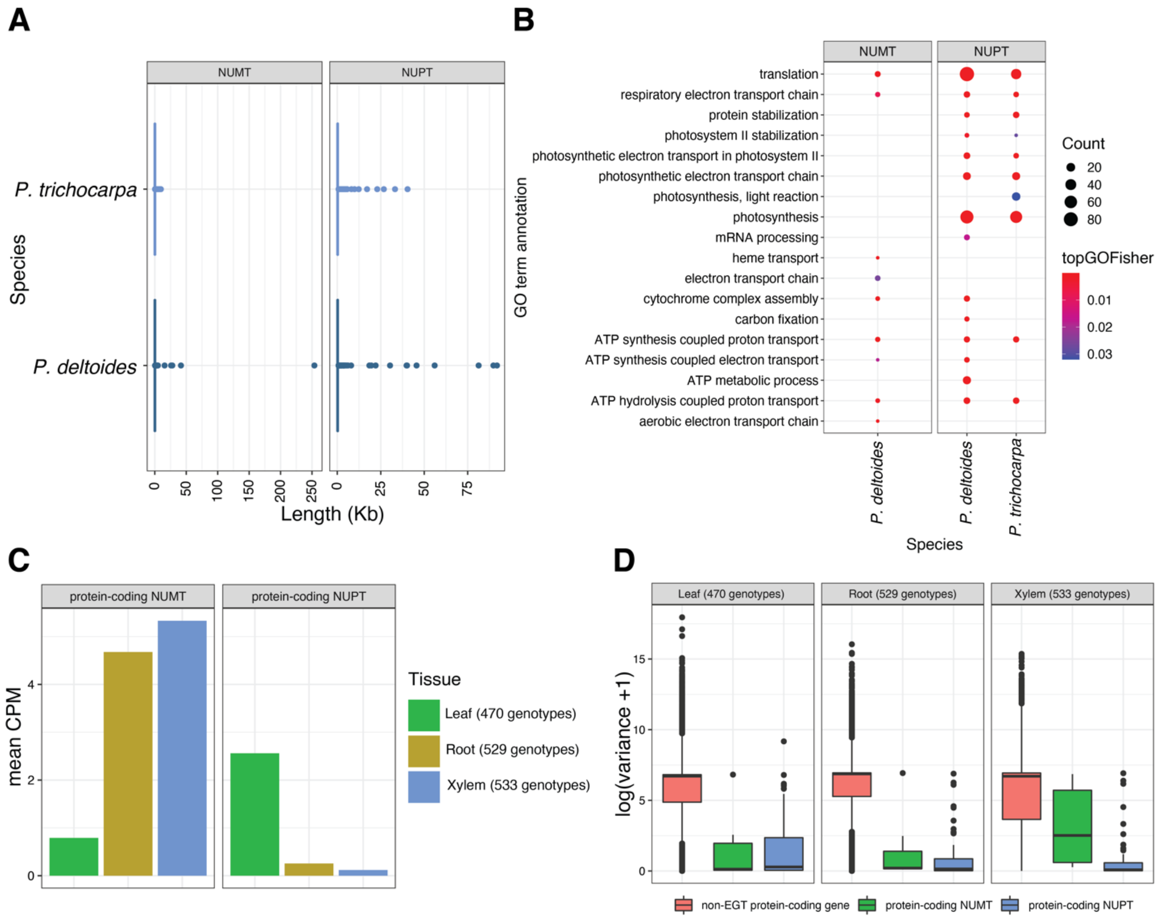

Figure III-2. NUPT and NUMTs and organellar derived protein-coding genes are identified in the *P. trichocarpa* and *P. deltoides* nuclear genomes. (A) Size distributions of NUPTs and NUMTs in *P. trichocarpa* and *P. deltoides.* (B) Functional enrichment with the 'biological process' ontology of protein-coding NUPTs and NUMTs in *P. trichocarpa* and *P. deltoides*. (C) Mean expression of uniquely mapping transcript levels represented as counts per million (CPM). Expression levels are derived from genotypes in the *P. trichocarpa* GWAS mapping panel. The number of genes used in this figure was 8 and 127 for protein-coding NUMTs and NUPTs, respectively. (D) Variance was used to describe variation of genes that are not derived from organellar transfer events (non-EGT protein-coding gene) and protein-coding NUPTs and NUMTs which had expression in the *P. trichocarpa* GWAS mapping panel for leaf, root, and xylem tissues.

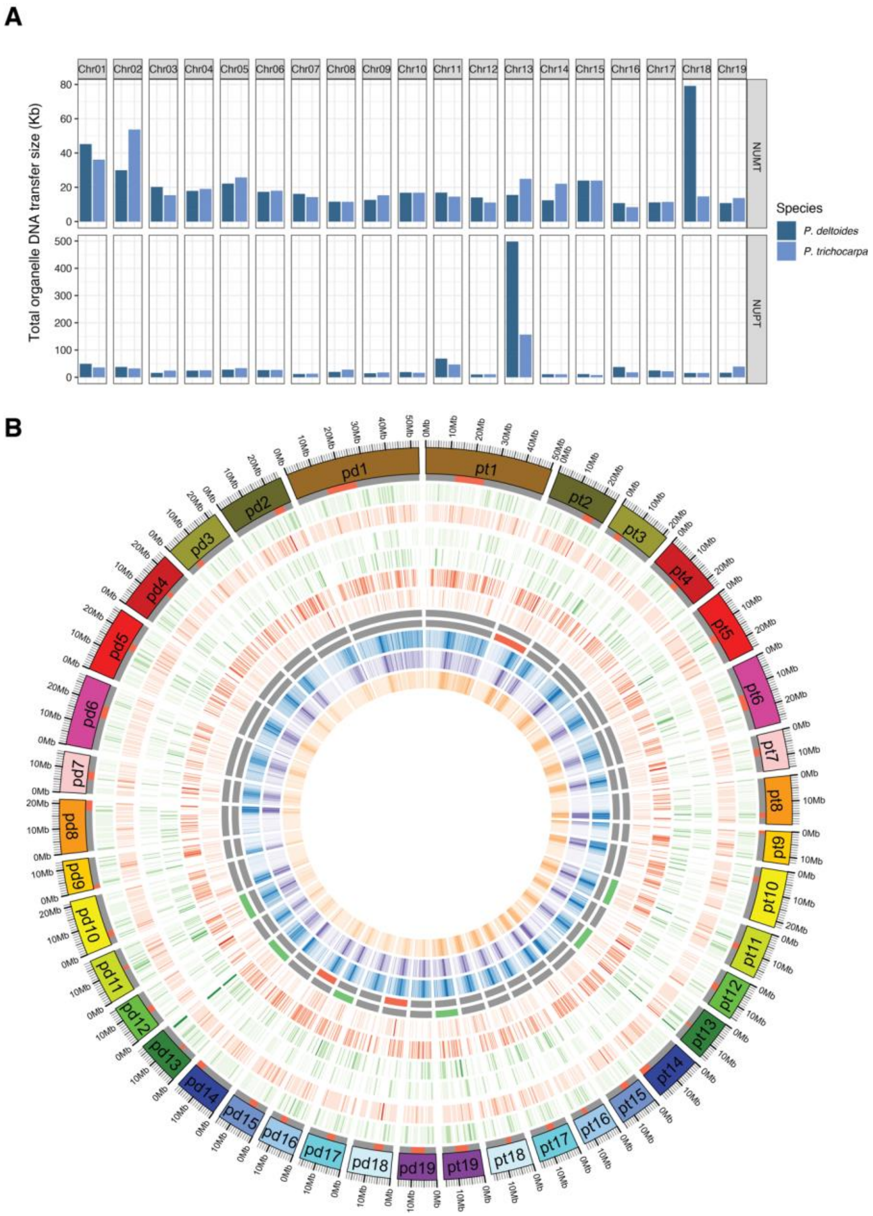

Figure III-3. NUPT and NUMT content varies between *P. trichocarpa* and *P. deltoides.* (A) NUPT and NUMT content by chromosome in *P. trichocarpa* and *P. deltoides* . (B) Circos plot depicting NUPT and NUMT content and presence and absence differences between *P. trichocarpa* and *P. deltoides*. The prefixes of 'pt' and 'pd' in the chromosome names refer to *P. trichocarpa* and *P. deltoides,* respectively. All heatmaps depicting total, present and absent NUPT and NUMT content represent the sum of the size of the NUPT and NUMT features in 500kb windows. The Circos figure from outer to inner depicts: centromere position (red band), total NUPT content, total NUMT content, shared NUPT content , absent NUPT content, shared NUMT content, absent NUMT content, green highlights indicate the chromosome is an outlier (total content is > (upper quartile + 1.5 * IQR) for total NUPT content, and red highlights indicate the chromosome is an outlier for total NUMT content, CG methylation frequency in 100kb windows, CHG methylation frequency in 100 kb windows, and CHH methylation frequency in 100kb windows.

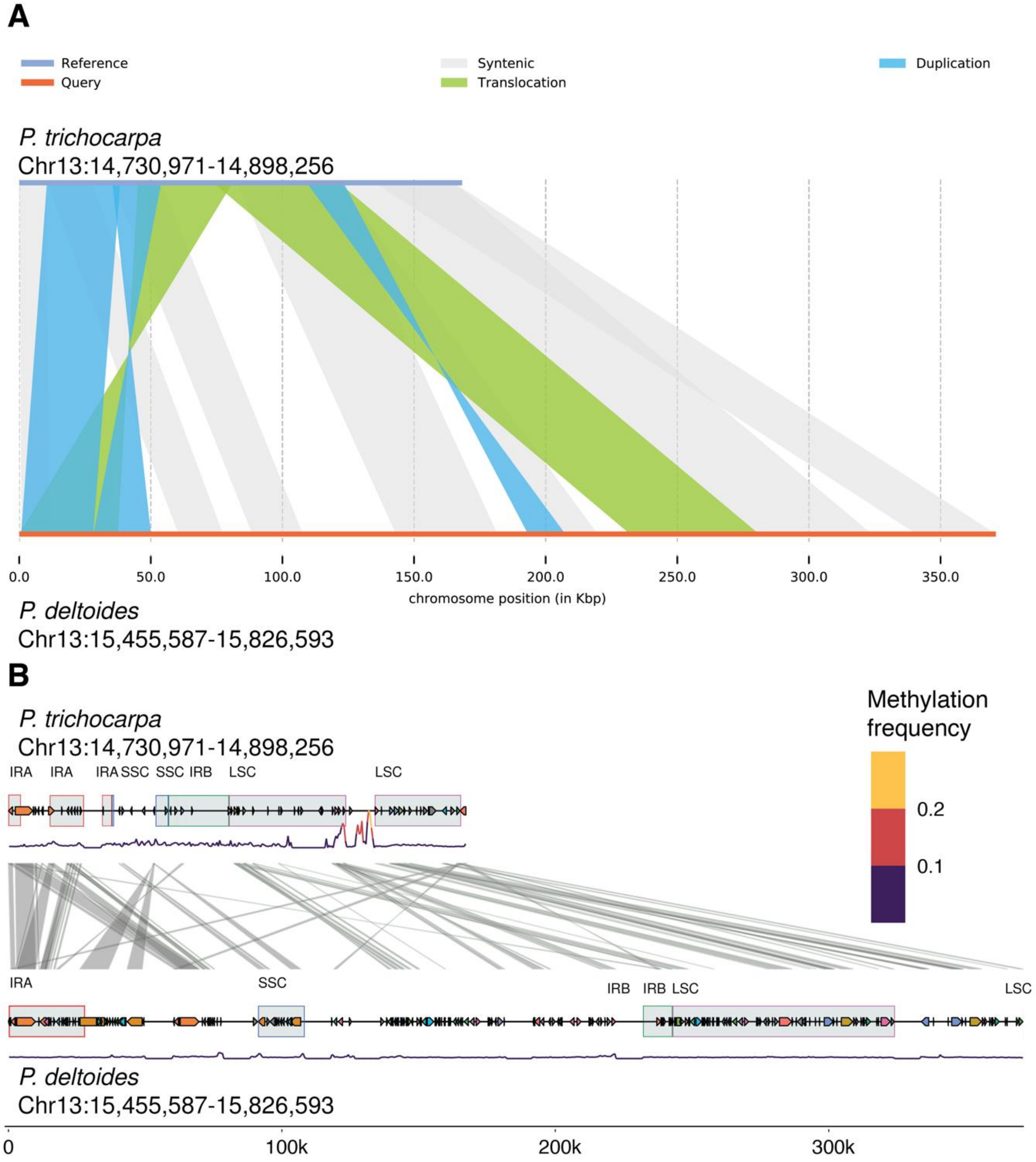

Figure III-4. The *P. trichocarpa* and *P. deltoides* whole chloroplast integration events. (A) Overview of the syntenic structure of the between *P. trichocarpa* and *P. deltoides* whole chloroplast integration events. *P. trichocarpa* was used as the reference sequence while *P. deltoides* was used as the query sequence in the alignment (B) Orthology and methylation frequency in *P. trichocarpa* and *P. deltoides* whole chloroplast integration events. Gray boxes are labeled with the organelle feature these regions are derived from (IR = inverted repeat region, SSC = short single copy region, LSC = large single copy region), links between tracks indicate orthology, Methylation frequency including all methylation contexts was averaged over 500 bp windows.

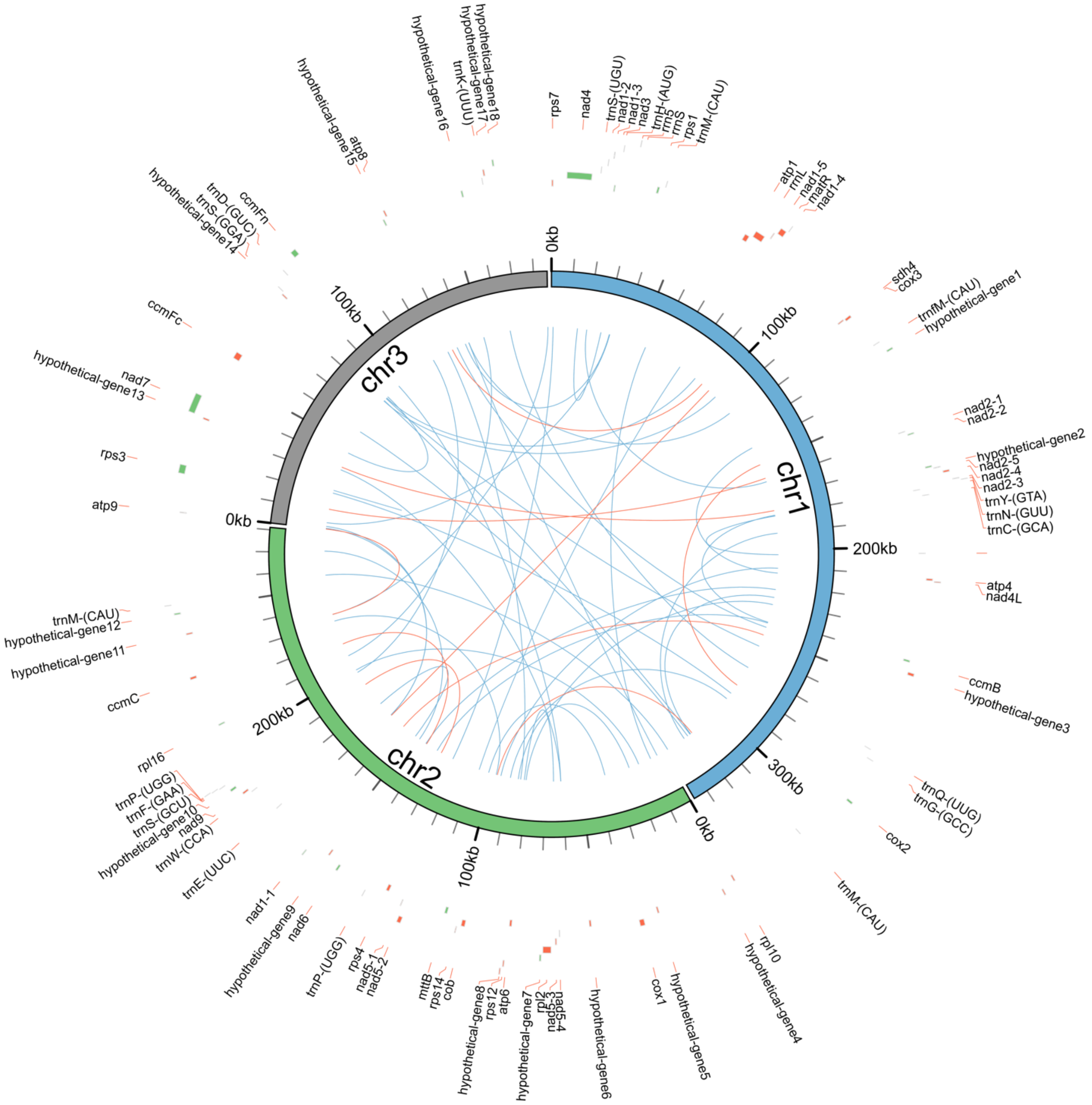

Figure III-S1. *P. trichocarpa* Nisqually-1 mitogenome assembly. Red tiles indicate a feature on the reverse strand and green tiles indicate a feature on the forward strand. Red links indicate repeats greater than 100 bp and blue links indicate repeats less than 100 bp.

