Adaptive Introgression Shapes the Pan-genome of Populus Hybrid Zones

Baxter Worthing[1]

[1]The University of Vermont,

Abstract

Poplar trees…

0.1 Introduction

Hybridization between distinct species can lead to the exchange of genetic variation across species boundaries, a process known as introgression. Introgression is a key source of genetic novelty that shapes the evolution of natural systems (*CITE*) and humans (*cite*) and contributes to the development of crop varietals [Gao et al. (2019); Kou et al. (2020); Cheng et al. (2019); Sun et al., 2020; Qiao et al.,2021; Zanini et al., 2021]. Admixed individuals, the recipients of intorgressed sequence, present novel combinations of alleles originating from both parental species and serve as tests of the environmental and molecular factors that limit the adaptive exchange of genetic variation between species. Indeed, research on natural admixed populations suggests that introgression can be an adaptive process,through which genetic variation related to locally-adaptive phenotypes is passed from one species to the other [Suarez-Gonzalez et al. (2018); Leroy et al. (2020); Rendón-Anaya et al. (2021); Hamilton & Miller (2016); savolainen2007gene; kremer2020oaks] (*CITE MORE THAN JUST TREES*). Closer and more extensive analysis of natural mainfestations of adaptive introgression can reveal the types of genetic variation that are able to adaptively cross species boundaries, the regions of the genome that are receptive to the introduction of foreign sequence and the degree to which environmental context influences the adaptive nature pf introgression. Such research is valuable not only becasue it sheds lights on important questions in evolutionary biology, such as the evolution of species boundaries and molecular basiss of adaptive traits, but also because it gauges the overall potential for the introduction of foreign genetic variation into a novel genetic background, which is a key area of exploration in the field of biotechnology.

The size and molecular origin of the genetic variation that is exchanged between species through adaptive introgression is important to define, as a growing body of research suggests that genomic structural variation (SV), here defined as genetic variation larger than 50bp resulting for insertion, deletion, translocation or inversion of genomic sequence is often the causal variation underlying ecologically and economically important traits in many taxa. For instance, SV may play a role in local adaptation to climate, and past research shows divergence between populations for SV genotypes resulting from both local adaptation and neutral population structure (Hämälä et al., 2021; Y. Li et al., 2024; Z. Li et al., 2023; Songsomboon et al., 2021). At a broader scale, research also shows that SV maintain genetic differentiation between species ((L. Zhang et al., 2021). However, less is known about the adaptive exchange of SV genotypes between divergent populations and species. Introgression between crop varietals and wild relatives has been a key factor in the breeding history of many crops, and recent work shows that SV are often the casual variants involved in this process [Gao et al. (2019); Kou et al. (2020); Cheng et al. (2019); Sun et al., 2020; Qiao et al.,2021; Zanini et al., 2021]. Introgression is also common in natrual systmes and recent work highlights introgression as an important source of adaptive genetic novelty in many species, particularly forest trees (Hamilton & Miller, 2016; Leroy et al., 2020; Rendón-Anaya et al., 2021; Suarez-Gonzalez et al., 2018). However, past work on the genetics of such adaptive introgression in natural systems focused primarily on SNP genotypes. While hybrid genomes may be porous to variation at the singl -nucleotide scale, and introgressed SNP alleles may even present an adaptive advantage in admixed individuals, it remains unclear if the same true for SV genotypes, which have greater potential for delterious phenotypic effects. Recently, research has highlighted a potential role for SV in adaptive intorgression. This work suggests that SV may play an important role in both adaptive introgression (Almarri et al., 2020; Xia et al., 2023; X. Zhang et al., 2021) and in the maintenance of species boundaries (L. Zhang et al., 2021). If SV is indeed involved in adaptive introgression, it is worth

investigating the molecular nature (eg size, frequency, variant class, genomic location) of the SV alleles involved, as well as the consistency and directionality of intrgression involving SV. Admixed individuals may present mosaics of SV alleles derived from both parental sepecies, or introgression may be directionally biased. Similarly, natural hybrid zones may be hotspots for the generation of novel SV alleles through processes such as unequal crossing over. Here, we leverage extensive sampling of natural forest tree hybrid zones, cutting edge techniques for genotyping SV in admixed individuals and established landscape genomics analyses to investigate these areas of uncertainty.

****CHOP*** Hybridization between distinct species can lead to the exchange of genetic variation across species boundaries, a process known as introgression. Introgression is common among inter-fertile species of forest trees, and in many cases, is hypothesized to be an adaptive process, through which genetic variation related to locally-adaptive phenotypes is passed from one species to the other (Hamilton & Miller, 2016; Leroy et al., 2020; Rendón-Anaya et al., 2021; Suarez-Gonzalez et al., 2018) Support for this hypothesis requires a clearer understanding of the nature of genetic variation that is exchanged between species, and the degree to which this variation is involved in adaptive processes in natural populations. Characterization of the genetic variation involved in adaptive introgression would shed light on the evolutionary forces that influence the content of admixed genomes and help reveal the loci that underlie both the adaptive traits and genetic incompatibilities that shape species boundaries.

****CHOP*** The majority of past research on adaptive introgression has focused on single nucleotide polymorphisms (SNPs) and small indels, but structural variation (SV) is also likely to play a role in adaptive introgression. SV, which includes deletions, duplications, inversions, and translocations of DNA segments, can have a significant impact on gene expression, gene function, and ultimately, adaptive phenotypic variation. For this reason, SV is known to be a major source of genetic variation in many species, and has been implicated in the evolution of adaptive traits in a variety of taxa (Hämälä et al., 2021; Y. Li et al., 2024; Z. Li et al., 2023; Songsomboon et al., 2021). As the notion that SV is involved in adaptive evolution gains support, it is only logical to ask is how SV is involved in adaptive introgression. Recent research suggests that SV may play an important role in both adaptive introgression (Almarri et al., 2020; Xia et al., 2023; X. Zhang et al., 2021) and in the maintenance of species boundaries (L. Zhang et al., 2021) in natural systems. Introgression between crop varietals and wild relatives has been a key factor in the breeding history of many crops, and recent work shows that SV are often the casual variants involved in this process [Gao et al. (2019); Kou et al. (2020); Cheng et al. (2019); Sun et al., 2020; Qiao et al., 2021; Zanini et al., 2021]. While the significant effect of introgressed SV on economically important crop traits is well recognized, far less in known about the relationship between introgression, SV and ecologically important traits in natural hybrid zones.

