The western redcedar genome reveals low genetic diversity in a self-compatible conifer

Tal J. Shalev[1], Omnia Gamal El-Dien[1,2], Macaire M.S. Yuen[1], Shu Shengqiang[3], Shaun D. Jackman[4], René L. Warren[4], Lauren Coombe[4], Lise van der Merwe[5], Ada Stewart[6], Lori B. Boston[6], Christopher Plott[6], Jerry Jenkins[6], Guifen He[3], Juying Yan[3], Mi Yan[3], Jie Guo[3], Jesse W. Breinholt[7,8], Leandro G. Neves[7], Jane Grimwood[6], Loren H. Rieseberg[9], Jeremy Schmutz[3,6], Inanc Birol[4], Matias Kirst[10], Alvin D. Yanchuk[5], Carol Ritland[1], John H. Russell[5], Joerg Bohlmann[1]

Author affiliations: 1. Michael Smith Laboratories, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; **2.** Pharmacognosy Department, Faculty of Pharmacy, Alexandria University, Alexandria, 21521, Egypt; **3.** Department of Energy Joint Genome Institute, Lawrence Berkeley National Lab, Berkeley, CA, 94720, USA; **4.** Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, V5Z 4S6, Canada; **5.** British Columbia Ministry of Forests, Victoria, BC, V8W 9E2, Canada; **6.** HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806, USA; **7.** Rapid Genomics, Gainesville, FL, 32601, USA; **8.** Intermountain Healthcare, Intermountain Precision Genomics, St. George, UT, 84790, USA; **9.** Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; **10.** School of Forest, Fisheries and Geomatic Sciences, University of Florida, Gainesville, FL, 32603, USA

Corresponding authors: Tal J. Shalev and Joerg Bohlmann

Address: Michael Smith Laboratories, 2185 East Mall, Vancouver, BC, V6T 1Z4, Canada

Phone: 1-604-822-9673

Email: tal.shalev@msl.ubc.ca; bohlmann@msl.ubc.ca

Keywords: genome, genetic diversity, inbreeding, western redcedar, conifer, selfing

Running title: Genomics of western redcedar

Abstract: We assembled the 9.8 Gbp genome of western redcedar (WRC, *Thuja plicata*), an ecologically and economically important conifer species of the Cupressaceae. The genome assembly, derived from a uniquely inbred tree produced through five generations of self-fertilization (selfing), was determined to be 86% complete by BUSCO analysis – one of the most complete genome assemblies for a conifer. Population genomic analysis revealed WRC to be one of the most genetically depauperate wild plant species, with an effective population size of approximately 300 and no significant genetic differentiation across its geographic range. Nucleotide diversity, π, is low for a continuous tree species, with many loci exhibiting zero diversity, and the ratio of π at zero- to four-fold degenerate sites is relatively high (~ 0.33), suggestive of weak purifying selection. Using an array of genetic lines derived from up to five generations of selfing, we explored the relationship between genetic diversity and mating system. While overall heterozygosity was found to decline faster than expected during selfing, heterozygosity persisted at many loci, and nearly 100 loci were found to deviate from expectations of genetic drift, suggestive of associative overdominance. Non-reference alleles at such loci often harbor deleterious mutations and are rare in natural populations, implying that balanced polymorphisms are maintained by linkage to dominant beneficial alleles. This may account for how WRC remains responsive to natural and artificial selection, despite low genetic diversity.

Introduction

Gymnosperms are an ancient group of plants, with fossil records dating over 300 million years ago (MYA) (Stewart 1983). Conifers are by far the largest group of gymnosperms with approximately 615 known species (Christenhusz et al. 2011; Farjon 2018). The Pinaceae form the largest conifer family, and genomes of numerous members of the Pinaceae, such as white spruce (*Picea glauca*), Norway spruce (*Picea abies*), loblolly pine (*Pinus taeda*), sugar pine (*Pinus lambertiana*), and Douglas-fir (*Pseudotsuga menziesii*), have been sequenced (Birol et al. 2013; Nystedt et al. 2013; De La Torre et al. 2014; Zimin et al. 2014, 2017; Warren et al. 2015; Stevens et al. 2016; Neale et al. 2017). Such efforts revealed the notoriously complex nature of their immense genomes, which are rife with repetitive sequences, transposable elements, long introns, gene duplications, pseudogenes, and gene fragments.

However, little genomic research has been completed on conifers of other families. In particular, the Cupressaceae, such as cypresses, junipers, and redwoods, are thought to have undergone a whole genome duplication unique from the Pinaceae (Li et al. 2015). There is also evidence for substantial rearrangements of orthologous linkage groups during the evolutionary history of the two families, that resulted in differences in karyotype (*n* = 11 in Cupressaceae; *n* = 12 in Pinaceae) (De Miguel et al. 2015), genome size (9 – 20 Gbp in Cupressaceae; 18 – 31 Gbp in Pinaceae) (Hizume et al. 2001; De La Torre et al. 2014; Stevens et al. 2016), and likely other genomic differences. However, the genomes of only two Cupressaceae species, giant sequoia (*Sequoiadendron giganteum*) and the hexaploid coast redwood (*Sequoia sempervirens*), have been published (Scott et al. 2020; Neale et al. 2022).

Western redcedar (WRC, *Thuja plicata*) is an ecologically, economically, and culturally important species in the Cupressaceae. Endemic to the Pacific Northwest of North America and ranging from Northern California to Southern Alaska, WRC is a stress-tolerant, slow-growing tree prized for its durable, lightweight and rot-resistant wood (Grime 1977). WRC is one of only five extant *Thuja* species and is estimated to have diverged from its North American sister species *Thuja occidentalis* around 26 MYA (Li and Xiang 2005). Genetic studies have indicated low diversity in WRC (Copes 1981; Glaubitz et al. 2000; O'Connell et al. 2008). Microsatellite data suggest that all WRC originated from an isolated refugium near the southern end of its current distribution and radiated north and inland following the last glacial period (O'Connell et al. 2008). Current climate models predict that its range will increase over the next century, particularly in the interior of British Columbia (BC) (Gray and Hamann 2013), thus making it a priority for genome analysis and expediting of traditional breeding cycles via genomic selection (GS).

Uniquely among conifers, WRC employs a mixed mating system of outcrossing and self-fertilization (selfing), with a mean outcrossing rate of around 70% (El-Kassaby et al. 1994; O'Connell et al. 2001, 2004), and appears to suffer very little inbreeding depression for fitness growth traits (Wang and Russell 2006; Russell and Ferguson 2008). Mating systems in plants, particularly rates of selfing (*s*) and its complement, outcrossing ($1 – s$), are of interest to evolutionary biologists due to their implications for genetic diversity and fitness and have been investigated extensively over the past century (Stebbins 1957; Lande and Schemske 1985; Barrett and Eckert 1990; Barrett et al. 2003; Wright et al. 2013). Though inbreeding depression resulting from selfing can lead to negative fitness impacts, a benefit of selfing may include reproductive assurance (Fisher 1941; Baker 1955), which, in the absence of strong inbreeding depression, can allow self-compatible populations to expand their geographic range faster than obligate outcrossers (Lande and Schemske 1985). Research on inbreeding in plants has mostly focused on mating strategies in angiosperms (Barrett and Eckert 1990; Jarne and Charlesworth 1993; Vogler and Kalisz 2001; Barrett et al. 2003; Kalisz et al. 2004; Wright et al. 2013). Characterization of mixed mating systems in conifers has been limited, largely by their long generation times and generally high self-incompatibility (Sorensen 1982; Bishir and Namkoong 1987; Remington and O'Malley 2000; Williams et al. 2003; Williams 2008). The exceptional ability of WRC to maintain such a mating system has allowed for successful selfing for up to five generations in experimental trials, making WRC a potential model for the study of inbreeding in conifers and more broadly in gymnosperms (Russell and Ferguson 2008).

Here we introduce the first genome sequence for WRC and present unique features of the genome in the context of genetic diversity and the evolutionary history of WRC. We further explore the effects of extreme selfing on heterozygosity and selective pressures in multiple selfing lines (SLs).

Results

The WRC genome assembly represents a highly complete conifer genome

The assembly of conifer genomes remains challenging partly due to their large size (Nystedt et al. 2013; Warren et al. 2015; Stevens et al. 2016; Zimin et al. 2017) and high heterozygosity (Prunier et al. 2016). Given WRC's unique selfing abilities, we were able to facilitate assembly of a WRC reference genome using a fifth-generation SL tree (2323-211-S5; **Table S1**) expected to be > 98% homozygous. The S5 reference genome assembly was generated from a combination of short fragment paired-end reads, large fragment mate-pair reads, and linked-reads from large molecules, using 13 libraries and 28 lanes of Illumina sequencing, with a sequencing read length of 2×151 bp (**Table S2**). Overall genome depth of coverage was estimated to be 77×.

The WRC genome was previously estimated to be 12.5 Gbp in size across 11 chromosomes (Ohri and Khoshoo 1986; Hizume et al. 2001). GenomeScope (Vurture et al. 2017) estimated the genome size at 9.8 Gbp (**Figure S1**). We calculated approximately one single nucleotide variant (SNV) every 4.6 kbp, for an estimated genome-wide heterozygosity of 0.000216 – an exceptionally low estimate, highlighting the value of SLs for genome sequencing and assembly.

We assembled 7.95 Gbp of the estimated 9.8 Gbp genome to produce a draft assembly with an N50 of 2.31 Mbp, the largest scaffold being 16.3 Mbp. This assembly comprises 67,895 scaffolds > 1 kbp (**Table 1; Table S3**). Benchmarking Universal Single Copy Ortholog (BUSCO) analysis (Simão et al. 2015) determined the genome assembly to be 86% complete in the gene space. This is one of the highest completeness estimates for a conifer genome (**Table 2; Table S4A**).

Table 1. Assembly metrics and statistics for each version of the WRC draft genome.

Assembly version	N50 (Mbp)	NG50 (Mbp)	Largest scaffold (Mbp)	Size (Gbp)	L50 (bp)	LG50 (bp)	Scaffolds > 1kbp
redcedar-v1	1.45	1.07	9.79	7.95	1,642	2,463	94,166
redcedar-v2	2.23	1.63	15.3	7.95	1,067	1,605	90,083
redcedar-v3	2.31	1.71	16.3	7.95	1,035	1,551	67,895

Genome annotation using evidence from Iso-Seq full-length cDNAs resulted in the identification of 39,659 gene models supported by the alignment of unique primary transcripts (**Dataset S1**), and an additional 26,150 alternative transcripts. A total of 25,984 gene models had Pfam protein family annotation, 31,537 had transcriptome support over their full length (100%), and 19,506 had peptide homology coverage support 90% or greater (**Table S5; Dataset S2**). Intron length ranged from 20 bp to 148.3 kbp, which is consistent with estimates from other conifers, with maximum lengths ranging from 68 kbp (Norway spruce) (Nystedt et al. 2013) up to 579 kbp (sugar pine) (Stevens et al. 2016). Scott et al. (2020) reported a maximum intron length of 1.4 Mb in the highly contiguous giant sequoia genome assembly. Repeat elements comprised 60% of the WRC genome, which is low compared to other conifers. Repeats comprised 79% of the sugar pine (Stevens et al. 2016) and giant sequoia genomes (Scott et al. 2020). Single copy orthologs (SCOs) were detected by orthogroup comparison to the giant sequoia gene set, yielding 11,937 SCOs (**Dataset S3**).