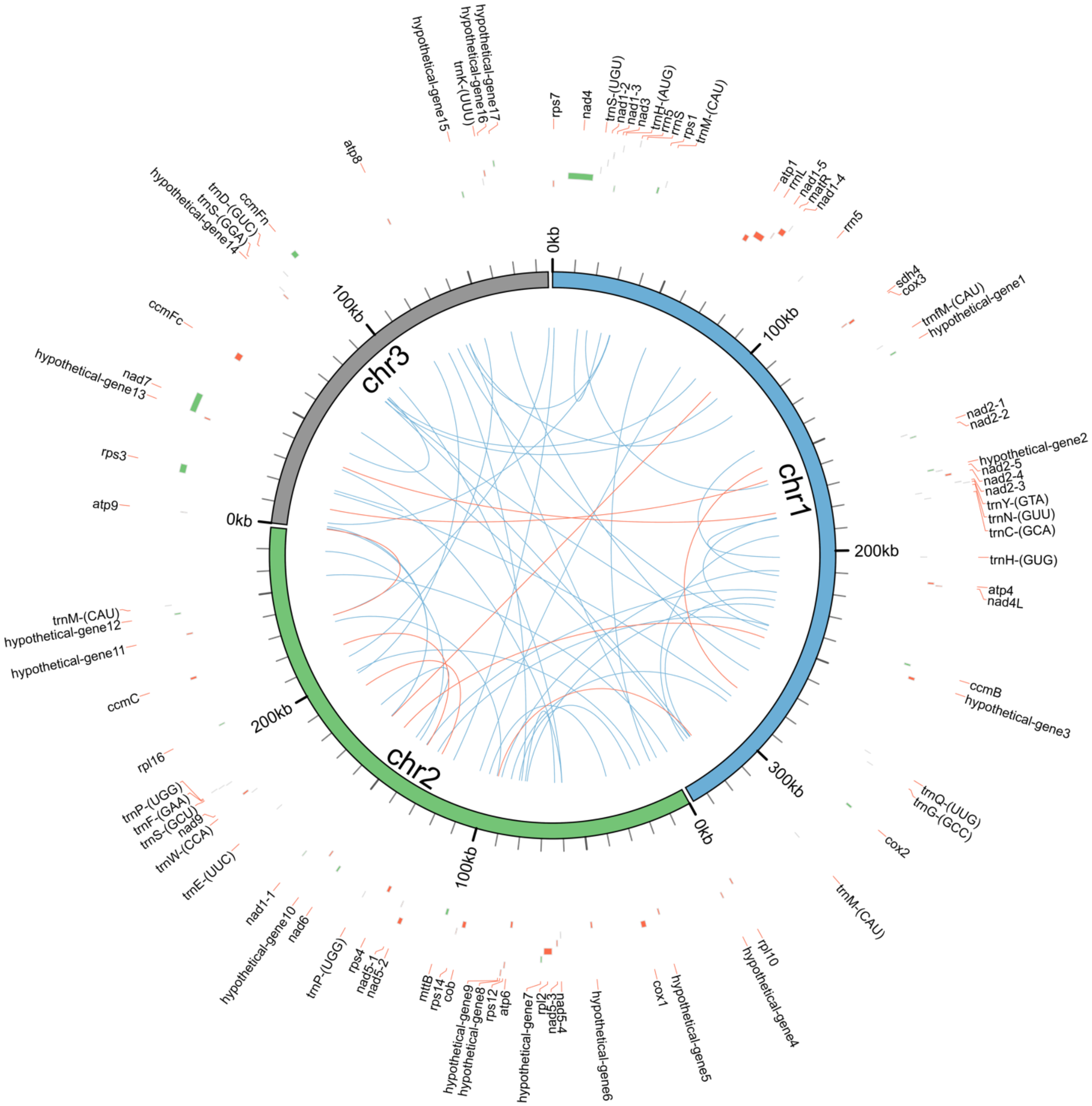

Figure III-S2. *P. deltoides* WV94 mitogenome assembly. Red tiles indicate a feature on the reverse strand and green tiles indicate a feature on the forward strand. Red links indicate repeats greater than 100 bp and blue links indicate repeats less than 100 bp.

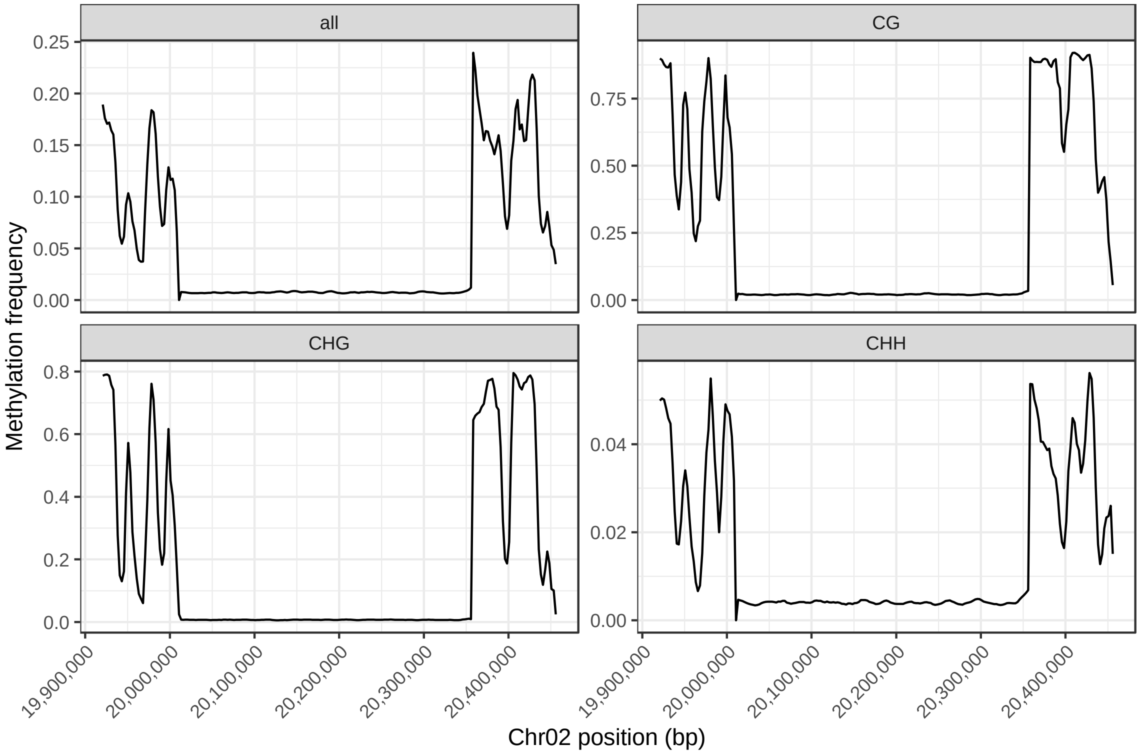

Figure III-S3. Methylation frequency of a large 336 kb possible misassembly (Chr02:20,020,773-20,357,166) in *P. deltoides*. Methylation frequencies were averaged in 10kb windows, a sliding distance of 2.5 kb was used. Regions highlighted in red indicate the target region.

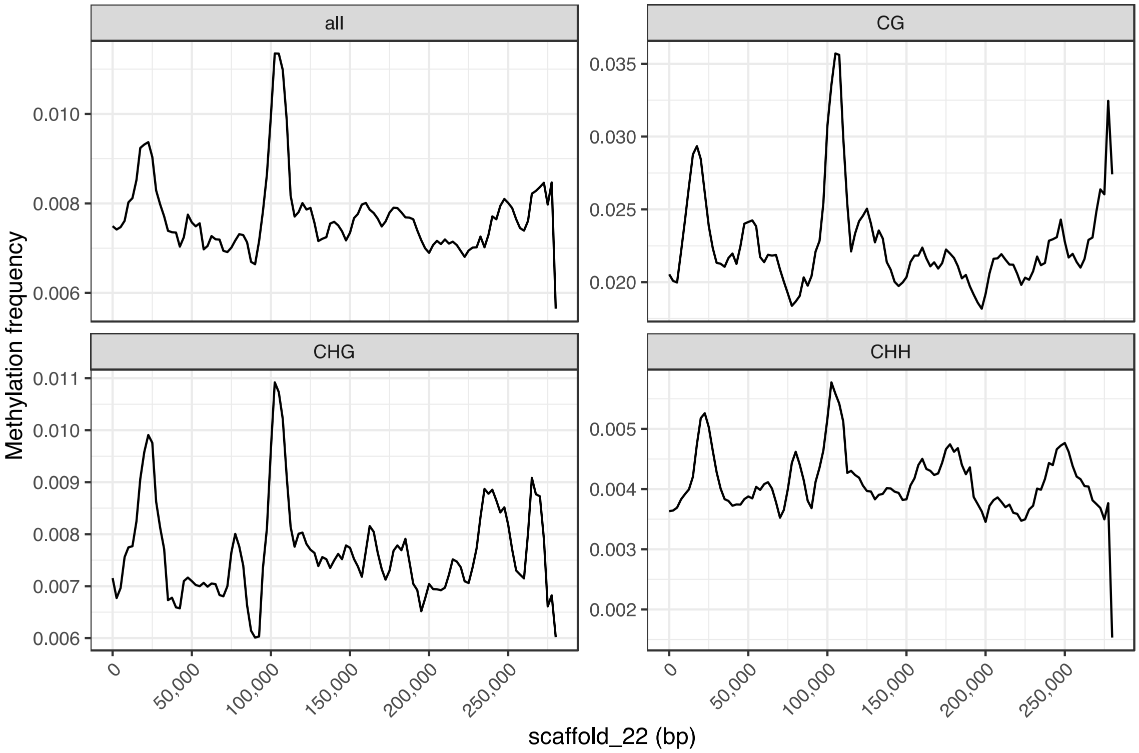

Figure III-S4. Methylation frequency of scaffold_22 in *P. deltoides.* Methylation frequencies were averaged in 10kb windows, a sliding distance of 2.5 kb was used.

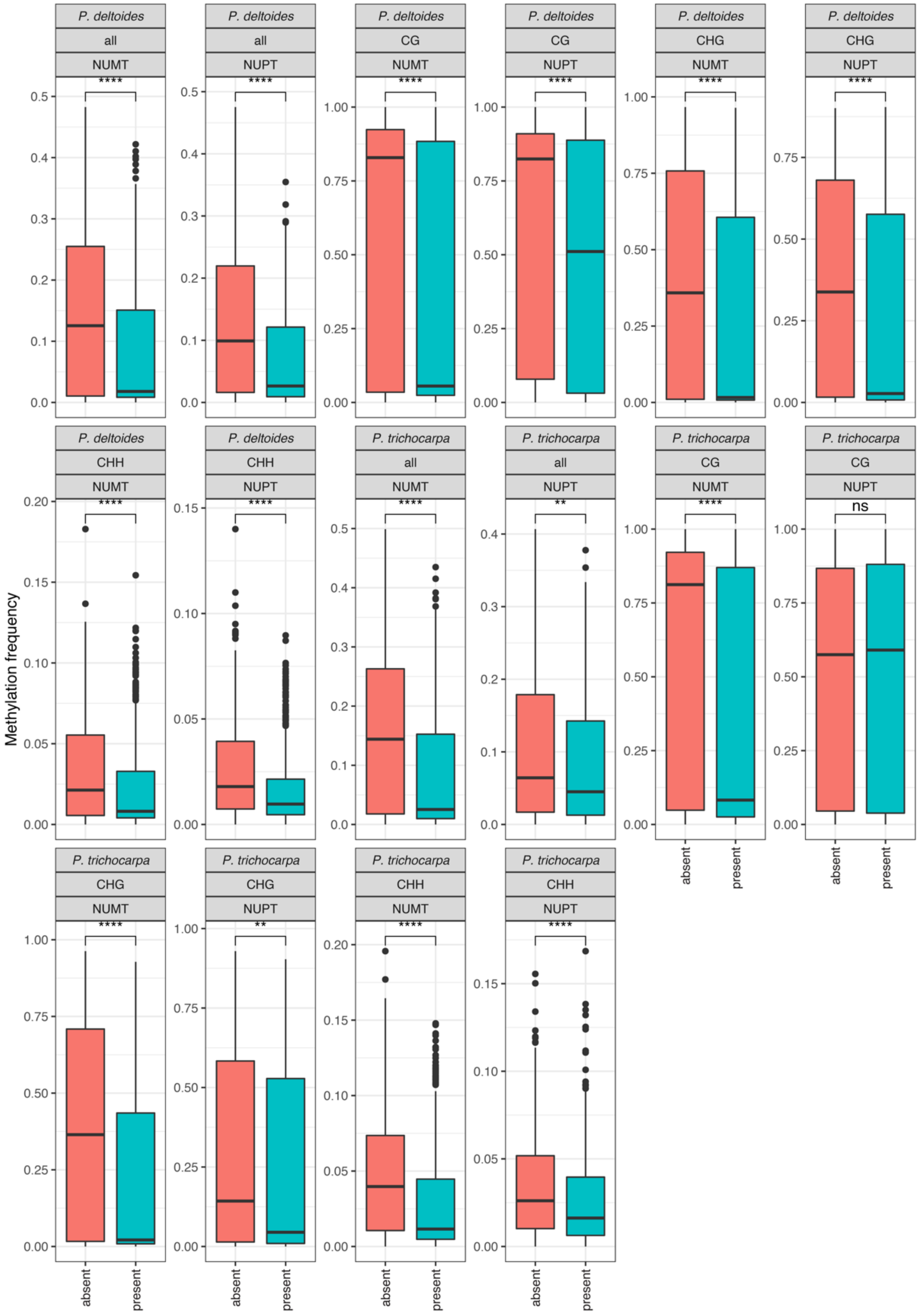

Figure III-S5. Methylation frequency of *P. trichocarpa* and *P. detloides* NUPTs and NUMTs in present and absent contexts, 'all' represents a combined average of all methylation contexts. ns: p > 0.05 ,*: p <= 0.05 ,**: p <= 0.01 ,***: p <= 0.001 ,****: p <= 0.0001, Wilcoxon signed rank test.

