

**1 Adaptive Introgression Shapes the Pan-genome of
2 *Populus* Hybrid Zones**

3 Baxter Worthing¹

4 ¹The University of Vermont,

5 **Abstract**

6 Poplar trees...

7 **0.1 Introduction**

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

20 The majority of past research on adaptive introgression has focused on single nu-
 21 cleotide polymorphisms (SNPs) and small indels, but structural variation (SV) is
 22 also likely to play a role in adaptive introgression. SV, which includes deletions, du-
 23 plications, inversions, and translocations of DNA segments, can have a significant
 24 impact on gene expression, gene function, and ultimately, adaptive phenotypic varia-
 25 tion. For this reason, SV is known to be a major source of genetic variation in many
 26 species, and has been implicated in the evolution of adaptive traits in a variety of
 27 taxa (Hämälä et al., 2021; Y. Li et al., 2024; Z. Li et al., 2023; Songsomboon et al.,
 28 2021). As the notion that SV is involved in adaptive evolution gains support, it is
 29 only logical to ask is how SV is involved in adaptive introgression. Recent research
 30 suggests that SV may play an important role in both adaptive introgression (Al-
 31 marri et al., 2020; Xia et al., 2023; X. Zhang et al., 2021) and in the maintenance of
 32 species boundaries (L. Zhang et al., 2021) in natural systems. Introgression between
 33 crop varietals and wild relatives has been a key factor in the breeding history of
 34 many crops, and recent work shows that SV are often the casual variants involved
 35 in this process [Gao et al. (2019); Kou et al. (2020); Cheng et al. (2019); Sun et
 36 al., 2020; Qiao et al., 2021; Zanini et al., 2021]. While the significant effect of in-
 37 trogressed SV on economically important crop traits is well recognized, far less is
 38 known about the relationship between introgression, SV and ecologically important
 39 traits in natural hybrid zones.

40 The study of SV in natural hybrid zones is challenging, as it requires the ability to
 41 accurately genotype SV in a large number of admixed individuals. The majority of
 42 past research on SV in natural populations has relied on reference genomes that rep-
 43 resent only one species, which can lead to reference bias. Reference bias occurs when
 44 the genomes of re sequenced individuals contain regions that are highly diverged
 45 from the reference genome , causing reads originating from these regions to align
 46 incorrectly, or not align at all to the reference genome. This can lead to misinter-
 47 pretation or under-representation of variation resulting from admixture (Secomandi
 48 et al., 2025). Reference bias is particularly problematic when studying SV, as large
 49 insertions and deletions may be longer than sequencing reads, making variation in
 50 their presence and absence (PAV) challenging to detect. Furthermore, SV is often
 51 species-specific and, because admixed individuals represent a mosaic of genetic vari-
 52 ation from two or more species, the presence of reference bias can obscure the effect
 53 that the presence (or absence) of genetic variation has on traits, and can hinder the
 54 identification of SV that is involved in introgression.

55 Recently, an increasing number of studies have relied on pan-genome assembly to
 56 overcome reference bias and to accurately genotype SV (Gao et al., 2019; Y. Li et

al., 2024; Z. Li et al., 2023; 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. 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 the adaptive evolution in natural populations (Fang & Edwards, 2024; Secomandi et al., 2025). However, to our knowledge, no studies have used pan-genome assembly to examine introgression in the context of forest tree hybrid zones, which are often dynamic and geographically broad. Pan-genome assembly could help overcome the challenges of studying SV in admixed tree genotypes, and could provide new insights into the role of SV in adaptive introgression in natural populations of forest trees.

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, through which genetic variation related to locally-adaptive phenotypes is passed from one species to the other through adaptive introgression (Suarez-Gonzalez et al., 2016; Suarez-Gonzalez et al., 2018). However, studies on 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 adaptiveness of introgression between *P. balsamifera* and *P. trichocarpa* is dependent on 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, 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 from these zones 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.

Here, we take advantage of recent advances in pan-genome assembly methods to produce a pan-genomic 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 genomic assembly, we genotype structural variation across 575 individuals sampled from within and outside of natural *P. balsamifera* and *P. trichocarpa* hybrid zones. We assess the true 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

111 role that introgression may play in shaping the pan-genome of these species, and the
 112 degree to which tree genomes are porous to the inter-specific exchange of structural
 113 variant alleles.