BUSCO analysis found the predicted gene set to be 90.5% complete, much higher than any other conifer gene set to date (**Table 2; Table S4B**). We further validated the completeness of the genome assembly and annotation using a panel of 59 full-length WRC sequences from GenBank, of which 48 were reliably identified in the genome annotation. We also searched for a set of 33 WRC terpene synthase (TPS) transcripts, of which we reliably (>90% identity) identified 15 (Shalev et al. 2018) **(Table S6**; **Dataset S4**). This confirms the completeness of the gene space and quality of the draft genome annotation, while suggesting that BUSCO core genes may somewhat overestimate gene space completeness when considering family or species-specific genes.

Table 2. BUSCO genome assembly and predicted gene set completeness of seven currently available conifer genome assemblies.

Taxon	***Thuja plicata***	***Sequoiadendron giganteum*** (Scott et al. 2020)	***Picea glauca*** (Warren et al. 2015)	***Picea abies*** (Nystedt et al. 2013)	***Pinus lambertiana*** (Stevens et al. 2016)	***Pinus taeda*** (Zimin et al. 2014)	***Pseudotsuga menziesii*** (Neale et al. 2017)
	Western redcedar	**Giant sequoia**	**White spruce**	**Norway spruce**	**Sugar pine**	**Loblolly pine**	**Douglas- fir**
Family	Cupressaceae	Cupressaceae	Pinaceae	Pinaceae	Pinaceae	Pinaceae	Pinaceae
Genome completeness (%)	86.0	84.3	49.7	34.9	61.5	49.7	74.1
Gene set completeness (%)	90.5	49.9	18.0	28.1	73.3	41.7	68.5

Genome completeness and gene set completeness were estimated in genome mode and protein mode, respectively, on the Embryophyta OrthoDB v10 database. MetaEuk was used for gene prediction in genome mode.

Population genomic analysis reveals extremely low levels of genetic diversity in WRC

We estimated nucleotide diversity, short-range linkage disequilibrium (LD), population structure, genetic differentiation, and effective population size (N_e) in n =112 unrelated trees from across the geographic range of WRC (range-wide population; RWP) (**Table S7**). Trees were grouped into three subpopulations: Northern-Coastal (n = 77), Central (n = 26), and Southern-Interior (n = 9) (**Figure 1A**). Using a panel of single nucleotide polymorphisms (SNPs) that were genotyped via targeted sequence capture approach, we identified 2,454,925 variant and invariant sites, which were filtered separately and resulted in sets of 18,371 SNPs (**Dataset S5**) and 2,186,998 invariant sites (see **Materials and Methods**). Total mean SNP depth was 34.3×.

We annotated 17,728 SNPs across 2,886 genomic scaffolds using the Ensembl Variant Effect Predictor (VEP) (McLaren et al. 2016) (**Dataset S6**). We detected 13,097 SNPs within 5,045 genes, 3,288 of which were SCOs. Intergenic loci made up 25.2% of all annotated SNPs (4,631), 1,105 of which were in regions 0.5 to 2 kbp up- or downstream of coding regions. Within coding regions, 50.0% (3,002) were missense variants while 46.7% (2,807) were synonymous variants (**Table S8**).

Linkage Disequilibrium

Decay of linkage disequilibrium (LD), the non-random association of alleles at different loci in a population, can inform on how likely different loci are to be assorted together during recombination. We assessed short-range LD as represented by the squared correlation coefficient r^2 in the RWP, at a minor allele frequency (MAF) threshold of 0.05 to avoid bias due to rare alleles (n = 16,202 SNPs). The mean of all pairwise r^2 estimates was 0.299 with a median of 0.151. The half-decay value (the distance in which r^2 decays to half of the 90th percentile value) was 0.118 Mbp. LD decayed to an r^2 of 0.2 at 0.751 Mbp, and an r^2 of 0.1 at 2.17 Mbp (**Figure 2A**). Further, high LD ($r^2 > 0.8$) appears to exist for SNPs millions of bp apart (**Figure 2B**). These estimates for LD decay are several orders of magnitude greater than those found in other conifers, as well as many other tree species, where LD has been reported to decay rapidly within tens to a few thousand bp (Krutovsky and Neale 2005; Heuertz et al. 2006; Pyhäjärvi et al. 2011; Pavy et al. 2012; Fahrenkrog et al. 2017).

Population structure and genetic differentiation

We analysed STRUCTURE (Pritchard et al. 2000) results using two post-hoc cluster identification methods on a filtered set of n = 4,765 SNPs (see **Methods and Materials**). The ΔK method of Evanno et al. (2005) identified an optimal K of two; this approach may return a K of two more often than expected when genetic structure is weak (Janes et al. 2017). The approach of Puechmaille (2016), which can help resolve K when subsampling is uneven, identified an optimal K of two as well. Analysis of fastStructure (Raj et al. 2014) results using cross-validation suggested that optimal K may lie between one and three. These results suggest genetic structure is exceptionally weak in our RWP. Indeed, there is apparent gene flow between trees in all three subpopulations across all three STRUCTURE clusters (**Figure 1B**).

We applied non-parametric approaches of Discriminant Analysis of Principal Components (DAPC) and Principal Component Analysis (PCA) using a set of n = 13,427 SNPs from SCO and intergenic regions. Cross-validation for DAPC with *a priori* cluster definitions optimally retained 22 PCs and two discriminant functions capturing 30.3% of the conserved variance; however, *de novo k*-means clustering failed to resolve any clusters, identifying an optimal K of one (**Figure S2**). PCA revealed a latitudinal gradient of differentiation along the first principal component (PC), with some separation of the Southern-Interior subpopulation along the second PC (**Figure 1C)**, mostly for trees originating from California and Oregon. However, the first PC only explains 3.73% of the variance in the data, and the second explains 1.63%. These results are consistent with DAPC and suggest that gene flow has been prevalent across the range of WRC. No significant differentiation was found between trees from different subpopulations based on a hierarchical F_{ST} test (F_{ST} = 0.0334, p = 0.726) (**Table S9**), and no significant isolation by distance was found by our Mantel test for subpopulations (r = -0.241, p = 0.672) and individuals (r = 0.0833, p = 0.121) (**Figure S3**).

Nucleotide diversity in WRC

We estimated nucleotide diversity π (Nei and Li 1979) in the RWP and absolute nucleotide divergence d_{XY} between subpopulations using all SCO and intergenic SNPs. Average π (SD) across 10,631 SCOs was 0.00272 (0.0122) (**Figure 3A**; **Figure S4**; **Table S10**); 1,411 genes had a π of zero. Across 10 kb windows, average π was 0.00204 (0.0141), indicating diversity is similar in coding and noncoding regions. Average d_{XY} was not significantly different between any pair of subpopulations nor was it significantly different from π in SCOs ($p > 0.05$, Kruskal-Wallis rank sum test) (**Figure 3B**; **Table S11**).

To assess the efficacy of purifying selection, we estimated π_0/π_4, the ratio of π in 0-fold to 4-fold degenerate sites. We found a π_0 of 0.00158 (0.0146) and a π_4 of 0.00485 (0.0147), yielding a π_0/π_4 of 0.325. The site frequency spectrum (SFS) for 4-fold SNPs appeared to decay slower than the SFS for 0-

fold and for all SNPs, supporting evidence of a recent bottleneck and indicating that there may be stronger positive selection at these sites (**Figure S5**).

Effective population size (N_e)

N_e can be defined as the idealized population size expected to experience the same rate of loss of genetic diversity as the population under observation (Wright 1931). We estimated N_e using the LD method of NeEstimator (Do et al. 2014). SNPs were mapped to the giant sequoia genome (Scott et al. 2020) (**Dataset S7**) and SNPs estimated to be at least 2.17 MB apart were isolated for the analysis, resulting in a set of $n = 412$ SNPs. N_e was estimated to be 270.3 (JackKnife 95% CI: 205.5, 384.6).

We further explored demography using Stairway Plot 2 (Liu and Fu 2020). We observe a decline in N_e from ~ 500,000 beginning ca. 2 MYA, accelerating from ~40 KYA down to under 300 by present day, consistent with one or more bottleneck events during the recent glacial maximum (**Figure S6**). Our estimates of N_e are extremely low for a continuous tree population; for example, species in *Picea* (Chen et al. 2010), *Pinus* (Brown et al. 2004), and *Populus* (Fahrenkrog et al. 2017) have estimated N_e in the range of $10^4 – 10^5$.

Persistent heterozygosity during complete selfing highlights genomic regions under selection

To examine the effect of complete selfing on heterozygosity and selection in WRC, we selected 189 trees from the 15 FS families, forming 41 SLs for SNP genotyping. The process of SNP calling is error-prone, and despite filtering for multiple quality criteria, errors are likely to remain in any SNP data set. Using SLs, we were able to correct for erroneous genotyping calls and impute missing genotypes for SLs up to S4 ($n = 28$) or S5 ($n = 11$), retaining $n = 151$ trees (**Dataset S8**). We used all filtered SNPs for these analyses ($n = 18{,}371$ SNPs).

Under selfing in diploids, heterozygosity is expected to decline by 50% in each generation purely through genetic drift. Mean heterozygosity declined slower than expected (**Figure 4A**; **Table S12; Table S13A**), while median observed heterozygosity was significantly lower than expected beginning in generation S3 ($p < 0.05$, pairwise Sign test; **Table S13B**). Mean heterozygosity at FS was 0.296, while in the RWP it was 0.219 (**Table S12**). In comparison, Chen et al. (2013) found mean heterozygosities of 0.33 and 0.36 in lodgepole pine (*Pinus contorta*) and white spruce, respectively.

The inbreeding coefficient F is the probability of any two alleles being identical by descent (IBD) and is a measure of reduction in heterozygosity due to inbreeding. We estimated F for each sample in the SLs and the RWP using the approach of Yang et al. (2011) (F_{UNI}). Mean F increased from 0.00569 at FS to 0.801 at S5 – significantly less than expected ($p < 0.05$, one-sample t-test), indicating that the observed reduction in heterozygosity cannot entirely be attributed to inbreeding (**Figure 4B**; **Table S12**; **Table S13C**). Mean F in the RWP was 0.331, further emphasizing the degree of inbreeding in wild populations (**Table S12**).

Following the expectation of a 50% decline in heterozygosity per generation under complete selfing and assuming a model of only genetic drift, we anticipated that 25% of SLs would become fixed for one allele at any given locus and 25% for the other in each generation. Thus, by generation S4, 46.875% of SLs should fix for each allele, and 6.25% should remain heterozygous. We identified 83 SNPs that deviated from expected proportions of fixation at a false discovery rate threshold of 0.05 (hereafter: outlier SNPs) (**Dataset S9**). Of these, 15 fixed for the reference allele and 2 fixed for the alternate allele more often than would be expected under drift alone; meanwhile, all outlier SNPs had a higher proportion of heterozygous alleles by S4 than expected under drift. Outlier SNPs were present on all putative LGs in the genome, and mean depth was similar to the mean depth of the total SNP set (29.3× vs 34.3×, respectively). VEP predicted effects for 67 outlier SNPs, 14 (16.9%) of which were in coding regions (**Dataset S10**). Gene Ontology (GO) annotation was available for 30 genes containing outlier

SNPs; ten GO categories were over-represented in outliers when compared to the entire SNP set ($p <$ 0.05, Fisher's Exact Test) (**Table S14**).

When comparing SNP effect categories, we found intergenic variants (1.76×10^{-5}) to be over-represented in outlier SNPs, while synonymous variants ($p = 0.0274$) and 3' UTR variants ($p = 0.00993$) were under-represented (**Figure S7**). In the RWP, the minor allele for these SNPs was nearly always the alternate allele, i.e., the allele inducing the change. This pattern suggests that balanced polymorphisms may be maintained by selection favouring linked dominant alleles, i.e., associative overdominance (Bierne et al. 2000).