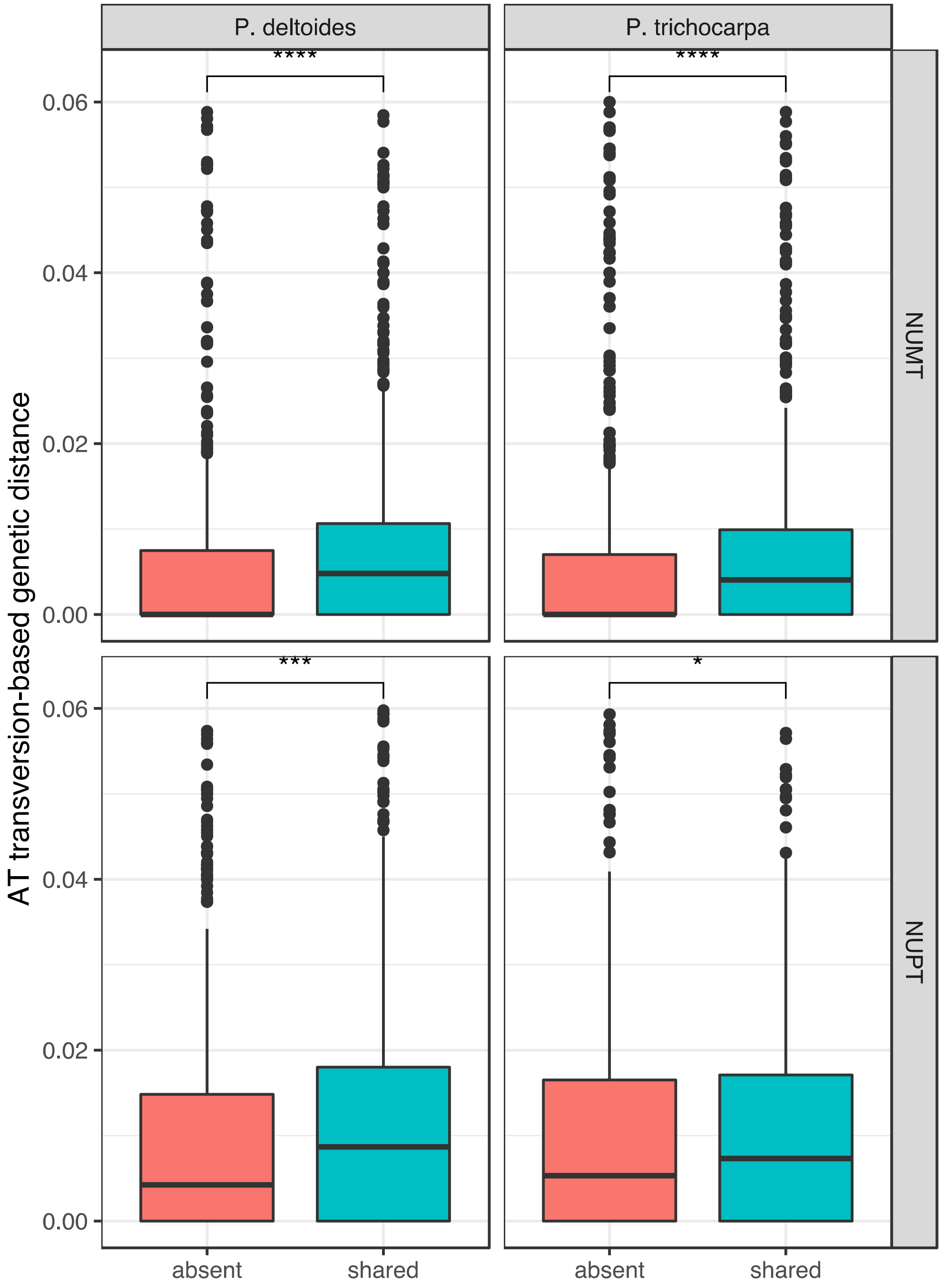

Figure III-S6. AT transversion-based genetic distance of of *P. trichocarpa* and *P. deltoides* NUPTs and NUMTs in present and absent contexts. ns: p > 0.05 ,*: p <= 0.05 ,**: p <= 0.01 ,***: p <= 0.001 ,****: p <= 0.0001, Wilcoxon signed rank test.

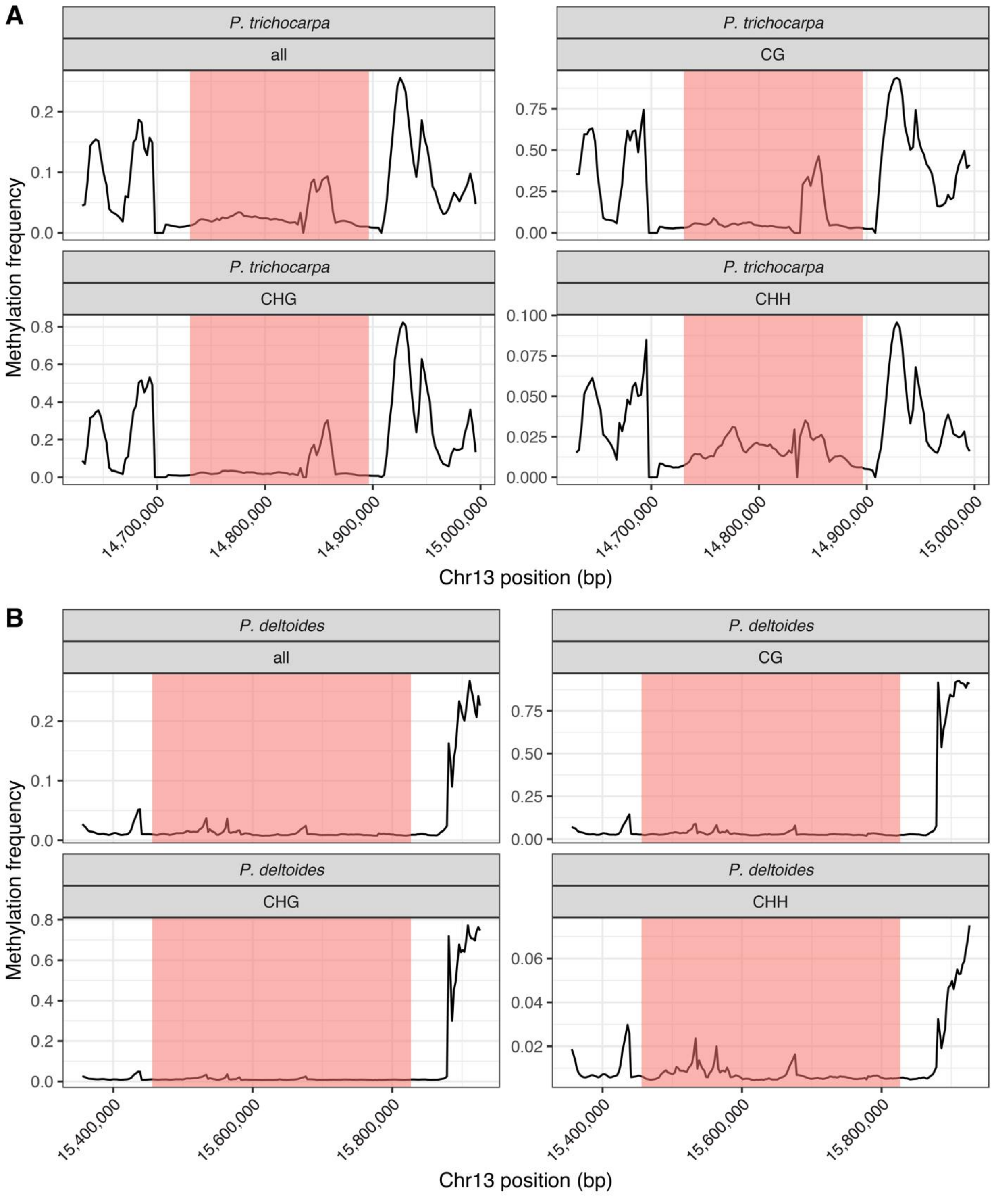

Figure III-S7. Methylation frequency in the *P. trichocarpa* (A) and *P. deltoides* (B) whole chloroplast integration events with 100 kb flanking regions. Regions highlighted in red indicate the whole chloroplast integration event. 10kb windows with a sliding distance of 2.5kb was used.

CHAPTER IV
Assembly, annotation, and characterization of 11 *Salix* organelle genomes

A version of this chapter is in preparation for publication by:

Timothy B. Yates, Kai Feng, Chelsea N. Cereghino, Sara S. Jawdy, Lee E. Gunter, Larry B. Smart, and Wellington Muchero

Author contributions: T.B.Y. and W.M. designed this study. T.B.Y conducted data analysis and wrote the manuscript. S.J.S., L.E.G., and C.N.C generated data. K.F. performed data analysis. L.B.S. provided materials. T.B.Y., L.B.S. and W.M. revised the manuscript.

Abstract

The genus *Salix* consists of more than 350 species, has an expansive natural range, and is economically and ecologically important. Here, we provided 11 assembled and annotated chloroplast and mitogenomes which consisted of six different species. We also characterized structural variation in these genomes and observed highly homogenous chloroplast genomes and heterogenous mitogenomes. Furthermore, protein-coding gene content was very conserved in the chloroplast genomes which had 77 protein-coding genes and one pseudogene. The mitogenomes were also fairly conserved as most genomes had 33 protein-coding genes and one pseudogene. Interestingly, two genotypes (*S. integra*-P336 and *S. viminalis*-Jorr) were missing the large subunit ribosomal protein *rpl10*. Repeat content was highly similar in the chloroplast genomes and varied considerably in the mitogenomes. Notably, *S. udensis*-04-BN-051, had a large 11.5kb repeat and *S. suchowensis*-P63, *S. viminalis*-Jorr, and *S. integra*-P336, had a large 7kb repeat which was conserved between them. Phylogenomic analyses in both chloroplast and mitogenome contexts proved to be challenging due to the low variability in *Salix*. Although relationships at the subgenera level could be resolved, species could not be classified correctly. Future *Salix* phylogenomic studies should utilize nuclear genome data. Lastly, we assessed selection on genes in the 11 organelle genomes and determined *rps7* and *rpl23* in the chloroplast genomes were under positive selection, and the remaining genes were under purifying selection. In the mitogenomes, we identified six genes (*atp4*, *ccmB*, *nad3*, *nad4*, *rps4*, and *sdh4*) under positive selection. Overall, these highly curated *Salix* organelle genomic resources should aid in future domestication and conservation efforts.

Introduction

The family Salicaceae consists of three genera (*Salix, Populus,* and *Chosenia)* and is placed in the order *Malpighiales*. *Salix* and *Populus* share many characteristics and diverged from a common ancestor approximately 52 million years ago (Dai et al., 2014). Additionally, the genus *Salix* consists of approximately 400-500 species of trees and shrubs which are both ecologically and economically important (Argus, 2010; Argus, 1997).

In addition to nuclear genomes, plants have haploid chloroplast and mitochondrial genomes which perform photosynthesis and supply energy to the cell, respectively. Chloroplast genomes have a highly conserved quadripartite structure which consists of a large single-copy (LSC), small single-copy (SSC), and two inverted repeat (IR) regions. Almost all chloroplast genomes range in size from 120kb to 160kb and contain between 110-130 genes (Fan et al., 2018; Palmer, 1985). In contrast, land plant mitochondria genomes are highly heterogeneous even among closely related species with most ranging in size from 200kb to 800kb (S. Wang et al., 2018). They contain 41 variable protein-coding genes, 3 ribosomal RNA (rRNA) genes, and tRNA genes which are insufficient in number to carry out the translation of all codons. These missing tRNAs are compensated by tRNAs encoded in the nuclear genome (Adams & Palmer, 2003).

Although chloroplast genomes can be readily assembled with next-generation sequencing (NGS) data, genome assembly of mitogenomes with short-read data remains a challenge due to repeats and chloroplast DNA transfer to the mitogenome which may be longer than the short read length (Cole, et al., 2018; Sloan, 2013). Currently, there are only 290 land plant mitogenomes, which consists of 204 genera that have been deposited at NCBI Genbank, of which there are seven assembled *Salix* mitogenomes (accessed 01/2022). Single molecule based assembly of mitogenomes with technologies such as PacBio or Oxford Nanopore have already proven to be highly useful for the characterization of complex mitogenomes and should facilitate the assembly of more plant mitogenomes (Jackman et al., 2020; Li et al., 2021).

Previous studies have explored *Salix* phylogenomics through whole chloroplast or mitochondria genome analysis (Chen et al., 2020; Huang et al., 2017; Wagner et al., 2021; Zhou et al., 2021). Organelle phylogenomics in *Salix* is challenging due to the low variation in the genus and resolution at the species level is unreliable (Percy et al., 2014). However, resolution in

Salix is sufficient phylogenetic inference at the subgenera level. In *Salix,* two main subgenera include *Salix s.l.* (tree willows) and *Chamaetia/Vetrix* (shrub willows) (Wagner et al., 2021).

In addition to phylogenomics analyses, structural variation as well as gene presence and absence was assessed within the 11 *Salix* organelle genomes and publicly available organelle genomes from the *Salicaceae*. The levels of horizontal transfer from the chloroplast to the mitogenomes were also determined for each of the 11 genotypes. We also characterize repeat content and assess selection on protein-coding genes in both chloroplast and mitogenomes. From these analyses, we describe and characterize the diversity present in 11 *Salix* organelle genomes and provide high quality organelle assemblies and genomic resources for *Salix*.

Results

Assembly and annotation of the 11 *Salix* chloroplast and mitochondrial genomes

The assembly of the mitogenomes and chloroplast genomes was performed with Unicycler and GetOrganelle, respectively (Jin et al., 2020; Wick et al., 2017). The chloroplast genome assemblies were highly similar in both size and structure and showed little divergence (Table IV-S1, Figure IV-1). Pairwise identity was calculated between all chloroplast genomes and mean identity was 99.77% (Table IV-S2). The most similar chloroplast genomes based on sequence identity were two *S. suchowensis* genotypes P294 and P295 (99.998%) while the most dissimilar were between P336 (*S. integra*) and 07-MBG (*S. viminalis*) (99.491%). To further assess structural differences between the chloroplast assemblies we explored the contraction and expansion of sequence boundaries around the inverted repeat regions (IRs) (Figure IV-S1). The short single-copy region (SSC) and large single-copy (LSC) region showed small variation in size across the 11 *Salix* chloroplast genomes and the inverted repeats were identical further indicating a highly homogenous structure. Gene numbers were identical across all 11 *Salix* chloroplast genomes with 77 protein-coding genes, one pseudogene (*infA*), 4 rRNAs, and 27 tRNAs. To further assess the reliability of the annotations we examined protein-coding gene presence and absence between publicly available *Populus* and *Salix* chloroplast genomes and the 11 *Salix* chloroplast genomes generated here (Figure IV-2A, Table IV-S3). Although variation in gene presence and absence was low across the chloroplast genomes some differences were apparent. For example, the *rps7* gene is absent in the publicly available *S. acutifolia, S. helvetica,* and *S. myrsinifolia*. These three *Salix* genomes are also missing the ycf15 annotation which should be corrected. Additionally, the *rps16* gene which is present in *S. acutifolia, S. helvetica,* and *S. myrsinifolia* has been shown to be lost in the *Salicaceae* and examination of the annotation of *rps16* in these three genomes suggested this gene was annotated incorrectly (Huang et al., 2017).

In contrast to the chloroplast assemblies, the mitogenome assemblies were more heterogenous in nature and showed evidence of differences in gene number, variation in size, and extensive rearrangement (Figure IV-S2). First, gene presence and absence was assessed in the 11 *Salix* mitogenomes with five publicly available *Salix* and two *Populus* mitogenomes (Figure IV-2B, Table IV-S3,). Although gene content was similar within the 11 *Salix* genomes as nine out of the 11 genomes had 34 protein-coding genes both *S. integra* (P336) and *S. viminalis* (Jorr) were missing the large subunit ribosomal protein *rpl10* gene and had 33 protein-coding genes. Interestingly, the mitogenome annotations of *S. paraflabellaris* (MK575518), *S. cardiophylla* (MT806745), and *S. polaris* (NC_052709) also indicated the absence of *rpl10*. After reannotation of these three genomes, it was apparent that *rpl10* and the three trans-spliced NADH genes (*nad1*, *nad2*, *nad5*) may be present in these genomes but were missing from the annotation (Table IV-S5). Furthermore, *rps14* was missing in all five publicly available *Salix* mitogenomes. We retained the *rps14* annotation in the 11 *Salix* mitogenome annotations as this gene had expression evidence based on RNA-Seq data from 8 different tissues (Figure IV-S3). RNA gene numbers were consistent for ribosomal RNAs as all genomes had three, however tRNA numbers were more variable as both *S. purpurea* (94006) and *S. integra* (P336) were missing trnH-AUG. All other mitogenomes had 21 tRNAs, and some differences in copy number were evident (Table IV-S5).

Although large amounts of variation were present in the 11 mitogenomes, two *S. suchowensis* (P294 and P295) mitogenome assemblies were highly homogenous which suggested they are siblings. They showed identical gene numbers, no gene rearrangements, and their genome structures consisted of one large circular subgenome (~556kb) and one smaller linear subgenome (~83 kb). Furthermore, whole genome alignment between these two genomes, also indicated no structural rearrangement (Figure IV-S2). Additionally, mitogenome conformation was conserved at the species level (Table IV-S1). For example, the *S. purpurea* (94001 and 94006) and *S. koriyanagi* (SH3 and 04-FF-016) mitogenomes assembled as single circles while the *S. suchowensis* mitogenomes (P294, P295, and P63) all had one subgenomic segment (Figure IV-2C). We also explored variation in mitogenome size among publicly available *Salix* mitogenomes and the 11 mitogenomes assembled here (Figure IV-2D). The average size based on 18 genomes is 635 kb, the species with the smallest mitogenome is *S. polaris* (562 kb) while the species with the largest mitogenome is *S. cardiophylla* (735 kb).