114 **0.2 Methods**

115 **0.2.1 Sample collection and whole genome sequencing**

116 In January 2020, we collected dormant branch cuttings from 546 poplar trees along
 117 7 distinct latitudinal transects spanning natural contact zones between *Populus*
 118 *trichocarpa* and *Populus balsamifera* Figure 1. Short read whole genome sequenc-
 119 ing and subsequent bioinformatic analyses were performed as described in Bolte
 120 et al. (2024). Briefly, Genomic DNA libraries for all sampled individuals were con-
 121 structed at the Duke University Center for Genomic and Computational Biology
 122 and sequenced using an Illumina NovaSeq 6000 instrument with an S4 flow cell with
 123 64 samples per lane (Illumina Inc., San Diego, USA). De-indexing, QC, trimming
 124 adapter sequences, and sequence pre-processing were completed by the sequenc-
 125 ing facility. In addition to short read sequencing, we also sequenced a subset of 40
 126 sampled individuals with PacBio HiFi long reads. We harvested tissue from newly-
 127 expanded leaves grown under low light conditions to use for high molecular weight
 128 DNA extraction. After confirming extraction quality through gel electrophoresis
 129 and bioanalysis, we submitted HMW DNA to the University of Maryland Center
 130 for Integrative Biosciences Genomics Core Facility for library preparation and se-
 131 quencing on the the PacBio Sequel system with two samples per SMRT flow cell.
 132 We sequenced one *P. balsamifera* sample (939) for an additional run on a full SMRT
 133 cell. We also sequenced this individual with Oxford Nanopore Technology (ONT)
 134 platform. We submitted high molecular weight DNA to the Vermont Integrative Ge-
 135 nomics Resource Sequencing Center for library preparation and sequencing on XXX
 136 flowcell (two runs) and XXX flowcell (one run)

137 **0.2.2 Genome assembly**

138 Of the 40 individuals sequenced with HiFi long reads, we selected 16 for whole
 139 genome assembly Figure 1. These samples ranged from 20 to 35x long read coverage.
 140 We performed de-novo genome assembly of the HiFi reads for these samples with
 141 HiCanu (Nurk et al., 2020). We set HiCanu parameters as follows: gSize=“400m”,
 142 lc=“5”, lcer=“0.055”, ovrlp=“350”, mincov=“9”. To detect potential contaminants
 143 in the raw assemblies produced by HiCanu, we used the program Kraken2 to com-
 144 pare the k-mer content of assembled contigs to the “PlusPFP” database of known
 145 taxon-specific k-mers representing archaea, bacteria, viral, human, protozoa, fungi
 146 and plant taxa (Lu et al., 2022). We then used a custom bioawk script to remove
 147 any assembled contig that Kraken2 assigned to a taxonomic unit other than *Populus*,
 148 leaving only unassigned or *Populus*- assigned contigs in each assembly. We assessed
 149 the accuracy and completeness of the decontaminated assemblies using QUAST and
 150 BUSCO (Gurevich et al., 2013; Simão et al., 2015). To further assess the contiguity
 151 and quality of these assemblies, we used minimap2 and BWA to map the original
 152 HiFi reads and additional Illumina short reads back to each assembled haplotype (H.
 153 Li, 2018; H. Li & Durbin, 2010). We passed these alignments to the program CRAQ
 154 (K. Li et al., 2023), which leverages read depth along assembled contigs for quality
 155 assessment. We repeated these quality assessment checks at each subsequent stage of
 156 genome assembly. Phasing, or the separation and concatenation of contigs belonging
 157 to the same parental haplotype is a key step in genome assembly for hybrid individ-
 158 uals, as divergent haplotypes likely contain important genetic information. We used
 159 minimap2 to map reads back to assembled contigs and to map assembled contigs
 160 to themselves. We then used the purge haplotigs pipeline (Roach et al., 2018) to
 161 split diploid assemblies into two assembled haplotypes for each individual. We used
 162 RagTag (Alonge et al. 2022) to connect decontaminated, phased contigs into pseudo-
 163 chromosomal scaffolds, guided by alignments of assembled contigs to the Nisqually1
 164 *P. trichocarpa* reference genome. We visually inspected minimap2 alignments of

165 scaffolded assemblies to the reference genome to identify potential scaffolding er-
 166 rors. We used RepeatMasker to annotate and mask repetitive regions and annotated
 167 possible coding domains and predicted protein sequences for each assembly with
 168 AUGUSTUS (Smit et al., 2015; Hoff et al., 2019).

169 ***0.2.3 Pan-genome assembly***

170 In a pan-genome graph, a non-redundant collection of all input sequences is rep-
 171 resented as “nodes”, while the linear sequence of each input genome is stored as a
 172 “path” connecting a series of nodes. In this approach to graph construction, each
 173 node of the graph w describes the alignment between at least two sequences, given
 174 an expected level of sequence divergence. We constructed a pan-genome graph from
 175 an all-by all alignment of the 24 assembled haplotypes. Any contigs shorter than
 176 100kb were dropped before alignment. We used wfmash (Guarracino et al., 2021)
 177 to perform all-by-all alignments between assembled chromosomes, and seqwish
 178 (Garrison and Guarracino, 2022) to project alignments into a graph pan-genome
 179 assembly. Pan-genome graphs constructed this way can have highly complex topog-
 180 raphy, which may hinder downstream applications such a sequence alignment. To
 181 combat this issue, we used the program smoothXG to smooth complex variation
 182 in these graphs. We also assembled a separate graph specifically for subsequent
 183 alignment-based analyses using the minigraph-caactus pipeline (Hickey et al., 2023).
 184 We used the program panacus to asses pan-genome growth and coverage statistics
 185 (Parmigiani et al., 2024). We used the program ODGI (Guarracino et al., 2022) to
 186 perform basic graph quality control and to partition the core (sequence present
 187 in all individuals), shell (sequence variably present or absent across individuals)
 188 and singleton sequence (present in only one individuals) content for each graph
 189 and individual represented in the graphs. We used the program vg (Garrison et
 190 al., 2018) to deconstruct each graph assembly into a VCF containing the varia-
 191 tion encapsulated in each graph. We passed this deconstructed VCF file to vcflub
 192 (github.com/pangenome/vcflub) to collapse nested SV sites into the top-level vari-
 193 ant for all SV less than 10kb in length. Using bcftools (Danecek et al., 2021) we
 194 split biallelic sites into multiallelic records for this deconstructed vcf.

