**Ancient Introgression in *Populus***

**Bailey Kretzler, Sarah Morris, Hannah Shafer, Greta Savitsky**

**15 December 2021**

**Ecological Genomics Final Project**

**Abstract**

Adaptive introgression, or the transfer of genes through hybridization and subsequent natural selection for these genes, can lead to increased resistance to disturbance. These introductions of genetic diversity are especially important in a rapidly changing environment. We sought to identify ancient introgression events between sister *Populus* taxa using whole genome sequences of *P. trichocarpa*, *P. davidiana*, *P. ilicifolia*, *P. euphratica*, *P. alba*, *P. deltoides*, *P. simonii*, *P. tomentosa*, and the outgroup, *Salix purpurea*. To do this, we downloaded whole genome data from Genbank and Phytozome, aligned these sequences using MUMmer and found the subsequent Patterson’s D statistics based on all unique trios of the aligned sequences. There were 14 significant trio comparisons, many of which contained a possible introgressive event between *P. trichocarpa* and *P. ilicifolia*. While these two species’ current ranges do not overlap geographically, long-distance wind-dispersed seeds or historically overlapping ranges could have facilitated this ancient introgression. Further study of these data could yield knowledge of more introgression events (using gene tree discordance) or more information about introgression events, such as direction and timing.

**Introduction**

Recent focus on hybridization and introgression has shed light on the porosity of species boundaries. Historically, hybridization and recent introgression events have been relatively easy to detect. However, it was harder to detect ancient introgression until recently when whole genome sequencing became more cost-efficient and widely available (Hibbins & Hahn, 2021). Finding ancient adaptive introgression events, in which an introgressed loci is selected for over years of backcrossing, is especially important in rapidly changing environments, as it offers an alternative to relying on mutation or existing genetic diversity for adaptation (Hedrick, 2013; Suarez-Gonzalez, Lexer, & Cronk, 2018). Understanding ancient introgression events can add clarity to current genetic topologies, as well as informing our understanding of rapid adaptation to changing environments.

There are several genomic methods that can be used to detect ancient introgression, such as Patterson’s D statistic and gene tree discordance. Patterson’s D statistic and related measures like fd or dxy detect shared alleles between taxa to identify candidate regions for introgression (Suarez-Gonzalez, Lexer, & Cronk, 2018). These statistics quantify biallelic patterns of sister and nonsister taxa as a proxy for gene tree frequencies and then count the occurrence of “ABBA” and “BABA” in three sister taxa and one outgroup, where A is the ancestral allele and B is the derived allele. If the number of ABBA and BABA sites differs, asymmetrical gene tree topologies are inferred, and introgression has occurred (Hibbins & Hahn, 2021). Additionally, introgression events can be identified by creating gene trees and identifying areas of the species tree topology with discordance (Hibbins & Hahn, 2021). The more discordance observed, the more likely it is that introgression has occurred.

Trees of the genus *Populus* (Saliaceae) are of particular interest in studies about adaptive introgression. In 2006, the US Department of Energy (DOE) chose *P. trichocarpa* to be the first tree to have its whole genome sequenced, and since then it has become a robust resource with thorough annotations (Pespeni and Keller, “Population Genomics Day One”). This completed sequence spurred many to redefine the class of model organisms in botany to include *P. trichocarpa* (Jansson & Douglas, 2007). The inclusion of poplar trees in model-species research not only lends new insights into tree-specific traits, but also helps enrich our understanding of adaptive evolution since trees have slower molecular clocks (Jansson & Douglas, 2007). Indeed, the long generation times of *Populus* allow for a stronger signal of ancient introgressed genes in contemporary genome sequences.

*Populus* has been characterized as having rampant hybridization (Zhou et al, 2021), but the mode and tempo of these hybridization events remains unclear. Recent studies found evidence for contemporary clear-cut hybridization and backcrossing (Chhatre, Evans, DiFazio, & Keller, 2018), but what about farther back in the evolution of this clade? Can we observe a more admixed, or anciently introgressed, pattern in this genus? Using whole-genome sequences, we conducted a preliminary investigation into ancient introgression events in the genus *Populus*. Our analysis focused on Patterson’s D statistic with recommendations for future investigation.