Repeat sequences in the 11 *Salix* chloroplast and mitogenomes

We identified long repeats and simple sequence repeats (SSRs) in the 11 chloroplast and mitogenomes with REPuter and MISA, respectively (Beier et al., 2017; Kurtz et al., 2001). Long or dispersed repeats are nucleotides sequences present in multiple instances at random positions in the genome (Fan et al., 2018). Recombination via dispersed repeats has been shown to be essential for organelle genome maintenance and is the primary mechanism underlying double-strand break repair in plant mitochondrial genomes (Gualberto & Newton, 2017; Sullivan et al., 2019). Additionally, recombination via dispersed repeats in plant mitogenomes is an important driver of genome structural diversity and changes in conformations may have important phenotypic impacts (Sloan, 2013; Woloszynska, 2010).

Microsatellites or SSRs are tandemly repeated sequences with unit sizes from one to six (monomer to hexamer) (Powell et al., 1996). SSRs have been used as markers for marker assisted selection and for phylogenetic inference due to their high polymorphism within species (Kalia et al., 2011). Across the 11 chloroplast and mitogenomes total SSR numbers were consistent (Figure IV-3A). Average counts were 86 and 100 SSRs in the chloroplast and mitogenomes, respectively. SSR distribution was biased towards monomeric repeats which on average accounted for 93% and 77% of SSRs in the chloroplast and mitogenomes, respectively (Figure IV-3B, Table IV-S6). The second most common SSR class were dimers and pentamers in the chloroplast and mitochondria at 4.7% and 9.7%, respectively. Surprisingly, there were no detected SSRs which had a repeat length of three or 4 in the chloroplast genomes. Additionally, there were no pentamer or hexamer SSRs located in coding regions in either the chloroplast or mitogenomes. Coordinates for all SSRs are available in Table IV-S7.

Dispersed repeats or long repeats were identified in forward, reverse, and palindromic contexts in the 11 chloroplast and mitogenomes (Figure IV-3C, Table IV-S8). Overall, forward (direct) repeats were most numerous in the chloroplast and mitogenomes except for four genotypes (94006, P63, P336, and 04-BN-051) where palindromic repeats were larger in quantity. The chloroplast genome of *S. integra* (P336) and the mitogenome of *S. koriyanagi* (SH3) had the largest percentage of repeats at 2.1% and 9.4%, respectively (Figure IV-S4). Additionally, mitogenome size was weakly correlated (*r* = 0.26) with repeat content while chloroplast genome size was moderately correlated (*r* = 0.61). We examined repeats by size classes (35-50 bp, 51-70 bp, 71-90 bp, and >90 bp) (Figure IV-S5). Repeats in the 35-50bp class had the largest number of counts and on average made up 89% and 75% of repeats in the chloroplast and mitogenomes, respectively. Large repeats were also variable in the 11 mitogenomes. There was some consistency at species level as both *S. purpurea* genotypes (94001 and 94006) and two of the *S. suchowensis* genotypes (P294 and P295) did not have repeats larger than 400 bp. However, the remaining genotypes had larger repeat sizes (Table IV-S8), the largest being *S. udensis* (04-BN-051) which had an 11.1 kb repeat that was shared between its two subgenomes. Additionally, differences in maximum repeat length were evident between the two *S. viminalis* genotypes (07-MBG and Jorr) which had repeat sizes of 1.5 kb and 7.4 kb, respectively. Also, three genotypes (P63, Jorr, and P336) from three different species (*S. suchowensis, S. viminalis,* and *S. integra*) had 7 kb repeats. Based on pairwise alignment, these 7kb repeats were identical between P336 and Jorr and had 80% coverage (5.7 kb) in P63.

Phylogenomic analysis of the 11 *Salix* chloroplast and mitogenomes with publicly available genomes

Chloroplast genomes have been frequently used for phylogenetic studies due their small size, conserved structure, and uniparental inheritance (Gitzendanner et al., 2018). To determine the phylogenetic relationships of the 11 genomes within the genus *Salix*, we constructed a phylogeny based on the alignment of whole chloroplast genomes (Figure IV-4A). This phylogenetic tree included the 11 chloroplast genomes assembled here and 61 publicly available *Salix* chloroplast genomes used in (Wagner et al., 2021) (Table IV-S9). We observed similar tree topologies to those in Wagner et al., 2021 which included distinct clades for *Salix* s.l. (tree willows), Amygdalinae, and *Chamaetia/Vetrix* (shrub willows). As expected, the 11 *Salix* were part of the *Chamaetia/Vetrix* clade. Additionally, we observed short branch lengths in the *Chamaetia/Vetrix* clade and low percentages of variable sites. From the alignment of 72 whole plastome sequences there were 2.3% variable sites and in the *Chamaetia/Vetrix* clade (59 sequences) there were 0.59% variable sites. Overall, as was observed in Wagner et al., 2021, resolution in the *Chamaetia/Vetrix* clade is poor and relationships at the species level in the 11 genomes are not accurate except for *S. koriyanagi* (SH3 and 04-FF-016) and *S. suchowensis* (P294 and P295).

In comparison to chloroplast genomes, mitogenome based phylogenies are less commonly used for phylogenetic inference in plants. This is primarily due to their poor conservation and slower evolutionary rates compared to the nuclear and chloroplast genomes. However, there have been examples where mitogenome derived phylogenies are informative (S. Wang et al., 2018) Therefore, we constructed a phylogeny based on the whole genome alignment of locally collinear blocks (LCBs) of seven publicly available *Salix* mitogenomes, the 11 *Salix* mitogenomes, and two *Populus* mitogenomes which were used as outgroups (Figure IV-4B, Table IV-S9). We observed two distinct *Populus* and *Salix* clades. As was observed in the chloroplast derived phylogeny, branch lengths were short and overall percentage of variable sites were low at 0.52% and when excluding the *Populus* genomes the percentage of variable sites were 0.34%. Phylogenetic relationships are not accurate in mitogenome derived phylogenetic tree for the 11 *Salix* genomes except for two *S. suchowensis* genotypes (P294 and P295).

Chloroplast DNA transfer to the mitogenome in gene and intergenic contexts

Chloroplast DNA transfer to the 11 mitogenomes was identified with BLASTN. Each of the 11 chloroplast genomes were searched against their corresponding mitogenome and variation in chloroplast DNA transfer was apparent (Figure IV-5A). By total size, the *S. koriyanagi* SH3 mitogenome had the largest amount of chloroplast DNA at 17.3 kb. However, by percentage of mitogenome size *S. purpurea* 94006 had the largest amount of chloroplast DNA at 2.86% (Figure IV-5B). Comparatively, *S. udensis* 04-BN-051 had the lowest amount of chloroplast DNA by size and percentage at 12.2 kb and 1.85%, respectively. At the species level, transfer rates were consistent between two *S. suchowensis* genotypes (P294 and P295) and the *S. koriyangi* genotypes (SH3 and 04-FF-016). Additionally, based on the comparison of query and subject alignment lengths all genotypes except for SH3 and 04-FF-016 had total chloroplast DNA transfer sizes which were smaller than the chloroplast progenitor sequence. In these genotypes this is likely due to indels in these regions after insertion. Interestingly, SH3 and 04-FF-016 both had 2.1 kb more total chloroplast DNA content. This was due to transfer from the same locus which occurred independently or duplication in the mitogenome after initial integration.

Gene transfer from the chloroplast genome to the mitogenome was also determined for the 11 mitogenomes with BLASTN. Although, there was no functional protein-coding gene transfer, across all 11 genotypes 9 tRNA genes were transferred. These included trnD-GUC, trnH-GUG-cp, trnN-GUU-cp, trnC-ACA, trnS-GGA-cp, trnS-GGA-cp, trnS-UGA, trnV-GAC and trnW-CCA-cp.

Selection on protein-coding genes in the 11 *Salix* chloroplast and mitogenomes

The ratio of nonsynonymous (Ka) to synonymous (Ks) mutation is a metric commonly used to detect selection in protein-coding genes and is an estimate of how much non-neutral evolution has occurred relative to neutral evolution. Ka/Ks values less than one indicate purifying or negative selection while values equal to one indicate neutral selection, and values greater than one indicate positive or diversifying selection. Selection on protein-coding genes in the 11 chloroplast and mitogenomes was estimated through comparison with protein-coding genes from species outside the *Salicacea*e family but within the order *Malpighiales*. Genes with evidence of positive selection are available in Table IV-S10.

For the chloroplast genomes, the protein-coding genes from *J. curcas* (NC_012224.1), *M. esculenta* (NC_010433.1), and *R. communis* (NC_016736.1) were used to assess selection on the protein-coding genes in the 11 chloroplast genomes (Figure IV-6A). Although most genes were under purifying selection, *rpl23* and *rps7*, both ribosomal proteins had average Ka/Ks values greater than one at 1.4 and 2.1, respectively. No genes showed evidence of neutral evolution. Levels of non-neutral (Ka) and neutral (Ks) evolution were also assessed across the chloroplast protein-coding genes (Figure IV-S6A). Notably, *ndhK* had a very high Ks value suggesting this gene is highly constrained.

For the mitogenomes, the protein-coding genes from *M. esculenta* (NC_045136.1), *P. edulis* (NC_050950.1), and *R. communis* (NC_015141.1) were used to assess selection on protein-coding genes in the 11 mitogenomes (Figure IV-6B). There were six protein-coding genes (*atp4, ccmB, nad3, nad4, rps4,* and *sdh4*) which showed evidence of positive selection. In contrast, two genes that were under strong purifying selection were two ATP synthases, *atp1* and *atp9*.

Moreover, *atp9* had the highest average Ks value further which indicated strong constraint (Figure IV-S6B).

Selection on protein-coding genes was in the chloroplast and mitogenomes was also explored by gene functional class of which there were ten and nine classes in the chloroplast and mitogenomes, respectively. In the chloroplast genomes all gene functional classes had median Ka/Ks values which were less than one (Figure IV-6C). However, some classes appear to be more constrained than others. The classes under very strong purifying selection based on median Ka/Ks value include ATP synthases, Cytochrome b/f complex, Photosystem I, Photosystem II, and the RubisCO large subunit. In contrast, the Ribosomal proteins appeared to be less constrained. In the mitogenome functional classes, Complex II (succinate dehydrogenase) had a median Ka/Ks value of 1.27. Additionally, Cytochrome C Biogenesis and small subunit ribosomal proteins had the next highest median Ka/Ks values at 0.78 and 0.77, respectively. In contrast, the gene functional class with the lowest median Ka/Ks value was the large subunit ribosomal proteins at 0.31.

Discussion and Conclusion

Here, we provided 11 high quality chloroplast and mitogenome assemblies from six different *Salix* species. Although there are large numbers of publicly available chloroplast genomes from a diverse set of *Salix* species, there are only seven publicly available *Salix* mitogenomes. Land plant mitogenomes are known be highly diverse even between closely related species (Mower, 2020). Therefore, the small number of publicly available *Salix* mitogenomes is a limiting factor prohibiting the understanding of mitogenome diversity within *Salix* and land plants overall. Although the 11 mitogenome assemblies provided here may slightly increase the understanding of mitogenome diversity in *Salix*, due to the large numbers of species in *Salix* (400-500), there is still much to be done.

Major structural variation was non-existent among the 11 chloroplast genomes and pairwise identity values were very high. In contrast, structural variation and rearrangement were widespread among the mitogenome assemblies which showed great variation in size. For example, the largest assembly (*S.* udensis-04-BN-051) was 655 Kb and the smallest assembly (*S. purpurea*-94006) was 599 Kb indicating that the *S. purpurea*-94006 mitogenome has lost 55 Kb of sequence since their divergence or duplication has occurred in *S. udensis*. The extensive structural variation and rearrangement observed in plant mitogenomes is a product of recombination and non-homologous end joining which is the repair mechanism used in non-coding regions (Christensen, 2013, 2017; Sullivan et al., 2019).

Gene content across the chloroplast and mitogenome assemblies was generally very conserved, especially in the chloroplast genomes. Initially, when gene content in the 11 mitogenomes were compared with other publicly available *Salix* mitogenomes, some variation in the presence and absence of NADH dehydrogenase genes was apparent. However, upon closer inspection, these genes (*nad1*, *nad2*, and *nad5*) all of which undergo *trans*-splicing were incorrectly annotated. Interestingly, the large ribosomal subunit protein *rpl10* was missing from the *S. integra* (P336) and *S. viminalis* (Jorr) mitogenome assemblies. Previously, *rpl10* has been shown to be absent from other mitogenome assemblies and ribosomal protein genes have been shown to be functionally transferred to the nuclear genome (Adams et al., 2002; Alverson et al., 2011). Future nuclear genome assemblies of these genotypes could validate the functional transfer of *rpl10* in *S. integra* (P336) and *S. viminalis* (Jorr).

Repeat-mediated recombination plays an essential role in replication, repair, and stability of chloroplast and mitogenomes (Maréchal & Brisson, 2010). Dispersed repeat lengths were considerably smaller and less variable in size in the chloroplast genomes compared to repeats in the mitogenomes. Small numbers of repeats have been observed in plastid genomes as they have been shown to cause instability and are therefore likely selected against (Staub & Maliga, 1994). Although large repeats are present in the mitogenomes assembled here, future work will address if they are active in facilitating recombination.

The percent of variable sites within the *Chamaetia/Vetrix* clade based on the alignment of whole plastomes were exceptionally low. Wagner et al., (2021) reported 0.4% variable sites and we determined similar levels of variable sites at 0.59%. In comparison, an alignment constructed from RAD sequencing data in Wagner et al., 2021 had 8.05% variable sites. Consequently, the low variability in the chloroplast genomes resulted in the poor resolution and incorrect phylogenetic assignment for some of 11 *Salix* species in the *Chamaetia/Vetrix* clade. Additionally, this same trend of low levels of variable sites were also observed in the mitogenome derived phylogeny. Land plant mitogenomes are not commonly used for phylogenetic inference due to their low rates of evolution relative to the chloroplast and mitogenomes. Even the use of a dedicated LCB alignment pipeline which should perform better than a phylogeny derived from the alignment of protein-coding genes resulted in short branch lengths and incorrect phylogenetic relationships. Future phylogenetic analyses in *Salix* should rely on data derived from the nuclear genomes.

The majority of genes were under purifying selection in the 11 chloroplast genome and only two ribosomal subunit genes (*rps7* and *rpl23*) showed evidence in positive selection. The *rps7* gene is located in the IR and has been previously shown to have high nucleotide diversity based on an analysis of 21 *Salix* chloroplast genomes and has been proposed as a potential molecular marker for species identification (Zhou et al., 2021). Furthermore, *rps7* has also been shown to be under positive selection in the *Salicaceae* (Huang et al., 2017). In *P. trichocarpa*, there is functional copy of *rps7* in the nuclear genome, which may explain the lack of constraint in the 11 *Salix* chloroplast genomes. Future nuclear genome assemblies should assess whether functional transfer of *rps7* has occurred. Positive selection on *rpl23* has also been detected in the fabids, which may suggest that it has also transferred successfully to the nuclear genome (Han et al., 2020). Protein-coding genes in the mitogenomes had larger numbers of genes under positive selection compared to the plastid genomes. Genes under positive selection were from functional classes such as ATP synthases, Cytochrome C Biogenesis, NADH dehydrogenases, succinate dehydrogenases, and small ribosomal subunit proteins. Although, the number of studies which have examine selection of mitogenome protein-coding genes in *Salix* is limited, one study which assembled the *S. suchowensis* mitogenome has identified *ccmB* with a Ka/Ks greater than 1 (Ye et al., 2017). However, they did not identify *atp4*, *nad3*, *nad4*, *rps3*, or *sdh4* as under positive selection, although *atp4* and *sdh4* did have Ka/Ks values relatively close to 1.

In this study, we provide novel high quality organelle genomic resources for 6 *Salix* species. Furthermore, we identified and characterized structural variation in these genomes in both coding and intergenic contexts. We also identified variable rates of chloroplast DNA transfer to the mitogenomes. Repeat content was highly variable in the mitogenomes and were generally homogenous in the chloroplast genomes. Additionally, we performed phylogenetic analyses which were reliable at the subgenera level. Lastly, we identified levels of selection on protein-coding genes and identified a subset which were under positive selection and may have functional copies in the nuclear genome.