195 ***0.2.4 Genotyping Pan-genomic Variation***

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

207 ***0.2.5 Analysis of Structural Variation and Introgression***

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

216 We used ODGI to flatten the Nisqually1 *P. trichocarpa* reference genome path in the
 217 cactus graph. We then used a bed file representation of this path and reference gene

218 annotations to genotype PAV of reference sequence across all paths of the cactus
 219 graph. We used R version 4.1.0 to identify core and shell gene sequences from this
 220 data and to perform PCA on the PAV genotypes (R Core Team, 2024). We also ran
 221 a similar PAV PCA with the merged vcf representing all 575 sequenced samples in
 222 this study, using plink2 (Chang et al., 2015). We used the --allele-weights flag to
 223 extract the loading of each allele for each SV on the first 3 principal components.

224 To find SV potentially involved in introgression, we employed custom scripts to iden-
 225 tify ancestry blocks along the chromosomes of all 335 admixed individuals in this
 226 study. An individual was considered admixed if its k=2 ADMIXTURE score indi-
 227 cated at least 5 percent admixture genome-wide. We used Bedtools (Quinlan & Hall,
 228 2010) to find SV that overlapped ancestry blocks in each admixed individual. We
 229 then scored each SV allele based on its prevalence in ancestry blocks that contrasted
 230 the average global ancestry of an individual.

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

234 0.3 Results

235 0.3.1 Whole genome sequencing and assembly.

236 Illumina sequencing yielded on average of 8358 MB of sequence, corresponding to an
 237 average sequencing depth of 21x. PacBio HiFi sequencing yielded on average 6368
 238 MB of sequence, corresponding to an average sequencing depth of 16x. Sample 939
 239 was sequenced to a depth of roughly 60X HiFi and roughly 20X ONT reads. Table
 240 1 shows summary statistics for the 16 whole genome assemblies included in the pan-
 241 genome. Assembly quality and contiguity were generally high, however alternate
 242 haplotypes were not always fully assembled (Table 1).

243 0.3.2 Pan-genome assembly

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

252 0.3.3 Pan-genomic Variation

253 3708 annotated genes were variably present or absent across the 17 genomes used to
 254 construct the pan- =genome. These genes were significantly enriched for various pro-
 255 tein families including NB-ARC domain, D-mannose binding lectin, sulfotransferases,
 256 cupins, and Cytochrome P450. Of these shell genes, 167 were present in all of the
 257 pure *P. trichocarpa* individuals and none of the pure *P. balsamifera* samples making
 258 up the pan-genome graph. These *P. trichocarpa* - specific genes were enriched for
 259 protein families, including sulfotransferases, serine carboxypeptidases, Mlo family
 260 proteins, D-mannose binding lectin, Glycosyl hydrolases and abscisic acid (ABA)
 261 induced proteins.

262 The cactus pan-genome graph contained 1,773,427 SV greater than 20bp in length.
 263 Figure 5 shows the distribution of these variants across the genome. The distribu-
 264 tion of frequencies of the non-reference allele for these SV is shown in Figure 6.

265 The first two principal components of a PCA on PAV genotyped from short read
 266 alignments to the pan-genome graph are shown in Figure 7. The loading of each
 267 allele for each SV on the first principal component of a PCA on PAV genotyped
 268 from short read alignments to the pan-genome graph is shown in Figure 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

haplotype	Assembly Statistics											
	Assembly size			Assembly size completeness				Assembly quality				
	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

278

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.
- 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.

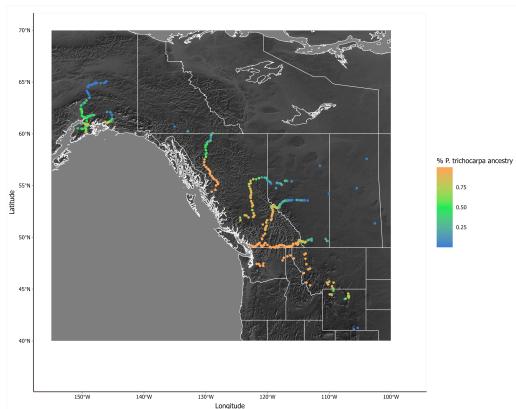


Figure 1: (Left): sampling locations for 575 individuals used in this study, color indicates ancestry based in ADMIXTURE analysis (Right): A subset of 16 individuals used for whole genome assembly