Discussion

Genome analysis of WRC, a self-compatible conifer, revealed low genetic diversity, high levels of LD, and low N_e across its geographic range. WRC emerges as a genetically depauperate wild plant species, providing insight into how selfing may have facilitated its expansion across its current geographic range, but at the expense of genetic variation.

The WRC genome

Sequence assembly of large and repetitive conifer genomes is becoming more feasible, with new technologies such as Single Molecule Real Time (SMRT) long-read sequencing or linked-reads (Zimin et al. 2017). WRC is one of only two conifers outside of the Pinaceae whose genome sequence has been published. Recently, Scott et al. (2020) reported the first genome sequence in Cupressaceae, giant sequoia, with a near-chromosome-scale assembly of 8.125 Gbp of the estimated 9 Gbp genome using a combination of Oxford Nanopore long reads and Illumina short reads together with a Dovetail HiRise Chicago and Hi-C statistical scaffolding and assembly approach. Though future chromosome-level assembly would be of value to improve contiguity, BUSCO completeness scores (86.0% and 84.3% for WRC and giant sequoia, respectively) and the very high completeness of the annotated gene set suggest that the WRC assembly is currently of very high quality for the gene space. Previous studies using flow cytometry estimated WRC's genome size at 12.0 – 12.5 Gbp (Ohri and Khoshoo 1986; Hizume et al. 2001). The WRC genome assembled at 7.95 Gbp. This discrepancy may be partially explained by filtering of *k*-mers with very high depth of coverage in GenomeScope to remove organelle-derived reads, which may also remove other heterochromatic sequences such as centromeres and telomeres; however, a recent study in maize (*Zea mays*) found that selfing over several generations can reduce genome size by up to 7.9% (Roessler et al. 2019), which may suggest that we are observing genome loss in WRC as well given the genome assembly source. Flow cytometry to assess genome loss during selfing would be a valuable future endeavour.

Genetic diversity in WRC

We estimated π to be 0.0027 in SCOs and 0.0020 across all sequenced space. These estimates are lower than many other plant species using comparable methods, for example, Norway spruce (π = 0.0049 – 0.0063) (Wang et al. 2020), weedy broomcorn millet (π = 0.14; *Panicum miliaceum*) (Li et al. 2021), and most *Populus* species (π = 0.0041 – 0.011) (Liu et al. 2022). We found lower π estimates only in highly cultivated plants, such as soybean (π = 0.0015; *Glycine* spp.) (Bayer et al. 2022), or rare, isolated species, such as *Populus qiongdaoensis* (π = 0.0014), which is restricted to a single small island and has an estimated N_e of ~500 (Liu et al. 2022). Additionally, our probe selection strategy for genotyping targeted regions of high variability due to the very low levels of polymorphism in initial sequencing runs. This may have led to inflated estimates of π, suggesting that genome-wide diversity may be even lower.

The relatively high observed π_0/π_4 ratio (0.33) may suggest weak purifying selection in WRC (Chen et al. 2017); it could also be indicative of demography, as π_0 returns to equilibrium quicker than π_4

following a bottleneck event (Brandvain and Wright 2016; Chen et al. 2019). The low N_e (~ 300) and general lack of population structure, genetic differentiation, or nucleotide divergence between geographic subpopulations despite a relatively wide geographic range and continuous population suggest that much of the variation in WRC was likely eliminated due to bottlenecks following the last glacial period, a pattern confirmed by our Stairway Plot results and affirming the conclusions of previous studies (Copes 1981; Glaubitz et al. 2000; O'Connell et al. 2008). Mating system likely plays a role as well in WRC's low diversity. It has been argued that selfing species should generally have a lower nucleotide diversity due to a reduction of the effective recombination rate (Buckler IV and Thornsberry 2002). Thus, the exceptionally slow rate of LD decay observed in our RWP is further evidence of a recent population bottleneck or long-term effects of inbreeding (Golding and Strobeck 1980; Zhang et al. 2004; Slatkin 2008). In future studies, more extensive sampling, in particular for the Southern-Interior region, could help in gaining more accurate estimates of genetic differentiation across populations of WRC.

Selfing in WRC

Heterozygosity declined faster than expected under complete selfing. Despite starting at an F of nearly 0, our FS generation had a low mean heterozygosity (0.296), and with each successive generation, F increased slower than expected, indicating that IBD does not fully explain the reduction in heterozygosity in WRC (Wright 1922; Slate et al. 2004). The lack of strong fitness costs associated with selfing in WRC (Wang and Russell 2006; Russell and Ferguson 2008) suggests that most strongly deleterious alleles have been purged from the genome, presumably due to past population bottlenecks and inbreeding. Nonetheless, even weak purifying selection could explain the faster than expected decline in heterozygosity.

Of greater interest, however, is that the majority of loci deviating from expectations of drift during selfing remained heterozygous, suggestive of balancing selection or associative overdominance at these loci, with high levels of LD promoting genetic hitch-hiking near loci under selection to remain heterozygous. The presence of missense variants in outlier loci coupled with the general rarity of missense mutations in natural populations offers further support for associative overdominance as an explanation for the retention of heterozygosity at these loci. This is congruent with relatively high π_0/π_4 in the RWP, suggesting strong positive selection is maintaining current allele frequencies. No outlier loci remained heterozygous in all lines, which suggests that these loci do not harbour strongly deleterious or lethal mutations. Further, all three genotypes exist for many of these loci in the RWP.

Excess heterozygosity can also occur from genotyping error due to the presence of paralogs. To address this source of error, we employed stringent filters for maximum mean depth, allele balance, excess heterozygosity, read-ratio deviations, and deviations from HWE. Further, the low estimated π in the RWP as well as similar mean depths (~30×) for outlier SNPs and the total SNP set suggest paralog content is minimal. Higher than expected heterozygosity was observed during selfing in eucalypts (*Eucalyptus grandis*) (Hedrick et al. 2016) and maize (Roessler et al. 2019). However, the average heterozygosity in these species is notably much higher (~ 0.65 in each for S1, compared with 0.15 at S1 for WRC). It is also possible that our genotyping strategy, in which probes were designed to capture highly variable sites, may have influenced heterozygosity estimates. Future analyses using whole-genome sequencing or comprehensive genotyping-by-sequencing (GBS) for comparison may be of value.

We recognize that use of SLs of single seed descent makes differentiating between patterns of selection and genetic drift difficult, as genetic drift is stronger when there are fewer individuals in a population. The use of multiple cloned seedlings for each SL in future studies could help improve our analysis, with the potential to find more SNPs under selection.

Implications for conservation, adaptation to climate change and breeding with genomic selection (GS)

Current breeding of WRC focuses on traits such as growth and herbivore and disease resistance; thus, low genetic diversity may have considerable ecological and potential economic consequences. When low genetic diversity is observed in plant or animal populations, conservation strategies may become necessary to maintain existing genetic variation and reduce the risk of extreme inbreeding depression, especially when census population size in the wild is small. Although ours and previous results (O'Connell et al. 2008) indicate its range was likely reduced to a single refugium during the last glaciation, WRC has since greatly expanded throughout the Pacific Northwest. We found genetic isolation by distance to be small, consistent with the low observed variation. Yet, successful selection of genetically superior families for these traits has been possible. Provenance trials have revealed significant local adaptation among natural populations of WRC (Cherry 1995), and WRC can be found in a variety of different climates, moisture levels, elevations, and light availabilities (Grime 1977; Antos et al. 2016). Resistance to cedar leaf blight, a foliar fungal pathogen, has been observed to be an adaptation to native climate, with trees from wetter climates showing greater resistance than those from drier climates, regardless of geographical distance (Russell et al. 2007). These observations suggest sufficient genetic variation exists within and between natural populations upon which selection can act. Furthermore, WRC is well known for its high phenotypic plasticity (El-Kassaby 1999), possibly due to epigenetic variation (Zhang et al. 2013), although the fraction of plasticity that is adaptive remains unknown. Our observation of balanced polymorphisms, due in part to associative overdominance, offers a potential explanation for WRC's reported adaptability and response to selection. Together with self-compatibility, which is known to facilitate range expansion (Baker 1955), WRC may be less threatened by climate change and other anthropogenic pressures than might be expected based on its low genetic diversity.

WRC's apparent adaptability and potential for range expansion make it an important forest tree in a time of changing climate and environments (Gray and Hamann 2013). Low genetic diversity and unique mating system need to be considered as WRC breeding adopts strategies of GS that largely rely on controlling relatedness in the population (Ritland et al. 2020). Recombination rate is another important consideration, as GS relies on the presence of LD between SNPs and causal regions for traits, in addition to relatedness between individuals (Meuwissen et al. 2001). WRC's high LD may be an advantage for finding linked SNPs, but may also increase the risk of unintentional selection for correlated traits not under selection. This could be mitigated by whole genome sequencing across breeding populations, similar to GS approaches in livestock breeding (Raymond et al. 2018; Georges et al. 2019).

WRC is a fascinating example of adaptation in a long-lived conifer, despite very low levels of genetic variation. As our understanding of the genome improves, we will be able to improve prospects for survival and maintenance of this tree as an ecologically and economically significant species and better understand and test how selfing behaviour evolves and can be advantageous in wild plant populations.

Methods and Materials

Plant materials

The WRC RWP represented n = 112 individuals originating from across the geographic range growing at the Cowichan Lake Research Station (CLRS) at Mesachie Lake, BC, Canada. Trees were separated into three geographic subpopulations, Northern-Coastal, Central, and Southern-Interior, based on UPGMA clustering of genetic distances (O'Connell et al. 2008). SLs were produced over 12 years (1995 – 2007) using an accelerated breeding approach (Russell and Ferguson 2008) at CLRS. Briefly, 15 pairs (30 individuals) of unrelated parents from across coastal BC and Vancouver Island (**Table S15**) were crossed to create 15 FS families and ensure an initial inbreeding coefficient of F = 0. Each FS line was then selfed for up to five generations (S1 – S5) with one generation every two years, facilitated by GA_3 hormone treatment. A single S5 seedling of SL 23 (2323-211-S5) was used for genome sequencing. We selected 189 individuals from the 15 FS selfing families for genotyping for the SL analysis (**Table S1**).

Genome sequencing and assembly

Foliar tissue was used for DNA extraction for genome sequencing. Purified nuclear genomic DNA was extracted at BioS&T (http://www.biost.com/, Montreal, Canada) (Birol et al. 2013) and sequenced at the Joint Genome Institute (JGI; Berkeley, USA).

Genome sequencing was executed using three types of libraries: short fragment paired-end, large fragment mate-pair, and linked-reads from large molecules using 10× Genomics Chromium. Depth of *k*-mer coverage profiles were computed for multiple values of *k* using ntCard v1.0.1 (Mohamadi et al. 2017) (**Figure S8**). The largest value of *k* providing a *k*-mer coverage of at least 15 was selected, based on an estimated coverage of > 99.9%, yielding *k* = 128 (Lander and Waterman 1988). We analyzed and visualized *k*-mer profiles using GenomeScope v1.0.0 (Vurture et al. 2017). Paired-end reads were assembled using ABySS v2.1.4 (parameters: k=128; kc=3) and scaffolded using the mate-pair reads with ABySS-Scaffold (Jackman et al. 2017) (**Figure S9**). Linked-reads were aligned and misassemblies were identified and corrected with Tigmint v1.1.2 (Jackman et al. 2018). The assembly was scaffolded using the linked-reads with ARCS v1.0.5 (Yeo et al. 2018) (-c 2; -m 4-20000) and ABySS-Scaffold (-n 5-7; -s 5000-20000). Molecule size of the linked read libraries was estimated using ChromeQC v1.0.4 (https://bcgsc.github.io/chromeqc). Detailed DNA extraction, sequencing, and assembly methods can be found in **Supplemental Methods**.