The study of SV in natural hybrid zones is challenging, as it requires the ability to accurately genotype SV in a large number of admixed individuals. The majority of past research on SV in natural hybrid zones has relied on the use of reference genomes derived from a single individual from only one of the parental species, which can lead to reference bias [Bock et al. (2023); song2023plant; feng2024integrating]. Reference bias occurs when the genomes of re sequenced individuals contain regions that are highly diverged from the reference genome sequence, causing reads originating from these regions to align incorrectly, or not align at all. This can lead to misinterpretation or under-representation of variation resulting from admixture (Secomandi et al., 2025). Reference bias is particularly problematic when studying SV, as large insertions and deletions may be longer than sequencing reads, making variation in their presence and absence (PAV) challenging to detect. Furthermore, SV is often species-specific and, because admixed individuals represent a mixture of genetic

variation from two or more species, the presence of reference bias can obscure the effect that the presence (or absence) of genetic variation has on traits, and can hinder the identification of SV that is involved in introgression.

Recently, an increasing number of studies have relied on pan-genome assembly to overcome reference bias and to accurately genotype SV (Gao et al., 2019; Kang et al., 2023; Y. Li et al., 2024; Z. Li et al., 2023; Liang et al., 2025; Secomandi et al., 2025; Songsomboon et al., 2021). A pan-genome assembly is a non-redundant collection of sequences originating from multiple individuals. This genetic information can be represented as "nodes" in a pan-genome graph, while the linear sequence of each input genome is stored as a "path" connecting a series of nodes. In a pan-genome graph, each node describes the alignment between at least two sequences, given an expected level of sequence divergence (Kang et al., 2023). Pan-genome graphs can capture complex variation that is not present in a single reference genome, and can provide a more accurate representation of the genetic variation present in admixed individuals. Pan-genome assembly has been used to identify and genotype SV related to ecologically and economically important traits in a variety of taxa, and has been shown to be an effective tool for studying adaptive evolution in natural populations (Fang & Edwards, 2024; Kang et al., 2023; Secomandi et al., 2025). Pan-genome assembly could help overcome the challenges of studying SV in admixed tree populations, and could provide new insights into the role of SV in adaptive introgression in natural populations of forest trees [secomandi2025pangenome]. Despite this potentail, we know of only one study that used pan-genome assembly to examine introgression in the context of forest tree hybrid zones. Liang et al. (2025) use a pan-genome based approach to genotype SV in a range-wide sampling of the interfertile oak species *Quercus variabilis* and *Q. acutissima*, showing that SV show signals of adaptive intorgression that differs from that detected from SNP data and identifying adaptive intorgression of SV genotyps in a gene rich region of oak chromosome 9 (Liang et al., 2025). Here, we add to this naicent line of research by exploring the relationship between structural variation and introgression between *Populus balsamifera* and *Populus trichocarpa*, two forest tree species that diverged much more recently than the oak species studied by Liang et al. (2025). We leverage a larger sample size and finer geographic scale to study the role of SV in local adaptation and adaptive introgression across repeated hybrid zones.

Populus balsamifera and *Populus trichocarpa* are two species of forest tree that readily interbreed in nature, and early and advanced generation hybrids have been described in hybrid zones located where their ranges overlap (Geraldes et al., 2014; Suarez-Gonzalez et al.,2016; 2018; Chhatre et al. 2019). It has previously been hypothesized that hybridization between these two species contributes to an adaptive process, facilitating introgression of genetic variation related to locally-adaptive phenotypes from one species to the other (Suarez-Gonzalez et al., 2016; Suarez-Gonzalez et al., 2018). However, studies on controlled inter-specific crosses in *Populus* have made it clear that introgression is often biased toward particular regions of the genome, particular *Populus* species, or specific environments (Lexer et al., 2005; Thompson et al., 2010; Meirmans et al., 2017). Therefore, the porosity of the genome to adaptive introgression between hybridizing species may be constrained by incompatibilities between the genomic variants involved as well as environmental context, two factors which warrant further investigation. The environmental context of adaptive introgression may be of particular importance in the case of *P. balsamifera* and *P. trichocarpa*. *P. balsamifera* is more commonly found in colder, continental climates while *P. trichocarpa* is more often associated with milder, coastal climates with longer growing seasons (Geraldes et al., 2014; Suarez-Gonzalez et al., 2018). Considering hybrid zones between these species often fall along the boundaries of their ranges (Chhatre et al., 2019; Fetter and Keller, 2023), it seems possible that adaptive introgression helps hybrid individuals persist in environments that are outside of the optimum for either parental

species. Suarez-Gonzalez et al. (2018) showed support for this hypothesis, finding that introgressed regions in the genomes of hybrids harbored signs of selection and variation associated with adaptive traits, including phenology. However, the specific adaptive variants within these genomic regions, and the nature of their effects on fitness remain, for the most part, undiscovered. Identification of SV involved in adaptive introgression between these two species would not only contirbute to a broader undersatanidng of the molecular basis of adaptive introgression, but would also provide helpful insight into the conservation of these ecologically important species. In trees, long generations times are thought to limit the potential contribution of de novo mutation to adaptive evolution in response to rapid environmental change (Feng et al., 2024; Savolainen et al., 2007). Therefore, adaptive intorgression is viewed as one of the few routs for trees to undergo rapid evolutionin repsonse top cliamte change (Feng et al., 2024). If adaptive introgression does help admixed individiuals persist in environments oitside of their optimum, then the varaition involved could guide geneic conservation efforts, such as assisted gene flow. Furthermore, industrial breeding of *Populus* varietals generally involves interspecific corsses, and breeding programs would benefit from a clearer understanding of the genetic variation that is involved in adaptive introgression between these two species.