**Methods**

*Biological background on species of interest*

Poplars are generally diecious species that favor areas of high moisture such as river banks and wetlands. However there are important differences in the geographic range and habitat of the species selected for our analysis.

*P. trichocarpa*

Also known as Balsam Poplar, *Populus trichocarpa* is the largest hardwood tree in the western United States. Its range extends from the Cook Inlet in Alaska to the mountains of southern California. They are normally diecious and seeds are disseminated from May-June (although they can reproduce vegetatively as well). They grow best in high humidity (Debell, n.d.).

*P. davidiana*

Korean Poplar is endemic to western Asia and is among the most ecologically and economically important tree species in China. It has a large geographic range and a high rate of intraspecific polymorphisms (Hou, Li, & Zhang, 2020).

*P. ilicifolia*

The only poplar species endemic to eastern Kenya, the Tana River Poplar is diecious and either semi-deciduous or evergreen. It commonly occurs along riverbanks and prefers moist soil and high humidity (Populus ilicifolia, n.d.).

*P. euphratica*

Also called Desert Poplar, *P. euphratica* inhabit the Tugia floodplain that spans northern Africa to western China. They are important components of riparian vegetation in the area, but their spatial extent varies with hydrologic events. Their population is currently suffering due to drought (Miao et al., 2020).

*P. alba*

*Populus alba*, or White Poplar, is endemic to central, southern, and eastern Europe, the Mediterranean islands, temperate Asia, and northern Africa. It was first reported in the United States as a planting in New England in 1785, and popped up in Michigan, Texas, North Dakota, and Montana over the next hundred years. Because stands are often all female, sexual reproduction is often limited to hybrids. Thus, the current range in the United States is highly dependent on human planting (Gucker, 2010).

*P. deltoides*

The Eastern Cottonwood has a range from western Canada south to Florida, Texas, Arizona, and northern Mexico and is often the primary species of floodplains and riparian vegetative stands. It requires periodic flooding. They are valuable both for lumber and the critical habitat they provide, and they are diecious and drought tolerant (Taylor, 2001).

*P. simonii*

Simon’s poplar is endemic to northern China and tolerant of nutrient deficient and drought. They are fast growing and deciduous. They are important reseeding plants for recently disturbed areas (Zhou et al., 2010).

*P. tomentosa*

*Populus tomentosa* is one of the most economically important trees in China, where it is found endemically. This species also provides protective habitat, especially along the Yellow River (Du et al., 2012).

*S. purpurea*

*S. purpurea*, or shrub willow, is a commonly grown energy crop. They are fast-growing, high-yielding, genetically diverse, and pest-resistant. The *S. purpurea* genome is publicly available, and this species also often exhibits heterosis (Carlson et al., 2015).