We estimated completeness of the WRC genome assembly and other conifer genome assemblies using BUSCO v5.0.0 in genome mode, on OrthoDB Embryophyta v10 (Simão et al. 2015; Waterhouse et al. 2018; Kriventseva et al. 2019), which determines the proportion and completeness of single-copy genes from the Embryophtya database (1,614 models) present in the genome.

Genome annotation

For PacBio Iso-Seq, full-length cDNAs were synthesized from total RNA. We then generated transcript assemblies from 1.4B 2×150 and 50M 2×100 stranded paired-end Illumina RNA-seq reads using PERTRAN (Shengqiang et al. 2013), 18M PacBio Iso-Seq Circular Consensus Sequences (CCS), and previous RNA-seq assemblies (Shalev et al. 2018) (NCBI PRJNA704616). We determined gene loci by transcript assembly alignments and EXONERATE v2.4.0 (Slater and Birney 2005) alignments of proteins from *Arabidopsis thaliana*, *Glycine max*, *Populus trichocarpa*, *Oryza sativa*, *Vitis vinifera*, *Aquilegia coerulea*, *Solanum lycopersicum*, *Amborella trichopoda*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Sphagnum magellanicum*, UniProt Pinales and Cupressales, and Swiss-Prot proteomes to the repeat-soft-masked WRC genome using RepeatMasker v4.0.8 (Smit et al. 2015) with up to 2 kbp extension on both ends. Gene models were predicted using FGENESH+ v3.1.1 (Salamov and Solovyev 2000), FGENESH_EST v2.6, and EXONERATE and PASA (Haas et al. 2003) assembly ORFs. The best-scored predictions for each locus were selected and improved by PASA, adding untranslated regions (UTRs), splicing correction, and alternative transcripts. All software was run using default parameters. Detailed RNA extraction, sequencing, and genome annotation methods can be found in **Supplemental Methods**.

We estimated completeness of the WRC primary transcript gene set and other conifer gene sets using BUSCO v5.0.0 in protein mode on OrthoDB Embryophyta v10. We also assessed coverage of 59 complete WRC sequences found on NCBI and 33 WRC terpene synthase sequences (Shalev et al. 2018) . Sequences were searched against the genome using BLAST+ v2.10.0 (Altschul et al. 1990; Camacho et al. 2009), and presence or absence analyzed using EXONERATE. SCOs were identified using OrthoFinder v2.5.4 (Emms and Kelly 2019), isolating genes identified in one copy in the WRC gene set when compared against the giant sequoia gene set (Scott et al. 2020).

SNP genotyping

DNA was isolated from lyophilized tissue with a modified protocol of Xin and Chen (2012). Targeted sequencing-based genotyping was done by Capture-Seq methodology at Rapid Genomics (Gainesville FL, USA). Initially, probes were designed using only limited publicly transcriptome data and database matches for functionally characterized conifer genes, an approach that has worked previously for other organisms (e.g., Mukrimin et al. 2018; Vidalis et al. 2018; Acosta et al. 2019; Telfer et al. 2019); however, this approach yielded less than 2,000 polymorphic sites. Thus, we developed a specialized probe design approach targeting regions of putative high variability, specifically: previously identified differentially expressed regions from cold-tolerance, deer browse, wood durability, leaf blight, and growth trials, database matches for functionally characterized conifer genes, and whole transcriptome and genome data (NCBI Umbrella BioProject PRJNA704616). A set of 57,630 probes, 37,294 targeting genic regions and 19,706 targeting intergenic regions, was designed for marker discovery, from which a panel of 20,858 probes was selected for genotyping.

Putative SNPs were identified using FreeBayes v1.2.0 (Garrison and Marth 2012) in 150bp on either side of the probes and filtered probes that had more than 17 SNPs per 420 bp target region to prevent over-capture. Sequencing depth was used to select the final set, removing probes with low and high sequencing depth for Capture-Seq on the remainder of the samples. Detailed methods for SNP genotyping can be found in **Supplemental Methods**.

SNP filtering and annotation

Variant sites were filtered using VCFtools v0.1.17 (Danecek et al. 2011) with the following flags: --max-missing 0.95; --minQ 30; --min-meanDP 15; --max-meanDP 60. SNPs with an allele balance > 0.2 and < 0.8 or < 0.01 were retained to eliminate incorrectly called heterozygotes using vcffilter in vcflib v1.0.1 (Garrison 2016). To eliminate paralogous loci, we excluded: SNPs with a read-ratio deviation score *D* (McKinney et al. 2017) > 5 and < -5, SNPs with a heterozygosity greater than 0.55, and SNPs with excess heterozygosity and deviations from HWE in the RWP at a *p*-value cutoff of 0.05 and 1e-5, respectively. We also excluded SNPs with negative inbreeding coefficients (F_{IS}), using the formula:

$$F_{\mathrm{IS}} = 1 - \frac{H_{\mathrm{O}}}{H_{\mathrm{E}} \times (1 - 0.17)}$$

where H_{O} is the observed heterozygosity of the locus, H_{E} is the expected heterozygosity of the locus under HWE, and the factor of (1 – 0.17) accounts for the expected equilibrium fixation index of 0.17 in WRC based on the average outcrossing rate of 0.7 (El-Kassaby et al. 1994; O'Connell et al. 2001, 2004). Invariant sites were filtered using the following flags: --max-missing 0.95; --min-meanDP 15; --max-meanDP 60. Variant effect prediction was carried out using the Ensembl Variant Effect Predictor r103; one effect per SNP was selected, and for compound effects, only the most severe consequence was retained (McLaren et al. 2016). Relationships between trees were estimated by generating a genomic realized relationship matrix for all individuals using the 'A.mat' function of rrBLUP v4.6.1 in R (Endelman 2011; R Core Team 2021). For the RWP, five trees with a relatedness coefficient > 0.2 were excluded from analyses for a total of *n* = 112 trees. For the SLs, nine individuals whose relationships did not match the *a priori* pedigree were removed from analysis (**Table S1**).

Linkage disequilibrium

Pairwise LD (r^2) was estimated in the RWP using PLINK v1.9 (Chang et al. 2015). LD was calculated for all scaffolds containing at least two SNPs (--r2; --ld-window-r2 0; --ld-window-kb 999999; --ld-window 999999; --maf 0.05). Under drift-recombination equilibrium, our expectation of LD decay over distance will be:

$$E\left(r^2\right) = \frac{1}{(1 + C)}$$

Where C is the product of the population recombination parameter $\rho = 4N_e r$ and the distance in bp, N_e is the effective population size and r is the recombination rate per bp (Sved 1971). Adjusting for population size n and a low level of mutation, decay of LD was estimated as a factor of n, and C (Hill and Weir 1988; Remington et al. 2001; Marroni et al. 2011; Fahrenkrog et al. 2017).

$$E(r^2) = \left[\frac{10 + C}{(2 + C)(11 + C)}\right]\left[1 + \frac{(3 + C)(12 + 12C + C^2)}{n(2 + C)(11 + C)}\right]$$

C was estimated by non-linear regression, using the nls function in R.

Population structure and genetic differentiation

For STRUCTURE (Pritchard et al. 2000; Falush et al. 2007) analysis, SNPs were further filtered to include only SNPs with an r^2 threshold of 0.1 using a 2.17 Mb window, following the outcome of our LD analysis, a MAC threshold of 3 was used to remove singletons, and only SCO and intergenic SNPs were retained (n = 4,765). For all other genetic differentiation analyses, all SCO and intergenic SNPs were used (n = 13,427). Population structure of the RWP was estimated using STRUCTURE v2.3.4 and a DAPC using the adegenet package (v2.1.1) in R (Jombart 2008; Jombart et al. 2010; Jombart and Ahmed 2011). A hierarchical F_{ST} test as implemented in the hierfstat package (v0.04-22) in R (Goudet 2005) was used to assess genetic differentiation between subpopulations in the RWP and a Mantel test was executed using mantel.randtest for 9,999 permutations in ade4 v1.7-15 in R (Dray and Dufour 2007) to assess isolation by distance. PCA was performed on the genotype matrix of each subpopulation to visualise genetic distance between individuals using ade4.

STRUCTURE software was run using 10,000 MCMC repetitions with 10,000 repetitions of burn-in and 10 iterations of each K. Results were analyzed using methods of Evanno et al. (2005) and Puechmaille (2016) of cluster and admixture estimation and selection; fastStructure v1.0 (Raj et al. 2014) was also used with 10-fold cross-validation. For DAPC, find.clusters was used to select the optimal number of clusters based on Bayesian Information Criterion (BIC), and 10-fold cross-validation with 1,000 replicates was performed with xvaldapc to select the number of PCs and discriminant functions to retain.

Nucleotide diversity

To avoid downward bias due to missing data, SCO and intergenic variant sites (n = 13,427) and all invariant sites were used to estimate π and d_{XY} for n = 10,631 SCO genes and 10 kb windows in the RWP using pixy v1.2.7.beta1 (Korunes and Samuk 2021). Zero- and four-fold degenerate variant and invariant sites were identified using the NewAnnotateRef.py script (Williamson et al. 2014), and π_0/π_4 was calculated over all SCO genes. SFS was estimated using easySFS (https://github.com/isaacovercast/easySFS).

Effective population size (N_e)

We estimated N_e using the LD model estimation method under random mating as implemented in NeEstimator v2.1 (Do et al. 2014). This method uses background LD shared among samples to estimate N_e; thus, SNPs with as little LD as possible are required (Waples and Do 2008; Gilbert and Whitlock 2015). Due to the high LD observed in WRC, we first generated putative linkage groups (LGs) for the WRC genome by aligning all genomic scaffolds containing SNPs to the giant sequoia genome (Scott et al. 2020) using BLAST+. Scaffolds were assigned to their most likely LG based on bitscore. We then used the nucmer command from MUMmer v4 (Marçais et al. 2018) to determine the most likely alignment region for each scaffold in each LG. We retained SNPs estimated to be at least 2.17 Mbp apart. A MAF threshold of 0.05 was established to eliminate bias that may be introduced by rare alleles, and a 95% nonparametric JackKnife confidence interval was taken for the estimated value, as recommended by Waples and Do (2008) and Gilbert and Whitlock (2015).

Stairway Plot 2 (Liu and Fu 2020) was used on the folded SFS from intergenic and 4-fold degenerate positions to further assess N_e changes over time. We used the following parameters: nseq = 222; L = 238,557; pct_training = 0.67; nrand = 55, 110, 165, 220; ninput = 200; mu = 3.74e-9; year_per_generation = 50.

Genotype correction for SLs

Genotype correction in continuous SLs used two criteria: Individuals with homozygous calls in at least two consecutive generations were considered to be homozygous for that allele in all subsequent generations; and individuals with heterozygous calls in at least two consecutive generations were considered to be heterozygous for all preceding generations, up to and including the FS generation. SNPs that could not be corrected following these criteria were removed. We manually corrected genotypes for SLs which had been completely genotyped from either S1 – S4 or S1 – S5 (**Table S1**). For seven SLs in which only the S3 generation had not been sequenced, we imputed the genotypes for the S3 generation at each locus for SNPs where no other genotype was possible, and marked the rest as missing.