Here, we take advantage of recent advances in pan-genome assembly methods to produce a pan-genome reference, comprising 24 diverse haplotypes from *P. balsamifera*, *P. trichocarpa* and their hybrids, facilitating unbiased analysis of the sequence variation that is segregating within natural *Populus* hybrid zones. Using this new pan-genome assembly, we genotype structural variation across 566 individuals sampled from within and outside of 6 distinct *P. balsamifera* and *P. trichocarpa* hybrid zones. We assess the extent of genomic diversity present in these species and their hybrids, identifying genomic variation that is not present in the *P. trichocarpa* reference genome. Furthermore, we describe structural variation involved both in introgression and in genomic divergence between the two species. We shed light on the role that introgression may play in shaping the pan-genome of these species, and the degree to which tree genomes are porous to the inter-specific exchange of structural variant alleles.

0.2 Methods

0.2.1 Sample collection and whole genome sequencing

In January 2020, we collected dormant branch cuttings from 546 poplar trees along 7 distinct latitudinal transects spanning natural contact zones between Populus trichocarpa and Populus balsamifera Figure 1. Short read whole genome sequencing and subsequent bioinformatic analyses were performed as described in Bolte et al. (2024). Briefly, Genomic DNA libraries for all sampled individuals were constructed at the Duke University Center for Genomic and Computational Biology and sequenced using an Illumina NovaSeq 6000 instrument with an S4 flow cell with 64 samples per lane (Illumina Inc., San Diego, USA). De-indexing, QC, trimming adapter sequences, and sequence pre-processing were completed by the sequencing facility. In addition to short read sequencing, we also sequenced a subset of 40 sampled individuals with PacBio HiFi long reads. We harvested tissue from newly-expanded leaves grown under low light conditions to use for high molecular weight DNA extraction. After confirming extraction quality through gel electrophoresis and bioanalysis, we submitted HMW DNA to the University of Maryland Center for Integrative Biosciences Genomics Core Facility for library preparation and sequencing on the the PacBio Sequel system with two samples per SMRT flow cell. We sequenced one *P. balsamifera* sample (939) for an additional run on a full SMRT cell. We also sequenced this individual with Oxford Nanopore Technology (ONT) platform. We submitted high molecular weight DNA to the Vermont Integrative Genomics Resource Sequencing Center for library preparation and sequencing on XXX flowcell (two runs) and XXX flowcell (one run)

0.2.2 Genome assembly

Of the 40 individuals sequenced with HiFi long reads, we selected 16 for whole genome assembly Figure 1. These samples ranged from 20 to 35x long read coverage. We performed de-novo genome assembly of the HiFi reads for these samples with HiCanu (Nurk et al., 2020). We set HiCanu parameters as follows: gSize="400m", lc="5", lcer="0.055", ovrlp="350", mincov="9". To detect potential contaminants in the raw assemblies produced by HiCanu, we used the program Kraken2 to compare the k-mer content of assembled contigs to the "PlusPFP" database of known taxon-specific k-mers representing archaea, bacteria, viral, human, protozoa, fungi and plant taxa (Lu et al., 2022). We then used a custom bioawk script to remove any assembled contig that Kraken2 assigned to a taxonomic unit other than *Populus*, leaving only unassigned or *Populus*- assigned contigs in each assembly. We assessed the accuracy and completeness of the decontaminated assemblies using QUAST and BUSCO (Gurevich et al., 2013; Simão et al., 2015). To further assess the contiguity and quality of these assemblies, we used minimap2 and BWA to map the original HiFi reads and additional Illumina short reads back to each assembled haplotype (H. Li, 2018; H. Li & Durbin, 2010). We passed these alignments to the program CRAQ (K. Li et al., 2023), which leverages read depth along assembled contigs for quality assessment. We repeated these quality assessment checks at each subsequent stage of genome assembly. Phasing, or the separation and concatenation of contigs belonging to the same parental haplotype is a key step in genome assembly for hybrid individuals, as divergent haplotypes likely contain important genetic information. We used mininmap2 to map reads back to assembled contigs and to map assembled contigs to themselves. We then used the purge haplotigs pipeline (Roach et al., 2018) to split diploid assemblies into two assembled haploytpes for each individual. We used RagTag (Alonge et al. 2022) to connect decontaminated, phased contigs into pseudo-chromosomal scaffolds, guided by alignments of assembled contigs to the Nisqually1 *P. trichocarpa* reference genome. We visually inspected minimap2 alignments of scaffolded assemblies to the reference genome to identify potential scaffolding errors. We used RepeatMasker to annotate and mask repetitive regions and annotated possible coding domains and predicted protein sequences for each assembly with AUGUSTUS (Smit et al., 2015; Hoff et al., 2019).

0.2.3 Pan-genome assembly

In a pan-genome graph, a non-redundant collection of all input sequences is represented as "nodes", while the linear sequence of each input genome is stored as a "path" connecting a series of nodes. In this approach to graph construction, each node of the graph w describes the alignment between at least two sequences, given an expected level of sequence divergence. We constructed a pan-genome graph from an all-by all alignment of the 24 assembled haplotypes. Any contigs shorter than 100kb were dropped before alignment. We used wfmash (Guarracino et al., 2021) to perform all-by-all alignments between assembled chromosomes, and seqwish (Garrison and Guarracino, 2022) to project alignments into a graph pan-genome assembly. Pan-genome graphs constructed this way can have highly complex topography, which may hinder downstream applications such a sequence alignment. To combat this issue, we used the program smoothXG to smooth complex variation in these graphs. We also assembled a separate graph specifically for subsequent alignment-based analyses using the minigraph-cactus pipeline (Hickey et al., 2023). We used the program panacus to asses pan-genome growth and coverage statistics (Parmigiani et al., 2024). We used the program ODGI (Guarracino et al., 2022) to perform basic graph quality control and to partition the core (sequence present in all individuals), shell (sequence variably present or absent across individuals) and singleton sequence (present in only one individuals) content for each graph and individual represented in the graphs. We used the program vg (Garrison et al., 2018) to deconstruct each graph assembly into a VCF containing the variation encapsulated in each graph. We passed this deconstructed VCF file to vcfbub (github.com/pangenome/vcfbub) to collapse

nested SV sites into the top-level variant for all SV less than 10kb in length. Using bcftools (Danecek et al., 2021) we split biallelic sites into multiallelic records for this deconstructed vcf.