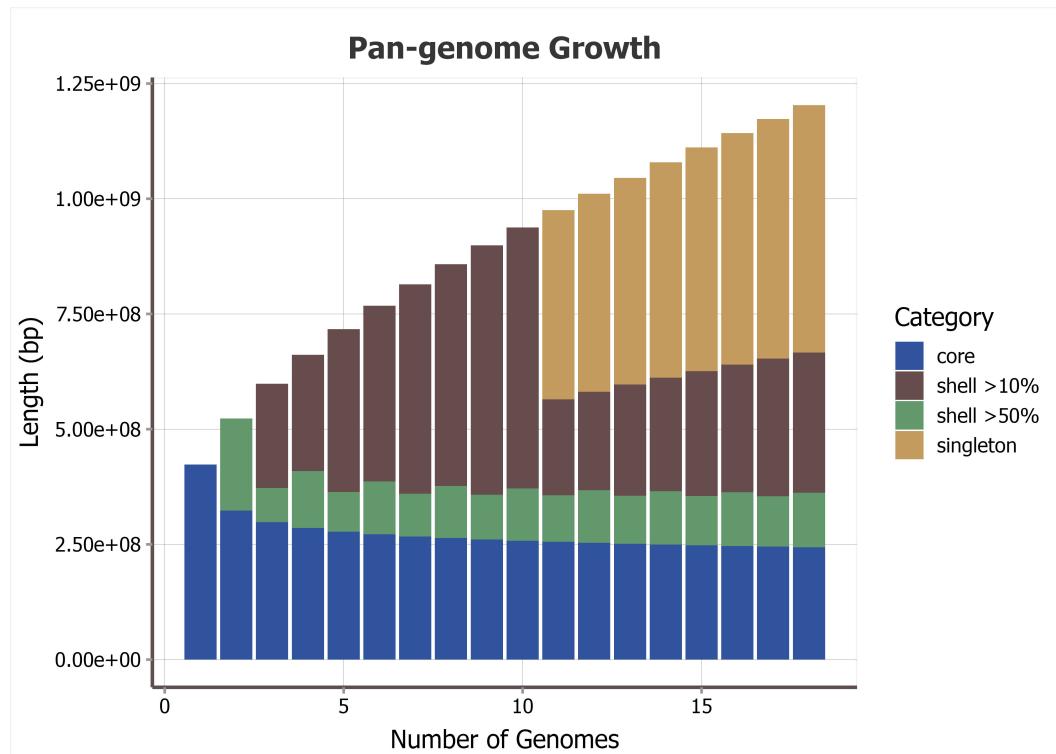


Figure 2: Pan-genome growth curve visualizes the change in core and shell genome size as samples are added

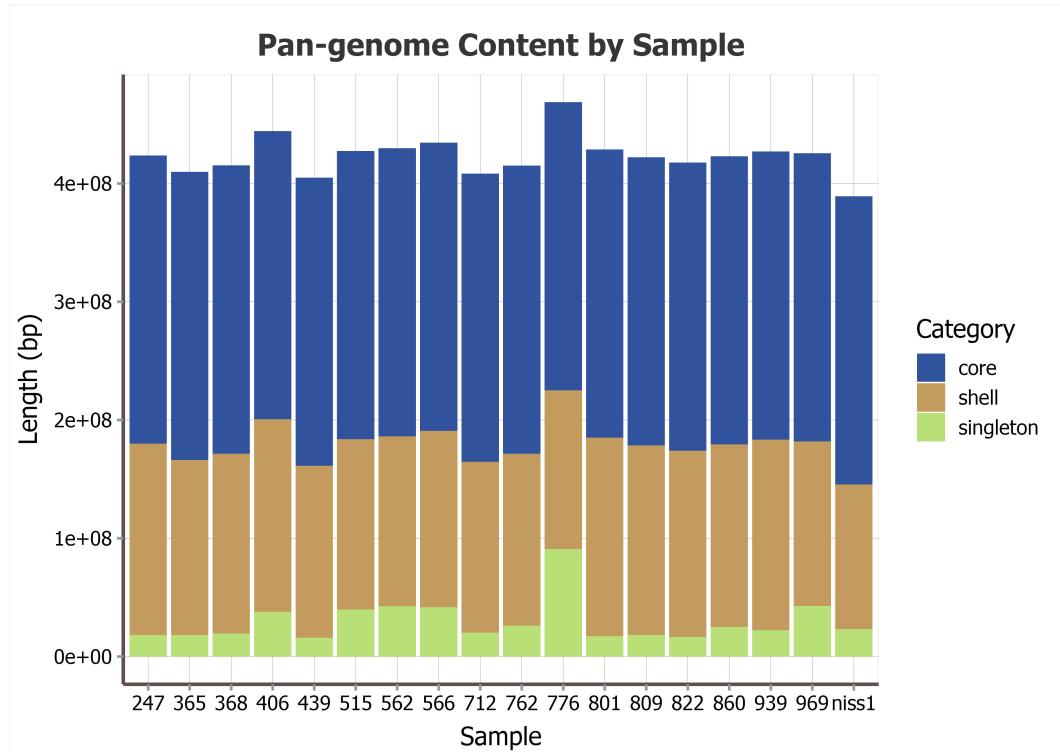


Figure 3: The relative length of the core, shell and singleton portions of the pan-genome for each sample represented

- 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.
- 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.
- 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.