Change in heterozygosity and inbreeding coefficients over time

Corrected genotypes for all filtered SNPs (n = 18,371) were used to calculate the observed and expected changes in heterozygosity and inbreeding coefficients over time in the SLs, and in the RWP. Observed heterozygosity was calculated at each SNP locus for each generation across all SLs using the adegenet package in R. Expected heterozygosities in the S1 – S5 generations were calculated for each SNP locus as half the observed heterozygosity of the previous generation. We calculated inbreeding coefficients in PLINK using the --ibc flag to obtain a measure for inbreeding based on the correlation between uniting gametes (F_{UNI}). This metric is defined by Yang et al. (2011) for each ith SNP and each jth individual as

$$F_{\mathrm{UNI}} = \frac{x_{ij}^2 - (1 + 2p_i)x_{ij} + 2p_i^2}{2p_i(1 - p_i)}$$

Where x is the number of copies of the reference allele and p is the population-wide allele frequency at that locus. Calculations of F_{UNI} do not consider LD; thus, we used SNPs filtered for LD and MAC (n = 6,123). In a diploid population, F should increase as $F_{t+1} = \frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right) F_t$ each generation; thus, under complete selfing we expect F to increase by a factor of $F_{t+1} = \frac{1}{2}(1 + F_t)$. F was calculated using corrected genotypes in SLs and uncorrected genotypes in the RWP.

Isolating SNPs significantly deviating from expectations of drift

To identify loci that diverge from patterns expected under genetic drift, we evaluated all SNPs for which the FS generation was heterozygous and then observed whether the SNP went to fixation or not by the S4 generation in our 28 complete, corrected SLs. The S5 generation was excluded from this analysis due to small sample size. For statistical analysis, each SL was considered an independent replicate. SLs were categorized at each generation as 'fixed for reference allele', 'fixed for alternate allele', or 'not fixed'. The observed number of SLs in each category was tabulated for the S4 generation. The expected number of SLs in each category was calculated following the expectation of a 50% reduction in heterozygosity in each generation, resulting in an expectation of 6.25% of the SLs being heterozygous, 46.875% being homozygous for the reference allele, and 46.875% being homozygous for the alternate allele in the S4 generation. A χ^2 test was performed for SNPs with genotyping data present in at least 3 SLs to test for significant differentiation from this expectation, and a Benjamini-Hochberg false discovery rate of 0.05 was used to correct for multiple hypothesis testing across all SNPs. Variant effects were predicted for significant SNPs, and a Fisher's Exact Test was used to determine the presence of over or under-representation of significant SNPs and over-representation of GO categories in the significant SNPs.

Data access

The genome sequence reads, assembly and annotation, and transcriptomes used in annotation generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA704616.

The SNP data generated in this study has been submitted to the Zenodo data repository under DOI 10.5281/zenodo.6562381, and is available in the **Supplemental Datasets**.

Supplemental Code and raw data used for generating data files and figures, including all filtered SNP sets for each step of the study, are available as **Supplemental Code** files, and have been uploaded together with copies of the genome annotation and **Supplemental Dataset** files to the following GitHub repository: https://github.com/tshalev/WRC-genome-paper. Summaries of **Supplemental Code** are available in the **Supplemental Information.**

Acknowledgements

The research was supported with funds from Genome Canada and Genome British Columbia (CEDaR GAPP [184CED] to JB, ADY, JHR; 281ANV to IB); the Natural Sciences and Engineering Research Council of Canada (to JB, TJS, IB, OG); and the Province of British Columbia Land-Based Investment Strategy Tree Improvement Program to the western redcedar breeding program (JHR, ADY, LvdM). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank Allyson Miscampbell for SL DNA extraction, Angela Chiang for total RNA extraction, Lori Handley at HudsonAlpha for genomic data quality control and the HudsonAlpha Genomic Services Laboratory for the 10x library preparation, Eugene Goltsman at the JGI for assistance with HipMer, Kermit Ritland and Michael Whitlock for their careful help with the analyses and review of the paper, Annette Fahrenkrog for population genomics data analysis support.

Author Contributions: TJS conceived the project, performed research, analyzed data, and wrote the manuscript. SS and SDJ contributed to data analysis and writing of the manuscript. OG, MMSY, RLW and LC contributed to data analysis. AS, LBB, CP, JJ, GH, JY, MY, JGu and JWB generated materials and data. LvdM contributed to study design and provided essential materials. LGN and JGr contributed to study design and generated materials and data. LHR contributed to interpretation of the results and writing of the manuscript. JS, IB, MK and ADY contributed to study design and interpretation of the results. CR contributed to study design, generated materials and data, contributed to interpretation of the results, and managed and coordinated the overall project. JHR conceived the project and provided essential materials. JB conceived the project, managed and coordinated the overall project, and wrote the manuscript. All authors reviewed and edited the manuscript.

Competing Interest Statement: The authors declare no competing interests.

References

Acosta JJ, Fahrenkrog AM, Neves LG, Resende MFR, Dervinis C, Davis JM, Holliday JA, Kirst M. 2019. Exome resequencing reveals evolutionary history, genomic diversity, and targets of selection in the conifers *Pinus taeda* and *Pinus elliottii*. *Genome Biol Evol* **11** doi: 10.1093/gbe/evz016.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol* **215**: 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Antos JA, Filipescu CN, Negrave RW. 2016. Ecology of western redcedar (*Thuja plicata*): Implications for management of a high-value multiple-use resource. *For Ecol Manage* **375**: 211–222. doi: 10.1016/j.foreco.2016.05.043.

Baker HG. 1955. Self-Compatibility and Establishment After "Long-Distance" Dispersal. *Evolution* **9**: 347–348. doi: 10.2307/2405656.

Barrett SCH, Eckert CG. 1990. Variation and Evolution of Mating Systems in Seed Plants. In *Biological Approaches and Evolutionary Trends in Plants*, pp. 229–254.

Barrett SCH, Richards AJ, Bayliss MW, Charlesworth D, Abbott RJ. 2003. Mating strategies in flowering plants: The outcrossing-selfing paradigm and beyond. *Philos Trans R Soc B Biol Sci* **358**: 991–1004. doi: 10.1098/rstb.2003.1301.

Bayer PE, Valliyodan B, Hu H, Marsh JI, Yuan Y, Vuong TD, Patil G, Song Q, Batley J, Varshney RK, et al. 2022. Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome* **15**: e20109. doi: 10.1002/tpg2.20109.

Bierne N, Tsitrone A, David P. 2000. An inbreeding model of associative overdominance during a population bottleneck. *Genetics* **155**: 1981–1990. doi: 10.1093/genetics/155.4.1981.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MM, Keeling CI, Brand D, Vandervalk BP, et al. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**: 1492–1497. doi: 10.1093/bioinformatics/btt178.

Bishir J, Namkoong G. 1987. Unsound seeds in conifers: estimation of numbers of lethal alleles and of magnitudes of effects associated with the maternal parent. *Silvae Genet* **36**: 180–185.

Brandvain Y, Wright SI. 2016. The Limits of Natural Selection in a Nonequilibrium World. *Trends Genet* **32** doi: 10.1016/j.tig.2016.01.004.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci U S A* **101**: 15255–15260. doi: 10.1073/pnas.0404231101.

Buckler IV ES, Thornsberry JM. 2002. Plant molecular diversity and applications to genomics. *Curr Opin Plant Biol* **5**: 107–111. doi: 10.1016/S1369-5266(02)00238-8.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10** doi: 10.1186/1471-2105-10-421.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4** doi: 10.1186/s13742-015-0047-8.

Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA. 2013. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. *Tree Genet Genomes* **9**: 1537–1544. doi: 10.1007/s11295-013-0657-1.

Chen J, Glémin S, Lascoux M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Mol Biol Evol* **34**: 1417–1428. doi: 10.1093/molbev/msx088.

Chen J, Källman T, Gyllenstrand N, Lascoux M. 2010. New insights on the speciation history and nucleotide diversity of three boreal spruce species and a Tertiary relict. *Heredity (Edinb)* **104**: 3–14.

doi: 10.1038/hdy.2009.88.

Chen J, Li L, Milesi P, Jansson G, Berlin M, Karlsson B, Aleksic J, Vendramin GG, Lascoux M. 2019. Genomic data provide new insights on the demographic history and the extent of recent material transfers in Norway spruce. *Evol Appl* **12**: 1539–1551. doi: 10.1111/eva.12801.

Cherry ML. 1995. Genetic variation in western red cedar (*Thuja plicata* Donn) seedlings. The University of British Columbia.

Christenhusz MJM, Reveal JL, Farjon A, Gardner MF, Mill RR, Chase MW. 2011. A new classification and linear sequence of extant gymnosperms. *Phytotaxa* **19**: 55–70. doi: 10.11646/phytotaxa.19.1.3.

Copes DL. 1981. Isoenzyme uniformity in western red cedar seedlings from Oregon and Washington. *Can J For Res* **11**: 451–453. doi: 10.1139/x81-060.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158. doi: 10.1093/bioinformatics/btr330.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J, Nilsson O, Ritland K, et al. 2014. Insights into conifer giga-genomes. *Plant Physiol* **166**: 1724–1732. doi: 10.1104/pp.114.248708.

De Miguel M, Bartholomé J, Ehrenmann F, Murat F, Moriguchi Y, Uchiyama K, Ueno S, Tsumura Y, Lagraulet H, De Maria N, et al. 2015. Evidence of intense chromosomal shuffling during conifer evolution. *Genome Biol Evol* **7**: 2799–2809. doi: 10.1093/gbe/evv185.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2014. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. *Mol Ecol Resour* **14**: 209–214. doi: 10.1111/1755-0998.12157.

Dray S, Dufour AB. 2007. The ade4 package: Implementing the duality diagram for ecologists. *J Stat*

Softw **22**: 1–20. doi: 10.18637/jss.v022.i04.

El-Kassaby YA. 1999. Phenotypic plasticity in western redcedar. *For Genet* **6**: 235–240. citeulike-article-id:7424246.

El-Kassaby YA, Russell J, Ritland K. 1994. Mixed mating in an experimental population of western red cedar, *Thuja plicata*. *J Hered* **85**: 227–231. doi: 10.1093/oxfordjournals.jhered.a111441.

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* **20** doi: 10.1186/s13059-019-1832-y.

Endelman JB. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* **4**: 250–255. doi: 10.3835/plantgenome2011.08.0024.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol* **14**: 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x.

Fahrenkrog AM, Neves LG, Resende MFR, Dervinis C, Davenport R, Barbazuk WB, Kirst M. 2017. Population genomics of the eastern cottonwood (*Populus deltoides*). *Ecol Evol* **7**: 9426–9440. doi: 10.1002/ece3.3466.

Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Mol Ecol Notes* **7**: 574–578. doi: 10.1111/j.1471-8286.2007.01758.x.

Farjon A. 2018. Conifers of the World. *Kew Bull* **73** doi: 10.1007/s12225-018-9738-5.

Fisher RA. 1941. Average Excess and Average Effect of a Gene Substitution. *Ann Eugen* **11**: 53–63. doi: 10.1111/j.1469-1809.1941.tb02272.x.

Garrison E. 2016. A C++ library for parsing and manipulating VCF files. *GitHub*.

https://github.com/vcflib/vcflib.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv Prepr* **1207.3907**. http://arxiv.org/abs/1207.3907 (Accessed April 9, 2021).

Georges M, Charlier C, Hayes B. 2019. Harnessing genomic information for livestock improvement. *Nat Rev Genet* **20**: 135–156. doi: 10.1038/s41576-018-0082-2.

Gilbert KJ, Whitlock MC. 2015. Evaluating methods for estimating local effective population size with and without migration. *Evolution* **69**: 2154–2166. doi: 10.1111/evo.12713.