0.2.4 Genotyping Pan-genomic Variation

We aligned sequencing reads from additional samples to pan-genome graph assembly to genotype SNPs and SVs. We aligned PacBio HiFi reads for 40 total individuals (including the 16 used to construct the pan-genome graph) to the minigraph-cactus graph using the program GraphAligner (Rautiainen & Marschall, 2020). To characterize sequence variation across a broader set of individuals, we also mapped Illumina short reads for 575 individuals to the pan-genome graph using the program giraffe (Sirén et al., 2021). We then called SV and SNPs from these alignments using the vg call algorithm (Garrison et al., 2018). We used bcftools to split multialleleic calls into separate records, normalize alleles against the *P. trichocarpa* reference and merge calls for all samples into one VCF representing SV and SNP genotypes in the *P. trichocarpa* reference coordinate space.

0.2.5 Analysis of Structural Variation and Introgression

We used ADMIXTURE results from Bolte et al (2024) to assign each individual an admixture proportion at K=2 and K=4. The genetic data used for this ADMIXTURE analysis came from SNP calls from mappings of the same illlumina reads used in this study, however they were mapped to the *P. trichocarpa* reference genome. We used this same genetic data set to perform local ancestry inference (LAI) using LOTR (Dias-Alves et al., 2018). These ADMIXTURE and LAI data sets were used to visualize and analyze patterns of local and global genomic ancestry throughout this study.

We used ODGI to flatten the Nisqually1 *P. trichocarpa* reference genome path in the cactus graph. We then used a bed file representation of this path and reference gene annotations to genotype PAV of reference sequence across all paths of the cactus graph. We used R version 4.1.0 to identify core and shell gene sequences from this data and to perform PCA on the PAV genotypes (R Core Team, 2024). We also ran a similar PAV PCA with the merged vcf representing all 575 sequenced samples in this study, using plink2 (Chang et al., 2015). We used the -–allele-weights flag to extract the loading of each allele for each SV on the first 3 principal components.

To find SV potentially involved in introgression, we employed custom scripts to identify ancestry blocks along the chromosomes of all 335 admixed individuals in this study. An individual was considered admixed if its k=2 ADMIXTURE score indicated at least 5 percent admixture genome-wide. We used Bedtools (Quinlan & Hall, 2010) to find SV that overlapped ancestry blocks in each admixed individual. We then scored each SV allele based on its prevalence in ancestry blocks that contrasted the average global ancestry of an individual.

For all SV that were outliers in principal component or LAI analyses, we used Bedtools to check for overlap with annotated genes. We then used plantgenie to analyze PFAM enrichment of gene sets (Sundell et al., 2015).

0.3 Results

0.3.1 Whole genome sequencing and assembly.

Illumina sequencing yielded on average of 8358 MB of sequence, corresponding to an average sequencing depth of 21x. PacBio HiFi sequencing yielded on average 6368 MB of sequence, corresponding to an average sequencing depth of 16x. Sample 939 was sequenced to a depth of roughly 60X HiFi and roughly 20X ONT reads. Table 1 shows summary statistics for the 16 whole genome assemblies included in the pan-genome. Assembly quality and contiguity were generally high, however alternate haplotypes were not always fully assembled (Table 1).

0.3.2 Pan-genome assembly

The minigraph-cactus pangenome graph contained a total of 102,907,131 nodes comprising 1.2 Gbp of sequence. Of this, 243 Mbp was present across all individuals (core genome), while 423Mbp was present in some but not all individuals (shell genome). The remaining 537 Mbp represented singleton sequence. Figure 2 shows how the size of the sequence classes changes as individuals are added to the pan-genome. Figure 3 shows the proportions of core, shell and singleton sequence within each sample. **?@fig-4** shows each genome in the pan-genome graph plotted on the first two principal components of a PCA on PAV of reference sequence.

0.3.3 Pan-genomic Variation

3708 annotated genes were variably present or absent across the 17 genomes used to construct the pan- =genome. These genes were significantly enriched for various protein families including NB-ARC domain, D-mannose binding lectin, sulfotransferases, cupins, and Cytochrome P450. Of these shell genes, 167 were present in all of the pure *P. trichocarpa* individuals and none of the pure *P. balsamifera* samples making up the pan-genome graph. These *P. trichocarpa* - specific genes were enriched for protein families, including sulfotransferases, serine carboxypeptidases, Mlo family proteins, D-mannose binding lectin, Glycosyl hydrolases and absistic acid (ABA) induced proteins.

The cactus pan-genome graph contained 1,773,427 SV greater than 20bp in length. Figure 4 shows the distribution of these variants across the genome. The distribution of frequencies of the non-reference allele for these SV is shown in Figure 5.

The first two principal components of a PCA on PAV genotyped from short read alignments to the pan-genome graph are shown in Figure 6. The loading of each allele for each SV on the first principal component of a PCA on PAV genotyped from short read alignments to the pan-genome graph is shown in **?@fig-8**. The SV that were outliers on PC1 overlapped 1362 annotated genes. These genes were significantly enriched for protein families involved in development and growth (K-Box proteins, KNOX proteins), cold tolerance (DEAD/DEAH box helicases), and phenology (SRF transcription factors).

0.3.4 SV and Introgression

We identified 126,973 SV that overlapped ancestry blocks in admixed individuals, and found that 30,721 of these were associated with blocks of either *P. trichocarpa* or *P. balsamifera* ancestry in admixed individuals.