Materials and Methods

DNA extraction and sequencing: Fresh young leaf tissue (approximately 100 mg) for all 11 *Salix* genotypes was collected and ground in liquid nitrogen using the Qiagen TissueLyser II with one 5mm stainless steel bead. DNA extraction was performed using a modified CTAB based protocol (Doyle & Doyle, 1987). Briefly, the organic and aqueous phase were extracted using chloroform:isoamyl alcohol 24:1. After separation, a SPRI bead solution was used to select for reads greater than 1kb (Mayjonade et al., 2016). For long read sequencing, 1 ug of DNA was used as input to Oxford Nanopore's genomic DNA by ligation sequencing kit (SQK-LSK109) and the subsequent library was sequenced on a R.9.4.1 flow cell. Short read sequencing of the same samples was performed on the illumina HiSeq X Ten platform. Raw sequencing data has been deposited at the NCBI and can be accessed with the BioProject ID PRJNA827350.

Organelle genome assembly and annotation: The chloroplast genomes were assembled with GetOrganelle v1.7.5 using default parameters with *S. purpurea* as a bait genome (NC_026722) (Jin et al., 2020). The chloroplast genomes were annotated with GeSeq (Tillich et al., 2017). Publicly available *Salix* chloroplast genomes (Table IV-S11) were used as references during annotation, and BLAT search identity thresholds were set at 85% for protein, rRNA, tRNA, and DNA. The chloroplast genomes have been deposited at NCBI, refer to Table IV-S9 for accession numbers. To enrich for mitogenome derived reads from Oxford Nanopore data, long reads were mapped with nucmer v4.0.0rc1 with an alignment length threshold of 1kb and the 'maxmatch' flag against five doubled *Salix* mitogenomes (*Salix cardiophylla* (MT806745), *Salix paraflabellaris* (MK575518), *Salix polaris* (NC_052709), *Salix purpurea* (NC_029693), and *Salix suchowensis* (NC_029317)). Doubled genomes were used to ensure reads mapped to the entire circular molecule. To enrich for organelle derived reads in illumina data, short reads were mapped with bowtie2 v2.4.2 against the same doubled organelle genomes (Langmead & Salzberg, 2012). Prior to assembly, both the long and short organelle enriched reads were randomly down sampled to 500X coverage. The mitogenomes were assembled with Unicycler v0.4.8 with default parameters using short illumina and long Oxford Nanopore reads (Wick et al., 2017). The *Salix* mitogenomes were annotated with Geneious Prime 2021.1.1 by transferring the mitochondria annotations with an identity threshold of 85% from the five publicly available

Salix mitogenomes used above (Kearse et al., 2012). The annotations were also verified further with OGAP (https://github.com/zhangrengang/OGAP).The mitogenomes have been deposited at NCBI, refer to Table IV-S9 for accession numbers.

Phylogenomic analysis of 11 *Salix* chloroplast and mitogenomes: 61 publicly available *Salix* chloroplast genomes were downloaded from NCBI (Table IV-S9). These 61 genomes were concatenated with the 11 *Salix* chloroplast genomes, rotated to a uniform start point, the inverted repeat A (IRA) was removed, and then aligned with MAFFT v7.487 (Katoh & Standley, 2013). The alignment was trimmed with Gblocks v0.91b, parameters used were '-b1 25 -b2 40 -b3 8'. IQ-TREE v2.1.2 was then used to construct the Maximum Likelihood (ML) tree with 1000 bootstraps and visualized with ggtree v3.0.4 (Minh et al., 2020; Yu, et al., 2017). For the mitogenome phylogeny, seven publicly available *Salix* and two *Populus* mitochondrial genomes were downloaded from NCBI (Table IV-S9). Due to the extensive rearrangement of mitogenomes, the HomBlocks pipeline with default parameters was used to construct a multiple sequence alignment (Bi et al. 2018) This alignment was used to construct a Maximum Likelihood tree was constructed with IQ-Tree v2.1.2 with 1000 bootstraps and visualized with ggtree v3.0.4 (Minh et al., 2020; Yu et al., 2017).

Structural variation analysis: The LAGAN mode of mVista was used to compare the 11 *Salix* chloroplast genomes and visualize differences in structural variation (Brudno et al., 2003; Frazer et al., 2004). Structural variation at the chloroplast genome inverted repeats was identified and visualized with IRscope (Amiryousefi et al., 2018). Structural variation in the 11 *Salix* mitogenomes was assessed and visualized with progressiveMauve with default parameters (Darling et al., 2010).

Plastid-derived content in the *Salix* mitogenomes: BLASTN 2.11.0+ with word size of 11, e-value less than 1E-5, percent identity greater than or equal to 75%, and minimum alignment length of 100 bp was used to detect plastid-derived content in the *Salix* mitogenomes (Altschul, 1977). The whole chloroplast genome minus one inverted repeat were used. Protein-coding and tRNA gene transfer were identified with BLASTN 2.11.0+ with a word size of 11, e-value less than 1E-5, percent identity greater than or equal to 90%, and subject coverage of greater than or equal to 90%.

Repeat analysis: Forward, palindromic, complement, and reverse repeats were identified with REPuter (Kurtz et al., 2001). Repeats with a minimum length of 30, an evalue of less than or equal to 1e-5, and a hamming distance of 3 were retained. Simple sequence repeats (SSRs) were identified with MISA (Beier et al., 2017; Thiel et al., 2003). Unit size of repeat by minimum repeat number were required as follows: monomer – 10, dimer – 6, trimer – 5, tetramer – 4, pentamer – 3, and hexamer – 3.

Selection analysis: Selection was assessed on protein-coding genes in between the 11 *Salix* chloroplast genomes and *J. curcas* (NC_012224.1), *M. esculenta* (NC_010433.1), and *R. communis* (NC_016736.1). In the 11 mitogenomes, selection was assessed on protein-coding genes in between *M. esculenta* (NC_045136.1), *P. edulis* (NC_050950.1), and *R. communis* (NC_015141.1). Protein alignments were first constructed with MAFFT v7.487 which was then converted into a codon alignment with pal2nal v14 with the '-nogap' flag (Katoh & Standley, 2013; Suyama et al., 2006). KaKs_Calculator 2.0 was used to calculate Ka/Ks with the 'NG' model (Wang et al., 2010).

RNA-Seq analysis: RNA was extracted from 8 *Salix* tissues for all 11 genotypes (Table IV-S12) following the protocol described in (Zhang et al., 2018). Strand-specific RNA-Seq libraries were prepared by BGI and sequenced on the DNB-Seq platform which generated 150 bp reads. The raw RNA-Seq data has been uploaded to NCBI and can be accessed with the BioProject ID PRJNA827350. RNA-Seq reads were mapped to a composite chloroplast and mitogenome of the same genotype with STAR v2.7.10a (Dobin et al., 2013). Expression levels (FPKM) of protein-coding genes in the chloroplast and mitogenomes were calculated with StringTie v2.2.1 (Pertea et al., 2015).

References

Adams, K. L., & Palmer, J. D. (2003). Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution*, *29*(3), 380–395. https://doi.org/10.1016/S1055-7903(03)00194-5

Adams, K. L., Qiu, Y. L., Stoutemyer, M., & Palmer, J. D. (2002). Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(15), 9905–9912. https://doi.org/10.1073/pnas.042694899

Altschul, S. (1977). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.*, *25*(17), 3389–3402.

Alverson, A. J., Zhuo, S., Rice, D. W., Sloan, D. B., & Palmer, J. D. (2011). The mitochondrial genome of the legume vigna radiata and the analysis of recombination across short mitochondrial repeats. *PLoS ONE*, *6*(1). https://doi.org/10.1371/journal.pone.0016404

Amiryousefi, A., Hyvönen, J., & Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics (Oxford, England)*, *34*(17), 3030–3031. https://doi.org/10.1093/bioinformatics/bty220

Argus, G. W. (2010). Flora of North America, Volume 7. Magnoliophyta: Salicaceae to Brassicaceae. Oxford University Press, New York, NY and Oxford, UK.

Argus, George W. (1997). Infrageneric Classification of Salix (Salicaceae) in the New World. *Systematic Botany Monographs*, *52*, 1–121.

Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics*, *33*(16), 2583–2585. https://doi.org/10.1093/bioinformatics/btx198

Bi, G., Mao, Y., Xing, Q., & Cao, M. (2018). HomBlocks: A multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genomics*, *110*(1), 18–22. https://doi.org/10.1016/j.ygeno.2017.08.001

Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., & Batzoglou, S. (2003). Glocal alignment: Finding rearrangements during alignment. *Bioinformatics*, *19*(SUPPL. 1). https://doi.org/10.1093/bioinformatics/btg1005

Chen, X., Zhang, L., Huang, Y., & Zhao, F. (2020). Mitochondrial genome of Salix cardiophylla and its implications for infrageneric division of the genus of Salix. *Mitochondrial DNA Part B: Resources*, *5*(3), 3503–3504. https://doi.org/10.1080/23802359.2020.1827065

Christensen, A. C. (2013). Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. *Genome Biology and Evolution*, *5*(6), 1079–1086. https://doi.org/10.1093/gbe/evt069

Christensen, A. C. (2017). Mitochondrial DNA repair and genome evolution. *Annual Plant Reviews*, *50*, 11–31.

Cole, L. W., Guo, W., Mower, J. P., & Palmer, J. D. (2018). High and Variable Rates of Repeat-Mediated Mitochondrial Genome Rearrangement in a Genus of Plants. *Molecular Biology and Evolution*, *35*(11), 2773–2785. https://doi.org/10.1093/molbev/msy176

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G. A., … Yin, T. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research*, *24*(10), 1274–1277. https://doi.org/10.1038/cr.2014.83

Darling, A. E., Mau, B., & Perna, N. T. (2010). Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, *5*(6).

https://doi.org/10.1371/journal.pone.0011147

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *PHYTOCHEMICAL BULLETIN*, (RESEARCH). Retrieved from http://worldveg.tind.io/record/33886

Fan, W. B., Wu, Y., Yang, J., Shahzad, K., & Li, Z. H. (2018). Comparative chloroplast genomics of dipsacales species: Insights into sequence variation, adaptive evolution, and phylogenetic relationships. *Frontiers in Plant Science*, *9*(May), 1–13. https://doi.org/10.3389/fpls.2018.00689

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Research*, *32*(WEB SERVER ISS.), 273–279. https://doi.org/10.1093/nar/gkh458

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., & Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany*, *105*(3), 291–301. https://doi.org/10.1002/ajb2.1048

Gualberto, J. M., & Newton, K. J. (2017). Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. *Annual Review of Plant Biology*, *68*, 225–252. https://doi.org/10.1146/annurev-arplant-043015-112232

Han, K., Shi, C., Li, L., Seim, I., Lee, S. M. Y., Xu, X., … Liu, X. (2020). Lineage-specific evolution of mangrove plastid genomes. *Plant Genome*, *13*(2), 1–13. https://doi.org/10.1002/tpg2.20019

Huang, Y., Wang, J., Yang, Y., Fan, C., & Chen, J. (2017). Phylogenomic analysis and dynamic evolution of chloroplast genomes in salicaceae. *Frontiers in Plant Science*, *8*(June), 1–13. https://doi.org/10.3389/fpls.2017.01050

Jackman, S. D., Coombe, L., Warren, R. L., Kirk, H., Trinh, E., MacLeod, T., … Birol, I. (2020). Complete Mitochondrial Genome of a Gymnosperm, Sitka Spruce (Picea sitchensis), Indicates a Complex Physical Structure. *Genome Biology and Evolution*, *12*(7), 1174–1179. https://doi.org/10.1093/gbe/evaa108

Jin, J. J., Yu, W. Bin, Yang, J. B., Song, Y., Depamphilis, C. W., Yi, T. S., & Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 1–31. https://doi.org/10.1186/s13059-020-02154-5

Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., & Dhawan, A. K. (2011). Microsatellite markers: An overview of the recent progress in plants. *Euphytica*, *177*(3), 309–334. https://doi.org/10.1007/s10681-010-0286-9

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kubo, N., & Arimura, S. I. (2010). Discovery of the rpl10 gene in diverse plant mitochondrial genomes and its probable replacement by the nuclear gene for chloroplast RPL10 in two lineages of angiosperms. *DNA Research*, *17*(1), 1–9. https://doi.org/10.1093/dnares/dsp024

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2001). REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, *29*(22), 4633–4642. https://doi.org/10.1093/nar/29.22.4633

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, J., Xu, Y., Shan, Y., Pei, X., Yong, S., Liu, C., & Yu, J. (2021). Assembly of the complete mitochondrial genome of an endemic plant, Scutellaria tsinyunensis, revealed the existence of two conformations generated by a repeat-mediated recombination. *Planta*, *254*(2), 1–16. https://doi.org/10.1007/s00425-021-03684-3

Maréchal, A., & Brisson, N. (2010). Recombination and the maintenance of plant organelle genome stability. *New Phytologist*, *186*(2), 299–317. https://doi.org/10.1111/j.1469-8137.2010.03195.x

Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., … Muños, S. (2016). Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *BioTechniques*, *61*(4), 203–205. https://doi.org/10.2144/000114460

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., … Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mower, J. P. (2020). Variation in protein gene and intron content among land plant mitogenomes. *Mitochondrion*, *53*(May), 203–213. https://doi.org/10.1016/j.mito.2020.06.002

Palmer, J. D. (1985). Comparative organization of chloroplast genomes. *Annual Review of Genetics*, *19*, 325–354. https://doi.org/10.1146/annurev.ge.19.120185.001545

Percy, D. M., Argus, G. W., Cronk, Q. C., Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., … Graham, S. W. (2014). Understanding the spectacular failure of DNA barcoding in willows (Salix): Does this result from a trans-specific selective sweep? *Molecular Ecology*, *23*(19), 4737–4756. https://doi.org/10.1111/mec.12837

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Powell, W., & Barkley, N. (1996). Polymorphism revealed by simple sequence repeats Polymorphism revealed by simple sequence r e p e a t s w.

Sloan, D. B. (2013). One ring to rule them all? Genome sequencing provides new insights into the "master circle" model of plant mitochondrial DNA structure. *New Phytologist*, *200*(4), 978–985. https://doi.org/10.1111/nph.12395

Staub, J. M., & Maliga, P. (1994). Extrachromosomal elements in tobacco plastids. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(16), 7468–7472. https://doi.org/10.1073/pnas.91.16.7468

Sullivan, A. R., Eldfjell, Y., Schiffthaler, B., Delhomme, N., Asp, T., Hebelstrup, K. H., … Wang, X. R. (2019). The Mitogenome of Norway Spruce and a Reappraisal of Mitochondrial Recombination in Plants. *Genome Biology and Evolution*, *12*(1), 3586–3598. https://doi.org/10.1093/gbe/evz263

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.), 609–612. https://doi.org/10.1093/nar/gkl315

Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*, *106*(3), 411–422. https://doi.org/10.1007/s00122-002-1031-0

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, *45*(W1), W6–W11. https://doi.org/10.1093/nar/gkx391

Wagner, N. D., Volf, M., & Hörandl, E. (2021). Highly Diverse Shrub Willows (Salix L.) Share Highly Similar Plastomes. *Frontiers in Plant Science*, *12*(September). https://doi.org/10.3389/fpls.2021.662715

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., & Yu, J. (2010). KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genomics, Proteomics and Bioinformatics*, *8*(1), 77–80. https://doi.org/10.1016/S1672-0229(10)60008-3

Wang, S., Song, Q., Li, S., Hu, Z., Dong, G., Song, C., … Liu, Y. (2018). Assembly of a complete mitogenome of chrysanthemum nankingense using oxford nanopore long reads and the diversity and evolution of asteraceae mitogenomes. *Genes*, *9*(11). https://doi.org/10.3390/genes9110547

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6), 1–22. https://doi.org/10.1371/journal.pcbi.1005595

Woloszynska, M. (2010). Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes-though this be madness, yet there's method in't. *Journal of Experimental Botany*, *61*(3), 657–671. https://doi.org/10.1093/jxb/erp361

Ye, N., Wang, X., Li, J., Bi, C., Xu, Y., Wu, D., & Ye, Q. (2017). Assembly and comparative analysis of complete mitochondrial genome sequence of an economic plant Salix suchowensis. *PeerJ*, *2017*(3). https://doi.org/10.7717/peerj.3148

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhang, J., Yang, Y., Zheng, K., Xie, M., Feng, K., Jawdy, S. S., … Muchero, W. (2018). Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in Populus. *New Phytologist*, *220*(2), 502–516. https://doi.org/10.1111/nph.15297