Glaubitz JC, El-Kassaby YA, Carlson JE. 2000. Nuclear restriction fragment length polymorphism analysis of genetic diversity in western redcedar. *Can J For Res* **30**: 379–389. doi: 10.1007/s00216-012-6081-9.

Golding GB, Strobeck C. 1980. Linkage disequilibrium in a finite population that is partially selfing. *Genetics* **94**: 777–789. doi: 10.1093/genetics/94.3.777.

Goudet J. 2005. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol Ecol Notes* **5**: 184–186. doi: 10.1111/j.1471-8286.2004.00828.x.

Gray LK, Hamann A. 2013. Tracking suitable habitat for tree populations under climate change in western North America. *Clim Change* **117**: 289–303. doi: 10.1007/s10584-012-0548-8.

Grime J. 1977. Evidence for the Existence of Three Primary Strategies in Plants and Its Relevance to Ecological and Evolutionary Theory. *Am Nat* **111**: 1169–1194. doi: 10.1086/283244.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**: 5654–5666. doi: 10.1093/nar/gkg770.

Hedrick PW, Hellsten U, Grattapaglia D. 2016. Examining the cause of high inbreeding depression:

Analysis of whole-genome sequence data in 28 selfed progeny of *Eucalyptus grandis*. *New Phytol* **209**: 600–611. doi: 10.1111/nph.13639.

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**: 2095–2105. doi: 10.1534/genetics.106.065102.

Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol* **33**: 54–78. doi: 10.1016/0040-5809(88)90004-4.

Hizume M, Kondo T, Shibata F, Ishizuka R. 2001. Flow cytometric determination of genome size in the Taxodiaceae, Cupressaceae *sensu stricto* and Sciadopityaceae. *Cytologia (Tokyo)* **66**: 307–311. doi: 10.1508/cytologia.66.307.

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJM, et al. 2018. Tigmint: Correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* **19** doi: 10.1186/s12859-018-2425-6.

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. 2017. ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome Res* **27**: 768–777. doi: 10.1101/gr.214346.116.

Janes JK, Miller JM, Dupuis JR, Malenfant RM, Gorrell JC, Cullingham CI, Andrew RL. 2017. The K = 2 conundrum. *Mol Ecol* **26** doi: 10.1111/mec.14187.

Jarne P, Charlesworth D. 1993. The evolution of the selfing rate in functionally hermaphrodite plants and animals. *Annu Rev Ecol Syst* **24**: 441–466. doi: 10.1146/annurev.es.24.110193.002301.

Jombart T. 2008. Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405. doi: 10.1093/bioinformatics/btn129.

Jombart T, Ahmed I. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**: 3070–3071. doi: 10.1093/bioinformatics/btr521.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet* **11** doi: 10.1186/1471-2156-11-94.

Kalisz S, Vogler DW, Hanley KM. 2004. Context-dependent autonomous self-fertilization yields reproductive assurance and mixed mating. *Nature* **430**: 884–887. doi: 10.1038/nature02776.

Korunes KL, Samuk K. 2021. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour* **21** doi: 10.1111/1755-0998.13326.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **47**: D807–D811. doi: 10.1093/nar/gky1053.

Krutovsky KV, Neale DB. 2005. Nucleotide Diversity and Linkage Disequilibrium in Cold-Hardiness- and Wood Quality-Related Candidate Genes in Douglas Fir. *Genetics* **171**: 2029–2041. doi: 10.1534/genetics.105.044420.

Lande R, Schemske DW. 1985. The Evolution of Self-Fertilization and Inbreeding Depression in Plants. I. Genetic Models. *Evolution* **39**: 24–40. doi: 10.2307/2408514.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**: 231–239. doi: 10.1016/0888-7543(88)90007-9.

Li C, Liu M, Sun F, Zhao X, He M, Li T, Lu P, Xu Y. 2021. Genetic Divergence and Population Structure in Weedy and Cultivated Broomcorn Millets (*Panicum miliaceum* L.) Revealed by Specific-Locus Amplified Fragment Sequencing (SLAF-Seq). *Front Plant Sci* **12** doi: 10.3389/fpls.2021.688444.

Li JH, Xiang QP. 2005. Phylogeny and biogeography of *Thuja* L. (Cupressaceae), an eastern Asian and North American disjunct genus. *J Integr Plant Biol* **47**: 651–659. doi: 10.1111/j.1744-7909.2005.00087.x.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Sci Adv* **1** doi: 10.1126/sciadv.1501084.

Liu S, Zhang L, Sang Y, Lai Q, Zhang X, Jia C, Long Z, Wu J, Ma T, Mao K, et al. 2022. Demographic History and Natural Selection Shape Patterns of Deleterious Mutation Load and Barriers to Introgression across *Populus* Genome. *Mol Biol Evol* **39** doi: 10.1093/molbev/msac008.

Liu X, Fu Y-X. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol* **21**: 280. doi: 10.1186/s13059-020-02196-9.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14** doi: 10.1371/journal.pcbi.1005944.

Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, Morgante M. 2011. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genet Genomes* **7**: 1011–1023. doi: 10.1007/s11295-011-0391-5.

McKinney GJ, Waples RK, Seeb LW, Seeb JE. 2017. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol Ecol Resour* **17** doi: 10.1111/1755-0998.12613.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17** doi: 10.1186/s13059-016-0974-4.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829. doi: 11290733.

Mohamadi H, Khan H, Birol I. 2017. ntCard: A streaming algorithm for cardinality estimation in genomics data. *Bioinformatics* **33**: 1324–1330. doi: 10.1093/bioinformatics/btw832.

Mukrimin M, Kovalchuk A, Neves LG, Jaber EHA, Haapanen M, Kirst M, Asiegbu FO. 2018. Genome-wide exon-capture approach identifies genetic variants of Norway spruce genes associated with susceptibility to *Heterobasidion parviporum* infection. *Front Plant Sci* **9** doi: 10.3389/fpls.2018.00793.

Neale DB, McGuire PE, Wheeler NC, Stevens KA, Crepeau MW, Cardeno C, Zimin AV, Puiu D, Pertea GM, Sezen UU, et al. 2017. The Douglas-Fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. *G3 Genes, Genomes, Genet* **7**: 3157–3167. doi: 10.1534/g3.117.300078.

Neale DB, Zimin AV, Zaman S, Scott AD, Shrestha B, Workman RE, Puiu D, Allen BJ, Moore ZJ, Sekhwal MK, et al. 2022. Assembled and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary adaptive potential and investigating hexaploid origin. *G3 Genes, Genomes, Genet* **12** doi: 10.1093/G3JOURNAL/JKAB380.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**: 5269–5273. doi: 10.1073/pnas.76.10.5269.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584. doi: 10.1038/nature12211.

O’Connell LM, Ritland K, Thompson SL. 2008. Patterns of post-glacial colonization by western redcedar (*Thuja plicata*, Cupressaceae) as revealed by microsatellite markers. *Botany* **86**: 194–203. doi: 10.1139/B07-124.

O’Connell LM, Russell J, Ritland K. 2004. Fine-scale estimation of outcrossing in western redcedar with microsatellite assay of bulked DNA. *Heredity (Edinb)* **93**: 443–449. doi: 10.1038/sj.hdy.6800521.

O'Connell LM, Viard F, Russell J, Ritland K. 2001. The mating system in natural populations of western redcedar (*Thuja plicata*). *Can J Bot* **79**: 753–756. doi: 10.1139/cjb-79-6-753.

Ohri D, Khoshoo TN. 1986. Genome size in gymnosperms. *Plant Syst Evol* **153**: 119–132. doi: 10.1007/BF00989421.

Pavy N, Namroud MC, Gagnon F, Isabel N, Bousquet J. 2012. The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity (Edinb)* **108**: 273–284. doi: 10.1038/hdy.2011.72.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959. doi: 10.1093/genetics/155.2.945.

Prunier J, Verta JP, Mackay JJ. 2016. Conifer genomics and adaptation: At the crossroads of genetic diversity and genome function. *New Phytol* **209**: 44–62. doi: 10.1111/nph.13565.

Puechmaille SJ. 2016. The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol Ecol Resour* **16**: 608–627. doi: 10.1111/1755-0998.12512.

Pyhäjärvi T, Kujala ST, Savolainen O. 2011. Revisiting protein heterozygosity in plants-nucleotide diversity in allozyme coding genes of conifer *Pinus sylvestris*. *Tree Genet Genomes* **7**: 385–397. doi: 10.1007/s11295-010-0340-8.

R Core Team. 2021. R: A language and environment for statistical computing.

Raj A, Stephens M, Pritchard JK. 2014. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589. doi: 10.1534/genetics.114.164350.

Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, Veerkamp RF. 2018. Utility of whole-genome sequence data for across-breed genomic prediction. *Genet Sel Evol* **50** doi: 10.1186/s12711-018-0396-8.

Remington DL, O'Malley DM. 2000. Whole-genome characterization of embryonic stage inbreeding depression in a selfed loblolly pine family. *Genetics* **155**: 337–348. doi: 10.1093/genetics/155.1.337.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler IV ES. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A* **98**: 11479–11484. doi: 10.1073/pnas.201394398.

Ritland K, Miscampbell A, van Niejenhuis A, Brown P, Russell J. 2020. Selfing and correlated paternity in relation to pollen management in western red cedar seed orchards. *Botany* **98**: 185–200. doi: 10.1139/cjb-2019-0123.

Roessler K, Muyle A, Diez CM, Gaut GRJ, Bousios A, Stitzer MC, Seymour DK, Doebley JF, Liu Q, Gaut BS. 2019. The genome-wide dynamics of purging during selfing in maize. *Nat Plants* **5**: 980–990. doi: 10.1038/s41477-019-0508-7.

Russell JH, Ferguson DC. 2008. Preliminary results from five generations of a western redcedar (*Thuja plicata*) selection study with self-mating. *Tree Genet Genomes* **4**: 509–518. doi: 10.1007/s11295-007-0127-8.

Russell JH, Kope HH, Ades P, Collinson H. 2007. Variation in cedar leaf blight (*Didymascella thujina*) resistance of western redcedar (*Thuja plicata*). *Can J For Res* **37** doi: 10.1139/X07-034.

Salamov AA, Solovyev VV. 2000. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* **10**: 516–522. doi: 10.1101/gr.10.4.516.

Scott AD, Zimin AV, Puiu D, Workman R, Britton M, Zaman S, Caballero M, Read AC, Bogdanove AJ, Burns E, et al. 2020. A reference genome sequence for giant sequoia. *G3 Genes, Genomes, Genet* **10**: 3907–3919. doi: 10.1534/g3.120.401612.

Shalev TJ, Yuen MMS, Gesell A, Yuen A, Russell JH, Bohlmann J. 2018. An annotated transcriptome of highly inbred *Thuja plicata* (Cupressaceae) and its utility for gene discovery of terpenoid biosynthesis and conifer defense. *Tree Genet Genomes* **14** doi: 10.1007/s11295-018-1248-y.

Shengqiang S, Goodstein D, Rokhsar D. 2013. PERTRAN: Genome-guided RNA-seq Read Assembler. In *Cold Spring Harbor Lab Genome Informatics*, New York, NY.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212. doi: 10.1093/bioinformatics/btv351.

Slate J, David P, Dodds KG, Veenvliet BA, Glass BC, Broad TE, McEwan JC. 2004. Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: Theoretical expectations and empirical data. *Heredity (Edinb)* **93**: 255–265. doi: 10.1038/sj.hdy.6800485.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6** doi: 10.1186/1471-2105-6-31.