*Genome Alignment and Data Cleaning*

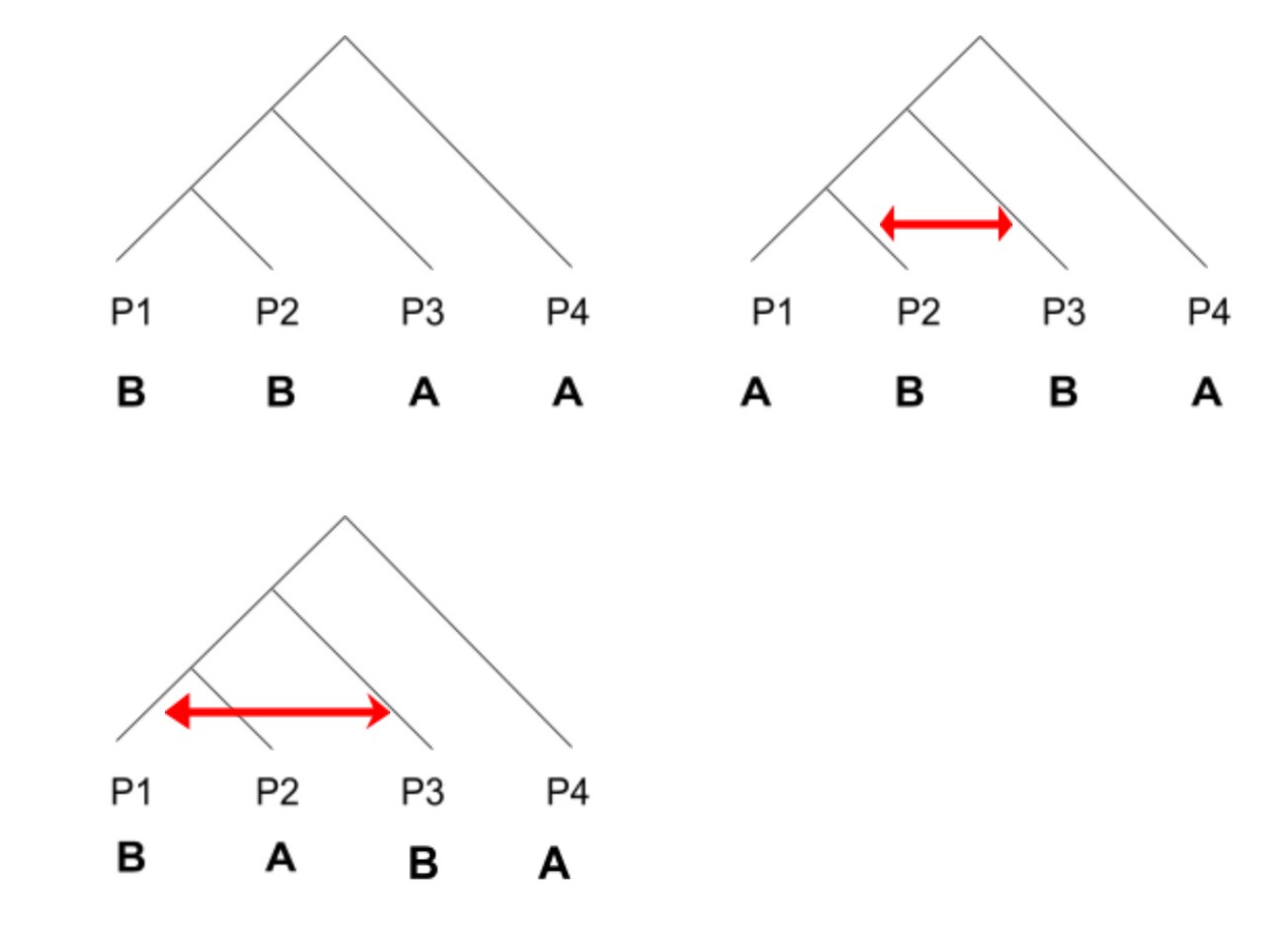
Whole genome sequences of *P. trichocarpa*, *P. davidiana*, *P. ilicifolia*, *P. euphratica*, *P. alba*, *P. deltoides*, *P. simonii*, *P. tomentosa*, and *S. purpurea* were downloaded from GenBank as zipped FASTA files and installed onto the PBIO381 server for remote access and greater computing power (Table 1). *P. trichocarpa* was selected as the reference genome as it is both well-annotated and well-studied. All sequences were first aligned to the *P. trichocarpa* genome using the nucmer function of MUMmer 4, and output to SAM file format (Marçais, Delcher, Coston, Salzburg, & Zimin, 2018). SAM file headers were reformatted using sed and then each file was converted to BAM format using ‘samtools view’ (Danecek et.al 2021). The resulting BAM files were sorted using ‘sambamba’ (Tarasov et.al. 2015). Sorted BAM files were then converted to a single vcf format using ‘samtools mpileup’ to compress the 9 files and ‘bcftools’ to convert to vcf (Danecek et.al. 2021). The output vcf file contained the genotypes for all individuals at each variable SNP location for each genome, which was then filtered to contain only the highest confidence loci.

***Table 1*** - Genome sequencing information for *Populus* species including genomic coverage and total sequence length.