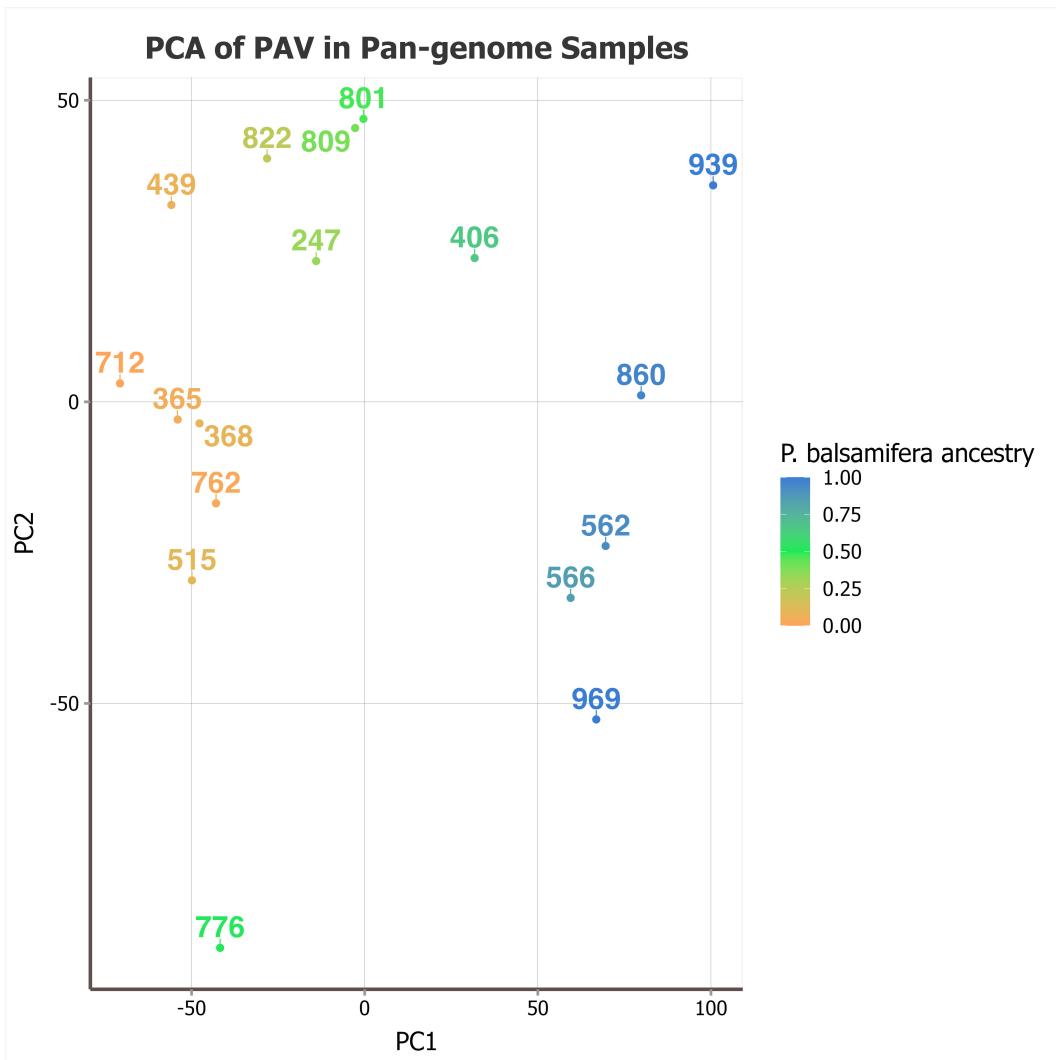


Figure 4: The first two principal components of a PCA on presence/absence variation in the pan-genome. Color indicates ancestry based in ADMIXTURE analysis.