Slatkin M. 2008. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**: 477–485. doi: 10.1038/nrg2361.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org*. http://www.repeatmasker.org.

Sorensen FC. 1982. The Roles of Polyembryony and Embryo Viability in the Genetic System of Conifers. *Evolution* **36**: 725–733. doi: 10.2307/2407885.

Stebbins GL. 1957. Self Fertilization and Population Variability in the Higher Plants. *Am Nat* **91**: 337–354. doi: 10.1086/281999.

Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, Cardeno C, Paul R, Gonzalez-Ibeas D, Koriabine M, Holtz-Morris AE, et al. 2016. Sequence of the sugar pine megagenome. *Genetics* **204**: 1613–1626. doi: 10.1534/genetics.116.193227.

Stewart WN. 1983. *Paleobotany and the evolution of plants*. 2nd ed. Cambridge University Press, Cambridge, UK.

Sved JA. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* **2**: 125–141. doi: 10.1016/0040-5809(71)90011-6.

Telfer E, Graham N, Macdonald L, Li Y, Klápště J, Resende M, Neves LG, Dungey H, Wilcox P. 2019. A high-density exome capture genotype-by-sequencing panel for forestry breeding in *Pinus radiata*. *PLoS One* **14** doi: 10.1371/journal.pone.0222640.

Vidalis A, Scofield DG, Neves LG, Bernhardsson C, García-Gil MR, Ingvarsson PK. 2018. Design and evaluation of a large sequence-capture probe set and associated SNPs for diploid and haploid samples of Norway spruce (*Picea abies*). *bioRxiv* doi: 10.1101/291716.

Vogler DW, Kalisz S. 2001. Sex among the flowers: The distribution of plant mating systems. *Evolution* **55**: 202–204. doi: 10.1111/j.0014-3820.2001.tb01285.x.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**: 2202–2204. doi: 10.1093/bioinformatics/btx153.

Wang T, Russell JH. 2006. Evaluation of selfing effects on western redcedar growth and yield in operational plantations using the tree and stand simulator (TASS). *For Sci* **52**: 281–289. doi: 10.5849/forsci.15-042.

Wang X, Bernhardsson C, Ingvarsson PK. 2020. Demography and Natural Selection Have Shaped Genetic Variation in the Widely Distributed Conifer Norway Spruce (*Picea abies*). *Genome Biol Evol* **12**: 3803–3817. doi: 10.1093/gbe/evaa005.

Waples RS, Do C. 2008. LDNE: A program for estimating effective population size from data on linkage disequilibrium. *Mol Ecol Resour* **8**: 753–756. doi: 10.1111/j.1755-0998.2007.02061.x.

Warren RL, Keeling CI, Yuen MM Saint, Raymond A, Taylor GA, Vandervalk BP, Mohamadi H, Paulino D, Chiu R, Jackman SD, et al. 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J* **83**: 189–212. doi: 10.1111/tpj.12886.

Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* **35**: 543–548. doi: 10.1093/molbev/msx319.

Williams CG. 2008. Selfed embryo death in *Pinus taeda*: A phenotypic profile. *New Phytol* **178**: 210–222. doi: 10.1111/j.1469-8137.2007.02359.x.

Williams CG, Auckland LD, Reynolds MM, Leach KA. 2003. Overdominant lethals as part of the conifer embryo lethal system. *Heredity (Edinb)* **91**: 584–592. doi: 10.1038/sj.hdy.6800354.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding Regions of *Capsella grandiflora*. *PLoS Genet* **10** doi: 10.1371/journal.pgen.1004622.

Wright S. 1922. Coefficients of Inbreeding and Relationship. *Am Nat* **56**: 330–338. doi: 10.1086/279872.

Wright S. 1931. Evolution in Mendelian Populations. *Genetics* **16**: 97–159. doi: 10.1007/BF02459575.

Wright SI, Kalisz S, Slotte T. 2013. Evolutionary consequences of self-fertilization in plants. *Proc R Soc B Biol Sci* **280** doi: 10.1098/rspb.2013.0133.

Xin Z, Chen J. 2012. A high throughput DNA extraction method with high yield and quality. *Plant Methods* **8** doi: 10.1186/1746-4811-8-26.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**: 76–82. doi: 10.1016/j.ajhg.2010.11.011.

Yeo S, Coombe L, Warren RL, Chu J, Birol I. 2018. ARCS: Scaffolding genome drafts with linked reads. *Bioinformatics* **34**: 725–731. doi: 10.1093/bioinformatics/btx675.

Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, Bentley DR, Morton NE. 2004. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc Natl Acad Sci U S A* **101**: 18075–18080. doi: 10.1073/pnas.0408251102.

Zhang YY, Fischer M, Colot V, Bossdorf O. 2013. Epigenetic variation creates potential for evolution of plant phenotypic plasticity. *New Phytol* **197**: 314–322. doi: 10.1111/nph.12010.

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, et al. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* **196**: 875–890. doi: 10.1534/genetics.113.159715.

Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, Langley CH, Neale DB, Salzberg SL. 2017. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience* **6** doi: 10.1093/gigascience/giw016.

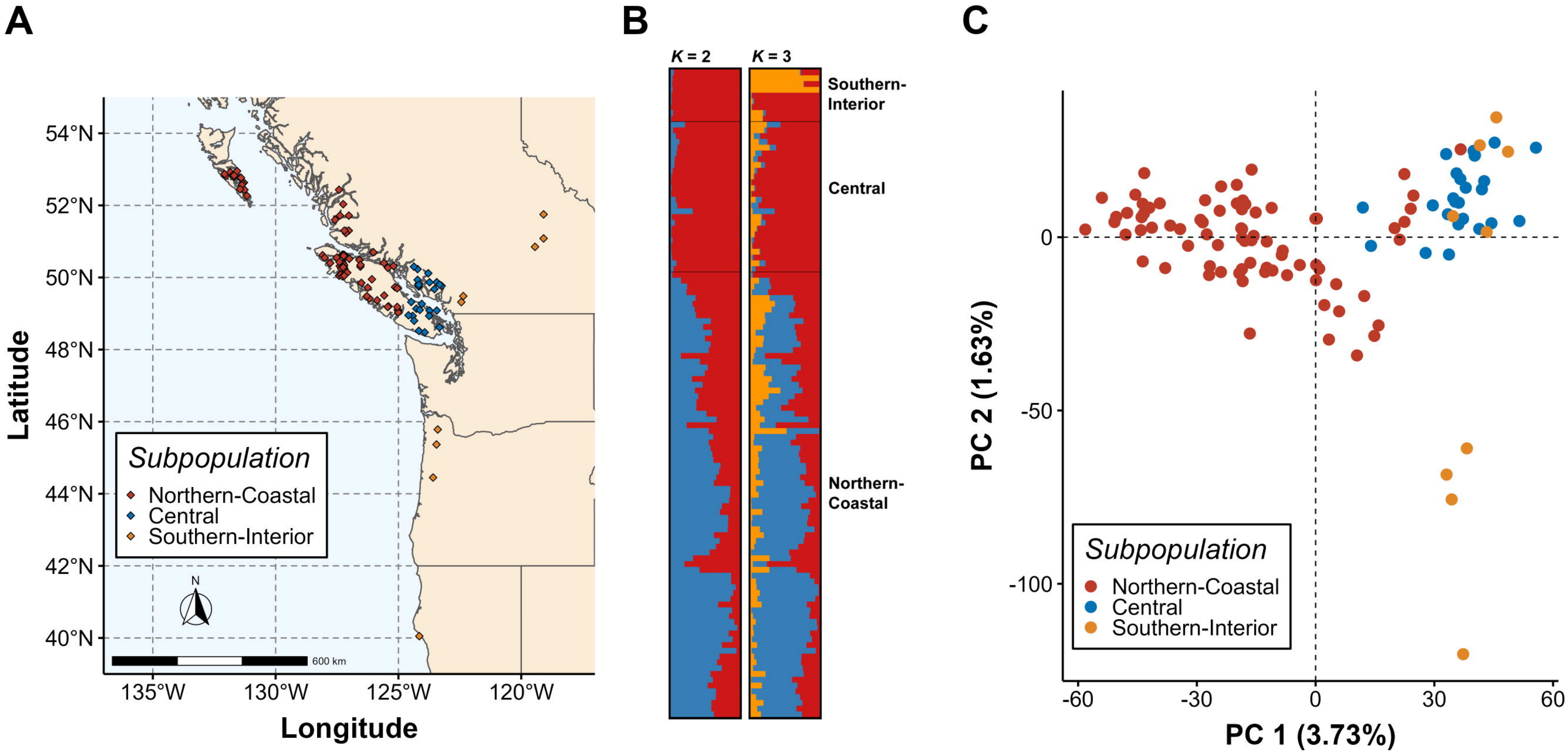
Figure Headings

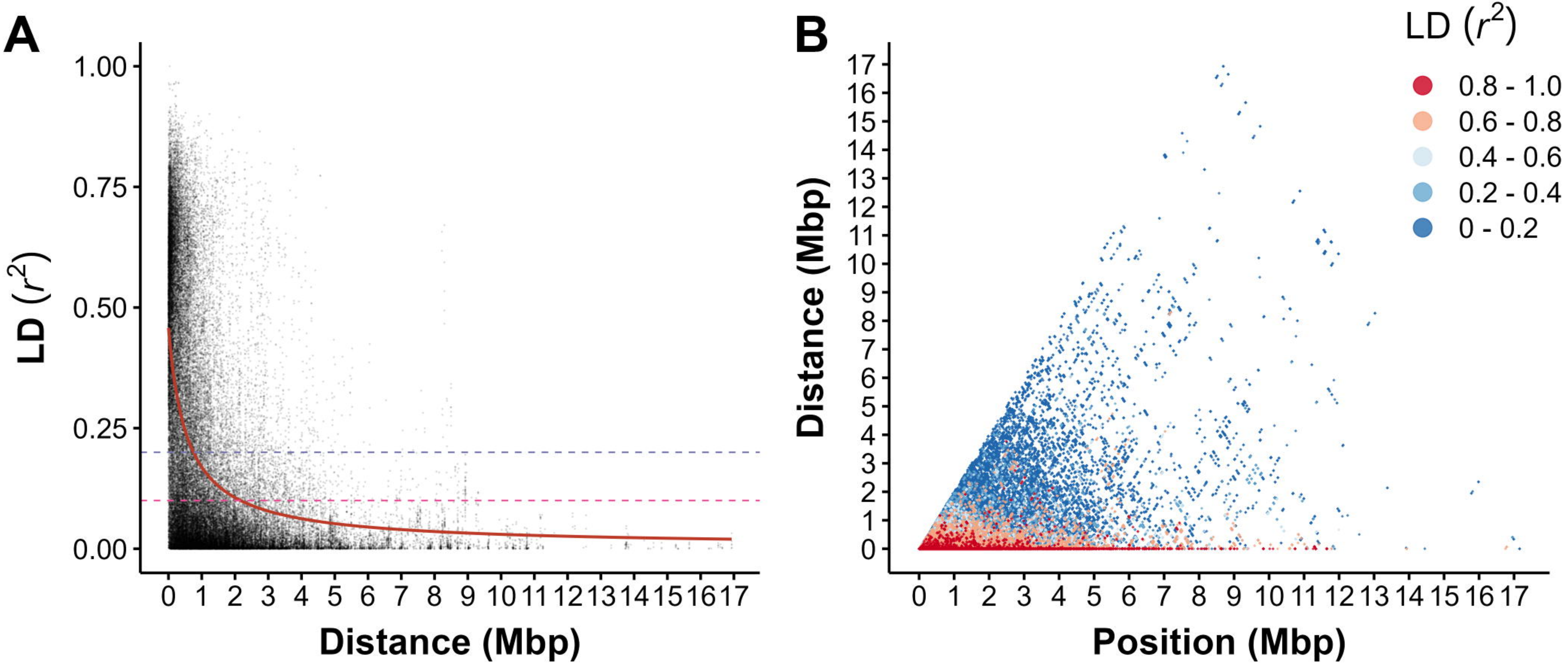
Figure 1: Genetic structure is weak across the geographic range of WRC. A) Map of geographic origin for trees in the range-wide population (RWP) (*n* = 112). Subpopulations were defined *a priori* based on analysis outcomes of O'Connell et al. (2008). Trees were separated into three main subpopulations: Northern-Coastal (*n* = 77); Central (*n* = 26); and Southern-Interior (*n* = 9). **B) STRUCTURE plot of the RWP for K = 2 and K = 3**. Optimal K was determined by evaluating STRUCTURE results using the methods of Evanno et al. (2005) and Puechmaille (2016), and by the approach of fastStructure (Raj et al. 2014). Gene flow is present throughout all three subpopulations. **C) Principal component analysis (PCA) of genetic distance between trees in the RWP.** Latitudinal separation of trees from different s can be observed, although each principal component only explains a very small proportion of the variation between individuals, indicating that genetic differentiation is low.