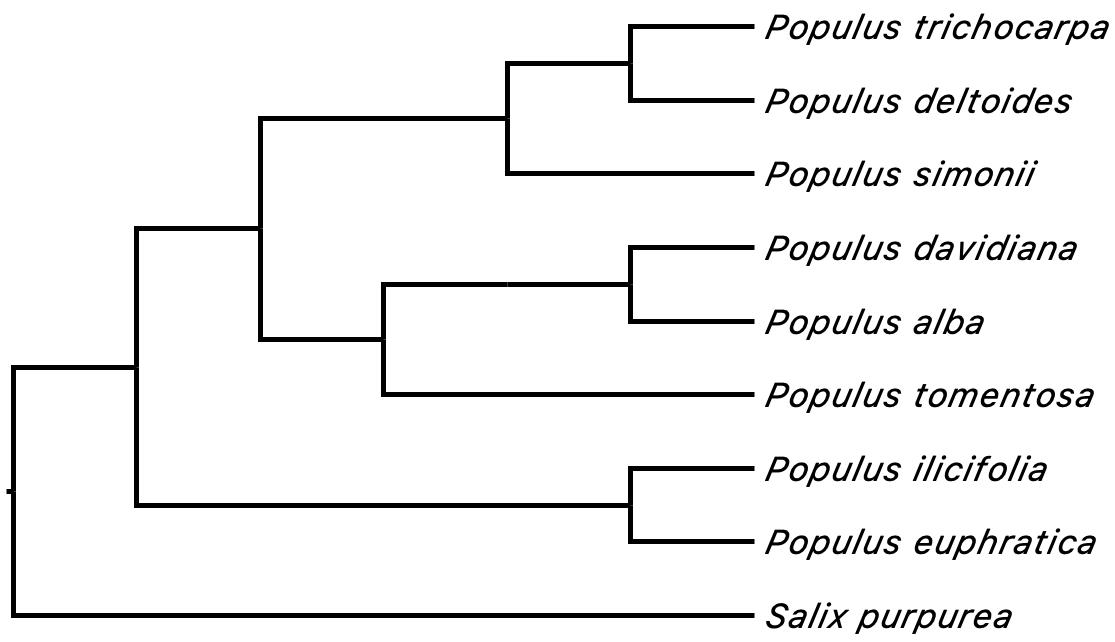
| Species | Genomic coverage | Total sequence length |
| --- | --- | --- |
| *P. trichocarpa* | 8x | 434,289,848 |
| *P. alba* | 130.17x | 416,985,780 |
| *P. davidiana* | 300.0x | 417,659,603 |
| *P. deltoides* | 100.0x | 428,270,921 |
| *P. euphratica* | 115x | 496,032,534 |
| *P. ilicifolia* | 266.0x | 399,016,331 |
| *P. tomentosa* | 75.0x | 739,803,048 |

*Calculating the D Statistic*

The filtered vcf file was used to calculate the Patterson's D statistic and the f4 ratio via ‘Dsuite Dtrios’ (Malinsky, Matschiner, & Svardal, 2021). The D statistic and f4 ratio serve as indicators of introgression and gene flow between a trio of species rooted in the outgroup. Dsuite determines the D statistic by inferring patterns of allele sharing (BBAA, ABBA, or BABA) at all loci for a given trio. The BBAA pattern is assumed to be normal, with P1 and P2 sharing the derived B allele and P3 and the outgroup sharing the ancestral A allele (Figure 1a). ABBA and BABA patterns indicate introgression between P2 and P3 and P1 and P3 respectively (Figure 1b,c). The d statistic is then calculated by the difference in the number of ABBA patterns vs the number of BABA patterns divided by the total number of BABA patterns (Malinsky, Matschiner, & Svardal, 2021). We specified a phylogeny to Dsuite in order for the program to account for the known relationships between species. We used the Poplar-Salix phylogeny identified by Olson et.al. (*in review*), as well as a chloroplast transcriptome phylogeny from Zhou et al. (2021) to place *P. tomentosa* in the tree. The resulting Newick tree of our eight *Populus* species of interest and outgroup were visualized in FigTree v1.4.4 (Rambaut, 2009) (Fig. 2). The file generated by Dsuite provided the D statistic, f4 ratio, z score, p-value, and count of BBAA, ABBA, and BABA patterns at the shared loci for all possible trios. This file was transferred to a local machine and checked for significant D statistics in R (R Core Team, 2021).



**Figure 1.** Allele sharing patterns used to calculate the D statistic where a) shows the expected BBAA pattern, b) shows the ABBA pattern - indicating introgression between P2 and P3 per the red arrows and c) shows the BABA pattern indicating introgression between P1 and P3 per the red arrows.