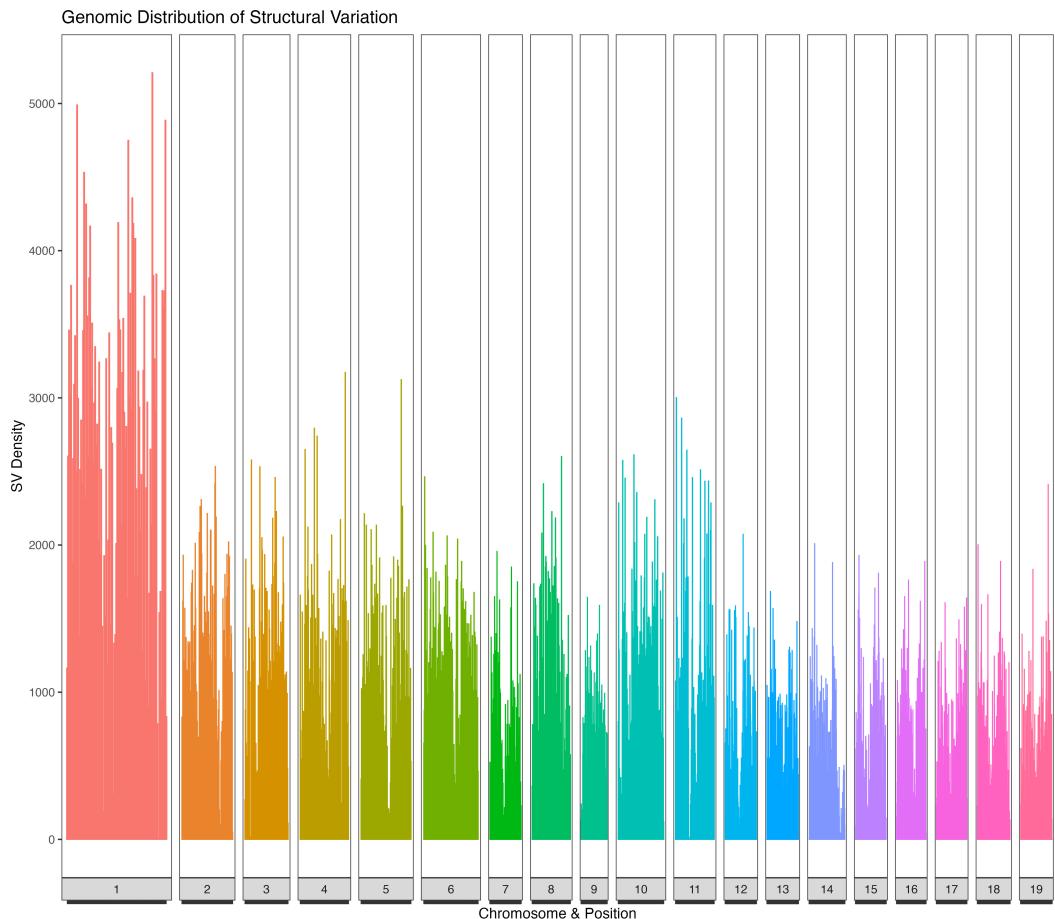


Figure 5: The distribution of SV larger than 20bp in length across the genome

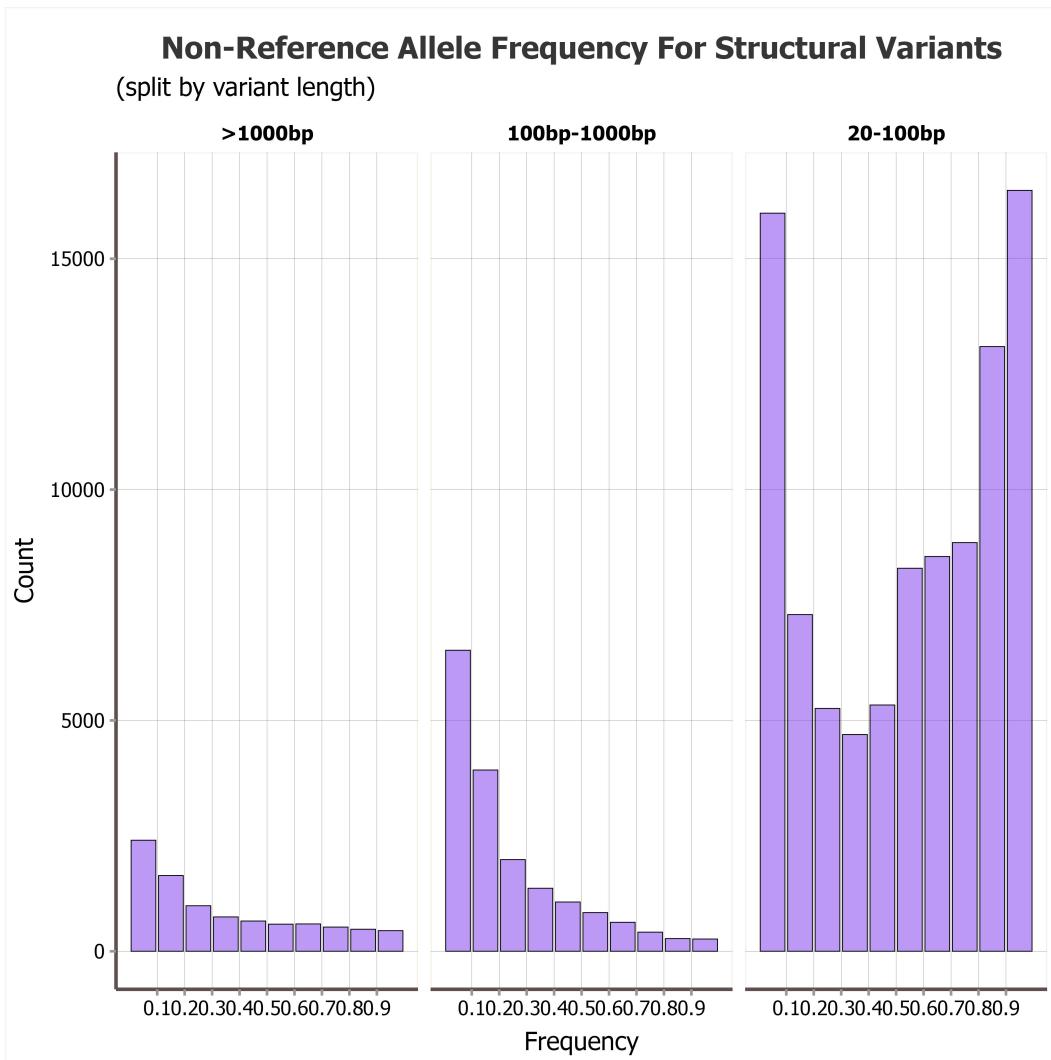


Figure 6: The frequency distribution of the non-reference allele for SV of different size classes