0.3.5 Tables

Assembly Statistics

Metrics of size, quality and completeness for the assmeblies making up the pan-genome

	Assembly size				Assembly size completeness						Assembly quality	
haplotype	length	Number of Scaffolds	N50	N50_n	gaps	N_count	Coverage Rate	busco_complete	busco_duplicated	busco_fragmented	Low Confidence Rate	Mean Assembly Quality Index
247	406044802	19	19100130	8	559	55900	0.9986	96.7	21.7	0.7	0.0000231000	90.450
247_alt	376222526	22	21729037	7	1470	147000	0.9970	92.7	16.8	0.9	0.0001421499	81.550
365	400343671	19	20009943	8	193	19300	0.9993	96.7	19.6	0.6	0.0000448000	93.800
365_alt	382988961	25	19752833	8	743	74300	0.9974	95.0	19.0	0.8	0.0001006478	86.900
368	408927241	19	20528059	8	479	47900	0.9984	96.6	20.6	0.7	0.0001896474	91.600
368_alt	376380057	26	19967418	8	1331	133100	0.9962	92.6	18.3	0.9	0.0001708698	82.850
406	406883189	20	21305652	8	934	93400	0.9986	97.6	19.4	0.6	0.0000080500	86.050
406_alt	343567265	23	17785992	8	5810	581000	0.9903	74.0	11.2	3.2	0.0001700919	58.550
439	399397382	19	19486512	8	566	56600	0.9985	96.7	19.4	0.7	0.0000117000	90.800
439_alt	356224784	21	19609853	8	2205	220500	0.9948	88.6	16.1	1.1	0.0000465000	76.050
515	395807708	19	21661901	7	1296	129600	0.9975	97.3	19.4	0.6	0.0000448000	81.700
515_alt	286914403	20	14090572	8	6632	663200	0.9881	60.1	8.0	3.5	0.0002017710	52.850
562	402113775	21	19407464	8	1535	153500	0.9977	97.2	19.0	0.6	0.0000486000	80.500
562_alt	294685596	22	15156518	8	7591	759100	0.9843	60.0	7.0	3.9	0.0001118243	53.050
566	404344743	20	20403993	8	1550	155000	0.9971	96.9	19.7	0.8	0.0000449000	80.200
566_alt	295750810	20	14501527	8	6773	677300	0.9873	62.3	8.3	4.1	0.0000791000	54.100
712	406603955	21	20302875	8	550	55000	0.9986	96.5	20.8	0.5	0.0000614000	91.350
712_alt	374451786	20	20382624	8	1553	155300	0.9961	92.4	18.2	0.9	0.0000495000	81.900
762	400616755	20	21889695	7	853	85300	0.9978	96.4	21.4	0.6	0.0003550476	60.400
762_alt	365455960	23	18068783	8	2445	244500	0.9946	90.5	17.8	1.2	0.0003571648	54.650
776	412523377	30	20099238	8	1192	119200	0.9974	97.4	20.3	0.5	0.0000362000	81.550
776_alt	407700287	24	20201693	8	4121	412100	0.9904	92.9	17.3	1.5	0.0000911000	65.400
801	410936669	21	19562430	8	323	32300	0.9988	96.2	20.3	0.6	0.0000448000	92.850
801_alt	381097682	22	19323978	8	1138	113800	0.9963	93.0	18.4	0.9	0.0001059650	84.350
809	408711020	22	20677021	7	796	79600	0.9981	96.9	20.9	0.9	0.0001127936	85.600
809_alt	365592262	24	18971862	8	2034	203400	0.9963	89.9	16.9	1.1	0.0000788000	73.950
822	405664496	21	19776368	8	740	74000	0.9981	96.7	20.5	0.7	0.0000317000	88.900
822_alt	347527765	21	18042639	8	2151	215100	0.9958	86.8	15.6	1.0	0.0001331520	74.800
860	416270270	20	23271044	7	941	94100	0.9970	97.1	21.0	0.8	0.0002180482	87.150
860_alt	377658316	22	19870789	8	2457	245700	0.9951	90.1	15.8	1.3	0.0001090960	73.150
969	399016342	20	20036743	8	1808	180800	0.9957	96.4	18.5	0.7	0.0000709000	56.750
969_alt	277863190	21	15795860	7	7380	738000	0.9841	58.9	7.5	3.9	0.0001591467	45.135
939	422799018	19	24674954	7	117	11700	0.9974	98.1	19.4	0.6	0.0000917000	95.950
939_alt	411479088	19	22188346	8	288	28800	0.9964	97.5	19.5	0.7	0.0002849355	94.200
nisqually_1	389204664	19	21678634	7	59	590000	NA	98.2	19.6	0.5	NA	NA

{figformat: svg}

1 Figures

References

Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., et al. (2020). Population structure, stratification, and introgression of human structural variation. *Cell*, *182*(1), 189–199.

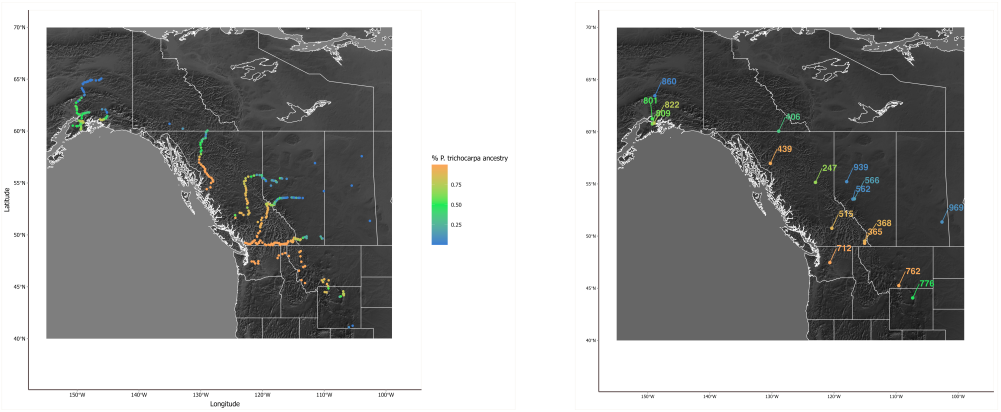
Bock, D. G., Cai, Z., Elphinstone, C., Gonzalez-Segovia, E., Hirabayashi, K., Huang, K., et al. (2023). Genomics of plant speciation. *Plant Communications*, *4*(5).