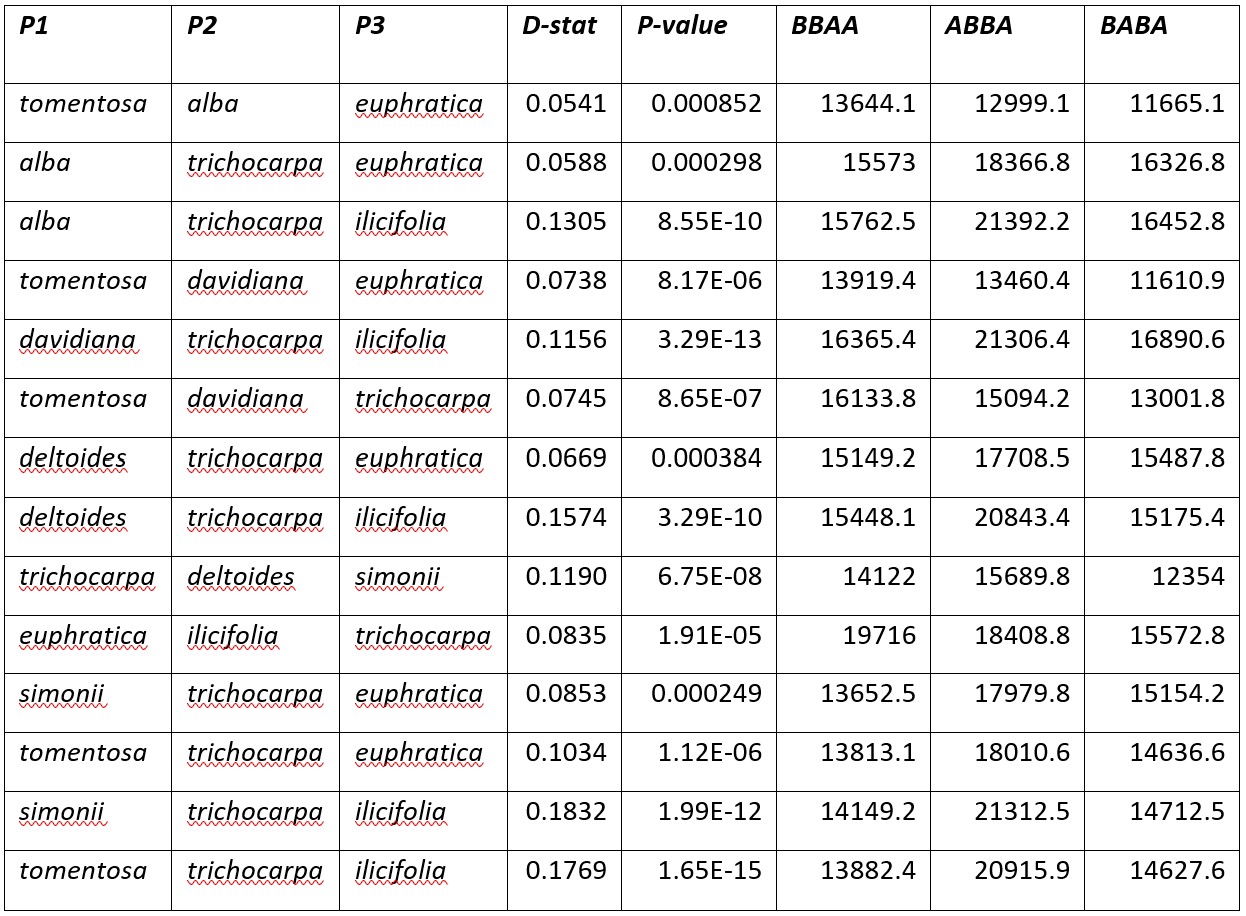
****

**Figure 2.** Species tree indicating phylogenetic relationships between all 9 species of interest, derived from Olson et.al (*in review*) and Zhou et al. (2021).