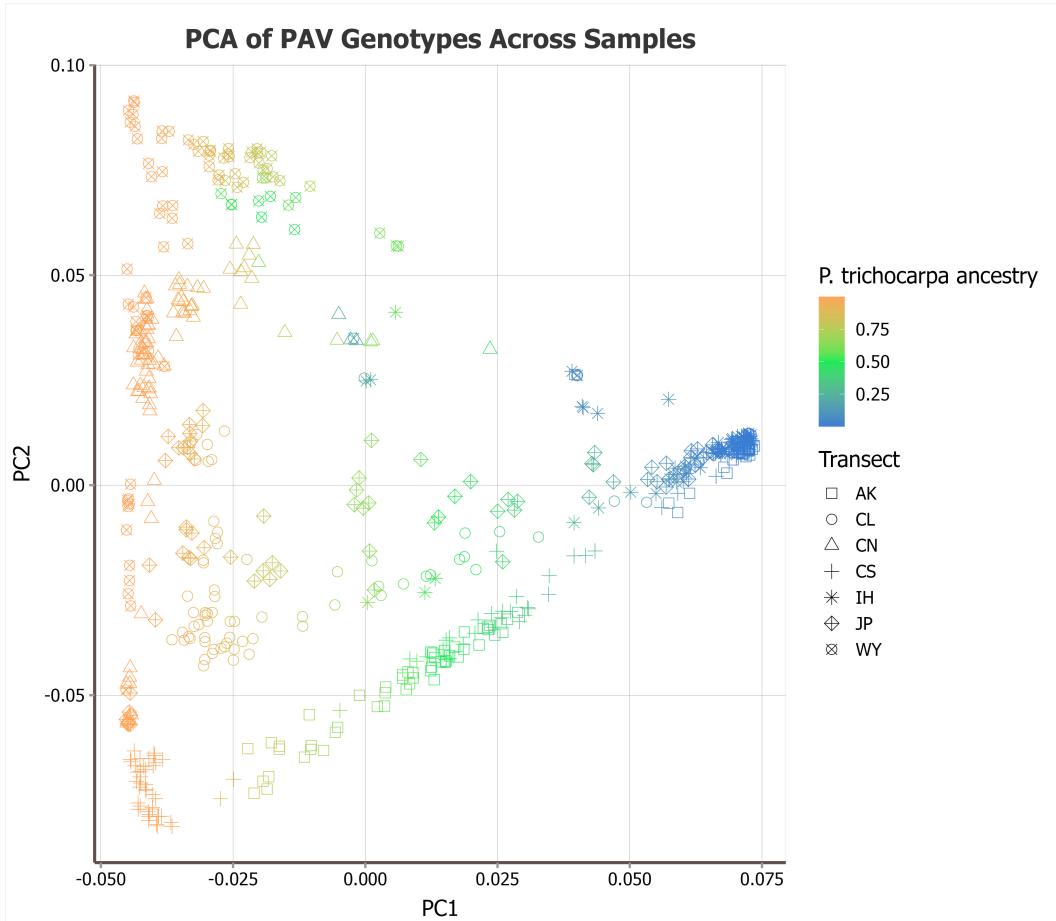


Figure 7: The first two principal components of a PCA of PAV genotyped from short read alignments to the pan-genome graph. Color indicates ancestry based in ADMIXTURE analysis. Shape indicates which of the latitudinal transects the individual was sampled from.