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742–015.

Cheng, H., Liu, J., Wen, J., Nie, X., Xu, L., Chen, N., et al. (2019). Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. *Genome Biology*, *20*, 1–16.

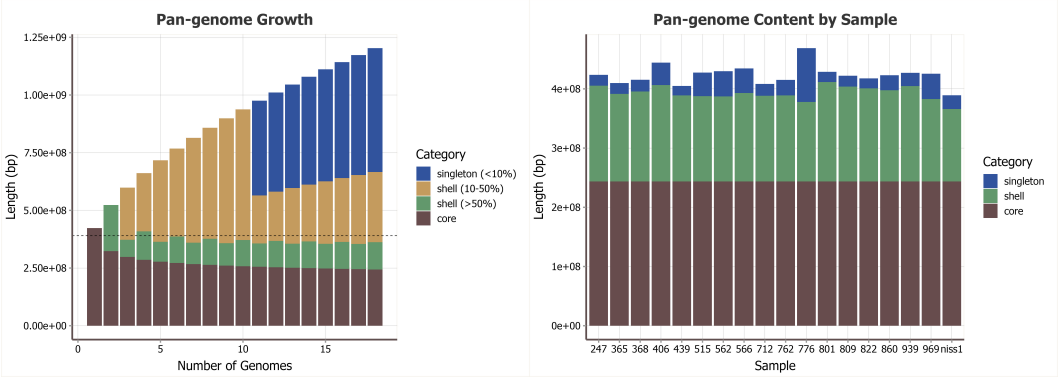
Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Dias-Alves, T., Mairal, J., & Blum, M. G. (2018). Loter: A software package to infer local ancestry for a wide range of species. *Molecular Biology and Evolution*, *35*(9), 2318–2326.



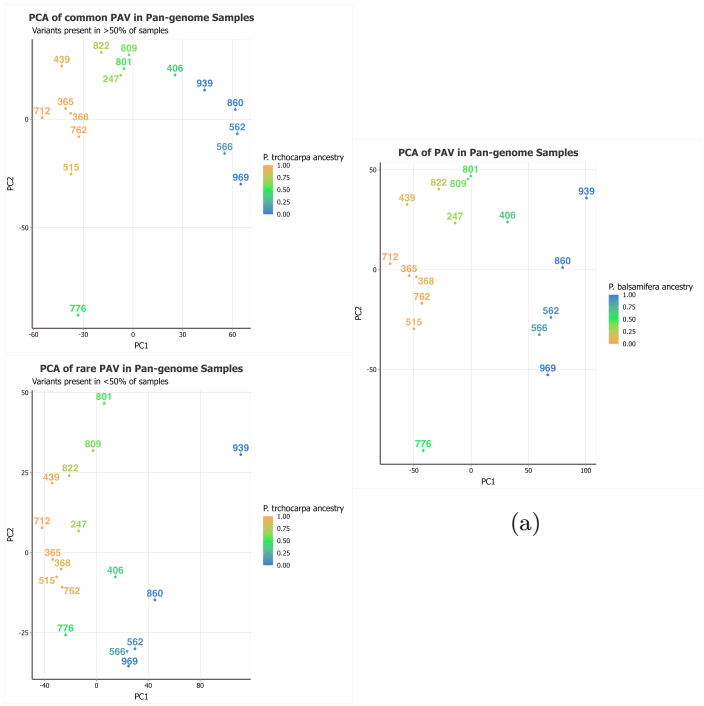
(a) (b)

Figure 1: (a): sampling locations for 575 individuals used in this study, color indicates ancestry based in ADMIXTURE analysis (b): A subset of 16 individuals used for whole genome assembly



(a) (b)

Figure 2: (a): Pan-genome growth curve visualizes the change in core and shell genome size as samples are added. (b): The relative length of the core, shell and singleton portions of the pan-genome for each sample represented



(a)

Figure 3: The first two principal components of a PCA on presence/absence variation in the pan-genome. Results of PCA on common (a), rare (b) and all (c) PAV are shown Color indicates ancestry based in ADMIXTURE analysis.

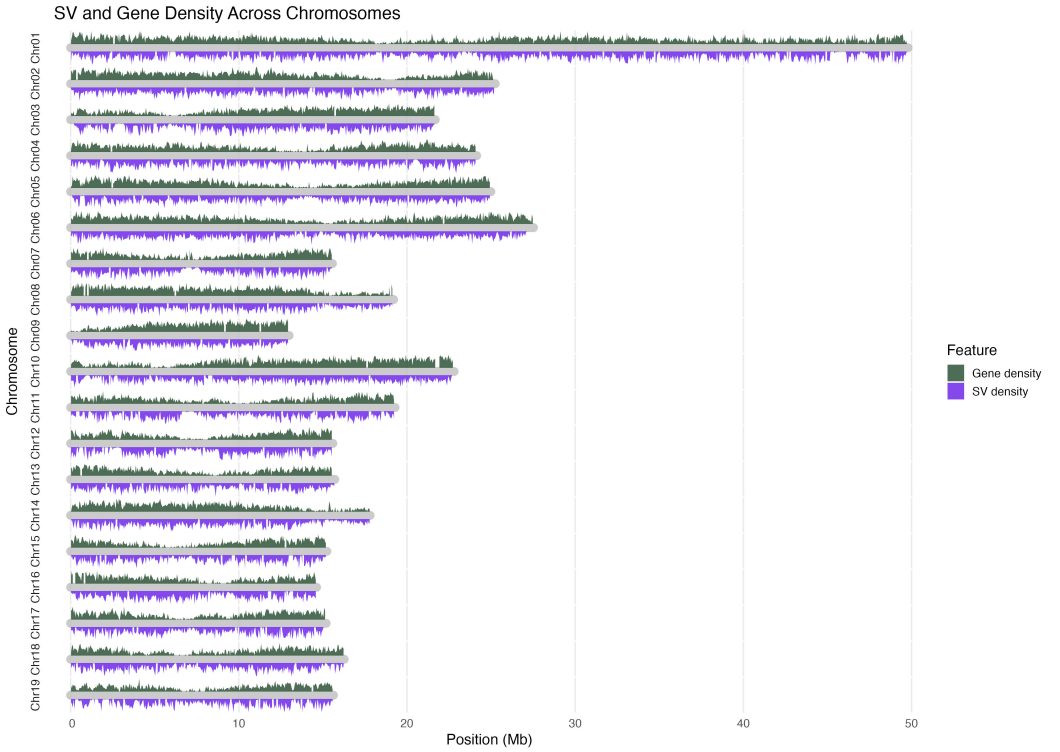


Figure 4: The density of SV larger than 20bp in length across the genome (purple) compared to the density of annotated genes (green)