**Results**

Out of the 56 trios examined using Dsuite, 14 showed significant d statistics. The shared loci and amongst all trios and significant D statistics ranged from 38,308 to 54,562 and 0.054 to 0.183 respectively (Table 2). ABBA was the dominant allele sharing pattern for all trios. The most shared loci were amongst the trio containing *P. davidiana P. trichocarpa P. ilicifolia*, which also showed a higher relative D statistic at 0.115 in comparison to other trios. The trio containing *P. simonii* *P. trichocarpa* *P. ilicifolia* had the highest D statistic at 0.183, suggesting high levels of allele sharing between these species (Table 2). The ABBA pattern was considerably more dominant than the BABA pattern indicating more shared alleles between *P. trichocarpa* and *P. ilicifolia.* Further, when looking at the trios with D statistics greater than 0.1, five out of seven contained  *P. trichocarpa* and *P. ilicifolia* as the P2 and P3 species.The other two trios*, P. tomentosa P. trichocarpa P. euphratica and P. trichocarpa P. deltoides P. simonii,* have a D statistic of 0.119 and 0.103 respectively and favor the ABBA pattern of allele sharing (Table 2).

***Table 2 -*** Significant D Statistics and Allele Sharing Patterns. The D statistic, P-value, and portion of BBAA, ABBA, and BABA allele sharing patterns across all shared loci amongst trios with significant d statistics.



**Discussion**

Six out of the 14 significant D statistics involved *P. trichocarpa* and *P. ilicifolia.* Given their disjunct distribution (*P. trichocarpa* is in North America and *P. ilicifolia* is in Africa), as well as their distant evolutionary relationship (Fig. 2), introgression between these species seems unlikely. Zhou et al. estimated the most recent common ancestor (MRCA) of *P. trichocarpa* and *P. ilicifolia* at 21.77 MYA (2021), in the early Miocene, when the Earth was generally warmer. It is possible that these two species’ ranges were historically more widespread than what we observe today due to a warmer climate. Additionally, pollen of poplar species are wind-dispersed, so a series of long distance pollen spread events may have occurred. Long distance dispersal has been observed in both orchids and ferns (Barrington, 1993; Sundue et al., 2014; Givnish et al., 2015; Barrington, 2020), making it a plausible explanation in poplar. Further, a historically broader geographic range may have placed *P.ilicifolia* on the western edge of Africa shortening the dispersal distance, making long distance dispersal all the more possible. Alternatively, the historical distribution of these two species could have been much broader during the Miocene. This could have been followed by range contraction and extinction events, which reduced their distribution to what we observe presently.

Another interesting result is the significant D statistic for the triplet including *P. trichocarpa, P. deltoides* and *P. simonii*. These three species form a clade in the species tree and are frequently hybridized in cultivation (Yang et al., 2021). While *P. trichocarpa* and *P. deltoides* have regions of range overlap in North America, *P. simonii* is native to China. Additionally, the MRCA for *P. trichocarpa* and *P. deltoides* is only a few MYA, but the MRCA for all three species is approximately 21.77 MYA (Zhou et al., 2021). This disjunct distribution pattern between east Asia and eastern North America is quite common amongst closely related plant groups; in fact, some 65 flowering plant genera have similar distribution patterns (Wen, 1999). Most of these disjunct distributions are attributed to the broader ranges of these lineages in the Cretaceous and the Bering land bridge allowing for movement between the two continents; the global cooling that followed these range expansions resulted in the range retraction of these lineages, leading to the distributions we observe today. Given the historical biogeographic precedent for the distribution patterns of these three species, ancient introgression events seem likely to have occured. However, it is possible that limitations in the genome alignment and deployment of the D statistic may have also resulted in the observed pattern.

One potential limitation is the use of *P. trichocarpa* as the reference genome during alignment. While *P .trichocarpa* is a commonly used reference due to its high-quality annotation, there are potential alignment biases that arise from this choice. Alignment biases occur as the chromosome structures for all other genomes are assumed to be that of the reference, *P. trichocarpa*. However, there may be differences in chromosome structure or loci that simply do not align well to the reference that end up being ignored. Many of our Dsuite trios with significant D statistics included *P. trichocarpa*, which may be due to the fact that each genome was aligned to *P. trichocarpa,* creating an inherent structural relationship to the reference. Recently, the concept of reference flow has been proposed which utilizes multiple references to determine the best alignment for a sequence (Chen et.al. 2021). The reference flow method should be explored in future studies. In addition to alignment bias, there are also taxonomic conflicts that impact the phylogeny of the *Populus*  genus.