Zhou, J., Jiao, Z., Guo, J., Wang, B. song, & Zheng, J. (2021). Complete chloroplast genome sequencing of five Salix species and its application in the phylogeny and taxonomy of the genus. *Mitochondrial DNA Part B: Resources*, *6*(8), 2348–2352. https://doi.org/10.1080/23802359.2021.1950055

Appendix

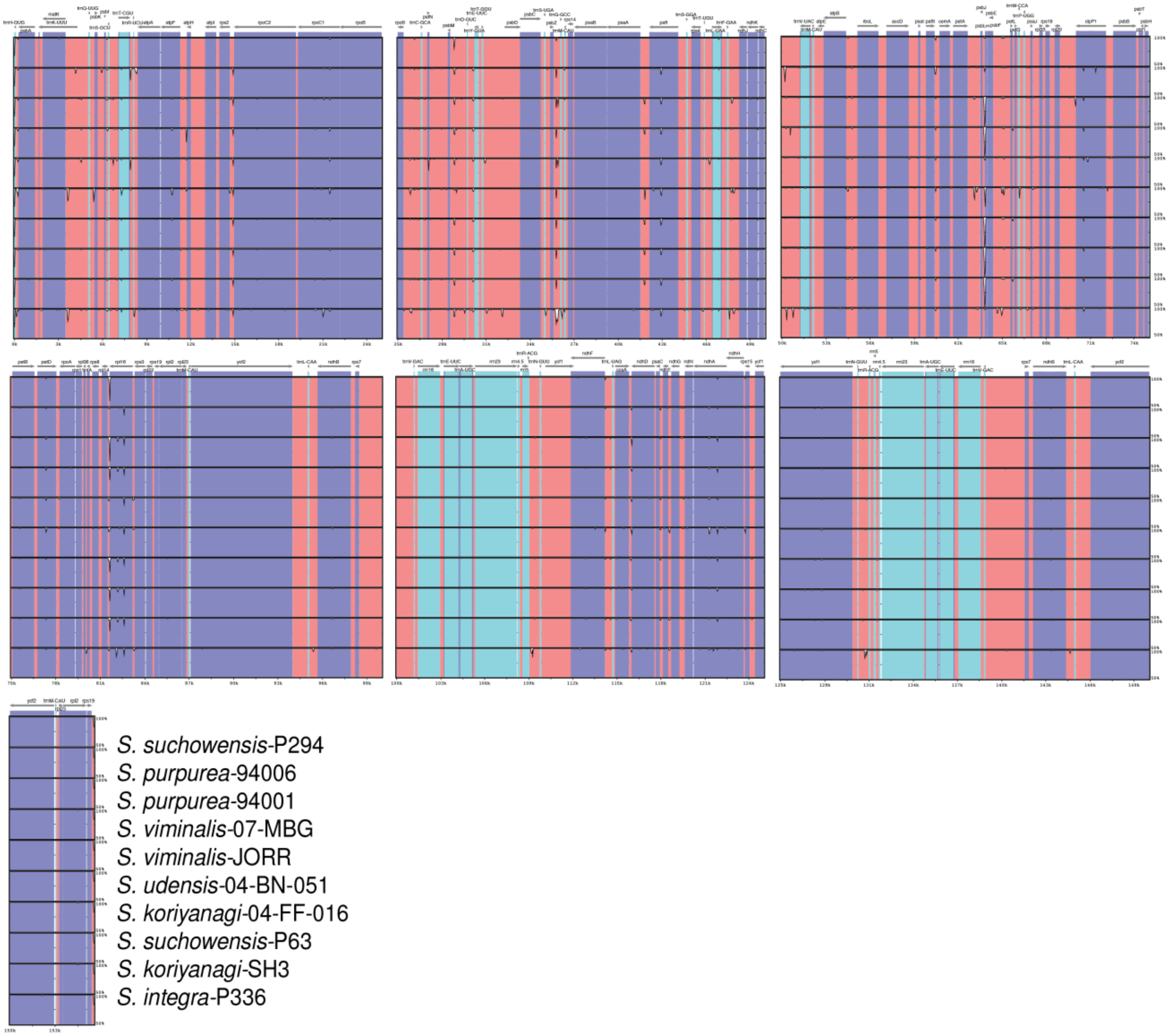

Figure IV-1. mVista plot describing structural variation in the 11 *Salix* chloroplast genomes, P295 (*S. suchowensis*) is used as a reference for alignment. Alignments were constructed in the LAGAN mode. Purple colored regions are genic, coral-colored regions are intergenic, and teal color regions represent RNA genes.

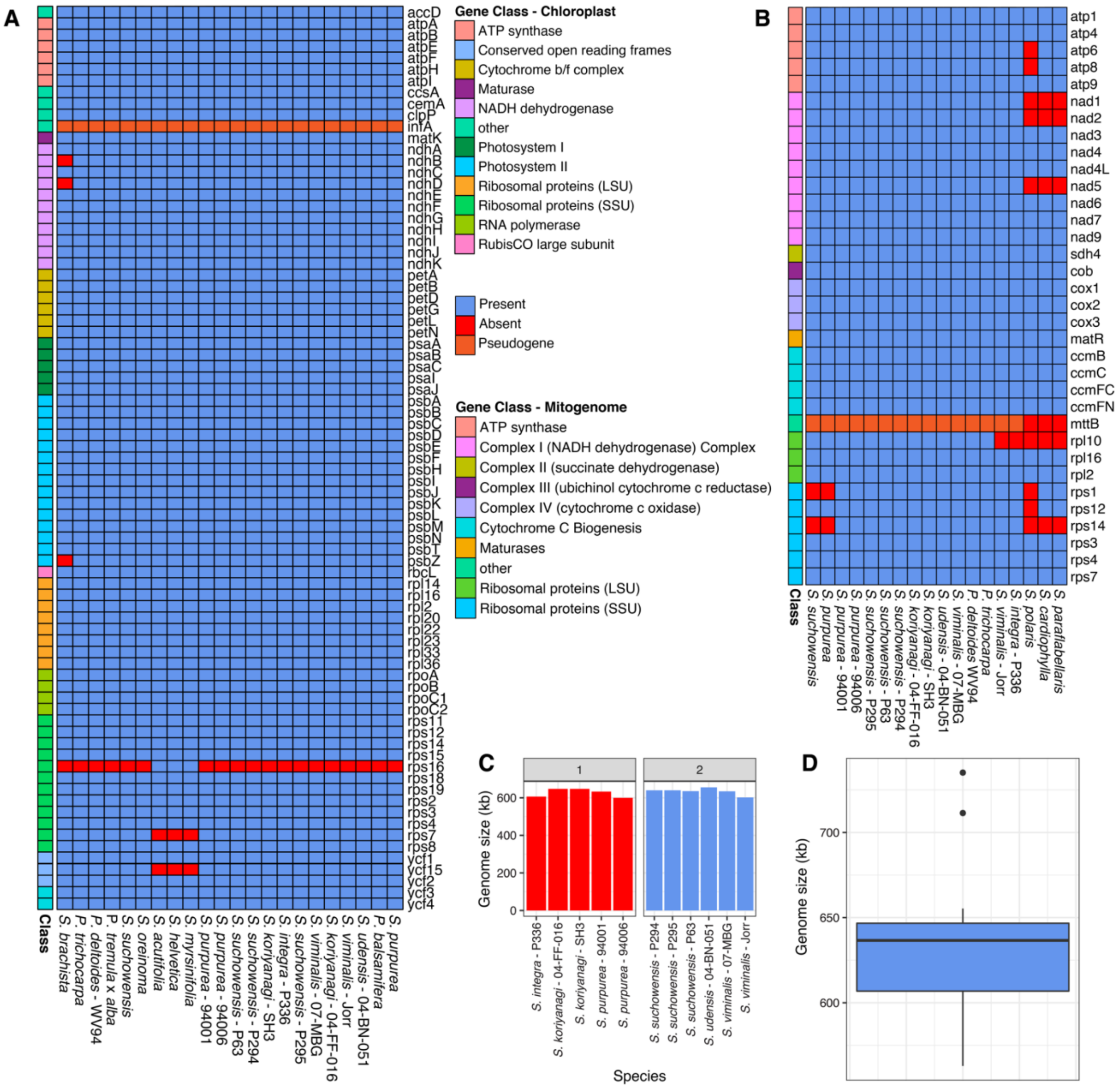

Figure IV-2. Gene presence and absence variation in the 11 *Salix* chloroplast and mitogenomes compared to publicly available *Salicaceae* organelle genomes. (A) Presence and absence matrix constructed from the 11 chloroplast genomes and eleven publicly available chloroplast genomes from the *Salicaceae*. (B) Presence and absence matrix constructed from the 11 mitogenomes and seven publicly available mitogenomes from the Salicaceae. (C) Mitogenome size, the facet labels indicate total genome number. (D) Mitogenome size distribution in *Salix* based on the 11 *Salix* mitogenomes and seven publicly available *Salix* mitogenomes.

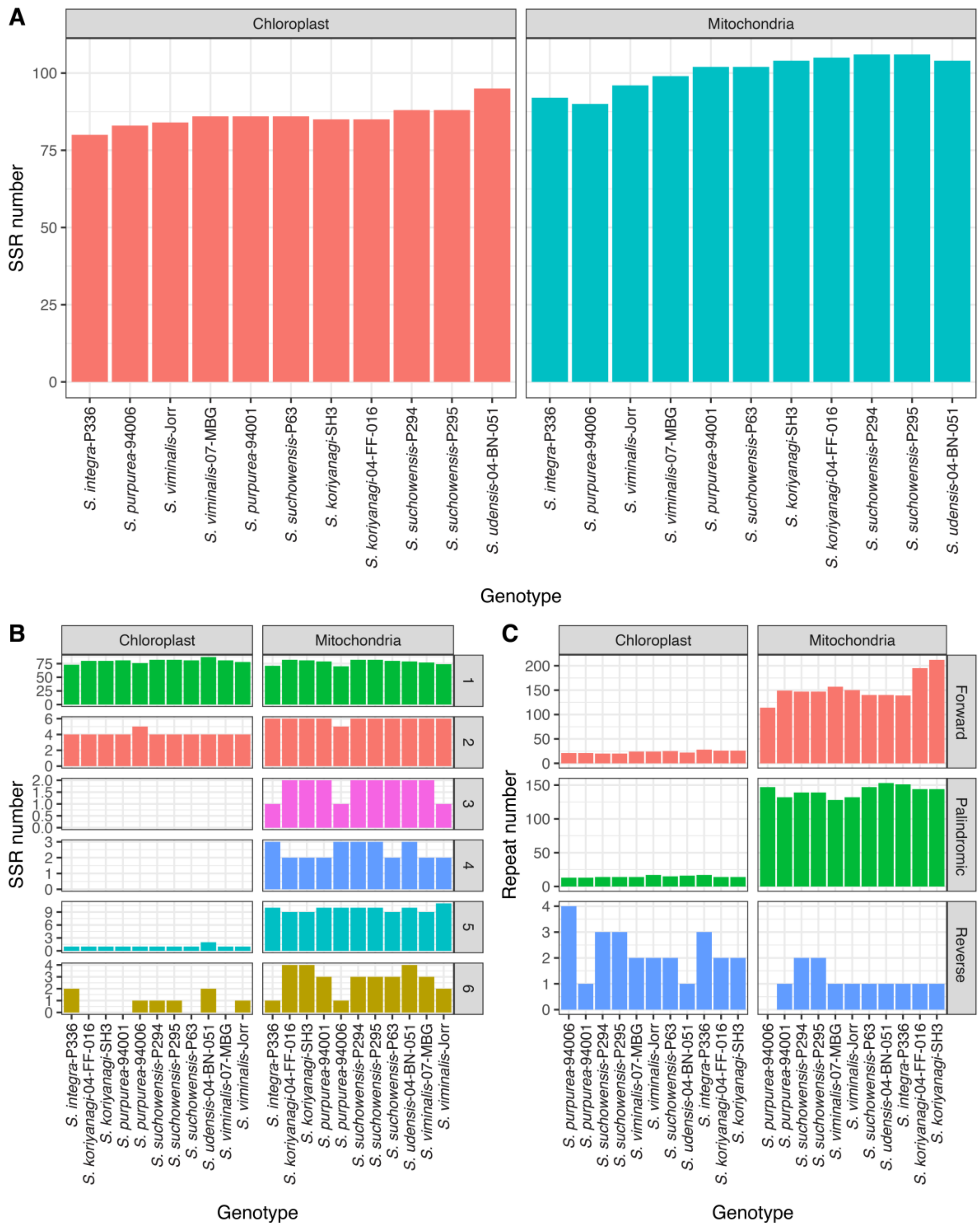

Figure IV-3. SSRs and dispersed repeats vary throughout the 11 *Salix* chloroplast and mitogenomes. (A) SSR count by genotype and organelle genome. (B) SSR count by unit size in the chloroplast and mitogenomes (C) Dispersed repeat counts (forward, palindromic, and reverse) by genotype in the chloroplast and mitogenomes.

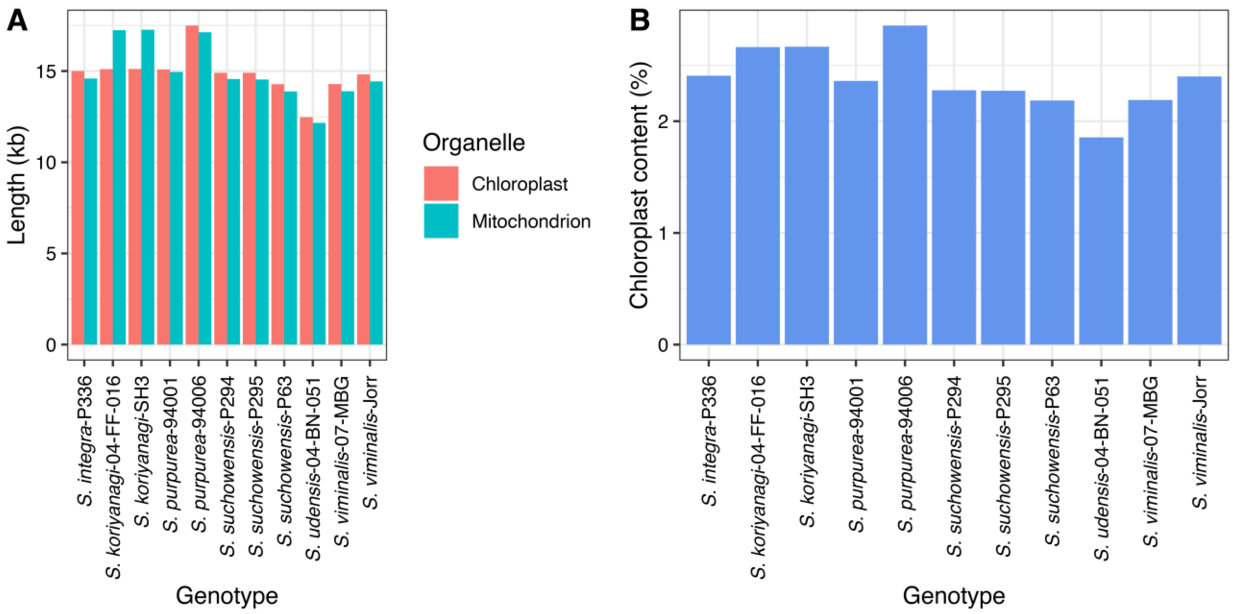

Figure IV-4. Chloroplast transfer to the mitogenomes is variable across the 11 *Salix* genotypes. (A) Total size of transfer in query (chloroplast) and subject (mitochondrion) perspectives. (B) Percentage of integrated chloroplast DNA content by genotype.