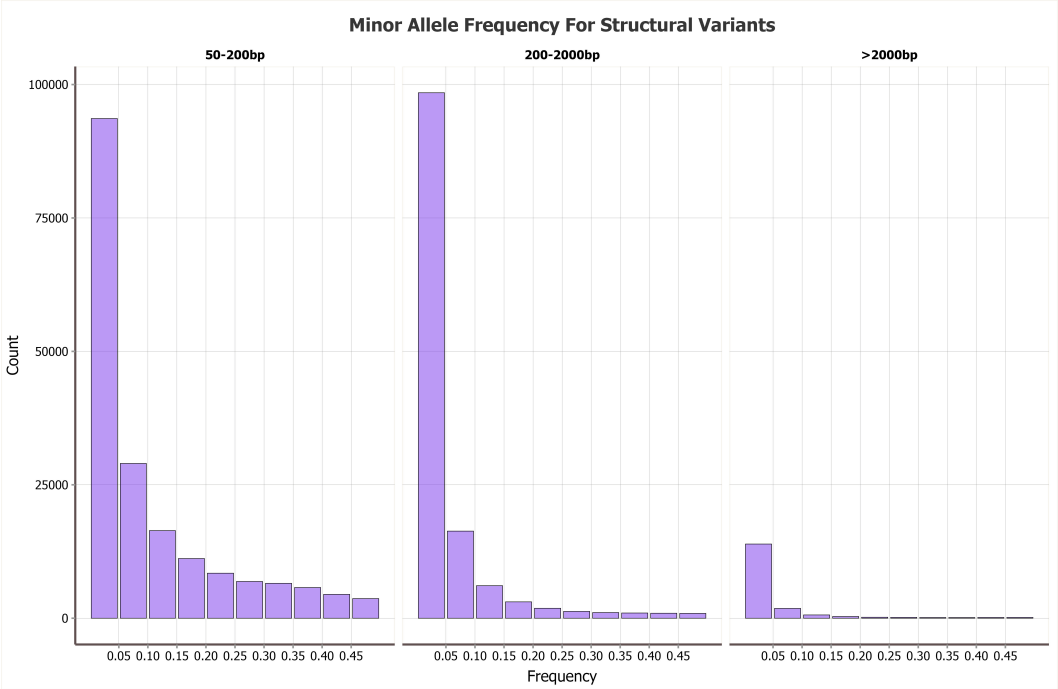


Figure 5: The frequency distribution of the minor allele for SV of different size classes

Fang, B., & Edwards, S. V. (2024). Fitness consequences of structural variation inferred from a house finch pangenome. *Proceedings of the National Academy of Sciences*, *121*(47), e2409943121.

Feng, J., Dan, X., Cui, Y., Gong, Y., Peng, M., Sang, Y., et al. (2024). Integrating evolutionary genomics of forest trees to inform future tree breeding amid rapid climate change. *Plant Communications*, *5*(10).

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*(6), 1044–1051.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–879. https://doi.org/10.1038/nbt.4227

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., Depamphilis, C. W., & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of theobroma cacao, the chocolate tree. *Proceedings of the National Academy of Sciences*, *118*(35), e2102914118.

Hamilton, J. A., & Miller, J. M. (2016). Adaptive introgression as a resource for management and genetic conservation in a changing climate. *Conservation Biology*, *30*(1), 33–41.

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., et al. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, *42*(4), 663–673. https://doi.org/10.1038/s41587-023-01793-w

Kang, M., Wu, H., Liu, H., Liu, W., Zhu, M., Han, Y., et al. (2023). The pan-genome and local adaptation of arabidopsis thaliana. *Nature Communications*, *14*(1), 6259.