Zhou et al. found 466 unresolved taxa in this genus (2021), and this taxonomic conflict only adds to the difficulty in estimating introgression. If *Populus* species are not well defined, any inference of their admixture is clouded by that uncertainty. However, the idea of discrete species is a debated topic, as genetic and phenotypic qualities lie on a spectrum rather than in distinct categories. Thus, species categorization, while sometimes a helpful tool for biological analysis, is arbitrary (Willis, 2017). Identification of introgression can still be helpful, though, in understanding certain traits, genes, or even alleles that correspond to adaptive attributes and changes.

We have identified potential points of ancient introgression in *Populus*. Specifically of interest is the evidence of allele sharing between *P. trichocarpa* and *P. ilicifolia* which share no geographic overlap in the present day. Here the D statistic is a useful starting point for identifying gene flow between species such as these, however there are other analyses that would be useful in assessing ancient introgression. Using a subset of informative loci could help assess gene tree discordance and potentially reveal introgression events. Using gene trees, we can further characterize the introgression events. For instance, differences in branch length helps determine the main direction of introgression between non-sister taxa (Hibbins, et al., 2021). Understanding more about introgression in *Populus* can shed light on how individuals might respond, or be better able to respond, to climate change and associated disturbances, like fire, flood, drought, or other loss or degradation of habitat and resources.

**Conclusion**

We used the publicly available whole genome sequence data of eight *Populus* taxa to identify ancient introgression events. With MUMmer and D-Suite, we aligned and analyzed these species, using *Salix purpurea* as an outgroup. We saw evidence for several possible ancient introgression events among the eight species of *Populus* that we based our analysis on, many including *P. trichocarpa* and *P. ilicifolia*, two currently geographically distinct species. While several factors could have contributed to these findings, such as aligning to one of our species of study or unclear lines between *Populus* species, there are also potential explanations for this event, like long-distance dispersal or historically overlapping ranges. Understanding ancient introgression events among *Populus* species can help us understand how they will respond to climate change.

**References**

Barrington DS. 1993. Ecological and historical factors in fern biogeography. *Journal of Biogeography* 20: 275–279.

Barrington DS. (2020). The Biogeography of Polyploid Ferns across Space and Time. *American Fern Journal* 110: 233–254.

Carlson, Craig H., et al. (2017). Dominance and Sexual Dimorphism Pervade the Salix Purpurea L. Transcriptome.” *Genome Biology and Evolution*, *9*, 2377–2394. https://doi.org/10.1093/gbe/evx174.

Chen, N., Solomon, B., Mun, T., Iyer, S., & Langmead, B. (2021). Reference flow: Reducing reference bias using multiple population genomes. Genome Biology, 22(1), 8.

Chhatre, V. E., Evans, L. M., DiFazio, S. P., & Keller, S. R. (2018). Adaptive introgression and maintenance of a trispecies hybrid complex in range-edge populations of Populus. *Molecular Ecology*, *27*, 4820-4838. DOI: 10.1111/mec.14820

Danecek, P., Bonfield, J., Liddle, J., Marshall, J., Ohan, V., Pollard, M., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience,* *10*, 1-4.

DeBell, Dean S. “Black Cottonwood.” *Populus Trichocarpa to RR*, USDA, https://www.srs.fs.usda.gov/pubs/misc/ag\_654/volume\_2/populus/trichocarpa.htm.

Du, Qingzhang, et al. (2012). Genetic Diversity and Population Structure of Chinese White Poplar (Populus Tomentosa) Revealed by SSR Markers.” *Journal of Heredity*, *103*, 853–862. https://doi.org/10.1093/jhered/ess061.

Givnish TJ, Spalink D, Ames M, Lyon SP, Hunter SJ, Zuluaga A, Iles WJD, Clements MA, Arroyo MTK, Leebens-Mack J, *et al.* (2015). Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proceedings of the Royal Society B: Biological Sciences* 282: 20151553.

Gucker, Corey L. 2010. Populus alba and hybrids. In: Fire Effects Information System, [Online]. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory (Producer). Available: https://www.fs.fed.us/database/feis/plants/tree/popspp/all.html [2021, December 12].

Hedrick, P. W. (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *2218*, 4606–4618.