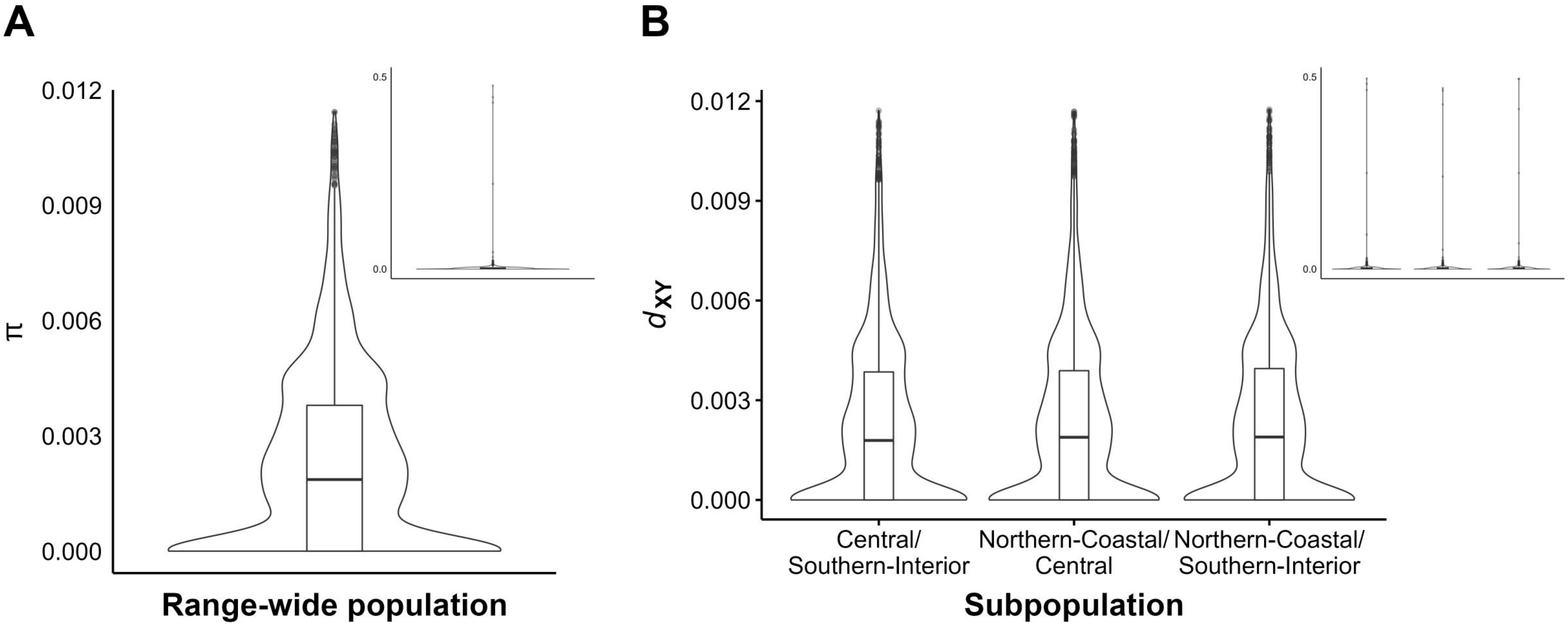
Figure 2: Within-scaffold linkage disequilibrium (LD) in the range-wide population. A) LD was assessed using SNPs with a minor allele frequency cutoff of 0.05 to reduce error associated with rare alleles (*n* = 16,202 SNPs). Decay was estimated using a non-linear model (red line); r^2 decayed to baseline thresholds of 0.2 (purple dotted line) and 0.1 (pink dotted line) at 0.751 and 2.17 Mbp, respectively. **B)** Pairwise LD for all pairs of SNPs (*n* = 16,202). Each point on the plot represents the LD between two SNPs at a given distance from one another and relative position on the scaffold. Colour indicates LD range, with red indicating SNPs in strong LD.

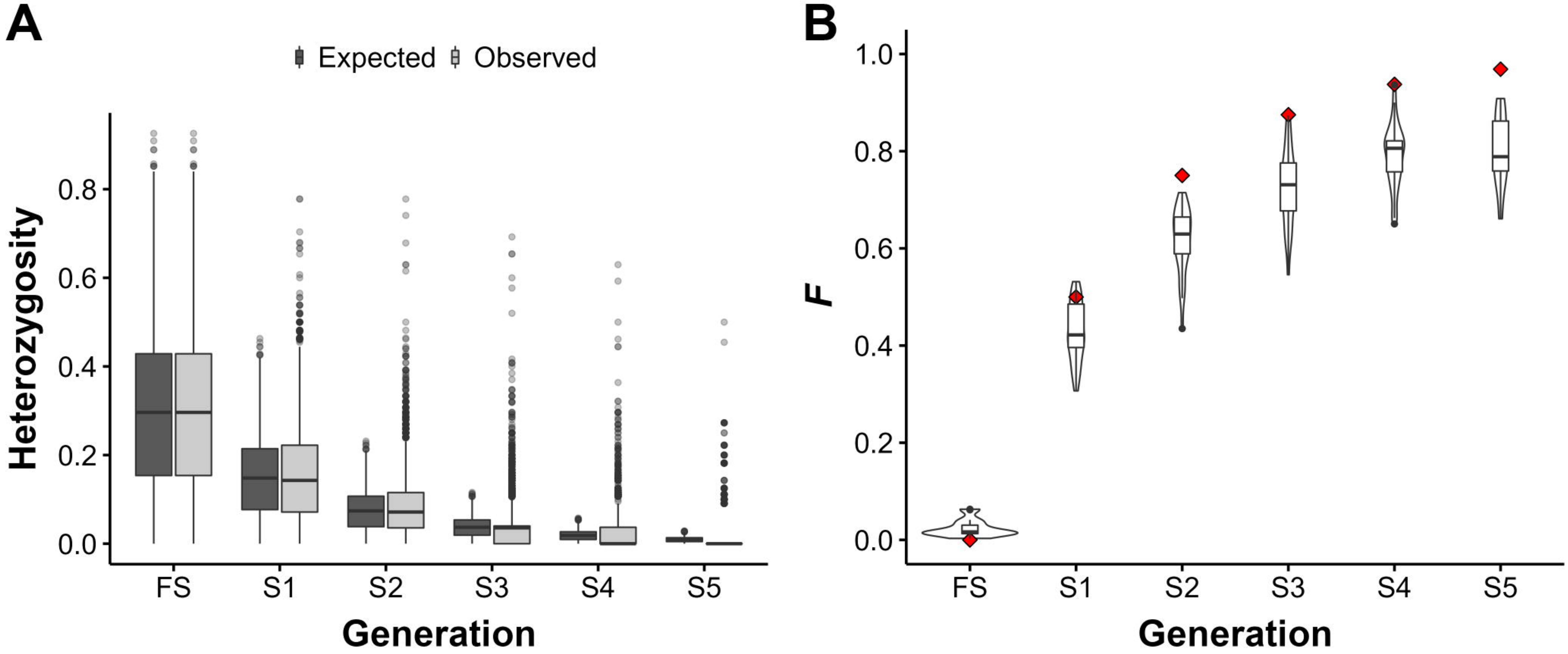
Figure 3: A) Overall distribution of average π of the range-wide population (RWP) in SCOs. We detected 1,411 SCOs with a π of zero, with an average π of 0.00272. **B) Overall distribution of average d_{XY} between each pair of geographic subpopulations.** No significant differences were observed between comparisons of different subpopulations. **Inlays** show all π estimates; **main plots** show π estimates with outliers in the top one percentile removed for clarity. The top 1% of π estimates account for 3% of the total estimated diversity, and the top 5% account for over 13% of the total.

Figure 4: Change in heterozygosity (H) and inbreeding coefficients (*F*) over five successive generations of complete selfing in WRC. A) Observed vs. expected change in heterozygosity over five successive generations of complete selfing in *n* = 28 (FS – S4) and *n* = 11 (S5) different selfing lines (SLs), at *n* = 18,371 SNP loci, after manual error correction. Each line at each generation is represented by a single tree. Black points indicate boxplot outliers. Observed median heterozygosity declines faster than expected under theoretical expectations during complete selfing, despite many SNP loci remaining heterozygous across all generations. **B)** Inbreeding coefficients (*F*) for *n* = 28 samples (FS – S4) and *n* = 11 samples (S5). Black points indicate boxplot outliers. Under complete selfing, *F* is expected to increase by a factor of ½(1+*F*) in the previous generation (red diamonds). *F* increases at a slower rate than expected in our SLs.


A
B
C
Latitude
54°N
52°N
50°N
48°N
46°N
44°N
42°N
40°N
135°W
130°W
125°W
120°W
Longitude
Subpopulation
Northern-Coastal
Central
Southern-Interior
N
600 km
K = 2
K = 3
Southern-Interior
Central
Northern-Coastal
PC 2 (1.63%)
0
-50
-100
-60
-30
0
30
60
PC 1 (3.73%)
Subpopulation
Northern-Coastal
Central
Southern-Interior


A
LD (r^2)
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Distance (Mbp)
B
Distance (Mbp)
17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Position (Mbp)
LD (r^2)
● 0.8 - 1.0
● 0.6 - 0.8
● 0.4 - 0.6
● 0.2 - 0.4
● 0 - 0.2


A
π
0.012
0.009
0.006
0.003
0.000
0.5
0.0
Range-wide population
B
d_{XY}
0.012
0.009
0.006
0.003
0.000
0.5
0.0
Central/
Southern-Interior
Northern-Coastal/
Central
Northern-Coastal/
Southern-Interior
Subpopulation


A
Expected
Observed
Heterozygosity
0.0
0.2
0.4
0.6
0.8
FS
S1
S2
S3
S4
S5
Generation
B
F
0.0
0.2
0.4
0.6
0.8
1.0
FS
S1
S2
S3
S4
S5
Generation

CSH PRESS
GENOME RESEARCH

The western redcedar genome reveals low genetic diversity in a self-compatible conifer

Tal J. Shalev, Omnia Gamal El-Dien, Macaire M.S. Yuen, et al.

Genome Res. published online September 15, 2022
Access the most recent version at doi:10.1101/gr.276358.121

Published by Cold Spring Harbor Laboratory Press