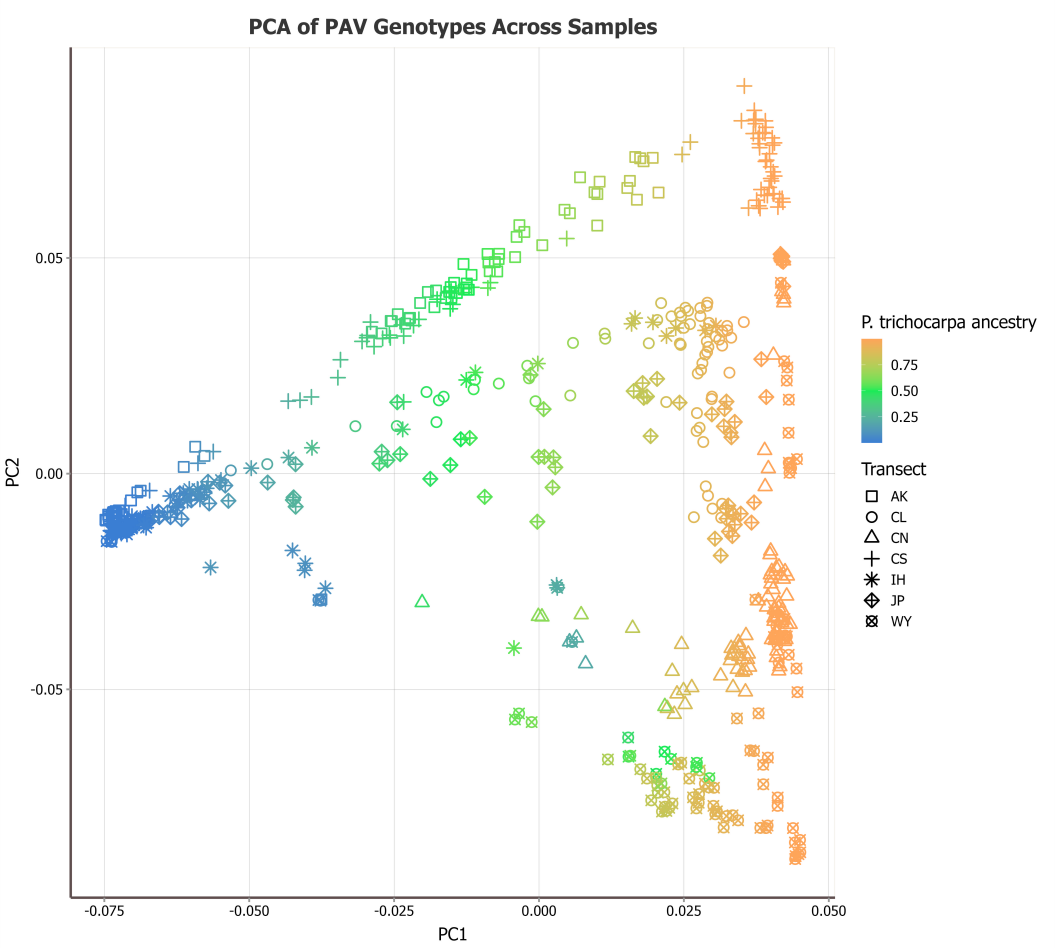


Figure 6: The first two principal components of a PCA of PAV genotyped from short read alignments to the pan-genome graph. Color indicates ancestry based on ADMIXTURE analysis. Shape indicates which of the latitudinal transects the individual was sampled from.

Kou, Y., Liao, Y., Toivainen, T., Lv, Y., Tian, X., Emerson, J., et al. (2020). Evolutionary genomics of structural variation in asian rice (oryza sativa) domestication. *Molecular Biology and Evolution*, *37*(12), 3507–3524.

Leroy, T., Louvet, J.-M., Lalanne, C., Le Provost, G., Labadie, K., Aury, J.-M., et al. (2020). Adaptive introgression as a driver of local adaptation to climate in european white oaks. *New Phytologist*, *226*(4), 1171–1182.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, *26*(5), 589–595.

Li, K., Xu, P., Wang, J., Yi, X., & Jiao, Y. (2023). Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature Communications*, *14*(1), 6556.

Li, Y., Yao, J., Sang, H., Wang, Q., Su, L., Zhao, X., et al. (2024). Pan-genome analysis highlights the role of structural variation in the evolution and environmental adaptation of asian honeybees. *Molecular Ecology Resources*, *24*(2), e13905.

Li, Z., Liu, X., Wang, C., Li, Z., Jiang, B., Zhang, R., et al. (2023). The pig pangenome provides insights into the roles of coding structural variations in genetic diversity and adaptation. *Genome Research*, *33*(10), 1833–1847.

Liang, Y.-Y., Liu, H., Lin, Q.-Q., Shi, Y., Zhou, B.-F., Wang, J.-S., et al. (2025). Pan-genome analysis reveals local adaptation to climate driven by introgression in oak species. *Molecular Biology and Evolution*, *42*(5), msaf088. https://doi.org/10.1093/molbev/msaf088

Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., et al. (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols*, *17*(12), 2815–2839. https://doi.org/10.1038/s41596-022-00738-y

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., et al. (2020). HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, *30*(9), 1291–1305.

Parmigiani, L., Garrison, E., Stoye, J., Marschall, T., & Doerr, D. (2024). Panacus: fast and exact pangenome growth and core size estimation. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btae720

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

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02157-2

Rendón-Anaya, M., Wilson, J., Sveinsson, S., Fedorkov, A., Cottrell, J., Bailey, M. E., et al. (2021). Adaptive introgression facilitates adaptation to high latitudes in european aspen (populus tremula l.). *Molecular Biology and Evolution*, *38*(11), 5034–5050.

Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007). Gene flow and local adaptation in trees. *Annu. Rev. Ecol. Evol. Syst.*, *38*(1), 595–619.

Secomandi, S., Gallo, G. R., Rossi, R., Rodríguez Fernandes, C., Jarvis, E. D., Bonisoli-Alquati, A., et al. (2025). Pangenome graphs and their applications in biodiversity genomics. *Nature Genetics*, 1–14.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., et al. (2021). Pangenomics enables genotyping of known structural variants

in 5202 diverse genomes. *Science*, *374*(6574). https://doi.org/10.1126/science.abg8871

Songsomboon, K., Brenton, Z., Heuser, J., Kresovich, S., Shakoor, N., Mockler, T., & Cooper, E. A. (2021). Genomic patterns of structural variation among diverse genotypes of sorghum bicolor and a potential role for deletions in local adaptation. *G3*, *11*(7), jkab154.

Suarez-Gonzalez, A., Lexer, C., & Cronk, Q. C. (2018). Adaptive introgression: A plant perspective. *Biology Letters*, *14*(3), 20170688.

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.-C., Sjodin, A., et al. (2015). The plant genome integrative explorer resource: PlantGenIE. org.

Xia, X., Zhang, F., Li, S., Luo, X., Peng, L., Dong, Z., et al. (2023). Structural variation and introgression from wild populations in east asian cattle genomes confer adaptation to local environment. *Genome Biology*, *24*(1), 211.

Zhang, L., Reifová, R., Halenková, Z., & Gompert, Z. (2021). How important are structural variants for speciation? *Genes*, *12*(7), 1084.

Zhang, X., Liu, T., Wang, J., Wang, P., Qiu, Y., Zhao, W., et al. (2021). Pan-genome of raphanus highlights genetic variation and introgression among domesticated, wild, and weedy radishes. *Molecular Plant*, *14*(12), 2032–2055.