Hou, Zhe, et al. (2020). Genetic Architecture, Demographic History, and Genomic Differentiation of Populus Davidiana Revealed by Whole-Genome Resequencing. *Evolutionary Applications*, *13,* 2582–2596. https://doi.org/10.1111/eva.13046.

Hibbins, M. S. & Hahn, M. W. (2021). Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, *0*, 1-17. https://doi.org/10.1093/genetics/iyab173

Jansson, S. & Douglas, C. J. (2007). Populus: A model system for plant biology. *Annual Review of Plant Biology*, *58*, 435-458. doi:10.1146/annurev.arplant.58.032806.103956

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite - fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*, 584–595. doi: <https://doi.org/10.1111/1755-0998.13265>

Miao, N., Jiao, P., Tao, W. *et al.* (2020). Structural dynamics of *Populus euphratica* forests in different stages in the upper reaches of the Tarim River in China. *Sci Rep,* *10*, 3196. https://doi.org/10.1038/s41598-020-60139-7

Marçais, G., Delcher, A. L., Coston, R., Salzburg, S. L., Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*, e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Olson, M., et al. (in review). Seeing the forest and the trees: Pylogenomics resolves taxonomic and differential diversification patterns in Populus and Salix. *Systematic Biology*.

Pespeni, M. & Keller, S. Population Genomics Day One. Github.io, https://pespenilab.github.io/ Ecological-Genomics/2021-10-18\_PopGenomics\_Day1.html. Accessed 6 Dec. 2021.

“Populus Ilicifolia.” *Useful Trees*, World Agroforestry, http://apps.worldagroforestry.org/usefultrees/pdflib/Populus\_ilicifolia\_KEN.pdf.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rambaut, A. (2009). FigTree v1.4.4. Computer Program. Available online at: https://github.com/rambaut/figtree/ (Accessed December, 2021).

Suarez-Gonzalez A., Lexer C., & Cronk Q. C. B. (2018). Adaptive introgression: a plant perspective. *Biology Letters*, *14*, 20170688. http://dx.doi.org/10.1098/rsbl.2017.0688

Sundue MA, Parris BS, Ranker TA, Smith AR, Fujimoto EL, Zamora-Crosby D, Morden CW, Chiou WL, Chen CW, Rouhan G, *et al.* (2014). Global phylogeny and biogeography of grammitid ferns (Polypodiaceae). *Molecular Phylogenetics and Evolution* 81: 195–206.

Tarasov, A., Vilella, A., Cuppen, E., Nijman, I., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats*. Bioinformatics (Oxford, England), 31(12)*, 2032-2034.

Taylor, J. L. (2001). Populus deltoides. In: Fire Effects Information System, [Online]. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory (Producer). Available: https://www.fs.fed.us/database/feis/plants/tree/popdel/all.html [2021, December 12].

Willis S. C. (2017). One species or four? Yes!...and, no. Or, arbitrary assignment of lineages to species obscures the diversification processes of Neotropical fishes. *PloS one*, *12*(2), e0172349. https://doi.org/10.1371/journal.pone.0172349

Yang, W., Yao, D., Wu, H., Zhao, W., Chen, Y., & Tong, C. (2021). Multivariate genome-wide association study of leaf shape in a Populus deltoides and P. simonii F1 pedigree. *PloS one*, *16*(10), e0259278.

Zhang, X., Gu, C., & Zhang, T. (2020). Chloroplast (Cp) Transcriptome of P. davidiana Dode×P. bolleana Lauch provides insight into the Cp drought response and Populus Cp phylogeny. *BMC Evolutionary Biology*, *20*, 1-14. https://doi.org/10.1186/s12862-020-01622-7

Zhou J, Zhang S, Wang J, Shen H, Ai B, Gao W, Zhang C, Fei Q, Yuan D, Wu Z, *et al.* (2021). Chloroplast genomes in Populus (Salicaceae): comparisons from an intensively sampled genus reveal dynamic patterns of evolution. *Scientific Reports,* *11*, 1–14.

Zhao, Qiong, et al. (2010). Nitrogen and Phosphorus Transformations in the Rhizospheres of Three Tree Species in a Nutrient-Poor Sandy Soil. *Applied Soil Ecology*, *46*, 341–346. https://doi.org/10.1016/j.apsoil.2010.10.007.