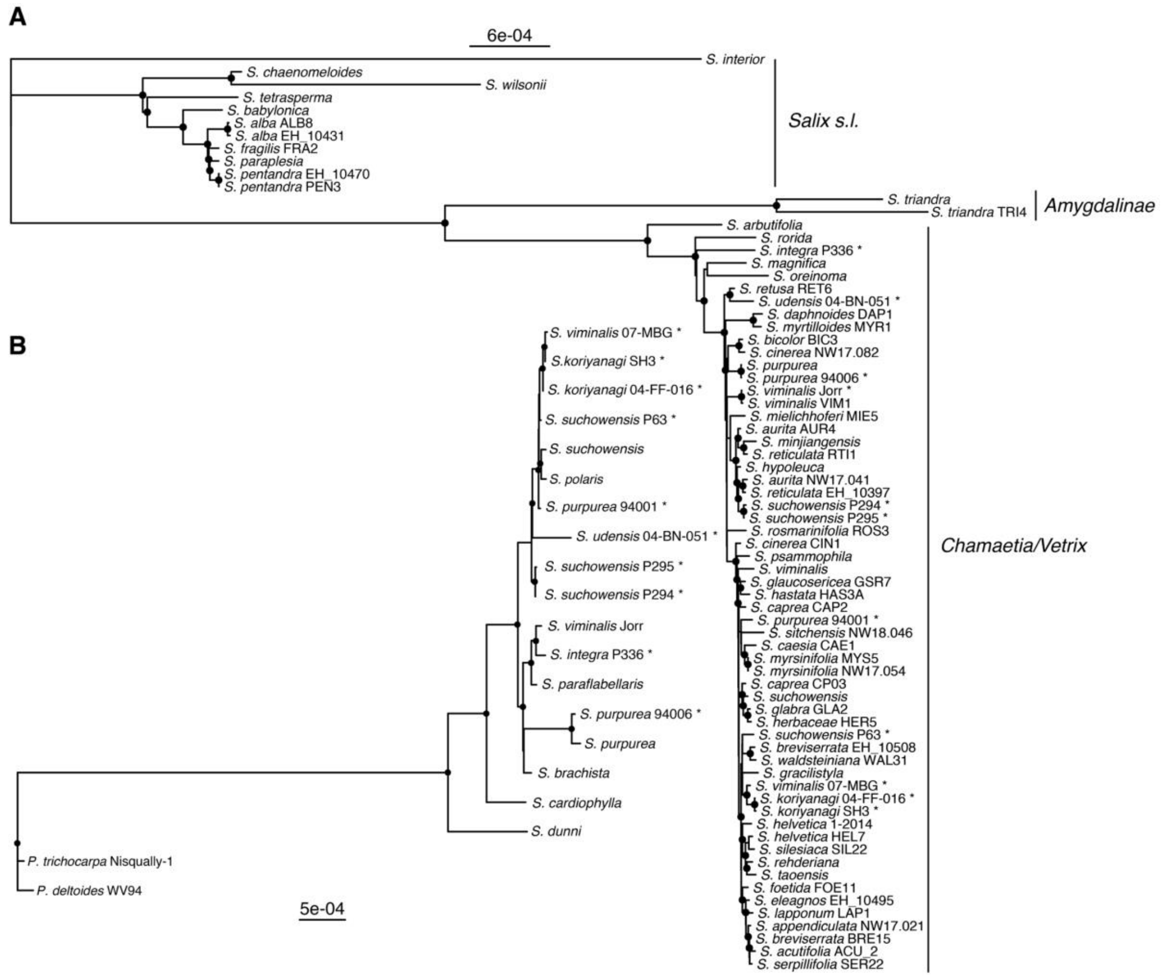

Figure IV-5. Phylogenomic analysis based on the Maximum-Likelihood (ML) method with publicly available *Salix* organelle genomes and the 11 organelle genomes. Black dots at nodes indicate a bootstrap support value greater than 90, and taxa labeled with an asterisk are genomes created in this study. (A) Chloroplast phylogeny based on the alignment of 72 whole plastomes minus one of the inverted repeats. *S. interior* was used as the outgroup. (B) Mitochondrion derived phylogeny with seven other publicly available *Salix* mitogenomes based on the alignment of LCBs. *P. trichocarpa* Nisqually-1 and *P. deltoides* WV94 were used as outgroups.

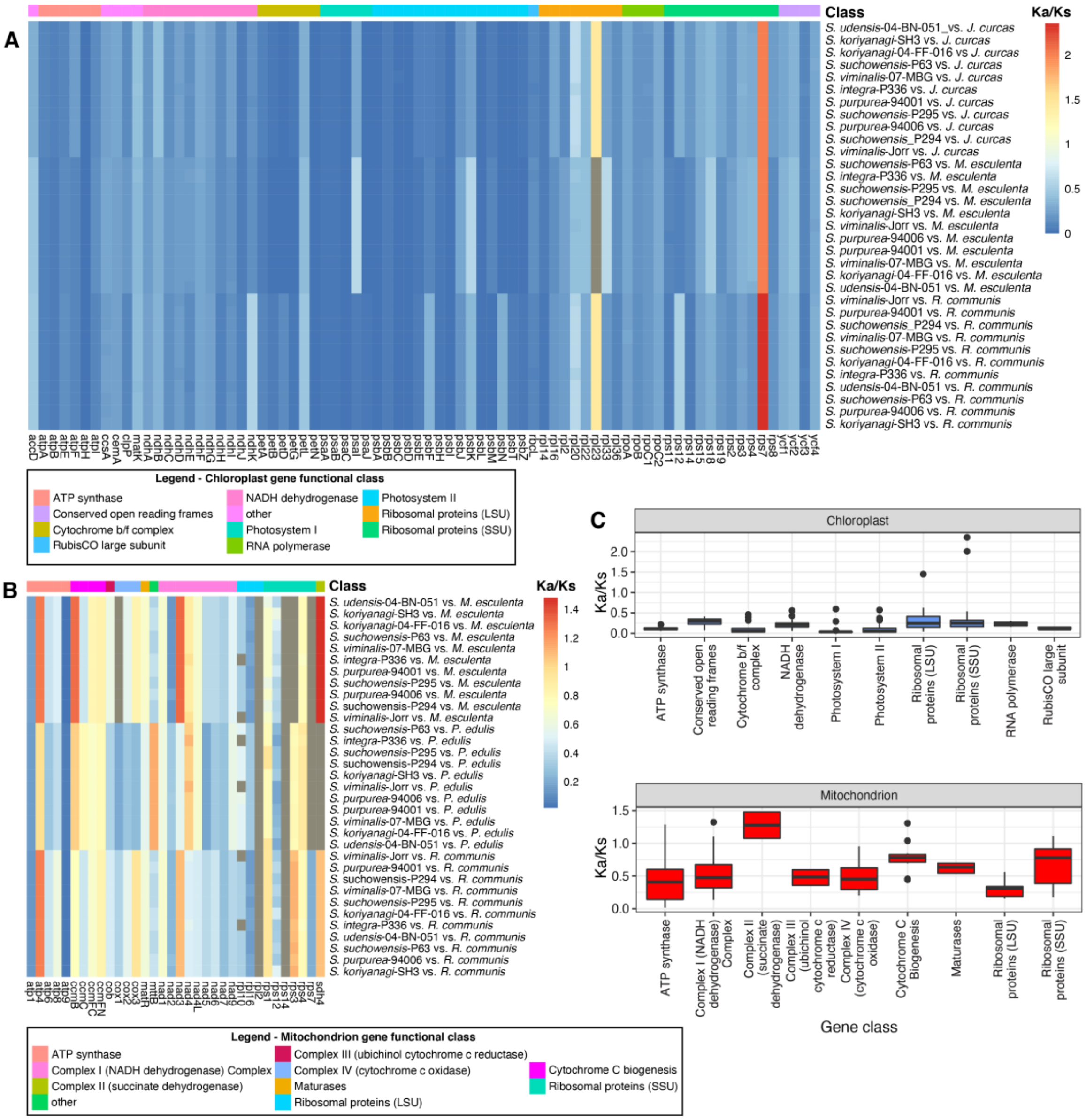

Figure IV-6. Ka/Ks values of the 11 *Salix* organelle protein-coding genes as determined through comparison with species in the order *Malpighiales*. (A) Ka/Ks values from the 11 chloroplast genomes derived from comparison with *J. curcas*, *M. esculenta*, and *R. communis*. (B) Ka/Ks values from the 11 chloroplast genomes derived from comparison with *M. esculenta*, *P. edulis*, and *R. communis*. (C) Ka/Ks values by gene functional class in the organelle genomes.

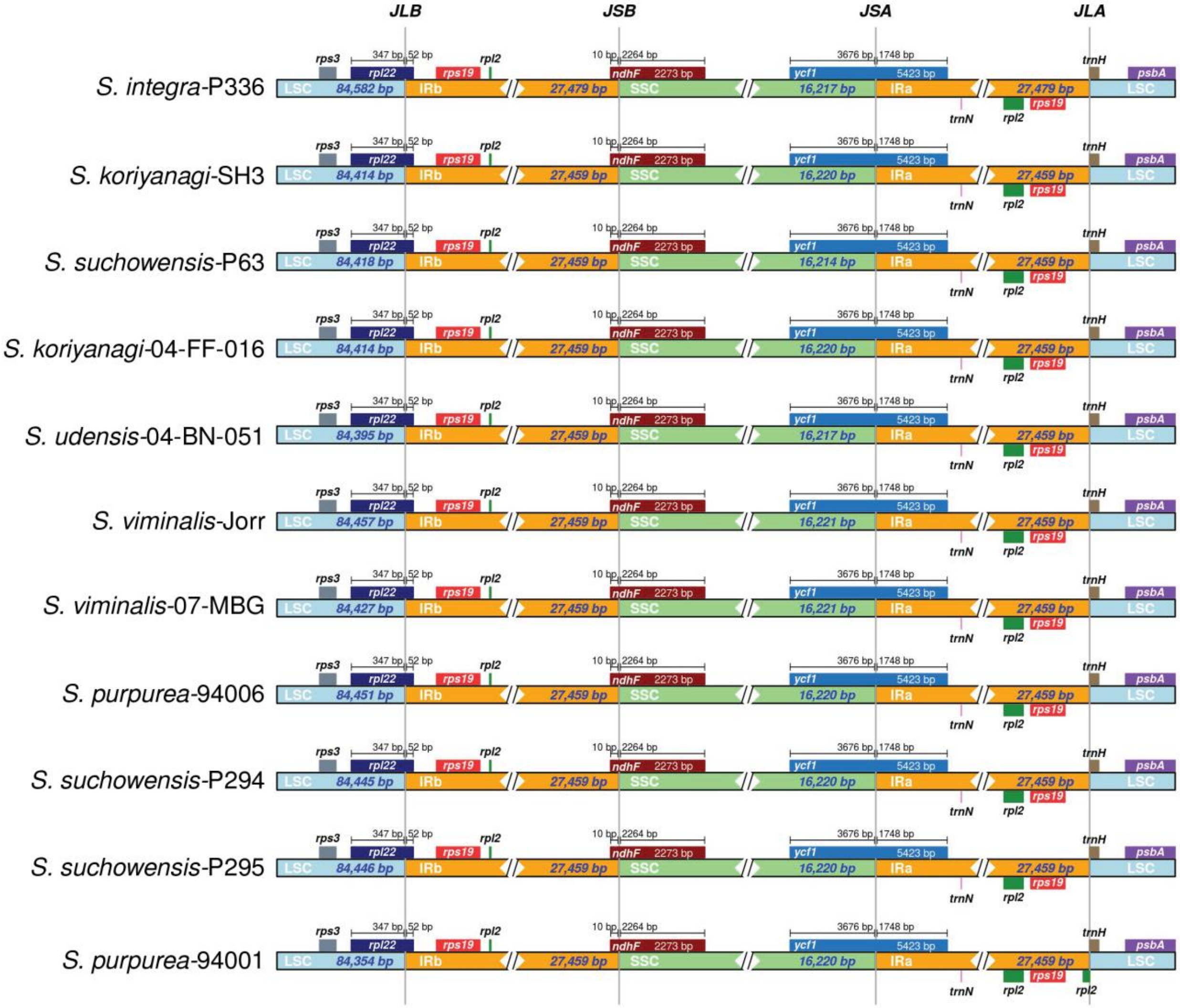

Figure IV-S1. Structural variation around IR junctions visualized with IRscope in the 11 *Salix* chloroplast genomes.

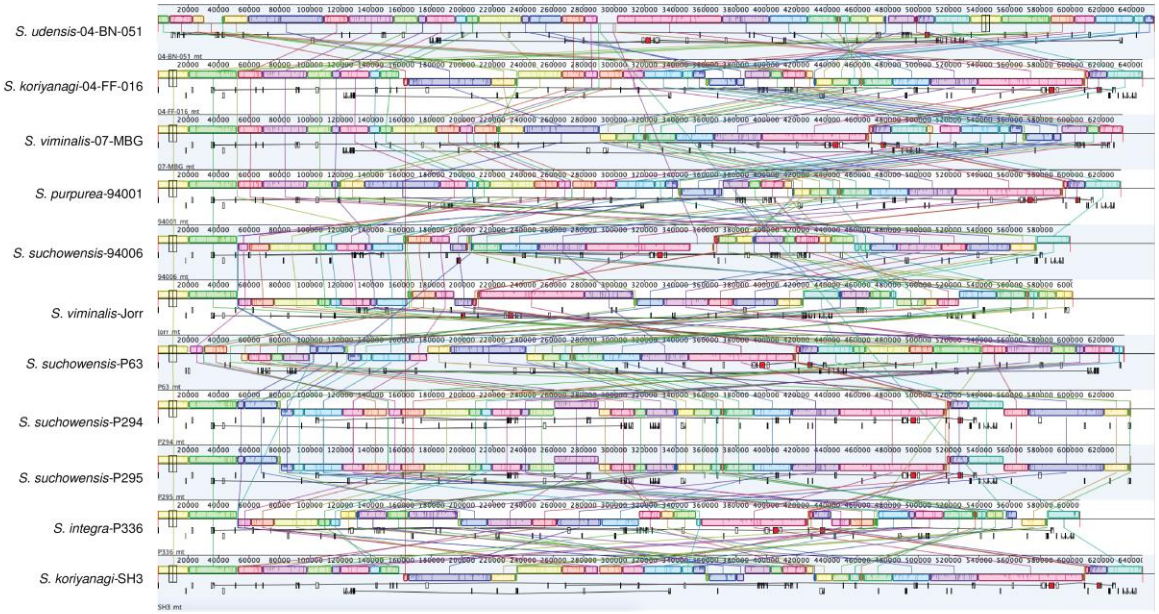

Figure IV-S2. progressiveMauve whole genome alignment depicting extensive rearrangement in the 11 *Salix* mitogenomes. Shared colors and links indicate shared locally collinear blocks (LCBs). Orientation of blocks are indicated by LCB orientation, blocks located below a genome's center line represent an inversion.

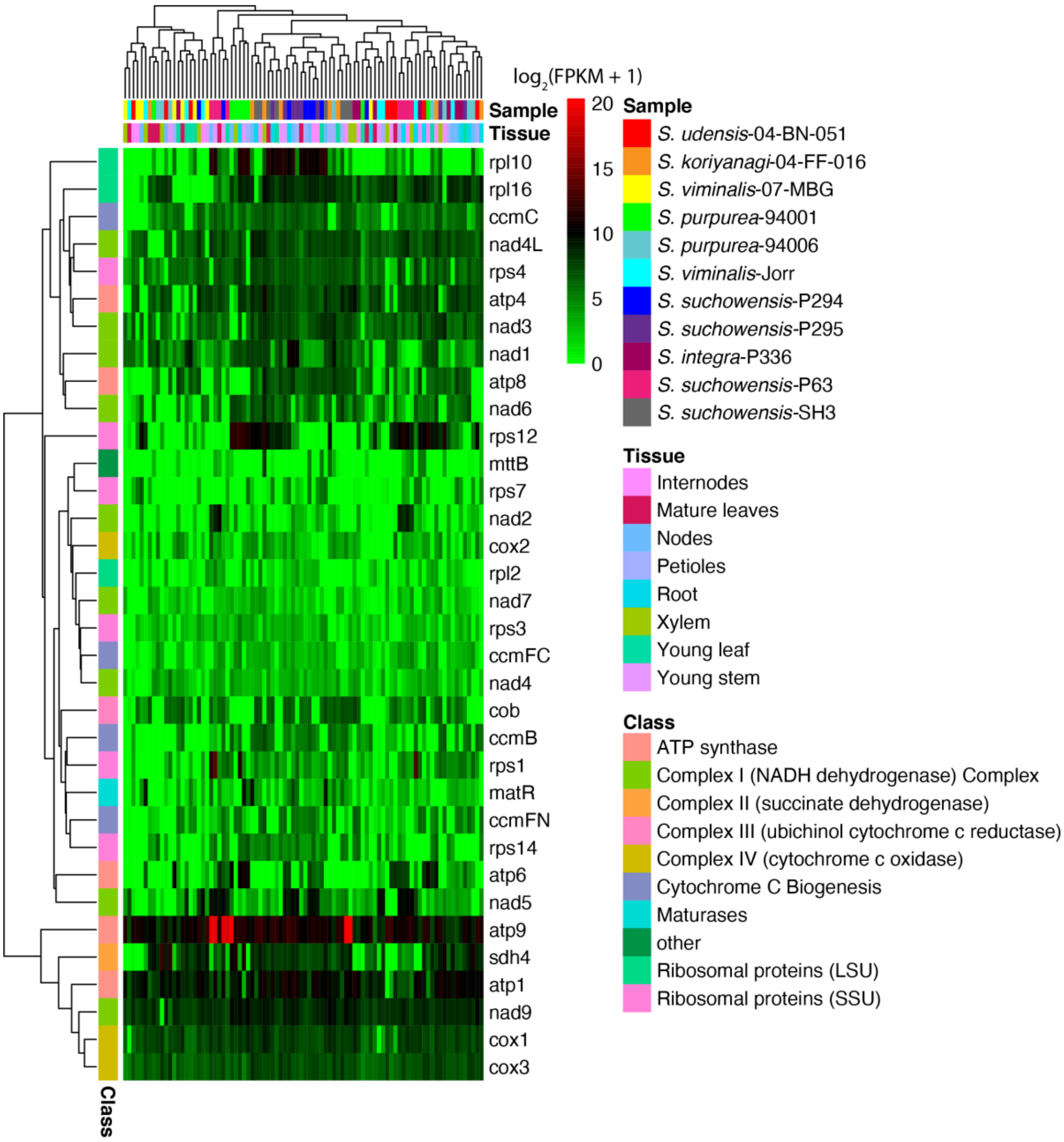

Figure IV-S3. Expression of protein-coding genes in the 11 *Salix* mitogenomes represented as $\log_2$(FPKM+1). Sample and tissue are depicted as column annotations while gene functional class are row annotations.