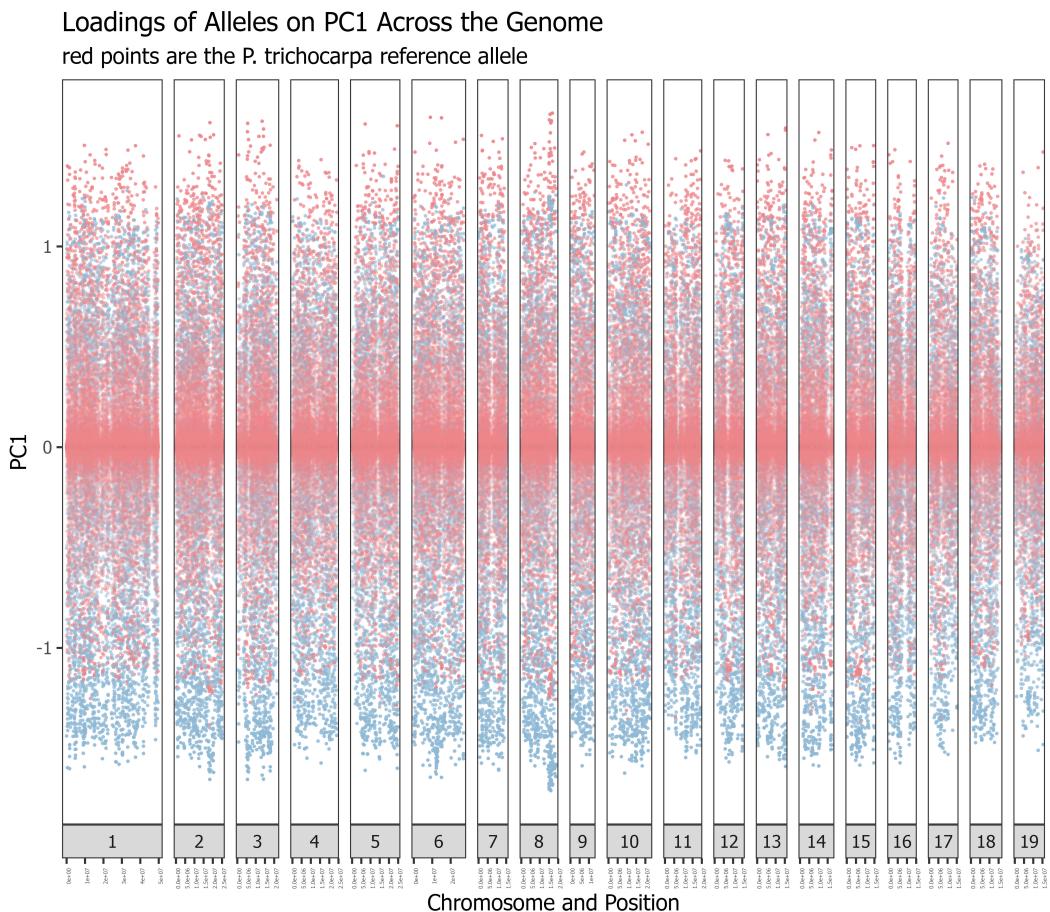


Figure 8: The loading of each allele for each SV on the first principle component of a PCA on PAV genotyped from short read alignments to the pan-genome graph. Red points are indicate *P. trichocarpa* reference alleles, blue points indicate non-reference alleles.

- 315 Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., et al.
 316 (2023). Pangenome graph construction from genome alignments with Minigraph-
 317 Cactus. *Nature Biotechnology*, 42(4), 663–673. <https://doi.org/10.1038/s41587-023-01793-w>
- 318 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.
 319 *Molecular Biology and Evolution*, 37(12), 3507–3524.
- 320 Leroy, T., Louvet, J.-M., Lalanne, C., Le Provost, G., Labadie, K., Aury, J.-M., et
 321 al. (2020). Adaptive introgression as a driver of local adaptation to climate in
 322 european white oaks. *New Phytologist*, 226(4), 1171–1182.
- 323 Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*,
 324 34(18), 3094–3100.
- 325 Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with burrows-
 326 wheeler transform. *Bioinformatics*, 26(5), 589–595.
- 327 Li, K., Xu, P., Wang, J., Yi, X., & Jiao, Y. (2023). Identification of errors in draft
 328 genome assemblies at single-nucleotide resolution for quality assessment and
 329 improvement. *Nature Communications*, 14(1), 6556.
- 330 Li, Y., Yao, J., Sang, H., Wang, Q., Su, L., Zhao, X., et al. (2024). Pan-genome
 331 analysis highlights the role of structural variation in the evolution and envi-
 332 ronmental adaptation of asian honeybees. *Molecular Ecology Resources*, 24(2),
 333 e13905.
- 334 Li, Z., Liu, X., Wang, C., Li, Z., Jiang, B., Zhang, R., et al. (2023). The pig pangenome
 335 provides insights into the roles of coding structural variations in genetic diversity
 336 and adaptation. *Genome Research*, 33(10), 1833–1847.
- 337 Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B.,
 338 et al. (2022). Metagenome analysis using the Kraken software suite. *Nature
 339 Protocols*, 17(12), 2815–2839. <https://doi.org/10.1038/s41596-022-00738-y>
- 340 Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., et
 341 al. (2020). HiCanu: Accurate assembly of segmental duplications, satellites, and
 342 allelic variants from high-fidelity long reads. *Genome Research*, 30(9), 1291–1305.
- 343 Parmigiani, L., Garrison, E., Stoye, J., Marschall, T., & Doerr, D. (2024). Pan-
 344 cus: fast and exact pangenome growth and core size estimation. *Bioinformatics*.
 345 <https://doi.org/10.1093/bioinformatics/btae720>
- 346 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for
 347 comparing genomic features. *Bioinformatics*, 26(6), 841–842.
- 348 R Core Team. (2024). *R: A language and environment for statistical computing*.
 349 Vienna, Austria: R Foundation for Statistical Computing. Retrieved from
 350 <https://www.R-project.org/>
- 351 Rautianinen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-
 352 to-graph alignment. *Genome Biology*, 21(1). <https://doi.org/10.1186/s13059-020-02157-2>
- 353 Rendón-Anaya, M., Wilson, J., Sveinsson, S., Fedorkov, A., Cottrell, J., Bailey, M.
 354 E., et al. (2021). Adaptive introgression facilitates adaptation to high latitudes
 355 in european aspen (*populus tremula l.*). *Molecular Biology and Evolution*, 38(11),
 356 5034–5050.
- 357 Secomandi, S., Gallo, G. R., Rossi, R., Rodríguez Fernandes, C., Jarvis, E. D.,
 358 Bonisoli-Alquati, A., et al. (2025). Pangenome graphs and their applications
 359 in biodiversity genomics. *Nature Genetics*, 1–14.
- 360 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E.
 361 M. (2015). BUSCO: Assessing genome assembly and annotation completeness
 362 with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212.
- 363 Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C.,
 364 et al. (2021). Pangenomics enables genotyping of known structural variants
 365 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.