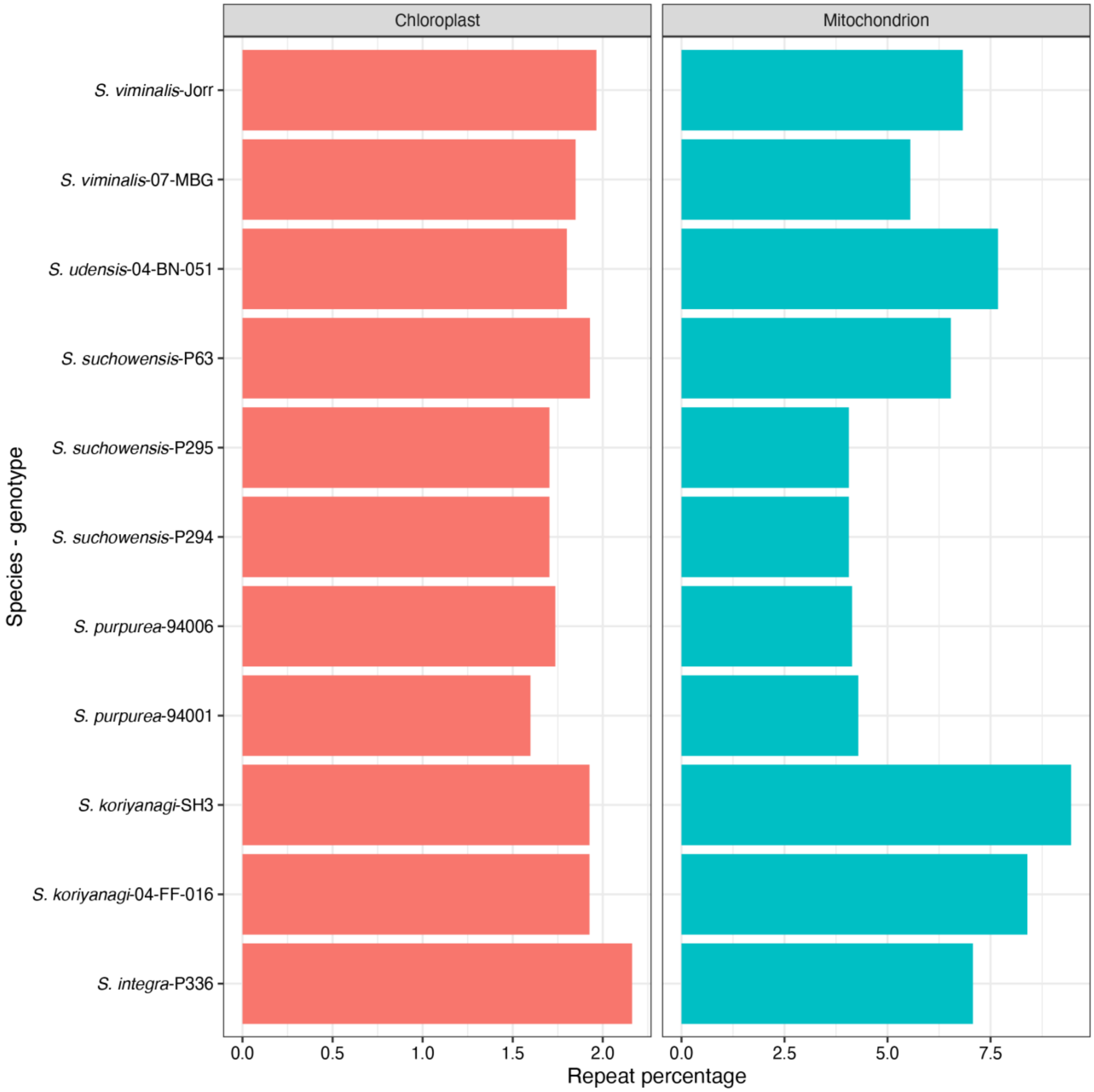

Figure IV-S4. Dispersed repeat percentage by genotype in the 11 *Salix* chloroplast and mitogenomes.

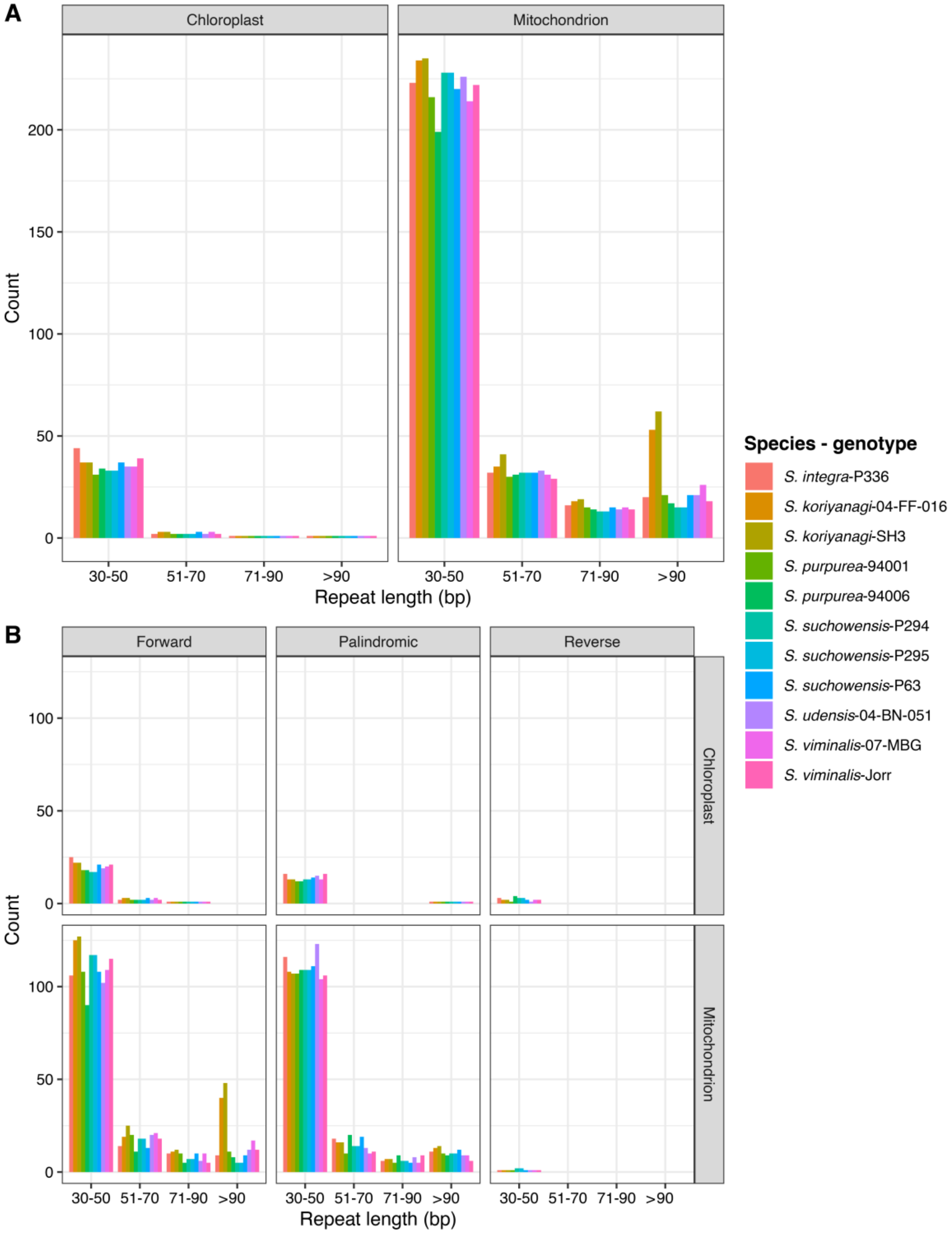

Figure IV-S5. (A) Dispersed repeat counts by size class (30-50 bp, 51-70 bp, 71-90 bp, and > 90 bp) by genotype in the 11 *Salix* chloroplast and mitogenomes. (B) Dispersed repeat counts by size class (30-50 bp, 51-70 bp, 71-90 bp, and > 90 bp) and context (forward, palindromic, and reverse) by genotype in the 11 *Salix* chloroplast and mitogenomes.

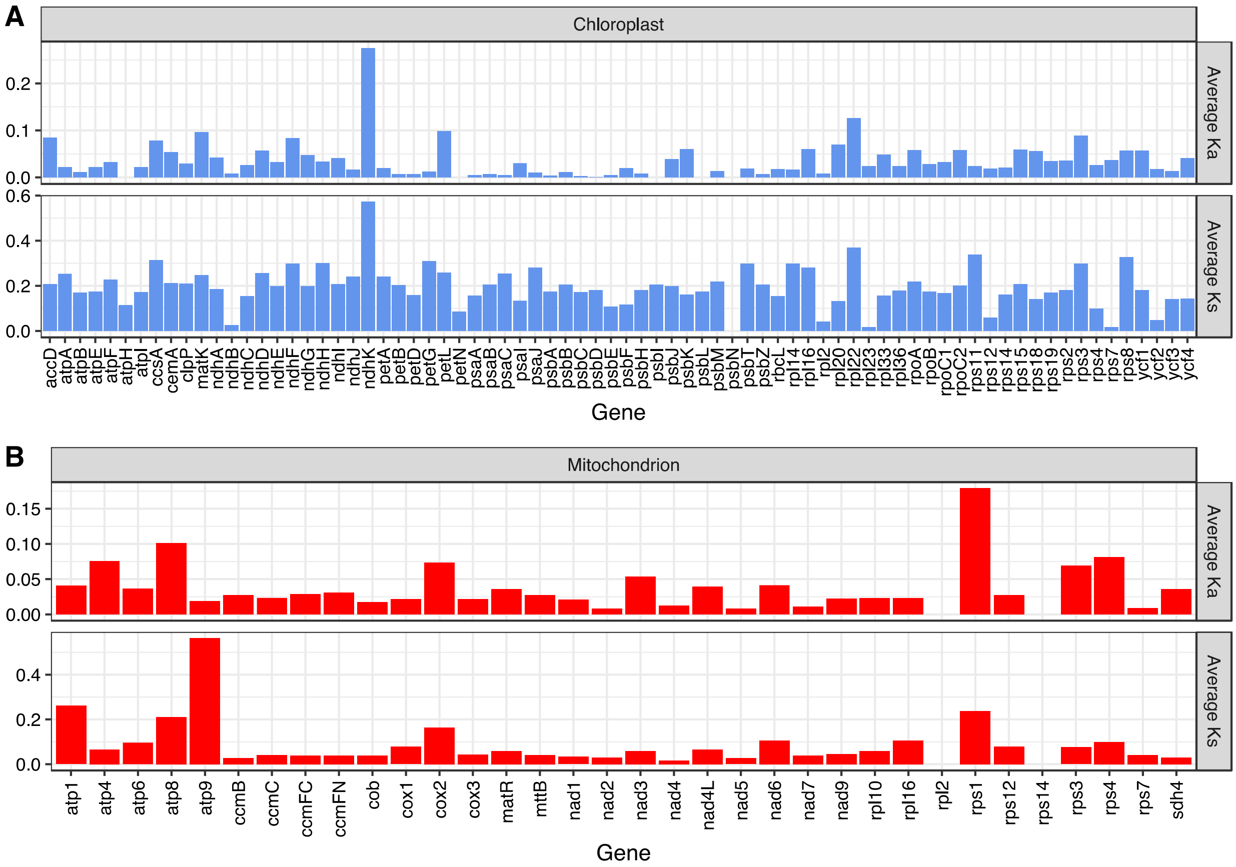

Figure IV-S6. Average Ka and Ks values by gene in the (A) chloroplast and (B) mitogenome across the 11 *Salix* genotypes.

Chapter V
Conclusion

CONCLUSION

The primary focus of this dissertation is to contribute to the understanding of genome evolution in the *Salicaceae* from both organelle and nuclear genome perspectives. Additionally, genomic resources generated here will aid future studies of the genetics and genomics of *Populus* and *Salix*. Ideally, these resources will contribute to their domestication and use as important biofuel and biomaterial feedstocks.

Chapter II leveraged the extensive omics data available to *P. trichocarpa* to generate a highly curated set of orphan genes. These analyses yielded information about their expression, polymorphism at the population scale, presence in regulatory networks, and their regulatory origins. As a result, a set of orphan genes with multiple layers of functional evidence was produced. Orphan genes which had evidence of *de novo* gene evolution were also identified and the highly conserved Salicoid whole genome duplication was used to reconstruct their origins from non-coding sequence. Future studies could explore this highly curated set of *P. trichocarpa de novo* and orphan genes in transgenic systems. Additionally, future gene annotation pipelines should be more inclusive of genes lacking homology in other species.

Chapter III explored non-coding and putative coding transfer of organelle derived sequence to the nuclear genomes of *P. trichocarpa* and *P. deltoides*. Methodology was developed to identify and curate these NUPTs and NUMTs and *P. deltoides* had considerably more organelle derived DNA content compared to *P. trichocarpa*. Future studies could examine this imbalance further. A presence absence variation pipeline which leveraged sequence identity and syntenic information was also developed and may be useful for future studies evaluating differences in NUPT and NUMT content in genome assemblies. Methylation levels and mutation rates were used to validate this pipeline. Lastly, we explored the whole chloroplast integration event which is shared between *P. trichocarpa* and *P. deltoides*. Future studies could explore the functionality of putative protein-coding transfer through transgenesis.

Chapter IV created new genome resources for six *Salix* species which included 11 genotypes and generated organelle assemblies and annotations. As of April 2022, there are only seven publicly available *Salix* mitogenomes. The 11 mitogenomes generated here contribute significantly to the total number of *Salix* mitogenomes and offer novel insights into *Salix* mitogenome diversity. Gene presence and absence was assessed within the chloroplast and mitogenomes and some heterogeneity was present. Additionally, variation in repeat content and chloroplast DNA content in the mitogenome was evident at the genotype and species levels. Lastly, selection analyses identified large numbers of genes under purifying selection and a smaller subset under positive selection. Future studies could investigate whether organelle genes which are under positive selection have functional copies in the nuclear genome.

Overall, this dissertation has developed new methods and investigated various facets of genome evolution in the *Salicaceae* and should aid future studies exploring the genetics and genomics of willow and poplar.

VITA

Timothy B. Yates was born in Vista, CA to Susan Lagasse and Brad Yates. He has one younger sister, Sydney Yates. Timothy grew up in Vista, CA, attended Rancho Buena Vista High School and graduated in 2012. He then enrolled at Loyola Marymount University where he met his fiancé Sonja Feck and obtained an undergraduate degree in biology and a minor in biochemistry in 2016. Timothy worked in the lab of Dr. Nancy Fujishige while at Loyola Marymount University. In 2016, Timothy began his doctoral studies in the Bredesen Center at the University of Tennessee, Knoxville. He worked in the lab of Dr. Wellington Muchero studying genome evolution in the *Salicaceae*, obtaining a Ph.D in May 2022.